Chapter 6: Difference in Difference Designs

Pre-requisites

- Chapter 1: Intro to STATA
- Chapter 2: Review of Regression
- Chapter 3: Experiments
- Chapter 4: Problems with Experiments
- Chapter 5: Regression Discontinuity Designs

Contents

1. Introduction

In previous chapters, we have covered techniques that are based on the knowledge of the assignment to treatment process. In experimental designs, the evaluators know the assignment rule to treatment since they participate in the selection process of treated and control units. In regression discontinuity designs, the evaluator takes advantage of an existing rule that defines treatment.

When the assignment rule is unknown, evaluators need to figure out what was the mechanism under which units select into treatment. Typically, this task is challenging and the assumptions required for computing an unbiased estimate of the impact are stronger. As a consequence, impact evaluation techniques under unknown (or imperfect knowledge of) assignment rule are usually more demanding in terms of data. You typically need several baseline and follow-up surveys (ideally panel data for several periods or many repeated cross-sectional surveys) and a rich set of control variables.

We will cover in this chapter one very popular non-experimental approach in impact evaluation that exploits temporal and cross-sectional variation to deliver a causal estimate of an intervention: the so-called differences in difference (DD) designs. DD designs compare the *change over time* in the outcome for the treatment group and contrasts that with the same change for the control group. In the context of PROGRESA, we can imagine a situation in which we collect data on enrollment for treatment and control villages before and after the implementation of the program. We know that a simple *before-after* comparison for treated villages are likely to give a biased estimate of PROGRESA since the impact of the intervention can be cofounded with any temporal shock (for instance, a natural disaster that affect schools). On the other hand, a simple comparison at follow-up between treatment and control villages is likely to suffer an omitted variable problem (like the inability to control for unobserved factors that affect treatment and outcome such that talent of the village mayor or social capital). Using a DD design, it is possible to control for temporal shocks and for those unobservable dimensions that are fixed over time. If villages' social capital affects enrollment and the access to the program and the level of social capital is constant over time, then we can control for this factor by simply differencing the outcome over time. We can also control for any temporal shock by using time-fixed effects. Under the critical assumption that there are no time-varying unobservable factors that explain treatment and outcomes, DD recover the true impact of the intervention.

We will cover the basic of DD and some extensions. We start with the simplest case in which we compare two units before and after the intervention. We present ways to estimate this simple case and well as a regression-based specification to estimate the same elements of this comparison. We then extend the basic model in several ways. We first accommodate the inclusion of covariates and expand the model to accommodate several groups and periods. We also incorporate matching in the DD design to reduce baseline heterogeneity before computing the differences in outcomes between treatment and control units before and after the intervention. Although there are several versions of matching that can be implemented here, we focus on the simplest case of matching via the propensity score. We conclude the chapter with a discussion of triple differences designs (DDD).

Specifically, we will cover the following issues:

- We consider the basics of DID designs for the simple case of before-after and treatment-control comparisons. We extend the model to incorporate covariates.
- We study extensions to the basic DID including its combination with propensity score matching and triple differences.

At the end of this chapter we expect students to be able to:

- Implement the estimation of the DID model using standard regression techniques.

- Improve the DID using matching techniques.
- Implement triple differences designs.

2. Preliminaries: the non-randomized Oportunidades dataset

In this section, we will be using a new dataset called `Panel_OPORTUNIDADES_00_07_year.dta` that we have prepared for this chapter. This is a panel dataset for a sample of households and individuals who were followed during three survey rounds for years 2000, 2003 and 2007. The original PROGRESA design studied 506 villages (320 assigned to treatment and 186 to control) and selection intro treatment was randomized. By the fall of 2000, all villages were incorporated into the program. In 2003, a new control group was incorporated with the goal of studying long-term impacts of the program. Since randomization was no longer feasible, the selection of this new control group was based on non-experimental techniques, particularly a matching algorithm. Each of the original 506 villages was matched to a comparison village from a pool of 14,000 villages no previously treated by the program using an algorithm based on location (state), villages' distance to school and clinics, and a propensity score constructed at household level using indexes of household assets and housing characteristics, household composition, labor market indicators, educational outcomes, distance to educational and health facilities and locality size.

Our dataset merges household and village level data for the final year of the randomized design (2000) with non-experimental samples for 2003 and 2007. For this section, we consider a basic case in which we compare outcomes between treatment and control villages in 2000 with the same outcomes in 2003. Therefore, we have in the initial period that already a fraction of villages were exposed to treatment.

This case is not the standard way in which DID models are discussed in introductory books. A standard intro to DID considers the case in which in the initial period (or baseline) there are no treated units and then a group of units gets treated in the final period. Then, a comparison between treatment and control, before and after the intervention, delivers the causal impact of the program.

We already covered this basic case in our discussion about the use of baseline data in experimental designs in Chapter 3 (section 9.3: Using baseline data). The only difference here is that now we are dealing with non-experimental data and some assumptions need to be imposed to interpret the DID estimates as causal. However, from a purely mechanical view, the techniques are the same.

To begin our discussion about DID, let's open the dataset:

```
set more off
clear all

use"$path/Panel_OPORTUNIDADES_00_07_year.dta", clear
```

Since this a new dataset, it is important to obtain some basic familiarity with it. We do that using the command `describe`. The result is the following:

```
Contains data from Panel_OPORTUNIDADES_00_07_year.dta
  obs:       466,838
 vars:            23                          4 Mar 2014 01:09
 size:    44,349,610
-------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------
year            int     %9.0g
villid          str9    %9s                   Village ID
geopolid        str2    %9s                   Federal entity
hogid2          str24   %24s                  Household ID/2000-2007
indexpov_HH     float   %9.0g                 HH Poverty score(Modelo 2003)
pov_HH          byte    %9.0g      pov        HH Poverty Status: 1= poor,0= Non poor
D_HH            byte    %9.0g      D_HH       Household-Level Treatment status
D               byte    %9.0g      D          Village-Level Treatment status
iid2            str26   %26s                  Individual ID
age             float   %9.0g                 Age: Years
sex             byte    %9.0g      sex        Gender: 1= Male, 0=Female
edu_child       byte    %9.0g                 Education (6-16): Years
enroll_child    byte    %9.0g                 enroll child (6-16): 1= Y, 0=N
labor            byte  %10.0g     labor condition (>8): 1:occupied; 0:unoccupied
Income_HH_per   float   %9.0g
Income_HH       float   %9.0g
sick            byte    %9.0g      sick       sick child(<5): 1= Y, 0=N
sick_child      byte    %9.0g                 sick child(<5): 1= Y, 0=N
Measure         byte    %9.0g      Measure    6 Months-measured child(<2): 1= Y, 0=N
famsize         int     %9.0g                 HH size
HH              byte    %9.0g                 HH head
agehead         int     %9.0g      age        Age of HH head: years
sexhead         byte    %9.0g      sex        Gender of HH head: 1= M, 0=F
Sorted by:
```

We refer the reader to the previous chapters for an explanation of these variables. We now inspect how the dataset is organized in the two years that are going to be part of the basic analysis. The result is the following:

```
. tab D year

Village-Le |
       vel |
 Treatment |          year
    status |      2000       2003 |     Total
-----------+----------------------+----------
   Control |    44,070     32,856 |    76,926
   Treated |    67,887    126,023 |   193,910
-----------+----------------------+----------
     Total |   111,957    158,879 |   270,836
```

In the dataset, `D` refers to the treatment status (which is given at a village level) with information defined at individual level. We observe a large increase in the treatment group for 2003 and a reduction in the size of the control group. This is consistent with the changes in the design of the evaluation sample of Oportunidades as described above. The treatment group sample increases because all of those who were part of the randomized control group are now being treated and a new comparison group was selected. Although this can be problematic if we want to analyze the data using standard panel data methods, this does represent a concern since the empirical exercise takes villages as unit of analysis.

