Corso di Metodi Applicati per la Valutazione dei Servizi in Economia e Finanza

-Applied Methods for Services Evaluation in Economics and Finance-

SEF-LM56. a.a. 2021-22

See “How Do we Know if a Program Made a Difference” Section 2.1

The focus of this course is to study the evaluation methods of policy interventions associated with welfare programs, training programs, wage subsidy programs, and tax credit programs.

The aim of program impact evaluation is to learn whether and to what degree a program influenced the outcomes from what otherwise might have been happen.

Our reasoning is based on the convincement that the heart of the program evaluation is a missing data problem. An individual may either be subject to the policy intervention or he/she may not, but no one individual can be in both states simultaneously.

In this context, there would be no evaluation problem if we could observe the counterfactual outcome for those included in the program such as had they not participated

The choice of evaluation method will depend on three broad concerns:

-the nature of the question to be answered,

-the type and quality of data available,

-the mechanism by which individuals are allocated to the program or receive the policy.

Last of these is typically labelled the "assignment rule" and will be a central component in the analysis we present.

Assignment Rule

In a perfectly designed experiment, assignment should be random. In a structural microeconomic model, assignment is assumed to obey some rules from economic theory.

Alternative methods exploit different assumptions concerning assignment and differ according to the type of assumption made.

In general, we consider six distinct, but related each other, approaches:

(i) social experiment methods,

(ii) natural experiment methods,

(iii) discontinuity design methods,

(iv) matching methods,

(v) instrumental variable methods,

(vi) control function methods.

The Social Experiment method

The social experiment method is the most convincing method of evaluation since it directly constructs a control (or comparison) group, which is a randomized subset of the eligible population.

Since programs are typically voluntary, those individuals "randomized in" may decide not to participate in the treatment (no-compliers). This circumstance may lead to the introduction of a non-random component into the composition of both treated and control groups. The measured program impact will therefore take into account an "intention to undergo treatment" parameter.

An example of a well-conducted social experiment is the Canadian Self Sufficiency Project (SSP), which was designed to measure the earnings and employment responses of single mothers on welfare to a time-limited earned income tax credit program. This study has produced relevant evidence on the effectiveness of financial incentives in inducing welfare recipients into work (see Card and Robbins, 1998).

The Natural Experiment approach

The natural experiment approach attempts to find a naturally occurring comparison group that can mimic the properties of the control group in the properly designed experiment. This method is also often labeled "difference-in-differences" because it is usually implemented by comparing the difference in average behavior before and after the treatment (reform, intervention) for the eligible group with the before and after difference of a comparison group.

The evaluation of the "New Deal for the Young Unemployed" in the United Kingdom is a good example of a policy design suited to this approach. It was an initiative to provide work incentives to unemployed individuals aged 18 to 24. The program is mandatory and was rolled out in selected pilot areas prior to the national roll out. The Blundell et al. (2004) study investigates the impact of this program by using similar 18-24 years old in nonpilot areas as a comparison group.

The Discontinuity Design method

The discontinuity design method can also be classified as a natural experiment approach but one that exploits situations where the probability of enrollment into treatment changes discontinuously with some continuous variable. For example, where eligibility to an educational scholarship depends on parental income falling below some cutoff.

It turns out to be convenient to discuss this approach after studying the properties of the instrumental variable (IV) estimator, since the parameter identified by discontinuity design is a sort of "local" average treatment effect similar to the parameter identified by IV, but not necessarily the same. Our aim is to compare the IV and discontinuity-design approaches.

The Matching method

The aim of matching is simple: to link individuals each other according to sufficient observable factors to remove systematic differences in the evaluation outcome between treated and untreated.

For this "selection on observables" approach, a clear understanding of the determinants of assignment rule on which the matching is based is essential.

The measurement of returns to education, where scores from prior ability tests are available in birth cohort studies, is a good example.

As we document below, matching methods have been extensively refined and their properties examined in the recent evaluation literature, and they are now a valuable part of the evaluation toolbox. Lalonde (1986) and Heckman, Ichimura, and Todd (1998) demonstrate that experimental data can help in evaluating the choice of matching variables.

The Instrumental Variable (*IV*) approach

The instrumental variable (IV) method is a standard econometric approach to control for endogeneity. It relies on finding one or more variables excluded from the outcome equation but which may be determinant of the assignment rule (moment conditions).

These variables (instruments) serves to randomly identify the selection of an individual into the treatment.

Work by Imbens and Angrist (1994) and Heckman and Vytlacil (1999) provided an ingenious interpretation of the *IV* estimator in terms of local treatment effect parameters. We will discuss these developments.

The Control Function method

The control function (CF) method directly characterizes the choice problem facing individuals deciding on program participation.

It is, therefore, closest to a structural microeconomic analysis. It uses the full specification of the assignment rule together with an excluded "instrument" to derive a control function which, when included in the outcome equation, controls for endogenous selection. This approach relates directly to the selectivity estimator of Heckman (1979).

In particular, given a more general Two-Regime model:

$y_i = d(\mathbf{x}'_i\beta_1)+(1-d)(\mathbf{x}'_i\beta_0)+u_i$,

d=1 indicates that the subject is undergone to treatment; d = 0 indicates, at the opposite, that the subject is not undergone to treatment.

The control function estimator (CF) considers the endogeneity of the treatment indicator, d, as a censored variable problem.

Treatment Effect Parameters

We now begin to consider some of the various types of program impact that might be of interest.

Following conventional terminology, first, we havethe Average Treatment Effect (ATE):

$$E(y_{1i}\text{-}y_{0i}) \qquad (1)$$

where $E(\ldots)$ is the expectations operator. The subscript 1 or 0 indicates, respectively, if the subject belongs to treated regime or to untreated regime. The Average Treatment Effect is the average impact of the program across all of the individuals in thepopulation of interest.

Treatment effects can be also conditional to some (observed) covariates. Then the "conditional" ATE effect is given by the following expected value:

$$E(y_{1i}\text{-}y_{0i})|\mathbf{x}'_i \qquad (2)$$

Another common parameter of interest is the Average Effect of Treatment on the Treated (ATT):

$E(y_{1i}\text{-}y_{0i})|d_i=1;\mathbf{x}'_i$ (3)

This is the impact of the program on those actually exposed to the program($d_i = 1$).

A third usual parameter is the is the Average Effect of Treatment on the Untreated (ATU):

$E(y_{1i}\text{-}y_{0i})|d_i= 0;\mathbf{x}'_i$ (4)

This is the hypothetical impact of the program on those actually are not exposed to the program ($1\text{-}d_i = 0$) as if they had been treated.

Brief recalls on basic preparatory topics

Truncated normal standard random variable(Johnson and Kotz, 1970; Maddala, 1983; Verbeek, 2017).

Let $z = (x-\mu)/\sigma$ a "left truncated" normal standard random variable, with z_0 as truncation point.

We have the following *pdf* function (observed):

$$\phi\left(z \middle| z > z_0\right) = \frac{\phi(z)}{1-\Phi(z_0)} \qquad (1)$$

and expected value:

$$E\left(z \middle| z > z_0\right) = \frac{\phi(z_0)}{1-\Phi(z_0)} \qquad (2)$$

In the case of right truncation, we have:

pdf function:

$$\phi\left(z\middle|z<z_0\right)=\frac{-\phi\left(z\right)}{\Phi\left(z_0\right)} \tag{3}$$

and expected value:

$$E\left(z\middle|z<z_0\right)=\frac{-\phi\left(z_0\right)}{\Phi\left(z_0\right)} \tag{4}$$

Note how the generalized residuals of Probit estimation in the two regimes are similar to the values of the density function of the observed part of two truncated standard normals, respectively to the right and to the left of the same truncation point.

Estimation of Logit and Probit models

Let's assume that the latent (continuous) variable y^*_i (given, for instance, by the propensity of a subject to be employed) is defined by the regression relationship:

$$y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + u_i \qquad (1)$$

In practice, y^*_i is unobservable. What we observe is a response (dummy) variable y_i defined by:

$y_i = 1$ if $y^*_i > 0$

$y_i = 0$ otherwise

from the Eq. (1), we get:

$$Pr(y_i = 1) = Pr\left(u_i > -\mathbf{x}'_i \boldsymbol{\beta}\right) = 1 - F\left(-\mathbf{x}'_i \boldsymbol{\beta}\right) = F\left(\mathbf{x}'_i \boldsymbol{\beta}\right) \qquad (2)$$

analogously, we have:

$$Pr(y_i = 0) = 1 - F\left(\mathbf{x}'_i \boldsymbol{\beta}\right) \qquad (3)$$

Hence, we specify the Likelihood function for both Logit and Probit:

$$L(u_i; \boldsymbol{\beta}) = \prod_{y=1} F(\mathbf{x}'_i \boldsymbol{\beta}) \prod_{y=0} \left[1 - F(\mathbf{x}'_i \boldsymbol{\beta})\right] = \prod_{y_i} F(\mathbf{x}'_i \boldsymbol{\beta})^y \left[1 - F(\mathbf{x}'_i \boldsymbol{\beta})\right]^{1-y} = \max \qquad (4)$$

while the Log-likelihood function is:

$$\ln L(u_i;\boldsymbol{\beta}) = \sum_n \left\{ y_i \ln F\left(\mathbf{x'}_i\boldsymbol{\beta}\right) + (1-y_i)\ln\left[1-F\left(\mathbf{x'}_i\boldsymbol{\beta}\right)\right]\right\} = \max \tag{5}$$

and the Score function is given by:

$$SCORE = \frac{\partial \ln L(u_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_n \left\{ \frac{[y_i - F(\mathbf{x'}_i\boldsymbol{\beta})] f(\mathbf{x'}_i\boldsymbol{\beta})}{F(\mathbf{x'}_i\boldsymbol{\beta})[1-F(\mathbf{x'}_i\boldsymbol{\beta})]} \right\} \mathbf{x}_i = 0 \tag{6}$$

If the cdf$F(\,) = \Phi(\,)$ is a Normal Standard, as in a Probit model, we have:

$$F(\mathbf{x'}_i\boldsymbol{\beta}) = \Phi(\mathbf{x'}_i\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x'}_i\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} e^{u^2/2} du \quad \text{and} \quad 1-F(\mathbf{x'}_i\boldsymbol{\beta}) = 1-\Phi(\mathbf{x'}_i\boldsymbol{\beta}) = 1-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mathbf{x'}_i\boldsymbol{\beta}} e^{u^2/2} du$$

and the Score function is:

$$\frac{\partial \ln L(u_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_n \left\{ \frac{[y_i - \Phi(\mathbf{x'}_i\boldsymbol{\beta})]\phi(\mathbf{x'}_i\boldsymbol{\beta})}{\Phi(\mathbf{x'}_i\boldsymbol{\beta})[1-\Phi(\mathbf{x'}_i\boldsymbol{\beta})]} \right\} \mathbf{x}_i = 0 \tag{7}$$

The part of the Score function of Probit model included in brackets {..} is known as "Generalized Residual" (GR). If the subject experienced the event, choosing the regime y_i =1, the value of *GR* is obtained by substituting y_i =1 into $\frac{\left[y_i - \Phi(\mathbf{x'}_i\boldsymbol{\beta})\right]\phi(\mathbf{x'}_i\boldsymbol{\beta})}{\Phi(\mathbf{x'}_i\boldsymbol{\beta})\left[1-\Phi(\mathbf{x'}_i\boldsymbol{\beta})\right]}$:

$$GR_{yi=1} = \frac{\phi(\mathbf{x'}_i\boldsymbol{\beta})}{\Phi(\mathbf{x'}_i\boldsymbol{\beta})} \quad (8)$$

At the opposite, if the subject did not experience the event, choosing the regime y_i =0, the value of *GR* is obtained by substituting y_i = 0:

$$GR_{yi=0} = \frac{-\phi(\mathbf{x'}_i\boldsymbol{\beta})}{1-\Phi(\mathbf{x'}_i\boldsymbol{\beta})} \quad (9)$$

Report in a table the number of both predicted and observed values of y_i:

		$\hat{y}_{0i} \leq 0.5$	$\hat{y}_{1i} > 0.5$	
		0	1	Tot
y_i	0	$\hat{n}_{00}$	$\hat{n}_{01}$	n_0
	1	$\hat{n}_{10}$	$\hat{n}_{11}$	n_1
		$\hat{n}_0$	$\hat{n}_1$	n

Censored Regression - Tobit II model

In this regression model, the (partial) observability of the dependent variable, y_i, depends on another variable, d_i^*, only observed in a dichotomous form, such as dummy (d_i= 0; 1) with 0 and 1= corner solutions. In practice, y_i is observed if d_i =1; y_i is censored if d_i = 0.

