[Discussion Paper 2/2012]

Simple Control Group Analysis by Matching: A Brief Introduction

Rolf Bergs

PLANUNG & FORSCHUNG
POLICY RESEARCH & CONSULTANCY

Planung & Forschung (PRAC), Bad Soden, Germany

Simple control group analysis by matching[1]

What is control group analysis?

A control group analysis is used to estimate impacts in situations where impacts are not directly observable. This is a characteristic problem when an item has been treated from one point of time onwards (by anything like medical treatment, funding, consultancy etc.). In this situation it is only possible to measure a status with treatment; it is not possible to compare it directly with the untreated status. This would necessitate a parallel existence of one item. The problem is solved by an approximation approach where an artificial benchmark of comparison is established to measure impact. This artificial benchmark is the control group. The use of a control group makes this approach quasi-experimental. The major requirement of a control group is a strong similarity (ideally an infinitesimally small difference) of the character of items with the only distinct exception that the control group items have not been treated.

Advantage of control group analysis

With a selection model it is possible to estimate the probability of being treated regardless the real status. This makes it theoretically possible to generate a control group to estimate the counterfactual situation.

The generation of two structurally homogenous groups (one of them treated, one not) for direct comparison leads to the following relationship describing the average treatment effect on the treated (ATT):

$$\Delta Y_{ATT} = E(Y_1 - Y_0|F = 1) = E(Y_1|F = 1) - E(Y_0|F = 1)$$

In this case the difference between the factually treated and the potentially treated - but factually untreated - items is considered to establish an approximation of the treatment impact - directly on the items participating.

[1] Handout belonging to the presentation on "Data requirements for evidence-based evaluation of EU funded interventions" convened by the European Evaluation Helpdesk for Rural Development at Vidékfejlesztési Minisztérium, Budapest, 8-9 October, 2012. The subject of this paper was considered as a facultative presentation in case of sufficient time and interest of participants. A Powerpoint presentation had been prepared for the seminar but not delivered during the session; this handout was produced later.

Limitations of control group analysis

Because of the requirement to generate a sample that is highly similar to the group of treated it is very likely that only a very large sample may contain items that are characteristically close to the group of treated. Otherwise there is the risk of a selection bias, so that in the end one would compare "apples" with "pears". The same problem would arise if the predictors to specify a control group are inadequate. Hence, in many cases one would need a larger number of explanatory variables (for that the corresponding data were needed) in addition to a sufficiently large sample.

Further to that the selection model chosen by the evaluator is based on the evaluator's subjective expectations and hypotheses, thus, it is not unequivocal as alternative specifications may exist. Results of the several alternative specifications might differ from the former one specified.

The most important caveat is the risk of biased results due to an inadequate control group (violation of the unconfoundedness assumption). This is always the case if there are significant differences in variance among the two groups (mean comparison test). If control groups are just established in taking a random sample or by simple visual inspection and selection of items from the universe, there is no sufficient control over adequacy. But even if control groups are generated statistically (i.e. through a logistic regression procedure as shown below) the difference in variance cannot always be reduced so that it is possible to match items.

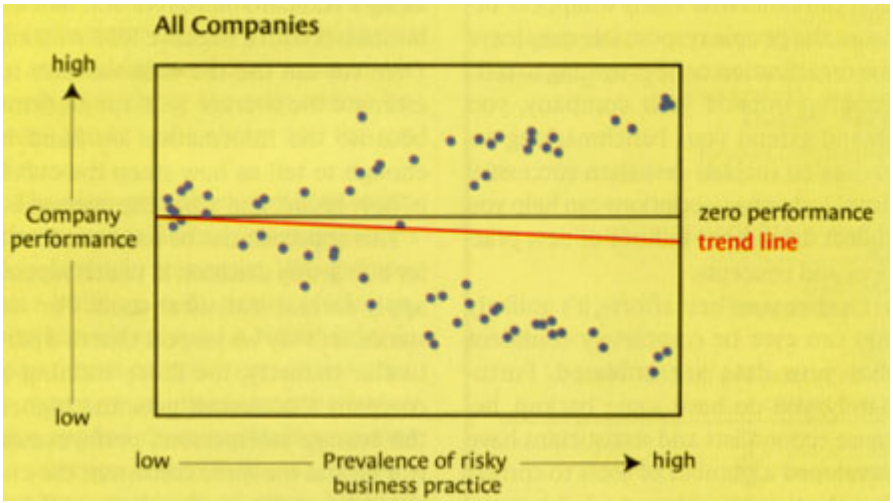
In many cases –especially those where the items exhibit a stong level of singularlity[2] - there is simply no way to generate an adequate control group. The larger the statistical universe the more likely will be the generation of an adequate control group, but a large universe alone still does not at all ensure that.

