The Problem of Causality in the Analysis of Educational Choices and Labor Market Outcomes

Slides for Lectures *

Andrea Ichino
(European University Institute and CEPR)

February 28, 2006

Abstract

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.

*Address correspondence to: Andrea Ichino, Department of Economics, I.U.E., I-50016, San Domenico di Fiesole, Firenze, Italia, e-mail: ichino@iue.it. This document can be downloaded from: http://www.iue.it/Personal/Ichino/Welcome.html

Contents

1 The Problem of Causality

1.1 Motivation

Consider the following questions.

- **Medical therapies**: how can we establish whether a drug is effective?
- **Fitness**: does exercising improve health? For everybody? Only for the (healthy) persons who do it?
- **College education**: does college increase your future earnings? Does college add anything to your innate abilities? Can we estimate in general the returns to schooling? For which group in the population?
- **Labor market programs**: does training increase the employment probability of jobless workers? Do work incentives for single mothers with children increase labor force participation?
- **Army**: does the military service increase or reduce earning and employment probabilities? Health? Life expectancy?
- **Educational policies**: can the offer of students' loans increase college education and earnings of poor highschool graduates?

The main goal of this course is to attract your attention to the problem of causality highlighted by these examples. This problem is evidently crucial for economic analysis.

We analyze this problem by focusing mainly on applications concerning education and the labor market, but the issue is more general and should be interesting for all economists independently of their field.

The outline of the course is described in the Contents.

1.2 A formal framework to think about causality

We have a population of individuals; for each individual we observe a variable D and a variable Y.

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

In general no, and 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 individuals in the population;
- D_i is the *treatment*, the potential cause of which we want to estimate the effects:

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

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

- $Y_i(D_i)$ is the outcome, the effect that we want to attribute to the treatment; the notation indicates that it (may) depend on D_i

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

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

- Note that the outcome for each individual can be written as:

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

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

Examples ...

- **Medical therapy**: population = cancer patients; D = therapy given by the doctor to the patient; Y = life or death, tumoral mass.
  We can run controlled experiments, but ethical issues often restrict the range of feasible experiments.
- **College**: population = highschool graduates; D = attending a college; Y = earnings, time to find a job.
  Typically, to study the causal effect of education we do not have data from controlled experiments and we have to rely on observational data.
- **Labor market programs**: ...
- **Army**: ...
- **Student loans**: ...

1.3 The fundamental problem of causal inference

Within this formal framework we can define the causality link in the following way.

Definition 1 *For every individual i*

The event $\{D_i = 1$ *instead of* $D_i = 0\}$ *causes the effect* $\Delta_i = Y_i(1) - Y_i(0)$

Given this (reasonable) definition, we would like to:

- establish whether the above causality link exists for an individual i;
- measure the dimension of the effect of D_i on Y_i.

It seems impossible to reach these goals because of the following proposition:

Proposition 1 *Fundamental Problem of Causal Inference.*
It is impossible to observe for the same individual 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 individual i (Holland, 1986).*

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

There are two classes of solutions to the Fundamental Problem of Causal Inference:

i. *The scientific solution.*
   Exploits various *homogeneity* and *invariance* assumptions to construct experiments on the causality link.

ii. *The statistical solution.*
   Approaches the problem by aiming at the identification of "average causal effects".

Within each of these solutions, different hypotheses lead to different interpretations of the causality link.

In our research we often assume specific interpretations of the causality link without paying sufficient attention to the hypotheses that are required for the validity of these interpretations.

1.4 The scientific solution

Consider the following assumptions

i. There is one i: a physical device, (e.g. an electric circuit);

ii. *Temporal stability:* the value of Y_i does not depend on *when* the sequence "apply $D = 1$ to i and then measure Y_i" takes place;

iii. *Causal transience:* the value of Y_i is not affected by the previous exposure of i to the above experimental sequence

iv. *Unit Homogeneity:* there exist other units $j \neq i$ such that $Y_i(D_i) = Y_j(D_j)$ for $D_i = D_j$

These are the assumptions behind scientific inference, but sometime also behind inference in our daily life. Examples ...

What is their relevance for economic analysis?

1.5 The statistical solution

Given that the causal effect for a single individual i cannot be observed, the statistical solution proposes methods to compute the average causal effect for the entire population or for some interesting sub-groups.

1.5.1 The effect of treatment on a random individual

Suppose you pick a person at random in the population and you expose him/her to treatment. What is the expected effect on the outcome for this person?

Formally this is given by:

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

Apparently we are not making progress, because we cannot observe the outcome in both counterfactual situations for all the individuals and therefore we cannot compute the expectations on the right-hand side.

Furthermore, the effect of treatment on a random person may not be an interesting treatment effect from the viewpoint of an economist.

1.5.2 The effect of treatment on the treated

This second type of average effect is often more interesting for economists.

Let's consider only the sub-population of those who are actually treated. What is the average treatement effect for these persons?

It is the difference between the average outcome in case of treatment (which we observe) minus the average outcome in the counterfactual situation of no-treatment (which we do not observe). Formally:

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

Why should the effect of treatment on the treated be more interesting for economists than the effect of treatment on a random person?

However, the problem is that both these average treatment effects cannot be easily identified and estimated with observational data.

Randomized experiments offer a way to solve the problem.

1.5.3 Randomized experiments

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

$$E\{Y_i(0)|i \in C\} = E\{Y_i(0)|i \in T\} = E\{Y_i(0)\} \tag{4}$$

and

$$E\{Y_i(1)|i \in C\} = E\{Y_i(1)|i \in T\} = E\{Y_i(1)\}. \tag{5}$$

Then substituting 4 and 5 in 2 it is immediate to obtain:

$$\begin{aligned} E\{\Delta_i\} &\equiv E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= E\{Y_i(1)|i \in T\} - E\{Y_i(0)|i \in C\}. \end{aligned} \tag{6}$$

In this way we can solve the Fundamental Problem of Causal Inference because we use the sample C (the *controls*) as an image of what would happen to the sample T (the *treated*) 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 data. We will repeatedly look at his results during the course.

However, randomized experiments are rarely a feasible solution for economists:

- ethical concerns;
- technical implementation;

In the rest of the course we will study different conventional and non-conventional alternatives to randomized experiments.

2 Conventional methods to estimate causal effects

This part of the course is devoted to conventional methods.

The goal is to explore in a deeper way the econometric problems raised by the identification and estimation of treatment effects.

We will consider the problems raised by:

- OLS estimation;
- IV estimation;
- Heckman "two stages" estimation;

2.1 Specification of the outcomes

Going back to the notation of Section 1, consider the following specification of outcomes, with or without treatment:

$$\begin{aligned} Y_i(1) &= \mu(1) + U_i(1) \\ Y_i(0) &= \mu(0) + U_i(0) \end{aligned} \tag{7}$$

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

$$\begin{aligned} \Delta_i &= Y_i(1) - Y_i(0) \\ &= [\mu(1) - \mu(0] + [U_i(1) - U_i(0)] \\ &= E\{\Delta_i\} + [U_i(1) - U_i(0)]. \end{aligned} \tag{8}$$

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.

(Figure: Differences between treated and control individuals.)

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

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

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.

2.2 Specification of the selection into treatment

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

$$D_i^* = \alpha + \beta Z_i + V_i \tag{10}$$

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

$$D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{cases} \tag{11}$$

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 (linearly) 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.

2.3 The model in compact form

$$Y_i = \mu(0) + \Delta_i D_i + U_i(0) \tag{12}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{13}$$

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{array} \right\} \tag{14}$$

$$\begin{aligned} \Delta_i &= \mu(1) - \mu(0) + U_i(1) - U_i(0) \\ &= E\{\Delta_i\} + U_i(1) - U_i(0) \end{aligned} \tag{15}$$

$$E\{U_i(1)\} = E\{U_i(0)\} = E\{V_i\} = 0 \tag{16}$$

Correlation between U_i and V_i is possible.

Examples:

- Cancer
- Education
- Training
- ...

We will first define the statistical effects of treatment in this model, and then we will discuss the identification and estimation problems.

2.4 The statistical effects of treatment in this model

Within this model the statistical effects of treatment considered by the conventional analysis are given by the following equations:

i. *The effect of treatment on a random individual.*

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1) - Y_i(0)\} \\ &= E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= \mu(1) - \mu(0) \end{aligned} \tag{17}$$

ii. *The effect of treatment on the treated*

$$\begin{aligned} E\{\Delta_i \mid D_i = 1\} &= E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &= \mu(1) - \mu(0) + E\{U_i(1) - U_i(0) \mid D_i = 1\} \end{aligned} \tag{18}$$

The two effects differ because of the term

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} \tag{19}$$

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

When these two treatment effects are equal?

i. When the idiosyncratic gain is zero for every individual:

$$U_i(1) = U_i(0) \qquad \forall i \tag{20}$$

In this case, the model has constant coefficients because

$$\Delta_i = E\{\Delta_i\} = \mu(1) - \mu(0) \qquad \forall i. \tag{21}$$

Therefore, we are assuming that the effect of treatment is identical for all individuals. And in particular for both a treated and a random person.

ii. When the average idiosyncratic gain for the treated is equal to zero:

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} = E\{U_i(1) - U_i(0)\} = 0 \tag{22}$$

In this case treatment is random and in particular is independent of the idiosyncratic gain. Therefore the average idiosyncratic gain for the treated is equal to the average idiosyncratic gain in the population that is equal to zero.

Examples:

- Cancer
- Education
- Training
- ...

2.5 Problems with OLS estimation

2.5.1 Bias for the effect of treatment on a random person

Using 15 we can rewrite equation 12 as:

$$\begin{aligned} Y_i &= \mu(0) + E\{\Delta_i\}D_i + U_i(0) + D_i[U_i(1) - U_i(0)] \\ &= \mu(0) + E\{\Delta_i\}D_i + \epsilon_i \end{aligned} \tag{23}$$

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

Problem:

$$E\{\epsilon_i D_i\} = E\{U_i(1) \mid D_i = 1\}Pr\{D_i = 1\} \neq 0 \tag{24}$$

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

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i\} + \tag{25}$$

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

The second line in 25 represents the OLS regression bias if we want to estimate the effect of treatment on a random person.

Readjusting the second line of 25, the bias in the estimation of $E\{\Delta_i\}$ can be written in the following form:

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i\} + \tag{26}$$
$$E\{U_i(1) \mid D_i = 1\} - E\{U_i(0) \mid D_i = 0\}$$

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:

$$\begin{aligned} E\{U_i(1)\} &= E\{U_i(1) \mid D_i = 1\} = 0 \\ E\{U_i(0)\} &= E\{U_i(0) \mid D_i = 0\} = 0 \end{aligned} \tag{27}$$

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

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ &= \mu(1) - \mu(0) \end{aligned} \tag{28}$$

Examples:

- Cancer

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

2.5.2 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 23 and remembering from 18 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 23 as:

$$\begin{aligned} Y_i &= \mu(0) + E\{\Delta_i \mid D_i = 1\}D_i + \\ &\quad U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D_i = 1\}] \\ &= \mu(0) + E\{\Delta_i \mid D_i = 1\}D_i + \eta_i \end{aligned} \tag{29}$$

Using 29 we can define the OLS bias in the estimation of $E\{\Delta_i \mid D_i = 1\}$. Note that this parameter is equal to the common effect *plus the average idiosyncratic gain.*

However, also in this case the error term is correlated with the treatment indicator D_i:

$$\begin{aligned} E\{\eta_i D_i\} &= E\{D_i U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D_i = 1\}]\} \\ &= E\{D_i U_i(0)\} \neq 0. \end{aligned} \tag{30}$$

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

$$\begin{aligned} E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i \mid D_i = 1\} + \\ E\{U_i(0) \mid D_i = 1\} - E\{U_i(0) \mid D_i = 0\} \end{aligned} \tag{31}$$

The second line in 31 represents the OLS regression bias if we want to estimate the effect of treatment on the treated.

The bias

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

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 effect of treatment on the treated?

Examples:

- Cancer
- Education
- Training
- ...

2.5.3 An important particular case: the Roy model

Consider the case in which the idiosyncratic gain from treatment exists and is one of the determinants of the participation into treatment, so that:

$$\begin{aligned} Pr\{D_i = 1 \mid U_i(1) - U_i(0)\} &\neq Pr\{D_i = 1\} \qquad \text{or equiv.} \\ E\{D_i \mid U_i(1) - U_i(0)\} &\neq E\{D_i\} \end{aligned} \tag{32}$$

In this case by Bayes Law, denoting with f the density of $U_i(1) - U_i(0)$ we have that

$$\begin{aligned} f(U_i(1) - U_i(0) \mid D_i = 1) Pr\{D_i = 1\} &= \\ Pr\{D_i = 1 \mid U_i(1) - U_i(0)\} f(U_i(1) - U_i(0)) \end{aligned} \tag{33}$$

Because of 32, from 33 descends that

$$f(U_i(1) - U_i(0) \mid D_i = 1) \neq f(U_i(1) - U_i(0)) \tag{34}$$

and therefore that

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} \neq E\{(U_i(1) - U_i(0)\} \tag{35}$$

This equation implies that in this case:

- the effect of treatment on a random person is different from the effect of treatment on the treated (see equation 22);
- OLS give seriously biased estimates of the effect on a random person (see equation 25);
- OLS appear to be more promising for the estimation of the effect of treatment on the treated, but the problem of the *mean selection bias* remains to be solved (see equation 31).

2.6 Conventional interpretation of Instrumental Variables

2.6.1 Assumptions for the IV estimation of the effect of treatment on a random person

We want to estimate equation 23, which is reported here for conveniece

$$Y_i = \mu(0) + E\{\Delta_i\}D_i + \epsilon_i.$$

Suppose that there exist a variable Z such that:

$$COV\{Z, D\} \neq 0 \tag{36}$$

$$COV\{Z, \epsilon\} = 0. \tag{37}$$

If this variable exists then (see the Appendix 6.1):

$$E\{\Delta_i\} = \frac{COV\{Y, Z\}}{COV\{D, Z\}}. \tag{38}$$

Substituting the appropriate sample covariances on the LHS of 38 we get a consistent estimate of $E\{\Delta_i\}$.

It is however crucial to understand what the two conditions 36 and 37 require in terms of our model.

The first condition that the instrument Z has to satisfy is:

$$Pr\{D_i = 1 \mid Z_i = 1\} \neq Pr\{D_i = 1 \mid Z_i = 0\} \tag{39}$$

This condition can be easily tested by estimating the participation equation 13 and checking that Z_i is a significant predictor of D_i.

Note that to do so we do not have to make functional assumptions on the error term V_i in the participation equation 13 (in contrast with the Heckman two step procedure that we will consider later).

The second condition is more problematic:

$$E\{\epsilon_i \mid Z_i\} = E\{U_i(0) + D_i[U_i(1) - U_i(0)] \mid Z_i\} = 0 \tag{40}$$

This (just-identifying) condition *cannot be tested*.

