Counterfactual approach 3

Prof. Roberto Gabriele

Methods for Empirical Economics

Department of Economics and Management, University of Trento,
Ph.D. programme in Economics and Finance

Outline

1. Characterizing bias
2. Regression adjustment
3. Matching

An example

Overlap assumption at work

1. 8 units observed and three variables: D, x, Y
2. We have just two units with x = 0 (unit 1 and 2), both in the untreated group (D = 0), and no units in the treated group having such a value of x. In a situation like this, $p(D = 1 \mid x = 0) = 0$ and **ATE cannot be identified**:

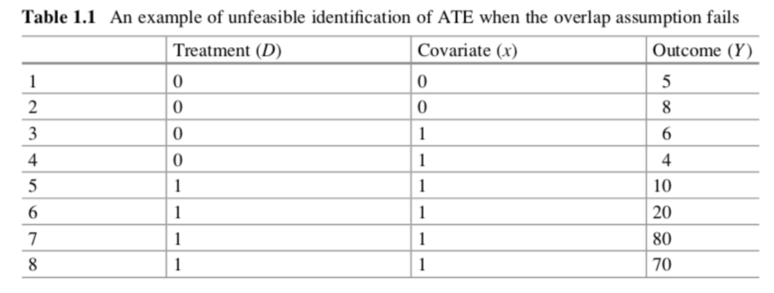
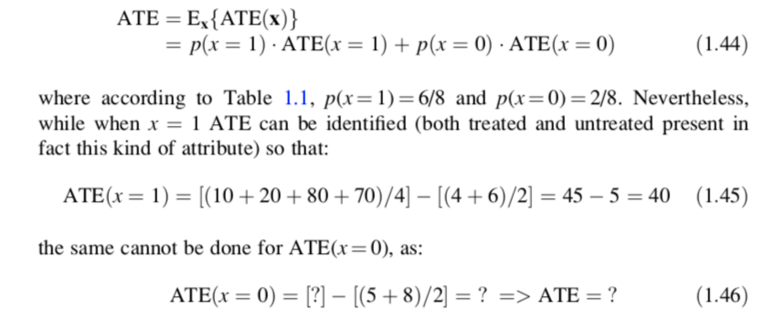


Table 1.1 An example of unfeasible identification of ATE when the overlap assumption fails

	Treatment (D)	Covariate (x)	Outcome (Y)
1	0	0	5
2	0	0	8
3	0	1	6
4	0	1	4
5	1	1	10
6	1	1	20
7	1	1	80
8	1	1	70

An example II

Indeed:



$$
\begin{aligned}
\text{ATE} &= \text{E}_{\mathbf{x}}\{\text{ATE}(\mathbf{x})\} \\
&= p(x=1)\cdot\text{ATE}(x=1) + p(x=0)\cdot\text{ATE}(x=0)
\end{aligned}
\tag{1.44}
$$

where according to Table 1.1, $p(x=1)=6/8$ and $p(x=0)=2/8$. Nevertheless, while when $x=1$ ATE can be identified (both treated and untreated present in fact this kind of attribute) so that:

$$
\text{ATE}(x=1) = [(10+20+80+70)/4] - [(4+6)/2] = 45 - 5 = 40 \tag{1.45}
$$

the same cannot be done for $\text{ATE}(x=0)$, as:

$$
\text{ATE}(x=0) = [?] - [(5+8)/2] = \,? \;=>\; \text{ATE} = \,? \tag{1.46}
$$

In order to identify all ATEs, each cell built by crossing the values taken by the various x-provided that they have finite discrete support- must have both treated and untreated units.

An example IIa

1. While for ATENT we have:
2. $ATENT = (6+4)/2 - (5+8)/2 = -1.5$
3. While ATET-for the same reason of ATE- is not identifiable!

An example III

Thus, as a general rule:

1. **The identification of ATET just requires that** $p(D = 1|x) < 1$
2. that of ATENT that $p(D = 1|x) > 0$ (or, equivalently, $p(D = 0|x) < 1$)
3. In other words, we can conclude that, in order to identify all ATEs: **each cell built by crossing the values taken by the various x -provided that they have finite discrete support- must have both treated and untreated units**

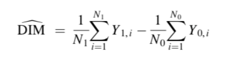
Selection bias characterization

Consider the linear model:

1. $Y = \mu + \alpha D + u$, under randomization we have that:
2. $E(Y|D = 1) = \mu + \alpha$
3. $E(Y|D = 0) = \mu$;
4. and: $\alpha = E(Y|D = 1) - E(Y|D = 0) = DIM$ (a)
5. Assume selection is driven by x, the outcome is also function of x:
6. $y = \mu_a + \alpha_a D + \beta x + u_a$

in these conditions: $Y^* = \mu_a + \alpha_a D + u_a$, where $Y^* = Y - \beta x$

(a)



$$\widehat{\text{DIM}} = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{1,i} - \frac{1}{N_0}\sum_{i=1}^{N_0} Y_{0,i}$$

Selection bias characterization II

1. We have that: $\alpha_a = E(Y^*|D=1) - E(Y^*|D=0)$
2. nonetheless: $\alpha_a = E(Y|D=1) - E(Y|D=0) - \beta[E(x|D=1) - E(x|D=0)]$
3. $\alpha_a = DIM - BIAS = \alpha - BIAS$
4. $BIAS = \alpha - \alpha_a = \beta[E(x|D=1) - E(x|D=0)]$
5. The **bias increases** either as soon as: (1) β is different from zero, and (2) the average value of x in the treated and untreated group is different.
6. The first **cause of bias variation** depends on the **degree of dependence of the outcome on factor x**
7. The second cause of bias variation depends on **how ''balanced'' are the two groups in terms of the factor x**

Example with BIAS

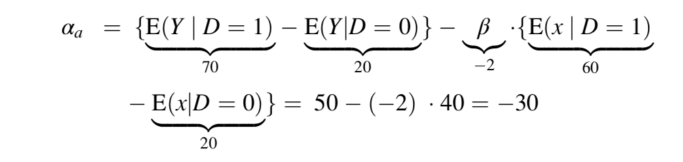
1. Suppose two groups of people, group 1 and group 0, are to be used as treated and control groups, respectively, in the evaluation of a given training program.
2. Suppose that, because of the underlying selection process, group 1 is made of young people (lets say, people with an average age of 20), whereas group 1 is made of older people (with an average age of 60)
3. We are interested in evaluating the effect of this training program on **individuals' comprehension capacity of a complex text**, measured by scores associated with a final exam

Example with BIAS II

1. We might find that group 1 is highly performing with, let's say, an average score of 70, and group 0 is poorly performing with an average of 20
2. The simple groups' DIM, equal to 50 in this case, would suggest that the training program was effective in fostering people's comprehension capabilities
3. Nevertheless, this result is misleading as the two groups are far from being balanced in terms of age

Example with BIAS III

1. Suppose that comprehension is significantly and negatively related to age, so that β is negative and equal to, let's say $\beta = -2$
2. In this case we have that:



$$\alpha_a = \{\underbrace{\mathrm{E}(Y \mid D=1)}_{70} - \underbrace{\mathrm{E}(Y|D=0)}_{20}\} - \underbrace{\beta}_{-2} \cdot \{\underbrace{\mathrm{E}(x \mid D=1)}_{60} - \underbrace{\mathrm{E}(x|D=0)}_{20}\} = 50 - (-2) \cdot 40 = -30$$

Example with BIAS

Observations

1. $DIM = ATET + B_1$
2. where: $B_1 = E(Y_0|D = 1) - E(Y_0|D = 0)$ is the selection bias
3.

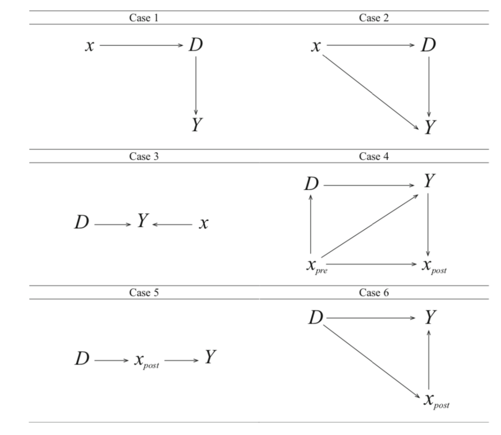
Choosing observables

We start with a question:

1. Which is however the rationale for choosing the variables to control for?
2. A guideline for establishing how one can wisely choose confounders
3. one should have as clear an understanding as possible of the causal relations linking the variables entering his model.
4. In other words, this suggests that one relies on a clear-cut "theoretical" representation of the relation among treatment, potential confounders, and outcomes
5. In this sense, **context's conditions, theoretical background, past evidence, and even personal beliefs** may play a fundamental role

Choosing observables II

The choice to include or exclude a given covariate x does depend on the **specific causal links assumed** among x (the potential confounder), D (the binary treatment), and Y (the outcome):



Choosing observables

Case 1 x->D->Y

1. x behaves as a pure pretreatment variable. Indeed, x determines D that in turn determines Y. No relation between x and Y is assumed
2. x does not need to be included as a confounder.
3. Case 1 can be represented by a system of two equations, the selection equation (assumed to be linear for simplicity) and the outcome equation, taking on this form:
4. $\begin{cases} D{=}a_1{+}a_x x{+}u \\ Y{=}b_1{+}b_w D{+}v \end{cases}$
5. $Under CMI : ATE = E_x\{E(Y|x, D = 1) - E(Y|x, D = 0)\}$ using the eq (4,2)
6. $E(Y|x, D = 1) = E(Y|D = 1) = b_1 + b_w$ so that:
7. $ATE = E_x\{E(Y|x, D = 1)\} - b_1 = E(Y, D = 1) - b_1 = b_w + b_1 - b_1 = b_w$

Choosing observables

Case 1 x->D->Y

1. Either if the outcome is balanced or not over x, this has no effect on the estimation of ATE
2. Therefore, conditioning on x is not necessary in this case

