***Impact Evaluation in Practice*, Second Edition**

Technical Companion[1] *(Version 1.0, September 2016)*

Introduction

Impact Evaluation in Practice (second edition) offers a comprehensive and accessible introduction to impact evaluation for policy makers and development practitioners. The book is divided into four parts. Part 1 reviews how to prepare for an impact evaluation, what to evaluate, and why. Part 2 presents the basic concepts of impact evaluation by relying mostly on intuition and graphical representations. Part 3 discusses how to choose an impact evaluation method in a given operational context, and how to manage impact evaluations. Part 4 reviews how to get data, including sampling, power calculations, and data sources for impact evaluation. The presentation in the book is nontechnical, and focuses on the intuition behind technical concepts and impact evaluation methods, as well as sampling and power calculation.

In this technical companion, we include additional material for readers with a background in statistics and econometrics. The technical companion assumes a basic understanding of statistics, in particular key concepts such as regression analysis and hypothesis testing. The technical companion presents an introduction to the analysis of impact evaluation data. It summarizes the basic potential outcome framework that underpins the econometrics of impact evaluation, discusses how to represent the methods in a regression framework, and provides some applications using Stata. The technical companion also provides examples of how to undertake power calculations in Stata and Optimal Design. Applications are based on the case study of the Health Insurance Subsidy Program (HISP) presented in the book. Supplementary data and related do-files can be found on the book website (www.worldbank.org/ieinpractice), and can be used to replicate the results presented in the companion.

While this technical companion presents an introduction to analyzing impact evaluation data, the objective is not to provide an in-depth discussion of the econometrics behind impact evaluation. If you would like additional details or a more comprehensive coverage, you are invited to read Angrist and Pischke (2009) or Angrist and Pischke (2014), on which this companion partly draws. Although we provide some examples and applications using Stata, this online companion is not a thorough empirical guide on how to apply the methods in practice. If you are interested in additional information and practical applications, you can consult, among other relevant material, the applied impact evaluation methods course at the University of California, Berkeley (http://aie.cega.org).

---

[1] Please cite this technical companion as: *"Gertler, Paul J.; Martinez, Sebastian; Premand, Patrick; Rawlings, Laura B.; Vermeersch, Christel M. J.. 2016. Impact Evaluation in Practice, Second Edition, Technical Companion (Version 1.0). Washington, DC: Inter-American Development Bank and World Bank.*

Marina Tolchinsky provided excellent research assistance in preparing this technical companion and related Stata material. Aakash Mohpal provided useful comments. This is version 1.0 (September 2016) of the technical companion. The authors welcome feedback and suggestions on how to improve this first version in the future.

Causal Inference and Counterfactuals: The Potential Outcome Framework

In chapter 3 of *Impact Evaluation in Practice, Second Edition*, we noted that the answer to the basic impact evaluation question—What is the impact or causal effect of a program *P* on an outcome of interest *Y*?— is given by a basic impact evaluation formula:

$$\delta = (Y \mid P = 1) - (Y \mid P = 0)$$

$$\delta = Y_1 - Y_0$$

This formula says that the causal impact (δ) of a program (*P*) on an outcome (*Y*) is the difference between the outcome (*Y*) with the program (in other words, when *P*=1) and the same outcome (*Y*) without the program (that is, when *P*=0). While this basic formula is quite simple and was presented intuitively in chapter 3, it captures deeper concepts in statistics and econometrics. The basic impact evaluation formula, and more generally the broader impact evaluation literature, trace back to the potential outcome framework.[2]

Let's assume that we are interested in measuring the impact of a program on an outcome in some population of interest. The population (or universe) constitutes a set (*I*) of units, each denoted by *i*. Depending on the context, the units may be individuals, households, or facilities. The outcome of interest is *Y*. For each unit (*i*) of the broader population, the outcome of interest takes a value Y_i. The population is usually comprised of a very large number of units, so that it is typically impossible to observe the value of the outcome of interest for the entire population.

In chapter 2 of the book[3], we introduced the Health Insurance Subsidy Program (HISP). HISP aims to subsidize the purchase of health insurance among poor rural households. In the context of this program, the population of interest (*I*) is comprised of poor rural households (*i*) in the country, and the outcome of interest (*Y*) for the impact evaluation is household yearly out-of-pocket health expenditures.

The impact evaluation problem is to assess how the outcome of interest (*Y*) responds to being exposed to the program (*P*). To make things easier, we assume that exposure to the program is a binary random variable: each individual *i* can either be exposed to the program (in which case P_i=1), or not exposed to the program (in which case P_i=0).[4] For now, we also assume that all units exposed to the program participate, and that all units not exposed to the program do not participate.

In this context, there are two potential outcomes for each unit: either s/he is exposed and participates in the program $[(Y_i|\ P_i = 1)$ or $Y_{i1}]$ or s/he is not exposed to the program and does not participate $[(Y_i|\ P_i = 0)$ or $Y_{i0}]$. For each unit in the population, the causal effect of the program is theoretically determined by a simple difference between the potential outcome with the program and the potential outcome without the program:

$$\delta = Y_{i1} - Y_{i0}$$

[2] See Rubin (1974); Holland (1986).
[3] See p. 39 in the book for a first description of the HISP case study.
[4] Issues related to noncompliance will be discussed further below.

The basic problem of causal inference is that we cannot observe the same unit in both states of the world at the same time, so it is impossible to observe program effects for each unit. We do not observe the counterfactual outcome for that same unit in another state of the world. For a unit *i* exposed to the program, we cannot observe what would have happened to that same unit *i* in the absence of that program. In the context of the Health Insurance Subsidy Program (HISP), a given poor household is either participating in the program or not participating in the program.

Since causal effects cannot be measured for each unit *i*, let us go back to the population (*I*) to see how to identify the average treatment effect (ATE). At the population level, the average treatment effect is the difference between the expected value of the outcome when the population is exposed to the program and the expected value of the outcome when the population is not exposed to the program:

$$ATE_I = E_I(Y_{i1} - Y_{i0}) = E_I(Y_{i1}) - E_I(Y_{i0})$$

We are interested in using a sample generated through an impact evaluation design to estimate the average treatment effect for the population. Econometric methods help to find consistent estimators of the average treatment effect. When applied to a sample, consistent estimators tend to generate accurate estimates of the average treatment effect for the population.

One potential estimator that can help infer the average treatment effect for the population is to take the difference in the average outcome of units in a sample exposed to the program and the average outcome of units in a sample not exposed to the program:

$$\delta = (\bar{Y} \mid P = 1) - (\bar{Y} \mid P = 0)$$

While the average treatment effect above (ATE_I) is for the whole population, the *estimator* is applied to an observed sample of units. The estimator produces an *estimate* of the average treatment effect by taking the difference between the mean sample outcomes among units exposed to the program and the mean sample outcomes among units not exposed to the program.

In the case of HISP[5], the average treatment effect obtained from this estimator would be the estimated difference in household yearly out-of-pocket health expenditures for poor rural households receiving the subsidies and poor rural households not receiving these subsidies.

This estimator of the average treatment effect is only consistent, meaning that, based on the sample, it tends to generate accurate estimates of the average treatment effect for the population, under specific conditions. The main condition is that exposure to the program should be independent of the distribution of potential outcomes. This condition has two parts:

- The average outcome of program beneficiaries and nonbeneficiaries should be the same if neither of them was exposed to the program.[6]

[5] For now, we still focus on the scenario where all units exposed to the program participate, and that all units not exposed to the program do not participate

[6] Formally, $(\overline{Y_1}|P = 1) = (\overline{Y_1}|P = 0)$.

- The average outcome of the program beneficiaries and nonbeneficiaries should be the same if they were both exposed to the program.[7]

Selection bias is one case when the condition of independence of potential outcomes does not hold. A selection bias may arise when the units that participate in the program are different or react differently to the program than units that do not participate in the program. In the case of the HISP program, a selection bias may occur if households deciding to sign up for the health insurance subsidy program are intrinsically different compared to households that do not sign up, and if these intrinsic differences are correlated with the outcome of interest. For instance, if households that decide to sign up are more prone to illness than households that decide not to sign up, comparing the average health expenditures between these two groups of households will mix the effect of participation in the program with the differences in expenditures driven by intrinsic differences between the two groups. Box 1 presents a theoretical discussion of the selection bias. We provide a more practical discussion of the selection bias below when introducing the regression framework.

When the condition of independence of potential outcomes holds, no baseline differences would generally be expected between the group exposed to the program and the group not exposed to the program. In addition, both groups should react to the program in the same way if they are exposed to it. If this is the case, the average causal effect of the program over the population can be estimated from the difference in average outcomes between the sample units exposed to the program and the sample units not exposed to the program. This means that we can replace the theoretical and unmeasurable treatment effect of the program on a specific individual unit *i* with the estimated average treatment effect of the program for a sample of units drawn from a population of such units.

In the case of HISP, if the condition of independence of potential outcomes holds and there is no selection bias, the average causal effect of providing health insurance subsidies to poor rural households is the difference in household yearly out-of-pocket health expenditures for poor rural households participating in the subsidy program and poor rural households not participating in the subsidy program.

How can we ensure that the condition of independence of potential outcomes is fulfilled, and rule out selection bias? Randomized assignment of the program provides one solution. When implemented properly with a large number of units, randomized assignment ensures that the average characteristics of the treatment group are similar to the average characteristics of the comparison group. As the number of units used for randomized assignment grows, on average the treatment and comparison groups will look more and more like the original population they are drawn from. The two groups will be expected to have the same baseline outcomes, and to react to the treatment in the same ways. As such, randomized assignment of the program (given a sufficiently large number of units) is one way to ensure that the condition of independence of potential outcomes is fulfilled, and rule out selection bias. In this case, the difference in average outcomes between the randomized treatment and comparison groups is a consistent estimator of the average treatment effect for the population.

[7] Formally, $(\overline{Y_0}|P=1) = (\overline{Y_0}|P=0)$.

Box 1. The Selection Bas

To see when the difference in average outcomes between the treatment and comparison groups can consistently estimate the average treatment effect, we can rewrite the average treatment effect of the program over the population as:

$$ATE_I = E_I(Y_{i1}) - E_I(Y_{i0}) = E_I(Y_{i1}|P = 1) - E_I(Y_{i0}|P = 0)$$

$$= E_I(Y_{i1}|P = 1) - \boldsymbol{E_I(Y_{i0}|P = 1)} + \boldsymbol{E_I(Y_{i0}|P = 1)} - E_I(Y_{i0}|P = 0)$$

$$= E_I((Y_{i1} - Y_{i0})|P = 1) + E_I(Y_{i0}|P = 1) - E_I(Y_{i0}|P = 0)$$

= (Average treatment effect on the treated) + (Selection bias)

In other words, the difference in average outcomes between a group exposed to the program and a group not exposed to the program is the sum of the average treatment effect on the treated (i.e. on those participating in the program) plus the selection bias. The selection bias is zero when there is no difference in average Y_{i0} between those who did and did not receive the program: namely, when $E_I(Y_{i0}|P = 1) - E_I(Y_{i0}|P = 0) = 0$. When there is no selection bias, the difference in average outcomes between groups provides a consistent estimate of the average treatment effect in the population. This is achieved under the conditional independence assumption, in which case:

$$ATE_I = E_I((Y_{i1} - Y_{i0})|P = 1)$$

Source: Angrist and Pischke (2009).