Note that it contains two requirements:

i. The instrument must be uncorrelated with the unobservable outcome in the base state; i.e. knowing the value of the instrument should not help to predict the outcome in the base state.

$$E\{U_i(0) \mid Z_i\} = 0 = E\{U_i(0)\} \tag{41}$$

ii. Conditioning on the instrument, the idiosyncratic gain must be uncorrelated with the treatment

$$\begin{aligned} E\{D_i[U_i(1) - U_i(0)] \mid Z_i\} &= E\{U_i(1) - U_i(0) \mid Z_i, D_i = 1\} Pr\{D_i = 1 \mid Z_i\} \\ &= 0 = E\{U_i(1) - U_i(0)\} \end{aligned} \tag{42}$$

For example, in the case of the Vietnam war lottery for the earning effect of the military service (Angrist, 1990), this condition requires that:

- the average gain of those who are not drafted and go and the average gain of those who are drafted and go must both be equal to the average gain of the entire population, which is equal to 0.

Other examples:

- Parental background for returns to schooling (Willis-Rosen, 1979).
- Quarter of birth for returns to schooling (Angrist and Krueger, 1994).
- Nearby college for returns to schooling (Card, 1995b)
- WWII for returns to schooling (Ichino and Winter-Ebmer, 1998)
- A random indicator of assignment to treatment.

It seems that if we really want to estimate the effect on a random person and there exists relevant idiosyncratic gains, we better go for randomization in a controlled experiment.

2.6.2 Assumptions for the IV estimation of the effect of treatment on a treated person

We want now to estimate equation 29, which is reported here for convenience

$$Y_i = \mu(0) + E\{\Delta_i \mid D_i = 1\}D_i + \eta_i.$$

We assume again that there exist a variable Z such that the two conditions 36 and 37 hold in this case:

$$COV\{Z, D\} \neq 0 \tag{43}$$

$$COV\{Z, \eta\} = 0. \tag{44}$$

If this variable exists then (see the Appendix 6.1):

$$E\{\Delta_i \mid D_i = 1\} = \frac{COV\{Y, Z\}}{COV\{D, Z\}}. \tag{45}$$

Substituting the appropriate sample covariances on the LHS of 45 we get a consistent estimate of $E\{\Delta_i \mid D_i = 1\}$.

Also in this case it is crucial to understand what the two conditions 36 and 37 require in terms of our model.

The first condition that the instrument Z has to satisfy is equal to the one that was needed for the IV estimation of the effect on a random person:

$$E\{D_i \mid Z_i\} = Pr\{D_i = 1 \mid Z_i\} \neq 0 \tag{46}$$

This condition can be easily tested by estimating the participation equation 13 and checking that Z_i is a significant predictor of D_i.

Note again that to do so we do not have to make functional assumptions on the error term V_i in the participation equation 13 (in contrast with Heckman procedure that we will consider later).

The second condition is different but still problematic:

$$E\{\eta \mid Z\} = E\{U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D_i = 1\}] \mid Z_i\} = 0 \tag{47}$$

There are again two requirements:

i. The instrument must be uncorrelated with the unobservable outcome in the base state; i.e. knowing the value of the instrument should not help predicting the outcome in the base state (like in the previous case).

$$E\{U_i(0) \mid Z_i\} = 0 = E\{U_i(0)\} \tag{48}$$

ii. The average idiosyncratic gain for the treated conditioning on the instrument, should be identical to the unconditional average idiosyncratic gain for the treated

$$E\{U_i(1) - U_i(0) \mid Z_i, D_i = 1\} = E\{U_i(1) - U_i(0) \mid D_i = 1\} \tag{49}$$

Using again the example of the Vietnam war lottery for the earning effect of the military service (Angrist, 1990), this condition requires that:

- the average gain of those who are not drafted and go and the average gain of those who are drafted and go must both be equal to the average gain of all those who go (i.e. the average gain of those who go is independent of the draft).

Keep in mind this condition because it will be crucial in the comparison between the Heckman (1997) interpretation of IV an the AIR interpretation of IV.

Other examples:

- Parental background for returns to schooling (Willis-Rosen, 1979).
- Quarter of birth for returns to schooling (Angrist and Krueger, 1994).
- Nearby college for returns to schooling (Card, 1995b)
- WWII for returns to schooling (Ichino and Winter-Ebmer, 1998)
- A random indicator of assignment to treatment.

2.6.3 Comments

Even if we are interested only in the effect of treatment on the treated and not in the effect of treatment on a random person, the IV estimation seems problematic.

Note that randomization does not solve the problem in the presence of non-compliance with the assignment.

Furthermore, it seems possible that using IV the estimated effect of treatment on the treated differs at different values of the instrument or for different instruments, in which case condition 49 would not be satisfied.

This intuition leads to the concept of Local Average Treatment Effect estimation on which we will focus later.

But first we look at another conventional approach to the estimation of treatment effects which applies to models with fixed coefficients.

2.7 Heckman procedure for endogenous dummy variable models

2.7.1 The basic model

Consider the case in which $U_i(1) = U_i(0)$ (no idiosyncratic gain from treatment) and let $\Delta = \mu(1) - \mu(0)$. Allow for the explicit consideration of covariates X_i Our model (see equation 12) simplifies to the following common coefficients model:

$$\begin{aligned} Y_i &= \mu(0) + \gamma X_i + \Delta D_i + U_i(0) \\ Y_i &= \mu + \gamma X_i + \Delta D_i + U_i \end{aligned} \tag{50}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{51}$$

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{array} \right\} \tag{52}$$

where $E\{U_i\} = E\{V_i\} = 0$ but $\text{COV}\{U_i, V_i\} \neq 0$ so that $E\{D_i U_i\} \neq 0$ and the OLS estimation of 50 is inconsistent. We will later make functional assumptions on these error terms.

This model is commonly called the *endogenous dummy variable* model (see Heckman (1978) and Maddala (1983). The OLS bias comes, for example, from the fact that 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 on unions on wages (Robinson, 1989)
- Wage equation for female workers (Heckman, 1978)
- ...

2.7.2 The model rewritten as a switching regression model

We can rewrite the model in the following way:

$$\text{Regime 1: if } D_i^* \geq 0 \qquad Y_i = \mu + \gamma X_i + \Delta + U_i \tag{53}$$

$$\text{Regime 0: if } D_i^* < 0 \qquad Y_i = \mu + \gamma X_i + U_i \tag{54}$$

or equivalently

$$\text{Regime 1: if } V_i \geq -\alpha - \beta Z_i \qquad Y_i = \mu + \gamma X_i + \Delta + U_i \tag{55}$$

$$\text{Regime 0: if } V_i < -\alpha - \beta Z_i \qquad Y_i = \mu + \gamma X_i + U_i \tag{56}$$

Note that 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 Δ from the difference between the two estimated constant terms.

However, if $\text{COV}\{U_i, V_i\} \neq 0$ the error terms U_i do not have zero mean within each regime.

$$\text{Regime 1:} \qquad E\{U_i \mid V_i \geq -\alpha - \beta Z_i\} \neq E\{U_i\} = 0 \tag{57}$$

$$\text{Regime 0:} \qquad E\{U_i \mid V_i < -\alpha - \beta Z_i\} \neq E\{U_i\} = 0 \tag{58}$$

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 57 and 58, we could include them in the two regressions and avoid the misspecification.

2.7.3 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 and with covariance equal to σ_{UV}. Denote with $\phi(.)$ the standard normal density and with $\Phi(.)$ the standard normal cumulative distribution.

The following results can be easily proved (see Appendix in Maddala, 1983).

$$E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} > k_1\right\} = \frac{\phi(k_1)}{1-\Phi(k_1)} \tag{59}$$

$$E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} < k_2\right\} = -\frac{\phi(k_2)}{\Phi(k_2)} \tag{60}$$

$$E\left\{\frac{U}{\sigma_U} \mid k_1 < \frac{U}{\sigma_U} < k_2\right\} = \frac{\phi(k_1)-\phi(k_2)}{\Phi(k_2)-\Phi(k_1)} \tag{61}$$

and similarly for V. The ratios between the normal density and its cumulative on the RHS are called *Inverse Mill's ratios.*

$$\begin{aligned} E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} > k\right\} &= \sigma_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > k\right\} \\ &= \sigma_{UV}\frac{\phi(k)}{1-\Phi(k)} \end{aligned} \tag{62}$$

$$\begin{aligned} E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} < k\right\} &= \sigma_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} < k\right\} \\ &= -\sigma_{UV}\frac{\phi(k)}{\Phi(k)} \end{aligned} \tag{63}$$

2.7.4 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 51 and assuming joint normality for U_i and V_i.

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 51, and retrieve the (consistently) estimated absolute values of the *Inverse Mill's Ratios*

$$M_{1i} = \frac{\phi(-\hat{\alpha} - \hat{\beta} Z_i)}{1 - \Phi(-\hat{\alpha} - \hat{\beta} Z_i)} = \frac{\phi(\hat{\alpha} + \hat{\beta} Z_i)}{\Phi(\hat{\alpha} + \hat{\beta} Z_i)} \tag{64}$$

$$M_{0i} = \frac{\phi(-\hat{\alpha} - \hat{\beta} Z_i)}{\Phi(-\hat{\alpha} - \hat{\beta} Z_i)} = \frac{\phi(\hat{\alpha} + \hat{\beta} Z_i)}{1 - \Phi(\hat{\alpha} + \hat{\beta} Z_i)} \tag{65}$$

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

$$\text{Regime 1:} \qquad Y_i = \mu + \gamma X_i + \Delta + \lambda_1 M_{1i} + \nu_i \tag{66}$$

$$\text{Regime 0:} \qquad Y_i = \mu + \gamma X_i + \lambda_0 M_{0i} + \nu_i \tag{67}$$

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.

iii. Get a consistent estimate of the treatment effect Δ by subtracting the estimated constant in 67 from the estimated constant in 66.

2.7.5 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 these 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.

- How is the Heckman method performing according to LaLonde's (1986) results?

3 The Angrist-Imbens-Rubin approach for the estimation of causal effects

3.1 Notation

Consider the following framework:

- N individuals denoted by i.
- They are subject 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 observe that:
  i. assignment to treatement may or may not be random;
  ii. the correspondence between assignment and treatment may not be perfect.

Examples:

- Parental background for returns to schooling (Willis-Rosen, 1979).
- Quarter of birth for returns to schooling (Angrist and Krueger, 1994).
- Nearby college for returns to schooling (Card, 1995b)
- WWII for returns to schooling (Ichino and Winter-Ebmer, 2001)
- Vietnam war lottery for the effect of the military service (Angrist, 1990).

3.2 Definition of potential outcomes

The participation into treatment for individual i is a function of the full N-dimensional vectors of assignments $\mathbf{Z}$

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

The outcome for individual i is a function of the full N-dimensional 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.

Our goal is to establish which of these effects can be identified and estimated, and whether this can be done for a random individual in the population or only for a random individual in a sub-group of the population.

To do so we need to begin with a set of assumptions and definitions.

3.3 Assumptions of the Angrist-Imbens-Rubin Causal model

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

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

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

where $\mathbf{Z}$ and $\mathbf{D}$ (note the bold face) are the N-dimensional vectors of assignments and treatments.

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

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

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

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

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

It is crucial to imagine that for each individual the full sets of

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

even if only one item for each set is actually observed; this implies thinking in terms of couterfactuals.

Implications for general equilibrium analysis?

Table 1: Classification of individuals according to assignment and treatment

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

Note that each individual i effectively falls in one and only one of these four cells, even if all the full sets of assignments, treatments and outcomes are conceivable.

Examples:

- Parental background for returns to schooling (Willis-Rosen, 1979).
- Quarter of birth for returns to schooling (Angrist and Krueger, 1994).
- Nearby college for returns to schooling (Card, 1995b)
- WWII for returns to schooling (Ichino and Winter-Ebmer, 2001)
- Vietnam war lottery for the effect of the military service (Angrist, 1990).

Assumption 2 *Random Assignment.*
All individuals have the same probability to be assigned to the treatment:

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

Given these first two assumptions we can consistently estimate the two *intention to treat* average effects by substituting sample statistics on the RHS of the following population equations:

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

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

Note that the ratio between the causal effect of Z_i on Y_i (eq. 71) and the causal effect of Z_i on D_i (eq. 70) gives the conventional IV estimator

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

The questions that we need to answer are:

- Under which assumptions this IV estimator gives an estimate of the the average causal effect of D_i on Y_i and for which (sub-)group in the population?
- Does the estimate depend on the instrument we use?

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

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

Note that this assumption is equivalent to the assumption 36 in the conventional approach to IV: i.e. the assumption that requires the instrument to be correlated with the endogenous regressor.

This assumption can be tested as in the conventional approach.

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

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

This assumption plays the same role as exclusion restrictions (assumption 37) in the conventional approach to IV.

It cannot be tested because it relates quantities that can never be observed jointly: we can never observe the two sides of the equation:

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

This assumption 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 individual i is*

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

As we know from the first lecture we cannot compute this causal effect because there is no individual for which we observe both its components.

We can, nevertheless, compare sample averages of the two components for individuals who are in the two treatment groups only because of different assignments , i.e. for *compliers* or *defiers*.

Provided that assignment affects outcomes only through treatment, the difference between these two sample averages seems to allow us to make inference on the causal effect of D on Y. But ...

Are the first four assumptions enough?

The four assumptions that we made so far allow us to establish the relation *at the individual level* between the *intention to treat* causal effects of Z on D and Y and the causal effect of D on Y.

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

Equation 73 states that at the individual level the causal effect of Z on Y (see Definition 3) is equal to the product of the the causal effect of Z on D (see Definition 2) times the causal effect of D on Y (see Definition 4).

At a first approximation it would seem that by taking expectations on both sides of 73 we could construct an estimator for the causal effect of D on Y. But ...

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

Equation 74 clearly shows that even with the four assumptions that were made so far we still have an identification problem: the average treatment effect for *compliers* may cancel with the average effect for *defiers*.

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

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

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

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

Note that the combination of Assumptions 3 and 5 implies:

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

This combination is called *Strong Monotonicity*, and ensures that:

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

Thanks to this assumption the average treatment effect for *defiers* is zero by assumption in equation 74

3.4 The Local Average Treatment Effect

3.4.1 Definition and relationship with IV

Given the monotonicity Assumption 5 equation 74 can be written as

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

Rearranging this equation we get the equation that defines the Local Average Treatment Effect:

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

Definition 5 *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.*

Substitution of the appropriate sample statistics in the expression on the RHS gives an estimate of the LATE.

The correct estimator of the covariance matrix for the LATE is the *White-Robust* estimator (see Angrist-Imbens, 1994)

Equivalent definitions of the LATE

$$
\begin{aligned}
E\{Y_i(1) - Y_i(0) \quad | \quad & D_i(1) = 1, D_i(0) = 0\} \\
&= \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i(1) = 1\} - Pr\{D_i(0) = 1\}} && (79) \\
&= \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i = 1 \mid Z_i = 1\} - Pr\{D_i = 1 \mid Z_i = 0\}} && (80) \\
&= \frac{COV\{Y, Z\}}{COV\{D, Z\}} && (81)
\end{aligned}
$$

Comments

- In order to go from 78 to 79 note that

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

  because there are no defiers.

- In order to go from 80 to 81 see the appendix 6.3

- The last expression 81 shows that the IV estimand is the LATE. In other words, under the assumptions made above IV estimates are estimates of Local Average Treatment Effects.

