Republika e Kosovës
Republika Kosova-Republic of Kosovo
Qeveria –Vlada-Government
Ministria e Tregtisë dhe Industrisë-Ministarstvo Trgovine i Industrije-
Ministry of Trade and Industry

Guideline on evaluating the impact of policies
-Quantitative approach-

Gratitude and Appreciations

Compilation of this guideline is funded by the Ministry for Foreign Affairs of Finland, in the framework of "Aid for Trade" project, implemented by the United Nations Development Program (UNDP).

Quality Assurance, United Nations Development Program (UNDP) in Kosovo:

"Aid for Trade" project
Artane Rizvanolli, External Consultant
Policy, Research, Gender and Communications team, UNDP

Author:

Alban Hashani

The original version is written in the Albanian language

The guideline is prepared as a response to the needs of participants of the training on industrial policies and economic analysis, and views expressed in the material belong exclusively to author.

Contents:

Introduction

Most policies are designed to achieve a particular objective, for example for improving business enterprises that are subject to **treatment**[1]. Furthermore, the process of designing foresees the expected outcomes of a certain policy. However, to assess whether a policy has achieved the expected effects, the impact of intervention must be taken into account. This is not an easy process, and remains a key concern of researchers and analysts. Such a thing becomes difficult in particular when analyzing the policy to be implemented in an environment that is subject to many reforms at once.

Over the recent decades, various approaches have been developed in evaluation of policies. Furthermore, now it dominates the opinion that evaluation of the impact of policies should be an integral part of policy making in order to enable decision-making based on evidence. However, although research methods for this purpose are advanced, and the level of data available has increased, still these studies face **major challenges during implementation**. Firstly, because in most cases, **the impact includes long-term changes** and it may take months or years that such changes to be visible. Secondly, **the effects of the impact of policies can be confused with the effects of other changes** that have occurred in the meantime and that attribution of impact to policy itself is not straightforward. However, beyond these challenges, many analysts now manage to successfully evaluate the impact of policies.

This guideline provides a brief summary of quantitative methods of evaluation of the impact of policies. In this guideline, the analytical framework and basic terminology of econometrics of policy impact assessment was initially introduced, and is afterwards used in the practical part. This guideline is designed based on the general literature of the policy impact assessment. Throughout the material, an effort was made to bring to light the essential points, by highlighting them in bold letters. The material is divided in three parts: **Part I,** focuses on the theoretical and conceptual aspect of evaluation of policies. In particular, it treats the problem and the method of evaluation. **Part II**, focuses on quantitative methods of evaluation of the impact of policies. **Part III**, is the part that

[1] The term "treatment" derives from the medical sciences and has more meaning when is used in that context. However, this term means any intervention that is as a result of certain policy. This term is used throughout this guideline in order to be in the same line with the dominant literature in this field.

practically guides the readers on how they should implement quantitative methods to evaluate the impact of policies, treated in Part II.

Targeted Audience

This guideline is dedicated for young analysts who deal with evaluation of the impact of policies. The same is an information source that helps the analysts to develop a structured approach in the evaluation process. This guideline is particularly important for analysts who are beginners in this field, but its content will be valuable for all those who like to expand the understanding of the evaluating process of the impact of policies. To get the most from this guideline, readers are advised to consider other materials that treat, in a more detailed way, various parts of the evaluation of the impact of policies.

Part I: Theoretical treatment

- The evaluation problem

The evaluation of the impact, in the current context, attempts to calculate changes in the outcomes that can be attributed to a certain policy. The relationship **cause-effect** between policy and outcome, which is the focus of this analysis, considered as the most essential characteristic and the biggest challenge of evaluation of the impact. Analyzing the impact of a certain policy requires evaluation of the outcome to be observed in the absence of policy; so, what would happen if a certain policy does not exist? If we mark with Y_0 the result that units would have if they are not subject of treatment and with Y_1 the result that units would have if they are subject of treatment, then the impact of policy would be:

$$\Delta = Y_1 - Y_0$$

If these data would be available, then the case-effect impact of a certain policy could easily be evaluated. However, as long as **only one of Y was observed** for each unit, (i.e. either Y_1 or Y_0), the difference or the impact of policy (Δ), cannot be evaluated directly for units that are subject of treatment. This problem of lack of data is the core of evaluation of the impact of policies. Consequently, the methodologies that attempt to make an estimation of the impact tend to evaluate these missing data. In lack of such data (counterfactual), analysts tend to compare the outcome of units that were subject to intervention with a **group that is comparatively similar** (control group). Concretely, how the control group is selected is of a great importance. In general, there are **two types of control groups** (surrogate counterfactuals): **(a) comparison of the outcome of the units that were subject to treatment with the outcome of those units that were not subject to treatment, and (b) comparison of outcome of the units that were subject to treatment before and after the treatment.**