Randomized Assignment in a Regression Framework

In the potential outcome framework discussed above, the observed outcome for a unit is either one of two potential outcomes. The potential outcome for unit *i* is Y_{i1} when exposed to the program, which happens with probability p_i. By contrast, the potential outcome for that same unit *i* is Y_{i0} when it is not exposed to the program, which happens with a probability 1-p_i. The observed outcome for unit *i* can be written as the following average of the two potential outcomes:

$$Y_i = p_i Y_{i1} + (1 - p_i)Y_{i0} = Y_{i0} + (Y_{i1} - Y_{i0})p_i$$

This relationship between the outcome of interest and exposure to the program can be rewritten in a linear regression framework:

$$Y_i = \alpha + \delta P_i + \varepsilon_i \qquad (1)$$

In the regression framework, α represents the average outcome for the group not exposed to the program (that is, when P_i=0). δ represents the difference in average outcomes between the group exposed to the program and the group not exposed to the program. Lastly, the error term (ε_i) captures any other individual factors that may affect the relationship between the program and the outcome of interest.

This reformulation is akin to testing the difference in average outcomes between two groups. To see why, consider the group of individuals exposed to the program (P_i=1), and the group of individuals not exposed to the program (P_i=0). We can measure the average outcomes for each group as follows:

- For the group of individuals not exposed to the program: $E(Y_i|P_i = 0) = \alpha + E(\varepsilon_i|\ P_i = 0)$
- For the group of individuals exposed to the program: $E(Y_i|P_i = 1) = \alpha + \delta + E(\varepsilon_i|\ P_i = 1)$

Taking the difference between the two groups gives the following:

$$E(Y_i|P_i = 1) - E(Y_i|P_i = 0) = \delta + E(\varepsilon_i|\ P_i = 1) - E(\varepsilon_i|\ P_i = 0)$$

This illustrates that δ provides an estimate of the average treatment effect through the difference in average outcomes between program beneficiaries and nonbeneficiaries. However, this estimate is consistent only if there is no selection bias: that is, when unobserved differences between treated and nontreated groups cancel out, in which case $E(\varepsilon_i|\ P_i = 1) = E(\varepsilon_i|\ P_i = 0)$. If there is a selection bias, then the difference in outcomes between beneficiaries and nonbeneficiaries will include the effect of the program plus a selection bias.

Through impact evaluations, we aim to estimate the causal effect of the program and rule out any potential selection bias. In cases when the program is randomly assigned, and randomization is performed on a sufficiently large number of units, then the error terms in the regression are expected to be distributed randomly. This is equivalent to saying that the difference in the error terms between treatment and comparison groups will cancel out. In other words, when the assumption of independence of potential outcomes holds, then the impact of the program can be estimated by a difference in average outcomes in the treatment and comparison groups. In a regression framework such as equation (1) above, this difference is captured by the parameter of interest (δ).

We now discuss how this can be implemented in practice in the presence of a randomized program (with full compliance) and follow-up data. In this case, the impact of a program can be estimated through an ordinary least squares (OLS) regression of the outcome of interest on a binary variable that takes the value of 1 if the individual is exposed to the program, and 0 otherwise.

In the HISP case, we can estimate the difference in average outcomes between two groups by running a regression of household yearly out-of-pocket health expenditures on a binary variable capturing

exposure to the (randomized) program. This can be done among the group of eligible households in the treatment and comparison groups.

The estimated regression coefficient of δ (see equation 1) provides an estimate of the program impact. The statistical significance of the program impact can be assessed by a standard t-test for that regression coefficient. Specifically, significance of the program impact is assessed by testing the null hypothesis that $\delta = 0$ against the alternative hypothesis that $\delta \neq 0$. This test yields a t-statistic, and related p-value, that can be used to report on the results.

To implement this in Stata (see Stata Example 1), we regress our outcome of interest (health expenditures) on a binary treatment variable (treatment_locality), which is equal to 1 if the household is located in a treatment area. We run this regression on the sample of households that were eligible for HISP and based on data collected after the program has been administered (round = 1). The estimated regression coefficient for δ is -10.14, indicating that eligible households exposed to HISP spent $10.14 less on health expenditures than eligible households in the comparison group. The standard error of the coefficient is 0.396.[8] The t-statistic calculated in the regression, -25.63, shows that this coefficient is statistically significant at the 1 percent level.

```
Stata Example 1. Randomized Assignment in a Regression Framework (Linear Regression)

* In this context, the program is randomized at the village level, and you compare
follow-up situation of eligible households in treatment and comparison villages.

*Select the relevant data
use "evaluation.dta", clear
keep if eligible==1

reg health_expenditures treatment_locality if round ==1, cl(locality_identifier)

Linear regression                               Number of obs =    5629
                                                F(  1,   196) =  656.77
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3004
                                                Root MSE      =  7.7283

                      (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------
                     |               Robust
health_expenditu~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |  -10.14037   .3956824   -25.63   0.000    -10.92071    -9.36003
             _cons |   17.98055   .3066373    58.64   0.000     17.37582    18.5852
```

[8] Because the treatment is assigned by village and outcomes are observed at the household level, we cluster the standard errors by village.

Testing Baseline Balance in a Regression Framework

As mentioned, independence of potential outcomes is one of the most crucial assumptions to ensure that the difference in average outcomes between program beneficiaries and nonbeneficiaries provides a consistent estimate of the average treatment effect. While this assumption cannot generally be verified, some falsification tests can be implemented to identify cases when it does not hold.

For instance, if data are available on program beneficiaries and nonbeneficiaries before the program, the same regression framework can be used to test that both groups are indeed similar prior to the program. We denote the time period prior to the program as *t*=0. In this context, the difference in means can be estimated between the two groups prior to the program through a regression similar to the one above, although for outcomes measured prior to the program:

$$Y_{i,t=0} = \alpha + \delta P_i + \varepsilon_i$$

In this regression, we focus on baseline data collected before the program is rolled out (round = 0) to compare pre-program outcomes among eligible households in the treatment and comparison groups. Stata Example 2 illustrates the approach and results. When the baseline outcome is regressed on a binary variable capturing exposure to treatment, the estimated coefficient is very small (-0.084) and not statistically significant (p value of 0.693). This indicates that eligible households in the treatment and comparison groups have similar levels of health expenditures prior to the intervention[9].

```
Stata Example 2. Testing for Balance in a Baseline Outcome

reg health_expenditures treatment_locality if round ==0, cl(locality_identifier)

Linear regression                                     Number of obs =    5628
                                                      F(  1,   196) =    0.16
                                                      Prob > F      =  0.6933
                                                      R-squared     =  0.0001
                                                      Root MSE      =  4.3012

                       (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------
                    |               Robust
health_expenditu~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |  -.0841528   .2130966    -0.39   0.693    -.5044093    .3361037
             _cons |   14.57385   .1560665    93.38   0.000     14.26606    14.88163
------------------------------------------------------------------------------
```

By the same token, the regression framework can be used to check that other characteristics observed prior to the program are indeed similar across beneficiaries and nonbeneficiaries. If *X* is another such characteristic, balance in this other observed pre-program characteristic can be estimated as follows:

$$X_{i,t=0} = \alpha + \delta P_i + \varepsilon_i$$

[9] You may notice that there is a small difference in the number of observations in Stata Example 2 compared to Stata Example 1. This is because there is one observation with data at follow-up but not at baseline, so it is dropped from Stata Example 2.

This regression is essentially the same as the one displayed in Stata Example 2, only with a different dependent variable. In example 3 below, the treatment variable is regressed on the age of the head of household. The estimated coefficient is again not significant, indicating that the age of the head of household is not statistically different between eligible households in the treatment and comparison groups.

```
Stata Example 3. Testing for Balance in a Baseline Covariate

reg age_hh treatment_locality if round ==0, cl(locality_identifier)

Linear regression                                      Number of obs =    5628
                                                       F(  1,   196) =    1.42
                                                       Prob > F      =  0.2341
                                                       R-squared     =  0.0005
                                                       Root MSE      =  14.044

                        (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------
             |               Robust
      age_hh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
treatment_~y |  -.6354625   .5324145    -1.19   0.234    -1.685459    .4145341
       _cons |   42.29204   .4300065    98.35   0.000     41.44401    43.14008
------------------------------------------------------------------------------
```

The estimation can be repeated over a large set of pre-program characteristics. An alternative is to run a regression of a variable capturing exposure to the program on a range of baseline characteristics.

Multivariate Regression

Let us assume that, in addition to the outcome variable (Y_i), you also observe a range of other characteristics for individuals (X_i) in the sample. These can be added as control variables in the regression above. To do so, you can run a multivariate regression including a set of control variables:

$$Y_i = \alpha + \delta P_i + \gamma X_i + \varepsilon_i$$

Stata Example 4 illustrates how to implement this regression. The dependent variable remains the outcome of interest, health expenditures. In addition, we include a range of control variables: age of household head (age_hh), age of spouse (age_sp), education level of household head (educ_hh), education level of spouse (educ_sp), whether the head of household is a female (female_hh = 1), whether the household is a member of an indigenous group (indigenous = 1), the household size (hhsize), whether the household home has a dirt floor (dirtfloor = 1), whether the home has a bathroom (bathroom = 1), how many hectares of land the household owns (land), and the distance to the closest hospital (hospital_distance). The estimated regression coefficient for δ is -10.01, indicating that eligible

households exposed to HISP in a treatment community spent $10.01 less on health expenditures than eligible households in communities not exposed to the program, holding all the control variables constant. The *t*-statistic calculated in the regression shows that this coefficient is statistically significant at the 1 percent level.

```
Stata Example 4. Randomized Assignment in a Regression Framework (Multivariate
Regression)

reg health_expenditures treatment_locality age_hh age_sp educ_hh educ_sp female_hh
indigenous hhsize dirtfloor bathroom land hospital_distance if round ==1,
cl(locality_identifier)

Linear regression                                      Number of obs =    5629
                                                       F( 12,   196) =  135.95
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4297
                                                       Root MSE      =  6.9844

                        (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------
                     |               Robust
health_expenditu~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |   -10.01032   .3412294   -29.34   0.000    -10.68327   -9.337363
            age_hh |    .0411975   .0146714     2.81   0.005     .0122635    .0701316
            age_sp |    .0031789   .0171833     0.18   0.853    -.0307089    .0370667
           educ_hh |   -.0389384   .0468306    -0.83   0.407    -.1312951    .0534182
           educ_sp |   -.0223849    .049144    -0.46   0.649    -.1193038    .0745339
         female_hh |    .6430682   .4442419     1.45   0.149    -.2330395    1.519176
        indigenous |   -1.905311   .3496098    -5.45   0.000    -2.594791   -1.215831
            hhsize |   -1.603432   .0655058   -24.48   0.000    -1.732619   -1.474245
          dirtfloor |   -1.849394   .2776092    -6.66   0.000    -2.396878   -1.301909
           bathroom |    .2850031   .2463569     1.16   0.249    -.2008474    .7708537
               land |    .0380483   .0376021     1.01   0.313    -.0361083     .112205
 hospital_distance |   -.0026034   .0042288    -0.62   0.539    -.0109431    .0057363
              _cons |    27.56544   .8635174    31.92   0.000     25.86246    29.26841
------------------------------------------------------------------------------
```

In the case of HISP, the coefficient estimated by multivariate regression in Stata Example 4 (-10.01) is very similar to the coefficient estimated by bivariate regression in Stata Example 1 (-10.14). This is because the HISP program was randomized in the first place, so that the additional covariates do not affect the consistency of the estimator.

