Methods to Estimate Causal Effects
Theory and Applications

Prof. Dr. Sascha O. Becker
U Stirling, Ifo, CESifo and IZA

last update: 21 August 2009

Preliminaries

Address

Prof. Dr. Sascha O. Becker
Stirling Management School
Division of Economics
University of Stirling
Stirling FK9 4LA
United Kingdom
Tel: +44 (1786) 46-7278
Fax: +44 (1786) 46-7469
email: `sascha.becker@stir.ac.uk`
http://www.sobecker.de

The main aim of this course is to provide an introduction to/review of the fundamental theoretical concepts and applications of modern econometric techniques used in empirical social sciences.
In addition to these lecture notes, the following textbooks are suggested as further reference:

- Angrist, Joshua D. and Jörn-Steffen Pischke (2009) Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.
- Cameron, Colin A. and Pravin K. Trivedi (2005) Microeconometrics: Methods and Application. Cambridge University Press, New York, 1st edition.
- Stock, James H. and Mark W. Watson (2007) Introduction to Econometrics, Pearson Education; 2nd edition
- Wooldridge, Jeffrey (2002) Econometric Analysis of Cross Section and Panel Data. The MIT Press, 1st edition.
- Wooldridge, Jeffrey (2003) Introductory Econometrics : A Modern Approach. South Western College Publishing, 2nd edition.

Additional references are listed at the end of these notes.

Contents

Chapter 1

A summary of OLS and IV estimation

Before entering the main lecture material, it is useful to recall the assumptions underlying ordinary least squares (OLS).

Also, these lecture notes assume familiarity with instrumental-variables (IV) estimation which will be important in great parts of these lectures.

The background material on OLS and IV is nicely treated in James H. Stock und Mark W. Watson, Introduction to Econometrics (2nd edition, Boston etc.: Pearson 2007).

Chapter 2

The Problem of Causality

> Causal parameters are easy to describe but hard to measure. (Angrist (2004), p.C55)

> ... statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings (Freedman (1991))

> Good econometrics cannot save a shaky research agenda, but the promiscuous use of fancy econometric techniques sometimes brings down a good one.(Angrist and Pischke (2009))

Parts of these lecture notes are based on Ichino (2006) and are used with his kind permission. More background reading is in

- Angrist and Krueger (2001) give a non-technical summary
- Wooldridge (2002), chapter 18, gives a textbook treatment of the issues involved

2.1 Motivation

Consider the following questions

- Does smoking cause lung cancer?
- Does aspirin reduce the risk of heart attacks?
- Does an additional year of schooling increase future earnings?
- Are temporary jobs a stepping stone to permanent employment?
- Does EPL increase unemployment?

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

We need a framework to study causality.

2.2 A formal framework to think about causality

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

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

In general no, because of:

- confounding factors;
- reverse causality.

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

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

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

- D_i is the *treatment* status:

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

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

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

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

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

The observed outcome for each unit can be written as:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{2.1}$$

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

2.3 The fundamental problem of causal inference

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

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

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

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

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

2.4 The statistical solution

Statistics proposes to approach the problem by focusing on the average causal effect for the entire population or for some interesting subgroups.

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

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1) - Y_i(0)\} \\ &= E\{Y_i(1)\} - E\{Y_i(0)\} \end{aligned} \tag{2.2}$$

The effect of treatment on the treated (ATT):

$$\begin{aligned} E\{\Delta_i \mid D_i = 1\} &= E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \end{aligned} \tag{2.3}$$

Are these effects interesting from the viewpoint of an economist?

Is this a progress towards the solution of the Fundamental Problem of Causality?

Is the comparison by treatment status informative?

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

$$\begin{aligned} E\{Y_i \mid D_i = 1\} \; - \; & E\{Y_i \mid D_i = 0\} \\ = \; & E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ = \; & E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ & + E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ = \; & \tau + E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \end{aligned} \tag{2.4}$$

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

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

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

2.5 Randomized experiments

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

$$E\{Y_i(0)|i \in C\} = E\{Y_i(0)|i \in T\} = E\{Y_i(0)\} \tag{2.5}$$

and

$$E\{Y_i(1)|i \in C\} = E\{Y_i(1)|i \in T\} = E\{Y_i(1)\}. \tag{2.6}$$

Substituting 2.5 and 2.6 in 2.2 it is immediate to obtain:

$$\begin{aligned} E\{\Delta_i\} &\equiv E\{Y_i(1)\} - E\{Y_i(0)\} \qquad (2.7) \\ &= E\{Y_i(1)|i \in T\} - E\{Y_i(0)|i \in C\}. \end{aligned}$$

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

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

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

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

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

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

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

Chapter 3

Conventional methods to estimate causal effects

This part of the course is devoted to conventional methods to estimate causal effects.

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 (1978) “two stages” estimation of the “dummy endogenous variables model”;

3.1 Specification of the outcomes

Going back to the notation of Section 2, consider the following specification of outcomes, with or without treatment:

$$\begin{aligned} Y_i(1) &= \mu(1) + U_i(1) \\ Y_i(0) &= \mu(0) + U_i(0) \end{aligned} \tag{3.1}$$

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

$$\begin{aligned} \Delta_i &= Y_i(1) - Y_i(0) \\ &= [\mu(1) - \mu(0)] + [U_i(1) - U_i(0)] \\ &= E\{\Delta_i\} + [U_i(1) - U_i(0)]. \end{aligned} \tag{3.2}$$

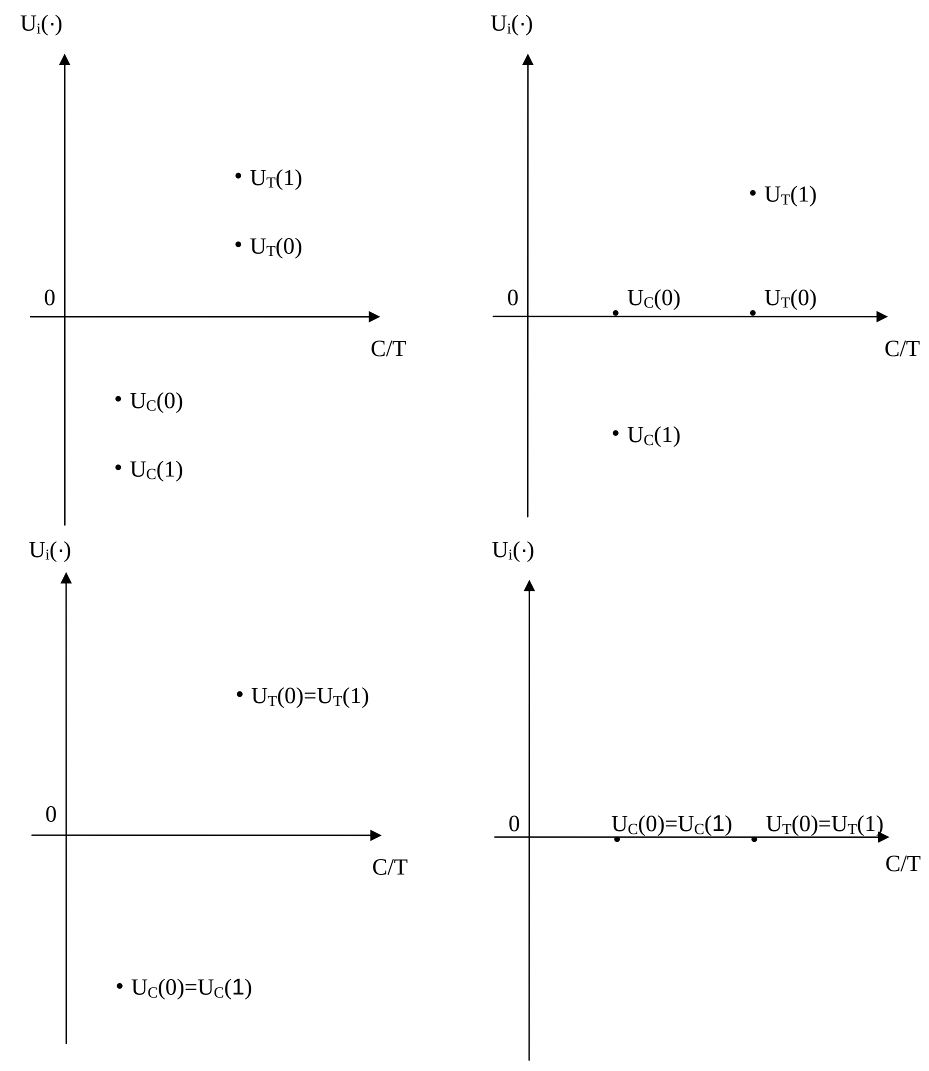
It is the sum of:

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

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

(Figure: Differences between treated and control individuals.)

Figure 3.1: Differences between treated and control individuals.


Unobservable Outcomes for Treated and Controls

C = Controls ➡ Not Treated
T = Treated

Let D_i indicate treatment: using equation 2.1 the outcome can be written as:

$$\begin{aligned} Y_i &= \mu(0) + [\mu(1) - \mu(0) + U_i(1) - U_i(0)]D_i + U_i(0) \qquad (3.3) \\ &= \mu(0) + \Delta_i D_i + U_i(0) \end{aligned}$$

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.

3.2 Specification of the selection into treatment

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

$$D_i^* = \alpha + \beta Z_i + V_i \tag{3.4}$$

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

$$D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{cases} \tag{3.5}$$

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.

3.3 The model in compact form

$$Y_i = \mu(0) + \Delta_i D_i + U_i(0) \tag{3.6}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{3.7}$$

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{array} \right\} \tag{3.8}$$

$$\begin{aligned} \Delta_i &= \mu(1) - \mu(0) + U_i(1) - U_i(0) \qquad (3.9) \\ &= E\{\Delta_i\} + U_i(1) - U_i(0) \end{aligned}$$

$$E\{U_i(1)\} = E\{U_i(0)\} = E\{V_i\} = 0 \tag{3.10}$$

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.

3.4 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:

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

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1) - Y_i(0)\} \\ &= E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= \mu(1) - \mu(0) \end{aligned} \tag{3.11}$$

2. *The effect of treatment on the treated*

$$\begin{aligned} E\{\Delta_i \mid D_i = 1\} &= E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &= \mu(1) - \mu(0) + E\{U_i(1) - U_i(0) \mid D_i = 1\} \end{aligned} \tag{3.12}$$

The two effects differ because of the term

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} \tag{3.13}$$

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 are these two treatment effects equal?

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

   $$U_i(1) = U_i(0) \qquad \forall i \tag{3.14}$$

   In this case, the model has **constant coefficients** because

   $$\Delta_i = E\{\Delta_i\} = \mu(1) - \mu(0) \qquad \forall i. \tag{3.15}$$

   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.

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

   $$E\{U_i(1) - U_i(0) \mid D_i = 1\} = E\{U_i(1) - U_i(0)\} = 0 \tag{3.16}$$

   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
- ...

3.5 Problems with OLS estimation

3.5.1 Bias for the effect of treatment on a random person

Using 3.9 we can rewrite equation 3.6 as:

$$\begin{aligned} Y_i &= \mu(0) + E\{\Delta_i\}D_i + U_i(0) + D_i[U_i(1) - U_i(0)] \\ &= \mu(0) + E\{\Delta_i\}D_i + \epsilon_i \end{aligned} \tag{3.17}$$

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

Problem:

$$\begin{aligned} E\{\epsilon_i D_i\} &= E\{\epsilon_i D_i | D_i = 1\}Pr\{D_i = 1\} + E\{\epsilon_i D_i | D_i = 0\}Pr\{D_i = 0\} \\ &= E\{U_i(1) \mid D_i = 1\}Pr\{D_i = 1\} \neq 0 \end{aligned} \tag{3.18}$$

using the *law of iterated expectations.*

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

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i\} + \tag{3.19}$$

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

The second line in 3.19 represents the OLS regression bias if we want to estimate the effect of treatment on a random person.

Readjusting the second line of 3.19, the bias in the estimation of $E\{\Delta_i\}$ can be written in the following form:

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i\} + \tag{3.20}$$

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

This bias is equal to the difference between two components:

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

$$\begin{aligned} E\{U_i(1)\} &= E\{U_i(1) \mid D_i = 1\} = 0 \\ E\{U_i(0)\} &= E\{U_i(0) \mid D_i = 0\} = 0 \end{aligned} \tag{3.21}$$

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

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ &= \mu(1) - \mu(0) \end{aligned} \tag{3.22}$$

Examples:

- Cancer

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

3.5.2 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 3.17 and remembering from 3.12 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 3.17 as:

$$\begin{aligned} Y_i &= \mu(0) + E\{\Delta_i \mid D = 1\}D_i + \\ &\quad U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D = 1\}] \\ &= \mu(0) + E\{\Delta_i \mid D_i = 1\}D_i + \eta_i \end{aligned} \tag{3.23}$$

Using 3.23 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:

$$\begin{aligned} E\{\eta_i D_i\} &= E\{D_i U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D = 1\}]\} \\ &= E\{D_i U_i(0)\} \neq 0. \end{aligned} \tag{3.24}$$

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

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i \mid D_i = 1\} + \tag{3.25}$$

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

The second line in 3.25 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?

It would help, but...

Examples:

- Cancer
- Education
- Training
- ...

3.5.3 An important particular case: the Roy (1951) 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:

$$\begin{aligned} Pr\{D_i = 1 \mid U_i(1) - U_i(0)\} &\neq Pr\{D_i = 1\} \qquad (3.26) \\ \text{or equiv.} \qquad E\{D_i \mid U_i(1) - U_i(0)\} &\neq E\{D_i\} \end{aligned}$$

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

$$\begin{aligned} f(U_i(1) - U_i(0) \mid D_i = 1) Pr\{D_i = 1\} &= \qquad (3.27) \\ Pr\{D_i = 1 \mid U_i(1) - U_i(0)\} f(U_i(1) - U_i(0)) & \end{aligned}$$

Because of 3.26, from 3.27 descends that

$$f(U_i(1) - U_i(0) \mid D_i = 1) \neq f(U_i(1) - U_i(0)) \qquad (3.28)$$

and therefore that

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} \neq E\{(U_i(1) - U_i(0)\} \qquad (3.29)$$

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 3.16);
- OLS gives seriously biased estimates of the effect on a random person (see equation 3.19);
- OLS appears 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 3.25).

3.6 Conventional interpretation of Instrumental Variables

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

We want to estimate equation 3.17, which is reported here for convenience

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

Suppose that there exists a variable Z such that:

$$COV\{Z, D\} \neq 0 \tag{3.30}$$

$$COV\{Z, \epsilon\} = 0. \tag{3.31}$$

If this variable exists then (see the Appendix 4.11.1):

$$E\{\Delta_i\} = \frac{COV\{Y, Z\}}{COV\{D, Z\}}. \tag{3.32}$$

Substituting the appropriate sample covariances on the RHS of 3.32 we get a consistent estimate of $E\{\Delta_i\}$.

It is however crucial to understand what the two conditions 3.30 and 3.31 require in terms of our model.

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

$$Pr\{D_i = 1 \mid Z_i = 1\} \neq Pr\{D_i = 1 \mid Z_i = 0\} \tag{3.33}$$

This condition can be easily tested by estimating the participation equation 3.7 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 3.7 (in contrast with the Heckman two step procedure that we will consider later).

The second condition is more problematic:

$$E\{\epsilon_i \mid Z_i\} = E\{U_i(0) + D_i[U_i(1) - U_i(0)] \mid Z_i\} = 0 \tag{3.34}$$

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

Note that it contains two requirements:

1. 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.

$$E\{U_i(0) \mid Z_i\} = 0 = E\{U_i(0)\} \tag{3.35}$$

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

$$\begin{aligned} E\{D_i[U_i(1) - U_i(0)] \mid Z_i\} &= E\{U_i(1) - U_i(0) \mid Z_i, D_i = 1\}Pr\{D_i = 1 \mid Z_i\} \\ &= 0 = E\{U_i(1) - U_i(0)\} \end{aligned} \tag{3.36}$$

   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.

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.

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

We want now to estimate equation 3.23, 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 3.30 and 3.31 hold in this case:

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

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

If this variable exists then (see the Appendix 4.11.1):

$$E\{\Delta_i \mid D_i = 1\} = \frac{COV\{Y, Z\}}{COV\{D, Z\}}. \tag{3.37}$$

Substituting the appropriate sample covariances on the LHS of 3.37 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 3.30 and 3.31 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:

$$E\{D_i \mid Z_i\} = Pr\{D_i = 1 \mid Z_i\} \neq 0 \tag{3.38}$$

This condition can be easily tested by estimating the participation equation 3.7 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 3.7 (in contrast with the Heckman procedure that we will consider later).

The second condition is different but still problematic:

$$E\{\eta \mid Z\} = E\{U_i(0)+D_i[U_i(1)-U_i(0)-E\{U_i(1)-U_i(0) \mid D = 1\}] \mid Z_i\} = 0 \tag{3.39}$$

There are again two requirements:

1. 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).

$$E\{U_i(0) \mid Z_i\} = 0 = E\{U_i(0)\} \tag{3.40}$$

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

$$E\{U_i(1)-U_i(0) \mid Z_i, D_i = 1\} = E\{U_i(1)-U_i(0) \mid D_i = 1\} \tag{3.41}$$

   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.

3.6.3 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 3.41 would not be satisfied.

This intuition leads to the concept of **Local Average Treatment Effect (LATE)** 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**.