In addition, the realization of d_i^* endogenously depends on the level of y_i, as in the case of the wage equation, in which the wage,y_i, is observable only if the subject is employed (d_i = 1), and censored if the subject is not employed (d_i = 0). On the other hand, the decision to be employed depends also on the level of y_i.

This model can be specified as follows:

$$y_i^* = \mathbf{x'}_i\boldsymbol{\beta} + u_i \qquad u_i \sim N\left(0;\sigma^2\right)$$
$$d_i^* = \mathbf{z'}_i\boldsymbol{\gamma} + v_i \qquad v_i \sim N\left(0;1\right) \qquad (10)$$

$y_i^* = \mathbf{x'}_i\boldsymbol{\beta} + u_i$ is the "outcome" equation, while $d_i^* = \mathbf{z'}_i\boldsymbol{\gamma} + v_i$ is the "selection" equation.

Censoring rule:

$y_i^*>0$ if$d_i= 1$ and $d_i= 1$ if $d_i^*> 0$

$y_i^*= 0$ if$d_i^*= 0$ and $d_i= 0$ if $d_i^* = 0$

Error terms are related each one according to the following covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_{uv} \\ \sigma_{uv} & 1 \end{bmatrix} \qquad (11)$$

Where the covariance σ_{uv} is a measure of endogeneity of censoring. In particular, we assume that error terms relation may be specified adopting a linear combination such as:

$$u_i = \sigma_{uv} v_i + \varepsilon_i \qquad \varepsilon_i \sim IID \qquad with \qquad mean = 0 \qquad (12)$$

Given the conditional expected value of the partially observed dependent variable y_i^*:

$$E\left(y_i^* \middle| d_i = 1\right) = E\left(y_i^* \middle| d_i^* > 0\right) = E\left(y_i^* \middle| \mathbf{z_i}'\boldsymbol{\gamma} + v_i > 0\right) = E\left(\mathbf{x_i}'\boldsymbol{\beta} + u_i \middle| v_i > -\mathbf{z_i}'\boldsymbol{\gamma}\right) = E\left(\mathbf{x_i}'\boldsymbol{\beta} + \sigma_{uv} v_i + \varepsilon_i \middle| v_i > -\mathbf{z_i}'\boldsymbol{\gamma}\right) \qquad (13)$$

and

$$E\left(\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}v_i+\varepsilon_i \,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right)=E\left(\mathbf{x_i}'\boldsymbol{\beta}\,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right)+E\left(\sigma_{uv}v_i \,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right)+E\left(\varepsilon_i \,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right)$$

Considering $\mathbf{x_i}'\boldsymbol{\beta}$ as a deterministic component of the outcome equation, and ε_i as an independent and unconditioned random disturbance, we obtain, as a result:

$$E\left(y_i^* \,|\, d_i=1\right)=\mathbf{x_i}'\boldsymbol{\beta}+E\left(\sigma_{uv}v_i \,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right)=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}E\left(v_i \,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right) \qquad (14)$$

v_i is a normal standard random variable, left truncated in $-\mathbf{z}_i'\boldsymbol{\gamma}$. Applying the Johnson-Kotz theorems, we have:

$$E\left(y_i^* \,|\, d_i=1\right)=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}E\left(v_i \,|\, v_i>-\mathbf{z_i}'\boldsymbol{\gamma}\right)=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}\frac{\phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)}{1-\Phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)}=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}\frac{\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)} \qquad (15)$$

Adopting the same rationale, we derive the expected value of y_i under right censoring:

$$E\left(y_i^* \,|\, d_i=0\right)=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}E\left(v_i \,|\, v_i<-\mathbf{z_i}'\boldsymbol{\gamma}\right)=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}\frac{-\phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)}{\Phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)}=\mathbf{x_i}'\boldsymbol{\beta}+\sigma_{uv}\frac{-\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{1-\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)} \qquad (16)$$

The ratios $\frac{\phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}$ and $\frac{-\phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}{1-\Phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}$ can be considered as correction terms for endogeneity due, respectively, to left and right censoring in dependent variable.

Following the Heckman approach (Heckman, 1979), we can adopt a simple two stage estimation procedure.

For example, in estimating Equation (15), at a first stage we derive the generalized residuals, $\frac{\phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}$, estimating by Probit the selection equation $d_i^* = \mathbf{z'}_i\boldsymbol{\gamma} + v_i$.

At a second stage we estimate by OLS the outcome equation, only on the sub-sample d_i =1, introducing the generalized residuals as an additional regressor, to correct the estimates for the endogeneity of the censoring:

$$\hat{y}_i = \mathbf{x'}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}_{uv}\frac{\phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z'}_i\hat{\boldsymbol{\gamma}})} \qquad (17)$$

INSTRUMENTAL VARIABLES (IV) ESTIMATOR

Linear Model: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$

$\underset{n\times 1}{\mathbf{y}}$ =vector of the dependent variable ; $\underset{n\times k}{\mathbf{X}}$ = matrix of *k*-1 explanatory variables plus a constant ; $\underset{k\times 1}{\boldsymbol{\beta}}$ = vector of *k*coefficients ; $\underset{n\times 1}{\mathbf{u}}$ = vector of errors

$$\underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_i \\ \dots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underset{n\times k}{\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1;k-1} & 1 \\ x_{21} & x_{22} & \dots & x_{2;k-1} & 1 \\ \dots & \dots & \dots & \dots & 1 \\ x_{i1} & x_{i2} & \dots & x_{i;k-1} & 1 \\ \dots & \dots & \dots & \dots & 1 \\ x_{n1} & x_{n2} & \dots & x_{n;k-1} & 1 \end{bmatrix}}_{\mathbf{X}} \underset{k\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_{k-1} \\ \beta_0 \end{bmatrix}}_{\boldsymbol{\beta}} + \underset{n\times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_i \\ \dots \\ u_n \end{bmatrix}}_{\mathbf{u}}$$

error terms distribution: $\mathbf{u} \sim N(0; \sigma^2\mathbf{I})$

Where $\sigma^2\mathbf{I}$ is a *n* x *n* covariance matrix:

$$\underset{n\times n}{\sigma^2\mathbf{I}} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix} \Rightarrow$$ omoskedastic and uncorrelated errors.

In addition, the condition $E(\mathbf{X'u})=0$ is assumed, and implies that the regressors matrix, **X**, is not stochastic or, at least, not correlated with the errors. For this reason, the column-vectors included as explanatory variables in **X**, are considered as "exogenous".

Specifying the model for a single *i-th* observation, we have :

$y_i = \mathbf{x}'_i\beta + u_i$, with $u_i \sim N(0; \sigma^2)$

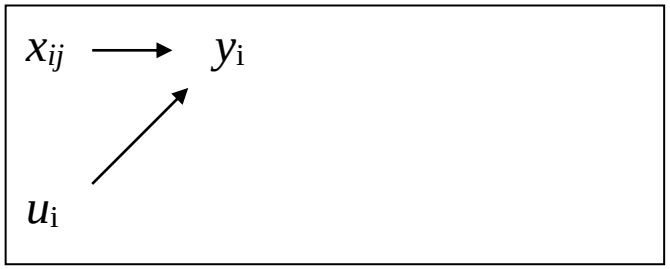
where $\mathbf{x}'_i$ is a *i-th* row vector of the matrix **X**,

while the scalar product $\mathbf{x}'_i\beta$ is equal to $\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_{k-1} x_{i;k-1} + \beta_0$,

The results of Standard OLS regression are based on the assumption that $E(x_{ij}u_i)=0$.

Namely, the regressors are not correlated with the errors: $cov(x_{ij}u_i)=0$

The assumption $E(\mathbf{X'u})=0$ (or $E(x_{ij}u_i)=0$) involves that the only the effect of x_{ij} on y_i is a "direct effect" via the term $\mathbf{x'}_i\boldsymbol{\beta}$, as represented in the following path analysis diagram:

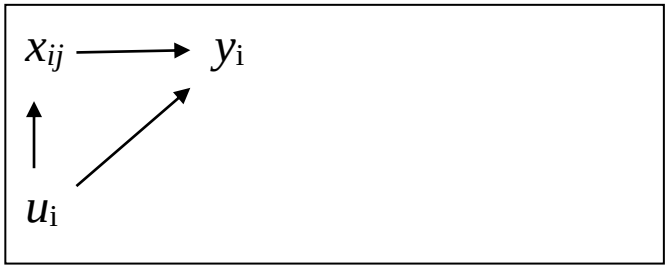


Where there is no association between x_{ij}andu_i. So x_{ij}andu_i are independent causes of y_i.

However, in some situations there may be an association between regressors and errors.

For example, consider the regression of individual earnings function (y_i). Education is included as a regressor, and it is measured by years of schooling (x_{ij}). The error term, u_i, includes latent factors (such as individual ability) that influence, jointly, individual earnings and education.

In this case, a more appropriate path diagram is the following:



where now there is an association between x_{ij}andu_{i}

What are the consequences of this correlation between x_{ij}andu_{i} , using *OLS* estimator?

<u>Using Ordinary Least Squares (*OLS*) estimator</u>:

What if $E(\mathbf{X'u})\neq 0$ or $cov(x_{ij}u_i)\neq 0$?