Comparison of the result of the units that were part of treatment and those that were not, in literature are usually referred as **with and without comparison**. Such techniques require additional efforts to ensure that the units that we compare between them are not systematically different. For example, we assume a policy that, through subsidies, attempts to support enterprises of a particular sector. If the selection is not random, and if, in order to become beneficiary of subsidies, only the best firms of the sector qualify, then other

remaining excluded enterprises do not represent a good comparative group. If not checked for these systematic differences between unit groups, then the outcomes can be diverted. In a similar situation, evaluation of the resulting difference of two unit groups, not necessarily is the result of policy; in fact, the same represents the fundamental differences that exists between two unit groups, regardless of policy.

The comparison of outcome of the units that were subject to treatment, before and after the treatment, represents another technique in determining the control group. In literature the evaluation of policies, is a technique referred to as **before-and-after** comparison. This technique attempts to evaluate the impact of policy by following changes of different indicators over time. Control group in this case is the situation before implementation of the policy. For example, we assume similarly a policy that, through subsidies, aims to support enterprises in a particular sector. This technique assumes that the business of an enterprise, which has not benefited from subsidies, after the end of the policy, will be the same with business before the implementation of the policy. However, we need to keep in mind that there are many factors, in addition to the policy itself, which may change over time. For example, if during the period the policy of subsidy is under implementation, there is a further change in tax rates, and the result of the subsidy policy may be confused with the effect of tax changes. Ignoring other factors might cause deviation of the results, unfairly attributed to differences in the policy itself. In such a situation, the results may be over- or underestimated.

However, analysts can combine the two techniques to improve the evaluation. A combination of the method "with and without" and "before and after" improves significantly the accuracy of evaluation. Such combined technique is explained in this guideline.

- ***The evaluation method***

There are different approaches in evaluation of policies. Two methods can be used, depending on the data that are used: (a) **experimental** and (b) **non-experimental.** It is important to note that both approaches assume that indirect effects are not significant (in other words, the influence of politics in a unit does not affect other units and that the impact of policy in the treated unit does not affect the untreated units). For example, continuing the analogy of the subsidy policy, if an enterprise benefits from subsidies, the impact of subsidies on this enterprise is not significant for other enterprises that have received

subsidies or even to those enterprises that have not benefited from subsidies. While this is a strong assumption, in most literature is accepted as such, as direct effects of the policies in most cases are much larger than the indirect impacts, and consequently the last ones can be ignored (Imbens and Wooldridge, 2009).

The experimental methods are generally considered as the best evaluation methods. In the evaluation of the impact of policies the same are usually referred as, **the golden standard**, because they eliminate the evaluation problem. That is because the same rely on the data of the control group that is a subgroup randomly selected by the population. If strictly applied, experimental methods identify the control group, which is identical based on the observed and unobserved characteristics with the treated group (Blundell and Dias, 2002). The only difference between these groups would be, being subject to the policy, the effect of which we want to evaluate. Consequently, **the evaluation problem eliminated** because the choice to be part of the training happens randomly (Bryson et al., 2002).

However, experimental data are usually very rare in the field of economy; this is because. in most of the cases it is impossible to isolate that certain units from the right to be subject to treatment. Furthermore, due to the nature of phenomena treated in the field of economy, the likelihood of having an experiment where participation is determined randomly, is minimal. Consequently, most of the researches that evaluate the impact of policies rely on **non-experimental methods**. Non-experimental studies try to imitate the experimental methods. In the following we will treat two of these methods: **Compliance based on Propensity Score Matching (PSM)** and **Difference in Difference (DID)**. In addition, in this guideline we shall treat the combination between these two methods.