Choosing observables

Case 2 x->(D,y) D->Y

1. In this case:
2. $D = a_1 + a_x x + u$
3. $Y = b_1 + b_w D + b_x x + v$
4. We have:
   $E(Y|D=1) = E_x\{E(Y|x, D=1) = b_1 + b_w + b_x E(x|D=1)$
5. $E(Y|D=0) = E_x\{E(Y|x, D=0) = b_1 + b_x E(x|D=0)$
6. and we have: $ATE = b_w + b_x(E(x|D=1) - E(x|D=0))$

Without balancing the treated and untreated group on x we would get that $E(x|D=1) - E(x|D=0) \neq 0$, thus conditioning (that is equivalent to "balancing") on x is required. Otherwise, a bias equal to $b_x(E(x|D=1) - E(x|D=0))$ would appear in the estimation of ATE

Choosing observables

Case 3 d->Y<-x

1. We assume that D affects Y, x affects Y too, but there is no relation between D and X
2. The corresponding structural model becomes (again under CMI):
3. $D = a_1 + u$
4. $Y = b_1 + b_w D + b_x x + v$
5. We have: $E(Y|D=1) = E_x\{E(Y|x, D=1) = b_1 + b_w + b_x E(x|D=1) = b_1 + b_w + b_x E(x))$
6. $E(Y|D=0) = E_x\{E(Y|x, D=0) = b_1 + b_x E(x|D=0) = b_1 + b_x E(x)$
7. $ATE = b_w$

x must be controlled for only if x affects at the same time both D and Y. If x affects either only the selection equation or only the outcome equation, then controlling for x is not strictly necessary

Choosing observables

Case 4 xpre->(Y,D,xpost) d->Y Y->xpost

Similar to case (2), but:

1. We consider also that: a pretreatment x may have an effect on its post treatment status (self- effect), and
2. the outcome Y can also affect the post treatment status of X
3. In such a situation, while it is clearly needed to control for the pretreatment X (as in Case 2), it is not necessary to control for its post treatment status.
4. This is because Xpost is, in this case, just the result of the whole causal chain not explaining any other variable

Choosing observables

Case 5 d->xpost->Y

1. In this case the treatment D affects x that in turn affects Y.
2. As such, x takes the form of a post treatment variable working as a mediating factor (i.e., a factor causally laying between the treatment D and the outcome Y)
3. $X = c_1 + c_w D + u_x$
4. $Y = b_1 + b_w D + v$
5. In this case, conditioning on X is not needed as X does not appear in the Y- equation. Thus, one does not need to control for this variable

Choosing observables

Case 6 D->(Y,xpost) xpost->Y

1. In this case, the treatment D affects both x and Y
2. Y, in turn is also affected by X
3. X takes the form of a post-treatment variable working as a mediator
4. $X = c_1 + c_w D + u_x$
5. $Y = b_1 + b_w D + b_x x + v$ (1.84), substituting:
6. $Y = b_1 + b_w D + b_x(c_1 + c_w D + u_x) + v$ which leads to:
7. $Y = (b_1 + b_x c_1) + (b_w + b_x c_w)D + \eta$ (1.85)

Choosing observables

Case 6

1. Y-equation is the reduced form of the previous system of two equations.
2. Within this framework, we can define **three types of effect of D on Y**:
3. **Direct effect**: $E(Y|D=1,x) - E(Y|D=0,x) = b_w$
4. **Indirect effect**: $[E(x|D=1) - E(x|D=0)][E(Y|D=1,x) - E(Y|D=0,x)] = c_w b_x$
5. **Total effect**: $[E(Y|D=1) - E(Y|D=0)] = b_w + c_w b_x$

Choosing observables

Case 6

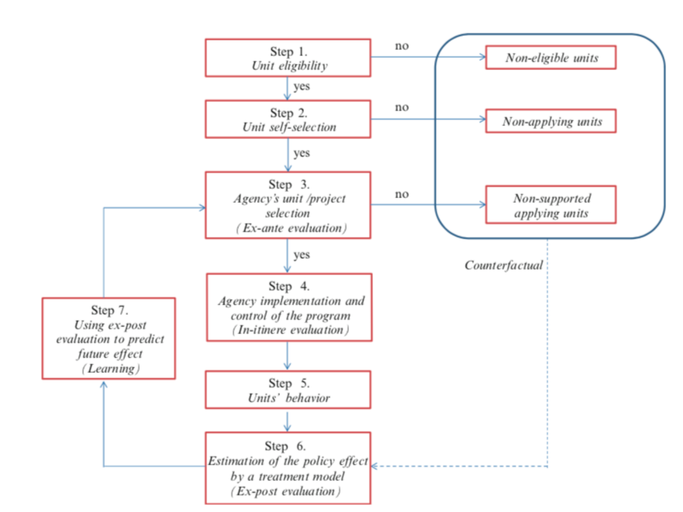
1. Total effect can be obtained -under CMI- by an OLS regression of the reduced form of the outcome Y.
2. Instead, the direct effect can be obtained by an OLS of (1.84), where both b_w and b_x are consistently estimated under CMI

Therefore, it is quite clear that:
if the analyst is interested in estimating the total effect of D on Y, then x should not need to be controlled for. Since we are interested in the direct effect of D on Y (i.e., the effect of D on Y "net of the effect of D on x"), then controlling for x is mandatory.

Policy Framework and the Statistical Design for Counterfactual Evaluation

1. A correct ex post program evaluation analysis first needs to draw upon a rich and qualified set of information
2. This information generally takes the form of: (1) **suitable indicators**, both qualitative and quantitative, and (2) availability of an **accurate sample** of (treated and untreated) units
3. The use of suitable indicators and of an appropriate sample of subjects are the product of the "framework" characterizing the functioning of the policy considered.
4. This framework may also suggest what econometric approach might be more suited for the specific context under scrutiny

Logical framework for an ex post assessment of a project-funding policy program



Introducing Regression Adjustment (RA)

We start introducing the RA method for evaluation
Assumptions:

1. Methods based on observables
2. we are ruling out any possible presence of loosely defined unobservables as hidden drivers of the selection process
3. Regression Adjustment (RA) is suitable only when the conditional independence assumption (CIA) holds:
4. $(Y_1, Y_0) \perp D|x$

Assumptions

Lets start observing that:

1. We assume CMI (less restrictive of CIA)
2. $E[Y_1|D,x] = E[Y_1|x]$
3. $E[Y_0|D,x] = E[Y_0|x]$
4. and we have:
5. $E[Y_0|D=1,x] = E[Y_0|D=0,x]$
6. $E[Y_1|D=0,x] = E[Y_1|D=1,x]$

right terms side are observable!

Assumptions II

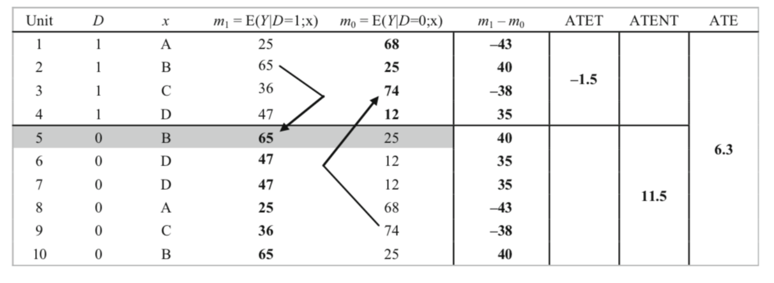
1. $ATE(x) = E[Y|D=1,x] - E[Y|D=0,x]$
2. We can interpret these quantities as:
3. $m_1(x) = E[Y_1|D=1,x]$, $m_0(x) = E[Y_0|D=0,x]$ are observed quantities!
4. $ATE(x) = m_1(x) - m_0(x)$
5. Getting:
6. $\widehat{ATE} = \frac{1}{N}\sum_i (\hat{m_{1i}}(x) - \hat{m_{0i}}(x))$
7. $\widehat{ATET} = \frac{1}{N}\sum_i D_i(\hat{m_{1i}}(x) - \hat{m_{0i}}(x))$
8. $\widehat{ATENT} = \frac{1}{N}\sum_i (1-D_i)(\hat{m_{1i}}(x) - \hat{m_{0i}}(x))$

Some observation about RA method

1. $m_1(x)$ and $m_0(x)$ can be estimated either parametrically, semi-parametrically, or nonparametrically: the choice depends on the assumption made on the form of the potential outcome, which can be modeled in a parametric as well as nonparametric or semi- parametric way

An example

1. Imputation is based on conditioning over the values of one single variable x, which is supposed to take on four values: A, B, C, D
2. The numbers reported in bold are those imputed according to the value assumed by x in the sample
3. For instance, consider m_1 for unit 5. In the sample, this unit is untreated: for such a unit, we observe m_0 but we do not observe the counterfactual m_1



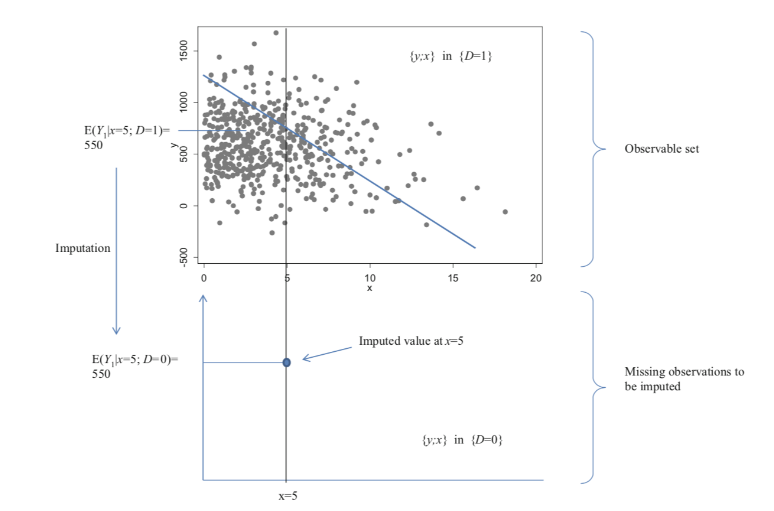
Unit	D	x	$m_1 = E(Y\|D=1;x)$	$m_0 = E(Y\|D=0;x)$	$m_1 - m_0$	ATET	ATENT	ATE
1	1	A	25	**68**	**–43**	**–1.5**		**6.3**
2	1	B	65	**25**	**40**			
3	1	C	36	**74**	**–38**			
4	1	D	47	**12**	**35**			
5	0	B	**65**	25	**40**		**11.5**	
6	0	D	**47**	12	**35**			
7	0	D	**47**	12	**35**			
8	0	A	**25**	68	**–43**			
9	0	C	**36**	74	**–38**			
10	0	B	**65**	25	**40**			

Discussion of the example

1. This example clearly proves that RA imputation works well only if we are able to "impute" $m_1(x_i)$ to each individual i belonging to the control group with $x = x_i$ and $m_0(x_i)$ to each individual i belonging to the treatment group with $x = x_i$.
2. Therefore, some minimal units' overlap over x is necessary for imputation to be achieved (and, thus, for identifying treatment effects)
3. Generally, however, perfect overlap between treated and untreated units (as in the previous example) may not occur in real contexts
4. Think what happen when x is continuos...