Consequences:

A relevant consequence of $E(\mathbf{X'u})\neq 0$ or$cov(x_{ij}u_i)\neq 0$ on *OLS* estimates is the following:

1) $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ is biased:

$$\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'u}; \qquad \Rightarrow E(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \neq \mathbf{0}$$

2) $\hat{\boldsymbol{\beta}}_{OLS}$ is inconsistent:

$$p\lim(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = \lim_{n\to\infty}\left(\frac{1}{n}\mathbf{X'X}\right)^{-1} \lim_{n\to\infty}\frac{1}{\sqrt{n}}\mathbf{X'u} = p\lim\left(\frac{1}{n}\mathbf{X'X}\right)^{-1} p\lim\frac{1}{\sqrt{n}}\mathbf{X'u} \neq 0$$

As a result, we obtain that the inequality $p\lim \frac{1}{\sqrt{n}} \mathbf{X'u} \neq \mathbf{0}$ or $\left(p\lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij} u_i \neq \mathbf{0} \right)$ converges to a non-zero value. This is a direct consequence of the Weak Law of large Numbers (WLLN), observing that *p*lim is the probability limit of the average of *n* vectors (*k* x 1), each of which has **non-zero** expected value, given $cov(\mathbf{Xu}) \neq 0$ **or** $cov(x_{ij}u_i) \neq 0$.

In general when $p\lim \frac{1}{\sqrt{n}} \mathbf{X'u} \neq \mathbf{0}$ $\left(\text{or} \quad p\lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij} u_i \neq \mathbf{0} \right)$ at least one among the explanatory variables in **X**, (e.g. x_j) is said to be **endogenous.**

<u>Endogeneity</u>of **X**[E(**X'u**) $\neq 0$ or $cov(x_{ij}u_i) \neq 0$] **usually occurs because of one of the following three causes:**

- **Omitted explanatory variables**: Additional variables that, because of data unavailability, cannot be included as regressors. An example is given by the omitted individual ability in a wage equation, where individual's years of schooling (observed explanatory variable) is correlated with the unobserved ability.

- **Measurement error:** When we can observe only an imperfect measure of an explanatory variable, x_j. An example is given by the individual income in a consumption equation estimated using microdata. The "reported"individual income is often systematically understated by the interviewed.

- **Simultaneity**: When an explanatory variable, x_j, is determined simultaneously along with the dependent variable y_i. For example, if y_i is the time that a subject devotes to domestic work and the explanatory variable x_j is his/her market working time. Market working time of the subject is partly determined by the his/her commitment in housework (reversed causality).

How to correct the influence of the endogenous explanatory variables on the estimation of $\mathbf{y = X\beta+u?}$

Instrumental Variables (*IV*) method

Let **Z**be a *n* x k matrix of the *n* column vectors, in which the second column reports the variable z_{i2} in substitution of the endogenous variable x_{i2}:

$$\underset{n\times k}{\mathbf{Z}} = \begin{bmatrix} x_{11} & z_{12} & \dots & x_{1;k-1} & 1 \\ x_{21} & z_{22} & \dots & x_{2;k-1} & 1 \\ \dots & \dots & \dots & \dots & 1 \\ x_{i1} & z_{i2} & \dots & x_{i;k-1} & 1 \\ \dots & \dots & \dots & \dots & 1 \\ x_{n1} & z_{n2} & \dots & x_{n;k-1} & 1 \end{bmatrix}$$

the variable z_{i2}, assumed as observable, is defined as an *instrumental variable* (*IV*) if it satisfies the following two conditions:

1) z_i must be not correlated with the errors u_i [$cov(z_{i2}u_i)= 0$];

2) z_{i2} must be correlated with the endogenous variable x_{i2} of the matrix $\mathbf{X}$[$cov(z_{i2}x_{i2}) \neq 0$].

In short, to conduct *IV* estimations, we need to have instrumental variables that are uncorrelated with the errors but partially and sufficiently correlated with the endogenous regressors. Then the matrix $\mathbf{Z}$ must include only independent variables x_i, not correlated with the error term, and instrumental variables, z_i.

IV estimator in the multivariate model

The *IV* estimator is given by: $\hat{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$

The covariance matrix, $Var\left(\hat{\boldsymbol{\beta}}_{IV}\right)$, is obtained by replacing the model specification, **y = Xβ+u** ,into $\hat{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ and deriving the expected value $E\left(\hat{\boldsymbol{\beta}}_{IV} - B\right)\left(\hat{\boldsymbol{\beta}}_{IV} - B\right)'$. The result is given by:

$$Var\left(\hat{\boldsymbol{\beta}}_{IV}\right) = \sigma^2\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1} \text{ with } \hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k}$$

The Two-Stage Least Squares (2SLS) Method

Again, let's consider a population model:

y = Xβ+u or $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_{k-1} x_{i;k-1} + u_i$

where x_{i1} is an endogenous regressor.
Suppose that further *m* exogenous variables (instruments) $z_i = z_1; z_2; \ldots; z_m$ are correlated with x_{i1} but not with the error u_i.

Then, let's consider $\hat{x}_{1i} = \gamma_0 + \gamma_2 x_{i2} + .. + \gamma_{k-1} x_{ik-1} + \pi_1 z_{1i} + \pi_2 z_{2i} + ... + \pi_m z_{mi}$ (or $\hat{x}_{1i} = \mathbf{x'_i}\,\boldsymbol{\gamma} + \mathbf{z'_i}\,\boldsymbol{\pi}$) as a linear projection of x_{i1} with all exogenous variables x_i and instruments z_i .

Thus, we can say that by estimating x_{i1}(for example, by OLS) with all exogenous regressors and instruments (reduced-form estimation), we obtain: $\hat{x}_{1i} = \mathbf{x'_i}\,\boldsymbol{\gamma}_{OLS} + \mathbf{z'_i}\,\boldsymbol{\pi}_{OLS}$ and $x_{1i} = \hat{x}_{1i} + \varepsilon_i$,

where $\hat{x}_{1i}$, estimated by OLS, is notcorrelated with u_i (while ε_i is correlated with u_i).

Then $\hat{x}_{1i}$, estimated by OLS, can be used as an instrumental variable of x_1 to estimate y_i.

SUMMARIZING THE TWO-STEP PROCEDURE:

Step 1: Obtain $\hat{\mathbf{x}}_\mathbf{1}$ estimating x_1 by OLS against all exogenous variables, including all of instruments (the first-stage regression)

Step 2: Use $\hat{\mathbf{x}}_\mathbf{1}$ in the place of $\mathbf{x_1}$ to estimate $\mathbf{y}$(by OLS) against $\hat{\mathbf{x}}_\mathbf{1}$ and all of exogenous independent variables, not instruments (the second stage regression)

SOCIAL EXPERIMENTS

Suppose that an evaluation is proposed in which it is possible to run a social experiment that randomly chooses individuals from a group to be administered the treatment.

By implementing randomization, one ensures that the treated and the nontreated groups are equal in all aspects apart from the treatment status.

In terms of the treatment effects model we consider the following simple Two-Equation model:

$y_{1i} = \mathbf{x}'_i\beta + \alpha + u_{1i}$

$y_{0i} = \mathbf{x}'_i\beta + u_{0i}$

or, alternatively:

$y_i = d_i(\mathbf{x}'_i\beta+\alpha)+(1-d_i)(\mathbf{x}'_i\beta) + d_iu_{1i} + (1-d_i)u_{0i} = d_i(\mathbf{x}'_i\beta+\alpha)+(1-d_i)(\mathbf{x}'_i\beta)+\varepsilon_i$

with d_i = 0;1, dummy indicator variable of exposure to treatment; and α = treatment parameter recognized as *ATE*.

Random assignment implies the following two key assumptions:

1) $E(u_{1i} | d_i = 1) = E(u_{0i} | d_i = 0) = E(\varepsilon_{1i})$
2) $E(\alpha | d_i = 1) = E(\alpha)$

One possible problem for randomization concerns dropout behaviour. For simplicity, suppose a proportion p of the eligible population used in the experiment prefer not to be treated and when drawn into the treatment group decide not to comply with treatment. Noncompliance influences the treatment parameter as follows:

$$\alpha^* = (1-p)E(\alpha)$$

which is a fraction of the ATE.
A relevant problem arises if compliance to the program is not observable. In this case, the treatment parameter, α, is not identifiable. If, on the other hand, compliance is observable, the ATE parameter can be identified obtaining an estimate p^* of p.

Another possible problem results from the complexity of contemporaneous policies in developed countries and the availability of similar alternative treatments accessible to experimental controls.

The experiment itself may affect experimental controls as, for instance, excluded individuals may obtain other available treatments, which, in some cases, is the same treatment but accessed through different channels.

This would amount to another form of noncompliance, whereby controls obtain the same treatment of treated administered through different channels.

NATURAL EXPERIMENTS

The natural experiment method makes use of naturally occurring phenomena that may induce some form of randomization across individuals in the eligibility or the assignment to treatment.

Typically, this method is implemented using a before and after comparison across groups. It is formally equivalent to a difference-in-differences (DID) approach which uses some naturally occurring event to create a "policy" shift for one group and not another.

The policy shift may refer to a change of law in one jurisdiction but not another, to some natural disaster, which changes a policy of interest in one area but not another, or to a change in policy that makes a certain group eligible to some treatment but keeps a similar group ineligible.

The difference between the two groups before and after the program change is compared, thereby creating a DID estimator of the program impact.

Difference-in-Differences (DID) approach to natural experiment

DID approach explores a change in policy occurring at some time period *t*, which introduces the possibility of receiving treatment for some individuals belonging to population, and not to receive any treatment for the remaining subjects.

It then uses longitudinal data, where the same individuals are followed over time, or repeated cross section data, where samples are drawn from the same population before and after the intervention, to identify some average impact of treatment.

We start by considering the evaluation problem when longitudinal data is available:

Each individual is observed before and after the policy change, at time t_0, before the treatment, and at t_1, after the treatment, d_i.

Let d_{it}(0;1 for *i*-th subject at time *t*) denote the treatment status of *i*-th individual at time *t* and d_i (0;1 without the time subscript) be the treatment group to which *i*-th individual belongs to.

Difference-in-Differences (DID) Estimator

[DID Method: Theory and Application pag. 72 “Handbook of Impact Evaluation]

The DID estimator adopts the assumption of no selection due to the transitory change, d_i, so that we can rewrite the previous Outcome Equation as follows:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha d_{it} + u_{it}$$

$$\mathrm{E}[u_{it1} - u_{it0} | d_i = 1] = \mathrm{E}[u_{it1} - u_{it0} | d_i = 0] = \mathrm{E}[u_{it1} - u_{it0}] = 0.$$

Then, we can write:

$$E\left[y_{it} \mid d_i; t\right] = \begin{cases} \mathbf{x}'_i \boldsymbol{\beta} + E\left(\alpha \mid d_i = 1\right) + E\left(u_{it} \mid d_i = 1\right) & \text{if } d_i = 1 \quad \text{and } t_i = t_{1i} \\ \mathbf{x}'_i \boldsymbol{\beta} + E\left(u_{it} \mid d_i = 0\right) & \text{otherwise} \end{cases}$$

If we can eliminate the error components, the ATT parameter is identified:

$$\alpha_{ATT} = E(\alpha \mid d_i = 1) = \{E[y_{it} \mid d_i = 1; t = t_1] - E[y_{it} \mid d_i = 1; t = t_0]\} - \{E[y_{it} \mid d_i = 0; t = t_1] - E[y_{it} \mid d_i = 0; t = t_0]\}$$

The corresponding sample estimator: $\hat{\alpha}_{ATT} = \hat{\alpha}_{DID} = \left[\bar{y}_{t1}^{1} - \bar{y}_{t0}^{1}\right] - \left[\bar{y}_{t1}^{0} - \bar{y}_{t0}^{0}\right]$

is the DID estimator.