In many other cases there are simply no or only inadequate data to specify a selection model.

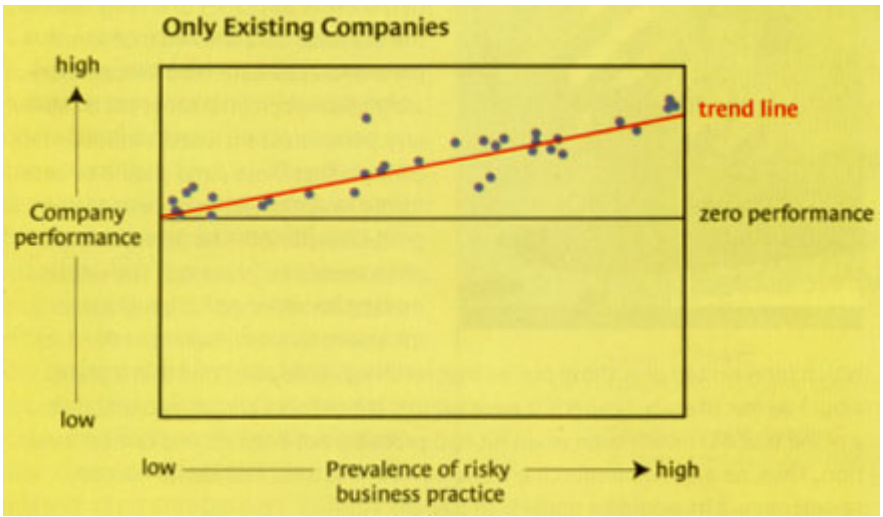
[2] in the Langenscheid German-English dictionary: a synonym for singularity is matchlessness, a perfectly fitting term to describe the problem addressed here (E Klatt and D. Roy, *Langenscheid's Pocket Dictionary on the English and German Languages*, sixth edition, Berlin and Munich, 1970)

The impact estimate with inadequate control group

To understand the problem of selection bias it is useful to demonstrate it by an example. The following two figures describe the relationship between the prevalence of risky business practice and the performance of the companies. If one looks at the universe of all companies – also those that already ceased to exist – there is a negative relationship between the two variables (trend line), hence, the riskier a company management the lower the performance appears on average.



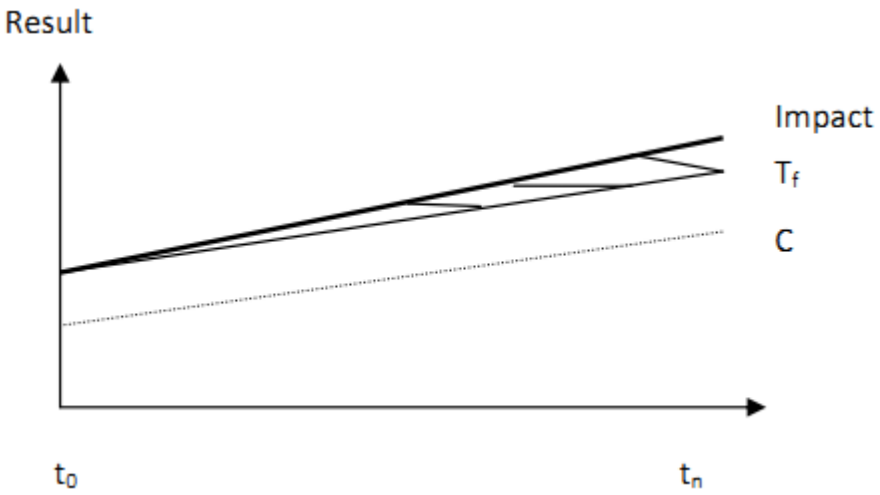
If one then just views the sample of existing companies we observe a completely different trend. For the existing companies it seems that riskiness is positively correlated with company success. Hence, omitting the formerly existing companies would lead to biased conclusions, because between the groups of existing and formerly existing companies there are systematic differences in the predictors.



