Chapter 4: Problems with Experiments

Pre-requisites

- Chapter 1: Intro to STATA
- Chapter 2: Review of Regression
- Chapter 3: Experiments

Contents

1. Introduction

In Chapter 3 we introduced the basics of experimental designs. We showed that, under ideal conditions, experimental designs provide a solution to the selection bias problem that is common in empirical research. In this scenario, a simple comparison of means between treatment and control units is usually enough to provide an unbiased estimate of the impact of an intervention.

However, experiments are not free of problems. Sometimes researchers fail to correctly randomize treatment across units and some imbalance persists. We discussed in the previous chapter how some methodological choices might be more efficient in dealing with this issue and how ex-post corrections using covariates can be helpful in mitigating it. So, we mostly covered this issue in the past chapter. A more difficult case in when units fail to comply with treatment protocol. In many social programs, individuals may refuse the assigned treatment. In the case of PROGRESA, individuals assigned to control group may find a way to get access to benefits using migration to a beneficiary village as a strategy. Since those who decide to move might differ from those who stay, this would introduce bias in the estimation of the causal effect. In this scenario, the assigned treatment status differs from the actual treatment status, an issue that would introduce some complexities in the empirical analysis as we will see in this chapter. This issue is known as **imperfect compliance**.

A related issue show up in the case of experiments in which baseline and follow-up data is available. Some units may drop out the sample after the implementation of the evaluation for reasons that might be related with the program. In the context of PROGRESA migration is again a good example. We selected the sample before the intervention but there is the chance that the cash transfer helps some families to migrate to urban areas. As a consequence, the original balance between treatment and control groups before the intervention is no longer guaranteed by the research design. This problem is known as **attrition**.

In some cases, the experiment itself can modify the behavior of units under study. If parents from non-poor households in villages treated by PROGRESA are aware that their children are no part of the program, they may develop a stronger interest on the academic performance of their children and provide more educational inputs to them. Although these households from PROGRESA villages do not receive benefits from the program, they do change their behavior in response to their original treatment assignment. As a consequence, a comparison between treatment and control units would be biased. This problem is known as **experimental effects**. When the behavior of the control group is affected as in the case of the example above, this effect is known as **John Henry effect**. When this happens in the treatment group, this is known as the **Hawthorne effect**.

One key assumption in the previous chapter was **SUTVA** (stable unit value assumption). This basically means that the treatment status of one unit should not affect the outcome of other unit. SUTVA is violated in cases in which externalities or general equilibrium effects are present. In the case of PROGRESA, externalities are present when the better health conditions of poor children due to their participation into the program reduce the infection rates or disease transmission to non-poor children from the same village. Therefore, comparing health outcomes between treatment and control children is subject to bias due to the presence of health externalities. On the other hand, general equilibrium effects might be present if the fraction of PROGRESA beneficiaries in a treatment village is high enough to affect relative prices. For instance, if half of the families in a village are PROGRESA recipients, it is possible that inflation might be an issue. This increase of prices affects all the families in the village.

In this chapter, students will learn how to deal with experiments that do not go as perfect as we planned. Specifically, we will cover three cases:

a) We will study a situation that arises when the treatment status does not apply to all of those that were initially assigned. This occurs when there is **imperfect compliance** to the random assignment process. In this case, those individuals that were initially assigned to the control group but are effectively being treated could be different from those that maintain their original treatment status, creating bias in the estimate of the causal relationship of interest.

b) We then study a situation in which there exist **externalities** from the treated to the control group. This happens when, for instance, untreated individuals interact with treated ones and by that way receive part of the benefits of treatment, causing a downward bias in the treatment effect estimate.

c) We finally cover a situation when our impact evaluation design considers a baseline and follow-up surveys, but from the second period we lose information from a subgroup of individuals. If the loss of information of these individuals is correlated with the outcome variable, treatment effect estimators could be biased. **Attrition** is a common issue in experimental and non-experimental designs.

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

- Implement instrumental variable solutions to solve imperfect compliance problems.
- Incorporate alternatives estimators that take into account externalities.
- Implement corrections for attrition in unbalanced panel dataset using re-weighting techniques and Lee's (2009) bounding methods.

2. Imperfect compliance

2.1. Why imperfect compliance matters?

Noncompliance arises when actual treatment status differs from treatment assignment. It happens -for example- when some individuals from the treatment group do not make use of the benefits of this status, or when some individuals from the control group get the benefits of the treatment even when they should not do according to the random assignment. If this happens, the comparison between actual treated and untreated individuals (by using an OLS regression, for example) should cause a bias in the estimates because both groups would differ in the unobservable dimensions. The control group no longer recovers the counterfactual (what would have happened to treatment group in absence of the program). In this scenario, there is a difference between the original **assignment to treatment** by the program and the **actual treatment status**. The random mechanism used to assign treatment that we studied in the previous chapter only identify the ideal allocation to the program and this variable may differ in significant ways with the actual treatment status. Thus, in this scenario it is necessary to implement a solution that corrects the estimates from the "non-compliance" behavior of individuals. The standard way to perform such correction is via **instrumental variables** techniques (IV).

Recall the basic notation from chapter 3. We defined a variable D_i as the treatment or potential cause and assumed a simple binary case in which $D_i = 1$ if individual i has been exposed to treatment and $D_i = 0$ if individual i has not been exposed to treatment. $Y_i(D_i)$ was defined as the outcome or the effect we want to attribute to the treatment, being $Y_i(1)$ is the outcome in case of treatment and $Y_i(0)$ is the outcome in case of no treatment. The individual outcomes were written as follows:

$$(1)\ Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

In this scenario, we assumed that all of those who were offered the treatment were effectively treated. We need to make now the distinction in order to understand how non-compliance works. To do that, we introduce a variable Z that reflects the **assignment to treatment status**. For simplicity, we assume Z is dichotomous, so $Z = 0$ if the individual is assigned to control group or $Z = 1$ if she or he is assigned to treatment group. Now, the treatment status or **participation into treatment** is defined as $D_i = D_i(Z)$. Therefore, the assignment to treatment Z affects the actual treatment status D but in presence of imperfect compliance there is no reason to assume that all of those who were offered treatment actually get it. Finally, the outcome of interest is a function of Z and D as follows: $Y_i = Y_i(Z, D)$.

Before moving into the conceptual details behind non-compliance, it is worth exploring the details of the datasets we will be using for the rest of the chapter. For this section, we will be using the `PanelPROGRESA_97_99year.dta` dataset. To illustrate the issue, we will be paying attention to the outcome enrollment with the goal of understanding the role of PROGRESA in improving educational outcomes among its beneficiaries. For simplicity, we assume that PROGRESA randomly selects villages and that all families in selected villages are offered the program. Some households accept to participate into the program and others reject the offer. Notice that this is not the actual case since the program only targets poor households within the selected villages. We assume this only for pedagogical purposes.

To open the database, we use the following STATA code:

```
set more off
clear all

global  path="C:\Users\Stanislao\Dropbox\Teaching\1.  Current\Econometrics\4.
Handbook\2. Data\"

use"$path/PanelPROGRESA_97_99year.dta", clear
```

We are interested in the variable `enroll`, the outcome of interest. We are also interested in the treatment status. Given our previous discussion, we need to distinguish between the actual treatment status D and the treatment assignment Z. In our dataset, the **actual treatment status is given by `D_HH`** (an indicator of whether the household in a PROGRESA beneficiary) whereas –given the simplification assumptions discussed above- the **assignment into treatment is given by the variable `D`** (an indicator of whether the village was selected by the program as a treated village). Since this variable recovers treatment status at village level, it implies –in our fictitious example- that all households in the village has been offered access to PROGRESA, so this variable is equal to 1 for all households in the same village. However, it is possible that households in a treated village (with `D=1`) refuse the offer (having a `D_HH=0` as a consequence).