Before continuing, we need to create some additional variables. First of all, we need to define a dummy variable equal to zero for the initial period (2000) and one for the final period (2003). Then, the variable `D` is not available for 2007 in the original dataset. We need to create it by using the information about treatment status at household level (`D_HH`). The code below creates both variables:

```
* Additional variables

gen period=0 if year==2000
replace period=1 if year==2003

gen aux=0 if year==2007
replace aux=1 if D_HH==1 & year==2007
egen aux2=max(aux) if year==2007,by(villid)

replace D=aux2 if year==2007
drop aux aux2
```

Since the code is similar to the ones used in previous chapters, we don't spend time discussing the details. The rest of the chapter introduces the basic characteristics of the DD design and some common extensions.

3. Difference in difference: the basics

Consider the following scenario: you are hired by Oportunidades to analyze the impact of this conditional cash transfer on children education. You have a longitudinal sample (2 periods) of villages targeted by the program and a group of non-treated villages. You have imperfect information regarding the selection process. Then, you are suggested to implement a DD design. Before implementing this strategy, it is critical to understand what the DD model does and the source of variation in the data that exploits to estimate the impact of the intervention. We cover these issues in the following section.

3.1. Conceptual remarks

So far, we have paid attention to the case in which a single cross-section was available. We need to introduce some additional notation to accommodate the fact that longitudinal data is now available. We return to the discussion about potential outcomes but in a context in which we have at least one extra period with information about potential outcomes for individuals. We first define the treatment variable in a given period in the following way:

$$(1)\quad D_i = \begin{cases} 1 \text{ if } t_{it_1} = 1 \\ 0 \text{ otherwise.} \end{cases}$$

Under the assumption of **common trends** and **no selection based on transitory shocks**, it is possible to write the following equation:

$$(2)\quad y_{it} = \alpha + \beta D_{it} + \varepsilon_{it};$$

where $E\left[\varepsilon_{it}/D_i, t\right]=E\left[p_i/D_i\right]+q_t$, being p_i a non-observable individual fixed effect and q_t an aggregated macro shock. Under the previous assumptions, we can write the following expression:

(3) $$E\left[y_{it}/D_i, t\right]=\begin{cases}\alpha+E\left[\beta/D_i=1\right]+E\left[p_i/D_i=1\right]+q_t \text{ if } \mathrm{D}_{it}=1, t=1\\ \alpha+E\left[p_i/D_i\right]+q_t\end{cases}$$

This implies that the constant and the error term can be dropped via a sequential use of differences in the following way:

(4) $$\begin{aligned}\beta^{DID}&=\left\{E\left[y_{it}/D_i=1,t=1\right]-E\left[y_{it}/D_i=1,t=0\right]\right\}\\ &-\left\{E\left[y_{it}/D_i=0,t=1\right]-E\left[y_{it}/D_i=0,t=0\right]\right\}.\end{aligned}$$

This expression can be computed from the simple analog. In this case, we can estimate $ATET$. It is important to keep in mind that we have considered only a simple case with two-units and two-periods. We will cover extensions to this basic design in next sections.

Typically researchers use interactions to estimate the DD model in a regression framework. For instance, consider the following equation:

(5) $$Y_{it}=\alpha+\gamma\phi_i+\lambda t_t+\beta(\phi_i * t_t)+\varepsilon_{it}.$$

where ϕ_i is a dummy variable equal to 1 for the treatment group and t_t is a period fixed effect. Notice that $D_{it}=\phi_i * t_t$. To show that this equivalent to our previous derivation, we estimate each element by taking the conditional expectation for each element of the model. Since we have two units and two periods, we define 4 potential outcomes. The first one is the outcome of the control group before the implementation of the program. This is represented in the following way:

(6) $$E[Y_{it}/i=0,t=0]=\gamma_0+\lambda_0=\alpha$$

A simple manipulation shows that this is equivalent to the constant of the regression. Recall that we already showed that in a simple cross-sectional regression the constant represents the mean outcome for the control group. The logic is essentially the same here. We also can estimate the other potential outcomes in the same way:

(7) $$\begin{aligned}&E[Y_{it}/i=1,t=0]=\gamma_1+\lambda_0=\alpha+\gamma\\ &E[Y_{it}/i=0,t=1]=\gamma_0+\lambda_1=\alpha+\lambda\\ &E[Y_{it}/i=1,t=1]=\gamma_1+\lambda_1=\alpha+\gamma+\lambda+\beta\end{aligned}$$

We have estimated each element. We then proceed to estimate the first difference for the case of the treatment and control groups. The result is the following:

(8) $$\begin{aligned}&E[Y_{it}/i=1,t=1]-E[Y_{it}/i=1,t=0]=\alpha+\gamma+\lambda+\beta-(\alpha+\gamma)=\lambda+\beta\\ &E[Y_{it}/i=0,t=1]-E[Y_{it}/i=0,t=0]=\alpha+\lambda-\alpha=\lambda\end{aligned}$$

Notice that this is consequence of the assumption regarding similar trends between treatment and control groups. Therefore, the coefficient for the time fixed effect in the equation recovers the before-after comparison for the control group.

We then estimate the difference between these two differences. The result is the following:

(9) $$\left(E[Y_{it}/i=1,t=1]-E[Y_{it}/i=1,t=0]\right)-\left(E[Y_{it}/i=0,t=1]-E[Y_{it}/i=0,t=0]\right)=\lambda+\beta-\lambda=\beta$$

Therefore, the coefficient associated to the interaction recovers the DD estimate. In this way, by looking at the coefficient of the interaction between treatment and period, we are able to implement a DD design in a regression framework.

Interactions are pretty common in empirical research. We will study more complex interactions in the final section of this chapter[1].