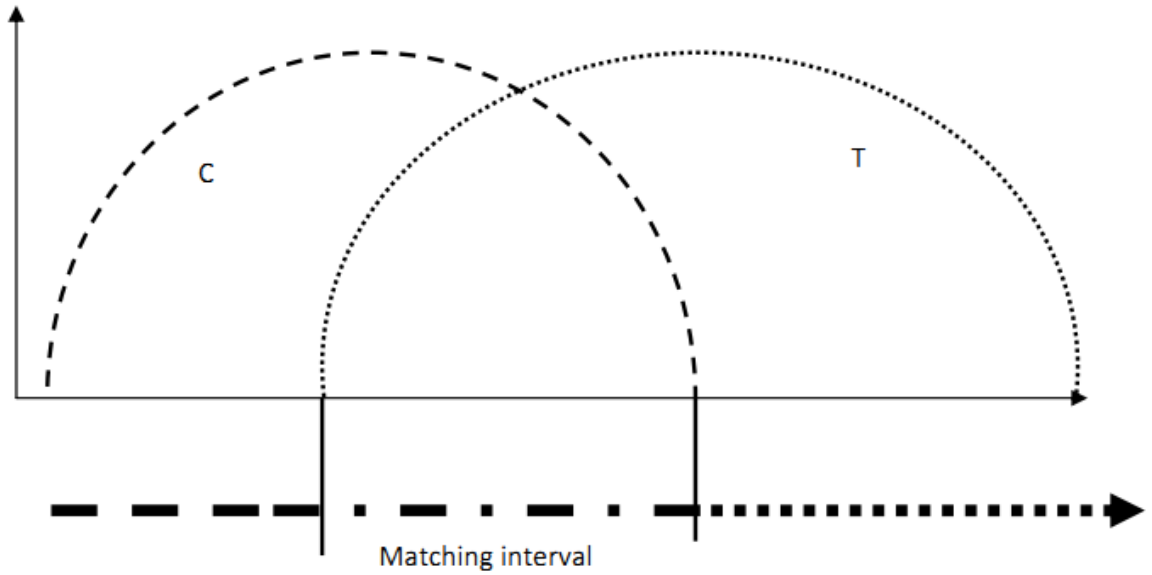
Source of the graphs: http://mahalanobis.twoday.net/stories/682756/

The impact estimate with an adequate control group

If the control group is correctly specified, i.e. if determining variables of treated items and the control group are significantly similar in their variance, a prior performance forecast (T_f) of the group of treated will be identical with the control group (C) or at least parallel. The impact is then the margin between T_f and the real outcome of the group of treated at the point of time t_n.



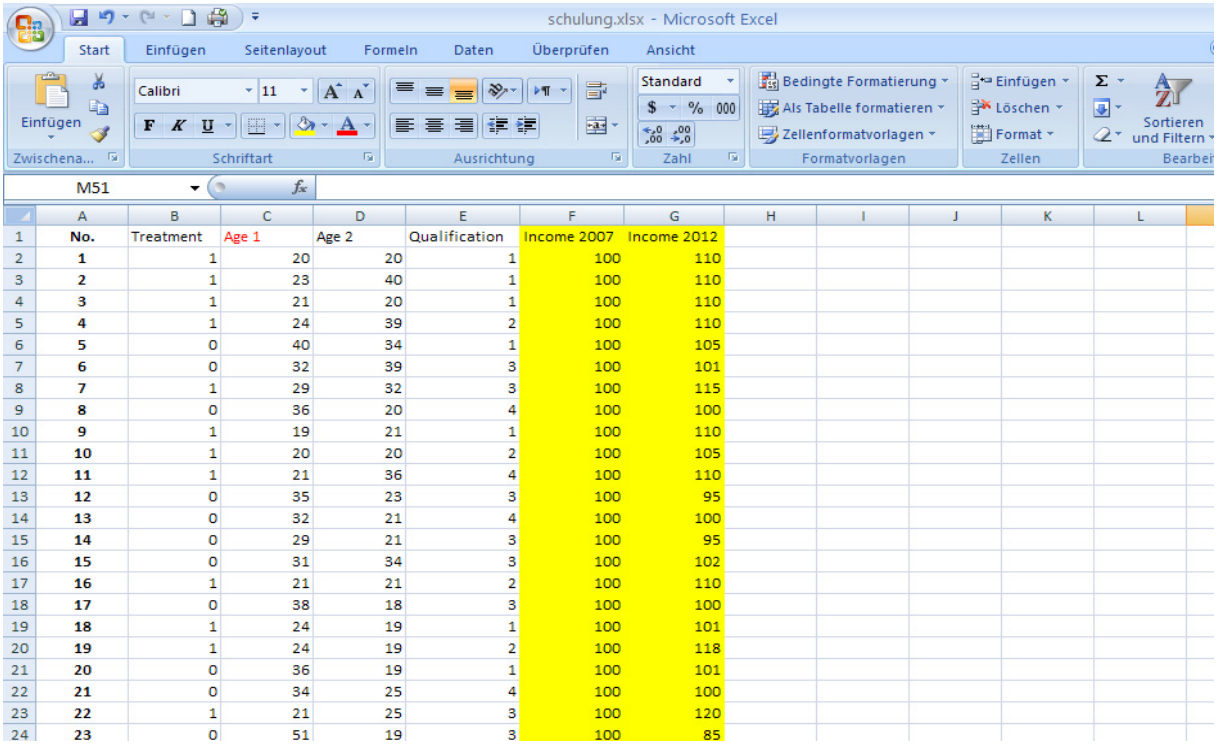
The selection of the control group will be taken from the interval of common support where the distributions of treated and untreated items overlap:



Outside the overlapping areas of C and T, selection of matches will be biased, such as matching existing with bankrupt companies.

A simple case study: Measure 111 RDP programme „Somewhere“

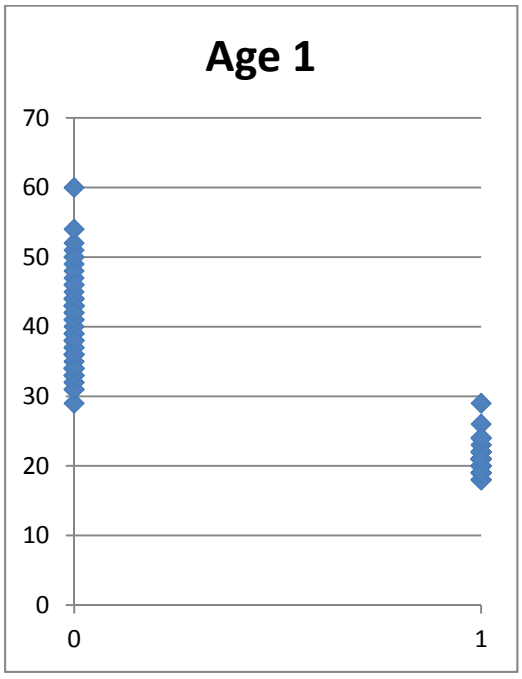
The following figure shows the snapshot of a database of farmers having received training under EAFRD measure 111 and a prospective control group. Treatment is labeled with 1, the non-treatment with 0 (column B). “Age” and “Qualification” are two candidates for predictor variables, and “Income” is the performance variable selected.



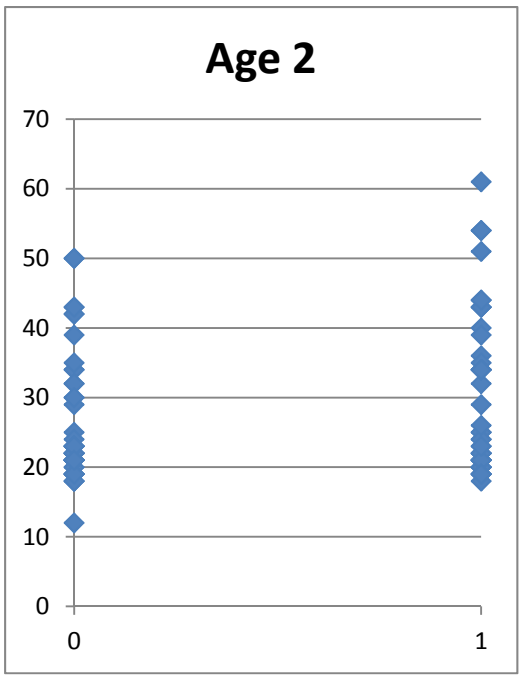
	A	B	C	D	E	F	G	H	I	J	K	L
1	No.	Treatment	Age 1	Age 2	Qualification	Income 2007	Income 2012					
2	1	1	20	20	1	100	110					
3	2	1	23	40	1	100	110					
4	3	1	21	20	1	100	110					
5	4	1	24	39	2	100	110					
6	5	0	40	34	1	100	105					
7	6	0	32	39	3	100	101					
8	7	1	29	32	3	100	115					
9	8	0	36	20	4	100	100					
10	9	1	19	21	1	100	110					
11	10	1	20	20	2	100	105					
12	11	1	21	36	4	100	110					
13	12	0	35	23	3	100	95					
14	13	0	32	21	4	100	100					
15	14	0	29	21	3	100	95					
16	15	0	31	34	3	100	102					
17	16	1	21	21	2	100	110					
18	17	0	38	18	3	100	100					
19	18	1	24	19	1	100	101					
20	19	1	24	19	2	100	118					
21	20	0	36	19	1	100	101					
22	21	0	34	25	4	100	100					
23	22	1	21	25	3	100	120					
24	23	0	51	19	3	100	85					