3.7 Heckman (1978) procedure for endogenous dummy variable models

3.7.1 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 3.6) simplifies to the following **common coefficients** model:

$$\begin{aligned} Y_i &= \mu(0) + \gamma X_i + \Delta D_i + U_i(0) \\ Y_i &= \mu + \gamma X_i + \Delta D_i + U_i \end{aligned} \tag{3.42}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{3.43}$$

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{array} \right\} \tag{3.44}$$

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 3.42 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 vice versa).

3.7.2 The model rewritten as a switching regression model

We can rewrite the model in the following way:

$$\begin{aligned} &\text{Regime 1: if } D_i^* \geq 0 \qquad Y_i &&= \mu + \gamma X_i + \Delta + U_i \qquad (3.45) \\ &\text{Regime 0: if } D_i^* < 0 \qquad Y_i &&= \mu + \gamma X_i + U_i \qquad (3.46) \end{aligned}$$

or equivalently

$$\begin{aligned} &\text{Regime 1: if } V_i \geq -\alpha - \beta Z_i \qquad Y_i &&= \mu + \gamma X_i + \Delta + U_i \qquad (3.47) \\ &\text{Regime 0: if } V_i < -\alpha - \beta Z_i \qquad Y_i &&= \mu + \gamma X_i + U_i \qquad (3.48) \end{aligned}$$

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.

$$\begin{aligned} &\text{Regime 1:} \qquad E\{U_i \mid V_i \geq -\alpha - \beta Z_i\} \neq E\{U_i\} = 0 \qquad (3.49) \\ &\text{Regime 0:} \qquad E\{U_i \mid V_i < -\alpha - \beta Z_i\} \neq E\{U_i\} = 0 \qquad (3.50) \end{aligned}$$

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 expectations in 3.49 and 3.50, we could include them in the two regressions and avoid the mis-specification.

3.7.3 Some useful results on truncated normal

Assume that U and V are jointly normally distributed with zero means, variances 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).

$$\begin{aligned}
E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} > k_1\right\} &= \frac{\phi(k_1)}{1-\Phi(k_1)} && (3.51)\\
E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} < k_2\right\} &= -\frac{\phi(k_2)}{\Phi(k_2)} && (3.52)\\
E\left\{\frac{U}{\sigma_U} \mid k_1 < \frac{U}{\sigma_U} < k_2\right\} &= \frac{\phi(k_1)-\phi(k_2)}{\Phi(k_2)-\Phi(k_1)} && (3.53)
\end{aligned}$$

and similarly for V. The ratios between the normal density and its cumulative on the RHS are called *Mills ratios.*

$$\begin{aligned}
E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} > k\right\} &= \sigma_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > k\right\} && (3.54)\\
&= \sigma_{UV}\frac{\phi(k)}{1-\Phi(k)}\\
E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} < k\right\} &= \sigma_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} < k\right\} && (3.55)\\
&= -\sigma_{UV}\frac{\phi(k)}{\Phi(k)}
\end{aligned}$$

3.7.4 The Heckman (1978) two-steps procedure

We cannot observe $E\{U_i \mid V_i \geq -\alpha - \delta Z_i\}$ and $E\{U_i \mid V_i < -\alpha - \delta Z_i\}$ but we can estimate them using the participation equation 3.43 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

1. Estimate a probit model for the participation into treatment using 3.43, and retrieve the (consistently) estimated absolute values of the *Mills Ratios*

$$M_{1i} = \frac{\phi(-\hat{\alpha} - \hat{\beta} Z_i)}{1 - \Phi(-\hat{\alpha} - \hat{\beta} Z_i)} = \frac{\phi(\hat{\alpha} + \hat{\beta} Z_i)}{\Phi(\hat{\alpha} + \hat{\beta} Z_i)} \qquad (3.56)$$

$$M_{0i} = \frac{\phi(-\hat{\alpha} - \hat{\beta} Z_i)}{\Phi(-\hat{\alpha} - \hat{\beta} Z_i)} = \frac{\phi(\hat{\alpha} + \hat{\beta} Z_i)}{1 - \Phi(\hat{\alpha} + \hat{\beta} Z_i)} \qquad (3.57)$$

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

2. Estimate using OLS the equations for the two regimes augmented with the appropriate *Mills Ratios* obtained in the first step

$$\text{Regime 1:} \qquad Y_i = \mu + \gamma X_i + \Delta + \lambda_1 M_{1i} + \nu_i \qquad (3.58)$$

$$\text{Regime 0:} \qquad Y_i = \mu + \gamma X_i + \lambda_0 M_{0i} + \nu_i \qquad (3.59)$$

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

3. Get a consistent estimate of the treatment effect Δ by subtracting the estimated constant in 3.59 from the estimated constant in 3.58.

3.7.5 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 treatment) we should expect a positive coefficient on the *Mills ratio* in Regime 1 and a negative coefficient in Regime 0.

- If the coefficients on the *Mills 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 *Mills 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 we could use IV and obtain estimates of treatment effects without making unnecessary distributional assumptions.

Chapter 4

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

4.1 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:
  1. assignment to treatment may or may not be random;
  2. the correspondence between assignment and treatment may not be perfect.

4.2 Definition of potential outcomes

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

$$D_i = D_i(\mathbf{Z}) \tag{4.1}$$

The outcome for individual i is a function of the full N-dimensional vector of assignments $\mathbf{Z}$ and treatments $\mathbf{D}$:

$$Y_i = Y_i(\mathbf{Z}, \mathbf{D}) \tag{4.2}$$

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.

4.3 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 $j \neq i$:*

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

where **Z** and **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 counterfactuals.

Table 4.1: 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 and Rosen (1979)).
- Quarter of birth for returns to schooling (Angrist and Krueger (1991)).
- Nearby college for returns to schooling (Card (1995b))
- WWII for returns to schooling (Ichino and Winter-Ebmer (2004))
- Vietnam war lottery for the effect of the military service (Angrist (1990)).

Assumption 2 *Random Assignment.*
Individuals have the same probability to be assigned to the treatment or the control group:

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

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:

$$E\{D_i \mid Z_i = 1\} - E\{D_i \mid Z_i = 0\} = \frac{COV\{D_i Z_i\}}{VAR\{Z_i\}} \tag{4.3}$$

$$E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\} = \frac{COV\{Y_i Z_i\}}{VAR\{Z_i\}} \tag{4.4}$$

Note that the ratio between the causal effect of Z_i on Y_i (eq. 4.4) and the causal effect of Z_i on D_i (eq. 4.3) gives the conventional IV estimator

$$\frac{COV\{Y, Z\}}{COV\{D, Z\}} \tag{4.5}$$

The question that we need to answer are:

- Under which assumptions this IV estimator gives an estimate of 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:

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

or equivalently

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

Note that this assumption is equivalent to the assumption 3.30 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 3.31) 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 an earlier 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.

$$\begin{aligned} Y_i(1, D_i(1)) &- Y_i(0, D_i(0)) \\ &= Y_i(D_i(1)) - Y_i(D_i(0)) \\ &= [Y_i(1)D_i(1) + Y_i(0)(1 - D_i(1))] - \\ &\quad [Y_i(1)D_i(0) + Y_i(0)(1 - D_i(0))] \\ &= Y_i(D_i(1)) - Y_i(D_i(0)) \\ &= (D_i(1) - D_i(0))(Y_i(1) - Y_i(0)) \end{aligned} \tag{4.6}$$

Equation 4.6 states that at the individual level the causal effect of Z on Y (see Definition 3) is equal to the product of 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 4.6 we could construct an estimator for the causal effect of D on Y. But ...

$$\begin{aligned} E\{Y_i(1, D_i(1)) &- Y_i(0, D_i(0))\} \\ &= E\{(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))\} \\ &= E\{Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1\}Pr\{D_i(1) - D_i(0) = 1\} - \\ &\quad E\{Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = -1\}Pr\{D_i(1) - D_i(0) = -1\} \end{aligned} \tag{4.7}$$

Equation 4.7 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:

$$D_i(1) \geq D_i(0) \qquad \forall i \tag{4.8}$$

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

Note that the combination of Assumptions 3 and 5 implies:

$$D_i(1) \geq D_i(0) \qquad \forall i \text{ with strong inequality for at least some } i \tag{4.9}$$

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 in equation 4.7

4.4 The Local Average Treatment Effect

4.4.1 Definition and relationship with IV

Given the monotonicity Assumption 5, equation 4.7 can be written as

$$\begin{aligned} E\{Y_i(1,D_i(1)) &- Y_i(0,D_i(0))\} \\ &= E\{Y_i(1)-Y_i(0) \mid D_i(1)-D_i(0)=1\}Pr\{D_i(1)-D_i(0)=1\} \end{aligned} \tag{4.10}$$

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

$$E\{Y_i(1)-Y_i(0) \mid D_i(1)-D_i(0)=1\} = \frac{E\{Y_i(1,D_i(1))-Y_i(0,D_i(0))\}}{Pr\{D_i(1)-D_i(0)=1\}} \tag{4.11}$$

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

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

$$= \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i(1) = 1\} - Pr\{D_i(0) = 1\}} \tag{4.12}$$

$$= \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i = 1 \mid Z_i = 1\} - Pr\{D_i = 1 \mid Z_i = 0\}} \tag{4.13}$$

$$= \frac{COV\{Y, Z\}}{COV\{D, Z\}} \tag{4.14}$$

Comments

- In order to go from 4.11 to 4.12 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 (see table 4.3 for illustration).

- In order to go from 4.13 to 4.14 see the appendix 4.11.3.

- The last expression 4.14 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 4.2: 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* $Y_i(1,1) - Y_i(0,1) = 0$

4.4.2 Causal interpretation of the LATE-IV estimator

- Each cell contains the causal effect of Z on Y (the numerator of the 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 4.3: 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 4.13):

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

- 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.

4.5 Effects of violations of the LATE assumptions

4.5.1 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:

$$E[H_i \mid i \text{ is a complier}] + E[H_i \mid i \text{ is a noncomplier}] \cdot \frac{P[i \text{ is a noncomplier}]}{P[i \text{ is a complier}]} \tag{4.15}$$

- 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).

4.5.2 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:

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

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.

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

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

1. *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.

2. *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.

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

4.7 Alternative and more informative ways to estimate the LATE

IV is not the only way to estimate the LATE. Imbens and Rubin (1997a, AnnStat), Imbens and Rubin (1997b, REStud) (← good to read) 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.

cannot go into more detail here :-(

4.8 Comments on the LATE and the conventional interpretation of IV

1. 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.

2. To identify the effect of treatment on the treated the conventional approach assumes (see eq. 3.41)

   $$E\{U_i(1) - U_i(0) \mid Z_i, D_i = 1\} = E\{U_i(1) - U_i(0) \mid D_i = 1\} \quad (4.16)$$

   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.

3. Translated in the AIR framework assumption 4.16 is (see the debate Heckman-AIR in AIR, 1996):

   $$E\{Y_i(1) - Y_i(0) \mid Z_i, D_i(Z_i) = 1\} = E\{Y_i(1) - Y_i(0) \mid D_i(Z_i) = 1\} \quad (4.17)$$

   $$\begin{aligned} E\{Y_i(1) - Y_i(0) \quad &| \quad D_i(1) = 1; D_i(0) = 1\} \quad (4.18) \\ &= \quad E\{Y_i(1) - Y_i(0) \mid D_i(1) = 1; D_i(0) = 0\} \end{aligned}$$

   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 - which we did not discuss here in detail - allows to obtain evidence on the validity of this assumption, while in the conventional approach there is no way to assess its validity.

4. 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 3.7).

5. 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.

6. 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.

7. The maximum likelihood approach to the estimation of the LATE - which we did not discuss in more detail here - 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.

8. 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 (*moving target*). 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.

9. Hence, the conventional approach seems to conclude that it is preferable to make additional assumptions like 4.16 or the ones required for the Heckman two steps procedure (in the context of dummy endogenous variables models, Heckman 1978), in order to answer more interesting and well posed policy questions.

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

4.9 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:

1. 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.

2. 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

4.9.1 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

$$VAR\{\Delta\} = \sigma^2 (Z'D)^{-1} Z'Z (Z'D)^{-1} \tag{4.19}$$

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.

4.9.2 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).

4.9.3 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:

$$Y_i = \mu + \Delta D_i + U_i \tag{4.20}$$

The IV estimand is

$$\begin{aligned} Plim\{\Delta^{IV}\} &= \frac{COV\{Z,Y\}}{COV\{Z,D\}} \\ &= \Delta + \frac{COV\{Z,U\}}{COV\{Z,D\}} \end{aligned} \tag{4.21}$$

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 4.5.1 we have seen that the bias due to exclusion restrictions violations increases with the weakness of the instrument.
- In section 4.5.2 we have seen that the bias due to monotonicity violations increases with the weakness of the instrument.

4.10 Application: 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):

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

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 interpret 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.

4.10.1 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:

$$Y = Y(S) \tag{4.22}$$

where:

- S is the number of years of schooling;
- $Y(S)$ 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 $Y'' < 0$).

4.10.2 The objective function

Individuals choose the optimal number of years of schooling S by trading off the benefits of schooling, $Y(S)$, and the costs of schooling, $h(S)$

We adopt a general expression for the utility function:

$$U(S,Y) = \log(Y) - h(S) \tag{4.23}$$

where $h(s)$ captures foregone earnings as well as other components of the cost of schooling.

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.

4.10.3 The optimization problem

The optimization problem for each individual is therefore:

$$\begin{aligned} \text{Max } U(Y,S) &= \log(Y) - h(S) \qquad (4.24) \\ \text{subject to } Y &= Y(S) \end{aligned}$$

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

$$\frac{Y'(S)}{Y(S)} = h'(S) \qquad (4.25)$$

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.

4.10.4 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.

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

$$\left[\frac{Y'(S)}{Y(S)}\right]_i = \beta_i(S) = b_i - k_b S \tag{4.26}$$

$$[h'(S)]_i = \delta_i(S) = r_i + k_r S \tag{4.27}$$

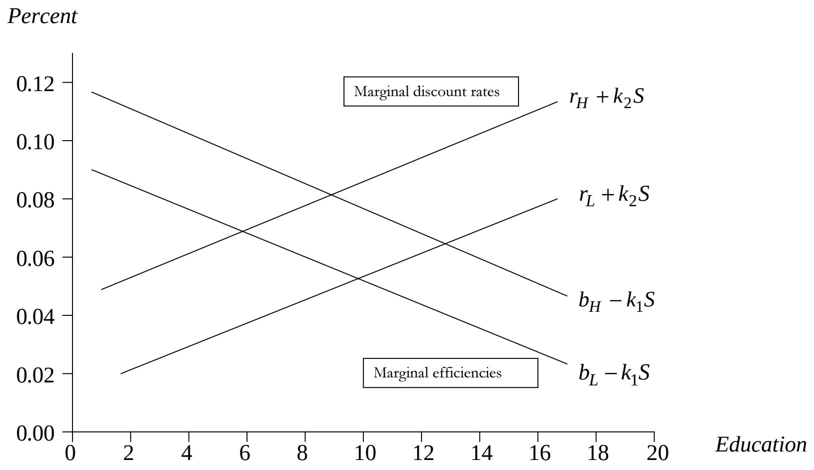
For example:

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.

Figure 4.1: Differences between treated and control individuals.

Figure 1: Marginal benefit and marginal cost schedules for different individuals



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:

$$[Y(S)]_i = ae^{(b_i S - (\frac{k_b}{2} S^2))} \tag{4.28}$$

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 schooling.

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.

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 schooling (i.e. the opportunity cost).

Examples:

1. $k_r = 0$ and $r_i = r$
   the opportunity cost of schooling does not increase with schooling and is equal across individuals which implies *linear* indifference curves with *equal* slopes for different individuals.

2. $k_r = 0$ 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.

3. $k_r > 0$ 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.

4. $k_r > 0$ 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.

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 4.26 and 4.27 in the first order condition 4.25, the optimal amount of schooling now differs across individuals:

$$S_i^* = \frac{(b_i - r_i)}{k_b + k_r} \tag{4.29}$$

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.

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.

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:*

$$\beta_i^* = b_i - k_b S_i^* = (1-\phi)b_i + \phi r_i \tag{4.30}$$

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.

4.10.5 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 66 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:

$$\begin{aligned} \text{Max } U_i(y, S) &= y - r_i S \\ \text{subject to } y &= b_i S - \frac{k_b}{2} S^2 \end{aligned} \tag{4.31}$$

$$\beta_i(S) = b_i - k_b S. \tag{4.32}$$

$$S_i^* = \frac{(b_i - r_i)}{k_b} \tag{4.33}$$

$$\beta_i^* = b_i - k_b S_i^* = r_i. \tag{4.34}$$

Note the difference between equation 4.34 and equation 4.30.

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

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 $g = \{LH, HH, LL, HL\}$. The first letter always refers to marginal benefits, whereas the second letter refers to marginal costs.

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

$$S_g \equiv S_{i,j} = \frac{(b_i - r_j)}{k_b}, \tag{4.35}$$

which implies the following optimal returns to schooling.

$$\begin{aligned} \beta_{LH} &= \beta_{HH} = r_H \\ \beta_{LL} &= \beta_{HL} = r_L. \end{aligned} \tag{4.36}$$

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:

$$\bar{\beta} = (P_{LH} + P_{HH})r_H + (P_{LL} + P_{HL})r_L = \bar{r}, \tag{4.37}$$

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

Note also that nothing on the right hand side of 4.37 is observable.

4.10.6 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.