Generally, if the program is randomized, most other characteristics would be expected to be balanced in the treatment and comparison groups. As such, controlling for them is not expected to affect the impact estimates. When randomized assignment is performed, a few rare baseline differences may still be observed. In this case, the multivariate regression can help control for them.

While adding control variables in the regression model should does not affect the consistency of the estimates under randomized assignment, their inclusion may increase the precision of the estimation

and the overall statistical power of the estimation.[10] For instance, the standard error in Stata Example 1 is 0.396, while the standard error in Stata Example 4 is 0.341. In this case, the estimates from the multivariate regression are slightly more precise.

When a program is not randomized, a common approach (mis)-used to attempt to estimate treatment effects is to run a multivariate regression with program participation as a dependent variable, and a range of control variables as regressors. However, this approach is often fraught with selection bias and is unreliable. Indeed, in this case, the regression analysis does not necessarily represent a causal relationship between participation in the program and the outcome variable. For instance, some differences may be observed between the treatment and comparison group at baseline. To give an example, households choosing to participate in the Health Insurance Subsidy Program may be more likely to be sick than households that choose not to participate. Yet such differences in health can also affect households' health expenditures, which is the outcome of interest. If there are such omitted variables that also explain part of the outcomes of interest, the regression coefficient will capture an association that may not be entirely causal: the estimated coefficient of the treatment variable confounds the effect of the program, along with the indirect effect of the omitted variable.

If the program is not randomized, controlling for observed characteristics may not be sufficient. Indeed, we can never be sure that proxy variables are included in the regression for all the unobserved or unobservable characteristics that can mediate the relationship between the program and final outcomes. In the HISP example, we may be able to control for household characteristics such as households' head and gender. However, we may not have a variable in the dataset that measures the household head's health status. This unobserved variable may confound program impacts in cases where the program has not been successfully randomized. As such, adding control variables to a regression does not provide assurance that all relevant variables have been accounted for. A potential selection bias cannot be fully ruled out. We return to these issues below. Specifically, we discuss what do to in cases when the program has not been randomized. For instance, one option is to use a difference-in-differences approach that accounts for unobserved time-invariant characteristics. Before turning to the difference-in-differences method, we discuss what happens when not all units eligible and exposed to a program take it up.

Instrumental Variables

Estimation of Intent-to-Treat and Local Average Treatment Effect in the Presence of Noncompliance

Up to now, we have not distinguished between exposure to the program and participation in the program. We have assumed that all units eligible and exposed to the program participate in the program. We have also assumed that we could identify the sets of units eligible for the program in both the treatment and comparison groups. If this is the case, and when the program is successfully randomized, an equation of the following form can be estimated:

[10] Bruhn and McKenzie (2009).

$$Y_i = \alpha + \delta P_i + \gamma X_i + \varepsilon_i$$

The estimate of δ provides a consistent estimate of the average treatment effect. In the case of the randomized Health Insurance Subsidy Program (HISP), we obtained this estimate by comparing average outcomes among eligible households that are exposed to the program and participate in the program in the treatment group, and eligible households in the comparison group that are not exposed and cannot participate in the program.

We now consider the case where there is noncompliance: that is, when some units exposed to the program do not participate in the program (noncompliance in the treatment group), or some units not exposed to the program participate (noncompliance in the comparison group). In the presence of noncompliance, different parameters can be estimated. An important distinction is between "intent-to-treat" (ITT) estimates and "local average treatment effect" (LATE) estimates. In the presence of full compliance, these two estimates are equivalent to the "average treatment effect" (ATE) discussed earlier. However, in the presence of noncompliance, they are different and need to be interpreted as such. In the rest of this section, we discuss how to obtain these two types of estimates, and how to interpret what they mean. In particular, we illustrate the difference between ITT and LATE estimates in the context of the HISP example.

In the second section, we illustrated how to estimate treatment effects in a specific case. We assumed that we could observe households that were eligible for the program in the treatment and comparison communities. In addition, we also assumed full compliance with treatment assignment: we assumed that all eligible households in the treatment group participated in the program, and no eligible households in the comparison group participated. In this context, we obtained an estimate of program impacts by comparing eligible households in the treatment and comparison groups. However, in real-life contexts, we often do not know exactly who in the treatment group will participate in the program, and who in the comparison group would participate in the program if that program was offered. In addition, when a program is randomized, not all households offered the program will take it up, and imperfect compliance with treatment assignment is common.

Let's assume that we randomize a program by randomly assigning villages to the treatment and comparison groups. Let's further assume that all households in treatment villages can participate in the program if they so wish, but that some decide to participate and others choose not to do so. In this context, there is some noncompliance in the treatment group where some households offered the program do not take it up. In the case of HISP, approximately 59 percent of households participate, and 41 percent do not participate in treatment village. For simplicity, we still assume there is no noncompliance in the comparison group, where no one participates in the program.

How can we estimate the treatment effects in this case? One option is to run the same regression as the one described above, where P_i describes exposure to the program, independently of whether a unit participates in the program or not. In this case, an estimate of the treatment effect δ can be obtained. By relying on randomization of the program, the condition of independence of potential outcomes remains, so that we still obtain a consistent estimate. However, the interpretation of the estimate will

change. Indeed, we are now comparing all units in the treatment groups to all units in the comparison group, independently of whether they participate in the program or not. This estimate represents an "intent-to-treat" estimate (δ_{ITT}) capturing the impact of offering a program to an average unit in the treatment group, whether this unit participates or not in the program.

Stata Example 5 illustrates how to obtain the ITT estimates in a regression framework for the HISP example. We regress our outcome of interest, health expenditures, on the binary variable treatment_locality (which equals 1 if the household is in a treatment area, independently of whether or not it took up the program). For the regression, we use data collected after the program has been administered (round = 1). The estimate for the regression coefficient (δ) is -6.4, indicating that households in villages where HISP was offered on average spent $6.4 less on health expenditures than households in villages where HISP was not offered. This is the ITT estimate for the impact of offering the program to the treatment group. Remember that the program "intended to treat" or offered the treatment to all households in treatment communities. However, 59 percent of households enrolled, but 41 percent of households did not participate. Therefore, when computing average outcomes in the treatment group, we obtain a weighted average of outcomes among the 59 percent of households that participated in the program and the 41 percent of households that chose not to participate in the program. The "intent-to-treat" (ITT) estimate is then obtained by taking the difference between the weighted average in the treatment group and the same weighted average of outcomes in the comparison group. As a result, it captures the "intent-to-treat" impacts of offering the program to an average household in a treatment locality, independently of who participates within the locality.

```
Stata Example 5. "Intent-to-Treat" (ITT) Estimates

* In this context, the program is randomized at the village level.
* While everyone is eligible for the program in treatment communities, not everyone
participates.

*Select the relevant data
use "evaluation.dta", clear
drop eligible

* You can estimate 'intent-to-treat estimates', i.e. program impact at the village-
level irrespective of who takes up the program or not.

reg health_expenditures treatment_locality if round ==1, cl(locality_identifier)

Linear regression                                      Number of obs =    9914
                                                       F(  1,   199) =  163.93
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0726
                                                       Root MSE      =  11.451

                        (Std. Err. adjusted for 200 clusters in locality_identifier)
------------------------------------------------------------------------------
                     |               Robust
health_expenditu~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------------+------------------------------------------------------------------
treatment_locality |  -6.406008   .5003313   -12.80   0.000     -7.39264   -5.419376
             _cons |   20.06416   .3763165    53.32   0.000     19.32208    20.80624
-------------------------------------------------------------------------------------
```

In contrast, the "local average treatment effect" (δ_{LATE}) seeks to estimate treatment effects only on the subgroup of units that actually comply with their treatment assignment. In the case above, in the treatment group the compliers are the 59 percent of households that were offered the program and indeed took it up. The "local average treatment effect" captures the estimated program impacts on this subgroup. It provides consistent estimates of the program impacts on the population whose take-up behavior is affected by exposure to the treatment.[11] In presence of noncompliance, the "local average treatment effect" estimates—that is, the average impact of the program if it was estimated on the groups actually participating in the program—are different than the "intent-to-treat estimates." We now provide the intuition on how to obtain the LATE estimates, before formally discussing how to obtain them in a regression framework in the next section.

In general, the estimate for the "local average treatment effect" (δ_{LATE}) will be larger in absolute terms than the "intent-to-treat" estimate (δ_{ITT}). This is easily explained by the fact that the magnitude of the ITT estimates is driven mostly by complier units that are exposed to the program and participating in it. Assuming that program impacts are positive, if fewer units comply and participate in the treatment group, the average outcomes over the whole treatment group will be relatively smaller, and so will be the ITT estimates. In fact, it can be shown that the "intent-to-treat" ITT estimate is equal to the share of compliers multiplied by the "local average treatment effect" LATE estimate. In presence of full compliance (100 percent take-up in the treatment group), the ITT and LATE estimates will be the same. The relationship between the "intent-to-treat" ITT estimates and the "local average treatment effect" LATE estimates can be written as follows:

$$\delta_{ITT} = \text{Take-up} * \delta_{LATE} \qquad (2)$$

Stata Example 5 can be used to illustrate the link between the LATE and ITT estimates. As mentioned , the local average treatment effect can be estimated by dividing the intent-to-treat estimates by the share of the group exposed to the program that effectively takes it up. The ITT estimate from Stata Example 5 is -6.4. The take-up rate among the population that was offered the program is 59 percent. The LATE estimate is conceptually the program impact if the entire population offered the program had taken it up. It can be retrieved from equation (2) by rescaling the ITT estimate by the share of the treatment group that took up the program: -6.4/0.59 = 10.7. This is equivalent to calculating what would have been the program impact if 100 percent of the treatment group had participated in it, instead of only 59 percent.