Although non-compliance can be analyzed in the context of panel data, we keep for simplicity only observations for 1999 to exploit the cross-sectional data for that year. We also focus our interest in children between 6 and 7 years old. We also create some auxiliary variables to define fixed effects for different levels of governments. The STATA code is the following:

```
keep if year==1999

* Keep information of children 6-7 years
```

```
keep if age>=6 & age<=7

* Creating Region Fixed Effects
gen entidad=substr(villid,1,2)
gen munici =substr(villid,3,3)
gen locali =substr(villid,6,3)
destring  entidad munici locali, replace
```

Let’s use some basic regressions to illustrate the distinction between actual treatment status and assignment to treatment discussed above. The code is the following:

```
areg enroll D_HH, absorb(entidad) vce(cluster villid)

areg enroll D, absorb(entidad) vce(cluster villid)
```

We use a simple fixed effects regression as the one discussed in the final part of the previous chapter clustering the standard errors at village level and using region fixed effects. In the first case, the actual treatment status is used as treatment variable whereas in the second it is the assignment to treatment the variable used as treatment. The results are below:

```
. areg enroll D_HH, absorb(entidad) vce(cluster villid)

Linear regression, absorbing indicators           Number of obs   =       6534
                                                  F(   1,    495) =       1.46
                                                  Prob > F        =     0.2268
                                                  R-squared       =     0.0032
                                                  Adj R-squared   =     0.0021
                                                  Root MSE        =     0.1262

                               (Std. Err. adjusted for 496 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |   .0044922    .003712     1.21   0.227     -.002801    .0117855
       _cons |   .9816775     .00283   346.88   0.000     .9761171    .9872379
-------------+----------------------------------------------------------------
     entidad |   absorbed                                       (7 categories)
```

We do not find a relationship between the actual treatment status and enrollment. Now, let’s analyze the results of regressing enrollment on assignment to treatment. The results are below:

```
. areg enroll D, absorb(entidad) vce(cluster villid)

Linear regression, absorbing indicators           Number of obs   =       6800
                                                  F(   1,    496) =       3.87
                                                  Prob > F        =     0.0498
                                                  R-squared       =     0.0040
                                                  Adj R-squared   =     0.0030
                                                  Root MSE        =     0.1254

                               (Std. Err. adjusted for 497 clusters in villid)
```

```
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0080458   .0040909     1.97   0.050     8.22e-06    .0160833
       _cons |   .9791124   .0035052   279.33   0.000     .9722255    .9859992
-------------+----------------------------------------------------------------
     entidad |   absorbed                                       (7 categories)
```

The results differ from the previous case. Now the relationship is significant from a statistical point of view and the coefficient is higher. These results are robust to the inclusion of additional control variables as it can be seen below. The code is the following:

```
areg enroll D_HH famsize langhead sexhead agehead, absorb(entidad)
vce(cluster villid)

areg enroll D famsize langhead sexhead agehead, absorb(entidad) vce(cluster
villid)
```

The results are below:

```
. areg enroll D_HH famsize langhead sexhead agehead, absorb(entidad) vce(cluster
villid)

Linear regression, absorbing indicators           Number of obs   =       6212
                                                  F(   5,    495) =       1.72
                                                  Prob > F        =     0.1292
                                                  R-squared       =     0.0057
                                                  Adj R-squared   =     0.0039
                                                  Root MSE        =     0.1268

                                  (Std. Err. adjusted for 496 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |   .0052632   .0039748     1.32   0.186    -.0025463    .0130727
     famsize |  -.0020047   .0007837    -2.56   0.011    -.0035445    -.000465
    langhead |  -.0046998   .0052301    -0.90   0.369    -.0149758    .0055762
     sexhead |   .0104189   .0074656     1.40   0.163    -.0042493    .0250871
     agehead |  -.0000295   .0001523    -0.19   0.847    -.0003288    .0002698
       _cons |   .9884432   .0112116    88.16   0.000      .966415    1.010471
-------------+----------------------------------------------------------------
     entidad |   absorbed                                       (7 categories)

.
. areg enroll D famsize langhead sexhead agehead, absorb(entidad) vce(cluster
villid)

Linear regression, absorbing indicators           Number of obs   =       6229
                                                  F(   5,    495) =       2.01
                                                  Prob > F        =     0.0765
                                                  R-squared       =     0.0063
                                                  Adj R-squared   =     0.0045
                                                  Root MSE        =     0.1266

                               (Std. Err. adjusted for 496 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0082069   .0044702     1.84   0.067    -.0005759    .0169898
     famsize |  -.0019104   .0007732    -2.47   0.014    -.0034296   -.0003912
    langhead |  -.0044413   .0051522    -0.86   0.389    -.0145642    .0056816
     sexhead |   .0103154   .0074075     1.39   0.164    -.0042386    .0248695
     agehead |   -.000045    .000151    -0.30   0.766    -.0003416    .0002516
       _cons |   .9860223   .0113163    87.13   0.000     .9637884    1.008256
-------------+----------------------------------------------------------------
     entidad |   absorbed                                       (7 categories)
```

At this point you may ask the reason of the difference between these results and what it means. The simple answer is that we are estimating different parameters in each case. In the first regression, we have estimated **the impact of the program on those who actually take the treatment** whereas in the second case we have estimated t**he impact on those who were offered the program regardless whether they take it or no**. We will introduce some additional terminology to conceptually differentiate between these two estimates.

With this evidence, those interested on the impact of the program on those treated will likely conclude that PROGRESA has increased in a very modest way the enrollment rate of children (less than 1 percent point). However, this may be wrong conclusion if we do not take into account that individuals differ in terms of compliance. To see that, let’s use the command tab for **participation into treatment** and **treatment assignment**:

```
Village-Le |

       vel |    Household-Level

 Treatment |   Treatment status

    status |         0          1 |     Total

-----------+----------------------+----------

   Control |     2,987          0 |     2,987

   Treated |       917      3,300 |     4,217

-----------+----------------------+----------

     Total |     3,904      3,300 |     7,204
```

According to the original assignment, about 4,217 children belong to households from villages assigned to the program. However, 917 belong to households that refused treatment. Probably, those children from households that changed their treatment status would have been enrolled in school no matter the exigencies of PROGRESA if they had been beneficiaries. If this were true, then the simple comparison through OLS would be pulling down the estimated parameter of the impact of PROGRESA. We will cover in the next sections techniques to address this issue.

2.2. Treatment on the treated (ATT) and intention to treat (IT)

We have shown so far that, when some individuals do not follow their treatment assignment, then the estimation of the impact of the program is more complex. With perfect compliance (all participants do what the program tells them to do), the assignment to treatment is the same as the actual treatment ($D = Z$), then a simple difference of means recovers the impact of the program D on the outcome of interest Y for the case of an experimental design. When this condition is no longer true, then the impact of D on Y (the impact of the program on those actually treated) is likely to differ to the impact of Z on Y (the impact of the program on those who were offered treatment).

The impact of Z on Y is usually called **intention to treat (ITT)**. Formally,

(2) $$ITT = E\left[\Delta_i / Z_i = 1\right] = E\left[Y_i(1) - Y_i(0) / Z_i = 1\right] = E\left[Y_i(1) / Z_i = 1\right] - E\left[Y_i(0) / Z_i = 1\right]$$

This is essentially a similar parameter that the average treatment effect on the treated (ATET) discussed in the previous chapter with the difference that is defined over the assignment to treatment Z rather than actual treatment status D. In many fields like education (see, for instance, Schochet 2009 for a discussion) researchers usually report the ITT estimates of the intervention. This is the case for interventions were enforcement is hard to pursue. For instance, governance programs that invites citizens to attend community meetings to improve advocacy and accountability are usually prone to low response rates, so effectiveness analysis for this kind of programs needs to take into account this low response rates. However, for policy purposes, it is usually the case that ATT is more interesting. In this context, ATT represents an average effect for those who comply with treatment. This group is usually known as **compliers**. More broadly, it is fair to say that both parameters are interesting and their combined use can be extremely useful to improve program implementation. The ATT can be useful to distinguish whether a low effect of an intervention is due to low compliance rates or a consequence of small treatment effects among compliers (Schochet, 2009).

We already compute the ITT for enrollment. We replicate the results below:

```
. areg enroll D, absorb(entidad) vce(cluster villid)

Linear regression, absorbing indicators           Number of obs   =       6800
                                                  F(   1,    496) =       3.87
                                                  Prob > F        =     0.0498
                                                  R-squared       =     0.0040
                                                  Adj R-squared   =     0.0030
                                                  Root MSE        =     0.1254

                               (Std. Err. adjusted for 497 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0080458   .0040909     1.97   0.050     8.22e-06    .0160833
```

```
       _cons |   .9791124   .0035052   279.33   0.000     .9722255    .9859992
-------------+----------------------------------------------------------------
     entidad |   absorbed                                       (7 categories)
```

We find a modest but significant effect of assignment to treatment on enrollment. Now the question is how to estimate ATT from ITT. We need to introduce some additional concepts before proceeding.

