Evaluating Program Impact using Panel or Longitudinal Data Models

Gustavo Angeles
David Guilkey

Carolina Population Center,
University of North Carolina at Chapel Hill

March 2016

I. The basic setup

Our basic empirical model:

$$Y_{ij} = X_{ij}\beta + W_j\delta + \alpha Prog_{ij} + (\mu_{ij} + \mu_j + \theta_{ij}) \qquad (1)$$

Where,

Y_{ij} = Outcome of interest (Use of FP)

X_{ij} = Individual-level observed characteristics (age, education, SES, gender)

W_j = Community-level observed characteristics (rural)

$Prog_{ij}$ = Program variable, at individual level

μ_{ij} = Unobserved individual-level characteristics (preferences, risk aversion, genetic background)

μ_j = Unobserved community-level characteristics (sanitation, schools)

θ_{ij} = Unobserved random shocks to Y, at the individual level

i= 1, 2, …, N_j individuals in community j

j= 1, 2, …, J communities

Estimation procedure:

It depends on the assumptions about the unobservables μ_{ij} and μ_j *.

1) If the explanatory variables X_{ij} , W_j and $Prog_{ij}$ are not correlated with μ_{ij} and μ_j , OLS will provide an unbiased and consistent estimator. You will need to correct the standard errors because the error terms will be correlated (use option `vce(cluster clustvar)` in Stata.)

2) If you suspect any of the explanatory variables X_{ij} , W_j or $Prog_{ij}$ are correlated with either μ_{ij} or μ_j , you should not use OLS, because it will generate biased and inconsistent estimates. This is the case of endogeneous $Prog_{ij}$.

*: Notice that θ_{ij} represents random shocks specific to individual i which influence the outcome of interest. So, they are assumed uncorrelated with the explanatory variables (X_{ij} , W_j and $Prog_{ij}$)

What can we do?

Remember that the source of endogeneity are unobserved factors (either μ_{ij} or μ_j) that influence both Prog (or any other explanatory variable) and Y (the outcome).

Question: What if we could observe the same individual at two points in time, say, two years apart, at time 0 and time 1?

First, rewrite model (1) to allow for variables that change over time and for variables that remain fixed between time 0 and time 1.

Then,

$$Y_{ijt} = X_{ijt}\beta_1 + X_{ij}^F \beta_2 + W_{jt}\delta_1 + W_j^F \delta_2 + \alpha Prog_{ijt} + (\mu_{ij} + \mu_j + \theta_{ijt}) \quad (2)$$

Where,

Y_{ijt} = Outcome of interest, time-varying (Use of FP)
X_{ijt} = Individual-level observed characteristics, time-varying (age, education, SES)
X_{ij}^F = Individual-level observed characteristics, time-invariant or fixed (sex)
W_{jt} = Community-level observed characteristics, time-varying
W_j^F = Community-level observed characteristics, time-invariant or fixed(rural)
$Prog_{ijt}$ = Program variable, at individual level, time-varying
μ_{ij} = Unobserved individual-level fixed characteristics (preferences, risk aversion, genetic background)
μ_j = Unobserved community-level fixed characteristics (sanitation, schools)
θ_{ijt} = Unobserved random shocks to Y, at the individual level, time-varying.

Key assumption: Unobserved factors μ_{ij} and μ_i do not change between time 0 and time 1. They remain "fixed" in the observation time interval.
All time-varying unobservables are summarized by θ_{ijt} which is uncorrelated with the observed explanatory variables.

Second, let's write the model for each point in time:

At t=0, $Y_{ij0} = X_{ij0}\beta_1 + X_{ij}^F\beta_2 + W_{j0}\delta_1 + W_j^F\delta_2 + \alpha Prog_{ij0} + (\mu_{ij} + \mu_j + \theta_{ij0})$

At t=1, $Y_{ij1} = X_{ij1}\beta_1 + X_{ij}^F\beta_2 + W_{j1}\delta_1 + W_j^F\delta_2 + \alpha Prog_{ij1} + (\mu_{ij} + \mu_j + \theta_{ij1})$

Then, take the difference between time 0 and time 1:

$$Y_{ij1} - Y_{ij0} = (X_{ij1} - X_{ij0})\beta_1 + (W_{j1} - W_{j0})\delta_1 + \alpha(Prog_{ij1} - Prog_{ij0}) + (\theta_{ij1} - \theta_{ij0})$$

All "fixed" or time-invariant variables, including μ_{ij} and μ_j, difference out. The sources of the endogeneity in model (1) have been eliminated in the new specification!
You can apply OLS to new model. It will generate unbiased and consistent estimates of α, the program impact.