Part II: Quantitative methods of evaluation of the impact of policies

a) Compliance based on Propensity Score Matching (PSM)

Compliance based on Propensity Score Matching (hereinafter, PSM), is a statistical method whereby **the units that are subject to treatment comply with one or more units which are not subject to treatment** (by the control group) based on propensity score matching to be subject to treatment. So, the results of the two groups of units that are comparably similar, are compared. This method evaluates the impact cause-effect, which can be attributed to treatment. The PSM can be applied in almost every context of evaluation of policies, as long as there is a treated and untreated group that can serve as an appropriate reference point. It **relies on the observed characteristics** to build a comparative group. The validity of this method depends on two basic assumptions. Firstly, the method assumes that **factors that cannot be observed did not affect the participation in treatment**. Secondly, the method assumes that there is a **sufficient overlap or common support** throughout the sample of the treated and untreated group.

Finding a suitable unit of the untreated group for each unit that is subject to treatment requires to approximate, as much as possible, all the characteristics between the two groups of units. The more number of features we want to approximate increases, the chances of finding a compatible relative unit get slimmer. We commonly refer to this phenomenon as "**curse of dimensionality**" and its effect increases exponentially with the increase of dimensions.

The "curse of dimensionality" is present in any case where the data has more dimensions, because in such cases, the data become very rare and distant to each other. Consequently, the number of compliant units decreases, which may lead to erroneous evaluations. For example, if someone wants to make compliance based on two dimensions (say, activity and size of enterprise), then it might be easier to find approximate units from the group of units that were not subject to treatment (control group). However, if the number of characteristics, based on which we want to perform the compliance, increases (say, to include exporting activity, the location of the enterprise), then the chances of finding similar units from control group are significantly reduced.

Table 1 is an example of compliance of treated units with those of the control group, based on four characteristics (location/region, export, activity and size of the enterprise). As can be seen in the table, for most of the treated units cannot be found units that fully comply

with the untreated units (only cells with the same shade represent exact compliance between the treated group and the control group). If the number of criteria increases, then **matching becomes even more difficult**. Therefore, in the absence of sufficient data of control group, the evaluation process becomes difficult.

Table 1. Compliance based on different characteristics

Treated group				Control group			
Region (1–capital city; 0–other)	**Export activity (from 1–5)**	**The activity (NACE)**	**The size (Small - S; Medium - M; Big – B)**	**Region (1–capital city; 0–other)**	**Export activity (from 1–5)**	**The activity (NACE)**	**The size (Small - S; Medium - M; Big – B)**
1	3	B	B	1	2	H	B
0	2	C	M	0	1	L	S
0	5	D	B	0	4	F	M
0	1	D	S	0	3	N	M
1	4	D	M	0	5	D	B
1	2	H	B	0	1	E	S
0	1	L	S	0	1	L	S

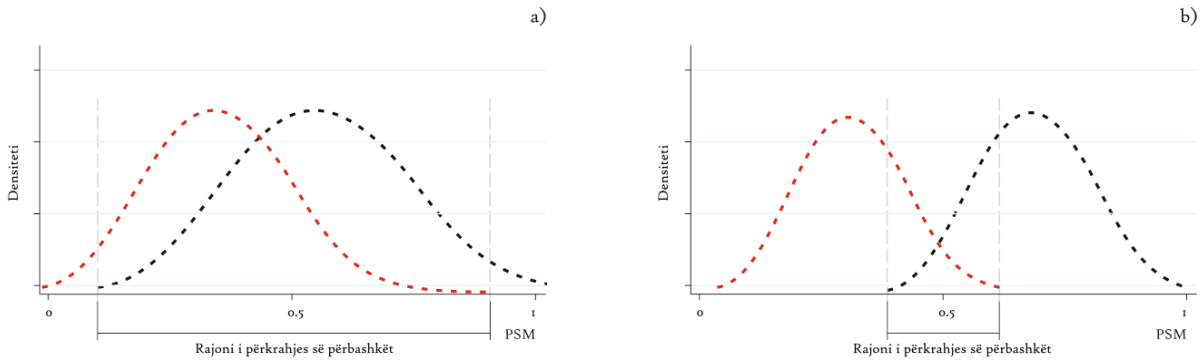
Source: Author
Remark: the cells with the same shade represent exact compliance between the treated group and the control group.