Individuals can differ in their responses to their original treatment assignment. Angrist et al (1996) classify individuals into four mutually exclusive categories: compliers, defiers, always-takers and never-takers. A **complier (c)** is someone that accepts the treatment status offered to him/she ($D_i(1)=1$ or $D_i(0)=0$) whereas a **defier (d)** does the opposite to his/her treatment assignment ($D_i(1)=0$ or $D_i(0)=1$). An **always-taker (a)** will always takes treatment whatever his/her original assignment ($D_i(1)=1$ or $D_i(0)=1$) whereas a **never-taker (n)** always refuses treatment no matter what his/her original assignment to treatment is ($D_i(1)=0$ or $D_i(0)=0$).

Although this formulation is very simple, it is important to keep in mind that these groups are not observable without additional restrictions. We only know that the ITT parameter recovers an average effect for these groups. Formally,

(3) $ITT=\delta_c ITT_c+\delta_d ITT_d+\delta_n ITT_n+\delta_a ITT_a$.

where δ is the fraction of a group in the population and $\sum \delta=1$. Angrist et al (1996) proposed a set of assumptions to rule out the existence of defiers, never-takers and always-takers. These assumptions are presented in Box 1. If these assumptions are valid, then ATT can be identified for a sub-population of compliers. Therefore,

(4) $ATT=ITT_c=E[Y_i(1,1)-Y_i(0,0)]=\dfrac{ITT}{\delta_c}$.

Box 1: Identification Assumptions for Complier Average Effect

Angrist et al (1996) show that the average effect for compliers can be identified under the following assumptions:

Stable Unit Treatment Value Assumption (SUTVA). This requires that potential outcomes and treatments for a particular unit are independent of assignments, treatments and outcomes of a different unit.

Monotonicity. It basically implies that no unit does the opposite to his/her original assignment regardless what the assignment is.

Exclusion Restriction. The assignment to treatment affects the outcome only through the treatment.

We will discuss how SUTVA matters in the next section. The intuition is quite simple thought: if SUTVA is violated, the outcomes of treated individuals can be affected by the treatment status of non-treated individuals or viceversa. Therefore, a comparison of means between treated and control groups are going to provide biased estimates of the impact of the intervention. Monotonicity basically helps to rule out the existence of difiers. The exclusion restriction is a typical assumption for instrumental variables and guarantees that the assignment plays only a role in terms of influencing the probability of being treated without having a direct effect on the outcome of interest.

An alternative approach is to think this problem under an instrumental variable framework. We want to estimate the impact of actual treatment (being a recipient of PROGRESA transfers) on our outcome of interest (enrollment). However, we know that the actual treatment status is endogenous due to non-compliance. In this case, we need a source of exogenous variation in actual treatment status that can be exploited in order to recover a causal relationship between the actual treatment and the outcome of interest. This source of variation is known as an "instrument". A good instrument is one that is correlated with the treatment status (satisfies the relevance condition) but not with the outcome of interest (satisfies the exclusion restriction). In this context, a good instrument is the random assignment to treatment `D`. As discussed above, this variable is equal to 1 for all households in a treated village. This implies that for all households in a village the program is offered. Since this is done in a random manner, those who are offered the program are statistically similar to those who were assigned to the comparison group. This guarantees that the assignment to treatment is balanced between those assigned to treatment and comparison groups.

2.3. Local average treatment effect (LATE)

So far, we have assumed that nobody in the control group is treated, so IV recovers ATT. Angrist et al (1996) show that, when this is no longer true, IV estimates an average effect for those who were induced by the instrument to take the treatment. This parameter is known as the **Local Average Treatment Effect (LATE)**. Therefore, when random assignment to treatment is imperfectly related to the actual treatment status, it is still possible to estimate a causal relationship between treatment and the outcome of interest, although this rarely recovers an average effect for the entire population. Instead, we are usually only able to estimate an average effect for a sub-population: compliers.

Implementing these estimators in STATA is straightforward. The basic command is `ivregress.` Notice that the endogenous variable, `D_HH`, and its instrument, `D`, have been written in parenthesis. After writing this command, we ask STATA to implement the Two Stage Least Squares (2SLS) estimator by using the option `2sls`. IV is just a special case of 2SLS when one endogenous variable and one instrument are available. IV can be estimated using alternative econometric techniques such as GMM or Maximum Likelihood. The code is the following:

```
tab entidad, gen(enti)

global entidad="enti2 enti3 enti4 enti5 enti6 enti7"

* IV
ivregress 2sls enroll (D_HH=D) $entidad, vce(cluster villid)

ivregress 2sls enroll (D_HH=D) sex langhead agehead sexhead $entidad, 
vce(cluster villid)
```

We have created regional fixed effects and run two specifications: a basic one with only fixed effects and one extended with socio-economic controls. The results for the basic regression are reported below:

```
. * IV
. ivregress 2sls enroll (D_HH=D) $entidad, vce(cluster villid)

Instrumental variables (2SLS) regression              Number of obs =     6534
                                                       Wald chi2(7)  =    13.93
                                                       Prob > chi2   =   0.0524
                                                       R-squared     =   0.0028
                                                       Root MSE      =   .12616

                               (Std. Err. adjusted for 496 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |   .0096473   .0051515     1.87   0.061    -.0004494     .019744
       enti2 |    .017795   .0097324     1.83   0.067    -.0012801    .0368702
       enti3 |    .012636   .0096392     1.31   0.190    -.0062565    .0315284
       enti4 |   .0103974   .0098714     1.05   0.292    -.0089502    .0297451
       enti5 |   .0060472   .0143203     0.42   0.673    -.0220201    .0341145
       enti6 |   .0214753   .0090739     2.37   0.018     .0036907    .0392599
       enti7 |   .0214444   .0089038     2.41   0.016     .0039933    .0388955
       _cons |   .9640773   .0093859   102.72   0.000     .9456814    .9824732
------------------------------------------------------------------------------
Instrumented:  D_HH
Instruments:   enti2 enti3 enti4 enti5 enti6 enti7 D
```

According to these results, the impact of PROGRESA is around 1 percentage points, significant at 10%. These results are essentially the same when controls are added to the basic specifications (results no reported). Consequently, we have estimated a positive relationship between actual treatment and the outcome of interest. Notice that these results differ from the previous evidence in which there was no evidence of impact of actual treatment on enrollment. This illustrates the idea of heterogeneity: whereas on average there is no direct impact due to the endogenous nature of actual treatment, we do find evidence of impact only for those who take the treatment due to exposure to the program.

One critical condition for IV to work is the existence of a strong relationship between the instrument (the assignment to treatment) and the endogenous variable (treatment status). This is known as the **first stage**. STATA report the first stage using the option `first` after a comma. As it is shown, the parameter associated to `D` in the first stage is positive and statistically significant. The first stage is strongly significant as shown below:

```
. * FIRST STAGE
. ivregress 2sls enroll (D_HH=D) $entidad, vce(cluster villid) first

First-stage regressions
-----------------------


                                                  Number of obs    =       6534
                                                  N. of clusters   =       1785
                                                  F(    7,    6526) =     1008.34
                                                  Prob > F          =      0.0000
                                                  R-squared         =      0.6189
                                                  Adj R-squared     =      0.6185
                                                  Root MSE          =      0.3082

------------------------------------------------------------------------------
             |               Robust
        D_HH |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       enti2 |   .0154049   .0495873     0.31   0.756    -.0818024    .1126123
       enti3 |   .0808586   .0434532     1.86   0.063    -.0043239     .166041
       enti4 |   .0459778   .0472933     0.97   0.331    -.0467325    .1386881
       enti5 |   .0000696   .0455824     0.00   0.999    -.0892869    .0894261
       enti6 |   .0343423    .044758     0.77   0.443    -.0533981    .1220826
       enti7 |   .0728169   .0425116     1.71   0.087    -.0105199    .1561536
           D |   .8020236   .0113574    70.62   0.000     .7797594    .8242878
       _cons |  -.0481402   .0399688    -1.20   0.228    -.1264922    .0302117
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression               Number of obs =     6534
                                                       Wald chi2(7)  =    13.93
                                                       Prob > chi2   =   0.0524
                                                       R-squared     =   0.0028
                                                       Root MSE      =   .12616

                               (Std. Err. adjusted for 496 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        D_HH |   .0096473   .0051515     1.87   0.061    -.0004494     .019744
       enti2 |    .017795   .0097324     1.83   0.067    -.0012801    .0368702
       enti3 |    .012636   .0096392     1.31   0.190    -.0062565    .0315284
       enti4 |   .0103974   .0098714     1.05   0.292    -.0089502    .0297451
       enti5 |   .0060472   .0143203     0.42   0.673    -.0220201    .0341145
       enti6 |   .0214753   .0090739     2.37   0.018     .0036907    .0392599
       enti7 |   .0214444   .0089038     2.41   0.016     .0039933    .0388955
       _cons |   .9640773   .0093859   102.72   0.000     .9456814    .9824732
------------------------------------------------------------------------------
Instrumented:  D_HH
Instruments:   enti2 enti3 enti4 enti5 enti6 enti7 D
```