3.2. Evaluating the DD sample

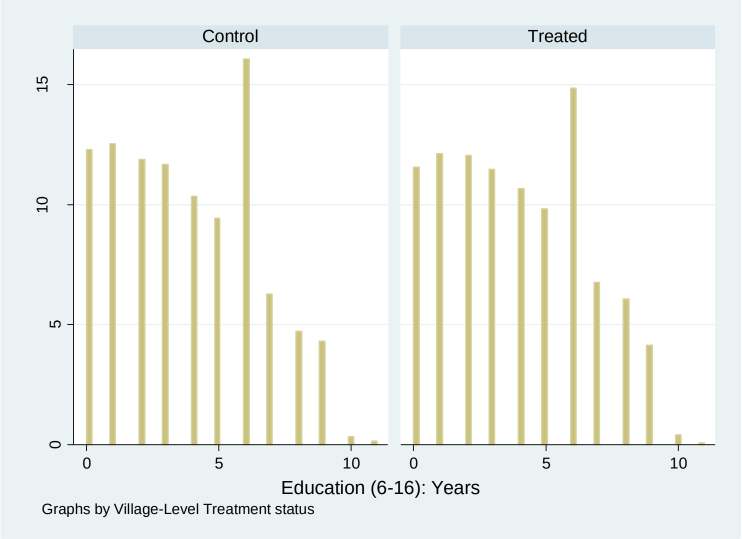
Now, we are in the position of implementing this technique in STATA. Since we are exploiting variation over-time between treatment and control groups, it is useful to show graphically that information. One simple way to do that consists in using histograms comparing both groups before and after the treatment. The code is the following:

```
* HISTOGRAM 1: Education levels before treatment

histogram edu_child if year==2000, bin(50) percent by(D)

* HISTOGRAM 2: Education levels after treatment

histogram edu_child if year==2003, bin(50) percent by(D)
```

The resulting graph for the pre-treatment case is below:

Figure 1: Education levels by treatment status (initial period)

[1] It is important to mention that interactions are useful for analyzing differential responses between groups in general. We saw a special case of this during the discussion on spillover effects.



The graphical evidence here seems to suggest that education levels for the treated group in the initial period are slightly better than those in control group, although this difference does not seem to be large enough. Since we are using as baseline a scenario in which the treatment group has been already exposed to PROGRESA for at least 3 years, this should not be surprising. In any case, the graphical evidence is not convincing enough since this small difference can still be consistent with a lack of rejection of the null hypothesis of no impact of the intervention. A formal test is needed to evaluate that.

Figure 2: Education levels by treatment status (final period)

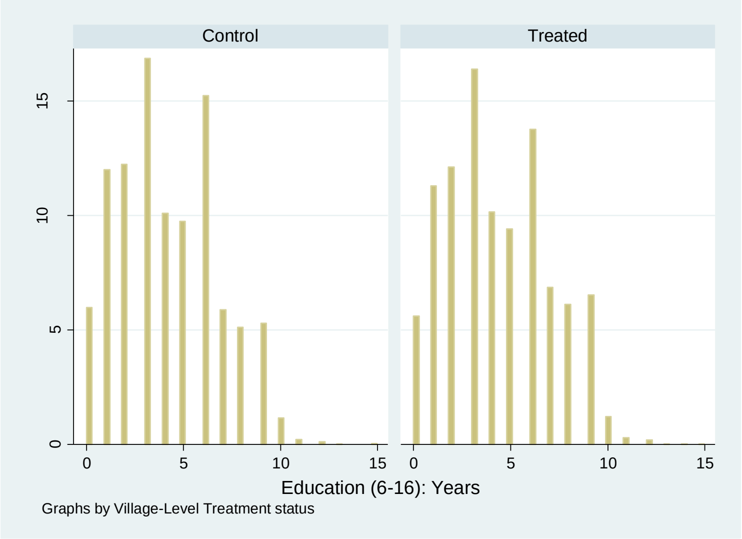


Figure 2 presents the education levels for treatment and control villages in the final period (2007). We see that treated villages seem to be slightly better in terms of children education years. Unless a formal test is used, it is not clear whether the difference in children education between treatment and control villages really exists.

In the ideal scenario in which the program or intervention is no yet implemented, the credibility of the DD design would rest in showing evidence of balance between treatment and control units before the intervention. This is the typical way experiments are analyzed and that's the reason DD is the typical approach used to analyze the so-called "natural experiments". Pre-treatment balance is a stronger requirement that is not necessary for DD designs and can be replaced with the weaker assumption about pre-trends between treatment and control units. In this case, since the treatment group was already exposed to treatment in our baseline year (2000), we know in advance that this estimate is biased but we still show them to introduce the logic of DD designs.

We now test balance formally using a t-test. We have already covered in many chapters how to perform such test. We will show you how to do this test in the context of DD. We can install a user-written program called `diff` (`ssc install diff`). We can implement a t-test for evaluating balance between treatment and control villages using the following routine in STATA:

```
* TEST OF DIFFERENCES IN MEANS

diff edu_child, t(D) p(period) cov(age sex agehead sexhead) test
```

The output is the following:

```
. diff edu_child, t(D) p(period) cov(age sex agehead sexhead) test

TWO-SAMPLE T TEST

Number of observations (baseline): 34906
            Baseline       Follow-up
   Control: 13699          -           13699
   Treated: 21207          -           21207
            34906          -

t-test at period = 0:
----------------------------------------------------------------------------------------------
 Variable(s)         |   Mean Control    | Mean Treated |    Diff.    |   |t|   |  Pr(|T|>|t|)
---------------------+------------------+--------------+------------+---------+---------------
edu_child            | 3.815            | 3.896        | 0.080      |  2.76   | 0.0057***
age                  | 25.459           | 25.226       | -0.233     |  1.84   | 0.0656*
sex                  | 0.494            | 0.503        | 0.008      |  2.63   | 0.0086***
agehead              | 47.334           | 46.855       | -0.479     |  5.36   | 0.0000***
sexhead              | 0.917            | 0.918        | 0.001      |  0.63   | 0.5266
----------------------------------------------------------------------------------------------
*** p<0.01; ** p<0.05; * p<0.1
```

The output displays the means for treatment and control villages before the intervention for the outcome of interest (child education) and a set of covariates (age and sex of the child and household head). We have for all the characteristics except sex of household head unbalance between treatment and control villages. Although the difference is not large in magnitude, the difference is significant in a statistical sense. This result is expected given the reasons explained above regarding the actual implementation of the program.

3.3. Implementing DD in STATA

We now proceed to estimate the DD model using a regression framework. To do that, we need to create the interaction between treatment status and the period dummy. The code is the following:

```
* Creating interaction

gen int_dXperiod=period*D
```

We then incorporate the interaction in a regression model. The specification is the following:

```
* Basic regression

reg edu_child int_dXperiod D period, r
```

The results are the following:

```
. reg edu_child int_dXperiod D period, r

Linear regression                                      Number of obs =   82619
                                                       F(  3, 82615) =  158.87
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0058
                                                       Root MSE      =  2.6468


------------------------------------------------------------------------------
             |               Robust
   edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
int_dXperiod |     .07501    .041586     1.80   0.071    -.0064983    .1565183
           D |   .0801867   .0289837     2.77   0.006     .0233789    .1369945
      period |   .3175853   .0348287     9.12   0.000     .2493213    .3858494
       _cons |   3.815461   .0225669   169.07   0.000      3.77123    3.859692
------------------------------------------------------------------------------
```

We focus on the coefficient of the interaction which delivers the causal effect of interest. We estimate an increase of 0.075 years of education for those in treatment villages with respect to control villages.