That is the **"First-Differences Model."**

Advantages:
1. Easy to implement.
2. It provides unbiased and consistent estimates of α even if X_{ij} , W_j or $Prog_{ij}$ are correlated with either μ_{ij} or μ_j in model (1). Individuals act as their own controls.

Disadvantages:
1. It does not generate estimates for fixed variables (education, sex, rural or even the program variable if it is time invariant.)
2. Large reduction in effective sample size makes it more difficult to obtain significant results when they really exist.
3. The new variables expressed in differences, [(X_{ij1} - X_{ij0}), (W_{j1} - W_{j0}), ($Prog_{ij1}$ - $Prog_{ij0}$)], have less variability than the original variables. Standard errors will be larger and confidence intervals wider. It will be harder to find significant program effects if they exist. So, the model is not efficient.
4. It does not control well for the "time" trend.

Important: We solved the problem of endogeneity without having to find IV variables.

You will need to control for clustering of error terms to obtain correct standard errors (option `vce(cluster clustvar)` in Stata).

Extension for multiple time observations T > 2:

Suppose you have T observations for each individual in the sample.

Take the average of all time observations for each individual and define the "between-quation":

$$\bar{Y}_{ij} = \bar{X}_{ij}\beta_1 + X_{ij}^F\beta_2 + \bar{W}_j\delta_1 + W_j^F\delta_2 + \alpha\overline{Prog}_{ij} + (\mu_{ij} + \mu_j + \bar{\theta}_{ij}) \qquad (3)$$

Where,
$\bar{Y}_{ij}$: Average of all time observations of time-varying individual-level outcome, for each individual
$\bar{X}_{ij}$: Average of all time observations of time-varying individual-level characteristics, for each individual
X_{ij}^F : Individual-level observed characteristics, time invariant or Fixed
$\bar{W}_j$: Average of all time observations of time-varying community-level characteristics, for each community
W_j^F : Community -level observed characteristics, time invariant or Fixed

We still assume that μ_{ij} and μ_j do not change over time.

Then, substract this equation from equation (2) for each time you observe:

$$Y_{ijt} - \bar{Y}_{ij} = (X_{ijt} - \bar{X}_{ij})\beta_1 + (W_{jt} - \bar{W}_j)\delta_1 + \alpha(Prog_{ijt} - \overline{Prog_{ij}}) + (\theta_{ijt} - \bar{\theta}_{ij})$$

All "Fixed" variables, including μ_{ij} and μ_j, were differenced out. OLS will generate unbiased and consistent estimates of α, the program impact.

Same advantages and disadvantages as in T=2 case. You need a "large" T.

Dummy variables model

Model:

$$Y_{ijt} = X_{ijt}\beta + W_{jt}\delta + \alpha Prog_{ijt} + (\mu_{ij} + \theta_{ijt}) \qquad (4)$$

μ_{ij} represents omitted individual-level unobserved characteristics. These are fixed over time.

You have several time observations for each individual.

Estimation
- Create a dummy variable for each individual
- Add all dummies minus one to the model
- Run the model with standard methods

Advantages
- Easy to estimate (if number of time observations is large)
- Provides control for bias due to omitted variables.

Disadvantage
- Loss of "degrees of freedom".
- Inefficient because you have to estimate a large number of parameters.