An alternative approach to obtain the ATT parameter is to estimate $\hat{\alpha}_3$ by running an OLS regression on the following DID model:

$y_{it} = \mathbf{x}'_{it}\beta + \alpha_1 t_i + \alpha_2 d_i + \alpha_3 (t_i d_{it}) + u_{it}$

where: t_i = dummy of the period (equal to 0, before the treatment; equal to 1 after the treatment)

d_i = dummy indicating the status, namely if the *i*-th subject has been treated (0 = untreated; 1 treated)

$t_i d_{it}$ = dummy indicating the status, namely if the *i*-th subject has been treated status

We can interpret estimated parameters of DID model as follows:

$$\hat{\beta}_0 = \hat{E}\left(y_i \mid t_i = 0; d_i = 0\right)$$

$$\hat{\alpha}_1 = \hat{E}\left(y_i \mid t_i = 1; d_i = 0\right) - \hat{E}\left(y_i \mid t_i = 0; d_i = 1\right)$$

$$\hat{\alpha}_2 = \hat{E}\left(y_i \mid t_i = 0; d_i = 1\right) - \hat{E}\left(y_i \mid t_i = 0; d_i = 0\right)$$

$$\hat{\alpha}_3 = \hat{\alpha}_{DID} = \left[\hat{E}\left(y_i \mid t_i = 1; d_i = 1\right) - \hat{E}\left(y_i \mid t_i = 0; d_i = 1\right)\right]$$

$$-\left[\hat{E}\left(y_i \mid t_i = 1; d_i = 0\right) - \hat{E}\left(y_i \mid t_i = 0; d_i = 0\right)\right]$$

Weaknesses of DID estimator

i) Selection on idiosyncratic temporary shock

The DID procedure does not control for unobserved temporary individual-specific shocks that influence the participation decision.

To illustrate the conditions such inconsistency might arise, suppose a training program is being evaluated in which enrolment is more likely if a temporary dip in earnings occurs just before the program takes place (see Ashenfelter l978; Heckman and Smith 1999).

As a reaction to the temporary decline, a faster earnings growth is expected among the treated, even without program participation. Thus, the DID estimator is likely to overestimate the impact of treatment

ii) Differential in macro-trends

Identification of ATT using DID relies on the assumption that both treated and untreated experience common trends or, in other words, the same macro shocks. If this is not the case, DID will not consistently estimate the ATT. Differential trends might arise in the evaluation of training programs if treated and untreated operate in different groups (For example, unemployment in different age groups is often found to respond differently to cyclical fluctuations).

Non-linear DID models

A restrictive feature of the DID method is the imposition of additive separability of the error term conditional on the status, d_i, and the observable covariates, $\mathbf{x}_i$:

$$E[u_{it1} - u_{it0}|d_i = 1; \mathbf{x}_i] = E[u_{it1} - u_{it0}|d_i = 0; \mathbf{x}_i] = E[u_{it1} - u_{it0}| \mathbf{x}_i] = 0$$

Blundell et al. (2004) noted that linearity in the error term can be particularly unrealistic when, for example, the outcome is given by a dummy variable $y_i(0;1)$.

Recently, Athey and Imbens (2006) developed a general nonlinear DID methodespecially suited for continuous outcomes: the "changes-in-changes" (CIC) estimator. An extension to the discrete case in which the outcome variable is observed as binary (0;1) and the estimated outcome is a probability is also considered by the authors.

DID estimation of a logarithmic function

Since the dependent variable is expressed as a logarithm, the impact in percentageterms of a dummy variable, such as the DID coefficient (α_3), on the logarithm of outcome is the exponential term [exp(α_3)−1]. Thus, we can compare those who underwent the treatment between the two waves and those who did not, obtaining the difference of changes in percentage terms (population average effect). In addition, we can evaluate as a percentage the impact of treatment only for those subjects who underwent to treatment (subject-specific effect) as [exp (α_1+α_3)−1].

DID estimation of a simultaneous-equation system

To estimate the effect of a treatment on a simultaneous variation of different outcomes, a longitudinal model using a Difference-in-Difference (DID) parametric structure may be implemented in a simultaneous-equation system specifying the error terms as in a Seemingly Unrelated Regressions (SUR) model.
An example of specification of SUR-DID model is the following:

$$\left\{\begin{array}{l} y_{1ti} = \mathbf{x}'_{ti}\,\boldsymbol{\beta}_1 + \alpha_{11}t_i + \alpha_{12}d_i + \alpha_{13}d_i * t_i + u_{1ti} \\ y_{2ti} = \mathbf{x}'_{ti}\,\boldsymbol{\beta}_2 + \alpha_{21}t_i + \alpha_{22}d_i + \alpha_{23}d_i * t_i + u_{2ti} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{mti} = \mathbf{x}'_{ti}\,\boldsymbol{\beta}_m + \alpha_{m1}t_i + \alpha_{m2}d_i + \alpha_{m3}d_i * t_i + u_{mti} \end{array}\right\} = \begin{bmatrix} \underset{n\times 1}{\mathbf{y}_1} \\ \underset{n\times 1}{\mathbf{y}_2} \\ \ldots \\ \underset{n\times 1}{\mathbf{y}_m} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 & 0 & 0 \\ 0 & \mathbf{X}_2 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \mathbf{X}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \ldots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} \underset{n\times 1}{\mathbf{u}_1} \\ \underset{n\times 1}{\mathbf{u}_2} \\ \ldots \\ \underset{n\times 1}{\mathbf{u}_m} \end{bmatrix} =$$

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta}_* + \mathbf{u}_*$$

Where $\mathbf{X}_1$, $\mathbf{X}_2$….$\mathbf{X}_m$ are n x k block matrices of regressors, DID dummies included.

Let us assume that latent variables common to the *m* equations influences the outcomes, we specify the stochastic component of the model as follows (for the sake of simplification we consider here a model with *m* = 2 simultaneous equations):

$$\underset{2n\times 2n}{\boldsymbol{\Omega}} = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 & \sigma_{12} & ... & 0 & 0 \\ 0 & \sigma_1^2 & ... & 0 & 0 & \sigma_{12} & ... & 0 \\ ... & 0 & ... & ... & 0 & 0 & ... & 0 \\ 0 & 0 & ... & \sigma_1^2 & 0 & 0 & ... & \sigma_{12} \\ \sigma_{12} & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & \sigma_{12} & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 \\ ... & 0 & ... & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & ... & \sigma_{12} & 0 & 0 & ... & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_1^2 \end{bmatrix} \otimes \underset{n\times n}{\mathbf{I}} = \underset{2\times 2}{\boldsymbol{\Sigma}} \otimes \underset{n\times n}{\mathbf{I}}$$

The covariance matrix of the error terms of a Two-Equation model reports the two variances of the errors, σ^2_1 and σ^2_2 in the diagonal, while the covariance $\sigma_{12} = \sigma_{21}$ between the error terms is introduced to specify the influence of latent common factors on both equations.
Adopting a Feasible Generalized Least Square (FGLS) estimator, we obtain the following estimation result:

$$\beta_{SUR} = \left(\mathbf{X}'_* \boldsymbol{\Omega}^{-1} \mathbf{X}_*\right)^{-1} \mathbf{X}'_* \boldsymbol{\Omega}^{-1} \mathbf{y}_* = \left[\mathbf{X}'_* \left(\boldsymbol{\Sigma} \otimes \mathbf{I}\right)^{-1} \mathbf{X}_*\right]^{-1} \mathbf{X}'_* \left(\boldsymbol{\Sigma} \otimes \mathbf{I}\right)^{-1} \mathbf{y}_*$$

The covariance matrix $\hat{\boldsymbol{\Omega}} = \left(\hat{\boldsymbol{\Sigma}} \otimes \mathbf{I}\right)$ can be preliminarily estimated computing variances and covariances of the residuals obtained by a first-stage OLS estimate of each equation. This procedurecan be iteratively replicated to improve efficiency in estimation results.

The SUR estimator, performing a FGLS procedure, is generally considered consistent and more efficient than the corresponding OLS estimator applied to each equation separately, if the correlation between the disturbances across the equations is high or, at least, moderate (see, among others, Srivastava and Giles 1987, pp. 70–71). The gain in efficiency using SUR, however, could be nullified if conditions for inconsistent estimates occur.

One thing to remember, in this regard, is that the FGLS–SUR estimator is consistent only if the explanatory variables in each equation are not correlated with the errors in each equation.
This means that if the specification of, say, the first equation suffers for omitted explanatory variables or for measurement errors, this also affects inconsistency in the estimates of the other equations. This implies that the problems due to misspecification may be amplified performing FGLS-SUR.

In practice, the SUR approach may perform better than other estimators only if the surveyed data used for the analysis allow a correct specification of the model.

INSTRUMENTAL VARIABLES (IV) METHOD

[SEE, ALSO HANDBOOK OF IMPACT EVALUATION: PAG 87 AND SUBSEQUENT]

Let's start by considering the following two-regime simultaneous equation model:

Eq. 1 (outcome of treated subjects)

$$y_{1i} = \mathbf{x}_i'\beta_1 + \alpha + u_{1i} u_1 \sim N(0, \sigma_1^2) \qquad (1)$$

Eq. 2 (outcome of untreated subjects)

$$y_{0i} = \mathbf{x}_i'\beta_0 + u_{0i} \qquad u_0 \sim N(0, \sigma_0^2) \qquad (2)$$

Selection Equation:

$$d_i^* = \mathbf{z}_i'\gamma + v_i \qquad (3)$$

with d_i= 1 if $d_i^* = \mathbf{z}_i'\gamma + v_i > 0$, and d_i= 0 otherwise.

The error term, v_i, is distributed as a normal standard, $v_i \sim N(0,1)$.

Endogeneity of the choice of the regime is specified by introducing proper assumptions on the relations between the error terms of the outcome equations, u_1 e u_0, and the error term of the selection equation, v_i :

$u_{1i} = \sigma_{1v} v_i + \varepsilon_{1i}$ error specification in eq. 1

$u_{0i} = \sigma_{0v} v_i + \varepsilon_{0i}$error specification in eq.2

where$\varepsilon_{1i}$and ε_{0i} are independent and identically distributed (IID) distrbances.σ_{1v} e σ_{0v} are the covariances between the error terms of each outcome equation and the error term of the selection equation.

The IV approach requires the existence of at least one regressor exclusive to the decision rule, say z_j. In our notation, this is included in the regressors set **z**. z_j is known as the instrument.