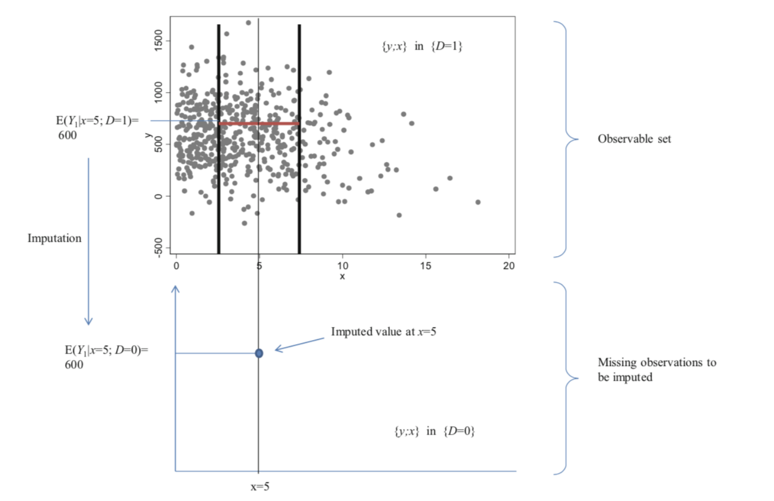
Grasping the intuition of RA

parametric RA



Grasping the intuition of RA II

Non parametric RA



Linear Regression adjustment

A parametric approach

1. We assume:
2. $m_0(x) = \mu_0 + x\beta_0$
3. $m_1(x) = \mu_1 + x\beta_1$
4. RA implies estimating two distinct OLS regressions: $Y_i = \mu_0 + x\beta_0$ only on untreated and $Y_i = \mu_1 + x\beta_1$ only on treated units,
5. thus getting the predicted values $\hat{m}_1(x)$ and $\hat{m}_0(x)$

A formal version of CFR

Control Function Regression

Consider the potential outcomes in a simple additive form as follows:

1. $Y_0 = \mu_0 + \nu_0$
2. $Y_1 = \mu_1 + \nu_1$
3. where ν_0 and ν_1 are random variables and μ_1 and μ_0 are scalars (we are assuming that outcomes consist of a constant term plus a random component)
4. Assume also:
5. $\nu_0 = g_0(x) + e_0$
6. $\nu_1 = g_1(x) + e_1$

making it explicit the dependence of the potential outcomes on the observable vector of covariates x

A formal version of CFR II

Starting from POM:

1. $Y = Y_0 + D(Y_1 - Y_0)$
2. $Y = \mu_0 + \nu_0 + D(\mu_1 + \nu_1 - \mu_0 - \nu_0) = \mu_0 + \nu_0 + D(\mu_1 - \mu_0) + D(\nu_1 - \nu_0)$
3. $Y = \mu_0 + D(\mu_1 - \mu_0) + g_0(x) + D(g_0(x) + e_0 + g_1(x) + e_1) + e_0 = \mu_0 + D(\mu_1 - \mu_0) + g_0(x) + D(g_0(x) + g_1(x)) + e$
4. let assume: $g_0(x) = x\beta_0$ and $g_1(x) = x\beta_1$
5. Taking expected value:
6. $E(Y|D,x) = \mu_0 + D(\mu_1 - \mu_0) + g_0(x) + D(g_1(x) - g_0(x))$

Homogeneous reaction function

1. HP: $g_0(x) = g_1(x)$ $(\beta_0 = \beta_1 = \beta)$
2. $ATE = ATE(x) = ATET = ATET(x) = ATENT = ATENT(x) = \mu_1 - \mu_0$
3. $E(Y|D, x) = \mu_0 + D \cdot ATE + x\beta$
4. indeed: $ATE = E(Y_1 - Y_0) = E[\mu_1 + g_1(x) + e_1 - (\mu_0 + g_0(x) + e_0)] = \mu_1 - \mu_0$
5. We can estimate OLS on:
6. $Y = \mu_0 + D \cdot \alpha + x\beta + \varepsilon$
7. so that: $\alpha = ATE$

Heterogeneous reaction function

1. HP: $g_0(x) \neq g_1(x)$
2. $ATE \neq ATE(x) \neq ATET \neq ATET(x) \neq ATENT \neq ATENT(x)$
3. $E(Y|D,x) = \mu_0 + D \cdot ATE + x\beta_0 + D(x - \mu_x)\beta$
4. where: $\mu_x = E(x)$, $\beta = \beta_1 - \beta_0$

Heterogeneous reaction function II

1. In this case, heterogeneous average treatment effects (over x) exist and the population causal parameters take on the following form:
2. $ATE = (\mu_1 - \mu_0) + \mu_x \beta$
3. $ATE(x) = ATE + (x - \mu_x)$
4. $ATET = ATE + E_X\{x - \mu_x | D = 1\}\beta$
5. $ATET(x) = [ATE + (x - \mu_x)\beta | D = 1]$
6. $ATENT = ATE + E_X\{x - \mu_x | D = 0\}\beta$
7. $ATENT(x) = [ATE + (x - \mu_x)\beta | D = 0]$

Heterogeneous reaction function III

Given these formulas for the population causal parameters, the sample estimates can be obtained by relying on the sample equivalents:

1. $\widehat{ATE} = \hat{\alpha}$
2. $\widehat{ATE(x)} = \hat{\alpha} + (x - \overline{x})\hat{\beta}$
3. $\widehat{ATET} = \hat{\alpha} + \frac{1}{N_1} \sum_i D_i(x_i - \overline{x})\hat{\beta}$
4. $\widehat{ATET(x)} = [\hat{\alpha} + (x - \overline{x})\hat{\beta}]_{(D=1)}$
5. $\widehat{ATENT} = \hat{\alpha} + \frac{1}{N_0} \sum_i (1 - D_i)(x_i - \overline{x})$
6. $\widehat{ATENT}(x) = [\hat{\alpha} + (x - \overline{x})\hat{\beta}]_{(D=0)}$

Heterogeneous reaction function IV

The estimated causal parameters of interest depend in turn on the unknown parameters: $\mu_1, \mu_0, \beta_1, \beta_0, \mu_x$ Procedure:

1. Estimate with OLS: $Y = \mu_0 + D \cdot \alpha + x\beta + D(x - \mu_x)\beta + \varepsilon$
2. Plug these estimated parameters into the sample formulas and recover all the causal effects
3. Obtain standard errors for ATET and ATENT via bootstrap

Observation

In some contexts assuming homogeneous response to confounders might be questionable

1. In many sociological environments, for instance, people's perception of the context may change according to a different state of the world (treated vs. untreated situations)
2. In the economic context, a company characterized by a weak propensity to bearing risks may become more prone to invest in a riskier business when public funding is available: for instance, such a company might change its reaction to, let's say, its stock of fixed capital when financed, by increasing its productive response to this asset
3. Similar conclusions can be reached from many psychological or sociological programs, as passing from the untreated to the treated status may produce different mental, relational, and environmental situations.

Observation II

1. How to test for heterogeneity:
2. Using: $E(Y|D,x) = \mu_0 + D \cdot ATE + x\beta_0 + D(x - \mu_x)\beta$
3. Test available:
4. a simple F-test of joint significance for the coefficients in vector β can be exploited to check the presence of heterogeneity; if the null hypothesis H0: $\beta = (\beta_1 - \beta_0) = 0$ is rejected, then it means that heterogeneity is at work, and vice versa.

A graphical representation

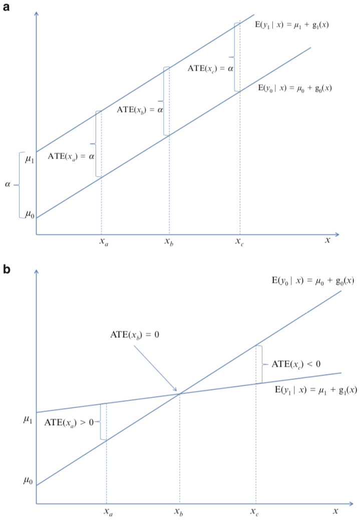


Fig. 2.3 A graphical representation of the potential outcomes function and of the corresponding ATE(x) under homogeneous (**a**) and heterogeneous (**b**) response to x

Nonlinear Parametric Regression-Adjustment

1. When the outcome is binary or count, however, the linearity assumption can be relaxed, and a proper parametric form of $m_0(x)$ and $m_1(x)$ can be assumed
2. By substituting previous formulas into the Regression-adjustment formulas, we can obtain the corresponding non linear Regression-adjustment estimators for ATEs. For instance, when the outcome variable is a count, a consistent estimation of ATET is:
3. $\widehat{ATET} = \frac{1}{N_1} \sum_i D_i [exp(x_i \hat{\beta}_1) - exp(x_i \hat{\beta}_0)]$