Difference-in-Differences Model

Basic setup:

- Two areas: Program area and Non-program area
- Two surveys: Baseline and Follow-up

Let's define:

P_{ij} = 1, individual i is in Program Area
0, individual i is in Non-program Area

T_{ijt} = 1, if observation at Follow-up (t=1)
0, if observation at Baseline (t=0)

The model: $Y_{ijt} = \beta_0 + \beta_1 P_{ij} + \beta_2 T_{ijt} + \beta_3 P_{ij} T_{ijt} + \mu_{ij} + \mu_j + \theta_{ijt}$ (5)

Let's examine different cases:

I. In Program Areas (P_{ij}=1)

- At Baseline (T_{ij0}=0): $Y_{ij0} = \beta_0 + \beta_1 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij0}$ (A)
- At Follow-up (T_{ij1}=1): $Y_{ij1} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \mu_{ij} + \mu_j + \theta_{ij1}$ (B)

Taking the difference: $(Y_{ij1} - Y_{ij0}) = \beta_2 + \beta_3 + (\theta_{ij1} - \theta_{ij0})$ (B-A)

Expected value: $E(Y_{ij1} - Y_{ij0} \mid P_{ij}=1) = \beta_2 + \beta_3$

assuming that $E(\theta_{ij1} - \theta_{ij0} \mid P_{ij}=1) = 0$

II. In Non-Program Areas (P_{ij}=0)

- At Baseline (T_{ij0}=0): $Y_{ij0} = \beta_0 + 0 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij0}$ (C)
- At Follow-up (T_{ij1}=1): $Y_{ij1} = \beta_0 + 0 + \beta_2 + 0 + \mu_{ij} + \mu_j + \theta_{ij1}$ (D)

Taking the difference: $(Y_{ij1} - Y_{ij0}) = \beta_2 + (\theta_{ij1} - \theta_{ij0})$ (D-C)

Expected value: $E(Y_{ij1} - Y_{ij0} \mid P_{ij}=0) = \beta_2$

assuming that $E(\theta_{ij1} - \theta_{ij0} \mid P_{ij}=0) = 0$

Now, the difference of the differences:

$E(Y_{ij1} - Y_{ij0} \mid P_{ij}=1) - E(Y_{ij1} - Y_{ij0} \mid P_{ij}=0) = (\beta_2 + \beta_3) - \beta_2 = \beta_3$

Notice that μ_{ij} and μ_j were differenced out.

Key Question: Is β_3 our program impact?

Answer: Yes, if the "Parallel assumption" holds.

- The “Parallel assumption” holds if the time-varying unobservables do not vary over time with program status.

  That is, $E(\theta_{ij1} - \theta_{ij0}|P_{ij}=1) = 0$ and $E(\theta_{ij1} - \theta_{ij0}|P_{ij}=0) = 0$

  Which means that assignment of the program does not influence the change in unobservables that vary over time.
  In other words, time-varying unobservables <u>are not</u> a potential source of endogeneity.

Also, the Program Area would have had the same change as the Non-program Area in the absence of the program. The time trend, β_2 , is the same for both groups.

Under the “Parallel Assumption”, β_3 is an estimate of Program Impact.

In some cases this could be a strong assumption. The time-varying unobservables could change depending on treatment status. The “time trend” could be different for unobserved reasons related to the program.

Graphically (see ppt file material)

***** <u>When is the Parallel Assumption valid?</u>**

- If the Non-program areas (comparison group) were selected to be “as similar as” possible as the Program Areas.
  <u>Best case</u>: Program was randomly assigned to the areas (experimental evaluation design).
  If you cannot implement an experiment, it is recommended to select Non-program areas by some matching procedure, using community-level aggregates collected before the baseline. You will have to define the extent of the “pool of areas” from which to obtain the matched Non-program areas to include in the baseline.
- There is a “relatively short” time interval between baseline and follow-up. But, the program might need more time to have an impact.

- Fixed factors μ_{ij} and μ_j are the only, or the main, source of potential endogeneity of the program variable. Even if there is correlation between time-varying unobservables and program, its effect is small.

Notice that you can easily add covariates (X_{ijt}) to the model to control for observed differences between the program and non-program areas. It is recommended that the covariates are measured at baseline. You need to make sure the covariates are not endogeneous.

Testing the Parallel Assumption

You need pre-baseline survey data.