The instrument z_j affects participation only, and so it is not included in **x**. This is known as the "exclusion restriction" rule. It implies that the potential outcomes do not vary with z_j and any difference in the mean observed outcomes of two groups, differing only with respect to z_j, can only be due to consequent differences in the participation rates and composition of the treatment group.

When the treatment effect is homogeneous, so that α = ATE = ATT, only differences in participation rates subsist and these can be used together with resulting differences in mean outcomes to identify the impact of treatment.

We formalize the following three assumptions below.

The first assumption states that the treatment effect is homogeneous across individuals, namely:

$\alpha = \alpha_i$ (IV1)

The second and the third assumption define the dependence of the outcome y_i on the participation status d_i and on the instrument z_j:

$P(d_i=1|z_j) \neq P(d_i=1)$ (IV2)

and

$E(u_i|z_j) = E(u_i)$ (IV3)

Under Conditions IV1 to IV3 the instrument z_j is the source of exogenous variation used to approximate randomization. It provides variation correlated with the participation decision only.

As a consequence of conditions IV1 and IV3, we have:

$E(y_i|z_j) = E[y_i|P(d_i=1|z_j)]$

IV estimation of an endogenous treatment effect

Following, for example, Heckman and Robb (1985) and Verbeek (2006), the IV estimator of a two-regime model can be given by:

$$\hat{y}_i = d_i\left(x_i'\hat{\boldsymbol{\beta}}_1\right) + _i\left(1-d_i\right)\left(x_i'\hat{\boldsymbol{\beta}}_0\right) + \hat{\alpha}\hat{d}_i$$

Or, in the (ideal) case in which no significant differences occur between the coefficients sets $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$:

$$\hat{y}_i = x_i'\hat{\boldsymbol{\beta}} + \hat{\alpha}\hat{d}_i .$$

The variable $\hat{d}_i$ is the prediction of the Probit estimation of the selection equation, $\hat{d}_i = \mathbf{z}_i'\hat{\boldsymbol{\gamma}}$, and identifies the probability of a subject to undergo the treatment, given the instruments z_j included in $\mathbf{z}_i$.

An alternative estimator (Hausman,1978) suggests to replace $\hat{d}_i$ with $\hat{\alpha} d_i + \hat{\phi}\hat{v}_i$ in the previous regression. $\hat{v}_i$ are the residuals of the *Probit* estimation of the selection equation

$$\hat{y}_i = x_i'\hat{\boldsymbol{\beta}} + \hat{\alpha} d_i + \hat{\phi}\hat{v}_i$$

The residuals, $\hat{v}_i$, reflect the unobserved heterogeneity affecting treatment not captured by the instruments and exogenous variables in the model. If the coefficient $\hat{\phi}$ of $\hat{v}_i$ is statistically different from zero, the impact of unobserved characteristics jointly affecting the treatment d_i and outcomes y_i is significant, as a consequnce the null that d_i is exogenous is rejected.

The Control Function Method

[see, also, Vella and Verbeek (1999)]

Consider a more general Two-Regime model as specified in Eqs. (1), (2) and (3):

Eq. 1 (outcome of treated subjects)

$$y_{1i} = \mathbf{x}_i'\boldsymbol{\beta}_1 + u_{1i} \qquad u_1 \sim N(0, \sigma_1^2) \tag{1}$$

Eq. 2 (outcome of untreated subjects)

$$y_{0i} = \mathbf{x}_i'\boldsymbol{\beta}_0 + u_{0i} \qquad u_0 \sim N(0, \sigma_0^2) \tag{2}$$

Selection Equation:

$$d_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + v_i \tag{3}$$

with d_i= 1 if $d_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + v_i > 0$, and I_i= 0 otherwise.

The control function estimator (CF) considers the endogeneity of the treatment indicator, d_i, as a censored variable (or selectivity)problem.In particular, CF approach takes simultaneously into account the potential endogenous selectivity of treatment in both treatment and control equations, given by the following expected outcomes:

$$E\left(y_i^* \middle| d_i = 1\right) = \mathbf{x_i}'\boldsymbol{\beta}_1 + \sigma_{1v} E\left(v_i \middle| v_i > -\mathbf{z_i}'\boldsymbol{\gamma}\right) = \mathbf{x_i}'\boldsymbol{\beta}_1 + \sigma_{1v} \frac{\phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)}{1-\Phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)} = \mathbf{x_i}'\boldsymbol{\beta}_1 + \sigma_{1v} \frac{\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}$$

for treated group, and:

$$E\left(y_i^* \middle| d_i = 0\right) = \mathbf{x_i}'\boldsymbol{\beta}_0 + \sigma_{0v} E\left(v_i \middle| v_i < -\mathbf{z_i}'\boldsymbol{\gamma}\right) = \mathbf{x_i}'\boldsymbol{\beta}_0 + \sigma_{0v} \frac{-\phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)}{\Phi\left(-\mathbf{z_i}'\boldsymbol{\gamma}\right)} = \mathbf{x_i}'\boldsymbol{\beta}_0 + \sigma_{0v} \frac{-\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{1-\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}$$

For untreated group.

The effect of treatment can be estimated by the difference of the intercept coefficients included, respectively, in $\boldsymbol{\beta}_1$ and in $\boldsymbol{\beta}_0$.

Analogously, the treatment effect can be obtained estimating the parameter α in the following model specified on the full sample:

$$E\left(y_i^* \middle| d_i; \mathbf{x_i}\right) = \mathbf{x_i}'\boldsymbol{\beta} + \alpha d_i + d_i \sigma_{1v} \frac{\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)} + \left(1-d_i\right)\sigma_{0v} \frac{-\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{1-\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}$$

d_i=1 indicates that the subject is undergone to treatment; d_i = 0 indicates, at the opposite, that the subject is not undergone to treatment.

The parameter σ_{1v} e σ_{0v} are the covariances of the error of the selection equation and the error terms of the outcome equations of, respectively, treated and untreated. If the condition $\sigma_{1v} = \sigma_{0v}$ occurs, this implies a "perfect randomization" condition, then the estimated coefficient $\hat{\alpha}$ corresponds to the average treatment effect (ATE). Estimating the control-function model, endogeneity of treatment can be verified if $\hat{\sigma}_{1v}$ and $\hat{\sigma}_{0v}$ significantly differ from zero.

Weaknesses of IV and CF estimators

A key issue in the implementation of IV is the choice of the instrument. It is frequently very difficult to find an observable instrumental variable, z_j, that satisfies the above reported Assumption IV3: $E(u_i|z_j) = E(u_i)$.

This will happen when the observables that determine participation are also the determinants of potential outcomes. In other cases, the instrument may have insufficient variation or may cause insufficient variation in the probability to undergo the treatment (*Propensity Score*).

Identification using classical IV still relies on the additional homogeneity assumption IV1 ($\alpha = \alpha_i$). If IV1 does not hold, the exclusion restriction is also unlikely to hold. To see why, notice that the unobservable in the outcome equation is now:

$$u_i^* = u_i + d_i\left(\alpha_i - \alpha\right)$$

and the new exclusion restriction needs to be expressed in terms of u^*_i:

$$E\left(u_i^* \mid z_j\right) = E\left(u_i \mid z_j\right) + P\left(z_j\right) E\left[\left(\alpha_i - \alpha\right) \mid d_i = 1; z_j\right] \neq E\left(u_i\right)$$

Only if no selection on the idiosyncratic gains occurs, it implies that the idiosyncratic gain, (α-α_i), and the unobservable in the selection equation, v_i, are not related. In such case, we have:

$$E\left[\left(\alpha_i - \alpha\right) \mid d_i = 1; z_j\right] = 0$$

and

$$E\left(u_i^* \mid z_j\right) = E\left(u_i\right) = 0$$

such as in assumption IV3.

An example

To illustrate the problem, consider the case on return to education.

Assume that the returns to education are partly determined by the student's unobservable ability. Suppose the instrument is some measure of the cost of education (say, distance to college and taxes) under the assumption that it is uncorrelated with the student's potential earnings and ability.

However, the selection process will create a relationship between distance to college and returns to college education in the data. This is because individuals facing a relatively low cost of education (live closer to college) may be more likely to invest in college education, even if expecting comparatively small returns.

Under our simplistic setup, this means that the distribution of ability among college graduates who live far from college is more concentrated on high ability levels than that for college graduates who live close to college. Such compositional differences will then affect the distribution of returns to college in the data for the two groups.

As a consequence the Homogeneity Assumption IV1 ($\alpha = \alpha_i$) fails to hold, and IV and CF will not generally identify ATE or ATT.

This happens because the average outcomes of any two groups, differing only on the particular z-realizations, are different for two reasons:

(i) different participation rates

(ii) compositional differences in the treated/nontreated groups with respect to the unobservables.

The later precludes identification of ATE or ATT.

However, can a different "local" average parameter be identified under slightly modified hypothesis? The "*Local Average Treatment Effect* (LATE)" parameter, to which we now turn

The Local Average Treatment Effect (LATE)

The solution advanced by Imbens and Angrist (1994) is to identify the impact of treatment from local changes in the instrument z_j when the treatment effect is heterogeneous.

The rationale is that, under certain conditions, a change in z_j reproduces random assignment locally by inducing individuals to alter their participation status without affecting the potential outcomes, (y_0 and y_1).

As with standard IV, the difference in average outcomes between two groups, differing only in the realization of z_j, results exclusively from the consequent difference in participation rates.

Unlike standard IV, the identifiable effect will not correspond to the ATE or the ATT. Instead, it will depend on the particular values of z_j used to make the comparison.

The identifiable effect is the average impact on individuals that change their participation status when faced with the change in z_j used to estimate the effect of treatment.

As with classical IV, the validity of an instrument z_j depends on whether it determines participation and can be excluded from the outcome equation conditional on participation.

In a heterogeneous effect framework, the exclusion condition requires that: (i) z_j has no joint variation with v_i and (ii) z_j is unrelated to the unobserved determinants of potential outcomes

The former condition is required or otherwisechanges in z_j would not separate changes in participation rates unrelated to outcomes as simultaneous changes in v_i could be related with changes in the unobservable components of the potential outcomes, particularly gains from treatment.

The LATE assumptions can now be formally established. The first two assumptions are identical to the classical IV Assumptions IV2 and IV3:

$P(d_i=1|z_j) \neq P(d_i=1)$ (LATE1 or IV2)

and

$E(u_i|z_j) = E(u_i)$ (LATE2 or IV3)

However LATE requires stronger identification assumptions than standard IV to allow for the relaxation of the homogeneity hypothesis.

The additional assumption pertains to the relationship between instrumentsz_j and the remaining unobservables included in v_i:

$$(\alpha_i ; \nu_i) \perp z_j \qquad \text{(LATE3)}$$

The last of the LATE assumptions is

$d_i(z_j)$ is a monotonic function of z_j. (LATE4)

The first assumption (LATE1) clarifies the meaning of the LATE parameter: it measures the impact of treatment on individuals that move from nontreated to treated when z_j changes.

The LATE approach can also be illustrated by the example on education return.

As before, suppose z_j is a measure of cost, say distance to college, with participation assumed to become less likely as z_j increases.