Given the non-randomized nature of treatment assignment and the lack of balance in socioeconomic characteristics previously found, there is the concern that our previous estimate might be biased. We can control for those characteristics by adding these variables in our previous regression model. The STATA code is the following:

```
* Adding controls

reg edu_child int_dXperiod D period age sex agehead sexhead, r
```

The results are the following:

```
. reg edu_child int_dXperiod D period age sex agehead sexhead, r

Linear regression                                      Number of obs =   77927
                                                       F(  7, 77919) =18421.73
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6802
                                                       Root MSE      =  1.4966


------------------------------------------------------------------------------
             |               Robust
   edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
int_dXperiod |   .0157462   .0239681     0.66   0.511    -.0312312    .0627235
           D |    .103376   .0148636     6.95   0.000     .0742434    .1325087
      period |   .2958656    .020499    14.43   0.000     .2556877    .3360435
         age |   .7102853   .0020115   353.11   0.000     .7063427    .7142279
         sex |  -.0966714   .0107214    -9.02   0.000    -.1176854   -.0756574
     agehead |  -.0013591   .0004906    -2.77   0.006    -.0023207   -.0003976
     sexhead |   .0940899   .0209877     4.48   0.000     .0529541    .1352256
       _cons |  -3.906229    .035701  -109.42   0.000    -3.976202   -3.836255
------------------------------------------------------------------------------
```

The coefficient associated to the interaction is no longer significant and experienced a dramatic reduction. This is evidence of substantial heterogeneity between treatment and control groups suggesting that the assumptions required for DD are not valid in this application. This calls for an alternative approach to deal with this heterogeneity. The combination of DD and propensity score matching can be a useful approach to minimize bias as we will see later.

We can obtain the same results using the user-written command `diff`. The STATA code is the following:

```
diff edu_child, t(D) p(period) cov(age sex agehead sexhead)  robust
```

The STATA output is the following:

```
. diff edu_child, t(D) p(period) cov(age sex agehead sexhead)  robust

DIFFERENCE-IN-DIFFERENCES WITH COVARIATES


Number of observations in the DIFF-IN-DIFF: 77927
            Baseline       Follow-up
   Control: 12337          9418         21755
   Treated: 18615          37557        56172
            30952          46975

R-square:  0.68022

                               DIFFERENCE IN DIFFERENCES ESTIMATION
-------------------- ------------ BASE LINE --------- ----------- FOLLOW UP ---------- --------------
 Outcome Variable     | Control |  Treated   | Diff(BL) | Control |  Treated   | Diff(FU)  | DIFF-IN-DIFF
--------------------+---------+-----------+----------+---------+-----------+----------+--------------
edu_child            | -3.906  | -3.803    | 0.103    | -3.610  | -3.491    | 0.119     | 0.016
Std. Error           | 0.036   | 0.035     | 0.015    | 0.037   | 0.034     | 0.019     | 0.024
t                    | -109.42 | -0.95     | 6.95     | 4.13    | -3.05     | 0.94      | 0.66
P>|t|                | 0.000   | 0.000     | 0.000*** | 0.000   | 0.000     | 0.000***  | 0.511
----------------------------------------------------------------------------------------------------
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

The advantage of using this command is that the difference in means for baseline and follow-up are presented. As expected, the estimate is exactly the same as the previously estimated.

4. Beyond the basics: generalizing DID

In the previous section, we have covered the basic 2 units-2 periods DD model. However, this model can be generalized for a case in which there are more than 2 units and several periods. In the context of our example, we can exploit the 3 periods available in the dataset and the fact that we have a large number of villages. To accommodate this case in a regression framework, we need to add fixed effects for each unit and each year. The basic equation is the following:

(10) $y_{it} = \alpha_i + \lambda_t + \beta D_{it} + X_{it}^{'}\delta + \varepsilon_{it};$

where y_{it} is the outcome of interest for unit i in period t. α_i and λ_t are respectively unit and period fixed effects. D_{it} is the treatment status of unit i in period t. $X_{it}^{'}\delta$ includes control variables and ε_{it} is an error term. The parameter of interest is β which recovers the effect of interest. The time fixed-effects accounts for the time-series changes in our dependent variable. The unit fixed-effects controls for time-invariant characteristics at unit level and the D_{it} accounts for changes in the dependent variable in treated units associated to the switch of treatment status across periods.

To implement this model, we need to make some changes in our dataset. We first need to create fixed effects for each year. We also need to create fixed effect for each village. Since we have about 1,000 villages in the sample, creating a single dummy variable for each village is not practical. We can accommodate these fixed effects using the command `xtreg` but the variable that identifies the id for villages needs to be in a numerical format. Since the original variable is in string format, we convert it to a numerical value using the command `destring`. The code for both cases is below:

```
tab year, gen(ydum)

gen villid2=villid
destring villid2, replace
```

After implementing these changes, we can proceed with the estimation of this generalized DD model. The STATA code is the following:

```
xtreg edu_child D ydum2 ydum3, fe i(villid2)
```

We added the time fixed effects with the inclusion of the dummy variables `ydum2 ydum3` (we exclude one dummy to avoid the dummy trap) and the village fixed effects are incorporated with the option `fe i(villid2).`

The output is the following:

```
. xtreg edu_child D ydum2 ydum3, fe i(villid2)

Fixed-effects (within) regression               Number of obs      =    123313
Group variable: villid2                         Number of groups   =       869

R-sq:  within  = 0.0256                         Obs per group: min =         1
       between = 0.1102                                        avg =     141.9
       overall = 0.0292                                        max =      1258

                                                F(3,122441)        =   1070.94
corr(u_i, Xb)  = -0.0072                        Prob > F           =    0.0000

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

```
   edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0677308   .0276197     2.45   0.014     .0135967    .1218649
       ydum2 |   .4529914   .0200191    22.63   0.000     .4137543    .4922284
       ydum3 |  -.5973277   .0246443   -24.24   0.000    -.6456303   -.5490252
       _cons |   3.774731   .0199313   189.39   0.000     3.735666    3.813796
-------------+----------------------------------------------------------------
     sigma_u |  .73578849
     sigma_e |  2.4596298
         rho |  .08213796   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(868, 122441) =     6.49         Prob > F = 0.0000
```

We focus on the coefficient for the village-level treatment variable D. We find an average impact of 0.06 years of education for treated villages with respect to control ones. This effect is similar to the basic estimate of the simplest DD model with only two periods-two groups. The advantage of using village-level fixed effects here is that we are able to control for time-invariant characteristics that can be related to treatment and outcome of interest. This differs from the basic model in which an aggregate comparison of means between treatment and control villages is performed without adequately controlling for heterogeneous characteristics at baseline.