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).

$$\Delta S_g = \Delta S \qquad \forall g. \tag{4.38}$$

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 2.7 we obtain:

$$E(y_i|i \in T)-E(y_i|i \in C) = (P_{LH}+P_{HH})r_H\Delta S+(P_{LL}+P_{HL})r_L\Delta S = \bar{r}\Delta S = \bar{\beta}\Delta S \tag{4.39}$$

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:

$$\begin{aligned}\frac{E\{y_i|i \in T\} - E\{y_i|i \in C\}}{E\{S_i|i \in T\} - E\{S_i|i \in C\}} &= \frac{E_g\{r_g \Delta S_g\}}{E_g\{\Delta S_g\}}. \qquad (4.40)\\ &= \frac{(P_{LH} + P_{HH})r_H \Delta S + (P_{LL} + P_{HL})r_L \Delta S}{\Delta S}\\ &= \bar{r}\\ &= \bar{\beta}.\end{aligned}$$

Note that, the expression on the left hand side of 4.40, 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?

4.10.7 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 4.28, 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:

$$y_i = a + b_i S_i - \frac{k_b}{2} S_i^2 \tag{4.41}$$

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 tends to be convex or concave depending on which type of heterogeneity has more variance.

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:

$$\text{plim } (\hat{\rho}^{OLS}) = \frac{COV(y_i, S_i)}{VAR(S_i)} \tag{4.42}$$

Following Card(1995a):

$$\text{plim } (\hat{\rho}^{OLS}) = (1 - \alpha)\bar{b} + \alpha\bar{r} \tag{4.43}$$

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 (loosely) 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

$$\text{plim } (\hat{\rho}^{OLS}) = \lambda\bar{b} + (1 - \lambda)\bar{r}. \tag{4.44}$$

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 4.44 the average marginal return to schooling, which using 4.34 can be written as:

$$E(\beta_i) = \bar{\beta} = \bar{b} - k_b\bar{S} \tag{4.45}$$

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 4.45, equation 4.44 can be rewritten as:

$$\text{plim } (\hat{\rho}^{OLS}) = \bar{\beta} + \lambda(\bar{b} - \bar{r}). \tag{4.46}$$

Equation 4.46 says that the OLS regression of log-earnings on schooling yield a biased 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 endogeneity 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.

4.10.8 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 4.11.1 and 4.11.3):

$$\text{plim } \rho_Z^{IV} = \frac{COV\ \{Y, Z\}}{COV\ \{S, Z\}} = \frac{E\{y_i|Z_i = 1\} - E\{y_i|Z_i = 0\}}{E\{S_i|Z_i = 1\} - E\{S_i|Z_i = 0\}} = \frac{E_g\{r_g \Delta S_{g|Z}\}}{E_g\{\Delta S_{g|Z}\}} \tag{4.47}$$

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 methods 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:

$$\text{plim } \rho_Z^{IV} = E_g(\beta_g) = \bar{r} = \bar{\beta} \tag{4.48}$$

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:

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

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, $\Delta S_{g|Z}$.

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

4.10.9 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

  ⇒ expected to affect the group HH with the highest return

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

  ⇒ 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:

$$\Delta S_{LL|Z} = \Delta S_{HL|Z} = \Delta S_{LH|Z} \approx 0$$

$$\text{plim } \rho_Z^{IV} \approx \beta_{HH} \tag{4.49}$$

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

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 in 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:

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

$$\text{plim } \rho_W^{IV} \approx \beta_{LL} + N \tag{4.50}$$

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 high school, 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}.

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
  1. $Z_i = 1$ if i had a father in the army during the war;
  2. $W_i = 1$ if i's father has more than high–school education

A potential problem leading to a richer specification

Bound and Jaeger (2000) 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 (1999) we therefore include also information on parental background as control variables.

$$\begin{aligned} lnW_i &= \beta_1 + \beta_2 EDU_i + \beta_3 AGE_i + \beta_4 AGE_i^2 + \beta_5 AGE_i^3 \\ &\quad + \beta_6 HIGHEDF_i + + \beta_7 BLUEF_i + \beta_8 SELFF_i + \varepsilon_i \end{aligned} \tag{4.51}$$

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 4.4: 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 (0.078)	0.048 (0.013)	0.117 (0.053)	0.048 (0. 014)	0.055 (0.005)
Age (years)	0.106 (0.101)	0.215 (0.039)	0.141 (0.070)	0.215 (0.039)	0.208 (0.033)
Age^2 /100	-0.183 (0.235)	-0.434 (0.093)	-0.263 (0.164)	-0.434 (0.094)	-0.418 (0.084)
Age^3 /10,000	0.106 (0.175)	0.291 (0.007)	0.165 (0.123)	0.290 (0.008)	0.279 (0.007)
Father is a blue–collar worker (0,1)	—	—	0.058 (0.051)	-0.001 (0.031)	0.004 (0.026)
Father is self–employed (0,1)	—	—	-0.032 (0.043)	-0.041 (0.042)	-0.041 (0.037)
Father has more than high–school education (0,1)	—	—	-0.209 (0.172)	—	-0.019 (0.052)
Constant	-0.684 (0.619)	-1.080 (0.483)	-0.909 (0.517)	-1.075 (0.484)	-1.060 (0.411)
$\bar{R}^2$	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 dependentvariable 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.

4.11 Appendix

4.11.1 Standard characterization of IV

Consider the model

$$Y = \alpha + \Delta D + \epsilon \tag{4.52}$$

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

$$\text{plim}\{\hat{\Delta}_{OLS}\} = \frac{COV\{Y, D\}}{V\{D\}} = \Delta + \frac{COV\{\epsilon, D\}}{V\{D\}} \neq \Delta \tag{4.53}$$

and OLS gives an inconsistent estimate of Δ.

Consider a variable Z such that:

$$E\{D \mid Z\} \neq 0 \;\; \Rightarrow \;\; COV\{Z, D\} \neq 0 \tag{4.54}$$

$$E\{\epsilon \mid Z\} = 0 \;\; \Rightarrow \;\; COV\{Z, \epsilon\} = 0. \tag{4.55}$$

If this variable exists, the following population equation holds (see also the Appendix 4.11.2 in the next page):

$$\frac{COV\{Y, Z\}}{COV\{D, Z\}} = \Delta + \frac{COV\{\epsilon, Z\}}{COV\{D, Z\}} = \Delta = \text{plim}\{\hat{\Delta}_{IV}\} \tag{4.56}$$

Substituting the appropriate sample covariances on the LHS of 4.56 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.

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

Consider the following model

$$\begin{aligned} Y &= D\Delta + \epsilon && (4.57)\\ D &= Z\gamma + u && (4.58) \end{aligned}$$

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

Note that

$$\hat{D} = Z(Z'Z)^{-1}Z'D = P_Z D \qquad (4.59)$$

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

$$P_Z Y = P_Z D\Delta + P_Z\epsilon \qquad (4.60)$$

gives

$$\begin{aligned} \hat{\Delta} &= (D'P_Z P_Z D)^{-1}D'P_Z P_Z Y && (4.61)\\ &= (D'P_Z D)^{-1}D'P_Z Y \\ &= (D'Z)^{-1}Z'Y \rightarrow \frac{COV\{Y,Z\}}{COV\{D,Z\}} \end{aligned}$$

which is the IV estimator.

4.11.3 Equivalence between IV and Wald estimators

Consider the setup of Section 3 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:

$$\frac{COV\{Y, Z\}}{COV\{D, Z\}} = \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i = 1 \mid Z_i = 1\} - Pr\{D_i = 1 \mid Z_i = 0\}} \tag{4.62}$$

The RHS of 4.62 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 3.17 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:

$$E\{Y_i \mid Z_i = 1\} = \mu(0) + \Delta^* E\{D_i \mid Z_i = 1\} + E\{\epsilon_i \mid Z_i = 1\} \tag{4.63}$$
$$E\{Y_i \mid Z_i = 0\} = \mu(0) + \Delta^* E\{D_i \mid Z_i = 0\} + E\{\epsilon_i \mid Z_i = 0\} \tag{4.64}$$

Assuming that the instrument Z satisfies the condition 4.55, so that the conditional expectations of the errors are zero:

$$E\{Y_i \mid Z_i = 1\} = \mu(0) + \Delta^* Pr\{D_i = 1 \mid Z_i = 1\} \tag{4.65}$$
$$E\{Y_i \mid Z_i = 0\} = \mu(0) + \Delta^* Pr\{D_i = 1 \mid Z_i = 0\} \tag{4.66}$$

Subtracting 4.66 from 4.65 and solving for Δ^* gives the Wald-IV estimator on the RHS of 4.62.

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

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

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

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

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

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

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

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

Q.E.D.

Chapter 5

Selection on Observables and Matching

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

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

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

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

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

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

However, matching methods have the following desirable features:

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

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

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

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

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

5.1 Notation

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

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{5.1}$$

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

$$\Delta_i = Y_i(1) - Y_i(0) \tag{5.2}$$

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

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

$$\tau = E\{\Delta_i | D_i = 1\} = E\{Y_i(1) - Y_i(0) | D_i = 1\} \tag{5.3}$$

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

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

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

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.

$$Y(1), Y(0) \quad \perp \quad D \tag{5.4}$$

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

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

$$E\{Y_i(0)|D_i = 0\} = E\{Y_i(0)|D_i = 1\} = E\{Y_i(0)\} \tag{5.5}$$

$$E\{Y_i(1)|D_i = 0\} = E\{Y_i(1)|D_i = 1\} = E\{Y_i(1)\} \tag{5.6}$$

and substituting 5.5 and 5.6 in 5.3 it is immediate to obtain:

$$\begin{aligned} \tau &\equiv E\{\Delta_i \mid D_i = 1\} \qquad (5.7)\\ &\equiv E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &\equiv E\{Y_i(1)|D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 0\} \\ &= E\{Y_i|D_i = 1\} - E\{Y_i|D_i = 0\}. \end{aligned}$$

Randomization ensures that the sample selection bias is zero:

$$E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} = 0 \tag{5.8}$$

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 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

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad X \tag{5.9}$$

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.

Remark that the assumption of unconfoundedness is also called *conditional independence assumption* or *CIA* for short.

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:

$$E\{Y_i(0)|D_i=0,X\} = E\{Y_i(0)|D_i=1,X\} = E\{Y_i(0)|X\} \tag{5.10}$$

$$E\{Y_i(1)|D_i=0,X\} = E\{Y_i(1)|D_i=1,X\} = E\{Y_i(1)|X\} \tag{5.11}$$

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

$$\begin{aligned} \delta_x &\equiv E\{\Delta_i|X\} \qquad (5.12) \\ &\equiv E\{Y_i(1)-Y_i(0)|X\} \\ &\equiv E\{Y_i(1)|X\} - E\{Y_i(0)|X\} \\ &= E\{Y_i(1)|D_i=1,X\} - E\{Y_i(0)|D_i=0,X\} \\ &= E\{Y_i|D_i=1,X\} - E\{Y_i|D_i=0,X\}. \end{aligned}$$

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

$$\begin{aligned} \tau &\equiv E\{\Delta_i|D_i=1\} \qquad (5.13) \\ &= E\{E\{\Delta_i|D_i=1,X\} \mid D_i=1\} \\ &= E\{\ E\{Y_i|D_i=1,X\} - E\{Y_i|D_i=0,X\}\ |D_i=1\} \\ &= E\{\delta_x|D_i=1\} \end{aligned}$$

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 5.12 and 5.13:

1. stratify the data into cells defined by each particular value of X;
2. within each cell (i.e. conditioning on X) compute the difference between the average outcomes of the treated and the controls;
3. 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?

In which sense do matching and regression differ?

Angrist (1998, p. 255): "Differences between regression and matching strategies for the estimation of treatment effects are partly cosmetic. While matching methods are often more transparent to nonspecialists, regression estimation is more straightforward to implement when covariates are continuously distributed because matching on continuous covariates requires stratification or pairing (Cochran (1968)). Note, however, that both methods require a similar sort of approximation since regression on continuous covariates in any finite sample requires functional form restrictions. The fact that both stratification and functional from approximations can be made increasingly accurate as the sample size grows suggests the manner in which continuous covariates are accommodated is not the most important difference between the two methods. The essential difference between regression and matching in evaluation research is the weighting scheme used to pool estimates at 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:

$$\begin{aligned}\delta_1 &= E\{Y_i(1) - Y_i(0)|D_i = 1, x_i = 1\} = E\{Y_i(1) - Y_i(0)|x_i = 1\} \\ &= E\{Y_i \mid D_i = 1, x_i = 1\} - E\{Y_i \mid D_i = 0, x_i = 1\} \qquad (5.14)\end{aligned}$$

$$\begin{aligned}\delta_0 &= E\{Y_i(1) - Y_i(0)|D_i = 1, x_i = 0\} = E\{Y_i(1) - Y_i(0)|x_i = 0\} \\ &= E\{Y_i \mid D_i = 1, x_i = 0\} - E\{Y_i \mid D_i = 0, x_i = 0\} \qquad (5.15)\end{aligned}$$

Using matching, the ATT is therefore

$$\begin{aligned}\Delta_M &= E\{Y_i(1) - Y_i(0)|D_i = 1\} \qquad (5.16)\\ &= \delta_0 P(x_i = 0 \mid D_i = 1) + \delta_1 P(x_i = 1 \mid D_i = 1) \\ &= \delta_0 \frac{P(D_i = 1 \mid x_i = 0) P(x_i = 0)}{P(D_i = 1)} + \delta_1 \frac{P(D_i = 1 \mid x_i = 1) P(x_i = 1)}{P(D_i = 1)}\end{aligned}$$

Note that

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

Suppose that we estimate instead the (fully saturated) model

$$Y_i = \alpha + \beta x_i + \Delta_r D_i + \epsilon_i. \tag{5.17}$$

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

$$\Delta_r = \frac{E\{[D_i - E\{D_i \mid x_i\}]Y_i\}}{E\{[D_i - E\{D_i \mid x_i\}]D_i\}}. \tag{5.18}$$

By unconfoundedness, Δ_r is free of selection bias.

We can also write that:

$$Y_i = E\{Y_i(0) \mid x_i\} + E\{Y_i(1) - Y_i(0) \mid x_i\}D_i + \epsilon \tag{5.19}$$

Substitute 5.14, 5.15 and 5.19 into 5.18, and iterating expectation with respect to x we obtain:

$$\begin{aligned} \Delta_r &= \delta_0 \frac{P(D_i = 1 \mid x_i = 0)[1 - P(D_i = 1 \mid x_i = 0)]P(x_i = 0)}{E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}} \\ &+ \delta_1 \frac{P(D_i = 1 \mid x_i = 1)[1 - P(D_i = 1 \mid x_i = 1)]P(x_i = 1)}{E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}} \end{aligned} \tag{5.20}$$

Note that

- the weights are *proportional 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.

In fact, 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, p.256f.) gives an interesting **example** of the differences between matching and regression:
Suppose that

- $P[D = 1|x = 0] = 0.9$
- $P[D = 1|x = 1] = 0.5$
- $P[x = 1] = 0.5$

Applying equations 5.16 and 5.20, we have

$$\begin{aligned}
\Delta_M &= E\{Y_i(1) - Y_i(0)|D_i = 1\} \\
&= \frac{P(D_i = 1 \mid x_i = 0)P(x_i = 0)}{P(D_i = 1)}\delta_0 + \frac{P(D_i = 1 \mid x_i = 1)P(x_i = 1)}{P(D_i = 1)}\delta_1 \\
&= \frac{0.9 * 0.5}{0.7}\delta_0 + \frac{0.5 * 0.5}{0.7}\delta_1 \\
&= 0.64\delta_0 + 0.36\delta_1 \qquad (5.21)
\end{aligned}$$

where $P[D = 1] = P[D = 1|x = 0]P[x = 0] + P[D = 1|x = 1]P[x = 1]$
$= 0.9 * 0.5 + 0.5 * 0.5 = 0.45 + 0.25 = 0.7.$

$$
\begin{aligned}
\Delta_r &= \frac{P(D_i = 1 \mid x_i = 0)[1 - P(D_i = 1 \mid x_i = 0)]P(x_i = 0)}{E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}}\delta_0 \\
&+ \frac{P(D_i = 1 \mid x_i = 1)[1 - P(D_i = 1 \mid x_i = 1)]P(x_i = 1)}{E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}}\delta_1 \\
&= \frac{0.9 * 0.1 * 0.5}{0.17}\delta_0 + \frac{0.5 * 0.5 * 0.5}{0.17}\delta_1 \\
&= 0.26\delta_0 + 0.74\delta_1 \qquad (5.22)
\end{aligned}
$$

where $E\{P(D_i = 1 \mid x_i)[1 - P(D_i = 1 \mid x_i)]\}$
$= P(D_i = 1 \mid x = 0)[1 - P(D_i = 1 \mid x = 0)]Pr(x = 0)$
$+ P(D_i = 1 \mid x = 1)[1 - P(D_i = 1 \mid x = 1)]Pr(x = 1)$
$= 0.9 * 0.1 * 0.5 + 0.5 * 0.5 * 0.5 = 0.045 + 0.125 = 0.17.$

Thus, while $E\{Y_i(1) - Y_i(0)|D_i = 1\}$ reflects the fact that veterans are much more likely to have $x = 0$, the regression parameter Δ_r puts more weight on the treatment effect for those with $x = 1$ because the variance of D is much larger for that group.