Nonlinear Parametric Regression-Adjustment

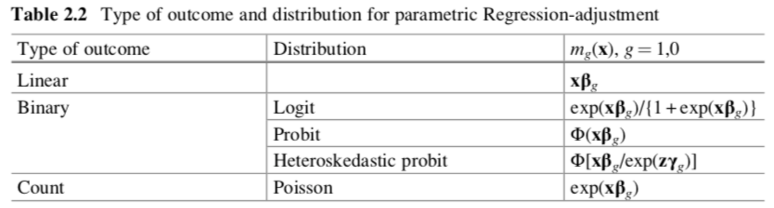


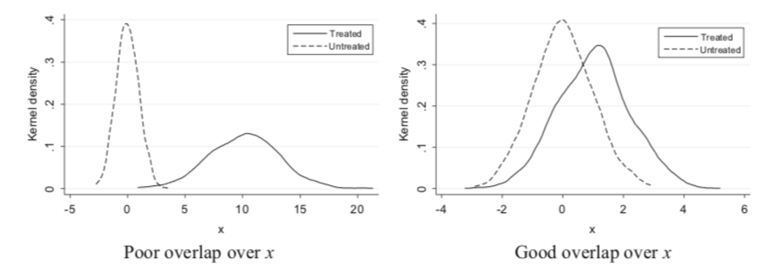
Table 2.2 Type of outcome and distribution for parametric Regression-adjustment

Type of outcome	Distribution	$m_g(\mathbf{x})$, $g = 1,0$
Linear		$\mathbf{x}\boldsymbol{\beta}_g$
Binary	Logit	$\exp(\mathbf{x}\boldsymbol{\beta}_g)/\{1 + \exp(\mathbf{x}\boldsymbol{\beta}_g)\}$
	Probit	$\Phi(\mathbf{x}\boldsymbol{\beta}_g)$
	Heteroskedastic probit	$\Phi[\mathbf{x}\boldsymbol{\beta}_g/\exp(\mathbf{z}\boldsymbol{\gamma}_g)]$
Count	Poisson	$\exp(\mathbf{x}\boldsymbol{\beta}_g)$

Nonparametric and Semi-parametric Regression-Adjustment

1. Local smoothing techniques such as kernel or local linear regression can be used to obtain nonparametric estimation of $m_1(x)$ and $m_0(x)$
2. these approaches are, however, unfeasible when no minimal overlap between treated and control group is present over x. This may occur in datasets where the support of the covariates x in the treated and untreated group is very different, and thus, the overlap is poor

Nonparametric and Semi-parametric Regression-Adjustment II



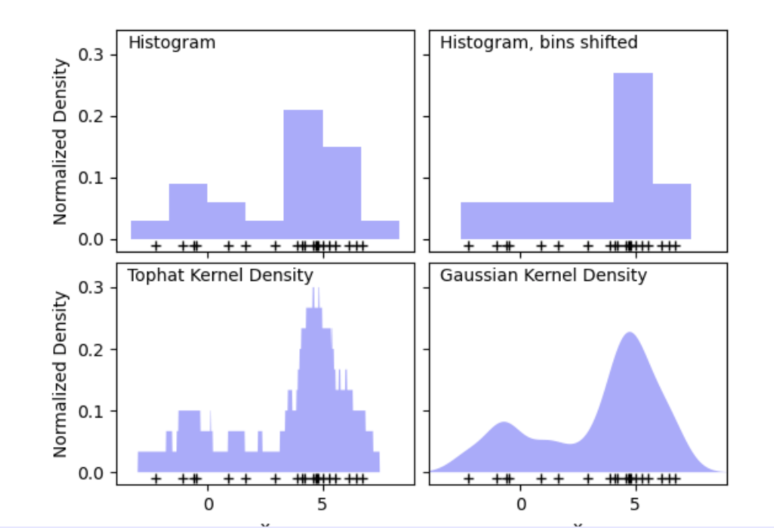
Poor overlap over *x*

Good overlap over *x*

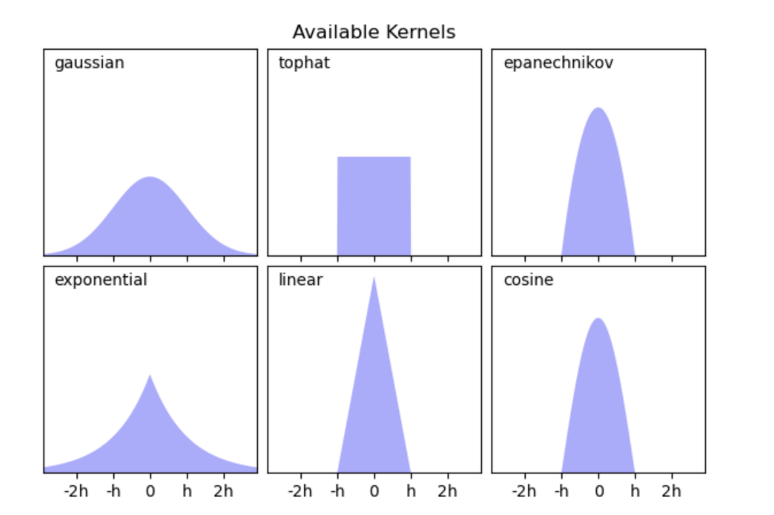
Nonparametric and Semi-parametric Regression-Adjustment III

1. $\hat{m}_g(x) = \sum_{i:D_{i=g}} Y_i K(\frac{x_i - x}{h}) / \sum_{i:D_{i=g}} K(\frac{x_i - x}{h})$
2. $K()$: kernel function
3. h: bandwidth parametrer

Nonparametric and Semi-parametric Regression-Adjustment VI



Nonparametric and Semi-parametric Regression-Adjustment VI



Matching

The idea behind

1. The idea behind Matching is simple, intuitive, and attractive, and this can partly explain its popularity
2. It can be summarized in the following statement: ”recovering the unobservable potential outcome of one unit using the observable outcome of similar units in the opposite status”

Matching

A special case of RA

Matching is equivalent to the nonparametric RA estimator seen above where, instead of using a nonparametric estimation of the observable conditional mean, one uses directly the observed outcome.

The Matching formulas for ATEs are:

1. $\widehat{ATET}_M = \frac{1}{N_1} \sum_i D_i[Y_i - m_0\hat{(}x_i)]$
2. $\widehat{ATENT}_M = \frac{1}{N_0} \sum_i (1 - D_i)[m_1(x_i\hat{)} - Y_i]$
3. $\widehat{ATE}_M = \frac{1}{N} \sum_i \{D_i[Y_i - m_0\hat{(}x_i)] + (1 - D_i)[m_1\hat{(}x_i) - Y_i]\}$

Matching also identifies ATEs under the CMI assumption

Matching

Features

1. It does not require to specify a specific parametric relation between potential outcomes and confounding variables.
2. In contrast to the CFR approach, a wide set of different Matching procedures can be employed, thus enabling one to compare various estimators and provide robustness to results.
3. It reduces the number of untreated to a subsample (the so-called selected controls) having structural characteristics more homogeneous to the those of treated units
4. It considers treated and untreated units to be compared only in the so-called common support, dropping out all those controls whose confounders values are either higher or smaller than that of the treated units.

Matching

1. Take the case of the estimation of ATET:
2. $ATET(x) = E[Y_1 - Y_0|D = 1, x]$, is not computable without any other assumption
3. Suppose, however, that Y_0 is perfectly estimated by using some average of the outcome of (matched) untreated individuals and call this quantity $\hat{Y}_0$
4. Y_0 imputed through a distance function f → $\hat{Y}_0$

Matching assumptions

1. The choice of the function f corresponds to a specific distance metric between treated and untreated units

Measuring such a distance can be done in two ways:

1. Based on the vector of covariates x, so that one can calculate, in a meaningful manner, how far x_i is from x_j, where unit j is assumed to be in the opposite treatment group, i.e., $D_j = 1 - D_i$ (covariates Matching or C Matching)
2. On the basis of only one single index-variable, the propensity-score $p(x_i)$, synthesizing all covariates in a one-dimension variable (propensity-score Matching or PS Matching).

Matching

Estimation of ATEs

Irrespective of the specific method chosen, the estimation of the $ATE_i(x_i)$ would be simply given by:

1. $\widehat{ATE}_i(x_i) = Y_{1i} - \hat{Y}_{0i}$
2. $\widehat{ATE} = \frac{1}{N}\sum_i(\hat{Y}_{1i} - \hat{Y}_{0i})$
3. $\widehat{ATET} = \frac{1}{N_1}\sum_i D_i(Y_{1i} - \hat{Y}_{0i})$
4. $\widehat{ATENT} = \frac{1}{N_0}\sum_i(1 - D_i)(\hat{Y}_{1i} - Y_{0i})$
5. RA for ATET: Setting $\hat{m}_1(x_i) = Y_{1i}$ and $\hat{m}_0(x_i) = \hat{Y}_{0i}$

Identification of ATEs Under Matching

Assumptions

Under specific assumptions, Matching is suited for eliminating biases due to weak overlap and to weak balancing:

1. Conditional mean independence (CMI): $E(Y_1|x, D) = E(Y_1|x)$ and $E(Y_0|x, D) = E(Y_0|x)$
2. Overlap: $0 < p(x) < 1$ where $p(x) = Pr(D = 1|x)$ (propensity score)

More precisely, however, ATEs are only identified under assumptions A.1 and A.2 if the Matching is exact, i.e., only if it is possible to build a finite number of cells based on crossing the values taken by the various x

Identification of ATEs Under Matching II

When this is not possible, as usually happens, when x contains at least one continuous variable, then we need a third hypothesis in order to identify ATEs:

1. Balancing: $[(D \perp x) \mid \text{Matching}]$, i.e., after matching, the covariates' distribution in the treated and control group has to be equal.