We can include additional individual and household level covariates to remove potential biases related to time-variant dimensions. For simplicity, we use children age and sex along the same variable for the household head. The results are reported below:

```
. xtreg edu_child D ydum2 ydum3 age sex agehead sexhead, fe i(villid2)

Fixed-effects (within) regression               Number of obs      =    118567
Group variable: villid2                         Number of groups   =       869

R-sq:  within  = 0.7362                         Obs per group: min =         1
       between = 0.6638                                        avg =     136.4
       overall = 0.7236                                        max =      1163

                                                F(7,117691)        =  46926.37
corr(u_i, Xb)  = 0.0354                         Prob > F           =    0.0000

------------------------------------------------------------------------------
   edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0762787    .014743     5.17   0.000     .0473827    .1051746
       ydum2 |   .3630784   .0108189    33.56   0.000     .3418736    .3842832
       ydum3 |   .3283461   .0133715    24.56   0.000     .3021381     .354554
         age |   .7443836   .0013419   554.72   0.000     .7417535    .7470138
         sex |  -.1158147   .0074551   -15.54   0.000    -.1304266   -.1012029
     agehead |  -.0039912   .0003357   -11.89   0.000    -.0046492   -.0033332
     sexhead |     .06751   .0118588     5.69   0.000     .0442669    .0907531
       _cons |  -4.148594   .0248182  -167.16   0.000    -4.197238   -4.099951
-------------+----------------------------------------------------------------
     sigma_u |  .46515741
     sigma_e |  1.2760442
         rho |  .11729608   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
F test that all u_i=0:      F(868, 117691) =    13.53          Prob > F = 0.0000
```

The coefficient is slightly higher than before (0.076) but remains essentially the same as the previous one. This gives credibility to this specification compared to the basic model discussed in section 3. Since DD is only able to remove time-invariant characteristics, it is still possible to be exposed to bias related to baseline heterogeneity. This calls for a strategy to minimize this source of bias. We discuss that in the next section.

5. Difference in difference meets matching

One concern with DD models in the presence of heterogeneity at baseline. This is true since we lack of perfect knowledge regarding the assignment of units into treatment. In experimental designs, we paid a lot of attention in providing evidence that there was not unbalance in pre-treatment characteristics between treatment and control units. Following the same logic, we may want to reduce heterogeneity by restricting the sample to units that are similar in terms of socioeconomic characteristics in a way that mimic pre-treatment balance in an experimental design. Using a an alternative technique known as matching on baseline characteristics is a powerful complement to a DD design since this technique minimizes bias making the estimate of DD less sensitive to this issue.

Matching is a statistical procedure that constructs an artificial control group based on a set of observable dimensions. Therefore, each treatment unit is matched to one or more control units that share the same observable characteristics. This can performed in several ways and there is a large literature of matching techniques. We focus here on propensity score matching, although the same logic can be followed with any matching technique since we just need to use matching to construct a matched sample; this is, a sample of treatment and control units that share similar characteristics.

5.1. Conceptual issues

We need to introduce some basic concepts before proceeding with the STATA implementation. The first one is the **propensity score**. The propensity score is just the probability of receiving treatment as a function of a set of characteristics. The advantage of using the propensity score instead of a matching based on the observed characteristics is that it helps to avoid the so-called **curse of dimensionality**. This refers to the fact that the introduction of large number of characteristics can derive in a situation in which some combination of characteristics would lack of treatment or control units. In propensity score matching, we match units in terms of their values for the propensity score, making this concern irrelevant. Roseanbaum and Rubin (1983) have shown that matching on the propensity score is equivalent to matching on characteristics.

One critical condition for propensity score matching is the validity of the **common support condition**. We want to restrict the comparison to units that share values of the propensity score for treatment and control units.

5.2. Preparing the dataset

We now turn to STATA to implement this technique. For simplicity, we return to our first example for before-after comparison between treatment and control units. We need to change the focus of the analysis to a comparison between treated versus non-treated households rather than villages since all villages in 2000 are eventually treated in 2003 as explained in section 2. Therefore, there is no variation to exploit in 2003 for estimating the propensity score at village-level analysis.

Since we need to estimate the propensity score for being treated in 2003 using household level socioeconomic characteristics from 2000, we need to assignment treatment status from 2003 to observations in 2000. To do that, we create an auxiliary variable for treated households in period 1 and the assign that information to period 0 using the `egen` command. The code is below:

```
* Preparing the dataset

gen auxp=0
replace auxp=1 if period==1 & D_HH==1

egen participation_03=max(auxp),by(hogid2)

keep if period==0
```

We have also dropped the observations for 2003. The dataset contains socioeconomic characteristics for 2000 along treatment status in 2003. This is all what we need to estimate the propensity score.

5.3. Estimating the propensity score

There are several options to compute the propensity score. The simplest is using the commands `probit` or `logit` from the official STATA version. We prefer the user-written program `psmatch2` due to the advantages it offers to evaluate the balance between treatment and control units after matching on the propensity score.

For simplicity, we consider a parsimonious model to estimating the propensity score. We assume that the probability of treatment depends of age, sex, family size, income per-capita and two household head characteristics: age and sex. The STATA code is the following:

```
* Estimating propensity score

psmatch2 participation_03 age sex famsize agehead sexhead Income_HH_per,
logit
```

The results of the estimation are below:

```
. psmatch2 participation_03 age sex famsize agehead sexhead Income_HH_per, logit

Logistic regression                               Number of obs   =      94470
                                                  LR chi2(6)      =    2511.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -57072.291                       Pseudo R2       =     0.0215

------------------------------------------------------------------------------
participa~03 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.001775   .0004025    -4.41   0.000    -.0025639   -.0009861
         sex |  -.0194013   .0143404    -1.35   0.176     -.047508    .0087054
```

```
     famsize |   .1059887   .0032057    33.06   0.000     .0997057    .1122716
     agehead |  -.0092888   .0005621   -16.52   0.000    -.0103905   -.0081871
     sexhead |   .3676487   .0252544    14.56   0.000     .3181509    .4171464
Income_HH_~r |  -4.56e-07   3.33e-07    -1.37   0.171    -1.11e-06    1.97e-07
       _cons |   .3264058   .0399982     8.16   0.000     .2480108    .4048008
------------------------------------------------------------------------------
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
```

We use a logit model to estimate the propensity score. Although the coefficients of the logit estimation do not have a causal interpretation, we can certainly establish an association between the probability of treatment and all the variables except sex and income per-capita.

5.4. Evaluating the matched sample

Computing the propensity score is usually a task that demands trying several specifications before reaching one in which there is a reasonable good balance between treatment and control groups. For this reason, it is important to evaluate whether the estimated propensity score leads to balance between treatment and control units. There are several statistical and graphical tools that can be implemented to perform this task.