Discussing the results of his Table II, Angrist (1998) states that: "The divergence between regression and matching estimates after 1984 is probably explained by differences in the long term impact of military service on men with covariate values that place them in low-probability-of-service and high-probability-of-service groups." Angrist's figure 4 shows a strong negative relationship between treatment effects and the probability of service for both whites and non-whites.
"The **matching estimator** gives the small covariate-specific estimates for men with high probabilities of service the most weight, while the larger covariate-specific estimates for men with low probability of service are given less weight.
The **regression estimator**, in contrast, gives more weight to covariate-specific estimates where the probability of military service conditional on covariates is close to one-half. This leads to a higher overall treatment effect."

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, continuous.

A rare exception of a situation where exact matching is feasible: Ichino, Schwerdt, Winter-Ebmer and Zweimüller (2008).

This dimensionality problem is likely to jeopardize also the matching strategy described by equations 5.12 and 5.13:

- 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 non-saturated model is not a solution and may lead to seriously misleading conclusions!

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:

$$p(X) \equiv Pr\{D = 1|X\} = E\{D|X\} \tag{5.23}$$

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

$$D \quad \perp \quad X \quad | \quad p(X) \tag{5.24}$$

Proof:
First:

$$\begin{aligned} Pr\{D = 1|X, p(X)\} &= E\{D|X, p(X)\} \qquad (5.25)\\ &= E\{D|X\} = Pr\{D = 1|X\} \\ &= p(X) \end{aligned}$$

Second:

$$\begin{aligned} Pr\{D = 1|p(X)\} &= E\{D|p(X)\} \qquad (5.26)\\ &= E\{E\{D|X, p(X)\}|p(X)\} = E\{p(X|p(X)\} \\ &= p(X) \end{aligned}$$

Hence:

$$Pr\{D = 1|X, p(X)\} = Pr\{D = 1|p(X)\} \tag{5.27}$$

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

$$Y(1), Y(0) \quad \perp \quad D \quad | \quad p(X) \tag{5.28}$$

Proof: First:

$$\begin{aligned} Pr\{D = 1|Y(1), Y(0), p(X)\} &= E\{D|Y(1), Y(0), p(X)\} \qquad (5.29) \\ &= E\{E\{D|X, Y(1), Y(0)\}|Y(1), Y(0), p(X)\} \\ &= E\{E\{D|X\}|Y(1), Y(0), p(X)\} \\ &= E\{p(X)|Y(1), Y(0), p(X)\} \\ &= p(X) \end{aligned}$$

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

$$Pr\{D = 1|p(X)\} \;=\; p(X) \tag{5.30}$$

Hence

$$Pr\{D = 1|Y(1), Y(0), p(X)\} = Pr\{D = 1|p(X)\} \tag{5.31}$$

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 one-dimensional variable (the propensity score) instead of the multidimensional vector of observables X.

$$E\{Y_i(0)|D_i = 0, p(X_i)\} = E\{Y_i(0)|D_i = 1, p(X_i)\} = E\{Y_i(0)|p(X_i)\} \tag{5.32}$$

$$E\{Y_i(1)|D_i = 0, p(X_i)\} = E\{Y_i(1)|D_i = 1, p(X_i)\} = E\{Y_i(1)|p(X_i)\} \tag{5.33}$$

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

$$\begin{aligned} \delta_{p(x)} &\equiv E\{\Delta_i|p(X_i)\} \qquad (5.34)\\ &\equiv E\{Y_i(1) - Y_i(0)|p(X_i)\}\\ &\equiv E\{Y_i(1)|p(X_i)\} - E\{Y_i(0)|p(X_i)\}\\ &= E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\}\\ &= E\{Y_i|D_i = 1, p(X_i)\} - E\{Y_i|D_i = 0, p(X_i)\}. \end{aligned}$$

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

$$\begin{aligned} \tau &= E\{\Delta_i|D_i = 1\} \qquad (5.35)\\ &= E\{E\{\Delta_i|D_i = 1, p(X_i)\}\}\\ &= E\{\quad E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\} \quad |D_i = 1\}\\ &= E\{\delta_{p(x)}|D_i = 1\} \end{aligned}$$

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

5.5.1 Implementation of matching based on the pscore

To implement the estimation strategy suggested by the propensity score and its properties two sequential steps are needed.

1. *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.

2. *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 5.34).
   - obtain the average of these conditional effects as in equation 5.35.

   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 neighbor 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:

$$Pr\{D_i = 1|X_i\} = \frac{e^{\lambda h(X_i)}}{1 + e^{\lambda h(X_i)}} \tag{5.36}$$

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 5.12 and 5.13), the propensity score reduces the dimensionality of the estimation problem.

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

An algorithm for estimating the propensity score

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

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

3. 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.

4. 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.sobecker.de

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

$$\tau_q^S = \frac{\sum_{i \in I(q)} Y_i^T}{N_q^T} - \frac{\sum_{j \in I(q)} Y_j^C}{N_q^C} \tag{5.37}$$

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 5.35 is computed with the following formula:

$$\tau^S = \sum_{q=1}^{Q} \tau_q^S \frac{\sum_{i \in I(q)} D_i}{\sum_{\forall i} D_i} \tag{5.38}$$

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

$$Var(\tau^S) = \frac{1}{N^T}\left[Var(Y_i^T) + \sum_{q=1}^{Q} \frac{N_q^T}{N^T}\frac{N_q^T}{N_q^C} Var(Y_j^C)\right] \tag{5.39}$$

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.sobecker.de

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:

  $$Var(\tau^S) = \frac{1}{N^T}\left[Var(Y_i^T) + Var(Y_j^C)\right] \tag{5.40}$$

- *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, 2002.)

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

Nearest and radius matching with replacement for the ATT

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

- For each treated unit find the nearest control unit.
- If the nearest control unit has already been used 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

$$C(i) = \min_j \parallel p_i - p_j \parallel, \tag{5.41}$$

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

In radius matching,

$$C(i) = \{p_j \mid \parallel p_i - p_j \parallel < r\}\,, \tag{5.42}$$

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):

$$\begin{aligned} \tau^M &= \frac{1}{N^T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C(i)} w_{ij} Y_j^C \right] && (5.43) \\ &= \frac{1}{N^T} \left[\sum_{i \in T} Y_i^T - \sum_{i \in T} \sum_{j \in C(i)} w_{ij} Y_j^C \right] && (5.44) \\ &= \frac{1}{N^T} \sum_{i \in T} Y_i^T - \frac{1}{N^T} \sum_{j \in C} w_j Y_j^C && (5.45) \end{aligned}$$

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.

$$\begin{aligned} Var(\tau^M) &= \frac{1}{(N^T)^2}\left[\sum_{i\in T} Var(Y_i^T) + \sum_{j\in C}(w_j)^2 Var(Y_j^C)\right] \quad (5.46)\\ &= \frac{1}{(N^T)^2}\left[N^T Var(Y_i^T) + \sum_{j\in C}(w_j)^2 Var(Y_j^C)\right] \quad (5.47)\\ &= \frac{1}{N^T} Var(Y_i^T) + \frac{1}{(N^T)^2}\sum_{j\in C}(w_j)^2 Var(Y_j^C). \quad (5.48)\end{aligned}$$

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.sobecker.de

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

$$\tau^K = \frac{1}{N^T} \sum_{i \in T} \left\{ Y_i^T - \frac{\sum_{j \in C} Y_j^C G(\frac{p_j - p_i}{h_n})}{\sum_{k \in C} G(\frac{p_k - p_i}{h_n})} \right\} \tag{5.49}$$

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

Under standard conditions on the bandwidth and kernel

$$\frac{\sum_{j \in C} Y_j^C G(\frac{p_j - p_i}{h_n})}{\sum_{k \in C} G(\frac{p_k - p_i}{h_n})} \tag{5.50}$$

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.sobecker.de

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. (see Hirano, Imbens, Ridder (2003))

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

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

Proof: Using the law of iterated expectations:

$$E\left\{\frac{Y_i D_i}{p(X_i)}\right\} - E\left\{\frac{Y_i(1-D_i)}{1-p(X_i)}\right\} = E\left\{E\left\{\frac{Y_i D_i}{p(X_i)}|X\right\} - E\left\{\frac{Y_i(1-D_i)}{1-p(X_i)}|X\right\}\right\} \tag{5.51}$$

which can be rewritten as:

$$E\left\{E\left\{\frac{Y_i(1)}{p(X_i)}|D_i = 1, X\right\} Pr\{D_i = 1|X\} - E\left\{\frac{Y_i(0)}{1-p(X_i)}|D_i = 0, X\right\} Pr\{D_i = 0|X\}\right\} \tag{5.52}$$

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:

$$E\{E\{Y_i(1)|X\} - E\{Y_i(0)|X\}\} = E\{Y_i(1)\} - E\{Y_i(0)\} \tag{5.53}$$

QED

Therefore, substituting sample statistics in the RHS of 5.51 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

$$\begin{aligned} \tau &= \{E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 1\}\} \qquad (5.54) \\ &= E\{Y_i D_i\} - E\left\{Y_i(1 - D_i)\frac{p(X_i)}{1 - p(X_i)}\right\} \end{aligned}$$

Proof: Using the law of iterated expectations:

$$E\{Y_i D_i\} - E\left\{Y_i(1 - D_i)\frac{p(X_i)}{1 - p(X_i)}\right\} = E\left\{E\{Y_i D_i|X\} - E\left\{Y_i(1 - D_i)\frac{p(X_i)}{1 - p(X_i)}|X\right\}\right\} \qquad (5.55)$$

which can be rewritten as:

$$E\left\{E\{Y_i(1)|D_i = 1, X\}Pr\{D_i = 1|X\} - E\left\{Y_i(0)\frac{p(X_i)}{1 - p(X_i)}|D_i = 0, X\right\}Pr\{D_i = 0|X\}\right\} \qquad (5.56)$$

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:

$$\begin{aligned} E\{E\{Y_i(1)|D_i = 1, X\} &- E\{Y_i(0)|D_i = 1, X\}|D_i = 1\} \qquad (5.57) \\ &= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 1\} \end{aligned}$$

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 5.54 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 (2003) show how to compute the standard error.

  See also Heckman, Ichimura and Todd (1998) and Hahn (1998).

5.6 Sensitivity Analysis of Matching Estimators to the CIA

The material of this section is based on

- Becker and Caliendo (2007)

5.6.1 Intro

- Matching is based on the conditional independence or unconfoundedness assumption.
- If there are unobserved variables which affect assignment into treatment and the outcome variable simultaneously, a *hidden bias* might arise to which matching estimators are not robust.
- bounding approach proposed by Rosenbaum (2002)
- Stata implementation: `mhbounds` allows the researcher to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of the matching analysis.

5.6.2 Sensitivity Analysis with Rosenbaum Bounds

assume that the participation probability is given by

$$P_i = P(x_i, u_i) = P(D_i = 1 \mid x_i, u_i) = F(\beta x_i + \gamma u_i) \tag{5.58}$$

where

- x_i are the observed characteristics for individual i
- u_i is the unobserved variable
- γ is the effect of u_i on the participation decision
- If the study is free of hidden bias, ...
- ... γ will be zero and the participation probability will solely be determined by x_i.
- However, if there is hidden bias, ...
- ... two individuals with the same observed covariates x have differing chances of receiving treatment.

Let us assume

- we have a matched pair of individuals i and j ...
- ... and further assume that F is the logistic distribution.
- The odds that individuals receive treatment are then given by
- $\frac{P_i}{(1-P_i)}$ and $\frac{P_j}{(1-P_j)}$, ...
- ... and the odds ratio is given by:

$$\frac{\frac{P_i}{1-P_i}}{\frac{P_j}{1-P_j}} = \frac{P_i(1-P_j)}{P_j(1-P_i)} = \frac{\exp(\beta x_i + \gamma u_i)}{\exp(\beta x_j + \gamma u_j)}. \tag{5.59}$$

If both units have identical observed covariates - as implied by the matching procedure -

- the x-vector cancels out implying that:

$$\frac{\exp(\beta x_i + \gamma u_i)}{\exp(\beta x_j + \gamma u_j)} = exp[\gamma(u_i - u_j)]. \tag{5.60}$$

- still, both individuals differ in their odds of receiving treatment by a factor that involves the parameter γ and the difference in their unobserved covariates u.

So, if there are either

- no differences in unobserved variables ($u_i = u_j$) or
- ... if unobserved variables have no influence on the probability of participating ($\gamma = 0$), ...
- ... the odds ratio is one, implying the absence of hidden or unobserved selection bias.

It is now the task of sensitivity analysis to evaluate how inference about the programme effect is altered by changing the values of γ and $(u_i - u_j)$.

- Aakvik (2001): assume that the unobserved covariate is a dummy variable with $u_i \in \{0, 1\}$.
- Rosenbaum (2002) shows that (5.59) implies the following bounds on the odds-ratio that either of the two matched individuals will receive treatment:

$$\frac{1}{e^\gamma} \leq \frac{P_i(1 - P_j)}{P_j(1 - P_i)} \leq e^\gamma. \tag{5.61}$$

- both matched individuals have the same probability of participating only if $e^\gamma = 1$.

- Otherwise, if for example $e^{\gamma} = 2$, individuals who appear to be similar (in terms of x) could differ in their odds of receiving the treatment by as much as a factor of 2.

- In this sense, e^{γ} is a measure of the degree of departure from a study that is free of hidden bias (Rosenbaum, 2002).[1]

The MH Test Statistic

For binary outcomes, Aakvik (2001) suggests using the Mantel and Haenszel (MH, 1959) test statistic. To do so, some additional notation is needed.

- we observe the outcome y for both participants and non-participants.

- If y is unaffected by different treatment assignments, treatment d is said to have no effect.

- If y is different for different assignments, then the treatment has some positive (or negative) effect.

- To be significant, the treatment effect has to cross some test statistic $t(d, y)$.

- The MH non-parametric test compares the successful number of individuals in the treatment group against the same expected number given the treatment effect is zero.

- Aakvik (2001) notes that the MH test can be used to test for no treatment effect both within different strata of the sample and as a weighted average between strata.

- Under the null-hypothesis of no treatment effect, the distribution of y is hypergeometric.

- We notate N_{1s} and N_{0s} as the numbers of treated and non-treated individuals in stratum s, where $N_s = N_{0s} + N_{1s}$.

[1] A related approach can be found in Manski (1990, 1995) who proposes 'worst-case bounds' which are somewhat analogous to letting $e^{\gamma} \to \infty$ in a sensitivity analysis.

- Y_{1s} is the number of successful participants, Y_{0s} is the number of successful non-participants, and Y_s is the number of total successes in stratum s.

- The test-statistic Q_{MH} follows asymptotically the standard normal distribution and is given by:

$$Q_{MH} = \frac{|Y_1 - \sum_{s=1}^{S} E(Y_{1s})| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(Y_{1s})}} = \frac{|Y_1 - \sum_{s=1}^{S}(\frac{N_{1s}Y_s}{N_s})| - 0.5}{\sqrt{\sum_{s=1}^{S} \frac{N_{1s}N_{0s}Y_s(N_s-Y_s)}{N_s^2(N_s-1)}}}. \tag{5.62}$$

To use such a test-statistic, we first have to make the individuals in the treatment and control groups as similar as possible, because this test is based on random sampling. Since this is done by our matching procedure, we can proceed to discuss the possible influences of $e^\gamma > 1$.

- for fixed $e^\gamma > 1$ and $u \in \{0, 1\}$, Rosenbaum (2002) shows that the test-statistic Q_{MH} can be bounded by two known distributions.

- As noted already, if $e^\gamma = 1$ the bounds are equal to the 'base' scenario of no hidden bias.

- With increasing e^γ, the bounds move apart reflecting uncertainty about the test-statistics in the presence of unobserved selection bias.

Two scenarios are especially useful:

- let Q_{MH}^+ be the test-statistic given that we have overestimated the treatment effect ...

- ... and Q_{MH}^- the case where we have underestimated the treatment effect.

The two bounds are then given by:

$$Q_{MH}^{+} = \frac{|Y_1 - \sum_{s=1}^{S} \widetilde{E}_s^{+}| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(\widetilde{E}_s^{+})}} \tag{5.63}$$

and

$$Q_{MH}^{-} = \frac{|Y_1 - \sum_{s=1}^{S} \widetilde{E}_s^{-}| - 0.5}{\sqrt{\sum_{s=1}^{S} Var(\widetilde{E}_s^{-})}} \tag{5.64}$$

where $\widetilde{E_s}$ and $Var(\widetilde{E_s})$ are the large sample approximations to the expectation and variance of the number of successful participants when u is binary and for given γ.[2]

5.6.3 Syntax of Stata module `mhbounds`

`mhbounds` computes Mantel-Haenszel bounds to check sensitivity of estimated average treatment effects on the treated.

`mhbounds` *outcome* [if], `gamma(numlist) [ treated(newvar) weight(newvar) support(newvar) stratum(newvar) stratamat`]

5.6.4 Options

`gamma(numlist)` is a compulsory option and asks users to specify the values of $\Gamma = e^{\gamma} \geq 1$ for which to carry out the sensitivity analysis. Estimates at $\Gamma = 1$ (no hidden bias) are included in the calculations by default.

`treated(varname)` specifies the name of the user-provided treatment variable; If no name is provided, mhbounds expects `_treated` from `psmatch` or `psmatch2`.