- The LATE is the only treatment effect that can be estimated by IV, and the causal interpretation of IV can only coincide with the causal interpretation of the LATE

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* $Y_i(1,1) - Y_i(0,1) = 0$

3.4.2 Causal interpretation of the LATE-IV estimator

- 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 individual independently of the others.
- The random assignment assumption allows us to estimate this average effect using sample statistics.
- Exclusion restrictions ensure this causal effect is zero for the *always-* and *never-takers*; it is non-zero only for *compliers* and *defiers* (via D).
- The assumptions of strong monotonicity ensure 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).

Table 3: Frequency of each type of individual 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* $Pr\{D_i(1) = 1, D_i(0) = 1\}$

- 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 80):

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

- 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.

3.5 Effects of violations of the LATE assumptions

3.5.1 Violations of Exclusion Restrictions

Suppose that all the assumptions hold except for the exclusion restrictions. Let the causal effect of Z on Y be

$$H_i = Y_i(1, d_1) - Y_i(0, d_0)$$

where $(d_1 = d_0 = 0)$ for *never takers*, $(d_1 = d_0 = 1)$ for *always takers* and $(d_1 = 1; d_0 = 0)$ for compliers.

Exclusion restrictions require

- for *non-compliers*: $H_i = 0$;
- Also for *compliers* $H_i = 0$ but H_i should be interpreted as the direct effect of Z on Y in addition to the indirect effect via D.

Then the IV estimand is equal to:

$$E[H_i \mid i \text{ is a complier}] + E[H_i \mid i \text{ is a noncomplier}] \cdot \frac{P[i \text{ is a noncomplier}]}{P[i \text{ is a complier}]} \tag{82}$$

- The first term is the LATE plus the bias due to violations of exclusion restrictions for *compliers*; the bias would exist even with perfect compliance.
- The second term is due to violations of exclusion restrictions for *non-compliers*; it decreases with compliance.

Note that the higher the correlation between assignment and treatment (i.e. the "stronger" the instrument), the smaller the odds of non-compliance and consequently IV is less sensitive to violations of exclusion restrictions, because the second term of the bias defined above decreases.

However, even the strongest instruments would suffer from violations of exclusion restrictions for compliers (the first term).

3.5.2 Violations of the Monotonicity Condition

Suppose that all the assumptions are satisfied except monotonicity. Then the IV estimand is equal to the LATE plus the following bias:

$$-\lambda \cdot \{E[Y_i(1) - Y_i(0) \mid i \text{ is a defier}] - E[Y_i(1) - Y_i(0) \mid i \text{ is a complier}]\} \quad (83)$$

where

$$\lambda = \frac{P(i \text{ is a defier})}{P(i \text{ is a complier}) - P(i \text{ is a defier})}$$

- The first multiplicative component of the bias is λ. This component is related to the probability of *defiers* and is zero if the monotonicity assumption is satisfied.
- Note that λ decreases with the proportion of *defiers* and its denominator is the average causal effect of Z on D. So again the "stronger" the instrument the smaller the bias.
- The second multiplicative component is the difference between the average causal effect of D on Y for *compliers* and *defiers*.
- Note that this second component could be close to zero, even if monotonicity is not satisfied.

3.6 LATE with multiple instruments, with Covariates and with non-binary treatments

Angrist and Imbens (1994) and (1995) show the following important results

i. *Multiple Instruments*

- The standard IV-TSLS estimator with multiple instruments gives an average of the LATE estimates that we would obtain using each instrument separately.
- In this case the weights are proportional to the "strength" of the instrument: the bigger the impact of the instrument on the regressor, the more weight it receives in the TSLS linear combination.

ii. *Covariates*
In the presence of covariates the interpretation of LATE is not so simple.

- One possibility is to assume that counterfactuals are additive in covariates which leaves things unchanged
- The other possibility is to think that the TSLS estimate is a variance-weighted average of the LATEs conditional on the covariates.

iii. *Non-binary treatments*
The LATE interpretation of IV-TSLS can be easily extended to the non-binary treatments (see Angrist and Imbens , 1995)

3.7 Alternative and more informative ways to estimate the LATE

IV is not the only way to estimate the LATE. Imbens and Rubin (1997a), Imbens and Rubin (1997b) and Hirano, Imbens, Rubin and Zhou (2000) propose a different estimation strategy which not only allows to estimate the LATE but also:

- allows to estimate the entire outcome distributions for the always takers, the never takers and the compliers;
- gives insights on the characteristics of these subgroups in the population
- offers a way to test a weaker version of the exclusion restrictions assumption.

The starting point of this alternative estimation strategy is the observation that, given the absence of defiers:

- units such that $Z_i = 0$ and $D_i = 1$ are certainly *always-takers*;
- units such that $Z_i = 1$ and $D_i = 0$ are certainly *never-takers*;
- units such that $Z_i = 1$ and $D_i = 1$ are a mixture of *always-takers* and *compliers*;
- units such that $Z_i = 0$ and $D_i = 0$ are a mixture of *never-takers* and *compliers*;

The impossibility to observe counterfactual events prevents the identification of the *compliers* but not the possibility to estimate the probability that a unit belongs to one of the three sub-populations.

The probability to be *always-taker*, *never-taker*, or *complier*

Denote with:

- ω_a the probability to be an *always-taker*,
- ω_n the probability to be a *never-taker*,
- ω_c the probability to be a *complier*

Given that assignment is random it follows that:

$$\phi_n \equiv Pr(D_i = 0|Z_i = 1) = \frac{\omega_n \cdot Pr(Z_i = 1)}{Pr(Z_i = 1)}, \tag{84}$$

and

$$\phi_a \equiv Pr(D_i = 1|Z_i = 0) = \frac{\omega_a \cdot Pr(Z_i = 0)}{Pr(Z_i = 0)}, \tag{85}$$

where note that ϕ_n and ϕ_a are directly estimable from the observed sample. Therefore the correspondent sample statistics can be used as estimates of the unobservable probabilities ω_n and ω_a.

Since there are no *defiers*, $\omega_n + \omega_a + \omega_c = 1$ and therefore an estimate of the probability to be a complier can be obtained using the fact that:

$$\phi_c = 1 - \phi_a - \phi_n. \tag{86}$$

Note that this is the denominator of the LATE.

The outcome distribution for the *always-takers*, the *never-takers* and the *compliers*

Denote with:

- $g_{n1}(y_i)$ the unobservable outcome distribution of the *never takers* assigned to treatment;
- $g_{n0}(y_i)$ the unobservable outcome distribution of the *never takers* not assigned to treatment;
- $g_n(y_i)$ the unobservable outcome distribution of the *never takers*.

If the exlusion restriction assumption holds:

$$g_{n1}(y_i) = g_{n0}(y_i) = g_n(y_i). \tag{87}$$

A similar notation can be used for the *always takers* and for them as well:

$$g_{a1}(y_i) = g_{a0}(y_i) = g_a(y_i). \tag{88}$$

Denote with $f_{zd}(y_i)$ the directly estimable outcome distribution for the units such that $Z_i = z$ and $D_i = d$. Note that:

- the units such that $Z_i = 1$ and $D_i = 0$ are certainly *never takers* assigned to treatment and therefore:

$$f_{10}(y_i) = g_{n1}(y_i) \tag{89}$$

- the units such that $Z_i = 0$ and $D_i = 1$ are certainly *always takers* not assigned to treatment and therefore:

$$f_{01}(y_i) = g_{a0}(y_i) \tag{90}$$

So, we can easily estimate, using the sample, the distribution of these two sub-populations.

We can also directly estimate the distribution $f_{00}(y_i)$ which corresponds to units who are a mixture of *compliers* and *never takers*. However, given the probabilities to be in one of these two sub-populations (see 84, 85 and 85), we can write that:

$$f_{00}(y_i) = \frac{\phi_n}{\phi_c + \phi_n} g_n(y_i) + \frac{\phi_c}{\phi_c + \phi_n} g_{c0}(y_i). \tag{91}$$

Similarly for the observed distribution $f_{11}(y_i)$ which can be written as:

$$f_{11}(y_i) = \frac{\phi_a}{\phi_c + \phi_a} g_a(y_i) + \frac{\phi_c}{\phi_c + \phi_a} g_{c1}(y_i). \tag{92}$$

Inverting these expressions and using the equations 87, 88, 90 and 89, we can write the four unoservable distributions of interest in terms of directly estimable distributions and parameters.

- *compliers* assigned to no treatment:

$$g_{c0}(y_i) = \frac{\phi_c + \phi_n}{\phi_c} f_{00}(y_i) - \frac{\phi_n}{\phi_c} f_{10}(y_i), \tag{93}$$

- *compliers* assigned to treatment:

$$g_{c1}(y_i) = \frac{\phi_c + \phi_a}{\phi_c} f_{11}(y_i) - \frac{\phi_a}{\phi_c} f_{01}(y_i) \tag{94}$$

- *always takers:*

$$g_a(y_i) = f_{01}(y_i) \tag{95}$$

- *never takers:*

$$g_n(y_i) = f_{10}(y_i) \tag{96}$$

These results are important in several ways.

3.7.1 Anatomy of IV estimates

Let $C_i = \{c, n, a, d\}$ if i is, respectively, a *complier, a never taker, an always taker or a defier* . From equations 78, 79, 80 and 81 we can rewrite the IV estimator as:

$$\begin{aligned} \hat{\Delta}_{IV} &= \frac{COV\{Y,Z\}}{COV\{D,Z\}} = \frac{\bar{y}_{.1} - \bar{y}_{.0}}{\bar{d}_1 - \bar{d}_0} \\ &= \frac{\bar{d}_1\bar{y}_{11} - \bar{d}_0\bar{y}_{10}}{\bar{d}_1 - \bar{d}_0} - \frac{(1-\bar{d}_0)\bar{y}_{00} - (1-\bar{d}_1)\bar{y}_{01}}{\bar{d}_1 - \bar{d}_0} \\ &= \int_{y_i} y_i \hat{g}_{c1}(y_i) - \int_{y_i} y_i \hat{g}_{c0}(y_i) \end{aligned} \tag{97}$$

where $\bar{y}_{zd}$ is the average outcome for units such that $(D_i, Z_i) = (d, z)$, $\bar{d}_z$ is the average of the treatment indicator for units such that $Z_i = z$, and $\hat{g}_{cz}(y_i)$ is the estimate of the distribution $g_{cz}(y_i)$ obtained: as

$$\hat{g}_{cz}(y_i) = (1-z)\left[\frac{\phi_c + \phi_n}{\phi_c}\hat{f}_{00}(y_i) - \frac{\phi_n}{\phi_c}\hat{f}_{10}(y_i)\right] + z\left[\frac{\phi_c + \phi_a}{\phi_c}\hat{f}_{11}(y_i) - \frac{\phi_a}{\phi_c}\hat{f}_{01}(y_i)\right], \tag{98}$$

where the $\hat{f}_{zd}(y_i)$ are the sample counterparts of the four directly estimable distributions $f_{zd}(y_i)$.

The above decomposition has two important implications:

i. IV can only give estimates of the difference on the LHS of

$$E\{Y_i(1) - Y_i(0)|C_i = c\} = E\{Y_i(1)|C_i = c\} - E\{Y_i(0)|C_i = c\} \tag{99}$$

while the estimation of the distribution $g_{cz}(y_i)$ allows to obtain estimates of the two terms on the RHS. These separate estimates are informative.

ii. The IV estimator does not take into account the fact that the two distributions $f_{00}(.)$ and $f_{11}(.)$ are mixtures of $g_n(.)$ and $g_{c0}(.)$, and $g_a(.)$ and $g_{c1}(.)$ respectively. Being densities, these mixtures should be non negative, but inspection of 98 shows that in small samples this constraint may not be satisfied. Imbens and Rubin (1997b) offer an interesting discussion (with an example) of the consequences of not imposing this constraint.

3.7.2 Maximum likelihood estimation

A more informative alternative to IV is a maximum likelihood approach in which the *the compliance status* becomes a parameter to be estimated, together with the outcome distributions for each type of unit in the population. As shown in Imbens and Rubin (1997a) and Mercatanti (1999), a likelihood function can be defined

- over the full set of actual and counterfactual "observations" for each unit, but
- assuming that the counterfactual "observations" are *missing at random* and
- integrating appropriately over the missing observations.

As a result, using SUTVA and random assignment the likelihood of the observed outcomes can be written as:

$$\begin{aligned} L\left(\theta|Y_{obs}\right) &= \Pi_{i\in(D_i=1,Z_i=0)}(\omega_a g^i_{a0} + \omega_d g^i_{d0}) \\ &\times \Pi_{i\in(D_i=0,Z_i=1)}(\omega_n g^i_{n1} + \omega_d g^i_{d1}) \\ &\times \Pi_{i\in(D_i=1,Z_i=1)}(\omega_a g^i_{a1} + \omega_c g^i_{c1}) \\ &\times \Pi_{i\in(D_i=0,Z_i=0)}(\omega_n g^i_{n0} + \omega_c g^i_{c0}). \end{aligned} \tag{100}$$

where Y_{obs} is the vector of observed outcomes and

$$\theta = \left(\omega_{\mathbf{a}}, \omega_{\mathbf{n}}, \omega_{\mathbf{c}}, \omega_{\mathbf{d}}, \eta_{\mathbf{a0}}, \eta_{\mathbf{a1}}, \eta_{\mathbf{n0}}, \eta_{\mathbf{n1}}, \eta_{\mathbf{c0}}, \eta_{\mathbf{c1}}, \eta_{\mathbf{d0}}, \eta_{\mathbf{d1}}\right), \tag{101}$$

is the parameters vectors composed by

- the proportions ω_t (with $t = c, a, n, d$) of *compliers, always takers, never takers and defiers* in the population
- the parameters η_{tz} are the parameters of the eight outcome distributions g_{tz} of the units assigned to treatment z and belonging to group t.

Given the presence of mixtures of distributions in this likelihood its maximization requires special algorithms (like the EM algorithm) for which see Imbens and Rubin (1997a,b,), Mercatanti (1999) and their references.

Maximum likelihood estimation of the LATE

Adding the other three assumptions for the identification of the LATE we obtain a likelihood from which the LATE can be estimated:

- *Monotonicity* imposes the absence of *defiers* which implies $\omega_d = 0$ and the irrelevance of the distributions g_{dz} and of their parameters η_{dz}.
- *The existence of a positive causal effect of Z on D* ensures the existence of some *compliers* which implies $\omega_c > 0$
- *Exclusion restrictions* require that, given D, Z has no effect on Y and therefore impose that
  - for the *compliers* $g_{c1} - g_{c0} \neq 0$ only because the treatment differs in the two groups;
  - for the *always takers* $g_{a0} = g_{a1} = g_a$;
  - for the *never takers* $g_{n0} = g_{n1} = g_n$.

With these assumption, the likelihood simplifies to

$$\begin{aligned} L_{LATE}\left(\theta|Y_{obs}\right) &= \Pi_{i\in(D_i=1,Z_i=0)}\omega_a g_a^i \\ &\times \Pi_{i\in(D_i=0,Z_i=1)}\omega_n g_n^i \\ &\times \Pi_{i\in(D_i=1,Z_i=1)}(\omega_a g_a^i + \omega_c g_{c1}^i) \\ &\times \Pi_{i\in(D_i=0,Z_i=0)}(\omega_n g_n^i + \omega_c g_{c0}^i), \end{aligned} \tag{102}$$

Where the vector of parameters now is

$$\theta = (\omega_{\mathbf{a}}, \omega_{\mathbf{n}}, \omega_{\mathbf{c}}, \eta_{\mathbf{a}}, \eta_{\mathbf{n}}, \eta_{\mathbf{c0}}, \eta_{\mathbf{c1}})\,. \tag{103}$$

Given maximum likelihood estimates of the parameters θ, an estimate of the LATE can be obtained substituting estimates of the distributions $g_{c1}(y)$ and $g_{c0}(y)$ in:

$$E\{Y_i(1) - Y_i(0) \mid C_i = c\} = \int y\, g_{c1}(y)\, dy - \int y\, g_{c0}(y)\, dy \tag{104}$$

Imbens and Rubin (1997a) call this parameter *CACE* (*Compliers Average Causal effect*) when the exclusion restriction assumption cannot be assumed to hold for the compliers.