The simplest way is to run t-tests comparing treatment and control groups for the original and the matched sample that results from the estimated propensity score. There are several ways to implement this using the standard `ttest` command in STATA for the original sample and for the sample for the common support defined by the previously estimated propensity score. Fortunately, there is a user-written command `pstest2` that does this for us with just a single line of command. The code is the following:

```
* Evaluating resulting matched sample

pstest2 age sex famsize agehead sexhead Income_HH_per, t(participation_03)
graph sum
```

We test the balance for the original and the matched sample based on age, sex, family size, income per-capita as well as household head age and sex. The treatment variable is whether the household is treated in 2003 (the final year). The output is the following:

```
. pstest2 age sex famsize agehead sexhead Income_HH_per, t(participation_03)
graph sum


----------------------------------------------------------------------------
                        |       Mean               %reduct |     t-test
    Variable     Sample | Treated Control    %bias  |bias| |    t    p>|t|
------------------------+----------------------------------+----------------
         age  Unmatched | 24.174   27.654    -16.6         | -26.52  0.000
                Matched | 24.774   25.337     -2.7    83.8 |  -5.12  0.000
                        |                                  |
         sex  Unmatched | .50013   .49798      0.4         |   0.63  0.526
                Matched | .50343   .51394     -2.1  -388.3 |  -3.80  0.000
```

```
                        |                                  |
     famsize  Unmatched | 6.4333   5.6675     30.9         |  48.83  0.000
                Matched | 6.4709   6.4931     -0.9    97.1 |  -1.61  0.108
                        |                                  |
     agehead  Unmatched | 46.246   48.717    -17.0         | -26.54  0.000
                Matched | 46.386   46.461     -0.5    97.0 |  -0.99  0.323
                        |                                  |
     sexhead  Unmatched | .93251   .88506     16.5         |  25.25  0.000
                Matched | .93374   .93192      0.6    96.2 |   1.31  0.189
                        |                                  |
Income_HH_~r  Unmatched |  698.5   938.16     -1.2         |  -2.01  0.045
                Matched | 712.95   683.37      0.1    87.7 |   0.37  0.710
                        |                                  |
----------------------------------------------------------------------------
-------------------------------------------------------------
         Summary of the distribution of the abs(bias)
-------------------------------------------------------------

                       BEFORE MATCHING
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .4303378       .4303378
 5%     .4303378       1.199926
10%     .4303378       16.53453       Obs                   6
25%     1.199926       16.63855       Sum of Wgt.           6

50%     16.58654                      Mean           13.78222
                        Largest       Std. Dev.      11.44427
75%     17.01781       16.53453
90%     30.87218       16.63855       Variance       130.9713
95%     30.87218       17.01781       Skewness       .1019097
99%     30.87218       30.87218       Kurtosis       1.992101
-------------------------------------------------------------

                       AFTER MATCHING
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .1480799       .1480799
 5%     .1480799       .5157014
10%     .1480799       .6343178       Obs                   6
25%     .5157014       .8928029       Sum of Wgt.           6

50%     .7635604                      Mean           1.164287
                        Largest       Std. Dev.      1.002417
75%     2.101512       .6343178
90%     2.693307       .8928029       Variance       1.004841
95%     2.693307       2.101512       Skewness       .6322602
99%     2.693307       2.693307       Kurtosis       1.798162
-------------------------------------------------------------
```

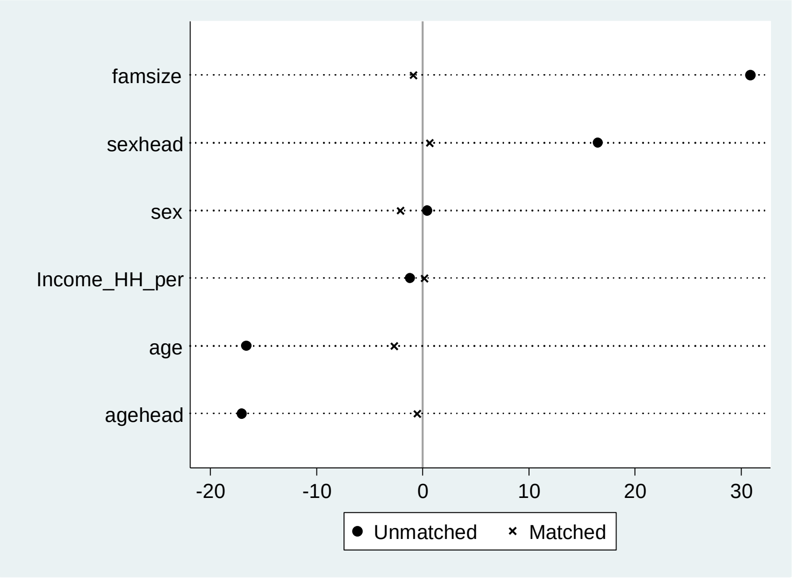
```
---------------------------------------------------------------
     Sample |    Pseudo R2      LR chi2        p>chi2
------------+--------------------------------------------------
  Unmatched |        0.022      2510.21          0.000
    Matched |        0.000        56.36          0.000
---------------------------------------------------------------
```

For each variable, the output includes the means for treatment and control groups for the unmatched and matched sample, the magnitude of bias, the reduction in bias after matching on the propensity score and the t-test for the unmatched and matched sample. To interpret the output, let’s consider the case of age. Before matching, the mean for the treatment group was 24.17 whereas the mean for control group was 27.65. This difference is statistically significant (t-statistic equal to -26.52 and p-value of 0.00, so we reject the null hypothesis of equal means). After matching, the difference between this means reduces (24.77 for treatment and 25.34 for control group). There is an important reduction in terms of bias (83%). Despite this dramatic reduction in terms of bias, we still find that we reject the null hypothesis of equal means between treatment and control groups (t-test of -5.12). In this case, matching helps to reduce bias in age but it is not able to fully remove it.

We can proceed in the same way for all the other variables. For the case of family size, per-capita income and household head’s age and sex, we find that there is balance between treatment and control groups after matching. In these cases, matching was good enough to remove the existing differences between both groups before matching. Interestingly, we find that matching actually worsens the pre-existing balance in the unmatched sample. Before matching, the difference in means was not significant (t-statistic of 0.63) but after matching this difference becomes significant (t-statistic of -3.8).

Figure 3 offers a graphical representation of the previous exercise. The black dots represent the level of bias for the unmatched sample whereas the small exes represent the bias for the matched sample. Consistent with our previous discussion, there is a systematic reduction in bias after matching.

Figure 3: Balance for unmatched and matched sample

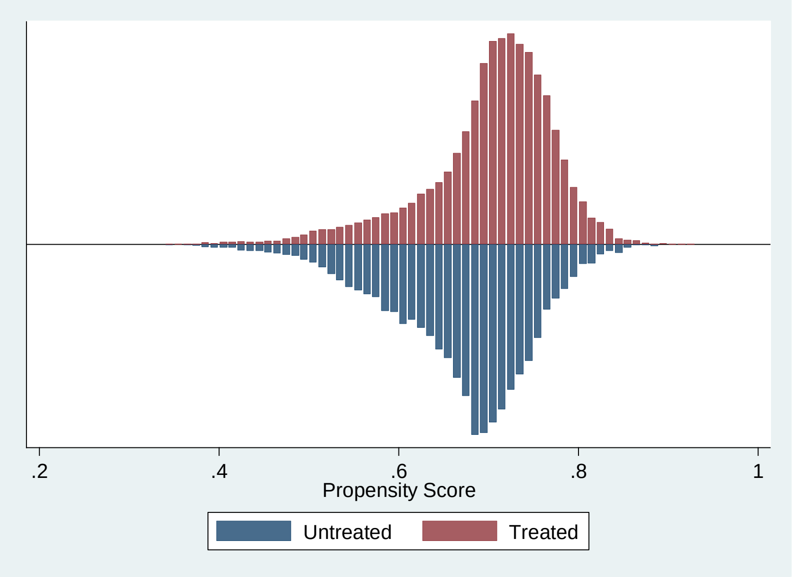


There are also graphical tools the common support using the estimated propensity score. One possibility is to use histograms of the propensity score for treatment and control groups. Then, the overlap of these histograms would give us an indication of the common support. The user-written program `psgraph`, a companion program of the `psmatch2` ado-file discussed above, offers a way to graph both histograms in the same graph. Using the propensity score previously estimated using `psmatch2 (_pscore)`, the code is the following:

```
psgraph, treated(D_HH) pscore(_pscore) bin(100)
```