The "curse of dimensionality" can be avoided by using **Propensity Score Matching (PSM),** proposed by Rosenbaum and Rubin (1983). The propensity score represents the likelihood of being subject to treatment. This method is not intended to match the units of the two groups based on all characteristics. Instead, for each unit that is the subject of treatment and for each unit that is not subject to treatment, chances are that the unit might be/was subject to treatment – **propensity score.** This way, the units that are subject to treatment and those that are not subject to treatment, can comply based on the propensity score, creating this way an approximate control group. The likelihood of being subject to treatment – propensity score – is based on the data that can be observed. Consequently, the propensity score simplifies the compliance by **decreasing the dimensions of**

characteristics to a single value. So the impact of all factors based on which we want to make compliance (such as location, export, activity and the size of the enterprise), we transform into a single value that represents the likelihood of being subject to the policy. This value is then used for units' conformity of the two groups. Once the data comply, the impact of the policy can be evaluated easily.

PSM relies on the data that can be observed, and the same initially assumes that participation in treatment is not influenced by factors that cannot be observed. If such assumption cannot be taken for granted, then this method is not appropriate to use. Nevertheless, this assumption cannot usually be tested. However, when combined with other methods, the risk that can divert the results is significantly minimized. The **PSM also assumes that there is sufficient overlap or mutual support** throughout the sample of the treated group and of the untreated one. In other words, it is assumed that for the units that were subject to treatment there are approximate comparable units in the distribution of the propensity score (Heckman et al., 1999; Blundell and Dias, 2002). The major concern in the PSM's application is the lack of common support (Imbens and Wooldridge, 2009). Figure 1, shows two cases where common support is (a) good and (b) poor. Finding a significant region of common support requires a large number of data. Consequently, the treated units should be similar to untreated units based on observable characteristics. In the absence of adequate regional of common support, the PSM cannot be used for evaluation of the impact of policies. The existence of a common support can be tested.

Figure 1. The region of common support. Examples when there is (a) good balancing and sufficient region of common support, and (b) poor balancing and low support of the common support.



Source: The author using various sources

Remark: The Figure represents independently distribution of the propensity score for the untreated group (threaded red line) and the treated group (threaded black line).

Based on the propensity score, **different parametric algorithms** can be used to match the units from the treated and untreated group. These algorithms include **nearest neighbor matching**; **radius matching** and **satisfaction matching**. Matching of the nearest neighbors finds the nearest unit from the control group, based on the propensity score. Matching of the nearest neighbors may result in poor matching if the nearest "neighbor" is very far from the treated unit. In such cases, matching based on the distribution range can provide solution. By establishing a certain level of tolerance, within which matching can be done, avoids the problem of matching the units that are far from each other. On the other hand, matching based on stratification separates the distribution of propensity score in different intervals (strata) and ensures that in each interval the propensity score of the treated units and untreated units is the same.

b) Difference in Difference - DID

The following section summarizes the method Difference in Difference (DID). DID initially used by Ashenfelter (1978), and Ashenfelter and Card (1985), has become the standard method of the evaluation of policies. DID method compares the changes over time in the outcomes of the treated and untreated units. In a standard case, available are data for at minimum **two time points**: one before and one after the implementation of policy of the treated and untreated group. These time points are marked as T$\in \{0,1\}$. Period 0 refers to the period before treatment, whereas period 1 refers to the period after treatment. The treated group exposes to the treatment in the period 1 and not in period 0. Control group is not exposed to treatment at any period. Assuming that After T (ATP) is the outcome after the treatment of the treated units (untreated) and Before T (BTP) is the outcome before treatment of the treated units (untreated); the difference After T – Before T (ATP – BTP) represents comparison before and after for the treated group (untreated group). The expression **(After T – Before T) - (ATP – BTP)** expresses the difference in difference, after insolating the effect of other factors, except treatment factor. This expression eliminates the aggregate factors, to which both groups are exposed in similarly. Table 2 formally shows this expression.