3.7.3 A test of a weak version of the exclusion restrictions assumption

- Consider again the unrestricted likelihood function 100 and impose strong monotonicity.
- Assume that that the exclusion restriction holds for *compliers*.
- Allow instead for the possibility that exclusion restrictions do not hold
  - for the *always takers*: $g_{a0} \neq g_{a1}$;
  - for the *never takers*: $g_{n0} \neq g_{n1}$.
- The likelihood under these assumptions is

$$\begin{aligned} L_{weak}\left(\theta|Y_{obs}\right) &= \Pi_{i\in(D_i=1,Z_i=0)}(\omega_a g_{a0}^i) \\ &\times \Pi_{i\in(D_i=0,Z_i=1)}(\omega_n g_{n1}^i) \\ &\times \Pi_{i\in(D_i=1,Z_i=1)}(\omega_a g_{a1}^i + \omega_c g_{c1}^i) \\ &\times \Pi_{i\in(D_i=0,Z_i=0)}(\omega_n g_{n0}^i + \omega_c g_{c0}^i). \end{aligned} \tag{105}$$

- A likelihood ratio test based on $L_{weak}\left(\theta|Y_{obs}\right)$ and $L_{LATE}\left(\theta|Y_{obs}\right)$ provides a test for the hypothesis that the exclusion restriction assumptions holds for the *always takers* and the *never takers*, although it may not hold for the *compliers*, for whom no test is possible

A similar testing strategy can also be used to test the monotonicity assumption.

3.7.4 LATE and Average Effect of Treatment on the Treated

Evidence that the outcome distribution is similar for the *compliers* treated and for the *always takers* suggests that the LATE may be close to the Average Effect of Treatment on the Treated (ATT).

In general the comparison between the outcome distributions may contain useful information.

3.8 Comments on the LATE and the conventional interpretation of IV

i. The AIR approach helps to clarify the set of assumptions under which IV may be interpreted as a way to estimate an average causal effect.

ii. To identify the effect of treatment on the treated the conventional approach assumes (see eq. 49)

$$E\{U_i(1) - U_i(0) \mid Z_i, D_i = 1\} = E\{U_i(1) - U_i(0) \mid D_i = 1\} \qquad (106)$$

This assumption says that the average idiosyncratic gain for the treated conditioning on the instrument, should be identical to the unconditional average idiosyncratic gain for the treated.

iii. Translated in the AIR framework assumption 106 is (see the debate Heckman-AIR in AIR, 1996):

$$E\{Y_i(1)-Y_i(0) \mid Z_i, D_i(Z_i) = 1\} = E\{Y_i(1)-Y_i(0) \mid D_i(Z_i) = 1\} \quad (107)$$

$$\begin{aligned} E\{Y_i(1) - Y_i(0) \quad &| \quad D_i(1) = 1; D_i(0) = 1\} \qquad (108) \\ &= \quad E\{Y_i(1) - Y_i(0) \mid D_i(1) = 1; D_i(0) = 0\} \end{aligned}$$

In words, the causal effect of D on Y must be the same for both *compliers* and *always-taker*, i.e. must be identical for all the treated. The maximum likelihood approach to the estimation of the LATE allows to obtain evidence on the validity of this assumption, while in the conventional approach there is no way to assess its validity.

iv. Note that in the conventional approach also the assumption of strong monotonicity is hidden. It is in fact implicit in the specification of the participation equation (more precisely: the common parameter β in equation 13).

v. If one does not want to assume that the effect of treatment is the same for both *compliers* and *always-taker* and given all the other assumptions, the AIR approach concludes that the only causal effect that one can identify and estimate is the causal effect for *compliers* that is the Local Average

Treatment Effect: the effect of treatment on those who would 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 individuals observed in both treatment status (given that *defiers* have been ruled out).
- *always takers* and *never-takers* are observed only in one of the two treatment status
- The LATE is analogous to a regression coefficient estimated in linear models with individual effects using panel data. The data can only be informative about the effect of regressors on individuals for whom the regressor change over the period of observation.

vii. The maximum likelihood approach to the estimation of the LATE provides additional valuable information with respect to IV. In particular it allows to get a better sense of who are the *compliers*, the *always-takers* and the *never-takers*, and even to test a weak version of the exclusion restrictions assumption.

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

ix. Hence, the conventional approach seems to conclude that it is preferable to make additional assumptions like 106 or the ones required for the Heckman two steps procedures, in order to answer more interesting and well posed policy questions.

3.9 Problems with IV when the instruments are weak

An instrument is "weak" when its correlation with the treatment is low. This situation has three important consequences:

i. If the assumptions that ensure consistency are satisfied,

   (a) the standard error of the IV estimate increases with the weakness of the instrument.

   (b) in finite samples the IV estimate is biased in the same way as the OLS estimate, and the weaker the instrument the closer the IV bias to the OLS bias.

ii. If the assumptions that ensure consistency are violated, the weakness of the instrument exacerbates the inconsistency of the IV estimate, so that even a mild violation leads to an inconsistency which is larger the weaker the instrument.

These consequences apply with some caveats to both the conventional and the AIR approach to IV

3.9.1 Weakness of the instrument and efficiency

Using a more general matrix notation, the covariance of the IV estimator using the conventional approach is given by

$$VAR\{\Delta\} = \sigma^2(Z'D)^{-1}Z'Z(Z'D)^{-1} \tag{109}$$

Clearly a weaker correlation between Z and D reduces efficiency of the IV estimator.

The correct estimator of the covariance matrix for the LATE is the *White-Robust* estimator (see Angrist-Imbens, 1994). But also in this case the weakness of the instrument generates a similar problem.

3.9.2 Weakness of the instrument and finite samples

Within the conventional approach,

- even if the instruments are legitimate and IV is consistent, in finite samples IV gives biased estimates.
- The weaker the instrument the closer is IV to OLS.

The intuition is:

- Consider the extreme case in which $COV\{D, Z\} = 0$.
- Nevertheless, in finite samples, the first stage provides estimates of the causal effect of Z on D.
- These estimates allow to obtain an arbitrary decomposition of D into an "exogenous" and an "endogenous" component.
- It is not surprising that the second stage regression of the outcome on the (arbitrary) exogenous component is similar to OLS.

Staiger and Stock, 1997 give a useful practical method to evaluate the seriousness of this problem (independently of distributional assumptions):

- Let F be the F-statistics on the excluded instruments in the first stage.
- $1/F$ is an estimate of the ratio between the finite sample bias of IV and the OLS bias.

Within the AIR approach, this finding implies that in finite samples, if the instrument is weak, IV may be closer to OLS than to the LATE.

See the discussion of Angrist and Krueger (1991) in Staiger and Stock (1997) and in Bound et al. (1995).

3.9.3 Weakness of the instrument and consistency

In the presence of violations of the exclusion restrictions (even if these are mild) the weakness of the instrument exaggerates the size of the related bias.

Consider the conventional version of our model:

$$Y_i = \mu + \Delta D_i + U_i \tag{110}$$

The IV estimand is

$$\begin{aligned} Plim\{\Delta^{IV}\} &= \frac{COV\{Z,Y\}}{COV\{Z,D\}} \\ &= \Delta + \frac{COV\{Z,U\}}{COV\{Z,D\}} \end{aligned} \tag{111}$$

Note that:

- if $COV\{Z,U\} \neq 0$ IV is inconsistent;
- the inconsistency is larger the smaller the $COV\{Z,D\}$;
- even if $COV\{Z,U\}$ is small the inconsistency can be very large.

See the discussion of Angrist and Krueger (1991) in Bound et al. (1995).

The same problem exists in the AIR approach, with the caveat that the bias has to be intended with respect to the LATE.

- section 3.5.1 we have seen that the bias due to exclusion restrictions violations increases with the weakness of the instrument.
- In section 3.5.2 we have seen that the bias due to monotonicity violations increases with the weakness of the instrument.

4 A Model of the Effect of Education on Earnings

In order to better understand the nature of the treatment effects studied so far, we will now define them in the context of the relationship between education and earnings.

Hundreds of studies from many different countries have estimated the following wage equation (see Mincer, 1974):

$$ln(W) = \alpha + \beta S + \gamma E + \delta E^2 + \epsilon \tag{112}$$

where W is the wage, S is years of schooling and E is years of labor market experience, finding that more educated workers earn higher wages (e.g. Psacharopoulos, 1985; Ashenfelter and Rouse, 1999; Card 1995a).

There are few similar regularities in economics and this is the reason why labor economists devoted so much attention to it.

Despite this evidence "most economists are reluctant to intepret the earning gap between more or less educated workers as an estimate of the causal effect of schooling". (Card, 1995a)

So far we have seen in general terms the problems connected to the definition and identification of causality.

In this part of the course we build on the canonical model of Becker (1967), as revisited by Card (1995a), to explore the counterpart of those general problems in the specific analysis of the causal effect of education on earnings.

4.1 The income generating function

We assume that going to school is a way to accumulate human capital and that a higher human capital generates higher earnings in the labor market:

$$Y = Y(S) \tag{113}$$

where:

- S is the number of years of scholing;
- $Y(S)$ is the income generated by the human capital accumulated in S years of schooling;
- the income generating function is assumed increasing and concave ($Y' > 0$ and $Y'' < 0$).

(Figure: The income generating function)

4.2 The objective function

Individuals choose the optimal number of years of schooling S to maximize the present discounted value of income

$$V(S,Y) = \int_S^\infty Y(S)e^{-rt}dt = \frac{Y(S)e^{-rS}}{r}, \tag{114}$$

where $Y(S)$ is income and r is the discount rate.

Taking logs, we can write the utility to be maximised as

$$\tilde{U}(S,Y) = \log(V(S,Y)) = \log(Y) - rS - log(r) \tag{115}$$

which formalizes the idea that:

- schooling is useful because it generates income
- but it is costly because of foregone earnings.

(Figure: earnings at different levels of S)

Here we adopt a more general expression for the utility function:

$$U(S,Y) = \log(Y) - h(S) \tag{116}$$

where $h(s)$ captures also other components of the cost of schooling in addition to foregone earnings.

Strict convexity of h implies that the marginal cost of each additional year of schooling rises by more than foregone earnings:

- tuition;
- foregone earnings;
- psychic costs;
- liquidity constraints.

(Figure: Indifference curves for utility functions 115 and 116)

4.3 The optimization problem

The optimization problem for each individual is therefore:

$$\begin{aligned} \text{Max } U(Y,S) &= \log(Y) - h(S) \\ \text{subject to } Y &= Y(S) \end{aligned} \tag{117}$$

The optimal number of years of schooling is given by the solution of the F.O.C:

$$\frac{Y'(S)}{Y(S)} = h'(S) \tag{118}$$

where:

- $\frac{Y'(S)}{Y(S)} =$
  - marginal rate of return of one year of schooling, or
  - marginal rate of transformation of schooling into income;
- $h'(S) =$
  - marginal cost of one year of schooling, or
  - marginal rate of substitution between schooling and income.

(Figure: The optimal choice of years of schooling)

4.4 From the model to the data

The model as described above does not allow for heterogeneity across individuals and therefore generates a single optimal combination of S and Y.

If we plot the combinations S and Y observed in the data (i.e. a sample of empirical observations) we obtain a cloud of points.
(Figure: The data)

This suggests that we need to introduce some form of heterogeneity in the model if we want the model to say something interesting on the data.

Card (1995a) assumes heterogeneity in the individual marginal returns to schooling and in the individual marginal costs of schooling

$$\left[\frac{Y'(S)}{Y(S)}\right]_i = \beta_i(S) = b_i - k_b S \tag{119}$$

$$[h'(S)]_i = \delta_i(S) = r_i + k_r S \tag{120}$$

For example:

b_i: differences in individual ability that generate heterogeneity of marginal returns to schooling.

r_i: differences in liquidity constraints that generate heterogeneity of marginal costs of schooling.

Understanding the Heterogeneity of Marginal Returns

The marginal return is a linear function of schooling with individual specific intercepts:

$$\left[\frac{Y'(S)}{Y(S)}\right]_i = \beta_i(S) = b_i - k_b S$$

We can interpret b_i as an indicator of "ability".

This assumption implies a specific functional form for the income generating function. By integration:

$$[Y(S)]_i = ae^{(b_i S - (\frac{k_b}{2}S^2))} \tag{121}$$

(Figure: Income generating functions for different abilities)

Note that this implies a specific characterization of ability:

- ability increases the slope of the income generating function, i.e. the marginal return to schooling

With standard homothetic preferences this assumption ensures that more able individuals choose more shooling.

We could have assumed alternatively that

- ability shifts up the income generating function in a parallel fashion, i.e. it increases incomes for each level of schooling leaving marginal returns unchanged

In this case with standard homothetic preference more able people choose less schooling.

Figure(Comparison of optimal choices in the two cases)

Understanding the Heterogeneity of Marginal Costs

Also the marginal cost is a linear function of schooling with individual specific intercepts:

$$[h'(S)]_i = \delta_i(S) = r_i + k_r S$$

We can interpret δ_i as the individual specific rate of return of the funds used to finance the Sth year of scholing (i.e. the opportunity cost).

Examples:

i. $k_r = 0$ and $r_i = r$
the opportunity cost of schooling does not increase with schooling and is equal across individuals. This the case of the utility function 115 and implies linear indifference curves with equal slopes for different individuals.

ii. $k_r = 0$ and $r_i \neq r_j$ for $i \neq j$
the opportunity cost of schooling does not increase with schooling but differs across individuals which implies linear indifference curves with different slopes for different individuals.

iii. $k_r > 0$ and $r_i = r$
The opportunity cost of schooling increases with schooling but is equal across individuals, which implies convex indifference curves with equal slopes for different individuals.

iv. $k_r > 0$ and $r_i \neq r_j$ for $i \neq j$
The opportunity cost of schooling increases with schooling and differs across individuals, which implies convex indifference curves with different slopes for different individuals.

(Figure: Comparison of the four examples)

To be focused, we will consider r_i as an indicator of the liquidity constraint faced by each individual.

Optimal schooling choices with heterogeneity

Substituting 119 and 120 in the first order condition 118, the optimal amount of schooling now differs across individuals:

$$S_i^* = \frac{(b_i - r_i)}{k_b + k_r} \tag{122}$$

The model can therefore generate data similar to what we observe. Note that:

- The optimal amount of schooling changes across individual because ability and discount rates differ.
- E.g., for given discount rate more able children choose more schooling.
- E.g., for given ability, less constrained children choose more schooling.

(Figure: Optimal choices of schooling with heterogeneity)

A controversial important correlation

The correlation between the individual ability b_i and the individual discount rate r_i can be expected to be negative if, for example:

- ability is partially inherited;
- more able parents have more education and higher incomes;
- higher income families have lower discount rates because
  - they are less liquidity constrained,
  - they like more education.