This estimate is quite close to the impact estimate obtained in Stata Example 1. This is not surprising, because in Stata Example 1 we had assumed that we could identify eligible households that would take up the program if offered in both the treatment and comparison groups. This is not information that is

[11] As such, it is typically not generalizable to the entire population of interest.

typically available, however. In practice, additional steps are needed to obtain the LATE estimates. Instrumental variable techniques provide one approach to obtain the LATE estimate. In the rest of this section, we discuss how instrumental variable techniques can help address cases of randomized programs with noncompliance, and in particular how they can be used to estimate the average program impact on the treated (δ_{LATE}) when some units offered the program do not participate. As we will explain at the end of the section, instrumental variable techniques also constitute the method to analyze data generated by a randomized promotion design.

Instrumental Variables, Local Average Treatment Effects, and Imperfect Compliance

Consider the original regression model:

$$Y_i = \alpha + \delta P_i + \gamma X_i + \varepsilon_i$$

P_i is now a binary variable describing actual participation in the program (not just exposure to a program being offered), taking the value of 1 for participants, and 0 for nonparticipants. X_i is a set of control variables.

Let us model the decision to participate in the program as follows:

$$P_i = \pi_0 + \pi_1 X_i + \pi_2 M_i + \vartheta_i$$

In this equation, participation in the program is modelled as a function of observed characteristics that are also controlled for in the original regression model (X_i), as well as an additional set of characteristics not included in the original model (M_i). In the HISP example, we may be able to control for household characteristics such as households' head and gender. However, there may be another variable, which cannot necessarily be accounted for in the original regression, but may affect participation program. In the example provided earlier in the case of HISP, the household head's health status could be one example of such a variable that is hard to observe.

If the original model is estimated without controlling for unobserved variables that affect program participation, it would lead to a biased estimate of the program impact when there is a correlation between M_i and P_i, that is when the unobserved variable also affects the output of interest. Indeed, the independence of potential outcomes would be violated in this case. A selection bias (or endogeneity bias) is introduced by the correlation between program participation and a third factor that is not included in the original regression. As such, the coefficient for program participation in the first equation would mix not only the causal effect of the program, but also the indirect effect of the additional factor that is not accounted for but correlated with the outcome. In this case, participation to the program is endogenous rather than being independent or exogenous to the potential outcomes.

In order to deal with this case of selection into the program, an instrumental variable (*Z)* is required. An instrumental variable (*Z)* has the following two essential properties:

1. Corr(Z,P)≠0: the instrumental variable affects the chance of an individual being offered the program to actually participate in the program. For an instrument to be valid, the correlation between the instrumental variable and program participation needs to be large. Instruments that are only weakly correlated are not appropriate.
2. Corr(Z, ε_i)=0. There is no correlation between the instrumental variable and the outcome of interest, apart from its effect on the probability to participate in the program.

If such an instrumental variable exists, the local average treatment effect (LATE) estimate can be obtained through the "two-stage least square" (2SLS) estimator.

In the first stage, we isolate the effect of the instrumental variable on program participation by estimating the following regression:

Stage 1: $$P_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \tau_i$$

In the second stage, we regress the outcome of interest on the predicted value of program take-up (P_i) from the first stage:

Stage 2: $$Y_i = \alpha + \delta_{LATE} \widehat{P_i} + \gamma X_i + \varepsilon_i$$

Given the properties of the instrumental variable, the first stage is used to "clean up" P_i of its endogeneity. In the second stage, the remaining exogenous variation in program participation driven by the instrumental variable is used to identify the impact of the program. It can be shown that the 2SLS estimator is a consistent estimator of the local average treatment effect.[12]

Instrumental variables are hard to come by in real-life settings. The randomization of a program provides one example of an instrumental variable. If a program is randomized, but there is imperfect compliance, the randomized exposure to the program (Z_i) can be used as an instrumental variable for actual program participation (P_i). In this case, the application of the 2SLS estimator generates an estimate of the local average treatment effect (δ_{LATE}). To summarize:

1. Stage 1 estimates the effect of offering the program on actual participation:

$$P_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \tau_i$$

2. Stage 2 provides the local average treatment effect estimates (δ_{LATE}):

$$Y_i = \alpha + \delta_{LATE} \widehat{P_i} + \gamma X_i + \varepsilon_i$$

[12] Given the estimation procedure, the standard errors of the second-stage OLS need to be corrected.

This is different from the estimate that would be obtained through a regression of the outcome on the randomized program (Z_i) in presence of non-compliance, which would provide an estimate of the intent-to-treat effects (δ_{ITT}) through the following regression.

$$Y_i = \pi_0 + \delta_{ITT} Z_i + \pi_2 X_i + \varepsilon_i$$

Stata Example 5 illustrated this approach for the HISP case study, and provided an estimate for δ_{IIT} (which was found to be -6.4).

By contrast, Stata Example 6 illustrates how to implement the instrumental variables (IV) estimator and obtain the LATE estimate in the context of HISP. Here, our instrumental variable (treatment_locality) takes the value of 1 if HISP was randomly offered to households in a given locality, and 0 otherwise. In the first-stage regression, a variable capturing whether a household was enrolled or not in the program is regressed on the randomization dummy (treatment_locality). The coefficient, 0.598 indicates that approximately 59.8% of households enrolled in HISP when the program was offered in their locality. The second stage regression uses the predicted enrollment from the first stage as a regressor to explain variation in the outcomes of interest. The estimated coefficient for δ_{LATE} suggests that participation in the HISP program lowers health expenditures by $10.7. This is the same number that we obtained by multiplying the ITT estimate by the share of units enrolled in the treatment group using equation (2).

Stata Example 6: "Local Average Treatment Effect" (LATE) 2SLS IV Estimates[13]

```
* You can back out 'local average treatment effect' estimates on complier units that
do take-up the program in treatment communities

ivreg health_expenditures (enrolled = treatment_locality) if round ==1, first

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

      Source |       SS       df       MS              Number of obs =    9914
-------------+------------------------------           F(  1,  9912) = 7361.23
       Model |  885.675858     1  885.675858           Prob > F      =  0.0000
    Residual |   1192.5756  9912  .120316344           R-squared     =  0.4262
-------------+------------------------------           Adj R-squared =  0.4261
       Total |  2078.25146  9913  .209649093           Root MSE      =  .34687

------------------------------------------------------------------------------
       enrolled_ro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |   .5977823   .0069674    85.80   0.000      .5841248    .6114397
             _cons |   2.12e-14   .0049282     0.00   1.000    -.0096602    .0096602
------------------------------------------------------------------------------
```

[13] The Stata command used here is ivreg. A more recent version of the command is called ivregress 2sls. The do-file for the technical companion provides the syntax for both commands. Note that the option "first" added at the end of the command ensures that the results of the first-stage of the two-stage estimation process are also displayed.

```
Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =    9914
-------------+------------------------------           F(  1,  9912) =  944.81
       Model |  334474.116     1  334474.116           Prob > F      =  0.0000
    Residual |  1067039.56  9912  107.651288           R-squared     =  0.2387
-------------+------------------------------           Adj R-squared =  0.2386
       Total |  1401513.68  9913  141.381386           Root MSE      =  10.376

------------------------------------------------------------------------------
health_exp~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   Enrolled  |  -10.71629    .348636   -30.74   0.000    -11.39969   -10.03289
       _cons |   20.06416   .1474116   136.11   0.000      19.7752    20.35312
------------------------------------------------------------------------------
Instrumented:  enrolled
Instruments:   treatment_locality
------------------------------------------------------------------------------
```

Instrumental Variables and Randomized Promotion

Randomized promotion is another example of an instrumental variable. In this case, it is not the program itself that is randomized, but rather an information campaign or other types of encouragement that can increase program take-up. When randomized promotion is used, the approach essentially seeks to generate a valid instrumental variable that is correlated with program participation, but uncorrelated with the outcome of interest, aside from its effect through participation. As such, the same two-step estimation methodology applies as the one outlined above. In the first stage, we isolate the effect of randomized promotion on program participation. In the second stage, we regress the outcome of interest on the predicted value of program take-up from the first stage: that is, the variation in program take-up that is driven by the randomized promotion campaign.

We now illustrate the use of randomized promotion to evaluate the impact of HISP. To do this, let's assume that we are in a different context than the one we have used for HISP so far. Let's now assume that HISP is offered universally throughout the country. In this case, the program itself is not randomized. However, let's assume that a randomized promotion campaign takes place. An intensive promotion effort is undertaken in a random subsample of villages, including communication and social marketing campaigns aimed at increasing awareness of HISP. This promotion campaign is an instrumental variable, with the two necessary properties mentioned earlier. First, it seeks to increase enrollment in HISP. Second it does not directly affect the outcome indicator of interest (health expenditures).

With an instrumental variable created by randomized promotion, the local average treatment effect (LATE) estimate can be obtained through the "two-stage least square" estimator. Stata Example 7 illustrates the two stages involved. The first stage identifies the effects of the promotion activities on program take-up. In this case, promotion activities increase program take-up by 40.8 percent. In the second stage, we regress the outcome variable on the predicted program participation from the first

stage to obtain the LATE estimates. In this case, the results suggest that participation in the HISP program lowers health expenditures by $9.5.

Stata Example 7. 2SLS IV Estimates for Randomized Promotion[14]

```
* In this context, everyone is eligible for the program. You compare what happens in
promoted and non-promoted villages.

*Select the relevant data
use "evaluation.dta", clear
drop eligible
drop treatment_locality
drop enrolled

ivreg health_expenditures (enrolled_rp = promotion_locality) if round ==1, first

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

      Source |       SS       df       MS              Number of obs =    9914
-------------+------------------------------           F(  1,  9912) = 2484.60
       Model |  411.879408     1  411.879408           Prob > F      =  0.0000
    Residual |  1643.13855  9912  .165772654           R-squared     =  0.2004
-------------+------------------------------           Adj R-squared =  0.2003
       Total |  2055.01795  9913  .207305352           Root MSE      =  .40715

------------------------------------------------------------------------------
       enrolled_rp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
promotion_locality |   .4077847   .0081809    49.85   0.000     .3917484     .423821
             _cons |   .0842476   .0058578    14.38   0.000      .072765    .0957301
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =    9914
-------------+------------------------------           F(  1,  9912) =  337.77
       Model |  310737.314     1  310737.314           Prob > F      =  0.0000
    Residual |  1090776.36  9912  110.046042           R-squared     =  0.2217
-------------+------------------------------           Adj R-squared =  0.2216
       Total |  1401513.68  9913  141.381386           Root MSE      =   10.49

------------------------------------------------------------------------------
health_exp~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 enrolled_rp |  -9.499769   .5168948   -18.38   0.000    -10.51299    -8.48655
       _cons |   19.64571   .1846287   106.41   0.000      19.2838    20.00762
------------------------------------------------------------------------------
Instrumented:  enrolled_rp
Instruments:   promotion_locality
```

[14] The Stata command used here is ivreg. A more recent version of the command is called ivregress 2sls. The do-file for the technical companion provides the syntax for both commands. Note that the option "first" added at the end of the command ensures that the results of the first-stage of the two-stage estimation process are also displayed.