[2] The large sample approximation of $\widetilde{E_s^{+}}$ is the unique root of the following quadratic equation: $\widetilde{E}_s^2(e^{\gamma} - 1) - \widetilde{E}_s[(e^{\gamma} - 1)(N_{1s} + Y_s) + N_s] + e^{\gamma} Y_s N_{1s}$, with the addition of $max(0, Y_s + N_{1s} - N_s \leq \widetilde{E}_s \leq min(Y_s, N_{1s}))$ to decide which root to use. $\widetilde{E}_s^{-}$ is determined by replacing e^{γ} by $\frac{1}{e^{\gamma}}$. The large sample approximation of the variance is given by: $Var(\widetilde{E}_s) = \left(\frac{1}{\widetilde{E}_s} + \frac{1}{Y_s - \widetilde{E}_s} + \frac{1}{N_{1s} - \widetilde{E}_s} + \frac{1}{N_s - Y_s - N_{1s} + \widetilde{E}_s}\right)^{-1}$.

`weight(varname)` specifies the name of the user-provided variable containing the frequency with which the observation is used as a match; if no name is provided, mhbounds expects `_weight` from `psmatch` or `psmatch2`.
`support(varname)` specifies the name of the user-provided common support variable. If no name is provided, `mhbounds` expects `_support` from `psmatch` or `psmatch2`.
`stratum(varname)` specifies the name of the user-provided variable indicating strata. Aakvik (2001) notes that the Mantel-Haenszel test can be used to test for no treatment effect both within different strata of the sample and as a weighted average between strata. This option is particularly useful when used after stratification matching, using, e.g. `atts`.
`stratamat`, in combination with `stratum(varname)` keeps in memory not only the matrix `outmat` containing the overall/combined test statistics, but also the matrices `outmat_j` containing the strata-specific test statistics, $j = 1, ...,$#strata.

Typical Examples

1. Running `mhbounds` after `psamtch2`:
   `psmatch2` college, outcome(wage) pscore(pscore) caliper(.25) common noreplacement `mhbounds` wage, gamma(1 (0.05) 2) [performs sensitivity analysis at Gamma = 1,1.05,1.10,...,2.]

2. Running `mhbounds` with user-defined treatment-, weight- and support-indicators:
   `mhbounds` outcome, gamma(1 (0.05) 2) treated(mytreat) weight(myweight) support(mysupport)

3. Running `mhbounds` with user-defined treatment-, weight- and support-indicators with different strata in the population:
   `mhbounds` outcome, gamma(1 (0.05) 2) treated(mytreat) weight(myweight) support(mysupport) stratum(mystratum) stratamat

Please note that `mhbounds` is suited for k-nearest neighbor matching without replacement and for stratification matching.

5.6.5 Applications

See illustrations in Becker and Caliendo (2007):

1. Rosenbaum (2002, Table 4.11, p. 130): medical study of the possible effects of the drug allopurinol as a cause of rash
2. National Supported Work (NSW) training program with non-experimental comparison groups from surveys as the Panel Study of Income Dynamics (PSID) or the Current (CPS): LaLonde (1986), Dehejia and Wahba (1999) and Smith and Todd (2005)

5.7 The average causal effect with multi-valued or multiple treatment

5.7.1 Empirical framework

- use superscripts m and l as running indices for more than two treatments
- Lechner (2001) defines three different types of treatment effects

The expected average effect of treatment m relative to treatment l for a firm drawn randomly from the population is defined as

$$\gamma^{m,l} = E(Y^m - Y^l) = E(Y^m) - E(Y^l). \tag{5.65}$$

The expected average effect of treatment m relative to treatment l for a firm randomly selected from the group of firms participating in either m or l is defined as

$$\alpha^{m,l} = E(Y^m - Y^l|S = m, l) = E(Y^m|S = m, l) - E(Y^l|S = m, l), \tag{5.66}$$

where S is the assignment indicator, defining whether a firm receives treatment m or l.

Finally, the expected average effect of treatment m relative to treatment l for a unit that is randomly selected from the group of firms participating in m only is defined as

$$\theta^{m,l} = E(Y^m - Y^l|S = m) = E(Y^m|S = m) - E(Y^l|S = m). \tag{5.67}$$

Note that

- both $\gamma^{m,l}$ and $\alpha^{m,l}$ are symmetric ...
- ... in the sense that $\gamma^{m,l} = -\gamma^{l,m}$ and $\alpha^{m,l} = -\alpha^{l,m}$, ...

- ... whereas $\theta^{m,l}$ is not, so that $\theta^{m,l} \neq -\theta^{l,m}$.

Estimates of the average treatment effects can be obtained as follows:

1. the response probabilities for each treatment can be estimated either

   - by a bivariate probability model OR
   - by a multinomial logit model

   Denote the estimated response probabilities that are a function of the vector of observable variables $\mathbf{x}$ as $\hat{P}^m(\mathbf{x})$.

2. estimate the expectation
   $E(Y^m|S = m)$ by $E\{E[Y^m|\hat{P}^m(\mathbf{x})S = m]|S \neq m\}$
   and the expectation $E(Y^l|S = m)$ by
   $E\{E[Y^l|\hat{P}^l(\mathbf{x}), \hat{P}^m(\mathbf{x})S = l]|S = m\}$.

3. apply (propensity score) matching methods as in the bivariate case:

   - radius matching
   - nearest-neighbor matching
   - kernel matching etc.

4. The average treatment effect (i.e., the outer expectation above) is estimated as the average of the difference in outcomes between the treated and the control units.

5.7.2 Standard errors

Two alternative estimates of the standard error of each of the treatment effects.

1. analytic standard errors a la Lechner (2001)

2. standard errors from subsampling a la Politis, Romano, and Wolf (1999)

Analytic standard errors

$$Var(\hat{\theta}^{m,l}) = \frac{1}{N^m} Var(Y^m|S=m) + \frac{\sum_{i\in l}(w_i^m)^2}{(\sum_{i\in l} w_i^m)^2} Var(Y^l|S=l), \tag{5.68}$$

$$\begin{aligned} Var(\hat{\alpha}^{m,l}) &= \sum_{i\in m}\left[\frac{1+w_i^l}{N^m+N^l}\right]^2 Var(Y^m|S=m) \\ &+ \sum_{i\in l}\left[\frac{1+w_i^m}{N^m+N^l}\right]^2 Var(Y^l|S=l), \end{aligned} \tag{5.69}$$

$$\begin{aligned} Var(\hat{\gamma}^{m,l}) &= \sum_{i\in m}\left[\sum_{j=0}^{M}\frac{w_i^j}{n}\right]^2 Var(Y^m|S=m) \\ &+ \sum_{i\in l}\left[\sum_{j=0}^{M}\frac{w_i^j}{n}\right]^2 Var(Y^l|S=l). \end{aligned} \tag{5.70}$$

Standard errors from subsampling In empirical applications, these analytical standard errors may deviate considerably from their small-sample-counterparts. Abadie and Imbens (2006) show that also bootstrapped standard errors cannot be relied upon. They suggest that subsampling gives reliable variance estimates of treatment effects even in small samples.

5.7.3 Application

Becker and Egger (2007): Endogenous Product versus Process Innovation and a Firm's Propensity to Export

5.8 The average causal effect with continuous treatment

Hirano and Imbens (2004) have proposed an extension of the propensity score methodology that allows for estimation of average causal effects with continuous treatments:

- random sample of units, indexed by $i = 1, \ldots, N$
- $\forall i$, postulate potential outcomes $Y_i(t)$, for $t \in \mathscr{T}$, referred to as the **unit-level dose-response function**
- in the binary treatment case $\mathscr{T} = 0, 1$
- in the continuous case, we allow $\mathscr{T}$ to be an interval $[t_0, t_1]$
- we are interested in the average dose-response function, $\mu(t) = E[Y_i(t)]$
- for each observation i, there is also a vector of covariates X_i,
- ... and the level of the treatment received, $T_i \in [t_0, t_1]$
- we *observe* the vector X_i, the treatment T_i, and the potential outcome corresponding to the level of treatment received, $Y_i = Y_i(T_i)$
- drop index i from now on
- assume that $Y(t)_{t \in \mathscr{T}}, T, X$ are defined on a common probability space, that t is cont. distributed w.r.t. Lebesgue measure on $\mathscr{T}$, and that $Y = Y(T)$ is a well defined random variable

Now generalize the unconfoundedness assumption for the binary treatment case made by Rosenbaum and Rubin (1983) to the continuous case:

Assumption 6 `Weak unconfoundedness` $Y(t) \perp T | X$ *for all* $t \in \mathscr{T}$

In words: "Conditional on the covariates the (actual) treatment (level) is independent of the potential outcomes." Put differently: "Potential treatment outcomes are independent of the assignment mechanism for any given value of a vector of attributes (X)."

This is referred to as weak unconfoundedness because it does not require *joint* independence of all potential outcomes, $Y(t)_{t \in [t_0, t_1]}, T, X$. Instead, we require conditional independence to hold for each value of the treatment (one by one).

Next, define the generalized propensity score:

Definition 8 *Generalized propensity score*
Let r(t,x) be the conditional density of the treatment given the covariates:

$$r(t,x) \; = \; f_{T|X}(t|x)$$

Then the generalized propensity score (GPS) is $R = r(T, X)$.

The GPS has a balancing property similar to that of the standard pscore: within strata with the same value of $r(t, X)$, the probability that $T = t$ does not depend on the value of X. Loosely speaking, the GPS has the property that

$$X \quad \perp \quad 1\{T = t\} | r(t, X). \tag{5.71}$$

This is a mechanical implication of the GPS, and does not require unconfoundedness. In combination with unconfoundedness this implies that assignment to treatment is unconfounded given the generalized propensity score.

Theorem 1 *Weak unconfoundedness given the GPS*
Suppose that assignment to the treatment is weakly unconfounded given pre-treatment variables X. Then, for every t,

$$f_T(t|r(t,X), Y(T)) = f_T(t|r(t,X))$$

Proof: see Hirano and Imbens (2004).

Interpretation: when we consider the conditional density of the treatment level at t, we evaluate the GPS at the corresponding level of the treatment. In that sense we use as many propensity scores as there are levels of treatment. Nevertheless, we never use more than a single score at one time.

Bias removal using the GPS

Two steps:

1. estimate the conditional expectation of the outcome as a function of two scalar variables, the treatment level T and the GPS R, $\beta(t,r) = E[Y|T=t, R=r]$
2. to estimate the dose-response function at a particular level of the treatment we average this conditional expectation over the GPS at that particular level of the treatment, $\mu(t) = E[\beta(t, r(t,X))]$. It is important to note that we do not average over the GPS $R = r(t,X)$; rather we average over the score evaluated at the treatment level of interest, $r(t,X)$; in other words, we fix t and average over X_i respectively $r(t,X_i)$ $\forall i$

Theorem 2 *Bias removal with GPS*
Suppose that assignment to treatment is weakly unconfounded given pre-treatment variables X. Then

(i) $\beta(t,r) = E[Y(t)|r(t,X)=r] = E[Y|T=t, R=r]$

(ii) $\mu(t) = E[\beta(t, r(t,X))]$

Estimation and inference

A guide to practical implementation of the GPS methodology is as follows:

1. In the **first stage**, use a normal distribution for the treatment given the covariates:

   $$T_i | X_i \sim N(\beta_0 + \beta_1' X_i, \sigma)$$

   Note: more general models may be considered (e.g. mixtures of normals, or hetoroskedastic normal distributions with the variance a parametric function of the covariates)

   In the simple normal model, we can estimate β_0, β_1, and σ^2 by maximum likelihood.

   The estimated GPS is

   $$\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\sigma^2}(T_i - \hat{\beta}_0 + \hat{\beta}_1' X_i)^2 \right)$$

2. In the **second stage**, we model the conditional expectation of Y_i given T_i and R_i as a flexible function of its two arguments, e.g. a quadratic approximation:

   $$E[Y_i | T_i, R_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 R_i + \alpha_4 R_i^2 + \alpha_5 T_i R_i$$

   We estimate these parameters by OLS using the estimated GPS $\hat{R}_i$

3. Given the estimated parameters in the second stage, in the **third stage**, we estimate the outcome at treatment level t as

$$\widehat{E[Y(t)]} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{r}(t, X_i) + \hat{\alpha}_4 \hat{r}(t, X_i)^2 + \hat{\alpha}_5 t\hat{r}(t, X_i)$$

   We do this *for each level of the treatment we are interested in*, to obtain an estimate of the **entire dose-response function**.

In practice, bootstrap standard errors are used.

5.8.1 Application

1. Hirano and Imbens (2004): Imbens-Rubin-Sacerdote lottery sample

2. Becker and Muendler (2008): effect of foreign direct investment expansion on domestic worker displacement

5.9 Comments on matching methods

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

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

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

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

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

Chapter 6

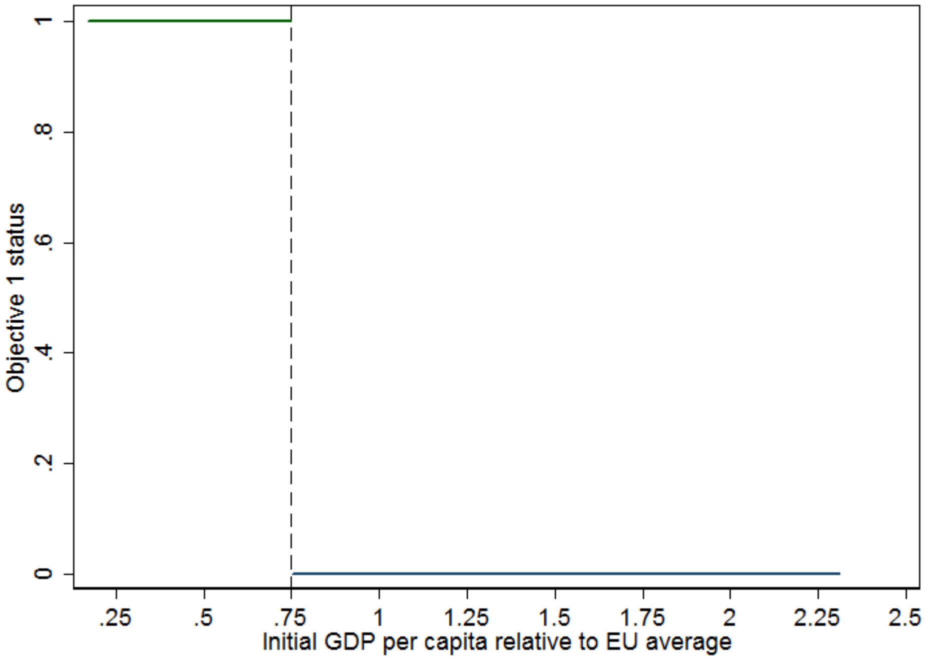
Regression Discontinuity Design

Useful overview articles and background material:

- Imbens and Lemieux (2008a)
- Imbens and Lemieux (2008b)
- Angrist and Pischke (2009), Chapter 6

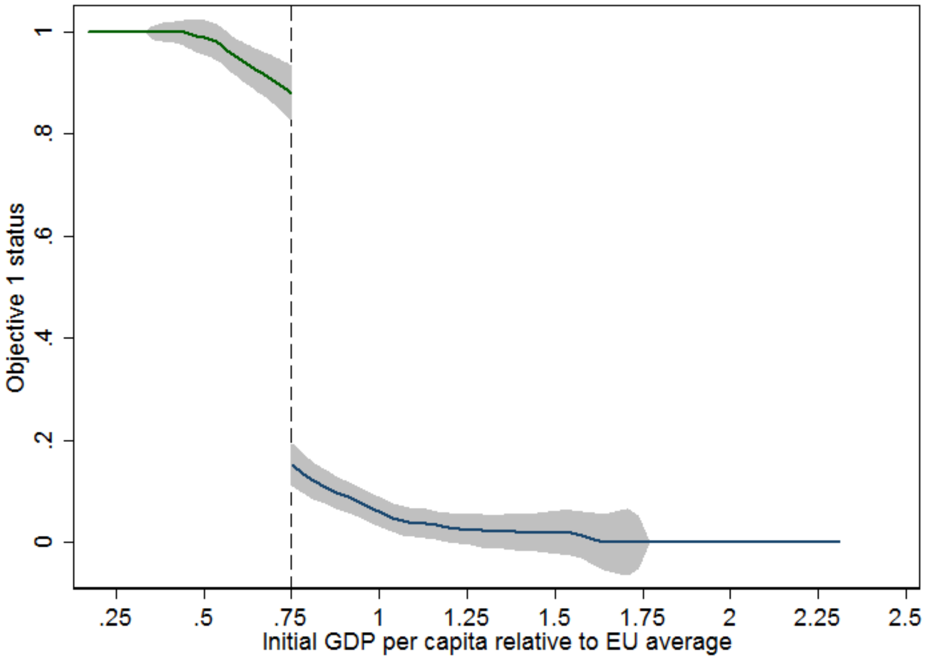
Let's start with an example:

- EU spends substantial amounts of money on structural funds
- Objective 1 (70% of structural fund budget):
  - Promote the development and structural adjustment of regions whose development is lagging behind.
  - Eligible: regions with a per-capita GDP less than 75% of the EU average



Regression discontinuity designs (RDD) exploit precise knowledge of the rules determining treatment. RDD identification is based on the idea that in a highly rule-based world, some rules are arbitrary and therefore provide good experiments. The above example is an example of a so-called *sharp* RDD.

However, as one might expect, there might be exceptions from the rule (*non-compliance*). In fact, the 75%-rule is not applied sharply:



This gives rise to a *fuzzy* RDD.

The *sharp* RDD can be seen as a selection-on-observables story:

- assignment to treatment solely depends on whether an observable pre-intervention variables satisfies a set of conditions *known* to the analyst.
- In a neighborhood of the threshold for selection a *sharp* RDD presents some features of a pure experiment.