Given this expectation, the solution implies that richer children are likely to choose more schooling because they are on average more able and have lower discount rates.

(Figure: given the model is the evidence consistent with this controversial correlation?)

The causal effect of education in this model

For each individual we can define the marginal return to schooling β_i *at the optimal choice:*

$$\beta_i^* = b_i - k_b S_i^* = (1 - \phi) b_i + \phi r_i \tag{123}$$

where $\phi = \frac{k_b}{k_b + k_r}$.

Note that this is the causal effect of schooling on earnings for person i and, because of the Fundamental Problem of Causal Inference (Holland, 1986), it cannot be identified and measured.

We are, therefore, interested in understanding which average causal effects can be identified and measured using some standard statistical methods:

- Randomized control experiments;
- OLS estimation;
- IV estimation.

We will study the outcome of these methods when they are applied to data generated by a simplified version of the model presented above, in which there are only four types of individuals.

4.5 Data generated by a simplified model with four types of individuals

Consider a simplified version of the model corresponding to the example 2 on page 69 in which we assume linear indifference curves with different intercepts across individuals ($k_r = 0$ and $r_i \neq r_j$ for $i \neq j$).

Denoting log-earnings with y, the model is:

$$\begin{aligned} \text{Max } U_i(y, S) &= y - r_i S \\ \text{subject to } y &= b_i S - \frac{k_b}{2} S^2 \end{aligned} \tag{124}$$

$$\beta_i(S) = b_i - k_b S. \tag{125}$$

$$S_i^* = \frac{(b_i - r_i)}{k_b} \tag{126}$$

$$\beta_i^* = b_i - k_b S_i^* = r_i. \tag{127}$$

Note the difference between equation 127 and equation 123.

In what follows, to simplify the notation, we will omit the * denoting values corresponding to optimal choices.

(Figure: The optimal choice for an individual in this model)

The four types

Assume that there are only two values for each heterogeneity parameter:

$$b_H > b_L$$

$$r_H > r_L$$

so that there are four possible combinations denoted by $g = \{LH, HH, LL, HL\}$.

Each group $g = \{i, j\}$ operates a different educational choice

$$S_g \equiv S_{i,j} = \frac{(b_i - r_j)}{k_b}, \tag{128}$$

which implies the following optimal returns to schooling.

$$\begin{aligned} \beta_{LH} &= \beta_{HH} = r_H \\ \beta_{LL} &= \beta_{HL} = r_L. \end{aligned} \tag{129}$$

(Figure: Optimal choices for the four groups.)

The distribution of the four types in the population is given by:

$$\{P_{LL}, P_{LH}, P_{HL}, P_{HH}\}$$

Note that with this data generating process, the average causal effect of education in the population is:

$$\bar{\beta} = (P_{LH} + P_{HH})r_H + (P_{LL} + P_{HL})r_L = \bar{r}, \tag{130}$$

which would reduce to $\bar{r} = \frac{r_H + r_L}{2}$ in case of a uniform distribution across groups ($P_g = P = 0,25 \ \forall \ g$).

4.6 What can we learn from a randomized controlled experiment?

Suppose that we can extract two random samples of the population, denoted by C and T.

Suppose also that we can offer to individuals in T a fellowship which induces them to increase their education. This implies for them a reduction of the marginal cost of education r_j.

(Figure: Optimal choices of the treated and the controls in a randomized experiment.)

To simplify the analysis, without loss of generality, we assume that the fellowship program is structured in a way such that every treated individual increases her education by the same amount ΔS (e.g. one year).

$$\Delta S_g = \Delta S \qquad \forall g. \tag{131}$$

Given the randomized design of the experiment the controls provide the counterfactual evidence of what would have happened to the treated in the absence of the fellowship, and viceversa. Hence adapting equation 6 we obtain:

$$E(y_i|i \in T) - E(y_i|i \in C) = (P_{LH}+P_{HH})r_H\Delta S + (P_{LL}+P_{HL})r_L\Delta S = \bar{r}\Delta S = \bar{\beta}\Delta S \tag{132}$$

Since we are interested in the average effect on income per unit of treatment we can divide both sides by the average increase in education, which gives:

$$\begin{aligned}\frac{E\{y_i|i \in T\} - E\{y_i|i \in C\}}{E\{S_i|i \in T\} - E\{S_i|i \in C\}} &= \frac{E_g\{r_g \Delta S_g\}}{E_g\{\Delta S_g\}}. \\ &= \frac{(P_{LH} + P_{HH}) r_H \Delta S + (P_{LL} + P_{HL}) r_L \Delta S}{\Delta S} \\ &= \bar{r} \\ &= \bar{\beta}.\end{aligned} \tag{133}$$

Note that, the expression on the left hand side of 133, is our estimand.

The estimand is equal to the value $\bar{r}$ assumed in equilibrium by the average return to education in the population, i.e. $\bar{\beta}$.

If we substitute appropriate sample averages in the estimand we obtain a consistent estimate of the average causal effect of education on earnings.

However:

- is such an experiment feasible?
  - Ethical problems.
  - Technical problems.
- Should we be interested in this theoretical parameter?

4.7 What can we learn from OLS estimation?

Since the model implies a relationship between log-earnings and schooling , and both these variables are observables, we may try to estimate this relationship by OLS using observational data

Let's first recall what is the equilibrium relationship between y and S implied by the model. Note that what follows holds in general and not only in the "four types" example.

This relationship can be derived taking the log of equation 121, evaluated at the optimal individual choice S_i:

$$[Y(S_i)]_i = ae^{(b_i S_i - (\frac{k_b}{2} S_i^2))}$$

which yield:

$$y_i = ln(a) + b_i S_i - \frac{k_b}{2} S_i^2 \tag{134}$$

where $y_i = \ln [Y(.)]_i$.

Note that even if the theoretical relationship is quadratic the data points generated by this model are likely to be aligned along a linear relationship because:

- Among individuals with the same ability, different discount rates trace a concave relationship between log earnings and schooling.
- Among individuals with the same discount rate, different abilities trace a convex relationship between log earnings and schooling.

In data generated by both types of variability we may get a close-to-linear relationship, which tend to convex or concave depending on which type of heterogeneity has more variance.

(Figure: linearity of the relationship between log-earnings and schooling)

Suppose now that we estimate the liner equation

$$y_i = a + \rho S_i + \epsilon_i.$$

The OLS estimator of ρ has a probability limit given by:

$$\text{plim } (\hat{\rho}^{OLS}) = \frac{COV(y_i, S_i)}{VAR(S_i)} \tag{135}$$

Following Card(1995a):

$$\text{plim } (\hat{\rho}^{OLS}) = (1-\alpha)\bar{b} + \alpha\bar{r} \tag{136}$$

where $\bar{b} = E(b_i)$, $\bar{r} = E(r_i)$,

$$\alpha = \frac{k_b}{k_b + k_r} - \lambda$$

and

$$\lambda = \frac{\sigma_b^2 - \sigma_{br}}{(\sigma_b^2 - \sigma_{br}) + (\sigma_r^2 - \sigma_{br})}$$

which "is the fraction of the variance of schooling attributable to variation in ability as opposed to variation in discount rates."

In the case of fixed individual discount rates, $k_r = 0$ implies $\delta_i = r_i$, so that $\alpha = 1 - \lambda$ and

$$\text{plim } (\hat{\rho}^{OLS}) = \lambda\bar{b} + (1-\lambda)\bar{r}. \tag{137}$$

The OLS coefficient can be interpreted as a weighted average of the average ability and the average discount rate with weights that depend, respectively, on the variance of schooling due to ability and the variance due to discount rates.

We would like to know if we can recover from 137 the average marginal return to schooling, which using 127 can be written as:

$$E(\beta_i) = \bar{\beta} = \bar{b} - k_b \bar{S} \tag{138}$$

Note again that this holds in general for a model with $k_r = 0$, even in the presence of more than four types of individuals.

Using 138, equation 137 can be rewritten as:

$$\text{plim } (\hat{\rho}^{OLS}) = \bar{\beta} + \lambda(\bar{b} - \bar{r}). \tag{139}$$

Equation 139 says that the OLS regression of log-earnings on schooling yield a inconsistent estimate of the average marginal return to schooling. The bias is larger

- the larger is λ, i.e. the larger is σ_b^2 (the variance in ability) relative to σ_r^2 (the variance in discount rates);
- the larger is $\bar{b} - \bar{r}$, which is the difference between the average ability and the average discount rate.

The expression $\lambda(\bar{b} - \bar{r})$ can be interpreted as the endogeniety bias due to the fact that more able persons choose more schooling.

It is important to understand that OLS estimates ρ consistently. The problem is that ρ is not equal $\bar{\beta}$.

To better understand what we get using OLS, let's go back to our "four types" example and consider how $\hat{\rho}^{OLS}$ changes with the distribution of individuals across types.

(Figure: OLS estimates of Y on S with different distributions of types in the population.)

4.8 What can we learn from IV estimation?

The estimated equation is again:

$$y_i = a + \rho S_i + \epsilon_i$$

Consider a dichotomous instrument Z_i such that

$$E(S_i|Z_i = 1) \neq E(S_i|Z_i = 0).$$

The IV estimator for the return to schooling has Plim (see the Appendix Sections 6.1 and 6.3):

$$\text{plim } \rho_Z^{IV} = \frac{COV\ \{Y, Z\}}{COV\ \{S, Z\}} = \frac{E\{y_i|Z_i = 1\} - E\{y_i|Z_i = 0\}}{E\{S_i|Z_i = 1\} - E\{S_i|Z_i = 0\}} = \frac{E_g\{\beta_g \Delta S_{g|Z}\}}{E_g\{\Delta S_{g|Z}\}} \tag{140}$$

which in the case of our four types becomes:

$$\text{plim } \rho_Z^{IV} == \frac{P_{LH} r_H \Delta S_{LH} + P_{HH} r_H \Delta S_{HH} + P_{LL} r_L \Delta S_{LL} + P_{HL} r_L \Delta S_{HL}}{P_{LH} \Delta S_{LH} + P_{HH} \Delta S_{HH} + P_{LL} \Delta S_{LL} + P_{HL} \Delta S_{HL}}$$

- E_g: expectation taken on the distribution of the four groups.
- $\Delta S_{g|Z}$: exogenous change in schooling induced by Z in each group.

The traditional interpretation of IV

According to this interpretation the IV method reproduces the outcome of a randomized experiment in which assignment to treatment is described by the instrument Z and is controlled by nature in a way such that

$$\Delta S_{g|Z} = \Delta S_Z$$

i.e. the instrument induces the same marginal change in schooling for all the four groups and therefore:

$$\text{plim } \rho_Z^{IV} = E_g(\beta_g) = \bar{r} = \bar{\beta} \tag{141}$$

IV estimates consistently the average return to schooling in the population.

In the absence of heterogeneity, i.e. if $\beta_g = \beta$ for all g, it estimates the true and unique return in the population because:

$$\text{plim } \rho_Z^{IV} = E_g(\beta_g) = \beta$$

(Figure: Optimal choices of the treated and the controls in a perfectly controlled experiment: an IV interpretation)

A non-orthodox interpretation of IV

Suppose instead that nature controls the treatment imperfectly. Then:

$$\Delta S_{g|Z} \neq \Delta S_{h|Z} \qquad \text{for} \qquad g \neq h$$

i.e. the instrument induces a different marginal change in schooling in different groups, and we obtain

$$\text{Plim } \rho_Z^{IV} = \frac{E_g(\beta_g \Delta S_{g|Z})}{E_g(\Delta S_{g|Z})} \neq \bar{r} = \bar{\beta}.$$

The IV estimator based on Z is a weighted average of the marginal returns to schooling in the four groups where the weights depend on the impact of Z on S, $\Delta S_{g|Z}$.

(Figure: Optimal choices of the treated and the controls in an imperfectly controlled experiment: an IV interpretation)

This is also the LATE interpretation of IV:

IV estimates only the average return of those who change schooling because of a change in the instrument, i.e the so called *compliers*.

Different instruments have different *compliers*:

- Distance to college
- Compulsory schooling age
- Liquidity constraints caused by World War 2

4.9 An application to German data

Using data from the German Socio Economic Panel, we search for two instruments each one likely to affect a different group in the population (see Ichino and Winter-Ebmer, 1999):

- $Z_i = 1$ if father took part in World War 2 at the time the student was 10 years old

  $\Rightarrow$ expected to affect the group HH with the highest return

- $W_i = 1$ if father has more than high–school education

  $\Rightarrow$ expected to affect the group LL with the lowest return

Who are the compliers of the father–in–war instrument Z?

Having a father in war causes a reduction in schooling for individuals in group $g = HH$:

- these are high-ability but liquidity constrained individuals who choose more schooling in the absence of the war constraint but drop out of school if constrained by the war.

For none of the other groups the schooling decision is likely to be affected by the war:

- The rich dynasties $g = LL$ and $g = HL$ suffer limited liquidity constraints: they are the *never takers* who never stop at lower education anyway ;
- The poor dynasty $g = LH$ suffers liquidity constraints and in addition has low ability; they are the *always takers* who always stop at lower education.

Hence we expect:

$$\Delta S_{LL|Z} = \Delta S_{HL|Z} = \Delta S_{LH|Z} \approx 0$$
$$\text{plim } \rho_Z^{IV} \approx \beta_{HH} \tag{142}$$

IV based on Z should estimate the *highest* return in the population.

(Figure: Choices of the treated and the controls in a natural experiment)

Evidence on the compliers of the father–in–war instrument Z

Having a father involved in the war reduces schooling:

- by 1.59 (0.39) years for those students whose father had *only compulsory education*,
- only by 0.49 (0.82) years for other students.

Standard errors on parenthesis.

Who are the compliers of the father's education instrument W?

Having a highly educated father causes an increase in schooling for individuals in group $g = LL$:

- these are rich individuals with limited ability who may be pushed to reach a higher education if their parents are highly educated, but would not do it otherwise.

For none of the other groups the schooling decision is likely to be affected by parental education:

- the groups $g = HL$ and $g = HH$ have high ability: they are the *always–takers* who continue into higher education independently of the education of the father.
- group $g = LH$ has low ability and is heavily liquidity constrained: they are the *never–takers* who don't continue into higher education independently of parental education

Hence we expect:

$$\Delta S_{HL|W} = \Delta S_{HH|W} = \Delta S_{LH|W} \approx 0$$

$$\text{plim } \rho_W^{IV} \approx \beta_{LL} + N \tag{143}$$

where $N > 0$ is the potential bias caused by the existence of a direct causal effect of family background on earnings.

Evidence on the compliers of the father's–education instrument W

If the father has a degree higher than highschool, the years of schooling of the child increase:

- by 3.84 (0.66) years in households with *self–employed heads*,
- by 2.98 (0.31) years in households with *white–collar heads*
- only by 0.49 (0.96) years in households with *blue–collar heads*.

Standard errors in parentheses.

What if each instrument affected more than one group?