To demonstrate confoundedness vs. unconfoundedness, two alternative distributions of the variable “Age” are compared in the following two figures:

Descriptive assessment I: Bias



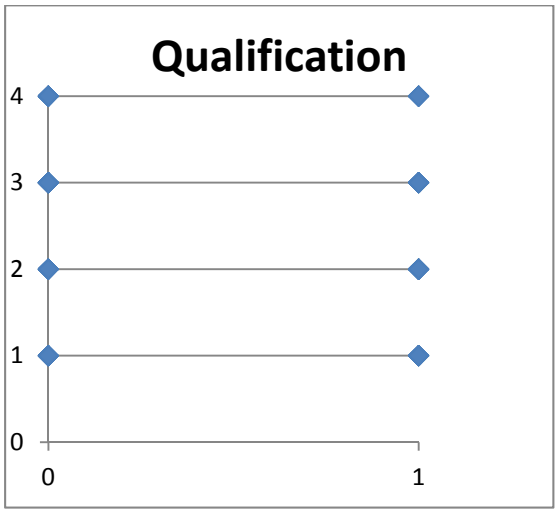
Descriptive assessment II: Unconfoundedness



While the comparison of distribution of Age1 among treated and untreated items shows a strong difference (without overlap), the distribution of Age2 is similar among the groups of treated and untreated. Visual inspection (descriptive analysis) suggests potential adequacy of Age2 as a predictor variable to generate a control group.

The second variable “Qualification” exhibits an identical distribution and thus also qualifies for a predictor variable – at a first glimpse.

Descriptive assessment III: Unconfoundedness



A statistical procedure to test the selection bias in the samples is to carry out a mean comparison test. The aim of the subsequent selection model (matching) is then to reduce the differences and to find a control group within a matching interval.

Propensity score matching

Propensity score matching is based on a Logistic or Probit regression analysis to estimate the probability of being treated regardless the status. Modern statistics software, such as SPSS, Stata, SAS or several others allow to estimate such logistic procedures.

With the procedure of a logit regression Y[yes/no]=f($x_1 \ldots x_n$) one will estimate the coefficients $b_1 \ldots b_n$. These determine the probability of yes or no respectively as an odd ratio.

The logit regression model is specified as:

$$Pr(Y=1 \mid x_1 \ldots x_n) = 1/(1+1/e^{a+b_1 x_1 + b_i x_i + b_n x_n})$$

If the researcher uses Stata for example, it is possible to choose the logit procedure from the pull-down menue. The variables entering the model have to be ticked. In

order to estimate the probabilities of being “yes” or “no” or “treated” or “not” (i.e. the dependent variable of the model) the user may type the Stata command

```
predict Y_prob
```

and then

```
browse.
```

This will generate the probability estimates for all observations.

The result will reveal important information about the adequacy of the control group selected. On the one hand it will show whether there is sufficiently large group of untreated that are potentially treated. On the other hand it helps to decide whether the choice of predictors is adequate. If there is for instance a very large share of treated items being potentially untreated it may indicate that there is no significant explanatory power of the predictor variables (or some of them).

Some introductory exercises with Stata

Stata allows a direct procedure of propensity score matching.

Excel spreadsheets can be processed by Stata. If one reads-in the Excel database on Measure 111 of the RDP programme “Somewhere” (see above) into Stata the following commands are necessary for a simple propensity score analysis.

Direct propensity score matching

Step 1: Start Stata

Step 2 (if not yet installed): download „psmatch2“ with the following command:
`ssc install psmatch2`

Step 3: Load the Excel database into Stata (via the Stata menu)