The *fuzzy* design leads to an instrumental-variables-type setup where the assignment rule is used as an instrument for the actual treatment status.

Examples for sharp and fuzzy RDDs:

- Thistlethwaite and Campbell (1960): Certificates of Merit
- Angrist and Lavy (1999): class size effects on student performance
- van der Klaauw (2002): effect of financial aid on college enrolment
- DiNardo and Lee (2004): impacts of new unionization on firms
- Lee (2008): U.S. House elections (the winner takes it all)
- Becker et al. (2009b): EU Objective 1 funds

Some authors exploit geographic features, i.e. borders:

- Black (1999): Parental Valuation of Elementary Education (using school district boundaries)
- Lalive (2008): How do extended benefits affect unemployment duration? (using boundary between districts with different rules)
- Becker et al. (2009a): Long-Run Effects of Institutions (using boundary between Habsburg Empire and neighboring Empires in Eastern Europe)

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

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

The *sharp* RDD features two main limitations:

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

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

6.1 Treatment effects in a RDD

- (Y_1, Y_0) are the two potential outcomes induced, respectively, by participation and non-participation.

- $\beta = Y_1 - Y_0$ is the causal effect of the treatment, which is not observable.

- We consider the general case in which β may vary across units.

- I is the binary variable that denotes treatment status, with $I = 1$ for participants and $I = 0$ for non-participants.

- If the assignment is determined by randomization and subjects comply with the assignment:

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

- Given randomization, we can identify the mean impact

$$E\{\beta\} = E\{Y_1 | I = 1\} - E\{Y_0 | I = 0\}, \tag{6.1}$$

Formal characterization of an RDD

Following Battistin and Rettore (2008) and Hahn et al.(2001), a RDD arises when:

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

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

$$Pr\{I = 1|\bar{s}^+\} \neq Pr\{I = 1|\bar{s}^-\}. \tag{6.2}$$

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

Without loss of generality, we also assume

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

Sharp and Fuzzy RDD

Following Trochim (1984), the distinction between *sharp* and *fuzzy* RDD depends on the size of the discontinuity in (6.2).

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

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

$$I = 1(S \geq \bar{s}). \tag{6.3}$$

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

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

6.2 Sharp RDD

6.2.1 Identification in a sharp RDD

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

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

$$\begin{aligned} E\{Y|\bar{s}^+\} &- E\{Y|\bar{s}^-\} \qquad (6.4) \\ &= E\{Y_0|\bar{s}^+\} - E\{Y_0|\bar{s}^-\} + E\{I(s)\beta|\bar{s}^+\} - E\{I(s)\beta|\bar{s}^-\} \\ &= E\{Y_0|\bar{s}^+\} - E\{Y_0|\bar{s}^-\} + E\{\beta|\bar{s}^+\} \end{aligned}$$

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

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

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

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

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

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

Thus, we obtain a local version of the average treatment effect in (6.1)

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

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

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

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

The sharp RDD represents a special case of selection on observables (which is also discussed in Section 5).

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

$$(Y_1, Y_0) \perp I | S = \bar{s}. \tag{6.5}$$

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

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

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

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

6.2.2 Implementing a sharp RDD in a regression framework

Assignment mechanism:

$$D_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0 \end{cases} \tag{6.6}$$

where x_0 is a known threshold or cutoff. This assignment mechanism is a deterministic function of x_i because once we know x_i, we know D_i. It's a discontinuous function because no matter how close x_i gets to x_0, treatment is unchanged until $x_i = x_0$.

An interesting and important feature of RDD, highlighted in the survey of RDD by Imbens and Lemieux (2008a), is that there is no value of x_i at which we get to observe both treatment and control observations. Unlike full-covariate matching strategies, which are based on treatment-control comparisons conditional on covariate values where there is some overlap, the validity of RD turns on our willingness to extrapolate across covariate values, at least in a neighborhood of the discontinuity. This is one reason why Sharp RD is usually seen as distinct from other control strategies. For this same reason, we cannot usually afford to be as agnostic about regression functional form in the RDD world.

A simple model formalizes the RDD idea. Suppose that in addition to the assignment mechanism, (6.6), potential outcomes can be described by a linear, constant-effects model

$$\begin{aligned} E[Y_{0i}|x_i] &= \alpha + \beta x_i && (6.7) \\ Y_{1i} &= Y_{0i} + \rho && (6.8) \end{aligned}$$

This leads to the regression,

$$Y_i = \alpha + \beta x_i + \rho D_i + \eta_i \tag{6.9}$$

where ρ is the causal effect of interest.

The key difference between this regression and others that are used to estimate treatment effects is that D_i, the regressor of interest, is not only correlated with x_i, it is a deterministic function of x_i. RDD captures causal effects by distinguishing the nonlinear and discontinuous function, $1(x_i \geq x_0)$, from the smooth and (in this case) linear function, x_i.

But what if the trend relation, $E[Y_{0i}|x_i]$, is nonlinear? To be precise, suppose that $E[Y_{0i}|x_i] = f(x_i)$ for some reasonably smooth function, $f(xi)$. Now we can construct RDD estimates by fitting

$$Y_i = f(x_i) + \rho D_i + \eta_i \tag{6.10}$$

where again, $D_i = 1(x_i \geq x_0)$ is discontinuous in x_i at x_0. As long as $f(x_i)$ is continuous in a neighborhood of x_0, it should be possible to estimate a model like (6.10), even with a flexible functional form for $f(x_i)$. For example, modeling $f(x_i)$ with a p^{th}-order polynomial, RDD estimates can be constructed from the regression

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \rho D_i + \eta_i \tag{6.11}$$

Allowing for interaction terms in the sharp RDD A generalization of RDD based on (6.11) allows for different trend functions for $E[Y_{0i}|x_i]$ and $E[Y_{1i}|x_i]$. Modeling both of these CEFs with p^{th}-order polynomials, we have

$$E[Y_{0i}|x_i] = f_0(x_i) = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + ... + \beta_{0p}\tilde{x}_i^p \quad (6.12)$$
$$E[Y_{1i}|x_i] = f_1(x_i) = \alpha + \rho + \beta_{11}\tilde{x}_i + \beta_{12}\tilde{x}_i^2 + ... + \beta_{1p}\tilde{x}_i^p \quad (6.13)$$

where $\tilde{x}_i \equiv x_i x_0$. Centering x_i at x_0 is just a normalization; it ensures that the treatment effect at $x_i = x_0$ is still the coefficient on D_i in the regression model with interactions.

To derive a regression model that can be used to estimate the effects interest in this case, we use the fact that D_i is a deterministic function of x_i to write

$$E[Y_i|x_i] = E[Y_{0i}|x_i] + E[Y_{1i} - Y_{0i}|x_i]D_i$$

Substituting polynomials for conditional expectations, we then have

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}^2 + ... + \beta_{0p}\tilde{x}^p \\ &+ \gamma D_i + \beta_1^* D_i\tilde{x}_i + \beta_2^* D_i\tilde{x}_i^2 + ... + +\beta_p^* D_i\tilde{x}_i^p + \eta_i \quad (6.14) \end{aligned}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{12} - \beta_{02}$, and $\beta_p^* = \beta_{1p} - \beta_{0p}$ and the error term, η_i, is the CEF residual.

Equation (6.11) is a special case of (6.14) where $\beta_1^* = \beta_2^* = ...\beta_p^* = 0$. In the more general model, the treatment effect at $x_i - x_0 = c > 0$ is $\rho + \beta_1^* c + \beta_2^* c^2 + ... + \beta_p^* c^p$, while the treatment effect at x_0 is ρ. The model with interactions has the attraction that it imposes no restrictions on the underlying conditional mean functions. But in many practical situations, RDD estimates of ρ based on the simpler model, (6.11), usually turn out to be similar to those based on (6.14).

Using a discontinuity sample in sharp RDD The validity of RD estimates based on (6.11) or (6.14) turns on whether polynomial models provide an adequate description of $E[Y_{0i}|x_i]$. If not, then what looks like a jump due to treatment might simply be an unaccounted-for nonlinearity in the counterfactual conditional mean function. To reduce the likelihood of such mistakes, we can look only at data in a neighborhood around the discontinuity, say the interval $[x_0 - \delta, x_0 + \delta]$ for some small number δ. Sometimes people call this a *discontinuity sample* (e.g. Angrist and Lavy (1999)).

6.2.3 Evidence on the validity of the identification condition

An attractive feature of a RDD is that it allows to test the validity of the identification condition (6.5).

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

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

Finding that the two groups of subjects present systematic differences in the values of these variables would cast serious doubts on the validity of the identification condition (6.5).

6.3 Fuzzy RDD

6.3.1 Identification in a fuzzy RDD

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

In this case, the continuity of Y_0 and Y_1 at $\bar{s}$ is no longer sufficient to ensure the orthogonality condition in (6.5).

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

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

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

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

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

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

If Condition 2 holds we are in the familiar IV framework of Section 4:

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

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

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

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

The right hand side is the difference between:

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

As in the IV framework:

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

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

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

$$E\{\beta | I(\bar{s}^+) \neq I(\bar{s}^-)\} = \frac{E\{Y|\bar{s}^+\} - E\{Y|\bar{s}^-\}}{E\{I|\bar{s}^+\} - E\{I|\bar{s}^-\}}, \qquad (6.15)$$

The right hand side of (6.15) is the mean impact on those subjects in a neighborhood of $\bar{s}$ who would switch their treatment status if the threshold for participation switched from just above their score to just below it.

It is the analog of the LATE in this context.

The denominator in the right-hand side of (6.15) identifies the proportion of *compliers* at $\bar{s}$.

6.3.2 Implementing a fuzzy RDD in a regression framework

Fuzzy RDD exploits discontinuities in the probability or expected value of treatment conditional on a covariate. The result is a research design where the discontinuity becomes an instrumental variable for treatment status instead of deterministically switching treatment on or off. To see how this works, let D_i denote the treatment as before, though here D_i is no longer deterministically related to the threshold-crossing rule, $x_i \geq x_0$: Rather, there is a jump in the probability of treatment at x_0, so that

$$P[D_i = 1|x_i] = \begin{cases} g_0(x_i) & \text{if } x_i \geq x_0 \\ g_1(x_i) & \text{if } x_i < x_0 \end{cases} \tag{6.16}$$

where $g_0(x_i) \neq g_1(x_0)$. The functions $g_0(x_i)$ and $g_1(x_i)$ can be anything as long as they differ (and the more the better) at x_0. We'll assume $g_1(x_0) > g_0(x_0)$, so $x_i \geq x_0$ makes treatment more likely. We can write the relation between the probability of treatment and x_i as

$$E[D_i|x_i] = P[D_i = 1|x_i] = g_0(x_i) + [g_1(x_i) - g_0(x_i)]T_i \tag{6.17}$$

where $T_i = 1(x_i \geq x_0)$. The dummy variable T_i indicates the point of discontinuity in $E[D_i|x_i]$. Fuzzy RDD leads naturally to a simple 2SLS estimation strategy.

The simplest fuzzy RD estimator uses only T_i as an instrument. [Again, one can allow for interactions between T_i and the polynomial in x_i (see below)] The resulting just-identified IV estimator has the virtues of transparency and good finite-sample properties. The first stage in this case is

$$D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + ... + \gamma_p x_i^p + \pi T_i + \xi_{1i} \tag{6.18}$$

where T_i is the excluded instrument that provides identifying power with a first-stage effect given by γ.

The fuzzy RDD reduced form is obtained by substituting (6.18) into (6.11):

$$Y_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + ... + \kappa_p x_i^p + \rho\pi T_i + \xi_{2i} \tag{6.19}$$

where $\mu = \alpha + \rho\gamma_0$ and $\kappa_j = \beta_1 + \rho\gamma_j$ for $j = 1, ...p$. As with sharp RDD, identification in the fuzzy case turns on the ability to distinguish the relation between Y_i and the discontinuous function, $T_i = 1(x_i \geq x_0)$, from the effect of polynomial controls included in the first and second stage.

Allowing for interaction terms in the fuzzy RDD Assuming that $g_0(x_i)$ and $g_1(x_i)$ can be described by p^{th}-order polynomials as in (6.11), we have

$$\begin{aligned} E[D_i|x_i] &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + ... + \gamma_{0p}x_i^p \\ &+ (\gamma_0^* + \gamma_1^* x_i + \gamma_2^* x_i^2 + ... + \gamma_p^* x_i^p)T_i \\ &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + ... + \gamma_{0p}x_i^p \\ &+ \gamma_0^* T_i + \gamma_1^* x_i T_i + \gamma_2^* x_i^2 T_i + ... + \gamma_p^* x_i^p T_i \end{aligned} \tag{6.20}$$

From this we see that T_i, as well as the interaction terms x_iT_i, x_2T_i, ..., x^pT_i can be used as instruments for D_i in (6.11).

Fuzzy RDD estimates with treatment effects that change as a function of x_i (as just assumed in (6.20)) can be constructed by 2SLS estimation of an equation with treatment-covariate interactions. Here, the second stage model with interaction terms is the same as (6.14), while the first stage is similar to (6.20), except that to match the second-stage parametrization, we center polynomial terms at x_0. In this case, the excluded instruments are $\{T_i, \tilde{x}_iT_i, \tilde{x}_i^2T_i, ..., \tilde{x}_i^pT_i\}$ while the variables $\{D_i, \tilde{x}_iD_i, \tilde{x}_i^2T_i, ..., \tilde{x}_i^pD_i\}$ are treated as endogenous.

The first stage for D_i becomes

$$\begin{aligned} D_i &= \gamma_{00} + \gamma_{01}\tilde{x}_i + \gamma_{02}\tilde{x}_i^2 + ...\gamma_{0p} \\ &+ \gamma_0^* T_i + \gamma_1^* \tilde{x}_i T_i + \gamma_2^* \tilde{x}_i^2 T_i + ...\gamma_p^* T_i \end{aligned} \tag{6.21}$$

An analogous first stage is constructed for each of the polynomial interaction terms in the set $\tilde{x}_iD_i, \tilde{x}_i^2D_i, ..., \tilde{x}_i^pD_i$.

Using a discontinuity sample in fuzzy RDD The idea of using discontinuity samples informally also applies in this context: start with a parametric 2SLS setup in the full sample, say, based on (6.11). Then restrict the sample to points near the discontinuity and get rid of most or all of the polynomial controls. Ideally, 2SLS estimates in the discontinuity samples with few controls will be broadly consistent with the more precise estimates constructed using the larger sample.

6.4 A partially *fuzzy* design

Battistin and Rettore (2008) consider an interesting particular case:

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

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

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

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

Condition 1 says that:

$$E\{Y_0|\bar{s}^+\} = E\{Y_0|\bar{s}^-\}. \tag{6.22}$$

and

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

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

The last expression combined with (6.22) yields

$$E\{Y_0|I = 1, \bar{s}^+\} = \frac{E\{Y_0|\bar{s}^-\}}{\phi} - E\{Y_0|I = 0, \bar{s}^+\}\frac{1 - \phi}{\phi}. \tag{6.23}$$

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

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

Hence, equation (6.23) implies that the mean impact on participants is identified:

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

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

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

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

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

6.5 Comments on RDD

- A *sharp* RDD identifies the mean impact of a treatment for a broader population than the one for which identification is granted by a *fuzzy* RDD.

- Whether the parameter identified by a *fuzzy* RDD is policy relevant depends on the specific case.

- A *fuzzy* RDD requires stronger identification conditions.

- Some of the simplicity of the RDD is lost moving to a *fuzzy* design.

- Both *sharp* and *fuzzy* designs cannot identify the impact for subjects far away from the discontinuity threshold.

- A RDD framework naturally suggests ways to test the validity of the identification assumptions.

- RDDs are promising tools for the identification of causal effects.

Bibliography

Aakvik, Arild, "Bounding a Matching Estimator: The Case of a Norwegian Training Program," *The Oxford Bulletin of Economics and Statistics*, 2001, *63* (1), 115–143.

Abadie, Alberto, Joshua D. Angrist, and Guido Imbens, "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 2002, *70* (1), 91–117.

Abowd, John, Francis Kramarz, and David Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, March 1999, *67* (2), 251–334.

Ahn, Hyungtaik and James Powell, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 1993, *58* (1/2), 3–29.

Altonji, Joseph G. and Charles R. Pierret, "Employer Learning and Statistical Discrimination," *The Quarterly Journal of Economics*, 2001, *116* (1), 313–350.

_ **and Christina H. Paxson**, "Labor Supply Preferences, Hours Constraints, and Hours-Wage Trade-offs," *Journal of Labor Economics*, April 1988, *6*, 254–276.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, June 1990, *80* (3), 313–336.

_ , "Short-Run Demand for Palestinian Labor," *Journal of Labor Economics*, July 1996, *14* (3), 425–453.

_ , "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, March 1998, *66* (2), 249–288.

_ , "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 2001, *19* (1), 2–16.

_ , "Treatment Effect Heterogeneity in Theory and Practice," *The Economic Journal*, March 2004, *114*, C52–C83.

_ **and Alan B. Krueger**, "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, November 1991, *106* (4), 979–1014.

_ **and** _ , "Estimating the Payoff to Schooling Using the Vietnam-Era Draft Lottery," *National Bureau of Economic Research Working Paper*, May 1992, *4067.*

_ **and** _ , "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, April 1995, *13* (2), 225–235.

_ **and** _ , "Empirical Strategies in Labor Economics," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Amsterdam, New York and Oxford: Elsevier Science, North-Holland, 1999, pp. 1277–1366.