Suppose that:

- the *father–in–war* instrument Z affected not only group $g = HH$ but also other groups. Then:

$$\text{Plim}\beta_Z^{IV} = \frac{E_g(\beta_g \Delta S_{g|Z})}{E_g(\Delta S_{g|Z})} \leq \beta_{HH}.$$

- the *educated–father* instrument W affected not only group $g = LL$ but also other groups. Then:

$$\text{Plim}\beta_W^{IV} = \frac{E_g(\beta_g \Delta S_{g|W})}{E_g(\Delta S_{g|W})} \geq \beta_{LL}.$$

As a result, the difference between the IV estimates obtained with the two instruments would *underestimate* the true range of variation between the highest return β_{LL} and the lowest return β_{HH}.

(Figure: Optimal choices and LATE estimates)

IV estimates with different instruments in Germany

$$lnW_i = \beta_1 + \beta_2 EDU_i + \beta_3 AGE_i + \beta_4 AGE_i^2 + \beta_5 AGE_i^3 + \varepsilon_i$$

- Data: Men in the 1986 wave of the Socio–Economic Panel.
- W_i: hourly wage
- EDU_i: years of education
- The instruments are
  i. $Z_i = 1$ if i had a father in the army during the war;
  ii. $W_i = 1$ if i's father has more than high–school education

A potential problem leading to a richer specification

Bound and Jaeger (1996) argue that IV estimates could be biased upward by unobserved differences between the characteristics of the treatment and the control groups implicit in the IV scheme.

This would happen if treatment and control groups came from different social backgrounds.

Following a suggestion by Card (1998) we therefore include also information on parental background as control variables.

$$\begin{aligned} lnW_i &= \beta_1 + \beta_2 EDU_i + \beta_3 AGE_i + \beta_4 AGE_i^2 + \beta_5 AGE_i^3 \\ &\quad + \beta_6 HIGHEDF_i + + \beta_7 BLUEF_i + \beta_8 SELFF_i + \varepsilon_i \end{aligned} \tag{144}$$

Empirical results

Returns for one further year of schooling are estimated to be:

- 11.7% for the father–in–war instrument
- 4.8% for the father's–education instrument

These two estimates can be considered as an approximation of the upper and lower bounds of the returns to schooling in Germany.

Further comments

- Father's education is likely to have a direct positive impact on earnings. Therefore, the IV estimate based on father's education is likely to overestimate the lowest return

- If the instruments affect the schooling choices of all the groups in the population, the true range of variations of returns to schooling is likely to be larger than the one implied by the above two estimates.

Conclusions

- Returns to one year of education in Germany vary at least between 4.8% and 11.7%.
- Several reasons suggest that, if anything, the true range is likely to be larger than the one estimated here.

These results are consistent with the following picture:

- Returns to schooling are heterogeneous in the population.
- IV estimates should be interpreted as estimates of Local Average Treatment Effects: they measure the average return to schooling of those who change schooling because of the instrument.
- Therefore, with different instruments we can estimate the returns of different groups in the population, and in particular the highest and the lowest returns
- In this way we can approximate the range of variation of returns to schooling in the population.

Table 4: IV estimates of returns to schooling with different instruments in Germany.

	IV: Instrument Father in war	IV: Instrument: Father highly ed.	IV: Instrument: Father in war	IV: Instrument: Father highly ed.	OLS
Years of education	0.140	0.048	0.117	0.048	0.055
	(0.078)	(0.013)	(0.053)	(0. 014)	(0.005)
Age (years)	0.106	0.215	0.141	0.215	0.208
	(0.101)	(0.039)	(0.070)	(0.039)	(0.033)
Age^2 /100	-0.183	-0.434	-0.263	-0.434	-0.418
	(0.235)	(0.093)	(0.164)	(0.094)	(0.084)
Age^3 /10,000	0.106	0.291	0.165	0.290	0.279
	(0.175)	(0.007)	(0.123)	(0.008)	(0.007)
Father is a blue–collar worker (0,1)	—	—	0.058	-0.001	0.004
			(0.051)	(0.031)	(0.026)
Father is self–employed (0,1)	—	—	-0.032	-0.041	-0.041
			(0.043)	(0.042)	(0.037)
Father has more than high–school education (0,1)	—	—	-0.209	—	-0.019
			(0.172)		(0.052)
Constant	-0.684	-1.080	-0.909	-1.075	-1.060
	(0.619)	(0.483)	(0.517)	(0.484)	(0.411)
R^2	0.071	0.207	0.148	0.207	0.205
# Observations	1822	1822	1822	1822	1822
Partial R^2 for instrument in 1^{st} stage	0.003	0.114	0.006	0.085	—
F-Test on instrument in 1^{st} stage	5.53	211.2	14.2	189.2	—

Standard errors in parentheses. The sample is taken from the 1986 wave of the German Socio–Economic Panel. The dependent variable is the log of hourly wages. The "father in war" instrument is an indicator that takes value 1 if the father has been involved in WWII. The "father highly ed." instrument takes value 1 if the father has obtained a degree higher than high–school.

5 Matching methods for the estimation of causal effects

Matching methods 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 (i.e. a good instrument).

The central idea of these methods is to base the estimation of treatment effects on a "very careful" matching of cases and controls.

The problem is that this careful matching can take place only on the basis of observables.

Hence, matching methods require the debatable assumption of *selection on observables* (or *unconfoundedness*). Intuitively, this assumption require that:

- 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.

Apparently it sounds like ... assuming away the problem. However, these methods

- offer interesting insights for a better understanding of the problem of the estimation of causal effects;
- the evidence in their favor is compelling (see Lalonde 1986, Dehejia and Wahba 1999 and Smith and Todd 2000).

5.1 Notation and the starting framework

Let:

- i denote a population of N individuals.
- $D_i \in \{0, 1\}$ be the treatment indicator for individual i.
- $Y_i(D_i)$ denote the outcomes for each individual in the two potential treatment situations
  - $Y_i(1)$ is the outcome in case of treatment;
  - $Y_i(0)$ is the outcome in case of no treatment. Hence, the observed outcome for individual i:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{145}$$

- Δ_i be the causal treatment effect for individual i defined as

$$\Delta_i = Y_i(1) - Y_i(0) \tag{146}$$

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

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

$$\tau = E\{\Delta_i | D_i = 1\} = E\{Y_i(1) - Y_i(0) | D_i = 1\} \tag{147}$$

The problem is the usual one: for each individual we do not observe the outcome in his/her counterfactual treatment situation.

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

Is the comparison by treatment status informative?

Let Y_i denote the observed outcome.

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

$$\begin{aligned} E\{Y_i \mid D_i = 1\} &- E\{Y_i \mid D_i = 0\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &\quad + E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ &= \tau + E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \end{aligned} \tag{148}$$

The difference between the left hand side (which we can estimate) and τ is the usual *sample selection bias* due to the fact that those who are not treated are not representative of what would have happened to the treated in the counterfactual situation of no treatment.

To put it differently, the outcome of the treated and the outcome of the non-treated are not identical in the no-treatment situation.

5.2 The case of random assignment to treatment

If assignment to treatment is random in the population, both potential outcomes are independent of the treatment status, i.e.

$$Y(1), Y(0) \quad \perp \quad D \tag{149}$$

where $Y(1)$, $Y(0)$ and D are the vectors of potential oucomes and treatment indicators in the population.

In this case the missing information does not create problems because:

$$E\{Y_i(0)|D_i = 0\} = E\{Y_i(0)|D_i = 1\} = E\{Y_i(0)\} \tag{150}$$

$$E\{Y_i(1)|D_i = 0\} = E\{Y_i(1)|D_i = 1\} = E\{Y_i(1)\} \tag{151}$$

and substituting 150 and 151 in 147 it is immediate to obtain:

$$\begin{aligned} \tau &\equiv E\{\Delta_i \mid D_i = 1\} \\ &\equiv E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &\equiv E\{Y_i(1)|D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 0\} \\ &= E\{Y_i|D_i = 1\} - E\{Y_i|D_i = 0\}. \end{aligned} \tag{152}$$

Randomization ensures that the sample selection bias is zero:

$$E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} = 0 \tag{153}$$

Note that randomization implies that the missing information is "missing completely at random" and for this reason it does not create problems.

If randomization is not possible and natural experiments are not available we need to start from a different set of hypotheses.

5.3 Unconfoundedness and selection on observables

Let X denote a matrix in which each row is a vector of pre-treatment observable variables for individual i.

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

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad X \tag{154}$$

Note that assuming unconfoundedness is equivalent to say that:

- within each cell defined by X treatment is random;
- the selection into treatment depends only on the observables X.

Examples ...

Note that the situation of pure randomization implies a particularly strong version of "unconfoundedness", in which the assignment to treatment is unconfounded independently of pre-treatment variables.

Average effects of treatment on the treated assuming unconfoundedness

If we are willing to assume unconfoundedness:

$$E\{Y_i(0)|D_i=0,X\} = E\{Y_i(0)|D_i=1,X\} = E\{Y_i(0)|X\} \tag{155}$$

$$E\{Y_i(1)|D_i=0,X\} = E\{Y_i(1)|D_i=1,X\} = E\{Y_i(1)|X\} \tag{156}$$

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

$$\begin{aligned} \delta_x &\equiv E\{\Delta_i|X\} \\ &\equiv E\{Y_i(1)-Y_i(0)|X\} \\ &\equiv E\{Y_i(1)|X\} - E\{Y_i(0)|X\} \\ &= E\{Y_i(1)|D_i=1,X\} - E\{Y_i(0)|D_i=0,X\} \\ &= E\{Y_i|D_i=1,X\} - E\{Y_i|D_i=0,X\}. \end{aligned} \tag{157}$$

Using the Law of Iterated expectations, the average effect of treatment on the treated is given by:

$$\begin{aligned} \tau &\equiv E\{\Delta_i|D_i=1\} \\ &= E\{E\{\Delta_i|D_i=1,X\} \mid D_i=1\} \\ &= E\{\quad E\{Y_i|D_i=1,X\} - E\{Y_i|D_i=0,X\} \quad |D_i=1\} \\ &= E\{\delta_x|D_i=1\} \end{aligned} \tag{158}$$

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

5.4 Matching and regression strategies for the estimation of average causal effects

Unconfoundedness suggests the following strategy for the estimation of the average treatment effect defined in equations 157 and 158:

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_i in the population of treated units.

This strategy raises the following questions:

- Is this strategy different from the estimation of a a linear regression of Y on D controlling non parametrically for the full set of main effects and interactions of the covariates X?
- Is this strategy feasible?

Here we are of course assuming that the crucial assumption of unconfoundedness (which raises the most foundamental question) is satisfied.

In which sense matching and regression differ?

The essential difference between regression and matching is the weighting scheme used to take the average of the treatment effects at the different values of the covariates.

Consider a simple example where there is a single binary covariate x and the probability of treatment is positive at each value of x.

If the treatment is unconfounded given x we can write:

$$\begin{aligned}\delta_1 &= E\{Y_i(1) - Y_i(0)|D_i = 1, x_i = 1\} = E\{Y_i(1) - Y_i(0)|x_i = 1\} \quad (159)\\ &= E\{Y_i \mid D_i = 1, x_i = 1\} - E\{Y_i \mid D_i = 0, x_i = 1\}\end{aligned}$$

$$\begin{aligned}\delta_0 &= E\{Y_i(1) - Y_i(0)|D_i = 1, x_i = 0\} = E\{Y_i(1) - Y_i(0)|x_i = 0\} \quad (160)\\ &= E\{Y_i \mid D_i = 1, x_i = 0\} - E\{Y_i \mid D_i = 0, x_i = 0\}\end{aligned}$$

Using matching, the ATT is therefore

$$\begin{aligned}\Delta_M = E\{Y_i(1) &- Y_i(0)|D_i = 1\} \quad (161)\\ &= \delta_0 P(x_i = 0 \mid D_i = 1) + \delta_1 P(x_i = 1 \mid D_i = 1)\\ &= \delta_0 \frac{P(D_i = 1 \mid x_i = 0)P(x_i = 0)}{P(D_i = 1)} + \delta_1 \frac{P(D_i = 1 \mid x_i = 1)P(x_i = 1)}{P(D_i = 1)}\end{aligned}$$

- The weights used by the matching estimator are proportional to the probability of treatment at each value of the covariates.
- Zero weight is given to cells in which the probability of treatment is zero.

Suppose that we estimate instead the (fully saturated) model

$$Y_i = \alpha + \beta x_i + \Delta_r D_i + \epsilon_i. \tag{162}$$

where $E\{\epsilon D\} = E\{\epsilon x\} = 0$, so that

$$\Delta_r = \frac{E\{[D_i - E\{D_i \mid x_i\}]Y_i\}}{E\{[D_i - E\{D_i \mid x_i\}]D_i\}}. \tag{163}$$

By unconfoundedness, Δ_r is free of selection bias.

We can also write that:

$$Y_i = E\{Y_i(0) \mid x_i\} + E\{Y_i(1) - Y_i(0) \mid x_i\}D_i + \epsilon \tag{164}$$

Substitute 159, 160 and 164 into 163, and iterating expectation with respect to x we obtain:

$$\begin{aligned} \Delta_r &= \delta_0 \frac{P(D_i = 1 \mid x_i = 0)[1 - P(D_i = 1 \mid x_i = 0)]P(x_i = 0)}{E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}} \\ &+ \delta_1 \frac{P(D_i = 1 \mid x_i = 1)[1 - P(D_i = 1 \mid x_i = 1)]P(x_i = 1)}{E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}}. \end{aligned} \tag{165}$$

- The weights are proportioal to the variance of treatment status at each value of the covariate.
- Zero weight is given to cells in which the probability of treatment is zero.

Note, in fact, that the variance of treatment given x is

$$(P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]$$

and is highest when the probability of treatment given x is 0.5.

- 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.

Are matching and regression feasible: the dimensionality problem

It is evident, however, that the inclusion in a regression of a full set of non-parametric interactions between all the observables may not be feasible when the sample is small, the set of covariates is large and many of them are multivalued, or, worse, continue.

This dimensionality problem is likely to jeopardize also the matching strategy described by equations 157 and 158:

- With K binary variables the number of cells is 2^K and grows exponentially with K.
- The number of cell increases further if some variables in X take more than two values.
- If the number of cells is very large with respect to the size of the sample it is very easy to encounter situations in which there are:
  - cells containing only treated and/or
  - cells containing only controls.

  Hence, the average treatment effect for these cells cannot be computed.

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 not saturated model is not a solution and may lead to seriously misleading conclusions.
(Figure: linear regression with non-overlapping samples of treated and controls).

5.5 Matching based on the Propensity Score

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

$$p(X) \equiv Pr\{D = 1|X\} = E\{D|X\} \tag{166}$$

The propensity score has two important properties:

Lemma 1 Balancing of pre-treatment variables given the propensity score (Rosenbaum and Rubin, 1983)
If $p(X)$ is the propensity score

$$D \quad \perp \quad X \quad | \quad p(X) \tag{167}$$

Proof:
First:

$$\begin{aligned} Pr\{D = 1|X, p(X)\} &= E\{D|X, p(X)\} \\ &= E\{D|X\} = Pr\{D = 1|X\} \\ &= p(X) \end{aligned} \tag{168}$$

Second:

$$\begin{aligned} Pr\{D = 1|p(X)\} &= E\{D|p(X)\} \\ &= E\{E\{D|X, p(X)\}|p(X)\} = E\{p(X)|p(X)\} \\ &= p(X) \end{aligned} \tag{169}$$

Hence:

$$Pr\{D = 1|X, p(X)\} = Pr\{D = 1|p(X)\} \tag{170}$$

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

Lemma 2 Unconfoundedness given the propensity score (Rosenbaum and Rubin, 1983)
Suppose that assignment to treatment is unconfounded, i.e.