Table 2. Difference in Difference

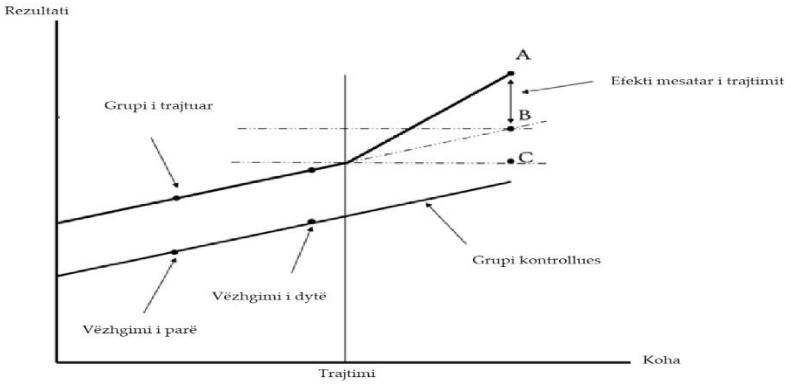
	Before	After	Difference
Treated group	Before T	After T	After T– Before T
Untreated group	Before TP	After TPT	After TP – Before TP
Difference in Difference (DID)			DID=(After T - Before T) - (After TP - Before TP)

Source: The author based on various sources

The DID method requires repeated data for the units that are subject of study. Availability of data, before and after intervention, for both groups of units is the main advantage of this method. The DID assumes that **external factors have similar effect in both groups**. The same allows the elimination of un-observed differences, which are supposedly constant over time. In this method, both groups of units (treated and untreated) not necessary should have same levels before the implementation of policy. **It is enough if the same follows similar trend** of development, which would enable the isolation of the effect of policy.

Figure 2 Represents the DID method. The treated group undergoes through treatment, whereas the control group does not. In the absence of control group, the distance A – C would be considered as an effect of treatment. Such evaluation would be based on the assumption that all pre-post difference is attributed to the treatment. However, if we have the control group, the effect of treatment reduces to distance A-B, since the distance B-C eliminates from the control group.

Figure 2. Graphic presentation of the difference in difference method



Source: The author using various sources

c) Matched Difference in Difference

Simple comparison of the result before and after policy implementation does not evaluate properly the impact of the policy itself. This because in the meantime many factors that can affect the outcome of the units that were subject to treatment change. Similarly, comparison of the units, that are subject to treatment, with units that are not subject to treatment, may divert the results, if both groups are systematically different from each other, and if no attempt is made to control for such a thing (as in the above example, where the best enterprises were supported by subsidies and the remaining enterprises, which have systematically poorer results, were used as comparative units).

However, combination of these two methods – namely, the comparison of the units that were subject to treatment before and after treatment, with units that were not subject to treatment before and after treatment – **significantly improves the results of the evaluation**. This is because the external factors that are constant eliminate by the presence of the control group, while it is assumed that **both groups are exposed to the approximate surrounding**. However, DID is unable to treat the problem of possible deviation if there are major differences between the treated and untreated group. In order to eliminate a potential difference, it is suggested to use initially the PSM, as a technique to approximate the group of treated and untreated units based on the propensity score. The PSM, as discussed above, is widely used as a methodology in itself, however, if such data is available, its combination with the DID significantly improves the results of the evaluation and controls for most of the problems that these two methods individually have (Blundell and Dias, 2002; Smith and Todd, 2005).

Part III: Practical part

Following are shown practical examples of the evaluation of policies; by using method (i) **PSM**, (ii) **DID** and combined (iii) **PSM with DID**. These examples conducted using STATA software as the research tool. **STATA** is a statistical software package that among others, can accomplish a large number of statistical and econometrical procedures of evaluation of the impact of policies. Consequently, the users recommend repeating the basic skills in using STATA in order to accomplish successfully practical examples that are included in this guideline. However, the guideline provides the **detailed syntax** (do files) of procedures to be followed in order to make an evaluation of the impact of policies. The data used for this purpose, are simulated by the author and reflects, hypothetically, the situation where the government has designed a grant scheme for enterprises, in order to encourage investment in innovation. The data can be downloaded from the following link (you must be connected to the internet):
https://www.dropbox.com/sh/3qsononynsemehn/AACTIj8NX5shMIC6KCd1zex2a?dl=0.
It is recommended that readers must download in their computers the whole content of the file, in order to follow the following steps. There you can download another document entitled "Variables.pdf" which describes variables used in these databases.

a) PSM

As discussed above, through this method it is intended to match the units that are subject to treatment and those that are not subject to treatment, and then to calculate the average difference in the results of the treated and the untreated group. The following exercise illustrates on how PSM, should be done in STATA. The command to evaluate the PSM in STATA is "pscore.ado" developed by Becker and Ichino (2002). If your STATA program has not been used previously for this purpose, you should install this package of commands, by pressing "findit pscore" in the command bar of STATA (have in mind that you must be connected to the internet in order to be able to install such package). The command "pscore" evaluates the propensity score, which represents the likelihood that certain unit shall be subject to treatment. This command also tests if there is a sufficient space of common support. If there is a sufficient common support, then other commands can be used to evaluate **the average treatment effect**, by using various matching algorithms (as discussed in the above sections).