There is an alternative way to compute LATE manually. Recall that the LATE parameter is composed by two terms: the ITT (the numerator of the former formula), which gives the impact of *assignment on the outcome*, and the compliance rate (the denominator of the former formula), which gives the change in the proportion treated owing to the change in assignment. The code below provides a manual approach to compute LATE:

```
INTENTION TO TREAT

* Intention to treat
reg enroll D $entidad, vce(cluster villid)
scalar ITT=_b[D]

* Compliance rate
* E[D_HH|D==1]
summ D_HH if D==1, meanonly
scalar D_HHD1=r(mean)
disp D_HHD1

* E[D_HH|D==0]
summ D_HH if D==0, meanonly
scalar D_HHD0=r(mean)
disp D_HHD0

scalar CR = D_HHD1-D_HHD0
disp CR

* LATE
scalar LATE=ITT/(D_HHD1-D_HHD0)
disp LATE
```

We have estimated the ITT and the compliance rates in order to compute the LATE parameter. Recall that in this context LATE is equal to the ATT. The results are reported below:

```
. * INTENTION TO TREAT
. reg enroll D $entidad, vce(cluster villid)

Linear regression                                      Number of obs =    6800
                                                       F(  7,   496) =    2.17
                                                       Prob > F      =  0.0351
                                                       R-squared     =  0.0040
                                                       Root MSE      =  .12541

                               (Std. Err. adjusted for 497 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      enroll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0080458   .0040909     1.97   0.050     8.22e-06    .0160833
       enti2 |   .0187018   .0098928     1.89   0.059    -.0007351    .0381387
       enti3 |   .0131712   .0099341     1.33   0.185    -.0063469    .0326894
       enti4 |   .0119691   .0100129     1.20   0.233    -.0077038    .0316421
       enti5 |   .0070125   .0142371     0.49   0.623    -.0209599     .034985
       enti6 |   .0224098   .0093339     2.40   0.017     .0040709    .0407488
       enti7 |   .0228149   .0092064     2.48   0.014     .0047266    .0409033
       _cons |     .96301   .0096454    99.84   0.000     .9440591     .981961
------------------------------------------------------------------------------

. scalar ITT=_b[D]

. * Compliance rate
. * E[D_HH|D==1]
. summ D_HH if D==1, meanonly

. scalar D_HHD1=r(mean)
. disp D_HHD1
.78254683

. * E[D_HH|D==0]
. summ D_HH if D==0, meanonly

. scalar D_HHD0=r(mean)

. disp D_HHD0
0

. scalar CR = D_HHD1-D_HHD0

. disp CR
.78254683

. * LATE
. scalar LATE=ITT/(D_HHD1-D_HHD0)

. disp LATE
.01028151
```

We have computed the ITT as before. The compliance rates for those who were offered the treatment in treated villages is 0.78 whereas it is 1 for those in control villages (no one in the control villages get treatment). The estimated LATE is 0.01, close to the estimate we got using `ivregress`.

3. Spillover Effects

SUTVA is an assumption that guarantees consistency of OLS under perfect compliance or that of IV under imperfect compliance. However, we should recognize that this assumption can fail in some situations. For instance, one in which the treatment impacts control units. Suppose, for example, that a conditional cash transfer is assigned selecting first some villages at random and then some poor households in the villages (as PROGRESA does). Then, the program requires parents to keep their children in school to continue being beneficiaries after the first transfer. This kind of treatment could spill its impact over non-beneficiaries in treated villages if parents from untreated households change their value of education and decide to maintain their children in school too as a response to the behavior of beneficiary parents.

Now, suppose that we do not take into account these potential externalities and use non-treated households from the same village as controls (for this to make sense, let suppose that the distinction of beneficiaries and controls was at random as well). If we naively do this comparison, we could get an estimate of the impact that is downward biased because of the externalities: since children from control households would increase their enrollment rate as a consequence of treatment, their outcome could not replicate the treated children's outcome in case of no treatment.

How can we design and manage experiments that are exposed to spillover effects? Some authors suggest that one should define the "geographic" levels at which externalities take place. This geographic level in the former example is the village, but could also be the school or even the household. Then, randomization should be made between the geographic levels selected and not within these units. This means that the comparison should be made between treated children from treated villages and untreated but eligible children from untreated villages. This procedure will allow us to estimate the impact of treatment among treated. On the other hand, to estimate the spillover effect we should compare untreated individuals from treated villages with ineligible individuals from unselected villages. This will allow us to estimate the impact of the treatment on untreated.

To illustrate the analysis of externalities, we will use the database `PanelPROGRESA_Enrollment_97_99.dta` we have created from ENCEL 1997-1999. This is a database with panel form that presents longitudinal information from a sample of children for these three years and was designed to focus on exploring enrollment with the goal to approximate the Bobonis and Finan's (2009) paper on externalities in PROGRESA.

To begin the exercise, let's open the database. We also need to create some auxiliary variables that we will need for the rest of the analysis. The code is below:

```
use"$path/PanelPROGRESA_Enrollment_97_99.dta", clear

destring year, replace

* Creating some auxiliary variables

gen aux=D if year==1998
egen D_assig=mean(aux),by(villid)

gen aux1=pov_HH if year==1998
egen aux2=mean(aux1),by(hogid)
replace pov_HH=aux2 if year==1997
drop aux1 aux2

gen Y8=year==1998
gen Y9=year==1999
```

We have created a new variable `D_assig` and change the variable `pov_HH` to have the same value in 1997 as the value in 1998. The original variables have missing values for 1997 since no treatment takes place in this year. We proceed in this way since we need the values for 1997 to reflect the original assignment to treatment at village and household level.

The goal of the exercise is to estimate the impact of PROGRESA on enrollment among treated comparing the outcomes of poor (it means, eligible) children in control and treated villages. Similarly, the externality will be estimated comparing the outcome of non-poor (it means, non-eligible) children in control and treated villages. We will compute these estimates for different models.

Before running regressions, it is important to clarify what we are estimating and how spillovers effects can be recovered from the data. Now, we need to consider that we have eligible and non-eligible households/individuals in our dataset. We define a new variable E that reflects the eligibility status of households or individuals. For simplicity, we assume E is a dummy variable equal to 1 for those eligible and 0 otherwise. Then, we redefine the ATT parameter as follows:

(5) $ATT = E\left[Y_i(1) - Y_i(0)/D_i = 1, E_i = 1\right] = E\left[Y_i(1)/D_i = 1, E_i = 1\right] - E\left[Y_i(0)/D_i = 1, E_i = 1\right]$

This is just the ATT defined over the eligible group. We can define also a similar parameter for the non-eligible group. Some scholars call this parameter the Indirect Treatment Effect (ITE).

(6) $ITE = E\left[Y_i(1) - Y_i(0)/D_i = 1, E_i = 0\right] = E\left[Y_i(1)/D_i = 1, E_i = 0\right] - E\left[Y_i(0)/D_i = 1, E_i = 0\right]$

These parameters will be the focus of our interest. The simplest specification is to run separate cross-section regressions for eligible and non-eligible households to recover ATT and ITE. We choose 1999 as the year of reference. The STATA code is the following:

```
reg enroll D if pov_HH==1 & year==1999, vce(cluster villid)

reg enroll D if pov_HH==0 & year==1999, vce(cluster villid)
```

The results are reported below:

```
. reg enroll D if pov_HH==1 & year==1999, vce(cluster villid)

Linear regression                                      Number of obs =    6268
                                                       F(  1,   482) =    8.76
                                                       Prob > F      =  0.0032
                                                       R-squared     =  0.0050
                                                       Root MSE      =  .49861

                               (Std. Err. adjusted for 483 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0731039   .0246976     2.96   0.003     .0245758    .1216321
       _cons |   .4690832   .0190887    24.57   0.000     .4315758    .5065905
------------------------------------------------------------------------------

. reg enroll D if pov_HH==0 & year==1999, vce(cluster villid)

Linear regression                                      Number of obs =    3810
                                                       F(  1,   441) =    2.62
                                                       Prob > F      =  0.1059
                                                       R-squared     =  0.0021
                                                       Root MSE      =  .49959

                               (Std. Err. adjusted for 442 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0465157   .0287132     1.62   0.106     -.009916    .1029473
       _cons |   .4687701   .0231031    20.29   0.000     .4233643    .5141759
------------------------------------------------------------------------------
```

We found a positive impact of PROGRESA in enrollment for those eligible as expected and a smaller impact for those non-eligible, although the coefficients are marginally no significant.