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad X$$

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

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad p(X) \tag{171}$$

Proof: First:

$$\begin{aligned} Pr\{D = 1|Y(1), Y(0), p(X)\} &= E\{D|Y(1), Y(0), p(X)\} \qquad (172)\\ &= E\{E\{D|X, Y(1), Y(0)\}|Y(1), Y(0), p(X)\} \\ &= E\{E\{D|X\}|Y(1), Y(0), p(X)\} \\ &= E\{p(X)|Y(1), Y(0), p(X)\} \\ &= p(X) \end{aligned}$$

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

$$Pr\{D = 1|p(X)\} = p(X) \tag{173}$$

Hence

$$Pr\{D = 1|Y(1), Y(0), p(X)\} = Pr\{D = 1|p(X)\} \tag{174}$$

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

Average effects of treatment and the propensity score

Using the propensity score and its properties we can now match cases and controls on the basis of a monodimensional variable (the propensity score) instead of the multidimensional vector of observables X.

$$E\{Y_i(0)|D_i = 0, p(X_i)\} = E\{Y_i(0)|D_i = 1, p(X_i)\} = E\{Y_i(0)|p(X_i)\} \quad (175)$$

$$E\{Y_i(1)|D_i = 0, p(X_i)\} = E\{Y_i(1)|D_i = 1, p(X_i)\} = E\{Y_i(1)|p(X_i)\} \quad (176)$$

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

$$\begin{aligned} \delta_{p(x)} &\equiv E\{\Delta_i|p(X_i)\} \quad (177)\\ &\equiv E\{Y_i(1) - Y_i(0)|p(X_i)\} \\ &\equiv E\{Y_i(1)|p(X_i)\} - E\{Y_i(0)|p(X_i)\} \\ &= E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\} \\ &= E\{Y_i|D_i = 1, p(X_i)\} - E\{Y_i|D_i = 0, p(X_i)\}. \end{aligned}$$

Using the Law of Iterated expectations, the average effect of treatment on the treated is given by:

$$\begin{aligned} \tau &= E\{\Delta_i|D_i = 1\} \quad (178)\\ &= E\{E\{\Delta_i|D_i = 1, p(X_i)\}|D_i = 1\} \\ &= E\{\quad E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\} \quad |D_i = 1\} \\ &= E\{\delta_{p(x)}|D_i = 1\} \end{aligned}$$

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

5.5.1 Implementation of the estimation strategy

To implement the estimation strategy suggested by the propensity score and its properties 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 177).
- obtain the average of these conditional effects as in equation 178.

This is unfeasible 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 neighbour matching on the Score;
- Radius matching on the Score;
- Kernel matching on the Score;
- Weighting on the basis of the Score.

5.5.2 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:

$$Pr\{D_i = 1|X_i\} = \frac{e^{\lambda h(X_i)}}{1 + e^{\lambda h(X_i)}} \tag{179}$$

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 157 and 158), 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 behavioural interpretation.

An algorithm for estimating 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

Some useful diagnostic tools

As we argued at the beginning of this section, 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, although they would be useful for this analysis as well. (See Dehejia and Wahba, 1999).

5.5.3 Estimation of the treatment effect by Stratification on the Score

This method is based on the same stratification procedure used for estimating the propensity score. By construction, in each stratum the 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

$$\tau_q^S = \frac{\sum_{i \in I(q)} Y_i^T}{N_q^T} - \frac{\sum_{j \in I(q)} Y_j^C}{N_q^C} \tag{180}$$

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 178 is computed with the following formula:

$$\tau^S = \sum_{q=1}^{Q} \tau_q^S \frac{\sum_{i \in I(q)} D_i}{\sum_{\forall i} D_i} \tag{181}$$

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

$$Var(\tau^S) = \frac{1}{N^T} \left[Var(Y_i^T) + \sum_{q=1}^{Q} \frac{N_q^T}{N^T} \frac{N_q^T}{N_q^C} Var(Y_j^C) \right] \tag{182}$$

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

Comments and extensions

- *Irrelevant controls*
  If the goal is to estimate the effect of treatment on the treated the procedure should be applied after having discarded all the controls with a propensity score higher than the maximum or lower than the minimum of the propensity scores of the treated.

- *Penalty for unequal number of treated and controls in a block*
  Note that if there is a block in which the number of controls is smaller than the number of treated, the variance increases and the penalty is larger the larger the fraction of treated in that block. If $N_q^T = N_q^C$ the variance simplifies to:

$$Var(\tau^S) = \frac{1}{N^T}\left[Var(Y_i^T) + Var(Y_j^C)\right] \tag{183}$$

- *Alternatives for the estimation of average outcomes within blocks*
  In the expressions above, the outcome in case of treatment in a block has been estimated as the average outcome of the treated in that block (and similarly for controls).

  Another possibility is to obtain these outcomes as predicted values from the estimation of linear (or more sophisticated) functions of propensity scores.

  The gains from using these more sophisticated techniques do not appear to be large. (See Dehejia and Wahba, 1996.)

5.5.4 Estimation of the treatment effect 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 neighbour 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 for a treated unit, use it again (replacement).
- Drop the unmatched controlled units.
- In the end you should have a sample of N^T pairs of treated and control units. Treated units appear only once while 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 from the score of the treated unit by less than a given tolerance level r chosen by the researcher.
- Allow for replacement of control units.
- When a treated unit has no control within the radius r take the nearest control.
- Drop the unmatched control units.
- In the end you should have a sample of N^T treated unites and N^C control units some of which are used more than once as matches .

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

$$C(i) = \min_j \parallel p_i - p_j \parallel, \tag{184}$$

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

In radius matching,

$$C(i) = \{p_j \mid \parallel p_i - p_j \parallel < r\}, \tag{185}$$

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):

$$\begin{aligned}
\tau^M &= \frac{1}{N^T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C(i)} w_{ij} Y_j^C \right] && (186) \\
&= \frac{1}{N^T} \left[\sum_{i \in T} Y_i^T - \sum_{i \in T} \sum_{j \in C(i)} w_{ij} Y_j^C \right] && (187) \\
&= \frac{1}{N^T} \sum_{i \in T} Y_i^T - \frac{1}{N^T} \sum_{j \in C} w_j Y_j^C && (188)
\end{aligned}$$

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.

$$Var(\tau^M) = \frac{1}{(N^T)^2}\left[\sum_{i\in T} Var(Y_i^T) + \sum_{j\in C}(w_j)^2 Var(Y_j^C)\right] \quad (189)$$

$$= \frac{1}{(N^T)^2}\left[N^T Var(Y_i^T) + \sum_{j\in C}(w_j)^2 Var(Y_j^C)\right] \quad (190)$$

$$= \frac{1}{N^T}Var(Y_i^T) + \frac{1}{(N^T)^2}\sum_{j\in C}(w_j)^2 Var(Y_j^C). \quad (191)$$

Note that there is a penalty for overusing 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

The difference between *attnd.ado* and *attnw.ado* has to do with the programming solutions adopted to compute the weights (see the documentation of the programs).

Estimation of the treatment effect by Kernel matching

The kernel matching estimator can be interpreted as a particular version of the radius method in which every treated unit is matched with a weighted average of all control units with weights that are inversely proportional to the distance between the treated and the control units.

Formally the kernel matching estimator is given by

$$\tau^K = \frac{1}{N^T} \sum_{i \in T} \left\{ Y_i^T - \frac{\sum_{j \in C} Y_j^C G(\frac{p_j - p_i}{h_n})}{\sum_{k \in C} G(\frac{p_k - p_i}{h_n})} \right\} \tag{192}$$

where $G(\dot{})$ is a kernel function and h_n is a bandwidth parameter.

Under standard conditions on the bandwidth and kernel

$$\frac{\sum_{j \in C} Y_j^C G(\frac{p_j - p_i}{h_n})}{\sum_{k \in C} G(\frac{p_k - p_i}{h_n})} \tag{193}$$

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

5.5.5 Estimation of the treatment effect by Weighting on the Score

This method for the estimation of treatment effects is suggested by the following lemma, where the ATE is the average effect of treatment in the population.

Lemma 3 ATE and Weighting on the propensity score
Suppose that assignment to treatment is unconfounded, i.e.

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad X$$

Then

$$\omega = E\{Y_i(1)\} - E\{Y_i(0)\} = E\left\{\frac{Y_i D_i}{p(X_i)}\right\} - E\left\{\frac{Y_i(1-D_i)}{1-p(X_i)}\right\} \tag{194}$$

Proof: Using the law of iterated expectations:

$$E\left\{\frac{Y_i D_i}{p(X_i)}\right\} - E\left\{\frac{Y_i(1-D_i)}{1-p(X_i)}\right\} = E\left\{E\left\{\frac{Y_i D_i}{p(X_i)}|X\right\} - E\left\{\frac{Y_i(1-D_i)}{1-p(X_i)}|X\right\}\right\} \tag{195}$$

which can be rewritten as:

$$E\left\{E\left\{\frac{Y_i(1)}{p(X_i)}|D_i=1,X\right\}Pr\{D_i=1|X\} - E\left\{\frac{Y_i(0)}{1-p(X_i)}|D_i=0,X\right\}Pr\{D_i=0|X\}\right\} \tag{196}$$

Using the definition of propensity score and the fact that unconfoundedness makes the conditioning on the treatment irrelevant in the two internal expectations, this is equal to:

$$E\{E\{Y_i(1)|X\} - E\{Y_i(0)|X\}\} = E\{Y_i(1)\} - E\{Y_i(0)\} \tag{197}$$

QED

Therefore, substituting sample statistics in the RHS of 194 we obtain an estimate of the ATE.

A similar lemma suggests a weighting estimator for the ATT.

Lemma 4 ATT and weighting on the propensity score
Suppose that assignment to treatment is unconfounded, i.e.

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad X$$

Then

$$\begin{aligned} \tau &= \{E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 1\}\} \\ &= E\{Y_i D_i\} - E\left\{Y_i(1 - D_i)\frac{p(X_i)}{1 - p(X_i)}\right\} \end{aligned} \tag{198}$$

Proof: Using the law of iterated expectations:

$$E\{Y_i D_i\} - E\left\{Y_i(1 - D_i)\frac{p(X_i)}{1 - p(X_i)}\right\} = E\left\{E\{Y_i D_i|X\} - E\left\{Y_i(1 - D_i)\frac{p(X_i)}{1 - p(X_i)}|X\right\}\right. \tag{199}$$

which can be rewritten as:

$$E\left\{E\{Y_i(1)|D_i = 1, X\}Pr\{D_i = 1|X\} - E\left\{Y_i(0)\frac{p(X_i)}{1 - p(X_i)}|D_i = 0, X\right\}Pr\{D_i = 0|X\right. \tag{200}$$

Using the definition of propensity score and the fact that unconfoundedness makes the conditioning on the treatment irrelevant in the two internal expectations, this is equal to:

$$\begin{aligned} E\{E\{Y_i(1)|D_i = 1, X\} &- E\{Y_i(0)|D_i = 1, X\}|D_i = 1\} \\ &= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 1\} \end{aligned} \tag{201}$$

where the outer expectation in the first line is over the distribution of $X_i|D_i = 1$.
QED

Substituting sample statistics in the RHS of 198 we obtain an estimate of the ATT. Note the different weighting function with respect to the ATE.

- A potential problem of the weighting method is that it is sensitive to the way the propensity score is estimated.
- The matching and stratification methods are instead not sensitive to the specification of the estimated propensity score.
- An advantage of the weighting method is instead that it does not rely on stratification or matching procedures.
- It is advisable to use all methods and compare them: big differences between them could be the result of
  - mis-specification of the propensity score;
  - failure of the unconfoundedness assumption;
- The computation of the standard error is problematic because the propensity score is estimated. Hirano, Imbens and Ridder (2000) show how to compute the standard error See also Heckman, Ichimura and Todd (1998) and Hahn (1998).

5.6 Recent developments

5.6.1 A panel-asymptotic framework to compare propensity score and covariate matching (Angrist and Hahn, 2000)

Propensity score matching may help to solve the dimensionality problem, but there appear to be no formal statistical theory to justify this method.

Standard asymptotic theory says that as long as $E\{Y \mid X\}$ varies with X efficient estimation should match on X not just on the propensity score.

This paper:

- Provides a framework to resolve the puzzle based on an asymptotic sequence that looks similar to the one used for "panel data": i.e. keeping fixed the number of observations per cells but increasing the number of cells.
- Provides guidelines to decide when propensity score matching is better than covariates matching.
- Proposes a more efficient random-effect-type estimator.

Note, however, that propensity score matching remains the only feasible solution when the dimensionality problem is otherwise unsolvable (e.g. in case of continues covariate).

Assume that:

- Covariates X_i define K possible cells (hence no continues covariate).
- (Y_{0ki}, Y_{1ki}) are the potential outcomes and D_{ki} is the treatment for individual i in cell k, so that

$$Y_{ki} \equiv D_{ki}Y_{1i} + (1 - D_{ki})Y_{0i}$$

  is the observed outcome for the same individual.
- The treatment is ignorable given covariates

$$(Y_{0i}, Y_{1i}) \perp D_i \mid X_i.$$

- The propensity score is constant: $P(D_i = 1 \mid X) = \pi$. Note that this assumption is meant to go to the heart of the comparison between the two methods which has to do with what happens when the score is constant but $E\{Y \mid X\}$ is not.
- Each cell has equal size M.

The model

$$Y_{ki} = \alpha_k + \beta D_{ki} + \epsilon_{ki}, \tag{202}$$

where $k = 1, ..., K$ and $i = 1, ..., M$, is a random effect model.

Note that because of unconfoundedness the individual specific error term ϵ_{ki} is white noise and random effect estimation of this model is consistent.

In this setting:

- covariates matching is equivalent to a panel estimation with fixed effects for each cell.
- propensity score matching is equivalent to an OLS estimation pooling all observations, because the propensity score is fixed and constant across individuals.

However, we know from the theory of panel estimation that none of these two methods is efficient, although they are both consistent.

Efficieny is achieved by a random effect estimator which is a weighted average of between (or pooled) and within estimators.

The paper:

- shows how this efficient random effect estimator can be constructed;
- compares the two inefficient estimators using Panel-asymptotic simulations and Monte Carlo experiments.

The comparison suggests that the relative efficiency of propensity score matching increases:

- if the R^2 of the regression on the covariates X_i decreases;
- if the cell size M falls;
- if the propensity score π falls;

and there are combinations of (π, M, R^2) for which propensity score matching is more efficient.

5.7 Comments on matching methods.

- The validity of matching methods depend on the quality of the observable covariates on which the matching can be constructed.

- It is crucial to be able to control in a convincing way for the pretreatment history of the units under studies.

- Matching methods should be wiewed as a *bias reducing* strategy.

- Matching methods offer also a wide range of useful self diagnostic tools.

- Propensity score matching is "philosophically" not different from standard matching, but is crucial to to solve the dimensionality problem.

The debate between the "Quasi-Experimental" and the "Non-Experimental" approaches to the estimation of causal effects is still open. Heckman and Hotz (1989) and the comments by Holland and Moffit in the same JASA issue present the terms of the debate in a very clear way.