Step 1:

Note: Use "PSM1.dta" data from the abovementioned file. Also, use the syntax "PSM_do_file.do" which is in the same location. The description of the variables is shown in the above-mentioned document.

The command to determine the propensity score and if there is a sufficient common support is as follows (Box 1):

Box 1.

```
pscore  trajtimi2011  region production services, pscore(ps) blockid(blockf1) comsup
level(0.001)
```

Although based on the data, there are observations for other periods as well, we will limit only to the latest data, as we could have in our disposition in cases when we must use the PSM as the evaluating strategy. The result (Box 2) includes data from the probit regression; estimation of the propensity score; number of blocks and stratifications, by using the propensity score; and the test regarding the common support.

Box 2.

```
************************************************
Algorithm to estimate the propensity score
************************************************


The treatment is trajtimi2011

trajtimi2011 |
          1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        600       81.86       81.86
          1 |        133       18.14      100.00
------------+-----------------------------------
      Total |        733      100.00




Estimation of the propensity score

Iteration 0:   log likelihood = -347.13357
Iteration 1:   log likelihood = -318.36806
Iteration 2:   log likelihood = -317.91701
Iteration 3:   log likelihood = -317.91638

Probit regression                                 Number of obs   =        733
                                                  LR chi2(3)      =      58.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -317.91638                       Pseudo R2       =     0.0842

------------------------------------------------------------------------------
trajtimi2011 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      region |  -.4696296   .1405443    -3.34   0.001    -.7450913   -.1941679
  production |   .9337271   .1484859     6.29   0.000     .6427001    1.224754
    services |   .3637375   .1427296     2.55   0.011     .0839926    .6434823
       _cons |  -1.233025   .1148712   -10.73   0.000    -1.458169   -1.007882
------------------------------------------------------------------------------
```

```
Note: the common support option has been selected
The region of common support is [.04431635, .38235632]



Description of the estimated propensity score
in region of common support

                Estimated propensity score
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0443164       .0443164
 5%     .0443164       .0443164
10%     .0902988       .0443164       Obs                 733
25%     .1087832       .0443164       Sum of Wgt.         733

50%     .1923449                      Mean           .1813212
                        Largest       Std. Dev.      .1112842
75%     .1923449       .3823563
90%     .3823563       .3823563       Variance       .0123842
95%     .3823563       .3823563       Skewness       .8313796
99%     .3823563       .3823563       Kurtosis       2.435619




*****************************************************
Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output
*****************************************************


The final number of blocks is 3

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks



***********************************************************
Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output
***********************************************************


The balancing property is satisfied


This table shows the inferior bound, the number of treated
and the number of controls for each block

  Inferior |
  of block |     trajtimi2011
of pscore  |         0          1 |     Total
-----------+----------------------+----------
  .0443164 |       165         11 |       176
        .1 |       321         60 |       381
        .2 |       114         62 |       176
-----------+----------------------+----------
     Total |       600        133 |       733

Note: the common support option has been selected


****************************************
End of the algorithm to estimate the pscore
****************************************
```

Results (Box 2), show that identified region of common support is between [0.044 and 0.382] and the number of blocks is 3. The assumption regarding the **sufficient common support**

stands (The balancing property is satisfied). If this assumption does not stand, then the model used should be re-specified.

Step 2:

Once we have evaluated the propensity score and if there is a sufficient common support, we can then evaluate the **average treatment effect** using various algorithms of compliance, as discussed in the sections above.

The average treatment effect using the nearest neighbor matching

The command to evaluate the average treatment effect using the nearest neighbor matching is 'attnd'. In the current example, evaluation of the average effect, by using this command as follows (Box 3):

Box 3.

```
attnd qarkullimi2011 trajtimi2011 , pscore (ps) comsup
```

Box 4.

```
ATT estimation with Nearest Neighbour Matching method
(random draw version)
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT    Std. Err.          t
------------------------------------------------------------

      133         600    8264.538     1840.036      4.492

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches
```

As seen from the results (Box 4), the policy seems to have had a positive impact on the average turnover of enterprises. On average, enterprises that have been treated increased the annual turnover for approximately 8.265 euro.