_ **and** _ , "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, Fall 2001, *15* (4), 69–85.

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

_ **and Jinyong Hahn**, "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, February 2004, *86* (1), 58–72.

_ **and Jörn-Steffen Pischke**, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

_ **and Victor Lavy**, "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, May 1999, *114* (2), 533–575.

_ **and** _ , "The Effect of High-Stakes High School Achievement Awards: Evidence from a School-Centered Randomized Trial," *American Economic Review*, December 2008, *forthcoming.*

_ **and William N. Evans**, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *The American Economic Review*, June 1998, *88* (3), 450–477.

_ **, Guido W. Imbens, and Donald B. Rubin**, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–472.

Arellano, Manuel, “Computing robust standard errors for withingroups estimators,” *Oxford Bulletin of Economics and Statistics*, 1987, *49*, 431–434.

Arrow, Kenneth J., “Higher Education as a Filter,” *Journal of Public Economics*, 1973, *2* (3), 193–216.

Ashenfelter, Orley, “Estimating the Effect of Training Programs on Earnings,” *Review of Economics and Statistics*, 1978, *6* (1), 47–57.

_ **and Alan B. Krueger**, “Estimates of the Economic Return to Schooling from a New Sample of Twins,” *American Economic Review*, December 1994, *84* (5), 1157–1173.

_ **and Cecilia Rouse**, “Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins,” *The Quarterly Journal of Economics*, February 1998, *113* (1), 253–284.

_ **and David Card**, “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs,” *Review of Economics and Statistics*, November 1985, *67* (4), 648–660.

_ **, Colm Harmon, and Hessel Oosterbeek**, “A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias,” *Labour Economics*, November 1999, *6* (4), 453–470.

Atkinson, Anthony and John Mickelwright, “Unemployment compensation and labor market transitions: a critical review,” *Journal of Economic Literature*, December 1991, *29* (4), 1679–1727.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger, “Computing Inequality: Have Computers Changed the Labor Market?,” *The Quarterly Journal of Economics*, November 1998, *113* (4), 1169–1213.

Battistin, Erich and Enrico Rettore, “Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs,” *Journal of Econometrics*, 2008, *142* (2), 715–730.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman, “Instrumental variables and GMM: Estimation and testing,” *Stata Journal*, 2003, *3* (1), 1–31.

Baumgartner, Hans J. and Viktor Steiner, “Student Aid, Repayment Obligations and Enrolment into Higher Education in Germany Evidence from a ”Natural Experiment”,” *Journal of Applied Social Science Studies*, 2005, *125* (1), 29–38.

Becker, Gary S., “A theory of the allocation of time,” *Economic Journal*, 1965, *75*, 493–517.

_ , *Human Capital and Personal Distribution of Income: An Analytical Approach*, Ann Arbor: University of Michigan Press, 1967.

_ , *Human Capital*, 3rd ed., New York: Columbia University Press (for National Bureau of Economic Research), 1993.

_ , "Nobel Lecture: The Economic Way of Looking at Life," *The Journal of Political Economy*, June 1993, *101* (3), 383–409.

_ **and Barry R. Chiswick**, "The Economics of Education - Eduaction and the Distribution of Earnings," *The American Economic Review*, March 1966, *56* (1/2), 358–369.

Becker, Sascha O. and Andrea Ichino, "Estimation of average treatment effects based on propensity scores," *Stata Journal*, 2002, *2* (4), 358–377.

_ **and Ludger Woessmann**, "Luther and the Girls: Religious Denomination and the Female Education Gap in 19th Century Prussia," *Scandinavian Journal of Economics*, 2008, *110* (4), 777–805.

_ **and** _ , "Was Weber Wrong? A Human Capital Theory of Protestant Economic History," *The Quarterly Journal of Economics*, May 2009, *124* (2), 531–596.

_ **and Marc-Andreas Muendler**, "The Effect of FDI on Job Security," *The B.E. Journal of Economic Analysis & Policy*, 2008, *8* (1). (Advances), Article 8. Available at: http://www.bepress.com/bejeap/vol8/iss1/art8.

_ **and Marco Caliendo**, "Sensitivity Analysis for Average Treatment Effects," *The Stata Journal*, 2007, *7* (1), 71–83.

_ **and Peter H. Egger**, "Endogenous Product versus Process Innovation and a Firm's Propensity to Export," *CESifo Working Paper 1906*, February 2007.

_ **and Robert Fenge**, "Gerechtigkeit und Effizienz nachgelagerter Studiengebühren," *ifo Schnelldienst*, 2005, *58* (2), 16–22.

_ **, Katrin Boeckh, Christa Hainz, and Ludger Woessmann**, "The Empire Is Dead, Long Live the Empire! Values and Human Interactions 90 Years after the Fall of the Habsburg Empire," *mimeo*, 2009.

_ **, Peter H. Egger, and Maximilian von Ehrlich**, "Going NUTS: The Effect of EU Structural Funds on Regional Performance," *updated version of CESifo Working Paper 2495*, 2009.

Bedard, Kelly, "Human Capital versus Signaling Models: University Access and High School Dropouts," *Journal of Political Economy*, August 2001, *109* (4), 749–775.

Behrman, Jere R. and Barbara L. Wolfe, "The Socioeconomic Impact of Schooling in a Developing Country," *The Review of Economics and Statistics*, 1984, *66* (2), 296–303.

Belzil, Christian and Jörgen Hansen, "Unobserved ability and the return to schooling," *Econometrica*, September 2002, *70* (5), 2078–2091.

Ben-Porath, Yoram, "The production of human capital and the life cycle of earnings," *The Journal of Political Economy*, 1967, *75*, 353–367.

Bergstrom, Theodore A., "Free Labor for Costly Journals?," *Journal of Economic Perspectives*, Fall 2001, *15* (4), 183–198.

Biddle, Jeff E. and Daniel S. Hamermesh, "Sleep and the Allocation of Time," *Journal of Political Economy*, October 1990, *98* (5, Part 1), 922–942.

Bjorklund, Anders and Robert Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 1987, *69* (1), 42–49.

Black, Sandra, "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, May 1999, *114* (2), 577 –599.

Blackburn, McKinley and David Neumark, "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *The Quarterly Journal of Economics*, 1992, *107* (4), 1421–36.

Bloom, Howard S., "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 1984, *8*, 225–246.

Blundell, Richard and Thomas MaCurdy, "Labor supply: A review of alternative approaches," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Elsevier Science, 1999, chapter 27.

Borjas, George J., "The Relationship between Wages and Weekly Hours of Work: The Role of Division Bias," *Journal of Human Resources*, 1980, *15* (3), 409–423.

Bound, John and David A. Jaeger, "Do Compulsory Attendance Laws Alone Explain the Association between Earnings and Quarter of Birth?," *Research in Labor Economics*, 2000, *19*, 83–108.

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

Breusch, Trevor and Adrian Rodney Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 1979, *47*, 1287–1294.

Browning, Martin, Francois Bourguignon, Pierre-Andre Chiappori, and Valerie Lechene, "Income and Outcomes: A Structural Model of Intrahousehold Allocation," *The Journal of Political Economy*, December 1994, *102* (5), 1067–1096.

Brownstone, David and Robert Valletta, "The Bootstrap and Multiple Imputations: Harnessing Increased Computing Power for Improved Statistical Tests," *Journal of Economic Perspectives*, Fall 2001, *15* (4), 129–141.

Caliendo, Marco and Sabine Kopeinig, "Some Practical Guidance for the Implementation of Propensity Score Matching," *Journal of Economic Surveys*, 2008, *22* (1), 31–72.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler, "Labor Supply of New York City Cabdrivers: One Day at a Time," *The Quarterly Journal of Economics*, May 1997, *112* (2), 407–441.

Campbell, Donald T., "Reforms as Experiments," *American Psychologist*, 1969, *XXIV*, 409–429.

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

_ , "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, eds., *Aspects of labour market behaviour: Essays in honour of John Vanderkamp*, Toronto, Buffalo and London: University of Toronto Press, 1995, pp. 201–222.

_ , "The Causal Effect of Education on Earnings," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Amsterdam, New York and Oxford: Elsevier Science, North-Holland, 1999, pp. 1801–1863.

_ **and Alan Krueger**, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, August 1994, *84* (4), 772–793.

_ **and** _ , "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply," *American Economic Review*, December 2000, *90* (5), 1397–1420.

_ **and Daniel G. Sullivan**, "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 1988, *56* (3), 497–530.

Carneiro, Pedro Manuel and James J. Heckman, "The Evidence of Credit Constraints in Post-Secondary Schooling," *The Economic Journal*, October 2002, *112*, 705–734.

Carrington, William J., "The Alaskan Labor Market during the Pipeline Era," *Journal of Political Economy*, February 1996, *104* (1), 186–218.

Cawlwy, J., J. Heckman, and E. Vytlacil, "Three observations on wages and measured cognitive ability," *Labour Economics*, September 2001, *8* (4), 419–442.

Chevalier, Arnaud, Colm Harmon, Ian Walker, and Yu Zhu, "Does Education Raise Productivity, or Just Reflect it?," *Economic Journal*, November 2004, *114* (499), F499–F517.

Chiappori, Pierre-Andre, "Rational Household Labor Supply," *Econometrica*, January 1988, *56* (1), 63–90.

_ , "Collective Labor Supply and Welfare," *The Journal of Political Economy*, June 1992, *100* (3), 437–467.

Chiswick, Barry R., "Interpreting the Coefficient of Schooling in the Human Capital Earnings Function.," *Journal of Educational Planning and Administration*, April 1998, *12* (2), 123–130.

Chou, Yuan, "Testing Alternative Models of Labour Supply: Evidence from Taxi Drivers in Singapore, Singapore Economic Review, vol.47, (1): 17-47.," *Singapore Economic Review*, 2002, *47* (1), 17–47.

Cochran, W.G., "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, June 1968, *24* (2), 295–313.

Cook, Thomas D. and Donald T. Campbell, *Quasi-Experimentation. Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company, 1979.

Cosslett, Stephen, "Semiparametric Estimation of Regression Model with Sample Selectivity," in William A. Barnett, James Powell, and George Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press, 1991.

Cox, David R., "Causality: Some Statistical Aspects," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 1992, *155* (2), 291–301.

Crépon, Bruno and Francis Kramarz, "Employed 40 Hours or Not Employed 39 Hours: Lessons from the 1982 Mandatory Reduction of the Workweek," *Journal of Political Economy*, December 2002, *110* (6), 1355–1389.

Cumby, Robert E., John Huizinga, and Maurice Obstfeld, "Two-step two-stage least squares estimation in models with rational expectations," *Journal of Econometrics*, 1983, pp. 333–355.

Davidson, Russell and James G. MacKinnon, *Estimation and inference in econometrics*, Oxford, New York, Toronto and Melbourne: Oxford University Press, 1993.

Davis, Steven J. and John Haltiwanger, "Gross Job Creation, Gross Job Destruction, and Employment Reallocation," *The Quarterly Journal of Economics*, August 1992, *107* (3), 819–863.

de la Fuente, Angel, "Human capital in a global and knowledge-based economy - Part II: assessment at the EU country level," *European Commission FINAL REPORT*, 2003.

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

_ **and** _ , "Propensity Score Matching Methods for Non-Experimental Causal Studies," *Review of Economics and Statistics*, February 2002, *84* (1), 151–161.

DellaVigna, Stefano and Daniele Paserman, "Job Search and Impatience," *Journal of Labor Economics*, July 2005, *23* (3), 527–588.

Dickens, William T. and Shelley J. Lundberg, "Hours Restrictions and Labor Supply," *International Economic Review*, February 1993, *34*, 169–192.

Diebolt, Claude, *L'evolution de longue periode du systeme educatif allemand : 19eme et 20eme siecles*, Vol. 23 of *Serie AF 23, Numero special de la revue Economies et Societes, Cahiers de l'ISMEA*, ISMEA, 1997.

DiNardo, John Enrico and David S. Lee, "The Impact of Unionization on Establishment Closure: A Regression Discontinuity Analysis of Representation Elections," *NBER Working Paper*, 2002, *8993.*

_ **and** _ , "Economic impacts of new unionization on private sector employers: 1984-2001," *Quarterly Journal of Economics*, November 2004, *119* (4), 1383–1441.

DiPrete, Thomas and Markus Gangl, "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments," *Sociological Methodology*, 2004, *34*, 271–310.

Dynarski, Susan M., "The Behavioral and Distributional Implications of Aid for College," *The American Economic Review Papers and Proceedings*, May 2002, *92* (2), 279–285.

_ , "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion," *The American Economic Review*, March 2003, *93* (1), 279–288.

Efron, Bradley, "Bootstrap Methods: Another Look at the Jackknive," *Annals of Statistics*, 1979, *7*, 1–26.

_ , *The Jackknife, the Bootstrap, and Other Resampling Plans*, Vol. 38 of *Cbms-Nsf Regional Conference Series in Applied Mathematics*, New York and London: Society for Industrial & Applied Mathematics, 1982.

_ , "The Bootstrap and Modern Statistics," *Journal of the American Statistical Association*, December 2000, *95* (452), 1293–1296.

_ **and Robert J. Tibshirani**, *An introduction to the bootstrap*, Vol. 57 of *Monographs on Statistics and Applied Probability*, New York and London: Chapman and Hall, 1993.

Ehrenberg, Ronald G., "Econometric studies of higher education," *Journal of Econometrics*, 2004, *121*, 19–37.

_ **and Robert S. Smith**, *Modern Labor Economics: Theory and Public Policy*, Addison Wesley Longman, 2000.

Eissa, Nada, "Comments on James Heckman's "Policies to Foster Human Capital"," *Research in Economics*, 2000, *54* (1), 75–80.

_ **and Jeffrey B. Liebman**, "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics*, May 1996, *111* (2), 605–637.

Fan, Jianqing, "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 1992, *87*, 998–1004.

_ **and Irène Gijbels**, *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall, 1996.

Farber, Henry S., "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers," *Journal of Political Economy*, February 2005, *113* (1), 46–82.

Fehr, Ernst and Lorenz Goette, "Do Workers Work More When Wages Are High? Evidence from a Randomized Field Experiment," *American Economic Review*, 2007, *97* (1), 298–317.

Freedman, David A., “Statistical Models and Shoe Leather,” *Sociological Methodology*, 1991, *21*, 291–313.

Freeman, Richard B., *The Overeducated American*, San Diego: Academic Press, 1976.

Gerfin, Michael and Michael Lechner, “A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland,” *The Economic Journal*, October 2002, *112*, 854–893.

Gómez-Salvadora, Ramón, Julián Messina, and Giovanna Vallanti, “Gross job flows and institutions in Europe,” *Labour Economics*, 2004, *11*, 469–485.

Goux, Dominique and Eric Maurin, “Education, experience et salaires,” *Economie et prevision*, 1994, *116*, 155–179.

Greene, William H., *Econometric analysis*, 4th ed., Prentice Hall International, 2000.

_ , *Econometric analysis*, 5th ed., Prentice Hall International, 2003.

Griliches, Zvi, “Wages of Very Young Men,” *The Journal of Political Economy*, August 1976, *84* (4, part 2), S69–S86.

_ , “Estimating the Returns to Schooling: Some Econometric Problems,” *Econometrica*, January 1977, *45*, 1–21.

Gronau, Reuben, “Home production,” in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Vol. 1, Elsevier Science / North-Holland, 1986, chapter 4.

_ , “The theory of home production: The past ten years,” *Journal of Labor Economics*, 1997, *15* (2), 197–205.

Hahn, J., P. Todd, and W. Van der Klaauw, “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design,” *Econometrica*, January 2001, *69* (1), 201–209.

Hahn, Jinyong, “On the role of the propensity score in efficient semi-parametric estimation of average treatment effects,” *Econometrica*, 1998, *66* (2), 315–331.

Haley, William J., “Human Capital: The Choice between Investment and Income,” *The American Economic Review*, December 1973, *63* (5), 929–944.

Hall, Peter, *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag, 1992.

_ , **Stephen J. Marron, and D. Michael Titterington**, "On Partial Local Smoothing Rules for Curve Estimation," *Biometrika*, September 1995, *82* (3), 575–587.

Ham, John C., "Estimation of a Labour Supply Model with Censoring Due to Unemployment and Underemployment," *Review of Economic Studies*, July 1982, *49*, 335–354.

Hamermesh, Daniel S., "Myth and Measurement: The New Economics of the Minimum Wage: Review Symposium: Comment," *Industrial and Labor Relations Review*, July 1995, *48* (4), 835–838.

_ **and Neal M. Soss**, "An Economic Theory of Suicide," *Journal of Political Economy*, January/February 1974, *82*, 83–98.

_ **and Stephen J. Trejo**, "The Demand for Hours of Labor: Direct Evidence from California," *The Review of Economics and Statistics*, February 2000, *82* (1), 38–47.

Hanoch, Giora, "An Economic Analysis of Earnings and Schooling.," *Journal of Human Resources*, Summer 1967, *2* (3), 310–329.

Hansen, Lars, John Heaton, and Amir Yaron, "Finite-sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, July 1996, *14* (3), 262–280.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, July 1982, *50* (4), 1029–1054.

Harmon, C. and I. Walker, "Estimates of the Economic Returns to Schooling for the United Kingdom," *American Economic Review*, 1995, *85* (5), 1278–1286.

Hausman, Jerry A., "Specification tests in econometrics," *Econometrica*, November 1978, *46* (6), 1251–1271.

Haveman, Robert H. and Barbara Wolfe, "Schooling and Economic Well-Being: The Role of Non-Market Effects," *Journal of Human Resources*, 1984, *19* (3), 377–407.