You may think that computing these estimates using separate regressions is inefficient. We can use interactions to accommodate these results in a single regression.

(5) $Y = \alpha_1 + \alpha_2 D + \alpha_3 PovHH + \alpha_4 (PovHH * D) + \mu$

This specification includes an interaction between the treatment assignment at village and household level. We create the interaction as follows:

```
gen int_D_PovHH=D_assig*pov_HH
```

How to identify ATT and ITE from the coefficients of this regression? Recall that under random assignment to treatment the following is true:

(6)
$$\begin{aligned} ATT &= E\left[Y_i / D_i = 1, E_i = 1\right] - E\left[Y_i / D_i = 0, E_i = 1\right] \\ ITE &= E\left[Y_i / D_i = 1, E_i = 0\right] - E\left[Y_i / D_i = 0, E_i = 0\right] \end{aligned}$$

We compare the treatment and control mean outcomes for eligible and non-eligible groups. We compute each expectation in order to recover the parameter of interest. In the case of ATT:

(7)
$$\begin{aligned} E\left[Y_i / D_i = 1, PovHH_i = 1\right] &= \alpha_2 + \alpha_3 + \alpha_4 \\ E\left[Y_i / D_i = 0, PovHH_i = 1\right] &= \alpha_3 \end{aligned}$$

Therefore:

(8) $ATT = E\left[Y_i / D_i = 1, E_i = 1\right] - E\left[Y_i / D_i = 0, E_i = 1\right] = \alpha_2 + \alpha_4$

Therefore, the coefficients for D and for the interaction recover ATT. We proceed in the same way for ITE:

(9)
$$\begin{aligned} E\left[Y_i / D_i = 1, PovHH_i = 0\right] &= \alpha_2 \\ E\left[Y_i / D_i = 0, PovHH_i = 0\right] &= 0 \end{aligned}$$

Therefore, ITE is recover with α_2.

The results are the following:

```
. reg enroll D pov_HH int_D_PovHH if year==1999, vce(cluster villid)

Linear regression                                      Number of obs =   10078
                                                       F(  3,    491) =    3.28
                                                       Prob > F      =  0.0207
                                                       R-squared     =  0.0042
                                                       Root MSE      =  .49898

                               (Std. Err. adjusted for 492 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
```

```
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .0465157   .0287104     1.62   0.106    -.0098947     .102926
      pov_HH |    .000313   .0199547     0.02   0.987    -.0388942    .0395202
 int_D_PovHH |   .0265883   .0260427     1.02   0.308    -.0245806    .0777572
       _cons |   .4687701   .0231008    20.29   0.000     .4233815    .5141587
------------------------------------------------------------------------------
```

We got the same results as before. ITE is 0.0465 and ATT is the sum of the coefficients for D and the interaction (0.0465+0.0266=0.0731).

You might be worry by the fact that the coefficients for ATT are not individually significant. D has a p-value of 0.106 whereas the interaction a p-value of 0.308. This differs from the original regression in which the coefficient for ATT in the regression for eligible was strongly significant (p-value=0.03). A no-so-smart scholar might incorrectly conclude that ATT is no significant by looking to the individual coefficients. One way to prevent that is by using a simple F-test to check whether the sum of coefficients is different from zero. The code is the following:

```
test D + int_D_PovHH = 0
```

The results are the following:

```
. test D + int_D_PovHH = 0

 ( 1)  D + int_D_PovHH = 0

       F(  1,   491) =    8.76
            Prob > F =    0.0032
```

We found that the sum of coefficients is strongly significant even though the individual coefficients are not. Therefore, we safely conclude that ATT is significant although the evidence of spillover effects (via ITE) is extremely weak.

We can also add control variables to the previous specification. The results are essentially the same.

These cross-sectional results can be extended by exploiting the panel structure of the original dataset. We can do that using a difference in difference (DD) approach applied to experimental data. We need to create interactions with a time dummy variable equal to 0 before the program started and 1 otherwise. The code is the following:

```
gen int_PovHH_Y = pov_HH if year==1998 | year==1999
replace int_PovHH_Y = 0 if year==1997

gen int_D_Y = D_assig if year==1998 | year==1999
replace int_D_Y = 0 if year==1997

gen int_D_Y_PovHH=pov_HH*int_D_Y
```

We have created two double interactions with time dummy and `pov_HH` and `D_assig`. We run the separate specification for eligible and non-eligible groups as before. The code is the following:

```
reg enroll int_D_Y D_assig Y8 Y9 if pov_HH==1, vce(cluster villid)

reg enroll int_D_Y D_assig Y8 Y9 if pov_HH==0, vce(cluster villid)
```

The results are reported below:

```
. reg enroll int_D_Y D_assig Y8 Y9 if pov_HH==1, vce(cluster villid)

Linear regression                                      Number of obs =   20445
                                                       F(  4,   492) =   72.30
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0189
                                                       Root MSE      =  .48429

                               (Std. Err. adjusted for 493 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     int_D_Y |   .0722822   .0142207     5.08   0.000     .0443414     .100223
     D_assig |   .0043284   .0224749     0.19   0.847    -.0398304    .0484871
          Y8 |  -.0689565   .0111681    -6.17   0.000    -.0908996   -.0470134
          Y9 |  -.1856037   .0125475   -14.79   0.000     -.210257   -.1609504
       _cons |   .6524922   .0174817    37.32   0.000     .6181442    .6868401
------------------------------------------------------------------------------

.
. reg enroll int_D_Y D_assig Y8 Y9 if pov_HH==0, vce(cluster villid)

Linear regression                                      Number of obs =   12523
                                                       F(  4,   458) =   86.91
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0266
                                                       Root MSE      =  .48144

                               (Std. Err. adjusted for 459 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     int_D_Y |   .0244131   .0177278     1.38   0.169    -.0104249     .059251
     D_assig |   .0279069   .0250824     1.11   0.266    -.0213839    .0771977
          Y8 |  -.0538757   .0137018    -3.93   0.000     -.080802   -.0269494
          Y9 |  -.1941972   .0154546   -12.57   0.000    -.2245679   -.1638265
       _cons |   .6595289   .0203424    32.42   0.000     .6195528     .699505
------------------------------------------------------------------------------
```

The variable of interest is the interaction `int_D_Y`. This recovers the impact of the intervention after the implementation of the program. We found a significant impact among those eligible is 0.07 percentage points. We find no impact among those non-eligible.

We also have created a triple interaction among these 3 variables. This variable is useful for a single regression that accommodates both groups. The estimating equation is the following:

$$\begin{aligned} & Y = \alpha_1 + \alpha_2 D + \alpha_3 PovHH + \alpha_4 t \\ (10) \quad & +\alpha_5 (PovHH * D) + \alpha_6 (PovHH * t) + \alpha_7 (D * t) \\ & +\alpha_8 (PovHH * D * t) + \mu \end{aligned}$$

This equation includes double interactions and triple interactions to recover the differential impacts of the treatment among eligible and non-eligible, before and after the treatment. Following our previous example, we find that ATT is recover by the sum of α_7 and α_8 and ITE is recovered by α_7. The results (along with the F-test for the sum) are reported below:

```
. reg enroll int_D_Y_PovHH int_D_Y int_D_PovHH int_PovHH_Y D_assig pov_HH Y8 Y9,
vce(cluster villid)

Linear regression                                      Number of obs =   32968
                                                       F(  8,   500) =   64.89
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0218
                                                       Root MSE      =  .48322

                                (Std. Err. adjusted for 501 clusters in villid)
-------------------------------------------------------------------------------
             |               Robust
 enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
int_D_Y_PovHH |   .0476086   .0198913     2.39   0.017     .0085277    .0866895
      int_D_Y |   .0246218   .0177411     1.39   0.166    -.0102346    .0594781
  int_D_PovHH |  -.0235785   .0232885    -1.01   0.312    -.0693339    .0221768
  int_PovHH_Y |  -.0032527   .0153222    -0.21   0.832    -.0333565    .0268512
      D_assig |   .0279069   .0250791     1.11   0.266    -.0213665    .0771803
       pov_HH |  -.0070368   .0182842    -0.38   0.701    -.0429601    .0288866
           Y8 |  -.0612501   .0140889    -4.35   0.000    -.0889308   -.0335694
           Y9 |  -.1868561   .0144929   -12.89   0.000    -.2153305   -.1583816
        _cons |   .6595289   .0203398    32.43   0.000     .6195669    .6994909
-------------------------------------------------------------------------------


. test int_D_Y + int_D_Y_PovHH = 0

 ( 1)  int_D_Y_PovHH + int_D_Y = 0

       F(  1,   500) =   25.80
            Prob > F =    0.0000
```

The results are consistent with the previous ones. We find no impact among those non-eligible.