Average treatment effect, by using matching based on the radius matching

The command to evaluate the average treatment effect, by using matching based on the radius matching "attr". In the current example, the evaluation of the average effect, by using this command is done as follows (Box 5):

Box 5.

```
attr qarkullimi2011 trajtimi2011 , pscore (ps) radius(0.001) comsup
```

Box 6.

```
ATT estimation with the Radius Matching method
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      133         600    8575.273    1830.149       4.686

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

As seen from the results (Box 6) through this estimation we determine that the effect of policy has a positive impact in the turnover of enterprise.

Average treatment effect by using stratification matching

The command to evaluate the average treatment effect by using stratification matching is "atts". In the current example, the estimate of the average effect using this command is done as follows (Box 7):

Box 7.

```
atts qarkullimi2011 trajtimi2011 , pscore (ps) blockid(blockf1) comsup
```

Box 8.

```
ATT estimation with the Stratification method
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      133         600    8201.809    1776.561       4.617

------------------------------------------------------------
```

As seen from the results (Box 8) and through this calculation, we determine that the effect of the policy has positive impact on the enterprise's turnover.

One of the ways to evaluate the validity of these calculations is to compare the results obtained, by using various algorithms. Theoretically, the results should have been **consistent in all cases**. In this concrete case, the results are approximate, which suggests that the same are sustainable.

b) DID

Step 1:

Note: Use the data "DID2.dta" from the file mentioned earlier. Also, use the syntax "DID2_do_file.do" which is found in the same location. The description of variables is shown in the above-mentioned document.

The DID is implemented by using regression method for data that have stretch of time. In this case, it should be used fixed effect panel regressions. In STATA, the command to execute such regression is "xtreg". In the current example, evaluation of the impact of policy is done as follows (Box 9):

Box 9.

```
xtreg  turnover treatment, fe i(id)
```

Box 10.

```
Fixed-effects (within) regression               Number of obs      =      5131
Group variable: id                              Number of groups   =       733

R-sq:  within  = 0.0169                         Obs per group: min =         7
       between = 0.0471                                        avg =       7.0
       overall = 0.0367                                        max =         7

                                                F(1,4397)          =     75.45
corr(u_i, Xb)  = 0.1371                         Prob > F           =    0.0000

------------------------------------------------------------------------------
  turnover |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    treatment |   2151.341   247.6716     8.69   0.000      1665.78    2636.902
       _cons |    6746.11    56.0889   120.28   0.000     6636.147    6856.072
-------------+----------------------------------------------------------------
     sigma_u |  11168.949
     sigma_e |  3346.5136
         rho |  .91761972   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(732, 4397) =    76.51           Prob > F = 0.0000
```

The results (Box 10), confirm that the impact of treatment was positive in the enterprise's turnover. On average, enterprises that have been treated, have increased annual turnover for approximately EUR 2,151. As seen, if we take into consideration the initial situation (since now we are using the data even before treatment of both groups of units), the effect of the policy turns out to be significantly smaller, although in both cases the direction is positive.

c) ***DID and PSM*** *(combined)*

Note: Use the data "DID1.dta' and 'DID2.dta' from the above mentioned file, according to the dynamic presented below. Also, use the syntax"DID1_do_file.do" which is found in the same location. Description of variables is shown in the above-mentioned document.

As discussed earlier, the combination of DID and PSM improves significantly the authenticity of the evaluation. This is because through this method we make sure that both groups are comparable and that they do not have major systematic differences.

Step 1:

Firstly, we implement the PSM as discussed above, by using the database "DID1.dta", according to the following formula (Box 11):

Box 11.

```
pscore  treatment region production services , pscore(ps) blockid(blockf1) comsup
level(0.001)
```

Box 12.