Implications of Assuming CIA

Under CIA we have:

1. $ATET(x) = E[Y_1|D=1,x] - E[Y_0|D=1,x] =$
2. $= E[Y_1|D=1,x] - E[Y_0|D=1,x] + E[Y_0|D=0,x] - E[Y_0|D=0,x]$
3. and also:
4. $E[Y_0|D=1,x] = E[Y_0|D=0,x]$

This relation suggests one should estimate (or impute) the unobservable (or missing) value using the observable quantity $ATET(x) = E[Y|D=1,x] - E[Y|D=0,x]$

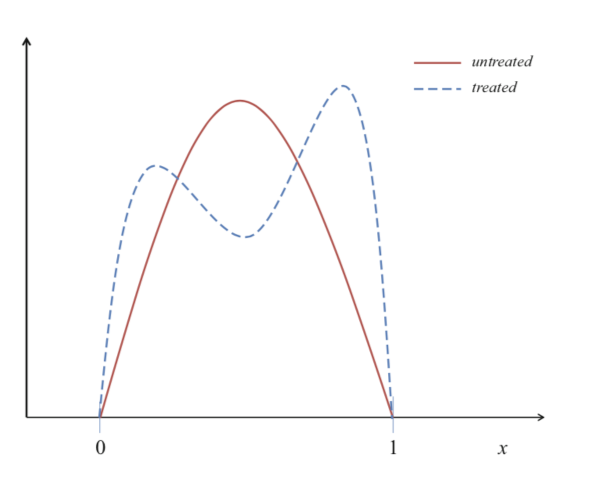
Implications of Assuming Overlap

1. If this assumption does not hold, there might exist units with specific characteristic x that either always receive treatment (i.e., $p(x) = 1$) or never receive treatment (i.e., $p(x) = 0$), thus **not permitting us to identify ATEs**
2. In empirical practice finding cases in which $p(x) = 1$ or $p(x) = 0$ is unlikely. in the case of Matching, **some imprecision in the capacity of x to explain all the variability of** $p(x)$ solves the identification problem
3. As a result, the model used to predict program participation should not be "too" good!

Implications of Assuming ''Balancing''

1. This assumption matters when Matching is not exact, a case typically occurring when **x presents at least one continuous variable**
2. Matching should help to restore some balancing over x
3. To be able to estimate ATEs we have to rely on a ''plausible degree'' of balancing over the observables; this should be possible to test using some suitable **test statistics after Matching** is completed
4. Observe that the overlap and the balancing **Hps are two distinct**, although partially linked, **assumptions**
5. Indeed, we might find a good degree of covariates' overlap, sometimes accompanied with some strong imbalance

Implications of Assuming "Balancing" II



Common Support

1. **Common support (S) HP**: states that Matching can be equally consistently estimated not only over the all support of x but also on the support of x common to both participant and comparison groups:
2. $S = Supp(x|D = 1) \bigcap Supp(x|D = 0)$
3. we may estimate Matching using a reduced sample by applying a **trimming rule**, which is a rule to reduce the number of units employed in estimation to the common support S
4. In general, the quality of the matches may be improved by imposing the common support restriction
5. However: high-quality matches may be lost at the boundaries of the common support and the sample may be considerably reduced

Exact Matching

This procedure exploits the idea that **within cells identified by x** the condition for **random assignment is restored** so that **intracell DIM is a consistent estimator**

Procedure to estimate ATEs:

1. The **data are stratified into cells** defined by each particular value of x
2. Within each cell (i.e., conditioning on x), one should compute the **difference between the average outcomes of the treated and that of the controls**
3. These differences should be averaged with respect to the distribution of x in the population of treated (for ATET) or untreated (for ATENT) units

Exact Matching II

This procedure leads to the following estimators of ATEs:

1. $\widehat{ATET} = E_x\{E(Y_1 - \hat{Y}_0|D = 1, x)\} = \sum_x \hat{TE}_x p(x_i = x|D_i = 1)$
2. $\widehat{ATENT} = E_x\{E(\hat{Y}_1 - Y_0|D = 0, x)\} = \sum_x \hat{TE}_x p(x_i = x|D_i = 0)$
3. $\widehat{ATE} = E_D\{E_x\{E(\hat{Y}_1 - \hat{Y}_0|D, x) =$
4. $= p(D = 1) \cdot \widehat{ATET} + p(D = 0) \cdot \widehat{ATENT}$

These ATEs estimators defines exact Matching, and it is feasible **only when x has a very small dimensionality** (taking, for instance, just three values).

But if the **sample is small**, the **set of covariates x is large** and many of them take discrete multi-values or, even worse, they are continuous variables, then **exact Matching is unfeasible**

The dimensionality problem

1. For example, if x is made of K binary variables, then the number of cells becomes 2K, and this number increases further if some variables take more than two values
2. If the number of cells (or blocks) is very large with respect to the size of the sample, it is possible that some cells contain only treated or only control subjects. Thus, the calculus of ATEs might become unfeasible and ATEs not identified
3. **If variables are all continuous, as happens in many socioeconomic applications, it would be even impossible to build cells**

This drawback, known as the **dimensionality problem**, Rosenbaum and Rubin (1983) have suggested that units are matched according to the propensity-score (defined, as said above, as the ”probability of being treated conditional on x”)

Using the **propensity-score** permits to reduce the multidimensionality to a single scalar dimension, p(x)

The Properties of the Propensity-Score

1. The **propensity-score is the conditional probability of receiving the treatment, given the confounding variables x**
2. Since D is binary, the following equalities apply:
3. $p(x) = Pr(D = 1|x) = E(D|x)$
4. that is, the propensity-score is the expectation of the treatment variable, conditional on x.

The propensity-score has **two important properties** which account for its appeal: the balancing and unconfoundedness properties.

Balancing of confounding variables

P1

1. If $p(x)$ is the propensity-score, then:
2. $D \perp x|p(x)$
3. (i.e. conditionally on p(x), the treatment and the observables are independent)
4. proof: $Pr(D = 1|x, p(x)) = E(D|x, p(x)) = E(D|x) = Pr(D = 1|x) = p(x)$
5. $Pr(D = 1|p(x)) = E(D|p(x)) = E_{p(x)}[E(D|x, p(x)|p(x)] = E_{p(x)}[p(x)|p(x)] = p(x)$

which entails that conditionally on p(x), the treatment D and the observables x are independent.

Unconfoundedness, given the propensity-score

P2

1. Suppose that the conditional independence assumption (CIA) holds: $(Y_1, Y_0) \perp D|x$
2. then assignment to treatment is random, also given the propensity-score, that is:
3. $(Y_1, Y_0) \perp D|p(x)$

(proof not given...)

Unconfoundedness, given the propensity-score II

1. Property P2 states that stratifying units according to p(x) produces the same orthogonal condition between the potential outcomes and the treatment that is stratifying on x, but with the advantage to rely just on one dimension variable
2. Property P1, additionally, states that if the propensity-score is correctly specified, then we should see that units stratified according to the propensity-score should be indistinguishable in terms of their x (i.e., they are balanced)
3. if the **propensity-score is correctly specified**, then we should see that units stratified according to the propensity-score should be indistinguishable in terms of their x (i.e., they are balanced). Thus, **testing empirically whether the balancing property holds is a way for assuring that the correct propensity-score is being used to stratify units.**

UNIVERSITÀ DI TRENTO

Quasi-Exact Matching Using the Propensity-Score

1. Assumption P2 suggests to match treated units and controls directly on the basis of the (estimated) propensity-score instead of using the larger set of variables in x
2. Exact Matching with a continuous variable is impossible, as none of the units have exactly the same value of such a variable.
3. Nevertheless, a **discretization procedure of the propensity-score may still be implemented to approximate the Exact-Matching approach**
4. Dehejia and Wahba (1999)

Estimating the propensity-score

1. Start with a **parsimonious specification** in order to estimate the propensity-score for each individual, using the following function: $P(x) = G[f(x)]$
2. where G[] can be probit, logit, or linear, and f(x) is a function of covariates with linear and higher order terms
3. **Order the units** according to the estimated propensity-score (from the lowest to the highest value)
4. **Stratify all observations into blocks** such that in each block, the estimated propensity-scores for the treated and the controls are not statistically different:
    - Start with five blocks of equal score range 0-0.2, . . ., 0.8-1
    - Test whether the means of the scores for the treated and the controls are statistically different in each block (balancing of the propensity-score)
    - If they are, increase the number of blocks and test again
    - If not, proceed to the next step

Estimating the propensity-score II

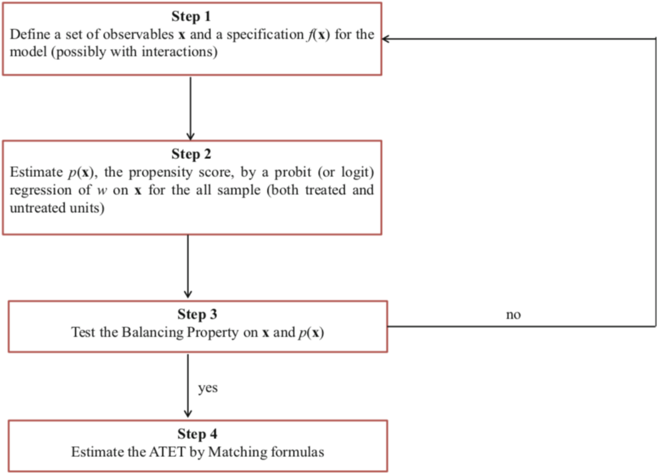
5. **Test whether the balancing property** holds in all strata for all covariates:
    - For each covariate, test whether the means for the treated and for the controls are statistically different in all strata (balancing for covariates)
    - If one covariate is not balanced in one block, split the block and test again within each finer block
    - If one covariate is not balanced in all blocks, modify the logit/probit/linear estimation of the propensity-score adding more interaction and higher order terms and then test the balancing property again.
6. Once the balancing property is satisfied and, thus, the **optimal number of strata** is found, then an (weighted) **average of the DIM estimators calculated in the final blocks provides an estimation of ATEs** (stratification Matching)

Estimating ATEs:

1. In standard applications, the quasi-exact-Matching procedure proposed by DW (1999) may be rather demanding, as it may be difficult to assure balancing for all covariates within all strata
2. Other Matching methods provide a less restrictive and, thus, easier way to obtain reliable estimates of ATEs, without requiring to build blocks

Estimating ATEs:

A typical procedure for estimating ATEs without blocks



Step 1
Define a set of observables **x** and a specification *f*(**x**) for the model (possibly with interactions)

Step 2
Estimate *p*(**x**), the propensity score, by a probit (or logit) regression of *w* on **x** for the all sample (both treated and untreated units)

Step 3
Test the Balancing Property on **x** and *p*(**x**)

Step 4
Estimate the ATET by Matching formulas

Fig. 2.6 Flow diagram of a Matching protocol

Estimating ATEs:

1. In this case, one should apply Matching estimation just when for each x and for p(x), no difference emerges in terms of the mean of treated and matched untreated units
2. The limits reside in the use of a less sophisticated test of the balancing property

Some diagnostic test

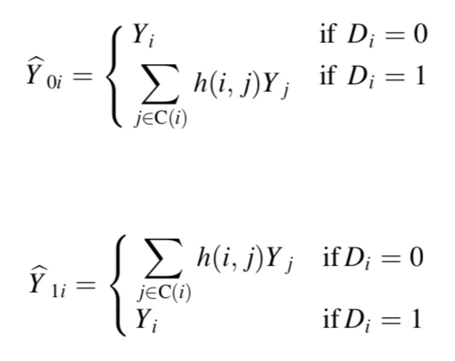
1. **Perfect balancing is impossible** due to the random nature of the data and even more importantly because the analyst rarely has access to the entire set of confounders explaining the selection-into-program
2. **if treated and control units are largely different** in terms of observables, the reached Matching is not sufficiently robust and it might be questionable.
3. **Comparison of the estimated propensity-scores across treated and controls** therefore provides a useful **diagnostic tool**

Some diagnostic test II

1. Calculate the frequency of matched untreated cases having a propensity-score lower than the minimum or higher than the maximum of the propensity-scores of the treated units (Preferably, one would hope that the range of variation of propensity-scores is the same in both groups)
2. Draw histograms and kernel densities of the estimated propensity-scores for the treated and the controls, before and after Matching when possible. In case of stratification Matching, one should use histogram bins corresponding to the strata constructed for the estimation of propensity-scores (one hopes to get an equal frequency of treated and untreated units in each bin)

Methods for Propensity-Score Matching

Smith and Todd (2005): conditional diff-in-diffs method

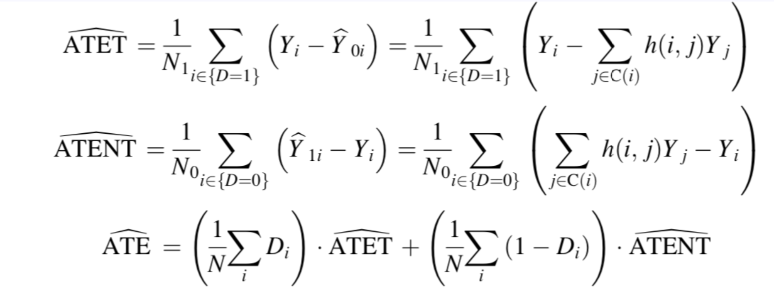


$$\widehat{Y}_{0i} = \begin{cases} Y_i & \text{if } D_i = 0 \\ \sum_{j \in \mathrm{C}(i)} h(i, j) Y_j & \text{if } D_i = 1 \end{cases}$$

$$\widehat{Y}_{1i} = \begin{cases} \sum_{j \in \mathrm{C}(i)} h(i, j) Y_j & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

Methods for Propensity-Score Matching II

Smith and Todd (2005): conditional diff-in-diffs method



$$\widehat{\text{ATET}} = \frac{1}{N_1} \sum_{i \in \{D=1\}} \left(Y_i - \widehat{Y}_{0i} \right) = \frac{1}{N_1} \sum_{i \in \{D=1\}} \left(Y_i - \sum_{j \in C(i)} h(i, j) Y_j \right)$$

$$\widehat{\text{ATENT}} = \frac{1}{N_0} \sum_{i \in \{D=0\}} \left(\widehat{Y}_{1i} - Y_i \right) = \frac{1}{N_0} \sum_{i \in \{D=0\}} \left(\sum_{j \in C(i)} h(i, j) Y_j - Y_i \right)$$

$$\widehat{\text{ATE}} = \left(\frac{1}{N} \sum_i D_i \right) \cdot \widehat{\text{ATET}} + \left(\frac{1}{N} \sum_i (1 - D_i) \right) \cdot \widehat{\text{ATENT}}$$

Different propensity-score Matching methods can be obtained by specifying different forms of the weights h(i, j) and of the set C(i)

Methods for Propensity-Score Matching

Smith and Todd (2005)

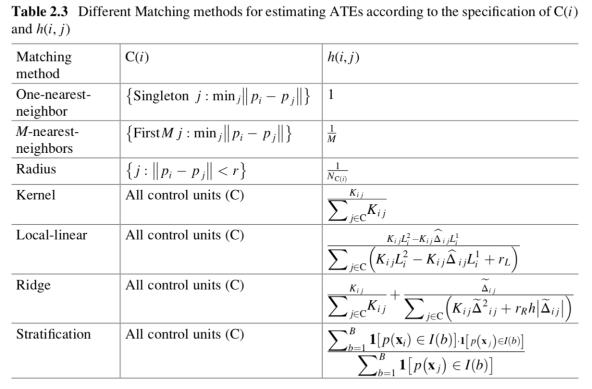


Table 2.3 Different Matching methods for estimating ATEs according to the specification of C(i) and $h(i, j)$

Matching method	C(i)	$h(i,j)$
One-nearest-neighbor	$\{\text{Singleton } \; j : \min_j \lVert p_i - p_j \rVert\}$	1
M-nearest-neighbors	$\{\text{First}\, M \; j : \min_j \lVert p_i - p_j \rVert\}$	$\frac{1}{M}$
Radius	$\{j : \lVert p_i - p_j \rVert < r\}$	$\frac{1}{N_{C(i)}}$
Kernel	All control units (C)	$\frac{K_{ij}}{\sum_{j \in C} K_{ij}}$
Local-linear	All control units (C)	$\frac{K_{ij} L_i^2 - K_{ij} \widehat{\Delta}_{ij} L_i^1}{\sum_{j \in C}\left(K_{ij} L_i^2 - K_{ij} \widehat{\Delta}_{ij} L_i^1 + r_L\right)}$
Ridge	All control units (C)	$\frac{K_{ij}}{\sum_{j \in C} K_{ij}} + \frac{\widetilde{\Delta}_{ij}}{\sum_{j \in C}\left(K_{ij} \widetilde{\Delta}^2{}_{ij} + r_R h \lvert \widetilde{\Delta}_{ij} \rvert\right)}$
Stratification	All control units (C)	$\frac{\sum_{b=1}^{B} \mathbf{1}[p(\mathbf{x}_i) \in I(b)] \cdot \mathbf{1}[p(\mathbf{x}_j) \in I(b)]}{\sum_{b=1}^{B} \mathbf{1}[p(\mathbf{x}_j) \in I(b)]}$

Nearest-neighbor Matching

The classical nearest-neighbor Matching suggests to match each treated unit with the closest untreated unit in the dataset, where ”closeness” is defined according to some distance metric over $p(x)$

1. When pair-wise matching is allowed, we have the so-called one-to-one nearest-neighbor Matching.
2. Generally, however, each unit in a given treatment status is matched with the closest M neighbors in the opposite status, and an average of them is thus produced as counterfactual.
3. Observe that matching may be done with and without replacement. When replacement is allowed, then the same unit can be used for more than one unit in the opposite status

Nearest-neighbor Matching II

Procedura to implement NNM:

1. or each treated unit i find the nearest control unit j using the Mahalanobis/ Euclidean distance
2. if the nearest control unit has already been used, use it again (replacement)
3. drop the unmatched controlled units
4. calculate ATEs applying formulas given

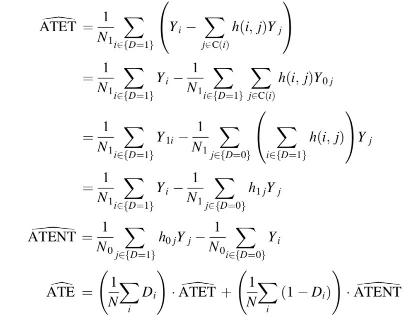
Radius(or caliper) Matching

1. A limit of the nearest-neighbor Matching is that it does not consider the distance between matches. This means that it could match pairs even when they are very different (as p_i and p_j are far)
2. It can be seen as a variant of the nearest-neighbor, trying to avoid the occurrence of bad matches by imposing a threshold on the maximum distance permitted between p_i and p_j
3. Two units are matched only when their distance in absolute terms is lower than a tolerance limit, identified by a prespecified caliper ”r”
4. Those treated units with no matches within the caliper are eliminated. Thus, radius Matching naturally imposes a common support restriction

Radius(or caliper) Matching II

1. for each treated unit i identify all the control units whose x differs by less than a given tolerance r (the caliper) chosen by the researcher
2. allow for replacement of control units
3. when a treated unit has no control closer than r, take the nearest control or delete it
4. estimate ATEs applying formulas

Radius(or caliper) Matching and NN



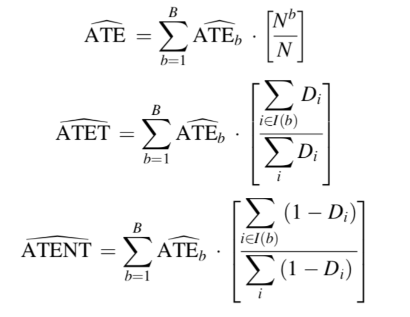
$$
\begin{aligned}
\widehat{\mathrm{ATET}} &= \frac{1}{N_1}\sum_{i\in\{D=1\}}\left(Y_i - \sum_{j\in \mathrm{C}(i)} h(i,j)Y_j\right)\\
&= \frac{1}{N_1}\sum_{i\in\{D=1\}} Y_i - \frac{1}{N_1}\sum_{i\in\{D=1\}}\sum_{j\in \mathrm{C}(i)} h(i,j)Y_{0j}\\
&= \frac{1}{N_1}\sum_{i\in\{D=1\}} Y_{1i} - \frac{1}{N_1}\sum_{j\in\{D=0\}}\left(\sum_{i\in\{D=1\}} h(i,j)\right)Y_j\\
&= \frac{1}{N_1}\sum_{i\in\{D=1\}} Y_i - \frac{1}{N_1}\sum_{j\in\{D=0\}} h_{1j}Y_j\\
\widehat{\mathrm{ATENT}} &= \frac{1}{N_0}\sum_{j\in\{D=1\}} h_{0j}Y_j - \frac{1}{N_0}\sum_{i\in\{D=0\}} Y_i\\
\widehat{\mathrm{ATE}} &= \left(\frac{1}{N}\sum_i D_i\right)\cdot \widehat{\mathrm{ATET}} + \left(\frac{1}{N}\sum_i (1-D_i)\right)\cdot \widehat{\mathrm{ATENT}}
\end{aligned}
$$

Kernel and local linear Matching

1. The kernel Matching estimator can be interpreted as a particular version of the radius Matching in which every treated unit is matched with a weighted average of all control units with weights that are inversely proportional to the distance
2.