4. Attrition

Attrition is a problem that arises in a context of panel data or when we have longitudinal observations of individuals. Our dataset suffers a problem of attrition, since some children that were observed in 1997 could not be surveyed in the following years. To start this section, we open again the dataset and create the additional variables we created before. The code is the following:

```
use"$path/PanelPROGRESA_Enrollment_97_99.dta", clear

destring year, replace

* Creation of auxiliary variables

gen aux=D if year==1998
egen D_assig=mean(aux),by(villid)

gen int_D_Y = D_assig if year==1998 | year==1999
replace int_D_Y = 0 if year==1997

gen aux1=pov_HH if year==1998
egen aux2=mean(aux1),by(hogid)
replace pov_HH=aux2 if year==1997
drop aux1 aux2

gen Y8=year==1998
gen Y9=year==1999

* Panel setting
egen id_=concat(iid)
encode id_, g(id)
xtset id year
```

The next table shows us the magnitude of attrition in the database among poor children. The variable `paths` has four categories, every one showing the order of appearance of observations during the three years: of the 13,146 children in the database in 1997, only 79% of them appear always, 17% appear in the first two years but not in the last one, and the remainder have other forms of appearance. It means that around 21% of the individuals have left the sample at least once.

```
. tab paths if year==1997

group(panel |
   7 panel8 |
    panel9) |      Freq.     Percent        Cum.
------------+-----------------------------------
      1 1 1 |     10,384       78.99       78.99
      1 1 . |      2,195       16.70       95.69
      1 . 1 |        244        1.86       97.54
      1 . . |        323        2.46      100.00
------------+-----------------------------------
      Total |     13,146      100.00
```

If this attrition had happened at random, then it would not mean a problem since would be balanced among treated and control group. Nevertheless, if this attrition were correlated with treatment, then a comparison between treated and untreated children could be biased. This may happen, for example, if those treated households for which education is a less valuable asset were also those that are more likely to drop the sample. Since this valuation of education is an unobserved variable, this kind of attrition drives the typical bias for omitted variable.

4.1. Testing non-random attrition

How to verify if attrition is random? There are no perfect procedures, but the most common is to verify if attrition is correlated with the treatment status or with other baseline covariates, and if it is different between treated and controls. This is done in the following table using the variable `attrit` (which was constructed as one for those left the sample at least once according to variable `paths`). As it is clear, a little more than 20% of treated individuals have left at least once period in the sample and the proportion is almost one percent point less in the control group.

```
. tab attrit D_assig if pov_HH==1 & year==1997, nof col

           |        D_assig
    attrit |         0          1 |     Total
-----------+----------------------+----------
         0 |     83.62      82.10 |     82.66 
         1 |     16.38      17.90 |     17.34 
-----------+----------------------+----------
     Total |    100.00     100.00 |    100.00 
```

The small difference observed in the last table could be driven just by sampling error. To assess this case, we run two regressions by OLS between attrition and treatment status: the first one without controlling for covariates and the second one controlling for all the baseline covariates that we have in our dataset. Additionally, we run two regressions separately for treated and control units in order to analyze whether attrition is driven differently in each group. The STATA code is the following:

```
reg attrit D_assig if pov_HH==1 & year==1997

reg attrit D_assig age sex lang eduHH sexHH yycali i.entidad if pov_HH==1 &
year==1997, vce(cluster villid)

reg attrit age sex lang eduHH sexHH yycali i.entidad if pov_HH==1 &
year==1997 & D_assig==1, vce(cluster villid)

reg attrit age sex lang eduHH sexHH yycali i.entidad if pov_HH==1 &
year==1997 & D_assig==0, vce(cluster villid)
```

The results are reported below for the simple specification:

```
. reg attrit D_assig if pov_HH==1 & year==1997

      Source |       SS       df       MS              Number of obs =    7792
-------------+------------------------------           F(  1,  7790) =    2.92
       Model |  .417844127     1  .417844127           Prob > F      =  0.0878
    Residual |  1116.34178  7790  .143304465           R-squared     =  0.0004
-------------+------------------------------           Adj R-squared =  0.0002
       Total |  1116.75963  7791  .143339703           Root MSE      =  .37856
```

```
------------------------------------------------------------------------------
      attrit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     D_assig |   .0151631      .00888     1.71   0.088     -.002244    .0325702
       _cons |   .1638379    .0070454    23.25   0.000      .150027    .1776488
------------------------------------------------------------------------------
```

We find in this case that attrition is weakly related to treatment. To see whether this relationship is strong with add some additional covariates:

```
. reg attrit D_assig age sex lang eduHH sexHH yycali i.entidad if pov_HH==1 &
year==1997, vce(cluster villid)

Linear regression                                      Number of obs =    7718
                                                       F( 13,   492) =   21.85
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0422
                                                       Root MSE      =  .37126

                               (Std. Err. adjusted for 493 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
      attrit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     D_assig |   .0143036    .0137077     1.04   0.297    -.0126292    .0412363
         age |   .0392863    .0026013    15.10   0.000     .0341754    .0443973
         sex |  -.0528829    .0087631    -6.03   0.000    -.0701007   -.0356651
        lang |   .0204527    .0152316     1.34   0.180    -.0094744    .0503797
       eduHH |  -.0000368    .0000216    -1.71   0.088    -.0000792    5.56e-06
       sexHH |  -.0454898    .0216678    -2.10   0.036    -.0880627   -.0029169
      yycali |  -.0000747    .0000726    -1.03   0.304    -.0002173    .0000679
             |
     entidad |
         13  |   .0716133    .0308333     2.32   0.021     .0110321    .1321945
         16  |  -.0212358    .0257075    -0.83   0.409    -.0717459    .0292742
         21  |   .0052605    .0272133     0.19   0.847    -.0482082    .0587292
         22  |   .0571334    .0292622     1.95   0.051    -.0003609    .1146276
         24  |   .0234319    .0276141     0.85   0.397    -.0308241    .0776879
         30  |   .0272734    .0259008     1.05   0.293    -.0236165    .0781633
             |
       _cons |  -.2644686    .0633971    -4.17   0.000    -.3890311   -.1399061
------------------------------------------------------------------------------
```

As it is clear, attrition is not related to the treatment variable when we control for covariates. It seems, however, that attrition is linked to some covariates. Age, gender and other household and regional characteristics are related to the probability of leaving the sample.

4.2. Re-weighting methods

The results above show that attrition seems not to be correlated with treatment but, since it is not uncorrelated with covariates, we are not sure it is random. That is why it is preferable to implement some kind of solution that tries to overcome the potential bias of estimations. This is what reweighting methods do. To implement this technique, we have to follow the next four steps (see Baulch and Agnes 2011 for details):

a) First: run a probit regression between a variable that take a value of one for observations that have not attrited (zero otherwise) and all covariates that you plan to include in your whole (less parsimonious) model and recover the fitted probabilities
b) Second: run a probit regression between the variable that take a value of one for observations that have not attrited (zero otherwise)and all significant variables that you encountered in the first regression and recover the fitted probabilities
c) Third: compute the weights for every i individual using the next formula and use them to reweight the final estimation:

$$weight_i = \frac{Pr[No\ attrit_i = 1|X_i^*]}{Pr[No\ attrit_i = 1|X_i]}$$

Where $X_i^* \subset X_i$

d) Four: estimate your final model linking the outcome variable and the treatment variable using the last variable created as weights.