Hayashi, Fumio, *Econometrics*, Princeton and Oxford: Princeton University Press, 2000.

Heckman, James J., "Shadow prices, market wages and labor supply," *Econometrica*, July 1974, *42* (4), 679–694.

_ , "A Life-Cycle Model of Earnings, Learning, and Consumption," *The Journal of Political Economy*, August 1976, *84* (4, Part 2: Essays in Labor Economics in Honor of H. Gregg Lewis), S11–S44.

_ , "Dummy Endogenous Variable in a Simultaneous Equations System," *Econometrica*, 1978, *46* (4), 931–960.

_ , "Sample Specification Bias as a Specification Error," *Econometrica*, 1979, *33*, 181–214.

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

_ , "What has been learned about labor supply in the past twenty years?," *American Economic Review, Papers and Proceedings*, 1993, *83* (2), 116–121.

_ , "Randomization as an Instrumental Variable," *Review of Economics and Statistics*, May 1996, *78* (2), 336–341.

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

_ , "Policies to foster human capital," *Research in Economics*, 2000, *54* (1), 3–56.

_ , "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, August 2001, *109* (4), 673–748.

_ **and Bo Honore**, "The Empirical Content of the Roy Model," *Econometrica*, September 1990, *58* (5), 1121–49.

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

_ **and G. Sedlacek**, "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self–Selection in the Labor Market," *Journal of Political Economy*, 1985, *93*, 1077–1125.

_ **and Richard Robb**, "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, 1985, pp. 156–245.

_ **and Salvador Navarro-Lozano**, "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics*, 2004, *86* (1), 30–57.

_ , **Hidehiko Ichimura, and Petra Todd**, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, October 1997, *64* (4), 605–654.

_ , _ , **and** _ , "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, April 1998, *65* (2), 261–294.

_ , _ , **Jeffrey Smith, and Petra Todd**, "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences*, November 1996, *93*, 13416–13420.

_ , _ , _ , **and** _ , "Characterizing Selection Bias Using Experimental Data," *Econometrica*, September 1998, *66* (5), 1017–1098.

_ , **Justin L. Tobias, and Edward Vytlacil**, "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, October 2001, *68* (2), 210–223.

_ , **Lance Lochner, and Christopher Taber**, "Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy," *Fiscal Studies*, March 1999, *20* (1), 25–40.

Hirano, Keisuke and Guido W. Imbens, "The propensity score with continuous treatments," in Andrew Gelman and Xiao-Li Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Wiley, 2004, chapter 7.

_ , _ , **and Geert Ridder**, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, July 2003, *71* (4), 1161–1189.

_ , _ , **Donald B. Rubin, and Xiao-Hua Zhou**, "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 2000, *1* (1), 69–88.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, December 1986, *81* (396), 945–970.

Holzer, Harry and David Neumark, "Assessing Affirmative Action," *Journal of Economic Literature*, September 2000, *38* (3), 483–568.

Horowitz, Stanley A. and Allan Sherman, "A Direct Measure of the Relationship Between Human Capital and Productivity.," *Journal of Human Resources*, Winter 1980, *15* (1), 67–76.

Hotz, V. Joseph, Finn E. Kydland, and Guilherme L. Sedlacek, "Intertemporal Preferences and Labor Supply," *Econometrica*, March 1988, *56* (2), 335–360.

Huffman, Wallace E., "Black-White Human Capital Differences: Impact on Agricultural Productivity in the U.S. South," *American Economic Review*, March 1981, *71* (1), 94–107.

Hujer, Reinhard, Marco Caliendo, and Stephan L. Thomsen, "New Evidence on the Effects of Job Creation Schemes in Germany - A Matching Approach with Threefold Heterogeneity," *Research in Economics*, 2004, *58*, 257–302.

Hunt, Jennifer, "Has Work-Sharing Worked in Germany?," *Quarterly Journal of Economics*, February 1999, *114* (1), 117–148.

Hyslop, Dean and Guido W. Imbens, "Bias from Classical and Other Forms of Measurement Error," *Journal of Business and Economic Statistics*, 2001, *19* (4), 475–481.

Ichino, Andrea, *The Problem of Causality in Microeconometrics*, European University Institute: http://www2.dse.unibo.it/ichino, 2006.

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

_ **and** _ , "The Long-Run Educational Cost of World War Two," *Journal of Labor Economics*, 2004, *22* (1), 57–86.

_ **, Guido Schwerdt, Rudolf Winter-Ebmer, and Josef Zweimüller**, "Too Old to Work, Too Young to Retire," *mimeo*, 2007.

Imbens, Guido W., "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 2000, *87* (3), 706–710.

_ **and Donald B. Rubin**, "Bayesian Inference for Causal Effects in Randomized Experiments with Noncomplianc," *The Annals of Statistics*, 1997a, *25*, 305–327.

_ **and** _ , "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 1997b, *64*, 555–574.

_ **and Joshua D. Angrist**, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62*, 467–475.

_ **and Thomas Lemieux**, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 2008, *142* (2), 615–635.

_ **and** _ , "Special issue editors introduction: The regression discontinuity design-Theory and applications," *Journal of Econometrics*, 2008, *142* (2), 611–614.

_ , **Donald Rubin, and Bruce Sacerdote**, "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review*, August 2001, *91* (4), 778–794.

Jaeger, David and Marianne Page, "Degrees matter: New evidence on sheepskin effects in the returns to education," *Review of Economics and Statistics*, November 1996, *78* (4), 733–740.

Johnes, Geraint, *The Economics of Education*, Macmillan, 1993.

Johnson, Richard W. and David Neumark, "Wage Declines among Older Men," *The Review of Economics and Statistics*, November 1996, *78* (4), 740–748.

Kane, Thomas J., *The Price of Admission: Rethinking How Americans Pay for College*, Brookings Institution Press, 1999.

_ **and Cecilia E. Rouse**, "Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?," *American Economic Review*, June 1995, *85* (3), 600–614.

Kling, Jeffrey R., "Interpreting Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business and Economic Statistics*, July 2001, *19* (3), 358–364.

Krueger, Alan B. and Lawrence H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, March 1988, *56* (2), 259–294.

Lalive, Rafael, "How do extended benefits affect unemployment duration? A regression discontinuity approach," *Journal of Econometrics*, 2008, *142* (2), 785–806.

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

Lang, Kevin, "Ability Bias, Discount Rate Bias and the Returns to Education," *Boston University*, 1993, *mimeo.*

_ **and David Kropp**, "Human Capital Versus Sorting: The Effects of Compulsory Attendance Laws," *The Quarterly Journal of Economics*, August 1986, *101* (3), 609–624.

Lavy, Victor, "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," *American Economic Review*, 2008, *forthcoming.*

Layard, Richard and George Psacharopoulos, "The Screening Hypothesis and the Returns to Education," *The Journal of Political Economy*, Sep. - Oct. 1974, *82* (5), 985–998.

Lazear, Edward Paul, "Education: Consumption or production?," *The Journal of Political Economy*, 1977, *85*, 569–597.

_ , "Firm-Specific Human Capital: A Skill-Weights Approach," *NBER Working Paper*, May 2003, *9679.*

Lechner, Michael, "Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption," in Michael Lechner and Friedhelm Pfeiffer, eds., *Econometric Evaluation of Active Labour Market Policies*, Heidelberg: Physica, 2001.

_ , "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *Review of Economics and Statistics*, May 2002, *84* (2), 205–220.

_ , "Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods," *Journal of the Royal Statistical Society - Series A*, 2002, *165* (1), 59–82.

Lee, David S., "Randomized Experiments from Non-random Selection in U. S. House Elections," *Journal of Econometrics*, 2008, *142* (2), 675–697.

_ **and David Card**, "Regression discontinuity inference with specification error," *Journal of Econometrics*, 2008, *142* (2), 655–674.

Lee, Lung-Fei, "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 1982, *49* (3), 335–372.

_ , "Generalized Econometric Models with Selectivity," *Econometrica*, 1983, *51* (2), 507–512.

Lemieux, Thomas and Kevin Milligan, "Incentive effects of social assistance: A regression discontinuity approach," *Journal of Econometrics*, 2008, *142* (2), 807–828.

Leuven, Edwin and Barbara Sianesi, "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing," *http://ideas.repec.org/c/boc/bocode/s432001.html*, 2003, *version 3.0.0.*

_ **, Hessel Oosterbeek, and Bas van der Klauuw**, "The effect of financial rewards on students' achievement: Evidence from a randomized experiment," *Journal of the European Economics Association*, 2009, *forthcoming.*

Lochner, Lance and Enrico Moretti, "The Effect of Education on Criminal Activity: Evidence from Prison Inmates, Arrests and Self-Reports," *American Economic Review*, 2004, *84* (4), 155–189.

Lundberg, Shelly, "The Added Worker Effect," *Journal of Labor Economics*, January 1985, *3* (1, Part 1), 11–37.

Maddala, G. S., *Limited dependent and qualitative variables in econometrics*, Cambridge, England: Cambridge University Press, 1983.

Mantel, N. and W. Haenszel, "Statistical Aspects of Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 1959, *22*, 719–748.

Matsudaira, Jordan D., "Mandatory summer school and student achievement," *Journal of Econometrics*, 2008, *142* (2), 829–850.

McCallum, B. T., "Relative Asymptotic Bias from Errors of Omission and Measurement," *Econometrica*, July 1972, *40* (4), 757–758.

McCrary, Justin, "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

McPherson, Michael S. and Morton Owen Schapiro, "Does Student Aid Affect College Enrollment? New Evidence on a Persistent Controversy," *The American Economic Review*, March 1991, *81* (1), 309–318.

Meltzer, David O., "Mortality Decline, the Demographic Transition and Economic Growth." PhD dissertation, University of Chicago 1992.

Mills, John P., "Table of the Ratio - Area to Bounding Ordinate, for Any Portion of Normal Curve," *Biometrika*, November 1926, *18* (3/4), 395–400.

Mincer, Jacob, *Schooling, Experience and Earnings*, New York: National Bureau of Economic Research, 1974.

_ , "Human capital and the labor market: A review of current research," *Educational Researcher*, 1989, *18* (4), 27–34.

Moretti, Enrico, "Estimating the Social Return to Higher Education: Evidence From Longitudinal and Repeated Cross-Sectional Data," *Journal of Econometrics*, 2004, *121* (1-2), 175–212.

Moulton, Brent, "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 1986, *32*, 385–397.

Murnane, Richard J., John B. Willett, and Frank Levy, "The Growing Importance of Cognitive Skills in Wage Determination," *Review of Economics and Statistics*, 1995, *77* (2), 251–266.

_ , _ , and _ , "The growing importance of cognitive skills in wage determination," *The Review of Economics and Statistics*, May 1995, *77* (2), 251–266.

Nelson, Charles R. and Richard Startz, "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimation," *Econometrica*, July 1990, *58* (4), 967–976.

Neumark, David and William Wascher, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment.," *American Economic Review*, December 2000, *90* (5), 1362–1396.

_ , Junfu Zhang, and Stephen Ciccarella, "The effects of Wal-Mart on local labor markets," *Journal of Urban Economics*, 2008, *63*, 405–430.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, May 1987, *55* (3), 703–708.

Oettinger, Gerald S., "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors," *Journal of Political Economy*, April 1999, *107* (2), 360–392.

Olsen, Randall, "A Least Squares Correction for Selectivity Bias," *Econometrica*, 1980, *48* (7), 1815–1820.

Pagan, Adrian Rodney and Alistar D. Hall, "Diagnostic tests as residual analysis," *Econometric Reviews*, 1983, *2* (2), 159–218.

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge: Cambridge University Press, 1986.

Psacharopoulos, George, "Returns to Education: A Further International Update and Implications," *Journal of Human Resources*, 1985, *XX*, 583–604.

_ , *Economics of education; Research and studies*, Oxford, UK: Pergamon Press., 1987.

_ , "Returns to investment in education: A global update," *World Development*, September 1994, *22* (9), 1325–1343.

_ and Harry Anthony Patrinos, "Returns to investment in education: a further update," *Education Economics*, August 2004, *12* (2), 111–134.

Robinson, Chris, "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy*, June 1989a, *97* (3), 639–667.

_ , "Union Endogeneity and Self-Selection," *Journal of Labor Economics*, January 1989b, *7* (1), 106–112.

Rosen, Sherwin, "A Theory of Life Earnings," *The Journal of Political Economy*, August 1976, *84* (4, Part 2: Essays in Labor Economics in Honor of H. Gregg Lewis), S45–S67.

_ , "Human Capital," in et al. John Eatwell, ed., *The New Palgrave: A Dictionary of Economics*, London: Macmillian, 1987, pp. 681–90.

Rosenbaum, P. R., "The Consequences of Adjustment for a Concomitant Variable that has been Affected by the Treatment," *Journal of the Royal Statistical Society. Series A*, 1984, *147*, 656–666.

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

_ , "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, June 1987, *82* (398), 387–394.

_ , "The Role of a Second Control Group in an Observational Study," *Statistical Science*, August 1987, *2* (3), 292–306.

_ , *Observational Studies*, 2nd ed., Springer-Verlag, 2002.

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

_ **and** _ , "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 1985, *39* (1), 33–38.

Roy, A. D., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 1951, *3*, 135–146.

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

_ , "Bias Reduction Using Mahalanobis-Metric Matching," *Biometrics*, June 1980, *36* (2), 293–298.

Schultz, Theodore W., "Capital Formation by Education," *Journal of Political Economy*, December 1960, *68* (6), 571–583.

_ , "Investment in Human Capital," *The American Economic Review*, March 1961, *51* (1), 1–17.

Sianesi, Barbara, "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s," *Review of Economics and Statistics*, February 2004, *86* (1), 133–155.

Sicherman, Nachum, ""Overeducation" in the Labor Market," *Journal of Labor Economics*, April 1991, *9* (2), 101–122.

Smith, Jeffrey and Petra Todd, "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, March-April 2005, *125* (1-2), 305–353.

Speed, T.P., "Introductory Remarks on Neyman (1923)," *Statistical Science*, 1990, *5* (4), 463–464.

Spence, Michael, "Job Market Signalling," *Quarterly Journal of Economics*, August 1973, *87*, 205–221.

Splawa-Neyman, J., "On the Application of Probability Theory to Agricultual Economics. Essay on Principles. Section 9," *Statistical Science*, 1990, *5* (4), 465–480.

Staiger, Douglas and J. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65* (3), 557–586.

Stock, James H. and Mark W. Watson, *Introduction to Econometrics*, Addison Wesley, 2003.

Thistlethwaite, Donald L. and Donald T. Campbell, "Regression discontinuity analysis: an alternative to the ex post facto experiment," *Journal of Educational Psycology*, December 1960, *51* (6), 309–317.

Topel, Robert, "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority," *The Journal of Political Economy*, February 1991, *99* (1), 145–176.

Topel, Robert H. and Michael P. Ward, "Job mobility and the careers of young men," *Quarterly Journal of Economics*, May 1992, *107* (2), 439–479.

Trochim, William M. K., *Research Design for Program Evaluation: the Regression-Discontinuity Approach*, Beverly Hills: Sage Publications, 1984.

van der Klaauw, Wilbert, "Estimating the Effect of Financial Aid Offers on College Enrollment: a Regression-Discontinuity Approach," *International Economic Review*, November 2002, *43* (4), 1249–1287.

Weiss, Andrew, "Human Capital vs. Signalling Explanations of Wages," *Journal of Economic Perspectives*, Winter 1995, *9* (4), 133–154.

Welch, Finis, "Myth and Measurement: The New Economics of the Minimum Wage: Review Symposium: Comment," *Industrial and Labor Relations Review*, July 1995, *48* (4), 842–849.

White, Halbert L., "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 1980, *48*, 817–838.

_ , "Instrumental variables regression with independent observations," *Econometrica*, 1982, *50* (2), 483–499.

_ , *Asymptotic Theory for Econometricians*, Orlando, FL: Academic Press, 1984.

Wickens, Michael R., "A Note on the Use of Proxy Variables," *Econometrica*, July 1972, *40* (4), 759–761.

Wigger, Berthold U. and Robert K. von Weizsaecker, "Risk, Resources, and Education: Public versus Private Financing of Higher Education," *IMF Staff Papers*, 2001, *48* (3), 547–560.

Willis, Robert J. and Sherwin Rosen, "Education and Self-Selection," *The Journal of Political Economy*, October 1979, *87* (5, Part 2: Education and Income Distribution), S7–S36.

Wolfe, Barbara and Robert H. Haveman, "Accounting for the Social and Non-Market Benefits of Education," in "The Contribution of Human and Social Capital to Sustained Economic Growth and Well-Being," International Symposium Report edited by the OECD and HRDC., 2001.

Woodbury, Stephen A. and Robert G. Spiegelman, "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois," *The American Economic Review*, September 1987, *77* (4), 513–530.

Wooldridge, Jeffrey M., "Applications of Generalized Method of Moments Estimation," *Journal of Economic Perspectives*, Fall 2001, *15* (4), 87–100.

_ , *Econometric analysis of cross section and panel data*, Cambridge and London: MIT Press, 2002.

_ , *Introductory Econometrics: A Modern Approach*, 2nd ed., South Western College Publishing, 2003.

Wu, De-Min, "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, July 1973, *41* (4), 733–750.

Yatchew, Adonis and Zvi Griliches, "Specification Error in Probit Models," *Review of Economics and Statistics*, February 1985, *67* (1), 134–139.