Step 4: View the descriptive statistics shown

Step 5: Type the command:
`psmatch2 Treatment Age2 Qualification`

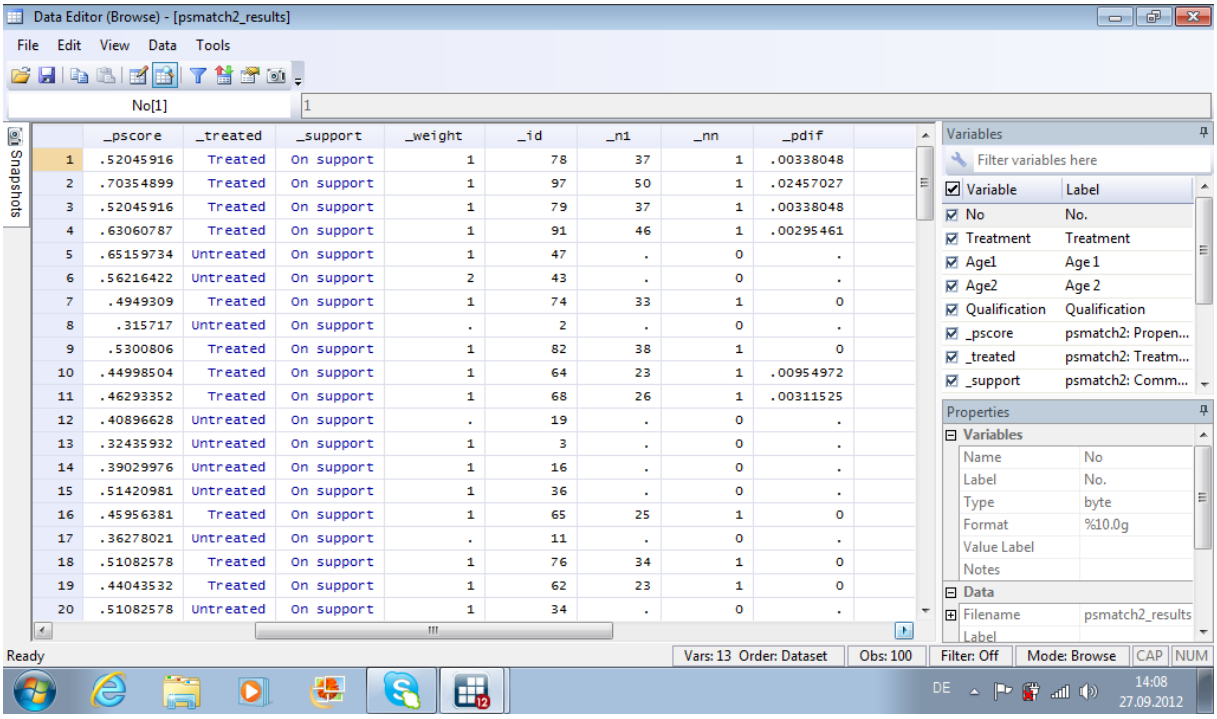
This command generates probit regression results. To view the results of the matching procedure a last command is necessary:

Step 6: Type:

```
browse
```

This step generates further columns including propensity scores and the treated/untreated pairs by ID numbers as shown below.

Matching results



Assessment of results

The resulting database indicates the optimum selection of matches by ID numbers. E.g. farmer 1 (ID-No 78) and farmer 3 (ID-No 79) can be both matched with farmer 37. Both show an identical probability difference.

In the subsequent steps it is possible to compare the performance variable(s) of the pairs of treated untreated at t_o and t_n. In this case the performance variable is "income". To get an estimate of the impact of Measure 111 on income the means of income at t_0 and t_n are compared for both groups. The comparison can be illustrated by the percentage change of the means for both groups. The margin in percentage points is then to be interpreted as the policy impact.

Instead of just viewing the average margin it is reasonable to extend the approach by using the individual margins for a subsequent difference-in-differences (DiD) regression analysis to statistically confirm the margin (policy impact) established by group comparison.

Apart from propensity score matching and DiD there are further quasi-experimental standard methods not covered in this paper, e.g. instrumental variables or regression discontinuity design. All of them have their *pros* and *cons* when deciding how the impact of policy should be estimated.

Rolf Bergs[3]

27 November, 2012

[3] PRAC, Im Hopfengarten 19b, D-65812 Bad Soden, Germany

References:

Heckman, J.J. (2008), “Econometric Causality”. *International Statistical Review*, Vol. 76

Heckman, J., H. Ichimura, and P. Todd (1997), “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme”, *Review of Economic Studies*, Vol. 64

Imbens, G.W. and J. M. Wooldridge (2009), “Recent Developments in the Econometrics of Program Evaluation”, *Journal of Economic Literature*, Vol. 47

Pearl, J. (2009), *Causal inference in statistics: An overview*, Technical Paper, Computer Science Department, University of California, Los Angeles

Roy, A. (1951), “Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, 3

Rubin, D. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”, *Journal of Educational Psychology*, Vol. 66