Stratification Matching

1. this method exploits directly the propensity-score property P2 i.e., independence conditional to the propensity-score.
2. If this assumption holds, then it suggests that within cells (or blocks), identified by splitting the sample according to the values assumed by x, the random assignment is restored



$$\widehat{\text{ATE}} = \sum_{b=1}^{B} \widehat{\text{ATE}}_b \cdot \left[\frac{N^b}{N}\right]$$

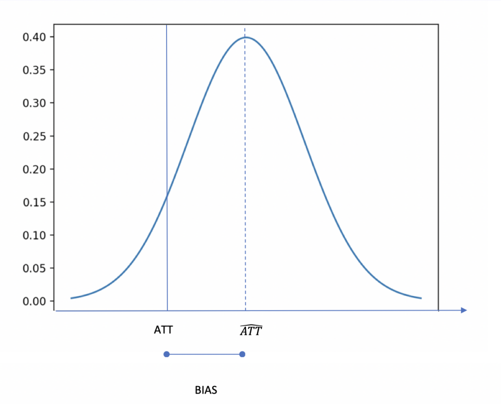
$$\widehat{\text{ATET}} = \sum_{b=1}^{B} \widehat{\text{ATE}}_b \cdot \left[\frac{\sum\limits_{i \in I(b)} D_i}{\sum\limits_{i} D_i}\right]$$

$$\widehat{\text{ATENT}} = \sum_{b=1}^{B} \widehat{\text{ATE}}_b \cdot \left[\frac{\sum\limits_{i \in I(b)} (1 - D_i)}{\sum\limits_{i} (1 - D_i)}\right]$$

Caliendo Koepinig

A premise:

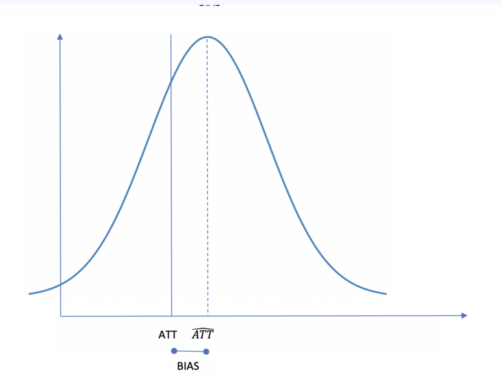
1. For each estimator of true ATT we have $\widehat{ATT}$ and $\widehat{\sigma_{ATT}}$



Caliendo Koepinig

less Bias

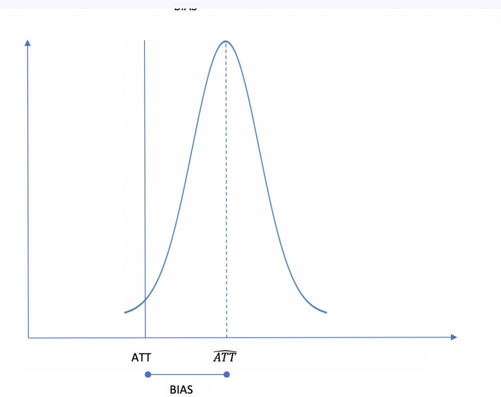
More precise point estimate (equal variance)



Caliendo Koepinig

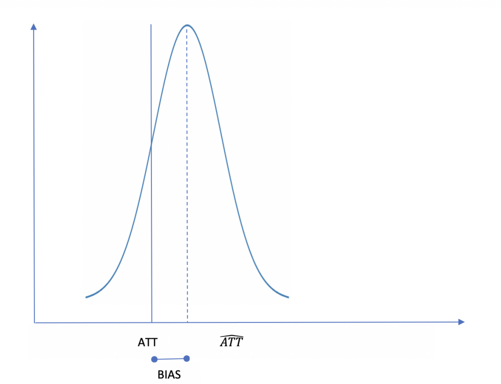
less Variance

equal precision less variance

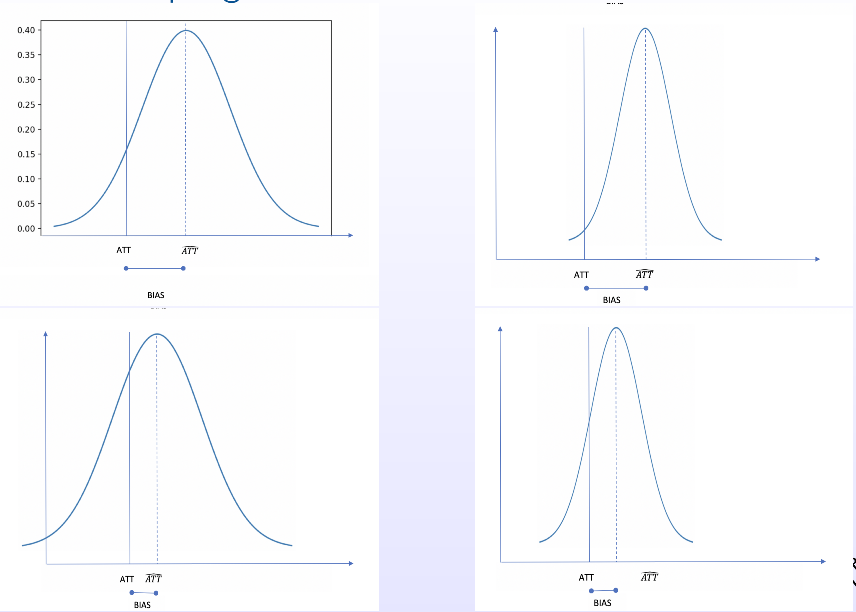


Caliendo Koepinig

less Variance+more precision



Caliendo Koepinig



Caliendo Koepinig

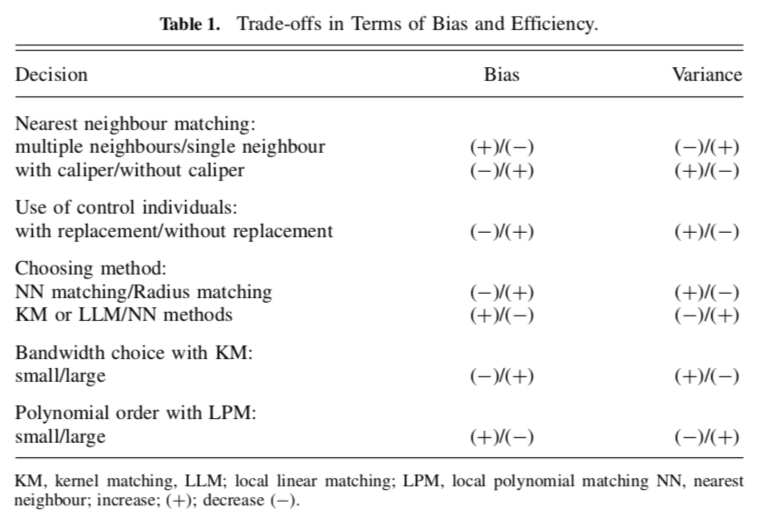


Table 1. Trade-offs in Terms of Bias and Efficiency.

Decision	Bias	Variance
Nearest neighbour matching:		
multiple neighbours/single neighbour	(+)/(−)	(−)/(+)
with caliper/without caliper	(−)/(+)	(+)/(−)
Use of control individuals:		
with replacement/without replacement	(−)/(+)	(+)/(−)
Choosing method:		
NN matching/Radius matching	(−)/(+)	(+)/(−)
KM or LLM/NN methods	(+)/(−)	(−)/(+)
Bandwidth choice with KM:		
small/large	(−)/(+)	(+)/(−)
Polynomial order with LPM:		
small/large	(+)/(−)	(−)/(+)

KM, kernel matching, LLM; local linear matching; LPM, local polynomial matching NN, nearest neighbour; increase; (+); decrease (−).

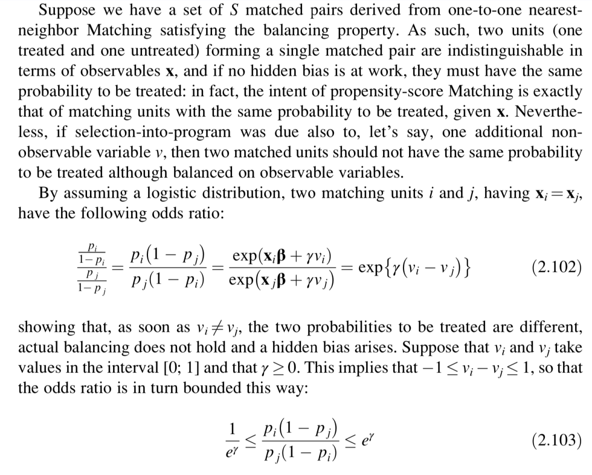
Observation about Inference

1. Abadie and Imbens (2006, 2012) show that for Matching with replacement, using the true propensity-score as the only matching variable, we have that:
2. $$\sqrt{N}\left(\widehat{\text{ATE}} - \text{ATE}\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

Sensitivity analysis

1. The aim of sensitivity analysis is that of assessing whether results obtained by applying a given estimation method are sufficiently reliable when the main assumptions under which the results are drawn may not be fully satisfied
2. Rosenbaum (2002, 2005) provides a powerful sensitivity analysis test when Matching is used in observational studies.
3. The aim of this test is that of assessing the reliability of ATEs estimations when unobservable selection (and thus hidden bias) might be present

Sensitivity analysis



Suppose we have a set of S matched pairs derived from one-to-one nearest-neighbor Matching satisfying the balancing property. As such, two units (one treated and one untreated) forming a single matched pair are indistinguishable in terms of observables $\mathbf{x}$, and if no hidden bias is at work, they must have the same probability to be treated: in fact, the intent of propensity-score Matching is exactly that of matching units with the same probability to be treated, given $\mathbf{x}$. Nevertheless, if selection-into-program was due also to, let's say, one additional non-observable variable v, then two matched units should not have the same probability to be treated although balanced on observable variables.