We chose 100 bins. Figure 4 presents the results:

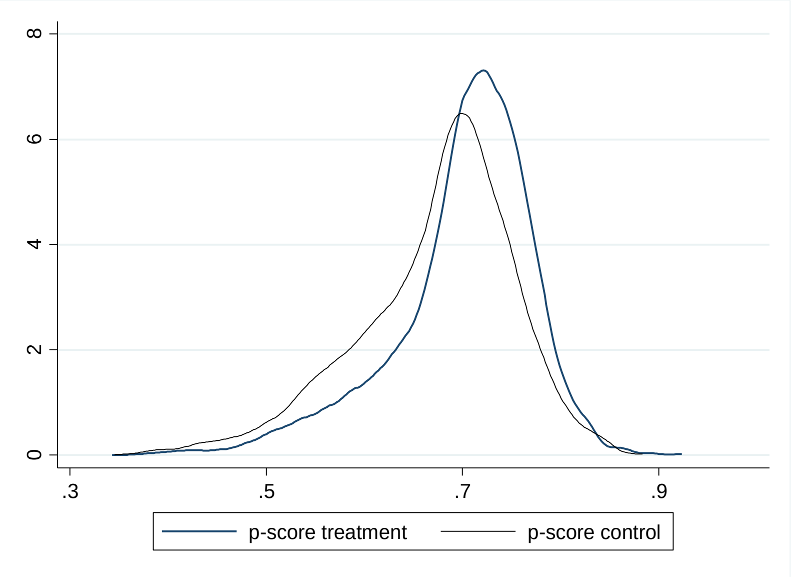
Figure 4: Histograms for the propensity score



We observe a significant overlap between the distributions for the propensity score for values between 0.4 and 0.85. This result is expected given the way in which the control group was constructed after the end of the PROGRESA experimental sample in 2000 which was precisely based on a matching approach. We can also use a non-parametric approach using kernel densities. The results are essentially the same. The code is the following:

```
twoway  (kdensity  _pscore  if  participation_03==1,  clwid(medium))  (kdensity
_pscore if participation_03==0, clwid(thin) clcolor(black)), ///
      xti("") yti("") title("") legend(order(1 "p-score treatment" 2 "p-score
control")) xlabel(0.3(.2)1) graphregion(color(white))
```

The resulting graph is the following:



To summarize the results of this section, it can be argued that we were able to reduce bias after the implementing the propensity score. Of course, it is important to keep in mind that matching helps to reduce bias due to observables. Still, we know that our sample is unbalanced in terms of age, although that is something that we can control by adding age in our DD model. In this sense, although in some dimensions the sample is still unbalanced, the combination of propensity score matching with DD is still robust to this fact if we can control for those observable dimensions in a regression framework.

5.5. Evaluating the impact of the intervention

We are now in the position of estimating the impact of the intervention for the matched sample with a DD design. After computing the propensity score with `psmatch2`, a new variable `_support` was created. This is an indicator variable equal to one for observations that belong to the common support. Therefore, we proceed by creating a new dataset with the sample in the common support. We merge this dataset to the original `Panel_OPORTUNIDADES_00_07_year.dta` and use the `_merge` variable to identify the matched sample. The code below performs these tasks:

```
* Merging observations of the common support
```

```
keep if _support==1

keep hogid2
sort hogid2

merge hogid2 using "$path/Panel_OPORTUNIDADES_00_07_year.dta"

tab _merge

gen matched=0
replace matched=1 if _merge==3
drop _merge
```

We need to create the auxiliary variables to identify the period and the interaction between treatment and period as above. The code is below:

```
* Creating additional variables

gen period=0 if year==2000
replace period=1 if year==2003

gen int_dhXperiod=period*D_HH
```

We can now implement the DD model in the matched sample defined by the estimated propensity score. We also compute the DD model for the unmatched sample for comparison. The code is below:

```
* Computing DD for standard DD and PSM DD

reg edu_child int_dhXperiod D_HH period, r
estimates store r1

reg edu_child int_dhXperiod D_HH period if matched==1, r
estimates store r2

xml_tab r1 r2, below stats(N r2)replace save("Table1.xml")
```

The output is the following:

```
. reg edu_child int_dhXperiod D_HH period, r

Linear regression                                      Number of obs =   82321
                                                       F(  3, 82317) =  249.47
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0089
                                                       Root MSE      =  2.6416

------------------------------------------------------------------------------
             |               Robust
   edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
int_dhXperiod |   .6422892   .0415167    15.47   0.000     .5609167    .7236618
        D_HH |  -.2759708   .0328308    -8.41   0.000    -.3403188   -.2116227
```

```
       period |   -.042306   .0351047    -1.21   0.228    -.1111111     .026499
        _cons |   4.067136   .0285133   142.64   0.000      4.01125    4.123022
------------------------------------------------------------------------------

. estimates store r1

.
. reg edu_child int_dhXperiod D_HH period if matched==1, r

Linear regression                                      Number of obs =   57788
                                                       F(  3, 57784) =  245.33
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0124
                                                       Root MSE      =  2.6536

------------------------------------------------------------------------------
              |               Robust
    edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
int_dhXperiod |     .43097   .0596055     7.23   0.000     .3141429     .547797
         D_HH |  -.2694897   .0354647    -7.60   0.000    -.3390007   -.1999787
       period |   .2252948   .0543454     4.15   0.000     .1187775     .331812
        _cons |   4.102166   .0309569   132.51   0.000     4.041491    4.162842
------------------------------------------------------------------------------

. estimates store r2
```

The first regression reports the result for the DD model for the full sample. The coefficient associated to the interaction is 0.64, significant at 1% significance level. When we focus on the matched sample, this coefficient is 0.43, also significant at 1% significance level. The evidence suggests that there was selection bias expressed in the dramatic reduction of the size of the coefficient. Since the matched sample has reduced bias in terms of baseline characteristics (or baseline heterogeneity), this result is expected.