This correction method tries to give more weight to observations that have similar initial characteristics to observations that subsequently attrite than to observations with characteristics that make them more likely to remain in the panel. The STATA code is the following:

```
* Construction of weights

gen noattrit=1-attrit

probit noattrit enroll age sex lang eduHH sexHH yycali D_assig i.entidad if
pov_HH==1 & year==1997, vce(cluster villid)
gen sample=e(sample)
predict prF if sample==1

probit noattrit enroll age sex D_assig if pov_HH==1 & year==1997, vce(cluster
villid)
predict prR if sample==1
```

Observe that we are using observations from year 1997 and have created the variable `sample` after fitting the model to identify all the observations used in the estimation of the whole model. Then we have restricted the prediction of fitted probabilities only to observations that were used in this estimation. The probabilities predicted are called `prF` for the full model and `prR` for the restricted one. The results for the probit full model are reported below:

```
. probit noattrit enroll age sex lang eduHH sexHH yycali D_assig i.entidad if
pov_HH==1 & year==1997, vce(cluster villid)

Iteration 0:   log pseudolikelihood = -3549.3946
Iteration 1:   log pseudolikelihood = -3374.0354
Iteration 2:   log pseudolikelihood = -3372.6537
Iteration 3:   log pseudolikelihood = -3372.6534
```

```
Probit regression                                 Number of obs   =       7674
                                                  Wald chi2(14)   =     275.80
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -3372.6534                 Pseudo R2       =     0.0498

                               (Std. Err. Adjusted for 493 clusters in villid)

             |               Robust
    noattrit |      Coef.   Std. Err.      Z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
enroll_child |   .2059888   .0469029     4.39   0.000     .1140609    .2979168
         age |  -.1334588   .0121984   -10.94   0.000    -.1573672   -.1095504
         sex |   .1847347   .0367303     5.03   0.000     .1127446    .2567248
        lang |  -.0912348   .0592722    -1.54   0.124    -.2074061    .0249366
       eduHH |   .0001287   .0000936     1.37   0.169    -.0000548    .0003121
       sexHH |   .1785207   .0785526     2.27   0.023     .0245604     .332481
      yycali |   .0002901   .0002842     1.02   0.307     -.000267    .0008472
     D_assig |  -.0607537   .0558964    -1.09   0.277    -.1703086    .0488011
             |
     entidad |
        13   |  -.2760336   .1144844    -2.41   0.016    -.5004189   -.0516483
        16   |   .1231614   .1074995     1.15   0.252    -.0875338    .3338565
        21   |   .0047961   .1134391     0.04   0.966    -.2175406    .2271327
        22   |  -.1971482   .1107041    -1.78   0.075    -.4141243    .0198279
        24   |  -.0864849   .1074652    -0.80   0.421    -.2971128     .124143
        30   |  -.1018132    .100502    -1.01   0.311    -.2987934    .0951671
             |
       _cons |   2.305226   .2874815     8.02   0.000     1.741773     2.86868
```

We also report the results for the restricted model:

```
. probit noattrit enroll age sex D_assig if pov_HH==1 & year==1997, vce(cluster
villid)

Iteration 0:   log pseudolikelihood = -3574.6624
Iteration 1:   log pseudolikelihood =  -3432.639
Iteration 2:   log pseudolikelihood = -3431.8387
Iteration 3:   log pseudolikelihood = -3431.8387

Probit regression                                 Number of obs   =       7741
                                                  Wald chi2(4)    =     250.18
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -3431.8387                 Pseudo R2       =     0.0400

                               (Std. Err. adjusted for 493 clusters in villid)
------------------------------------------------------------------------------
             |               Robust
    noattrit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
enroll_child |   .1897525   .0464106     4.09   0.000     .0987893    .2807157
         age |  -.1339186   .0125297   -10.69   0.000    -.1584765   -.1093608
         sex |   .1811639   .0365813     4.95   0.000     .1094658    .2528619
     D_assig |  -.0658357   .0556065    -1.18   0.236    -.1748225     .043151
       _cons |   2.590503   .1862264    13.91   0.000     2.225506      2.9555
------------------------------------------------------------------------------
```

Recall that these models are just opposite to those estimated previously by OLS (here, dependent variable is `noattrit`) and, even when they are not directly comparable, direction of parameters seems to confirm our previous findings.

To observe the differences in the range of probability estimates for `prR` and `prF`, we made use of the `kdensity` command to plot the kernel density estimates. We use just default options for the kernel (epanechnikov) and an optimal bandwidth. The resulting graph is presented in the left hand panel and the command line used is:

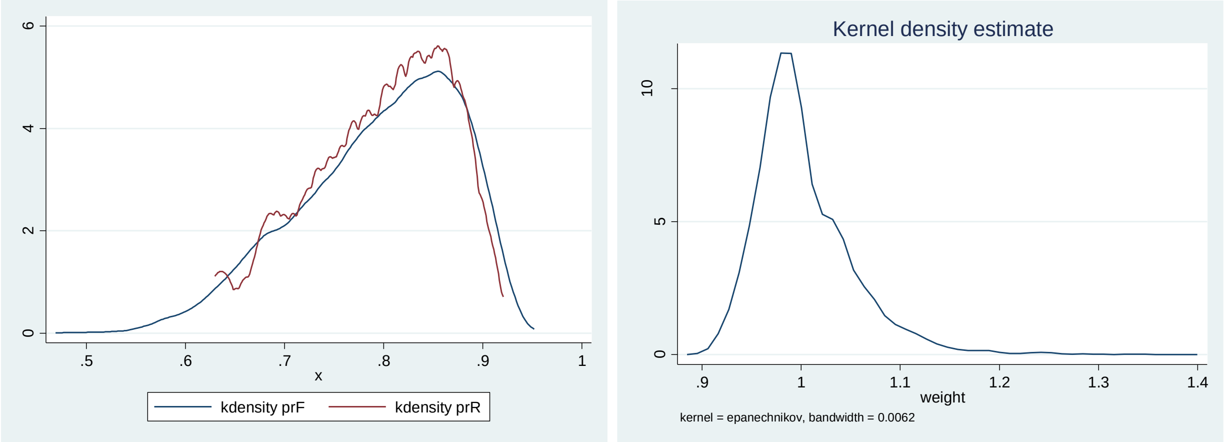
```
twoway (kdensity prF) (kdensity prR)
```

With these estimations, we computed the ratio of `prR` and `prF` to create the weights using the next command line:

```
* Weights
gen weight=prR/prF if sample==1
```

To observe the empirical distribution of the weights, we use the `kdensity` command again. This is shown in the right hand side graph below. The empirical distribution of this variable shows a mode around 1, and a range of values that runs from 0.85 to 1.4.

```
twoway (kdensity prF) (kdensity prR)    kdensity weight
```



kernel = epanechnikov, bandwidth = 0.0062

Now that we have already computed the non-attrition weights, we are able to estimate the model of interest between the outcome variable and treatment status. To do this, we first have to keep in mind that the weights just estimated are defined only for year 1997. It means, since it does not have observations for years 1998 and 1999 (they are missing), this variable is not usaful for a panel estimation yet. The strategy we follow from now on is to replicate the observations of 1997 for subsequent years for every individual. This is done using the next command line:

```
egen weight_y=mean(weight), by(id)
```

weight_y is now a useful variable for panel analysis.

For the estimation of the model of interest, we are going to use two estimators: the first one will be a simple difference estimator with the pool of data from 1998/1999 and a fixed effect estimator with the data from the three rounds of survey. The command lines used are shown in the next box:

```
* Simple Difference Estimator

reg enroll int_D_Y if pov_HH==1 & (year==1998 | year==1999) [w=weight_y]
estimates store WWNCS
gen samp_ws = e(sample)

* Fixed Effect Estimator

xtreg enroll int_D_Y Y8 Y9 if pov_HH==1 [w=weight_y], fe
estimates store WWNCF
gen samp_wf = e(sample)
```

The results for the simple difference model are reported below:

```
. * Simple Difference Estimator
.
. reg enroll int_D_Y if pov_HH==1 & (year==1998 | year==1999) [w=weight_y]
(analytic weights assumed)
(sum of wgt is   1.2520e+04)

      Source |       SS       df       MS              Number of obs =   12508
-------------+------------------------------           F(  1, 12506) =   68.70
       Model |  16.7101824     1  16.7101824           Prob > F      =  0.0000
    Residual |  3041.89449 12506  .243234807           R-squared     =  0.0055
-------------+------------------------------           Adj R-squared =  0.0054
       Total |  3058.60468 12507  .244551425           Root MSE      =  .49319

------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     int_D_Y |   .0756488   .0091269     8.29   0.000     .0577586    .0935389
       _cons |   .5263898   .0072365    72.74   0.000      .512205    .5405745
------------------------------------------------------------------------------
```