To estimate the effect of college education, consider a group of individuals that differ only in z_j. Among those that invest in further education when distance z_j equals z_j^* some would not do so if $z_j = z_j^{**}$ where $z_j^* < z_j^{**}$.

In this case, LATE measures the impact of college education on the "movers" by assigning any difference on the average outcomes of the two groups to the different enrollment rates caused by the difference in the cost of investing.

The monotonicity assumption is required for interpretation purposes.

Under monotonicity of d_i with respect to z_j, the LATE parameter measures the impact of treatment on individuals that move from nontreated to treated as z_j changes.

If monotonicity does not hold, LATE measures the change in average outcome caused by a change in the instrument, which is due to individuals moving in and out of participation, but cannot separate the effect of treatment on individuals that move in from that on individuals that move out as a consequence of a change in z_j (see Heckman, 1997).

Weaknesses of CF

The relative robustness of the classical parametric CF method comes from the structure it imposes on the selection process. However, this same feature has been strongly criticized for being overly restrictive.

There are two key assumptions underlying the selection model in CF approach:

(i) the parametric assumption on the joint distribution of unobservables;

(ii) the linear index assumption on the selection rule.

With regard to the point (i), important recent developments have proposed new semiparametric estimators that relax the assumptions of Normality of the errors distributions (see, for example, Powell, 1994).

More recently, Vytlacil (2002) has shown that the LATE approach can be seen as an application of a selection model. To see this, we first compare the two methods and then briefly discuss the equivalence result of Vytlacil.

The three CF assumptions can be equivalently written as:

-d_i is a nontrivial function of z_j;

-$\mathbf{z}_i$ is independent of (u_i, α, v_i)

-Index restriction: $d_i = 1[(\mathbf{z}_i'\boldsymbol{\gamma} + v_i) > 0]$

In turn, the LATE approach is based on the following regression model:

$$E\left(y_i \mid z_j\right) = \mathbf{x}_i'\boldsymbol{\beta} + P\left(d_i = 1 \mid z_j\right) E\left(\alpha_i \mid d_i = 1 \mid z_j\right)$$

The restricted CF estimator is based on an OLS regression, such as:

$$E\left(y_i^* \mid d_i; \mathbf{x_i}\right) = \mathbf{x_i}'\boldsymbol{\beta} + \alpha\left(d_i \mid z_j\right) + \sigma_{1v}\left(d_i \mid z_j\right)\frac{\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{\Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)} + \sigma_{0v}\left(1 - d_i \mid z_j\right)\frac{-\phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}{1 - \Phi\left(\mathbf{z_i}'\boldsymbol{\gamma}\right)}$$

based on the LATE assumptions discussed above.

We repeat them here:

d_i is a nontrivial function of z_j; (LATE1)

z_j is independent of (u_i, α_i, v_i) and (LATE2)

Monotonicity assumption: $d(z_j^*) > d(z_j^{**})$ or $d(z_j^*) < d(z_j^{**})$ for all individuals. (LATE3)

MATCHING METHODS

[SEE THE TEXTBOOK "How Do We Know If a Program Made a Difference?", PAG. 125]

The underlying motivation for the matching method is to reproduce the comparison group among the nontreated, this way re-establishing the experimental conditions in a nonexperimental setting.

The matching method constructs the correct sample counterpart of each treated subject by pairing each participant with members of the nontreated group.

The matching assumptions ensure that the only remaining relevant difference between the pairs of linked individuals is due to program participation.

Matching can be used with cross-sectional or longitudinal data. In its standard formulation, however, the longitudinal dimension is not considered.

We therefore abstract from time effect in this discussion.

We start by considering the potential outcome equations specified as follows:

Eq. 1 (outcome of treated subjects)

$$y_{1i} = \mathbf{x}_i'\boldsymbol{\beta} + \alpha_0 + \mathbf{x}_i'\boldsymbol{\alpha}_1 + u_{1i} + u_{0i} \qquad (1)$$

Eq. 2 (outcome of untreated subjects)

$$y_{0i} = \mathbf{x}_i'\boldsymbol{\beta} + u_{0i} \qquad (2)$$

Where the vector **x** includes the observed characteristics of the individual and $u_{1i} + u_{0i}$ are unobserved characteristics of the individual (in other words, $u_{1i} + u_{0i}$ take on different values for different individuals but only the variation in **x** is actually observed across individuals).

Program impact is then:

$$y_{1i} - y_{0i} = \alpha_0 + \mathbf{x}'_i\,\boldsymbol{\alpha}_1 + u_{1i}$$

This is thus a framework where program impact varies across individuals. It does so because individuals have different observed (captured by the term $\mathbf{x}_i'\boldsymbol{\alpha}_1$) and unobserved (captured by u_{1i}) determinants of program impact.

Variation in $\mathbf{x}_i'$ can be controlled adopting proper randomization techniques to homogenize sample composition for both treated and untreated groups and to identify the choice of be treated or untreated.

The consequent identifying assumption is known as "unconfoundedness" or conditional independence.

Variation in unobservable characteristics u_{1i}should be conformed to the following assumption:

- That u_{1i} plays no role in the participation decision (which in this context implies toassume that u_{1i}= 0);

This assumption implies the *absence of selection of unobservables* in the comparison between treated and untreated.

Matching estimates program impact for each individual by finding a similar individual who experienced the counterfactual outcome. For a participant the counterfactual outcome is y_{0i}, while for a non-participant it is y_{1i}.

For each individual, his/her counterfactual outcome is estimated by the outcome experienced by a similar person for whom that counterfactual is observed.

In this manner, an estimate of y_{1i} - y_{0i} can be formed for each observed individual. It is an estimate because the value of either y_{1i} or y_{0i} will have been estimated for each individual according to their participation status.

With the estimates of program impact y_{1i} - y_{0i} so obtained for each observed individual, estimating of average impact for whatever population the observed individuals simply involve suitably the computation of average across them.

Selection on observables (remedies)

In order to correct estimates for the term $\mathbf{x}_i'\boldsymbol{\alpha}_1$, we consider different possible values of $\mathbf{x}_i$ have associated probabilities of occurring in the population depending on $\boldsymbol{\alpha}_1$, $\Pr(\mathbf{x}_i = \mathbf{x}_i'\boldsymbol{\alpha}_1)$.

This probability can be easily estimated computing the ratio between the individuals included in the sample for whom $\mathbf{x}_i = \mathbf{x}_i'\boldsymbol{\alpha}_1$ and total of individuals included in the sample:

$$\Pr\left(\mathbf{x}_i = \mathbf{x}'_i\,\boldsymbol{\alpha}_1\right) = \frac{N_{\mathbf{x}'_i\boldsymbol{\alpha}_1}}{N}$$

The average treatment effect for the population is given by:

$$ATE = E\left(y_{1i} - y_{0i} \mid \mathbf{x}_i = \mathbf{x}'_i\,\boldsymbol{\alpha}_1\right)\Pr\left(\mathbf{x}_i = \mathbf{x}'_i\,\boldsymbol{\alpha}_1\right)$$

In other words, the average treatment effect is the sum of the expectations of the treatment effectforthe various types (as captured by $\mathbf{x}_i$) of individuals in the population, with each of

thoseexpectationsweighted by the frequency with which the individual with the type ($\mathbf{x}_i = \mathbf{x}_i'\boldsymbol{\alpha}_1$) occurs in thepopulation.

Parameters of Interest

The parameter that received the most attention in evaluation literature is the ‘average treatment effect on the treated’ (ATT), which isdefined as:

ATT = $E[y_{1i}|d_i = 1] - E[y_{0i}|d_i = 1]$

As the counterfactual mean for those being treated - $E[y_{0i}|d = 1]$- is not observed,one has to choose a proper substitute for it in order to estimate ATT.

Using themean outcome of untreated individuals $E[y_{0i}|d = 1]$ is in non-experimental studiesusually not a good idea, because it is most likely that components which determinethe treatment decision also determine the outcome variable of interest. Thus, theoutcomes of individuals from treatment and comparison group would differ even inthe absence of treatment leading, in this context, to a ‘self-selection bias’.

This concept can be formalized as follows.

For ATT it can be notedas:

$E[y_{1i}|d = 1] - E[y_{0i}|d = 0] = E[y_{1i}|d_i = 1] - E[y_{0i}|d_i = 1] + E[y_{0i}|d = 1] - E[y_{0i}|d = 0]$

$E[y_{1i}|d = 1] - E[y_{0i}|d = 0] = ATT + E[y_{0i}|d = 1] - E[y_{0i}|d = 0]$

The difference between the left hand side of previous equation and ATT is the so-called‘self-selection bias’. The true parameter ATT is only identified, if:

$E[y_{0i}|d = 1] - E[y_{0i}|d = 0] = 0.$

In social experiments where assignment to treatment is random this condition is ensured and the treatment effect, ATT, is identified.

In non-experimental studies one has to invoke some identifying assumptions to solve the selection problem stated above.

Another parameter of interest is the ‘average treatment effect’ (ATE), which is defined as:

ATE= $E[y_{1i}] - E[y_{0i}]$

The additional challenge when estimating ATE is that both counterfactual outcomes:

$E[y_{1i}|d = 0]$ and $E[y_{0i}|d = 1]$

have to be constructed.

How to remedies to bias occurring for self-selection?

Analysts suggest using these resources:

- **i)** Conditional Independence Assumption;
- **ii)** Common Support
- **iii)** Data Balancing
- **iv)** Estimation Strategy

Conditional Independence Assumption (CIA):
One possible identification strategyis to assume, that given a set of observable covariates $\mathbf{x}_i$ which are not affected bytreatment, potential outcomes are independent of treatment assignment.
This implies, that selection is solely based on observable characteristics and thatall variables that influence treatment assignment and potential outcomes simultaneously are observed by the researcher. Clearly, this is a strong assumption and has tobe justified by the data quality at hand.

Common Support:
A further requirement besides independence is the *common support* or *overlap condition*.
It rules out the phenomenon of perfect predictabilityof d_i given $\mathbf{x}_i$:

(Overlap) $0 < P(d_i = 1|X) < 1$

It ensures that persons with the same $\mathbf{x}_i$ values have a positive probability of being both participants and non-participants (Heckman, LaLonde, and Smith, 1999)

Data Balancing

Since conditioning on all relevant covariates is limited in the case of a highdimensional vector $\mathbf{x}_i$ (‘curse of dimensionality’), Rosenbaum and Rubin (1983)suggest the use of so-called balancing scores $b(\mathbf{x}_i)$, i.e. functions of the relevantobserved covariates $\mathbf{x}_i$ such that the conditionaldistribution of $\mathbf{x}_i$ given $b(\mathbf{x}_i)$ isindependent of assignment into treatment.

Estimation Strategy

Two approaches are currently adopted to manage the problem of multidimensionality of data in balancing: i) Covariate Matching and ii) Propensity Score Matching.

Covariate Matching (CVM)

CVM distance measures like the Mahalanobis distance are used to calculate similarityof two individuals in terms of covariate values and the matching is done on these distances. Theinterested reader is referred to Abadie and Imbens (2004a and 2004b) who develop covariateand bias-adjusted matching estimators.