6. Triple differences

Triple differences (DDD) is a natural extension to the basic DD analysis. Consider the case of PROGRESA. We know that the program only treats a fraction of those in treated villages. One might wonder whether the impact differ within villages and therefore proceed to estimate the impact of the intervention in this scenario. DDD simple extends the original DD to allow for more interaction to capture this potential differential response. The new equation would be the following:

(11) $Y_{ijt} = \alpha + \gamma\phi_i + \lambda t_t + \varphi c_j + \beta_1(\phi_i * t_t) + \beta_2(\phi_i * c_j) + \beta_3(c_{ji} * t_t) + \delta(\phi_i * c_{ji} * t_t) + \varepsilon_{it}$.

where y_{it} is the outcome of interest for unit i in period t. ϕ_i, c_j and t_t are respectively household, village and period fixed effects. The interaction $\phi_i * t_t$ captures changes over time for treatment households and the interaction $c_{ji} * t_t$ does the same for treatment villages. The interaction $\phi_i * c_j$ captures time invariant characteristics of treatment households in treatment villages. The triple interaction $\phi_i * c_{ji} * t_t$ captures all variation in the outcome related to treatment households in treatment villages after the implementation of the program. The coefficient for this triple interaction recovers the causal effect of interest. Finally ε_{it} is an error term.

To implement DDD in STATA, we need to create additional interaction between treatment at village level, treatment at household level and the period dummy. We also need to create the triple interaction among these variables to recover the impact of the intervention. The code is the following:

```
* Creating interactions

gen int_dXperiodXdh=period*D*D_HH

gen int_dXdh=D*D_HH

gen int_periodXdh=period*D_HH
```

We add these interactions to our basic DD model as follows:

```
* Regression

reg edu_child int_dXperiodXdh int_dXdh int_periodXdh int_dXperiod D period, r
```

The results are the following:

```
. reg edu_child int_dXperiodXdh int_dXdh int_periodXdh int_dXperiod D D_HH period,
r

Linear regression                                      Number of obs =   82321
                                                       F(  7, 82313) =  112.30
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0094
                                                       Root MSE      =  2.6411


------------------------------------------------------------------------------
                |               Robust
      edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
int_dXperiodXdh |   .4387161   .1205256     3.64   0.000     .2024868    .6749453
       int_dXdh |  -.0247702   .0682738    -0.36   0.717    -.1585863    .1090459
  int_periodXdh |   .3281588   .1081169     3.04   0.002     .1162505    .5400672
   int_dXperiod |   -.308184   .0724178    -4.26   0.000    -.4501224   -.1662456
              D |   .0866187   .0597012     1.45   0.147    -.0303952    .2036326
           D_HH |  -.2565798   .0544092    -4.72   0.000    -.3632215   -.1499382
         period |   .1215408   .0555608     2.19   0.029      .012642    .2304397
          _cons |   4.011057   .0480629    83.45   0.000     3.916854    4.105259
------------------------------------------------------------------------------
```

As discussed above, we pay attention to the triple interaction between D, D_HH and period to evaluate the impact of the intervention. This interaction captures the variation in children education related to treatment households (with respect to controls) in treatment villages (relative to control villages) after the introduction of the program (with respect to the initial period). In this case, this coefficient is 0.43 years of children education. The coefficient is strongly significant.

We can also add control variable to remove potential biases that can still persist. The STATA code is the following:

```
reg  edu_child  int_dXperiodXdh  int_dXdh  int_periodXdh  int_dXperiod  D  period
      age sex agehead sexhead, r
```

The output is below:

```
. reg edu_child int_dXperiodXdh int_dXdh int_periodXdh int_dXperiod D D_HH period
age sex agehead sexhead, r

Linear regression                                      Number of obs =   77661
                                                       F( 11, 77649) =11988.72
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6827
                                                       Root MSE      =  1.4905


------------------------------------------------------------------------------
                |               Robust
      edu_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
int_dXperiodXdh |   .2482412   .0707863     3.51   0.000     .1095004    .3869819
       int_dXdh |   .0227525   .0362631     0.63   0.530    -.048323     .0938281
  int_periodXdh |   .1450368   .0635909     2.28   0.023      .020399    .2696746
   int_dXperiod |  -.1953591    .042163    -4.63   0.000    -.2779984   -.1127198
              D |   .0813947   .0321427     2.53   0.011     .0183952    .1443942
           D_HH |  -.1157773   .0293439    -3.95   0.000    -.1732913   -.0582634
         period |   .2059721   .0317282     6.49   0.000     .1437849    .2681592
            age |   .7100122   .0020119   352.90   0.000     .7060688    .7139556
            sex |  -.0966853   .0106967    -9.04   0.000    -.1176507   -.0757199
        agehead |  -.0014299   .0004896    -2.92   0.003    -.0023895   -.0004703
        sexhead |   .0871413   .0209306     4.16   0.000     .0461175    .1281651
          _cons |  -3.803911   .0422029   -90.13   0.000    -3.886629   -3.721194
------------------------------------------------------------------------------
```

We observe an important drop in the size of the coefficient for the triple interaction, which suggests that the previous estimate was biased. This is consistent with the fact that all the control variables included in the specification are statistically significant.

7. Final remarks

We have covered the basics and some extension of DD designs. This technique is suitable for the case in which there exists only imperfect knowledge of assignment into treatment. Although we have covered some extensions, the literature about DD is large. Of particular interest are some recent techniques that analyze distributive impacts using DD designs (see for instance, Athey and Imbens 2006).

There are some aspects regarding DD that we were not able to cover due to data limitations. For instance, we did not cover specification checks based on placebo analysis of data before the implementation of the program. Since DD is based on the assumption that in the absence of treatment both groups behave in the same manner, one natural test to evaluate the validity of this design would be to create a pseudo-intervention for years previous to the implementation of the program. If the design is valid, we should observe no impact of this pseudo-intervention. Since our pre-treatment data contains household that were previously treated, we don’t have the ability of performing such as test. We recommend the reader to read McKinnish (2000) for details of this type of test. We also have not discussed aspects related to inference. See Bertrand et al (2004) for details.

8. Further readings

Athey, Susan and Guido Imbens (2006). “Identification and Inference in Nonlinear Difference-In-Differences Models,” Econometrica, 74, pp.431-497.

Bertrand, Marianne; Esther Duflo and Sendhil Mullainathan (2004). “How Much Should We Trust Differences-in-Differences Estimates?” Quarterly Journal of Economics, 119, 249-275.

Card, D. and A. Krueger (1994). “Minimum Wages and Employment: A Case Study of the Fast Food Industry”, American Economic Review, 84, 772-793.

Conley, Timothy and Christopher Taber (2011). “Inference with "Difference in Differences" with a Small Number of Policy Changes,” Review of Economics and Statistics, 93, 113-125.

Donald, S. G. and K. Lang (2007). “Inference with Difference-in-Differences and Other Panel Data”, Review of Economics and Statistics, 89, 221-233.

Hansen, Christian (2007). “Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects,” Journal of Econometrics, 140, 670-694.

Heckman, James and Jeffrey Smith (1999). “The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies,” Economic Journal, 109, 313-348.

McKinnish, T. (2000). “Model Sensitivity in Panel Data Analysis: Some Caveats About the Interpretation of Fixed Effects and Differences Estimators”, Mimeo, University of Colorado.

Meyer, Bruce (1995). “Natural and Quasi-Natural Experiments in Economics”, Journal of Business and Economic Statistics, 13, 151-161.

Rosenzweig, Mark and Kenneth Wolpin (2000). “Natural "Natural Experiments" in Economics”. Journal of Economic Literature, 38, 827-874.