Difference-in-Differences in a Regression Framework

So far, we have focused on contexts where a program can be randomized (either with or without perfect compliance), or when randomized promotion can be used to affect program participation. In some cases, such randomized approaches are not possible. When pre-program and post-program data are available for a treatment group of program beneficiaries and a comparison group of nonbeneficiaries, difference-in-differences provides an alternative strategy to estimate program impacts.

Table 1 illustrates the set-up. There are two time periods (before and after the program), as well as two groups of households (those exposed to the program and those not exposed to the program). We denote the time period by a binary variable taking the value t=0 at baseline, and t=1 at follow-up. We continue to denote exposure to the program by a binary variable taking the value P=1 for beneficiaries, and P=0 for nonbeneficiaries. Table 1 illustrates how the outcomes for each individual may be written depending on the time period and the group s/he belongs to.

Table 1. Summary of the Difference-in-Differences Approach in a Table

	After the program starts (t=1)	Before the program starts (t=0)	Before-after comparison
Group exposed to the program (P=1)	$Y_{i,t=1}\,|\,P_i=1$	$Y_{i,t=0}\,|\,P_i=1$	$(\overline{Y}_{t=1}|P=1)-(\overline{Y}_{t=0}|P=1)$
Group not exposed to the program (P=0)	$Y_{i,t=1}|P_i=0$	$Y_{i,t=0}|P_i=0$	$(\overline{Y}_{t=1}|P=0)-(\overline{Y}_{t=0}|P=0)$

In this context, the before-after difference in outcomes among the group participating in the program is: $(\overline{Y}_{t=1}|P=1)-(\overline{Y}_{t=0}|P=1)$. The before-after difference in outcomes among the group not participating in the program is $(\overline{Y}_{t=1}|P=0)-(\overline{Y}_{t=0}|P=0)$. The difference-in-differences approach uses the before-after difference among the comparison group as a counterfactual for the before-after difference in the treatment group. This means that, in the difference-in-differences framework, program impacts are estimated by:

$$DD=[(\overline{Y}_1|P=1)-(\overline{Y}_0|P=1)] - [(\overline{Y}_1|P=0)-(\overline{Y}_0|P=0)]$$

The difference-in-differences estimator can also be presented in a regression framework. In general terms, consider the outcome Y_{igt} for an individual i at time t in group g (treatment or comparison). The regression model in this case can be written as:

$$Y_{igt} = \beta_1 P_i + \beta_2 t + \delta P_i t + \alpha_g + \theta_t + \varepsilon_{igt} \qquad (3)$$

As discussed, P_i constitutes a binary variable denoting exposure to the program and taking the value P=1 for beneficiaries, and P=0 for nonbeneficiaries. t constitutes a binary variable taking the value of 0 for pre-program measures, and 1 for post-program measures. β_1, β_2, and δ are the regression coefficients to be estimated. α_g is a time-invariant group-level fixed effect capturing differences between the treatment and comparison group that are time-invariant. θ_t is the time-invariant fixed effect capturing constant effects related to the each period. ε_{igt} is the error term.

In the case with two time periods, the outcomes for each of the four cases can be rewritten based on the regression above. Table 2 denotes these individual outcomes in each of the four cases.

Table 2. Summary of the Difference-in-Differences Approach in a Regression Framework

	After the program starts (*t*=1)	**Before the program starts (*t*=0)**	**Before-after comparison**
Group exposed to the program (*P*=1)	$Y_{i11} = \beta_1.1 + \beta_2.1 + \delta.1.1 + \alpha_1 + \theta_1 + \varepsilon_{i11}$	$Y_{i10} = \beta_1.1 + \beta_2.0 + \delta.1.0 + \alpha_1 + \theta_0 + \varepsilon_{i10}$	$\beta_2 + \delta + (\theta_1 - \theta_0) + (\varepsilon_{i11} - \varepsilon_{i10})$
Group not exposed to the program (*P*=0)	$Y_{i01} = \beta_1.0 + \beta_2.1 + \delta.0.1 + \alpha_0 + \theta_1 + \varepsilon_{i01}$	$Y_{i00} = \beta_1.0 + \beta_2.0 + \delta.0.0 + \alpha_0 + \theta_0 + \varepsilon_{i00}$	$\beta_2 + (\theta_1 - \theta_0) + (\varepsilon_{i01} - \varepsilon_{i00})$

The difference-in-differences approach relies on differences in the before-after comparisons between the group participating in the program and the group not participating in the program.

For the group participating in the program, the before-after comparison ($Y_{i11} - Y_{i10}$) is:

$$Y_{i11} - Y_{i10} = \beta_2 + \delta + (\theta_1 - \theta_0) + (\varepsilon_{i11} - \varepsilon_{i10})$$

By taking the before-after comparison, the time-invariant group fixed effect α_1 cancels out. The same happens when taking before-after comparison for the group not participating in the program:

$$Y_{i11} - Y_{i10} = \beta_2 + (\theta_1 - \theta_0) + (\varepsilon_{i01} - \varepsilon_{i00})$$

As discussed in chapter 3 of the book, by themselves, each of these before-after comparisons do not allow the causal effects of the program to be identified. Indeed, as the two equations above illustrate, the before-after comparison captures not only differences between the two groups, but also time fixed effects related to each period ($\theta_1 - \theta_0$).

However, by taking the difference between the before-after comparisons in the treatment and comparison groups, the time fixed effects capturing constant effects related to the each period also cancel out:

$$(Y_{i11} - Y_{i00}) - (Y_{i01} - Y_{i00}) = \delta + (\varepsilon_{i11} - \varepsilon_{i10} - \varepsilon_{i01} + \varepsilon_{i00})$$

The treatment effect estimated through the difference-in-differences estimator is the difference between the before-after comparison of average outcomes for the group participating in the program, and the before-after comparison of average outcomes for the group not participating in the program. Overall, the difference-in-differences estimator controls for time-invariant group fixed effects, as well as time fixed effects for each of the periods.

As for the basic regression discussed earlier, whether this estimator provides a consistent estimate of the causal effect of the program for the population depends on the properties of the error tem. Specifically, if the mean of the error term is 0, and if there is no time-varying factors affecting outcomes differently in the treatment and comparison groups, then δ provides a consistent estimate of program impacts.

Stata Example 8 illustrates how to estimate equation (3) using Stata for the HISP case. Let's assume that we have data only from localities where the program has been offered. In these localities, we have data both for households that participate in the program, as well as households that do not to participate. Data are available for a baseline survey collected before the program, and a follow-up survey collected after the program. To obtain the difference-in-differences estimates, we first create a new variable (enrolled_round), which is the interaction between participation in the program and the time at which the data are measured. The outcome variable is then regressed on this new variable, along with dummy variables capturing whether or not the household participated in the program, and the time at which each data point is observed. The coefficient of the new interaction variable is our difference-in-differences impact estimate. The results below indicate that health expenditures for households that enrolled in the program were $8.16 lower than among households that did not enroll.

```
Stata Example 8. Difference-in-Differences in a Regression Framework

* In this method, you compare the change in health expenditures over time
* between enrolled and nonenrolled households in the treatment localities.

*Select the relevant data
use "evaluation.dta", clear
keep if treatment_locality==1

gen eligible_round=eligible*round

reg health_expenditures eligible_round round eligible, cl(locality_identifier)

Linear regression                               Number of obs =    9919
                                                F(  3,    99) =  813.98
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3436
                                                Root MSE      =  7.9128
                    (Std. Err. adjusted for 100 clusters in locality_identifier)
------------------------------------------------------------------------------
             |               Robust
health_expen~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
eligible_round |  -8.162931   .3191368   -25.58   0.000    -8.796168   -7.529695
         round |   1.513416   .3564533     4.25   0.000     .8061355    2.220697
      eligible |    -6.3018    .193082   -32.64   0.000    -6.684917   -5.918684
         _cons |   20.79149   .1722887   120.68   0.000     20.44964    21.13335
------------------------------------------------------------------------------
```

In practice, in addition to the outcome of interest, we may also observe other characteristics for the treatment and comparison groups in both time periods. In this case, these characteristics are observed for each unit in each group and time period (X_{igt}), and can be included in the regression model to be estimated:

$$Y_{igt} = \beta_1 P_i + \beta_2 t + \delta P_i.t + \beta_3 X_{igt} + \alpha_g + \theta_t + \varepsilon_{igt} \qquad (4)$$

Replicating the same procedure as above, the difference between the before-after comparisons in the treatment and comparison group becomes:

$$(Y_{i11} - Y_{i00}) - (Y_{i01} - Y_{i00}) = \delta + (X_{i11} - X_{i10} - X_{i01} + X_{i00}) + (\varepsilon_{i11} - \varepsilon_{i10} - \varepsilon_{i01} + \varepsilon_{i00})$$

Stata Example 9 illustrates how to implement this regression in Stata. The estimate obtained after adding control variables in Stata Example 9 (-8.16) is very close to the difference-in-differences estimate without the control variables in Stata Example 8.