Propensity Score Matching (PSM)

One possible balancing score is thepropensity score, i.e. the probability of participating in a programme given the observedcharacteristics of the subjects reported by the covariates $\mathbf{x}_i$. Matching procedures based on this balancing score are knownas propensity score matching (PSM).

Given that CIA holds and assuming additional that thereis overlap between both groups (called 'strong ignorability' by Rosenbaum and Rubin(1983)), the PSM estimator for ATT can be written in general as:

$$\text{ATT} = E\{[y_{1i}|\text{d}_\text{i} = 1;\text{P}(\mathbf{x}_i)] - [y_{0i}|\text{d}_\text{i} = 0;\text{P}(\mathbf{x}_i)]\} = E[y_{1i}|\text{d}_\text{i} = 1;\text{P}(\mathbf{x}_i)] - \text{E}[y_{0i}|\text{d}_\text{i} = 0;\text{P}(\mathbf{x}_i)]$$

To put it in words, the PSM estimator is simply the mean difference in outcomesover the common support, appropriately weighted by the propensity score distribution of participants.

Based on this brief outline of the matching estimator in thegeneral evaluation framework, we are now going to discuss the implementation of PSM in detail.

Once the researcher has decided to use PSM, he is confronted with a lot of questions regarding its implementation. summarises the necessary steps when implementing PSM:

Step 1: Propensity Score Estimation (Variable choice)

Step 2: Choose Matching Algorithm

Step 3: Check Overlap/Common Support

Step 4: Matching Quality/Effect Estimation

Step 5: Sensitivity Analysis

Variable Choice

More advice is available regarding the inclusion (or exclusion)of covariates in the propensity score model. The matching strategy builds on theCIA assumption, requiring that the outcome variable(s) must be independent of treatment conditional on the propensity score. Hence, implementing matching requires choosinga set of variables $\mathbf{x}_i$ that credibly satisfy this condition.

In particular, the omission of important variables can seriously increase bias inresulting estimates. The better andmore informative the data are, the easier it is to credibly justify the CIA and thematching procedure.

However, it should also be clear that 'too good' data is nothelpful either. If $P(\mathbf{x}_i) = 0$ or $P(\mathbf{x}_i)= 1$ for some values of $\mathbf{x}_i$, then we cannot use matching conditional on those $\mathbf{x}_i$ values to estimate a treatment effect, becausepersons with such characteristics either always or never receive treatment. Hence,the common support condition fails and matches cannot beperformed.

A commonly adopted approach in variables selection relies on statistical significanceand is very common in textbook econometrics. To do so, one starts with a parsimonious specification of the model, e.g. a constant, the age and some regionalinformation, and then 'tests up' by iteratively adding variables to the specification. A new variable is kept if it is statistically significant at conventional levels.

Choosing a Matching Algorithm

All matchingestimators contrast the outcome of a treated individual with outcomes of comparison group members. PSM estimators differ not only in the way the neighbourhood foreach treated individual is defined and the common support problem is handled, butalso with respect to the weights assigned to these neighbours.

We present the general ideas and the involved trade-offs between some different algorithms:

Nearest Neighbour Matching: The most straightforward matching estimatoris nearest neighbor (NN) matching. The individual from the comparison group ischosen as a matching partner for a treated individual that is closest in terms ofpropensity score. Several variants of NN matching are proposed, e.g. NN matching‘with replacement’ and ‘without replacement’. In the former case, an untreatedindividual can be used more than once as a match, whereas in the latter case itisconsidered only once. Matching with replacement involves a trade-off betweenbias and variance. If we allow replacement, the average quality of matching willincrease and the bias will decrease. This is of particular interest with data where thepropensity score distribution is very different in the treatment and the control group.

Caliper and Radius Matching: NN matching faces the risk of bad matches,if the closest neighbour is far away. This can be avoided by imposing a tolerancelevel on the maximum propensity score distance (caliper). Imposing a caliper worksin the same direction as allowing for replacement. Bad matches are avoided andhence the matching quality rises.
Applying caliper matching means that thoseindividual from the comparison group is chosen as a matching partner for a treatedindividual that lies within the caliper (‘propensity range’) and is closest in terms ofpropensity score.

Kernel and Local Linear Matching: The matching algorithms discussed so far have in common that only a few observations from the comparison group are used to construct the counterfactual outcome of a treated individual.
Kernel matching (KM) and local linear matching (LLM) are non-parametric matching estimators that use weighted averages of all individuals in the control group to construct the counterfactual outcome.
Thus, one major advantage of these approaches is the lower variance which is achieved because more information is used. A drawback of these methods is that possibly observations are used that are bad matches.

Overlap and Common Support

An important step is to check the overlap and the region of common support between treatment and comparison group.

Implementing the common support conditionensures that any combination of characteristics observed in the treatment group canalso be observed among the control group.

ForATT it is sufficient to ensure the existence of potential matches in the control group,whereas for ATE it is additionally required that the combinations of characteristicsin the comparison group may also be observed in the treatment group.

Several ways are suggested in the literature, where the most straightforward one is a visual analysis of the density distribution of the propensity score in both groups.

We will present two methods,where the first one is essentially based on comparing the minima and maxima ofthe propensity score in both groups and the second one is based on estimating the density distribution in both groups.

Minima and Maxima comparison: The basic criterion of this approach is to delete all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group.

To give an example let us assume for a moment that the propensity score lies within the interval [0.07, 0.94] in the treatment group and within [0.04, 0.89] in the control group. Hence, with the 'minima and maxima criterion', the common support is given by [0.07, 0.89].

Trimming to Determine the Common Support: A different way to overcome these possible problems is suggested by Smith and Todd (2005). They use a trimming procedure to determine the common support region and define the region of common support as those values of the estimated propensity whose density falls in the positive side of the density distribution, within both the D = 1 and D = 0 distributions.

Assessing the Matching Quality

Since we do not condition on all covariates but on the propensity score, it has to bechecked if the matching procedure is able to balance the distribution of the relevantvariables in both the control and treatment group.

The basic idea is to compare the situation beforeand after matching and check if there remain any differences after conditioning on the propensity score. If there are differences, matching on the score was not (completely) successful and remedial measures have to be done, e.g. by including interaction-terms in the estimation of the propensity score.

One suitable indicator to assess the distance in marginal distributions of the $\mathbf{x}_i$-variables is the standardised bias (SB) suggested by Rosenbaum and Rubin (1985). For each covariate $\mathbf{x}_i$ it is defined as the difference of sample means in the treated and matched control subsamples as a percentage of the square root of the average of sample variances in both groups.

The standardised bias (for each covariate x_i included in$\mathbf{x}_i$)before matching is given by:

$$SB_{before} = 100 \cdot \frac{\left(\overline{x}_{1i} - \overline{x}_{0i}\right)}{\sqrt{0.5 \cdot \left(\sigma_{x1}^2 + \sigma_{x0}^2\right)}}$$

The standardised bias after matching is given by:

$$SB_{after} = 100 \cdot \frac{\left(\bar{x}_{1M} - \bar{x}_{0M}\right)}{\sqrt{0.5 \cdot \left(\sigma^2_{x1M} + \sigma^2_{x0M}\right)}}$$

Where $\bar{x}_d$ and σ^2_{xd} are mean and variance in the treatment and control groups before matching; while $\bar{x}_{dM}$ and σ^2_{xdM} are mean and variance in treatment and control groups in the matched sample.

One possible problem with the standardised bias approach isthat we do not have a clear indication for the success of the matching procedure, even though in most empirical studies a bias reduction below 3% or 5% is seen as sufficient.

***t*-Test**: A similar approach uses a two-sample *t*-test to check if there are significant differences in covariate means for both groups (Rosenbaum and Rubin, 1985). Before matching differences are expected, but after matching the covariates should be balanced in both groups and hence no significant differences should be found.

The *t*-test might be preferred if the evaluator is concerned with the statistical significance of the results. The shortcoming here is that the bias reduction before and after matching is not clearly visible.

Joint significance and Pseudo-R2 : Additionally, Sianesi (2004) suggests to re-estimate the propensity score on the matched sample, that is only on participants and matched non-participants and compare the pseudo-R2's before and after matching. The pseudo-R2 indicates how well the regressors $\mathbf{x}_i$ explain the participation

probability. After matching there should be no systematic differences in the distribution of covariates between both groups and therefore, the pseudo-R2 should be fairly low. Furthermore, one can also perform an *F*-test on the joint significance of all regressors. The test should not be rejected before, and should be rejected after matching.

Stratification Test: Finally, Dehejia and Wahba (1999, 2002) divide observations into strata based on the estimated propensity score, such that no statistically significant difference between the mean of the estimated propensity score in both treatment and control group remain. Then they use *t*-tests and *F*-test within each stratus to test if the distribution of $\mathbf{x}_i$-variables is the same between both groups (for the first and second moments). If there are remaining differences, they add higher-order and interaction terms in the propensity score specification, until such differences no longer emerge.

Sensitivity Analysis

Unobserved Heterogeneity - Rosenbaum Bounds.

The estimation of treatment effects with matching estimators is based on the CIA, that is selection on observable characteristics. However, if there are unobserved variables which affect assignment into treatment and the outcome variable simultaneously, a ‘hidden bias’ might arise.

It should be clearthat matching estimators are not robust against this ‘hidden bias’. Since it is not possible to estimate the magnitude of selection bias with non-experimental data, we address this problem with the bounding approach proposed by Rosenbaum (2002).

The basic question to be answered is, if inference about treatment effects may be altered by unobserved factors. In other words, we want to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of matching analysis.

Let us assume that the participation probability is given by $P(\mathbf{x}_i) = P(d_i = 1; \mathbf{x}_i) = \boldsymbol{F}(\mathbf{x}_i\boldsymbol{\beta}+\gamma u_i)$,

where $\mathbf{x}_i$ are the observed characteristics for *i*-th individual, u_i is the unobserved variable and γ is the effect of u_i on the participation decision.

Clearly, if the matching is free of hidden bias, γ will be zero and the participation probability will solely be determined by $\mathbf{x}_i$. However, if there is hidden bias, two individuals with the same observed covariates $\mathbf{x}_i$ 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*cdf* of the logistics distribution. The odds that individuals receive treatment are then given by $P(\mathbf{x}_i)/(1- P(\mathbf{x}_i))$ and $P(\mathbf{x}_j)/(1- P(\mathbf{x}_j))$, and the odds ratio is given by:

$$\frac{P(\mathbf{x}_i)/(1-P(\mathbf{x}_i))}{P(\mathbf{x}_j)/(1-P(\mathbf{x}_j))} = \frac{P(\mathbf{x}_i)(1-P(\mathbf{x}_j))}{P(\mathbf{x}_j)(1-P(\mathbf{x}_i))} = \frac{\exp(\mathbf{x}'_j\boldsymbol{\beta}+\gamma u_j)}{\exp(\mathbf{x}'_i\boldsymbol{\beta}+\gamma u_i)} = \exp\left[\gamma(u_i-u_j)\right]$$

Thus, it follows that if there are no differences in the unobserved variables ($u_i = u_j$ or if the unobserved variables have no influence on the probability of participation ($\gamma = 0$), the odds ratio is one, which implies the absence of hidden or unobserved selection biases.