The results for the fixed effect estimator are presented below:

```
. * Fixed Effect Estimaror
.
. xtreg enroll int_D_Y Y8 Y9 if pov_HH==1 [w=weight_y], fe
(analytic weights assumed)

Fixed-effects (within) regression               Number of obs      =     20182
Group variable: id                              Number of groups   =      7674

R-sq:  within  = 0.0764                         Obs per group: min =         1
       between = 0.0089                                        avg =       2.6
       overall = 0.0172                                        max =         3
```

```
                                                F(3,12505)          =    344.94
corr(u_i, Xb)  = -0.0419                        Prob > F            =    0.0000

------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     int_D_Y |   .0773041   .0103192     7.49   0.000     .0570769    .0975313
          Y8 |  -.1332843   .0087407   -15.25   0.000    -.1504173   -.1161513
          Y9 |   -.232798   .0087616   -26.57   0.000    -.2499721   -.2156238
       _cons |   .6874478   .0038723   177.53   0.000     .6798576    .6950381
-------------+----------------------------------------------------------------
     sigma_u |  .42176945
     sigma_e |  .33237475
         rho |  .61689541   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(7673, 12505) =     3.95         Prob > F = 0.0000
```

To compare the results, we will also fit unweighted models. Notice that to ensure comparability, we have restricted the estimation of both models to observations used in the weighted estimations above. The code is below:

```
* UNWEIGHTED ESTIMATES

* Simple Difference Estimator

reg enroll int_D_Y if samp_ws==1 & pov_HH==1 &(year==1998|year==1999),
estimates store UWNCS

* Fixed Effect Estimator

xtreg enroll int_D_Y Y8 Y9 if samp_wf==1 & pov_HH==1, fe
estimates store UWNCF
```

The results are reported below:

```
. * UNWEIGHTED ESTIMATES
.
. * Simple Difference Estimator
.
. reg enroll int_D_Y if samp_ws==1 & pov_HH==1 &(year==1998|year==1999),

      Source |       SS       df       MS              Number of obs =   12508
-------------+------------------------------           F(  1, 12506) =   70.63
       Model |  17.1751418     1  17.1751418           Prob > F      =  0.0000
    Residual |   3040.9742 12506  .243161219           R-squared     =  0.0056
-------------+------------------------------           Adj R-squared =  0.0055
       Total |  3058.14934 12507  .244515019           Root MSE      =  .49311


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

```
     int_D_Y |   .0766764   .0091234     8.40   0.000      .058793    .0945597
       _cons |   .5260215   .0072314    72.74   0.000     .5118469    .5401961
------------------------------------------------------------------------------

. estimates store UWNCS

.
. * Fixed Effect Estimator

. xtreg enroll int_D_Y Y8 Y9 if samp_wf==1 & pov_HH==1, fe

Fixed-effects (within) regression               Number of obs      =     20182
Group variable: id                              Number of groups   =      7674

R-sq:  within  = 0.0760                         Obs per group: min =         1
       between = 0.0087                                        avg =       2.6
       overall = 0.0172                                        max =         3

                                                F(3,12505)         =    342.62
corr(u_i, Xb)  = -0.0419                        Prob > F           =    0.0000

------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     int_D_Y |   .0779189   .0103216     7.55   0.000      .057687    .0981508
          Y8 |  -.1331696   .0087411   -15.23   0.000    -.1503035   -.1160356
          Y9 |  -.2324124   .0087595   -26.53   0.000    -.2495824   -.2152423
       _cons |   .6873463   .0038758   177.34   0.000     .6797492    .6949435
-------------+----------------------------------------------------------------
     sigma_u |  .42173907
     sigma_e |  .33255822
         rho |  .61660047   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(7673, 12505) =     3.95         Prob > F = 0.0000

. estimates store UWNCF
```

The evidence shows that the interaction between treatment and time for the weighted estimation is a little smaller than that for the unweighted one. This is true also for fixed effects estimations.

To assess more formally the similarity of the parameters estimated by weighted and unweighted methods, we run Hausman tests using the following command line:

```
* COMPARING APPROACHES

* Simple difference
hausman WWNCS UWNCS

* Fixed effects
hausman WWNCF UWNCF
```

Results are presented below. For both estimations, we reject the null hypothesis of not systematic difference in parameters, although this test is not suitable for the case of the fixed effect model. This means that attrition should be a concern and, at least under the exercises developed in this section, it is preferable to choose the weighted models.

```
. * COMPARING APPROACHES
.
. * Simple difference
. hausman WWNCS UWNCS

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     WWNCS        UWNCS        Difference          S.E.
-------------+----------------------------------------------------------------
     int_D_Y |    .0756488     .0766764       -.0010276        .0002519
------------------------------------------------------------------------------
                         b = consistent under Ho and Ha; obtained from regress
          B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       16.64
                Prob>chi2 =      0.0000

.
. * Fixed effects
. hausman WWNCF UWNCF

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     WWNCF        UWNCF        Difference          S.E.
-------------+----------------------------------------------------------------
     int_D_Y |    .0773041     .0779189       -.0006148               .
          Y8 |   -.1332843    -.1331696       -.0001147               .
          Y9 |    -.232798    -.2324124       -.0003856        .0001905
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =    -8.38    chi2<0 ==> model fitted on these
                                        data fails to meet the asymptotic
                                        assumptions of the Hausman test;
                                        see suest for a generalized test
```

4.3. Lee's (2009) bounding methods

The procedure described and implemented above tries to reweight the observations in order to estimate an unbiased parameter of the treatment. However, it relies on selection on observables assumption, meaning that if attrition is caused by some unobservable factor, then the procedure could fail to correct the point estimates.

Contrarily to reweighting methods, Lee (2009) bounding method does not seek to solve the potential bias that attrition could cause. Instead, it tries to construct the "worst-case scenario" bounds of the treatment effect. The basic idea is to impute the missing data with either the largest or smallest possible values to compute the largest and smallest possible treatment effects consistent with the data that is observed. With this solution, even when we will not be able to recover a point estimate, we will have limits within which the impact should be placed.

To implement this procedure in STATA, we need first to download the user-written command `leebounds` from IdeasRepec web site. This is done by typing the following in the STATA command line:

```
ssc install leebounds
```

This STATA command implements directly the Lee (2009) procedure.

For the exercise, we estimate the impact of PROGRESA using a simple difference estimator with data of the outcome (enrollment) from 1998 and 1999, just as in the reweighting exercise. This is possible given the experimental nature of the data generated with PROGRESA.

To begin, the next table shows the point estimate of the basic model. As we can see, the point estimate is 0.778, very similar to those obtained above.

```
. reg enroll int_D_Y if pov_HH==1 & (year==1998|year==1999)

      Source |       SS       df       MS              Number of obs =   12701
-------------+------------------------------           F(  1, 12699) =   72.84
       Model |  17.7107215     1  17.7107215           Prob > F      =  0.0000
    Residual |  3087.74775 12699  .243148889           R-squared     =  0.0057
-------------+------------------------------           Adj R-squared =  0.0056
       Total |  3105.45847 12700  .244524289           Root MSE      =   .4931


------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     int_D_Y |   .0772848   .0090555     8.53   0.000     .0595346    .0950349
       _cons |   .5255459   .0071796    73.20   0.000     .5114727    .5396191
------------------------------------------------------------------------------
```

Now, let try to estimate the bounds of the impact by using the `leebounds` command. The syntax is shown in the following line:

```
leebounds   enroll   int_D_Y   if   pov_HH==1   &   (year==1998|year==1999),
select(noattrit)
```

We are asking STATA to estimate the bounds among poor people and considering the data from 1998 and 2000. Results are shown in the next table. According to them, the lower bound of the impact is around 7.5% and the upper one is almost 8.7%.

```
. leebounds enroll int_D_Y if pov_HH==1 & (year==1998|year==1999),
select(noattrit)

Lee (2009) treatment effect bounds

Number of obs.                  =    13132
Number of selected obs.         =    11803
Trimming porportion             =    0.0119

------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
int_D_Y      |
       lower |   .0752802   .0100066     7.52   0.000     .0556676    .0948928
       upper |   .0872957   .0098896     8.83   0.000     .0679124     .106679
------------------------------------------------------------------------------
```

5. Final Remarks

In this chapter, we have studied several problems that affect randomized designs. We discussed some ways to deal with these problems.

6. Further Readings

Angelucci , M. and V. Di Maro (2010). “Program Evaluation and Spillover Effects,” Inter-American Development Bank, Technical Note 136.

Angrist, Joshua and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

Baulch, B. and A. Quisumbing (2011). “Testing and adjusting for attrition in household panel data,” Mimeo.

Lee, David (2009). “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *Review of Economic Studies*, 76(3), 1071-1102.