Stata Example 9. Difference-in-Differences in a Multivariate Regression Framework

```
reg health_expenditures eligible_round round eligible age_hh age_sp educ_hh educ_sp
female_hh   indigenous   hhsize   dirtfloor   bathroom   land   hospital_distance,
cl(locality_identifier)
```

```
Linear regression                                   Number of obs =    9919
                                                    F( 14,    99) = 2410.28
                                                    Prob > F      =  0.0000
                                                    R-squared     =  0.5516
                                                    Root MSE      =  6.5437

                      (Std. Err. adjusted for 100 clusters in locality_identifier)
------------------------------------------------------------------------------
                    |               Robust
health_expendit~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
   eligible_round |  -8.161499   .3197482   -25.52   0.000    -8.795949   -7.527049
            round |   1.450526   .3558662     4.08   0.000       .74441    2.156641
         eligible |   -1.51276    .129937   -11.64   0.000    -1.770583   -1.254937
           age_hh |   .0804852   .0113711     7.08   0.000     .0579224     .103048
           age_sp |  -.0197229   .0129787    -1.52   0.132    -.0454754    .0060297
          educ_hh |   .0599944   .0290694     2.06   0.042     .0023144    .1176743
          educ_sp |  -.0765127   .0339694    -2.25   0.027    -.1439153   -.0091101
        female_hh |   1.103935   .3157136     3.50   0.001     .4774905    1.730379
       indigenous |  -2.311985   .2361846    -9.79   0.000    -2.780627   -1.843344
           hhsize |  -1.994729   .0391445   -50.96   0.000      -2.0724   -1.917058
         dirtfloor |  -2.299839   .1632436   -14.09   0.000     -2.62375   -1.975929
          bathroom |   .5000436    .157629     3.17   0.002     .1872735    .8128137
              land |   .0909001    .028528     3.19   0.002     .0342943    .1475058
hospital_distance |  -.0031917   .0030591    -1.04   0.299    -.0092617    .0028783
            _cons |   27.39458   .5526554    49.57   0.000     26.29799    28.49117
------------------------------------------------------------------------------
```

So far, we have assumed that we observed both the treatment and comparison groups in two time periods, which allows correcting for group fixed effects. Sometimes, we are able to observe all individual

units in both groups in both time periods. In these cases when panel data are available, we can not only control for group fixed effects, but also for unit fixed effects. Let's assume we observe a unit *i* in two time periods:

- At time *t*=0: $Y_{ig0} = \beta_1 P_i + \beta_2 0 + \delta P_i.0 + \alpha_i + \theta_0 + \varepsilon_{ig0}$
- At time *t*=1: $Y_{ig1} = \beta_1 P_i + \beta_2 1 + \delta P_i.1 + \alpha_i + \theta_1 + \varepsilon_{ig1}$

Taking the difference over time: $Y_{ig1} - Y_{ig0} = \beta_2 + \delta P_i + (\theta_1 - \theta_0) + (\varepsilon_{ig1} - \varepsilon_{ig0})$

There are two ways to estimate this single-difference equation in Stata. First, the single-difference equation can be estimated using the xtreg fixed-effect panel data command in Stata (see Stata Example 10). Before using the xtreg command in Stata, we need to adjust the settings to a panel data format. We can set the panel and time variables using the xtset command. In order to estimate a fixed-effect model, we create a variable that is equal to 1 only for enrolled households in the follow-up period. The coefficient of this variable is our impact estimate: in this case, -8.16.

```
Stata Example 10. Calculating Difference-in-Difference Estimates by Taking the
Difference between Before-After Differences in the Treatment and Comparison Groups

xtset household_identifier round
       panel variable:  household_identifier (unbalanced)
        time variable:  round, 0 to 1
                delta:  1 unit

gen xtenrolled=0

replace xtenrolled=1 if enrolled==1 & round==1

xtreg health_expenditures xtenrolled round if treatment_locality==1, fe vce(cluster
locality_identifier)

Fixed-effects (within) regression               Number of obs      =      9919
Group variable: household_~r                    Number of groups   =      4960

R-sq:  within  = 0.2437                         Obs per group: min =         1
       between = 0.3852                                        avg =       2.0
       overall = 0.2805                                        max =         2

                                                F(2,99)            =    633.70
corr(u_i, Xb)  = 0.2632                         Prob > F           =    0.0000

                  (Std. Err. adjusted for 100 clusters in locality_identifier)
------------------------------------------------------------------------------
             |               Robust
health_exp~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  xtenrolled |  -8.163337   .3190874   -25.58   0.000    -8.796476   -7.530198
       round |   1.513416   .3564353     4.25   0.000     .8061711    2.220661
       _cons |   17.02477   .1203165   141.50   0.000     16.78603     17.2635
-------------+----------------------------------------------------------------
     sigma_u |  7.1354916
     sigma_e |  6.5169842
         rho |  .54521089   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The same single-difference can be estimated manually by computing the differences in the variables over time (see Stata Example 11). One way to do this is to first reshape the dataset from long to wide, so that each row of data includes only one unit. In the example below, you can keep only the variables you need for the estimation. We set the program participation variable equal to 0 at baseline, and manually calculate the difference between health expenditures and program participation in the baseline and follow-up rounds. We then run a regression of the difference in the outcome variable over time on the treatment dummy. The impact estimate is exactly the same as with the fixed-effect panel estimate (-8.16).

```
Stata Example 11. Single-Differences Estimates for Difference-in-Differences

keep health_expenditures treatment_locality locality_identifier enrolled
household_identifier round

reshape wide health_expenditures enrolled, i(household_identifier) j(round)
(note: j = 0 1)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                     9919   ->    4960
Number of variables                   6   ->       7
j variable (2 values)             round   ->   (dropped)
xij variables:
                    health_expenditures   ->   health_expenditures0
health_expenditures1
                              enrolled   ->   enrolled0 enrolled1
-----------------------------------------------------------------------------
gen dy = health_expenditures1 - health_expenditures0

replace enrolled0=0

gen dp = enrolled1-enrolled0

reg dy dp if treatment_locality==1, cl(locality_identifier)

Linear regression                                      Number of obs =    4959
                                                       F(  1,    99) =  654.51
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1588
                                                       Root MSE      =  9.2164

                  (Std. Err. adjusted for 100 clusters in locality_identifier)
------------------------------------------------------------------------------
             |               Robust
          dy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          dp |  -8.163337   .3190874   -25.58   0.000    -8.796476   -7.530198
       _cons |   1.513416   .3564353     4.25   0.000     .8061711    2.220661
------------------------------------------------------------------------------
```

Regression Discontinuity Design

In many programs, eligibility is defined based on a continuous score and whether a given unit's score is below or above a certain cut-off. Regression discontinuity can be used in contexts where a continuous running variable and a fixed threshold are used to determine eligibility for a program.

Let's assume that the running variable is X_i, and that the threshold is set to determine eligibility at a specific value X_0. This means that the value of P_i (whether a unit participates in the program or not) is entirely determined by the value of the running variable X_i. If the running variable is smaller than or equal to the threshold, then the unit is exposed to the program (Pi=1 if $X_i \leq X_0$). If the running variable is greater than the threshold, then the unit is not exposed to the program (Pi=0 if $X_i > X_0$). A regression discontinuity design is called "sharp" when the threshold is set so that all units below or above have the same treatment status, without any exception. (For a discussion of "fuzzy" discontinuity, see box 2.)

In this context, identification for sharp discontinuity relies on the following linear regression:

$$Y_i = \beta + \delta P_i + f(X_i) + \varepsilon_i$$

where P_i=1 if 1 if unit i participates in the program, and P_i=0 if unit i does not participate in the program. f(X_i) is a continuous function around the threshold, so that the value of this function tends to asymptotically equal on both sides of the threshold. A continuous function of the running variable is used in order to account for nonlinearities in the relationship between the running variable and the outcome of interest. ε_i represents a random error term.

- For a unit just at the cut-off: $Y_{i0} = \beta + \delta 0 + f(X_0) + \varepsilon_{i0}$
- For a unit just τ_i below the cut-off: $Y_{i1} = \beta + \delta 1 + f(X_0 - \tau_i) + \varepsilon_{i1}$

Therefore, the difference between the two is: $Y_{i1} - Y_{i0} = \delta + f(X_0 - \tau_i) - f(X_0) + (\varepsilon_{i1} - \varepsilon_{i0})$

Since the function f is continuous, as we get closer and closer to the cut-off X_0, f($X_0 - \tau_i$) will tend to f(X_0), and their difference will tend to 0. Therefore, the local average treatment effect at the threshold is estimated by δ.

Let's return to the HISP case and assume that eligibility for the program depends on a proxy poverty index. Data for the poverty index are available only in localities where the program will be offered. Households with a score below a certain cut-off (in this case, a value of 58) are chosen to participate in the program. Households with a score above that cut-off do not participate. We assume that this program eligibility rule is strictly enforced, without any exceptions of either side of the cut-off. In addition to the value of the index measured at baseline and prior to the program roll-out, we also measure the outcome of interest (health expenditures) after the end of the program.

Stata Example 12 illustrates how to obtain regression discontinuity estimates in a regression framework. We start by normalizing the poverty index threshold to 0 and create dummy variables for households with a poverty-targeting index to the left or right of the threshold. By doing so, we allow the relationship between the outcome variables and the running variable (the poverty index) to have different slopes on either side of the threshold. We then run a regression of health expenditures on a dummy variable capturing exposure to the program, as well as the two dummies for whether households have a poverty index to the left or to the right of the threshold. Applying this approach, Stata Example 12 shows that the estimate of the treatment effect (δ) is -11.19.

```
Stata Example 12. Regression Discontinuity Design Estimates

* REGRESSION DISCONTINUITY DESIGN
* In this context, you compare health expenditures at follow-up between households
just above
* and just below the poverty index threshold, in the treatment localities.

*Select the relevant data
use "evaluation.dta", clear
keep if treatment_locality==1

*Normalize the poverty index
gen poverty_index_left=poverty_index-58 if poverty_index<=58
(8570 missing values generated)

replace poverty_index_left=0 if poverty_index>58
(8570 real changes made)

gen poverty_index_right=poverty_index-58 if poverty_index>58
(11257 missing values generated)

replace poverty_index_right=0 if poverty_index<=58
(11257 real changes made)

reg health_expenditures poverty_index_left poverty_index_right eligible if round ==1

      Source |       SS       df       MS              Number of obs =    4960
-------------+------------------------------           F(  3,  4956) =  843.52
       Model |  257911.257     3  85970.4191           Prob > F      =  0.0000
    Residual |  505111.412  4956  101.919171           R-squared     =  0.3380
-------------+------------------------------           Adj R-squared =  0.3376
       Total |   763022.67  4959  153.866237           Root MSE      =  10.096