Rosenbaum (2002) derived the following bounds on the odds-ratio that either of the two matched individuals will receive treatment:

$$\frac{1}{\exp[\gamma]} \leq \frac{P(\mathbf{x}_i)/(1-P(\mathbf{x}_i))}{P(\mathbf{x}_j)/(1-P(\mathbf{x}_j))} \leq \exp[\gamma]$$

Both matched individuals have the same probability of participating only if exp(γ) = 1.

If exp(γ) = 2, then individuals who appear to be similar (in terms of $\mathbf{x}_i$) could differ intheir odds of receiving the treatment by as much as a factor of 2. In this sense, exp(γ) is a measure of the degree of departure from a study that is free of hidden bias.

$P(\mathbf{x}_i)/(1- P(\mathbf{x}_i))$ and $P(\mathbf{x}_j)/ (1- P(\mathbf{x}_j))$ could be obtained by introducing covariates simulating hidden bias in the propensity score estimation. In this case one can adopt Rosenbaum's bounds to check the robustness of the matching results to departures from the CIA assumption.

We can calculate the results of the *p*-value from Wilcoxon sign-rank tests for the averaged treatment effect on the treated while setting the level of hidden bias to a certain value γ, which reflects our assumption about

unmeasured heterogeneity or endogeneity in treatment assignment expressed in terms of the odds ratio of differential treatment assignment due to an unobserved covariate.

At eachγ we calculate a hypothetical significance level "p-critical", which represents the bound onthe significance level of the treatment effect in the case of endogenous self-selection into treatment status.

By comparing the Rosenbaum bounds on treatment effects at different levelsof γ we can assess the strength such unmeasured influences would require in order that the estimated treatment effects from propensity score matching would have arisen purely through selection effects.

EXAMPLE (gretldata)

Rosenbaum Sensitivity Test for Wilcoxon Signed Rank P-Value

Unconfounded estimate 0

γ	Lower bound	Upper bound
1.0	0	0.0000
1.1	0	0.0000
1.2	0	0.0000
1.3	0	0.0000
1.4	0	0.0000
1.5	0	0.0004
1.6	0	0.0035
1.7	0	0.0176
1.8	**0**	**0.0602**
1.9	0	0.1512
2.0	0	0.2957

Note: Gamma is Odds of Differential Assignment To

Treatment Due to Unobserved Factors

After matching, we may still have covariates that we did not collect data on, observe and match, but nonetheless affect the treatment and outcome variable. Sensitivity analysis attempts to address how likely is this possibility.

Since only at the gamma of 1.8 the upper bound is 0.0602> 0.05, this is the critical value.

γ represents the ratio in odds of treatment for 2 subjects with the same observed covariates but a different unobserved covariate, and sensitivity analysis looks at how large γ can be before the conclusion of the study changes i.e. p-value is larger than 0.05.

So in this case, one subject needs to be 1.8 times more likely as another to receive the treatment due to an unobserved covariate before the study conclusion becomes non-significant. Since this is a large number ofγ (it is unlikely to find any unobserved covariate that affects odds of treatment this much) the study is quite robust to unobserved treatment.

DISCONTINUITY DESIGN (DD)

[VEDI IL MANUALE "How do we know if a program made a difference" PAG. 294]

However, a special case that has attracted recent attention occurs when the probability of enrolment into treatment changes discontinuously with some continuous variable z_i. The variable z_i is an observable instrument, typically used to determine eligibility. It is, therefore, included in the regressors set of the Selection Model. The discontinuity design estimator (DD) uses the discontinuous dependence of d_i on z_i to identify a local average treatment effect even when the instrument does not satisfy the IV assumptions discussed above.

DD relies on a continuous relationship between the instrument z_i and all the determinants of the outcome except participation in treatment. Any discontinuity in y_i as a function of z_i is, therefore, attributed to a discontinuous change in the participation rate as a function of z_i.

As a consequence, treated and nontreated are individuals with values of z_i, respectively, above and below the threshold (or cut-off point), z_c affecting the participation decision.

However, regression discontinuity design requires that all potentially relevant variables besides the treatment variable and outcome variable be continuous at the point where the treatment and outcome discontinuities

occur. One sufficient, though not necessary, condition is if the treatment assignment is "as good as random" at the threshold for treatment.

Sharp Design

Thus the probability of participation changes discontinuously at the treshold z_c from zero to one. The identification condition with sharp design can be stated as follows:

$$\lim_{z \to z_c^-} P\left(d_i = 1; z_i\right) = P\left(z_c^-\right) = 0$$
$$\lim_{z \to z_c^+} P\left(d_i = 1; z_i\right) = P\left(z_c^+\right) = 1 \qquad \text{(DD1)}$$

where, to simplify the notation, $P(z^-_c)$ and (P(z^+_c) represents the limit of the propensity score (P(d_i = l| z_i) = P(z_i)) as z_i approaches z_c, respectively, from below and from above. Both limits are assumed to exist.

The DD parameter is, in this case:

$$\alpha_{DD}\left(z_c\right) = E\left(y_i \mid z_c^+\right) - E\left(y_i \mid z_c^-\right)$$

z^+_c and z^-_c) are the limits of E[y_i;| z_i] when z_i approaches z_c from above and below, respectively. $\alpha_{DD}(z_c)$ measures the impact of treatment on a randomly selected individual with observable characteristics z_i just above z_c.

If we now impose this additional assumption:

$$E\left(y_i \mid z_c^+\right) - E\left(y_i \mid z_c^-\right) = E\left(y_i \mid z_c\right)$$

the DD parameter can be more naturally interpreted as being the impact of treatment on a randomly selected individual at the threshold point z_c:

$$\alpha_{DD}\left(z_c\right) = E\left(y_i \mid z_c\right)$$

Fuzzy Design

A fuzzy design occurs when dimensions other than z_i, (in particular, unobserved dimensions) also affect participation. In the general fuzzy design case, participation and nonparticipation occur on both sides of the threshold z_c. Thus, Assumption DD1 needs to be adjusted accordingly:

$$P\left(z_c^+\right) \neq P\left(z_c^-\right)$$ DD2

The additional problem here is that only a subpopulation moves treatment status at the discontinuity point and the selection of movers is likely to be related with potential outcomes.

Fuzzy DD relies on the following additional local (mean) independence assumption to identify a local treatment effect parameter:

$E\left(\alpha \mid d; z_i\right) = E\left(\alpha \mid z_i\right)$ in a small neighbourhood of z_i.

The DD parameter is identified as follows:

$$\alpha_{DD}\left(z_c\right) = \frac{E\left(y_i \mid z_c^+\right) - E\left(y_i \mid z_c^-\right)}{P\left(y_i \mid z_c^+\right) - P\left(y_i \mid z_c^-\right)}$$

Then a DD estimator is given by:

$$\hat{\alpha}_{DD}\left(z_c\right) = \frac{\bar{y}^+ - \bar{y}^-}{\hat{P}\left(z_c^+\right) - \hat{P}\left(z_c^-\right)}$$

where $\bar{y}^{+}$ and $\bar{y}^{-}$ are sample averages of the observed outcomes at each side of the threshold, and $\hat{P}\left(z_c^{+}\right)$ and $\hat{P}\left(z_c^{-}\right)$ are estimators of the participation probability at each side of the threshold.

A nonparametric version of DD is simple to implement. It only requires running nonparametric regressions of y_i and d_i on z_i locally, separately on each side of the discontinuity point. The predicted limits can then be used to estimate the impact of treatment using previous expression of estimator.

References

-Abadie, A., D. Drukker, J. Leber Herr, and G. W. Imbens (2004a): “Implementing Matching Estimators for Average Treatment Effects in STATA,” *The Stata Journal*, 4(3), 290–311.

Abadie, A., and G. Imbens (2004b): “Large Sample Properties of MatchingEstimators for Average Treatment Effects,” Working Paper,Harvard University.

-Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60(l):47-57.

Athey, Susan, and Guido Imbens. 2006. "Identification and Inference in Nonlinear Difference-In-Differences Models." *Econometrica* 74(2):431-97.

-Blundell, Richard, Monica Costa Dias, Costas Meghir, and John Van Reenen. 2004. "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program." *Journal of the European Economics Association* 2(4):596-606.

-Card, David, and Philip Robins, P. 1998. "Do Financial Incentives Encourage Welfare Recipients To Work?." *Research in Labor Economics* 17(1): 1-56.

-Hausman, J.A. (1978). “Specification Tests in Econometrics”. *Econometrica*, Vol. 46, No. 6, pp. 1251-1271

-Heckman, James,1979. "Sample Selection Bias as a Specification Error." *Econometrica*47(1): 153-611.

-Heckman, James.1997. “Instrumental Variables: A Study of the Implicit Assumptions underlying one Widely used Estimator for Program Evaluations." *Journal of Human Resources* 32(3):441-462.

-Heckman, James, HidenikoIchimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(4):605-54. _.

-Heckman, James, HidenikoIchimura, and Petra Todd, 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2):261-94.

-Heckman, J., R. LaLonde, and J. Smith (1999): “The Economics and Econometrics of Active Labor Market Programs,” in *Handbook of Labor Economics*Vol.III, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. Elsevier, Amsterdam.

-Heckman, James, and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In Longitudinal Analysis of Labour Market Data, ed. James Heckman and Burton Singer, 156-246. New York: Wiley.

-Heckman, James, and Jeffrey Smith. 1999. "The PreProgram Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies." *Economic Journal* 109(457):313-48.

-Heckman, James, and Edward Vytlacil. 1998. "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling." *Journal of Human Resources* 33(4):974-87.

-Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467-75.

-Johnson, N. L. and Kotz, S. (1970). Continuous univariate distributions-1, chapter 13. John Wiley & Sons.

-Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*. *Cambridge University Press*

-Powell, James. 1994. "Estimation of Semiparametric Models." In *Handbook of Econometrics*, eds. Robert Engle and Daniel McFadden, Volume 4: 2443-2521. Amsterdam: North Holland.

-Srivastava, K. V., & Giles, D. (1987). *Seemingly Unrelated Regression EquationsModels*. Statistics: textbooksand monographs 80. New York: Dekker.

-Rosenbaum, P. R. (2002): *Observational Studies*. Springer, New York.

-Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Scorein Observational Studies for Causal Effects," *Biometrika*, 70, 41–50.

-Smith, J., and P. Todd (2005): "Does Matching Overcome LaLonde's Critiqueof Nonexperimental Estimators?," *Journal of Econometrics*, 125(1-2), 305–353.

- Vella, F. and Verbeek M. (1999)"Estimating and Interpreting Models with Endogenous Treatment Effects".*Journal of Business & Economic Statistics*, Vol. 17, No. 4 (Oct., 1999), pp. 473-478

- Verbeek, M. (2017) *A guide to modern Econometrics*. (5th edition). Wiley.
-Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." Econometrica 70(1):331 -341.