- Causal inference in non-randomized studies requires more *data* than in randomized studies.
- Causal inference in non-randomized studies requires more *assumptions* than in randomized studies.

Is the future featuring more "cooperation" instead of "contrapposition" between approaches?

- An interesting example: the Difference in Difference Matching Estimator of Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998) and Smith and Todd (2000).

6 Appendix

6.1 Standard characteriziation of IV

Consider the model

$$Y = \alpha + \Delta D + \epsilon \tag{203}$$

in which $E\{\epsilon\} = 0$ but $COV\{\epsilon, D\} \neq 0$. In this situation,

$$\text{plim}\{\hat{\Delta}_{OLS}\} = \frac{COV\{Y, D\}}{V\{D\}} = \Delta + \frac{COV\{\epsilon, D\}}{V\{D\}} \neq \Delta \tag{204}$$

and OLS gives an inconsistent estimate of Δ.

Consider a variable Z such that:

$$E\{D \mid Z\} \neq 0 \;\Rightarrow\; COV\{Z, D\} \neq 0 \tag{205}$$

$$E\{\epsilon \mid Z\} = 0 \;\Rightarrow\; COV\{Z, \epsilon\} = 0. \tag{206}$$

If this variable exists, the following population equation holds (see also the Appendix 6.2 in the next page):

$$\frac{COV\{Y, Z\}}{COV\{D, Z\}} = \Delta + \frac{COV\{\epsilon, Z\}}{COV\{D, Z\}} = \Delta = \text{plim}\{\hat{\Delta}_{IV}\} \tag{207}$$

Substituting the appropriate sample covariances on the LHS of 207 we get the consistent estimator $\hat{\Delta}_{IV}$.

Examples:

- Estimation of supply and demand.
- Other simultaneous equations models.
- Omitted variables.
- Measurement error
- ...

The problem is to find the variable z.

6.2 Derivation of the IV-2SLS estimator in matrix notation

Consider the following model

$$\begin{aligned} Y &= D\Delta + \epsilon \quad (208) \\ D &= Z\gamma + u \quad (209) \end{aligned}$$

where D and Z are conformables matrices which include constant terms and $COV\{D, \epsilon\} \neq 0$ and $COV\{Z, \epsilon\} = COV\{Z, U\} = 0$.

Note that

$$\hat{D} = Z(Z'Z)^{-1}Z'D = P_Z D \quad (210)$$

is the predicted value of D given Z, where $P_Z = Z(Z'Z)^{-1}Z'$ is the corresponding projection matrix.

OLS estimation of the transformed equation

$$P_Z Y = P_Z D\Delta + P_Z \epsilon \quad (211)$$

gives

$$\begin{aligned} \hat{\Delta} &= (D'P_Z P_Z D)^{-1} D'P_Z P_Z Y \quad (212) \\ &= (D'P_Z D)^{-1} D'P_Z Y \\ &= (D'Z)^{-1} Z'Y \rightarrow \frac{COV\{Y, Z\}}{COV\{D, Z\}} \end{aligned}$$

which is the IV estimator.

6.3 Equivalence between IV and Wald estimators

Consider the setup of Section 2 in which the outcome is Y_i and the treatment is binary: $D_i = 0, 1$. Suppose also that the instrument is binary as well: $Z_i = 0, 1$. It can be easily checked (see next page) that:

$$\frac{COV\{Y, Z\}}{COV\{D, Z\}} = \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i = 1 \mid Z_i = 1\} - Pr\{D_i = 1 \mid Z_i = 0\}} \tag{213}$$

The RHS of 213 is also known as the *Wald estimator* (see Angrist, 1990) that is constructed on the basis of expectations of outcomes taken conditioning on different realizations of the instrument. Here is another way to derive it.

Suppose that we are trying to estimate $\Delta^* = E\{\Delta_i\}$ in equation 23 which is reported here for convenience

$$Y_i = \mu(0) + E\{\Delta_i\}D_i + \epsilon_i.$$

We can take the following two conditional expectations:

$$\begin{aligned} E\{Y_i \mid Z_i = 1\} &= \mu(0) + \Delta^* E\{D_i \mid Z_i = 1\} + E\{\epsilon_i \mid Z_i = 1\} \quad (214) \\ E\{Y_i \mid Z_i = 0\} &= \mu(0) + \Delta^* E\{D_i \mid Z_i = 0\} + E\{\epsilon_i \mid Z_i = 0\} \quad (215) \end{aligned}$$

Assuming that the instrument Z satisfies the condition 206, so that the conditional expectations of the errors are zero:

$$\begin{aligned} E\{Y_i \mid Z_i = 1\} &= \mu(0) + \Delta^* Pr\{D_i = 1 \mid Z_i = 1\} \quad (216) \\ E\{Y_i \mid Z_i = 0\} &= \mu(0) + \Delta^* Pr\{D_i = 1 \mid Z_i = 0\} \quad (217) \end{aligned}$$

Subtracting 217 from 216 and solving for Δ^* gives the Wald-IV estimator on the RHS of 213.

A formal proof of the result of the previous page follows.

$$\Delta_W = \frac{E\{Y \mid Z=1\} - E\{Y \mid Z=0\}}{Pr\{D=1 \mid Z=1\} - Pr\{D=1 \mid Z=0\}} = \textit{Wald estimator}$$

$$\Delta_{IV} = \frac{COV\ \{Y,Z\}}{COV\ \{D,Z\}} = \frac{E\{YZ\} - E\{Y\}E\{Z\}}{E\{DZ\} - E\{D\}E\{Z\}} = \textit{IV estimator} =$$

$$= \frac{E\{Y \mid Z=1\}\, Pr\{Z=1\} - E\{Y\} Pr\{Z=1\}}{Pr\{D=1, Z=1\} - Pr\{D=1\} Pr\{Z=1\}}$$

$$= Pr\{Z=1\}\, \frac{E\{Y \mid Z=1\} - E\{Y \mid Z=1\} Pr\{Z=1\} - E\{Y \mid Z=0\} Pr\{Z=0\}}{Pr\{D=1, Z=1\} - [Pr\{D=1, Z=1\} + Pr\{D=1, Z=0\}]\, Pr\{Z=1\}}$$

$$= Pr\{Z=1\}\, \frac{E\{Y \mid Z=1\}\,[1 - Pr\{Z=1\}] - E\{Y \mid Z=0\} Pr\{Z=0\}}{Pr\{D=1, Z=1\}\,[1 - Pr\{Z=1\}] - Pr\{D=1, Z=0\}\, Pr\{Z=1\}}$$

$$= Pr\{Z=1\}\, \frac{Pr\{Z=0\}\,[E\{Y \mid Z=1\} - E\{Y \mid Z=0\}]}{Pr\{D=1 \mid Z=1\}\, Pr\{Z=1\}\, Pr\{Z=0\} - Pr\{D=1 \mid Z=0\}\, Pr\{Z=0\} Pr\{Z=1\}}$$

$$= \frac{E\{Y \mid Z=1\} - E\{Y \mid Z=0\}}{Pr\{D=1 \mid Z=1\} - Pr\{D=1 \mid Z=0\}} = \Delta_W$$

Q.E.D.

7 References

Abadie, Alberto, Angrist, Joshua D. and Imbens, Guido (2000), "Instrumental Variables Estimates of the effect of Subsidized Training on the Quantile of Trainee Earnings", mimeo.

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 (1999), "When to Control for Covariates? Panel–Asymptotic Results for Estimates of Treatment Effects", *NBER Technical Working Paper* 241.

Angrist, Joshua D. and Jinyong Hahn (revised March 2001), "When to Control for Covariates? Panel–Asymptotic Results for Estimates of Treatment Effects", mimeo.

Angrist, Joshua D. and Krueger, Alan B. (1999), "Empirical Strategies in Laobr 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. (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. 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.

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

Ashenfelter, Orley (1997), "Estimating the Effect of Training Programs on Earnings", *The Review of Economics and Statistics* ??.

Ashenfelter, Orley and David Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *The Review of Economics and Statistics* ??.

Ashenfelter, Orley and Alan B. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins", *American Economic Review*.

Ashenfelter, Orley and Cecilia Rouse (1999), "The Payoff to Education", European Summer Symposium in Labour Economics, CEPR–IZA, Ammersee 14–18/9/1999, Draft.

Ashenfelter, Orley, Colm Harmon and Hessel Oosterbeek (1999), "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias", Draft 6.03 October 3rd.

Becker, Gary S. (1975), "Human Capital and the Personal Distribution of Income: An Analytical Approach" (W.S. Woytinsky Lecture), in: Gary S. Becker, *Human Capital*, 2nd edition. New York: Columbia University Press.

Dale, Stacy Berg and Alan B. Krueger "Estimating the Payoff to Attending a more Selective College: An Application of Selection on Observables and Unobservables", *NBER Working Paper* 7322.

Blundell, R. and M. Costa Dias (2000), "Evaluation Methods for Non-Experimental Data", mimeo.

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.

Bowden, Roger J. and Darrell A. Turkington (1984), *Instrumental Variables*. Econometric Society Monographs in Quantitative Economics. Cambridge University Press.

Card, David (1995a), "Earnings, Schooling, and Ability Revisited", *Research in Labor Economics* 14, 23–48.

Card, David (1995b), "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.

Card, David (2000), "Estimating the return to schooling: pregress on some persistent econometric problems" NBER WP 7769.

Carrasco R. (1998), "Binary Choice with Binary Endogenous Regressors in Panel Data: Estimating the Effect of Fertility on Female Labour Participation" *CEMFI Working Paper* 9805.

Cawlwy J., Heckman J. and Vytlacil E.(2001), "Three observations on wages and measured cognitive ability" *Labour Economics* Vol.8, Issue 4, September 2001.

Cox, D.R. (1992), "Causality: Some Statistical Aspects" *J.R. Statist. Soc.*, 2, 291–301.

Dawid, A.P. (1997), "Causal Inference Without Counterfactuals", University College London, Dept. of Statistical Science, Research Report No. 188.

Dehejia, R.H. and S. Wahba (1996), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", Harvard University, Mimeo.

Dehejia, R.H. and S. Wahba (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies", NBER WP 6829.

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.

Fisher, R.A. (19??), "The Principles of Experimentation, Illustrated by a Psycho–Physical Experiment", ch. 2 of: R. Fisher, The Design of Experiments.

Frhlich, M. Heshmati, A. and Lechner, M. (2000), "A Microeconomic Evaluation of Rehabilitation of Long-term Sickness in Sweden". discussion paper 2000-4, Department of Economics, niversity of St. Gallen, mimeo.

Griliches, Zvi (1977), "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica* 45, 1–21.

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.

Harmon, C. and I. Walker (1995), "Estimates of the Economic Returns to Schooling for the United Kingdom", *American Economic Review*.

Harmon, C. and I. Walker (1997), "The Marginal and Average Returns to Schooling", mimeo.

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. (1999a), "Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programs", *NBER Working Paper* 7230.

Heckman, James J. (1999b), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *NBER Working Paper* 7333.

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. 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", *NBER Working Paper* 6983.

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. 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.

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", European Summer Symposium in Labour Economics, CEPR–IZA, Ammersee 14–18/9/1999, Draft.

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", mimeo.

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.

Ichimura H. and C. Taber (2000), "Direct Estimation of Policy Impacts", mimeo.

Ichino, Andrea and Rudolf Winter–Ebmer (2001), "The Long Run Educational Cost of World War II: An Example of Local Average Treatment Effect Estimation", mimeo eui.

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.

Imbens, G.W. (1999), "The Role of Propensity Score in Estimating Dose–Response Functions", *NBER Technical Working Paper* 237.

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 (1997a), " Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance", *The Annals of Statistics* 25, 305–327.

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

Imbens, Guido W. and Donald B. Rubin (2000), "Unconfounded assignment", in *Notes on Unconfoundedness* ch. 9.

Kane, Thomas J., Cecilia E. Rouse and Douglas Staiger (1999), "Estimating Returns to Schooling when Schooling is Misreported", NBER Working Paper # 7235.

Kling, Jeffrey R. (1998), "Identifying Causal Effects of Public Policies", Ph.D. thesis, MIT.

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

Lechner, Michael and Pfeiffer, Friedhelm(2001), "Economic Evaluation of Labour Market Policies", ZEW Economic Studies 13, Physica-Verlag.

Lechner, Michael (2000), "Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods", mimeo.

Lechner, Michael (2000), "A note on the common support problem in applied evaluation studies", mimeo.

Lechner, Michael (2000), "Identification and estimation of causal treatments under the conditional independence assumption", mimeo.

Maddala, G.S. (1986), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Meghir, Costas and Marten Palme (1999), "Assessing the Effect of Schooling on Earnings Using a Social Experiment", mimeo.

Meyer, Bruce D. (1994), "Natural and Quasi-Experiments in Economics", *NBER Technical Working Paper*, 170.

Mincer, J. (1974), *Schooling, Experience and Earnings*, Columbia University Press, New York.

Mercatanti, A. (1999), *Inferenza causale in presenza di non-compliance e dati mancanti: aspetti metodologici e applicativi* , Doctoral Thesis, Dipartimento di Statistica, Università di Firenze.

Mooney, Christopher Z. and Duval, Robert D., "Bootstrapping, a Nonparametric Approach to Statistical Inference", Series: Quantitative Applications in Social Sciences, a Sage University Paper 95, 1993.

Pearl, Judea (1995), "On the Testability of Causal Modelswith Latent and Instrumental Variables". In: P. Besnard and S. Hanks (eds.), *Uncertainty in Artificial Intelligence II*. San Francisco, CA: Morgan Kaufmann Publishers.

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

Persson, T., G. Tabellini and F. Trebbi (2000), "Electoral Rules and Corruption", mimeo.

Persson, T. (2001), "Currency Unions and trade: How Large is the Treatment Effect?"" mimeo.

Psacharopoulos, George (1994), "Returns to Investment in Education: A Global Update", *World Development* 22, 1325–1343.

Psacharopoulos, George (1985), "Returns to Education: A Further International Update and Implications", *Journal of Human Resources* XX, 583–604.

Rettore, Enrico (1997), "The Impact of Training Programs on Duration of the First Job-Search Spell", mimeo.

Riccio, James A. (2000), "Extending the Reach of Randomized Social Experiments: New Directions in Evaluations of Americn Welfare–to–work and Employment Initiatives", mimeo.

Robinson (1989a), "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models", *Journal of Political Economy*.

Robinson (1989b), "Union Endogeneity and Self Selection", *Journal of Labor Economics*.

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.

Rubin, D.B. (????), "Statistical Inference for Causal Effects in Epidemiological Studies via Potential Outcomes", mimeo.

Schmidt, Christophe M. (1999), "Knowing what Works: The Case for Rigorous Program Evaluation", *IZA Discussion Paper Series* 88.

Speed, T.P. (1990), "Introductory Remarks on Neyman (1923)", *Statistical Science*, 5, 4, 463–464.

Splawa–Neyman, J. (1990), "On the Application of Probability Theory to Agricultual Economics. Essay on Principles. Section 9", *Statistical Science* 5, 4, 465–480.

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

Willis, R., "Wage Determinants: A Survey and Reinterpretation of Human Capital Earning Functions", in: Orley Ashenfelter and R. Layard, *Handbook of Labor Economics* ch. 10. Amsterdam: North–Holland, 1986.

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