------------------------------------------------------------------------------
health_expenditures |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+---------------------------------------------------------
 poverty_index_left |   .1755687   .0302796     5.80   0.000     .1162073      .23493
poverty_index_right |   .2202764   .0315631     6.98   0.000     .1583987    .2821541
           eligible |  -11.19171   .4661828   -24.01   0.000    -12.10564   -10.27779
              _cons |   20.55449   .3376327    60.88   0.000     19.89258     21.2164
------------------------------------------------------------------------------
```

Box 2. Fuzzy Regression Discontinuity

Under fuzzy regression discontinuity, the eligibility threshold does not fully determine participation in the program. Let's assume that the running variable is X_i, and that a threshold X_0 contributes to determine eligibility. Under fuzzy discontinuity, the threshold only partially determines program participation. For example, units above the threshold are more likely to participate in the program than units with as score X_i lower than the threshold. However, there are units on both sides of the threshold that participate in the program.

The case of fuzzy discontinuity can be analyzed in an instrumental variable framework. Intuitively, this is similar to the case of randomized offering with imperfect compliance described earlier in this technical companion.

In the first stage, a dummy variable $I(Xi \geq X_0)$ is created, taking the value of 1 for a value of the running variable equal or above the threshold, and the value of 0 for a value of the running variable below the threshold. In this case, the dummy variable does not fully determine whether the unit will participate in the program or not. However, it strongly influences program participation. This dummy variable is used as an instrumental variable to predict program participation:

Stage 1: $P_i = \gamma_0 + \gamma_1 I(X_i > X_0) + \eta_i$

In a second stage, the predicted participation from the first stage is used to estimate the program impact at the threshold:

Stage 2: $Y = \beta_0 + \delta \widehat{P}_i + f(score_i) + \varepsilon_i$

In this case, the instrumental variable approach allows us to deal with the fuzziness of program participation, and estimate local average treatment effects around the threshold.

Propensity Score Matching

Let's turn to a context where we have a group of beneficiaries and nonbeneficiaries, for which we measure a set of characteristics at baseline before a program is rolled out. Suppose that we observe imbalances in characteristics X_i between the two groups at baseline. One option to address this would be to control for the imbalanced variables in a multivariate regression approach:

$$Y_i = \alpha + \delta P_i + \gamma X_i + \varepsilon_i$$

As discussed, this approach is often fraught with selection bias and can be unreliable. Indeed, the regression analysis does not necessarily represent a causal relationship between participation in the

program and the outcome variable. It would be appropriate if the observed covariate X_i were the only characteristics correlated with both program participation and outcomes. This is a very strong requirement, and we can never be sure that the observed covariates are comprehensive and there are no omitted variables leading to potential bias.

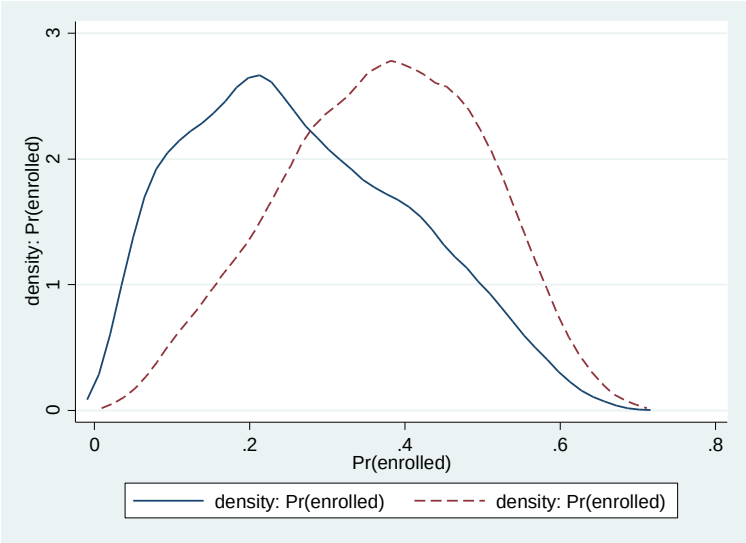
The matching approach is not very different from attempting to control for observed covariates in a regression framework. Rather than including the imbalanced characteristics in regression, units with similar characteristics are matched based on these characteristics. In the propensity score matching approach, a propensity score is obtained by a regression of program participation on a set of observed (pre-program) covariates. Then units in the treatment and comparison group with the closest propensity scores are matched, and differences in outcomes are calculated within each matched pair. The matching procedure is then repeated for all individuals in the treatment group, and averages in differences in outcomes within pairs are computed. A range of matching estimators can be used to calculate the "closest match," before calculating averages in the differences between treatment units and their matched comparison.

Stata Example 13 illustrates how to implement propensity score matching in the HISP example. First, a probit model is estimated by running a regression of program participation on a range of pre-program characteristics. The output of the probit regression shows which variables are the strongest predictors of program participation. For example, in the example below, among other variables, household size and dwellings with dirt floors are strongly associated with participation in HISP. The propensity score is obtained by computing the predicted values from this first stage. Figure 2 plots the propensity scores for the treatment and comparison group. Once the propensity scores are calculated, the next step is to find for each unit in the treatment group, a comparable "match" with a similar propensity score in the comparison group: that is, a unit with characteristics such that its likelihood of participating in the program are the same as the treatment unit for which a match is sought. As figure 2 shows, some units in the treatment group have high propensity scores; in these cases, there may not be comparison units with similar scores. In practice, matching generally occurs in the area of common support where the propensity scores for the treatment and comparison groups overlap. Once matching is performed, differences in outcomes within pairs are computed and the averages of these differences are obtained to provide estimates of treatment effects. The Stata propensity-score command (psmatch2) performs these different steps. As we can see below, the treatment effect estimated from propensity score matching is -9.97 : that is, a $9.97 reduction in household health expenditures.

```
Stata Example 13. Propensity Score Matching Estimates

* MATCHING
* In this context, you compare health expenditures at follow-up between enrolled
*  households and a set of matched nonenrolled households from both treament and
comparison villages.

*Select the relevant data
```

```
use "evaluation.dta", clear

* reshape the database
reshape wide health_expenditures age_hh age_sp educ_hh educ_sp hospital,
i(household_identifier) j(round)

probit enrolled age_hh age_sp educ_hh educ_sp female_hh indigenous hhsize dirtfloor
bathroom land hospital_distance

Iteration 0:   log likelihood =  -6047.086
Iteration 1:   log likelihood = -5510.3753
Iteration 2:   log likelihood = -5506.5201
Iteration 3:   log likelihood = -5506.5196

Probit regression                                 Number of obs     =      9,913
                                                  LR chi2(11)       =    1081.13
                                                  Prob > chi2       =     0.0000
Log likelihood = -5506.5196                       Pseudo R2         =     0.0894

------------------------------------------------------------------------------
     enrolled |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
      age_hh |  -.0131411   .0017648    -7.45   0.000    -.0166001   -.0096822
      age_sp |  -.0078841   .0020507    -3.84   0.000    -.0119035   -.0038648
     educ_hh |  -.0215019   .0062476    -3.44   0.001     -.033747   -.0092568
     educ_sp |  -.0155054   .0067557    -2.30   0.022    -.0287462   -.0022645
   female_hh |  -.0204807   .0518766    -0.39   0.693    -.1221569    .0811955
  indigenous |   .1613552    .031199     5.17   0.000     .1002062    .2225041
      hhsize |   .1188953   .0067088    17.72   0.000     .1057462    .1320443
    dirtfloor |   .3758706   .0308276    12.19   0.000     .3154496    .4362916
     bathroom |  -.1245256   .0289856    -4.30   0.000    -.1813364   -.0677149
         land |  -.0277659   .0049886    -5.57   0.000    -.0375435   -.0179884
hospital_distance |   .0015885   .0003514     4.52   0.000     .0008998    .0022772
        _cons |  -.4974732   .0904964    -5.50   0.000    -.6748429   -.3201035
------------------------------------------------------------------------------


predict pscore

kdensity pscore if enrolled ==1, gen(take1 den1)

kdensity pscore if enrolled ==0, gen(take0 den0)

graph twoway (line den0 take0, lpattern(solid)) (line den1 take1, lpattern(dash))
```

Figure 2. Common Support when Applying Matching to HISP case




```
set seed 100
generate u=runiform()
sort u

psmatch2 enrolled age_hh age_sp educ_hh educ_sp female_hh indigenous hhsize dirtfloor
bathroom land hospital_distance, out(health_expenditures1)

Probit regression                                 Number of obs     =      9,913
                                                  LR chi2(11)       =    1081.13
                                                  Prob > chi2       =     0.0000
Log likelihood = -5506.5196                       Pseudo R2         =     0.0894

------------------------------------------------------------------------------
         enrolled |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
           age_hh |  -.0131411   .0017648    -7.45   0.000    -.0166001   -.0096822
           age_sp |  -.0078841   .0020507    -3.84   0.000    -.0119035   -.0038648
          educ_hh |  -.0215019   .0062476    -3.44   0.001     -.033747   -.0092568
          educ_sp |  -.0155054   .0067557    -2.30   0.022    -.0287462   -.0022645
        female_hh |  -.0204807   .0518766    -0.39   0.693    -.1221569    .0811955
       indigenous |   .1613552    .031199     5.17   0.000     .1002062    .2225041
           hhsize |   .1188953   .0067088    17.72   0.000     .1057462    .1320443
         dirtfloor |   .3758706   .0308276    12.19   0.000     .3154496    .4362916
          bathroom |  -.1245256   .0289856    -4.30   0.000    -.1813364   -.0677149
              land |  -.0277659   .0049886    -5.57   0.000    -.0375435   -.0179884
hospital_distance |   .0015885   .0003514     4.52   0.000     .0008998    .0022772
             _cons |  -.4974732   .0904964    -5.50   0.000    -.6748429   -.3201035
------------------------------------------------------------------------------
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.
-------------------------------------------------------------------------------
       Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
----------------------------+--------------------------------------------------------
health_expendi~1  Unmatched | 7.83977335     20.70746  -12.8676866   .226604141   -56.78
                      ATT | 7.83977335   17.8088716  -9.96909828   .263484213   -37.84
----------------------------+--------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
```

```
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |     6,949 |     6,949
   Treated |     2,964 |     2,964
-----------+-----------+----------
     Total |     9,913 |     9,913
```

Independently of the matching estimator chosen, the fundamental issue with the matching approach is that it is valid only under an assumption of "uncounfoundedness", meaning if participation in the program is unconfounded conditional on the observed variables X_i on which the matching is performed. This is akin to ruling out selection bias due to unobserved or unobservable variables, and prespecifying that the condition of independence of the distribution of potential outcomes mentioned above holds. As such, the matching approach does not provide a solution to the endogeneity or selection bias issue. Rather, it assumes it away.

Generally, we advise against the use of matching estimators based on post-treatment data only. The matching approach is best used in combination with a difference-in-differences approach, when matching is performed based on pre-treatment variables, when a large set of observed characteristics is available, or when many rounds of pre-treatment data and pre-treatment trends can be matched. As such, the data requirements for matching are substantial.

Power Calculations

Power Calculations in Stata

We now provide some examples of how to undertake power calculations in Stata. We do so for the examples presented in chapter 15 of the book. The question is to determine what would be the sample size required to measure the impact of a modified version of the HISP program (called HISP+, which also covers hospitalization beyond primary care expenditures covered under HISP) on out-of-pocket health expenditures. As discussed in chapter 15, power calculations require assessing if the design creates clusters, defining the outcome indicator, as well as its mean and variance, and finally determining the minimum detectable effect. We focus on a benchmark case of an impact evaluation relying on a randomized assignment design at the unit level (that is, without clusters).

Power calculations can be implemented in Stata using the *sampsi* command. First, we must find the benchmark mean and standard deviation for the outcome indicator. We can do that with the *sum* command. We save the results in a scalar and use them to calculate a follow-up mean. We then calculate potential follow-up means with alternative minimum detectable effects of $1, $2, and $3. These can be used, along with the standard deviation saved previously, in the *sampsi* command to calculate the required sample size. The Stata output that follows provides example on how to compute for the first minimum detectable effect. (Supplementary do-file and data files can be found on the book website and provide the full code to construct the entire table 15.3 in book chapter 15).