By assuming a logistic distribution, two matching units i and j, having $\mathbf{x}_i = \mathbf{x}_j$, have the following odds ratio:

$$\frac{\frac{p_i}{1-p_i}}{\frac{p_j}{1-p_j}} = \frac{p_i\left(1-p_j\right)}{p_j(1-p_i)} = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta} + \gamma v_i)}{\exp\left(\mathbf{x}_j\boldsymbol{\beta} + \gamma v_j\right)} = \exp\left\{\gamma\left(v_i - v_j\right)\right\} \tag{2.102}$$

showing that, as soon as $v_i \neq v_j$, the two probabilities to be treated are different, actual balancing does not hold and a hidden bias arises. Suppose that v_i and v_j take values in the interval [0; 1] and that $\gamma \geq 0$. This implies that $-1 \leq v_i - v_j \leq 1$, so that the odds ratio is in turn bounded this way:

$$\frac{1}{e^{\gamma}} \leq \frac{p_i\left(1-p_j\right)}{p_j(1-p_i)} \leq e^{\gamma} \tag{2.103}$$

Sensitivity analysis II

Rosenbaum proposes a sensitivity analysis test based on the Wilcoxon's signed rank statistic: The null hypothesis asserts that the treatment is without effect

1. Consider S matched pairs, with $s = 1, . . ., S$, where each pair is formed by one treated and one untreated unit
2. For each pair, calculate the treated-minus-control difference (DIM) in outcomes and call it D_s, thus getting the absolute differences $|D_s|$
3. Eliminate from the sample any absolute difference score taking value zero, thereby yielding a set of S_0 nonzero absolute differences, where S_0 becomes the new sample size.
4. Assign ranks R_s ranging from 1 to S_0 to each $|D_s|$, so that the smallest absolute difference gets rank 1 and the largest one rank S_0. If ties occur, assign the average rank.
5. The Wilcoxon test statistic W is obtained as the sum of the positive ranks: $W = \sum_s R_s^+$

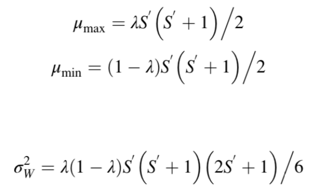
Sensitivity analysis III

The Wilcoxon test statistic W varies from a minimum of 0 -where all the observed differences are negative- to a maximum of $S'(S'+1)/2$ where all the observed difference scores are positive.

1. For a quite large randomized experiment and under the null hypothesis of equality in the two (treated and untreated) populations' medians (i.e., no-effect assumption), the W statistic is approximately normal distributed: $N((S'(S'-1)/4), S'(S'+1)(2S'+1)/24)$
2. If the null hypothesis is true, the test statistic W should take on a value approximately close to its mean $(S'(S'+1)/4)$

Sensitivity analysis IV

Rosenbaum shows that under the null hypothesis of equality in the populations' medians, the distribution of W is approximately bounded between two normal distributions with the following expectations:



$$\mu_{\max} = \lambda S' \left(S' + 1\right) \Big/ 2$$

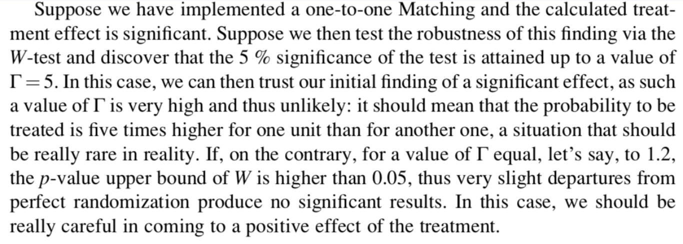
$$\mu_{\min} = (1 - \lambda) S' \left(S' + 1\right) \Big/ 2$$

$$\sigma_W^2 = \lambda(1 - \lambda) S' \left(S' + 1\right)\left(2S' + 1\right) \Big/ 6$$

with $\lambda = \Gamma/(1 + \Gamma)$
in the randomization case $\Gamma = 1$, the two formulas become the same and are equal to the case of randomized experiment

Sensitivity analysis V

For $\Gamma \geq 2$, the p-value is bounded between a minimum and a maximum and one can use the upper bound to see up to which value of Γ the usual 5% significance is maintained in the experiment.



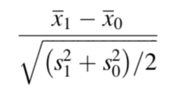
Suppose we have implemented a one-to-one Matching and the calculated treatment effect is significant. Suppose we then test the robustness of this finding via the W-test and discover that the 5 % significance of the test is attained up to a value of $\Gamma = 5$. In this case, we can then trust our initial finding of a significant effect, as such a value of Γ is very high and thus unlikely: it should mean that the probability to be treated is five times higher for one unit than for another one, a situation that should be really rare in reality. If, on the contrary, for a value of Γ equal, let's say, to 1.2, the p-value upper bound of W is higher than 0.05, thus very slight departures from perfect randomization produce no significant results. In this case, we should be really careful in coming to a positive effect of the treatment.

Assessing overlap

A good overlap of treated and control units over the covariates' support is required in order to obtain reliable estimates for ATEs.
How can we assess the goodness of overlap in a given dataset?
Imbens and Rubin suggest three types of overlap measures:

1. standardized difference in averages;
2. logarithm of the ratio of standard deviations
3. Frequency coverage

Assessing overlap II

1. standardized difference in averages;
2. 

$$\frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\left(s_1^2 + s_0^2\right)/2}}$$

3. This measure is scale-free (it does not depend on the unit of measure of x), but it has the limit to refer to a specific moment of the distribution, the average.

Assessing overlap II

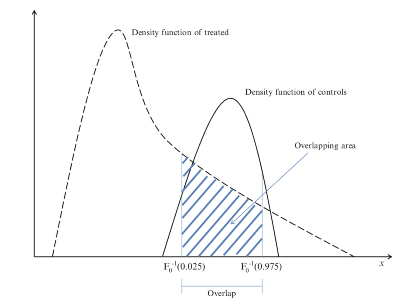
1. Logarithm of the ratio of standard deviations
2. the differences in the dispersion of the treated and control distribution over x, by computing the logarithm of the ratio of standard deviations:
3. $ln(s_1) - ln(s_0)$
4. This approach is straightforward, but it fails to take into account the overall shape of the two distributions

Assessing overlap II

Frequency coverage

1. A more reliable way to assess overlap is that of computing the share of the treated (control) units taking covariate values that are near the center of the distribution of the covariate values of the controls (treated):

2. $$\pi_1^{0.95} = F_1\left\{F_0^{-1}(0.975)\right\} - F_1\left\{F_0^{-1}(0.025)\right\}$$
$$\pi_0^{0.95} = F_0\left\{F_1^{-1}(0.975)\right\} - F_0\left\{F_1^{-1}(0.025)\right\}$$



Coarsened-Exact Matching

The basic idea behind CEM is that of allowing the analyst to choose ex ante the degree of the balancing of covariates, thus avoiding the necessity for its ex post assessment and repeatedly reestimating the propensity-score until balancing is satisfied.

1. discretize continuous variables, as well as reduce the number of values that a discrete covariate can take.
2. Such a procedure, which the authors call coarsening mechanism, enables one to build a tractable number of cells by: crossing all covariates' values, deleting cells that do not contain at least one treated and one control unit, and estimating ATEs on the remaining cells

Coarsened-Exact Matching II

The CEM algorithm is as follows:

1. start with the covariates x and generate a copy, which we indicate by x^c
2. ''coarsen'' x^c according to user-defined cut points (the CEM's automatic binning algorithm can also be exploited)
3. produce cells by crossing all values of x^c and place each observation in its corresponding cell
4. drop any observation whose cell does not contain at least one treated and one control unit
5. estimate ATEs by stratification Matching on the remaining cells (or, equivalently, run a WLS regression of Y on D using the remaining cells' weights).

Coarsened-Exact Matching III

1. increasing degree of coarsening is generally accompanied by higher imbalance in the covariates
2. To assess CEM quality, Iacus et al. (2012) suggest to examine a specific measure of (global) imbalance:
3. $$L_1(f,g) = \frac{1}{2}\sum_{b=1}^{B} |f_b - g_b|$$
4. a value of L_1 equal to zero signals perfect global balance; vice versa, the larger the L_1 is, the larger the extent of imbalance, until reaching a maximum of one which occurs when there is complete separation of treated and control units in each cell

f_b and g_b are the relative frequencies for the treated and control units within cell b.

Coarsened-Exact Matching IV

1. The authors suggest to take the value of L_1 obtained after coarsening (but without trimming) as a benchmark to be compared with the value of L_1 obtained when observations with cells not containing at least one treated and one control unit are dropped (trimming).
2. By calling the first $L_{1,unmacth}$ and the second $L_{1,match}$, we expect that CEM has worked well if:
3. $L_{1,unmatch} > L_{1,match}$
4. (some improvement in balancing occurs)

Bibliography

Cerulli, G. (2015). Econometric evaluation of socio-economic programs. Advanced Studies in Theoretical and Applied Econometrics Series, 49, Springer.