```
**************************************************
Algorithm to estimate the propensity score
**************************************************


The treatment is trajtimi

   trajtimi |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      4,488       87.47       87.47
          1 |        643       12.53      100.00
------------+-----------------------------------
      Total |      5,131      100.00


Estimation of the propensity score

Iteration 0:   log likelihood = -1936.3675
Iteration 1:   log likelihood = -1780.1912
Iteration 2:   log likelihood = -1776.4168
Iteration 3:   log likelihood = -1776.4076

Probit regression                                 Number of obs   =       5131
                                                  LR chi2(3)      =     319.92
                                                  Prob > chi2     =     0.0000
Log likelihood = -1776.4076                       Pseudo R2       =     0.0826

------------------------------------------------------------------------------
    treatment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      region |  -.4348009   .0597453    -7.28   0.000    -.5518996   -.3177023
  production |   .9239112   .0626643    14.74   0.000     .8010914    1.046731
  services |   .3784522   .0616153     6.14   0.000     .2576883     .499216
     _cons |  -1.501919   .0503025   -29.86   0.000     -1.60051   -1.403328
------------------------------------------------------------------------------


Note: the common support option has been selected
The region of common support is [.02638981, .28162951]
```

```
Description of the estimated propensity score
in region of common support

                Estimated propensity score
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0263898       .0263898
 5%     .0263898       .0263898
10%     .0595849       .0263898       Obs                5131
25%      .066559       .0263898       Sum of Wgt.        5131

50%     .1306197                      Mean           .1252824
                        Largest       Std. Dev.      .0853157
75%     .1306197       .2816295
90%     .2816295       .2816295       Variance       .0072788
95%     .2816295       .2816295       Skewness       .9119863
99%     .2816295       .2816295       Kurtosis       2.501651



*****************************************************
Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output
*****************************************************


The final number of blocks is 4

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks




***********************************************************
Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output
***********************************************************


The balancing property is satisfied


This table shows the inferior bound, the number of treated
and the number of controls for each block

  Inferior |
  of block |       treatment
of pscore  |         0          1 |     Total
-----------+----------------------+----------
  .0263898 |       502          9 |       511
       .05 |     1,851        130 |     1,981
        .1 |     1,402        229 |     1,631
        .2 |       733        275 |     1,008
-----------+----------------------+----------
     Total |     4,488        643 |     5,131

Note: the common support option has been selected


****************************************
End of the algorithm to estimate the pscore
****************************************
```

Similarly, the results (Box 12) show that the identified region of common support is sufficient.

Step 2:

The following command holds as a part of analysis, only the data that match (Box 13).

Box 12.

```
keep  id
sort  id
merge id using C:\......\DID_2.dta (to improve, depending on what path you keep the data)
keep if _merge==3
```

Step 3:

Once we have created the database with data that are already matching, we implement the DID as above, through the following command (Box 13):

Box 13.

```
xtreg  turnover treatment, fe i(id)
```

Box 14.

```
Fixed-effects (within) regression               Number of obs      =      5131
Group variable: id                              Number of groups   =       733

R-sq:  within  = 0.0169                         Obs per group: min =         7
       between = 0.0471                                        avg =       7.0
       overall = 0.0367                                        max =         7

                                                F(1,4397)          =     75.45
corr(u_i, Xb)  = 0.1371                         Prob > F           =    0.0000

------------------------------------------------------------------------------
  turnover |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Treatment |   2151.341   247.6716     8.69   0.000      1665.78    2636.902
       _cons |    6746.11    56.0889   120.28   0.000     6636.147    6856.072
-------------+----------------------------------------------------------------
     sigma_u |  11168.949
     sigma_e |  3346.5136
         rho |  .91761972   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(732, 4397) =    76.51           Prob > F = 0.0000
```

Similarly, the results (Box 14) shows that the effects are positive, but are smaller than when we evaluated based on the last year's data. As you may notice, the last results do not differ from those of the DID, this because during compliancy no unit is eliminated since all were approximate based on the propensity score.

References

Ashenfelter, O. 1978. "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics* 60 , 47–57.

Ashenfelter, O., and D. Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics* 67, 648–660.

Becker, Sascha, and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *Stata Journal* 2 (4): 358–77.

Blundell, R., and Monica Costa-Dias. 2002. "Alternative Approaches to Evaluation in Empirical Microeconomics, " Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Bryson, Alex, Richard Dorsett, and Susan Purdon. 2002. "The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies." Working Paper 4, Department for Work and Pensions, London.

Heckman, James J., Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, vol. 3, ed. Orley Ashenfelter and David Card, 1865–2097. Amsterdam: North-Holland.

Imbens, G. and Wooldridge, J. 2009. "Recent Developments in the Econometrics of Program Evaluation", Journal of Economic Literature 47, 5–86. (also available as IZA-DP)

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Smith, Jeffrey, and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–53.