```
Stata Example 14. Power Calculations without Clusters (out-of-pocket
expenditures)

sum health_expenditures if round==1 & treatment_locality==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
health_exp~s |      2965    7.840179    7.994495          0   87.38017

local m1 = `r(mean)'  /*This saves the mean which will be used as m1 in power
calculations below */

local sd = `r(sd)' /*This saves the standard deviation which will be used as sd1
and sd2 in power calclulations below */

local mde_1 = `m1'-1

local mde_2 = `m1'-2

local mde_3 = `m1'-3

sampsi `m1' `mde_1', p(0.8) r(1) sd1(`sd') sd2(`sd')

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
```

```
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =  7.84018
            m2 =  6.84018
           sd1 =  7.99449
           sd2 =  7.99449
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =     1004
            n2 =     1004
```

This shows that the required sample size is 1,004 for each treatment and comparison group, so 2,008 observations in total.

To obtain sample sizes for a context when the program operates through clusters (table 15.5 in handbook), we use the *sampclus* command in Stata. First, we need to find the intra-cluster correlation, using the *iclassr* command. We save the intra-cluster correlation, or rho, in the same way as we did the mean and standard deviation previously. Then we run the *sampsi* command, followed with *sampclus* to correct for a cluster design. An example for the first minimum detectable effect follows.

```
Stata Example 15. Power Calculations with Clusters (out-of-pocket expenditures)

iclassr health_expenditures  locality_identifier if round==1 & treatment_locality==1,
noisily

                       Analysis of Variance
    Source             SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      13520.0642     97   139.382105      2.27     0.0000
 Within groups      175914.979   2867   61.3585556
------------------------------------------------------------------------
    Total           189435.043   2964   63.9119579

Bartlett's test for equal variances:  chi2(96) = 487.8703  Prob>chi2 = 0.000

note: Bartlett's test performed on cells with positive variance:
      1 single-observation cells not used

Intra-locality_identifier r = 0.0403
Estimated reliability of a locality_identifier mean (n=30.26) = 0.5598

local rho = $S_1  /*This saves the intra-cluster correlation, or rho, which will be
used in clustered power calculations below*/

display `rho'
.04033407

sampsi `m1' `mde_1', p(0.8) r(1) sd1(`sd') sd2(`sd')

Estimated sample size for two-sample comparison of means
```

```
Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =  7.84018
            m2 =  6.84018
           sd1 =  7.99449
           sd2 =  7.99449
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =     1004
            n2 =     1004
```

```
sampclus, numclus(100) rho(`rho')

Sample Size Adjusted for Cluster Design

   n1 (uncorrected) = 1004
   n2 (uncorrected) = 1004

   Intraclass correlation     = .0403341

   Average obs. per cluster   = 102
   Minimum number of clusters = 100


   Estimated sample size per group:

      n1 (corrected) = 5095
      n2 (corrected) = 5095
```

This shows that we need a sample of 100 clusters, with an average 102 observations by cluster. In total, this gives 10,200 observations. Stata gives a more precise number: in this case, 5,095 observations for each treatment and comparison group, so 10,190 observations. This is a little under 10,200, so a few clusters could have 101 observations instead of 102.[15]

When an impact evaluation design includes clusters, there is a trade-off between the number of clusters and number of observations per cluster. We now illustrate that trade-off (see table 15.6 in handbook). For instance, to look into the necessary sample size with 30 clusters, we change the parameters in the *numclus* command from 100 to 30. Note, for table 15.6 in chapter 15, we used the minimum detectable effect of $2.

```
Stata Example 16.Power Calculations with Clusters (trade-off between number of
clusters and number of observations per cluster)

sampsi  `m1' `mde_2', p(0.8) r(1) sd1(`sd') sd2(`sd')
```

[15] In table 15.5 in the book, we provided the rounded number to avoid confusion.

```
Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =  7.84018
            m2 =  5.84018
           sd1 =  7.99449
           sd2 =  7.99449
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =      251
            n2 =      251

sampclus, numclus(30) rho(`rho')

Sample Size Adjusted for Cluster Design

   n1 (uncorrected) = 251
   n2 (uncorrected) = 251

   Intraclass correlation     = .0403341

   Average obs. per cluster   = 50
   Minimum number of clusters = 30


   Estimated sample size per group:

      n1 (corrected) = 748
      n2 (corrected) = 748
```

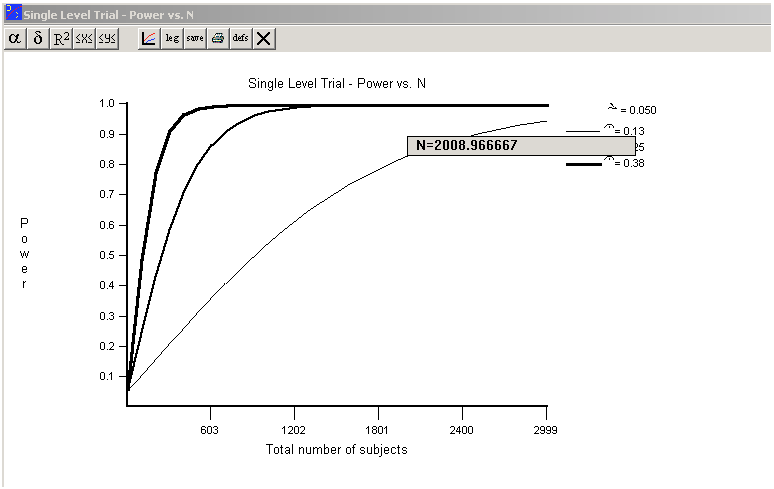
This shows that the required sample size is 30 clusters with 50 observations by cluster, so approximately 1,500 observations in total.

Power Calculations in Optimal Design

Optimal Design is a program that assists researchers with power calculations by graphing the relationships between power, sample, and effect size (Spybrook et al., 2009). To produce the same results as in Stata Example 14 (or table 15.3 in the book), you can create a figure in Optimal Design looking at a single level trial with power versus total sample size. We plot three given effect sizes, which must be in standardized effect form.[16] A minimum detectable effect of $1 is equivalent to a standardized effect of 0.13 standard deviation of the outcome of interest. A minimum detectable effect of $2 is equivalent to a standardized effect of 0.25 standard deviation of the outcome of interest. A minimum detectable effect of $3 is equivalent to a standardized effect of 0.38 standard deviation of the outcome of interest.

Note that *N* on the x-axis is the total sample with two treatment arms, so you must divide by two to get the size by arm. The plot shows that, for instance, for a standardized effect of 0.13 and a power of 0.9, the total number of subjects required is approximately 2,688 (same as table 15.2 in the handbook). For a power of 0.8, approximately 2,008 observations would be needed (same as table 15.3 in the book, or Stata Example 14 above). You can see the number of observations needed by clicking on the line in optimal design.

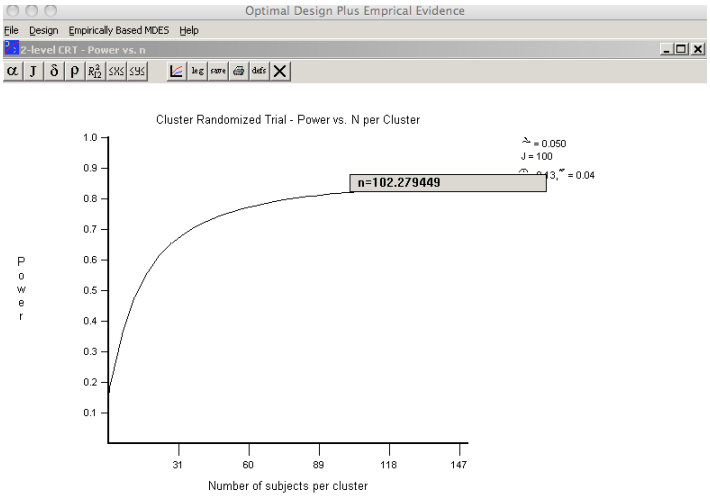
```
Optimal Design Example 1. Power Calculations without Clusters (out-of-pocket
expenditures)
```



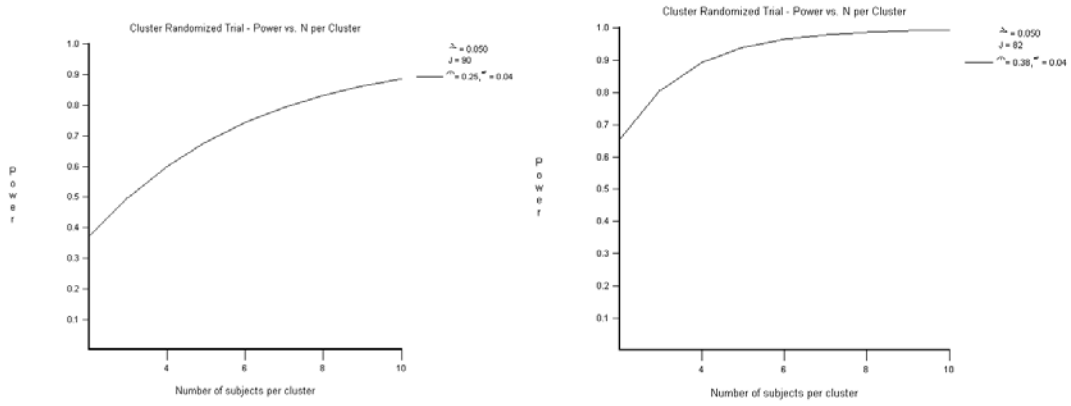
We can also plot cluster randomized trials in *Optimal Design*. Here, we plot power versus total number of clusters for a given number of units per cluster. We can see that for *n*=102, the ideal number of clusters for 0.8 power is 100 (this is the result for the first row in table 15.5 of the book).

[16] The standardized effect size, commonly denoted as δ, is calculated by dividing the effect size by the standard deviation.

Optimal Design Example 3. Power Calculations with Clusters (out-of-pocket expenditures)



Alternatively, we can plot the ideal number of units per cluster, or we can plot the power versus cluster size for a given number of clusters. We can see from the figure below that for 100 clusters, a cluster size of 102 is needed for a power level of 0.8. (Stata Example 16 above; table 15.5 in the book).



References

Angrist, J. D. and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bruhn, M., and D. McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–32.

Gertler, Paul J.; Martinez, Sebastian; Premand, Patrick; Rawlings, Laura B.; Vermeersch, Christel M. J.. 2016. *Impact Evaluation in Practice, Second Edition.* Washington, DC: Inter-American Development Bank and World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/25030 License: CC BY 3.0 IGO.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945–70.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Experiments.", *Journal of Educational Psychology* 66: 688–701.

Spybrook, Jessaca, Stephen Raudenbush, Xiaofeng Liu, Richard Congdon, and Andrés Martinez. 2008. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software.* New York: William T. Grant Foundation.

University of California, Berkeley, 2016. Applied Impact Evaluation Methods Course Curriculum (http://aie.cega.org).