Setup:
- Two areas: Program area and Non-program area
- Two surveys: Pre-baseline and Baseline

Define:
P_{ij} = 1, individual i in Program Area
0, individual i in Non-program Area

T_{ijt} = 1, if observation at Baseline (t=0)
0, if observation at Pre-Baseline (t=-1)

The basic model:

$$Y_{ijt} = \beta_0 + \beta_1 P_{ij} + \beta_2 T_{ijt} + \beta_3 P_{ij} T_{ijt} + \mu_{ij} + \mu_j + \theta_{ijt} \quad (6)$$

Let's examine the different cases:

I. In Program Areas ($P_{ij} = 1$)

- At Pre-baseline ($T_{ij-1}=0$): $Y_{ij-1} = \beta_0 + \beta_1 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij-1}$ (E)
- At Baseline ($T_{ij0}=1$): $Y_{ij0} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \mu_{ij} + \mu_j + \theta_{ij0}$ (A)

Taking the difference: $(Y_{ij0} - Y_{ij-1}) = \beta_2 + \beta_3 + (\theta_{ij0} - \theta_{ij-1})$ (A-E)

Expected value: $E(Y_{ij0} - Y_{ij-1} | P_{ij}=1) = \beta_2 + \beta_3$

assuming: $E(\theta_{ij0} - \theta_{ij-1} | P_{ij}=1) = 0$

II. In Non-Program Areas ($P_{ij} = 0$)

- At Pre-baseline ($T_{ij-1}=0$): $Y_{ij-1} = \beta_0 + 0 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij-1}$ (F)
- At Baseline ($T_{ij0}=1$): $Y_{ij0} = \beta_0 + 0 + \beta_2 + 0 + \mu_{ij} + \mu_j + \theta_{ij0}$ (C)

Taking the difference: $(Y_{ij0} - Y_{ij-1}) = \beta_2 + (\theta_{ij0} - \theta_{ij-1})$ (C-F)

Expected value: $E(Y_{ij0} - Y_{ij-1} | P_{ij} = 0) = \beta_2$

assuming: $E(\theta_{ij0} - \theta_{ij-1} | P_{ij} = 0) = 0$

Then, the difference of the differences is:

$$E(Y_{ij0} - Y_{ij-1}|P_{ij}=1) - E(Y_{ij0}-Y_{ij-1}|P_{ij}=0) =$$
$$(\beta_2 + \beta_3) \quad - \quad \beta_2 \quad = \beta_3$$

The test for the Parallel Assumption is: Ho: $\beta_3 = 0$.

Graphically (see ppt material)

If $\beta_3 \neq 0$, there is no evidence that the parallel assumption should hold in the following time period (baseline to follow-up).

Other ways to examine the Parallel Assumption
Duflo, E. (2002) "Empirical methods" proposes two other ways:

+ Use other control groups
+ Use a different outcome Y which is not supposed to be affected by the program.

Dif-in-Dif with three observations

Let's define:

P_{ij} = 1, individual i in Program Area
0, individual i in Non-program Area

$T1_{ijt}$ = 1, if observation at Follow-up 1 (t=1)
0, if observation at other time (baseline or follow-up 2)

$T2_{ijt}$ = 1, if observation at Follow-up 2 (t=2)
0, if observation at other time (baseline or follow-up 1)

The basic model:

$$Y_{ijt} = \beta_0 + \beta_1 P_{ij} + \beta_2 T1_{ijt} + \beta_3 T2_{ijt} + \beta_4 P_{ij} T1_{ijt} + \beta_5 P_{ij} T2_{ijt} + \mu_{ij} + \mu_j + \theta_{ijt}$$

I. In Program Areas ($P_{ij}=1$)

- At baseline (T1=0 and T2=0): $Y_{ij0} = \beta_0+\beta_1+0 + 0 + 0+0+ \mu_{ij} + \mu_j + \theta_{ij0}$ (A)
- At follow-up 1 (T1=1 and T2=0): $Y_{ij1} = \beta_0+\beta_1+\beta_2 + 0 + \beta_4 +0+ \mu_{ij} + \mu_j + \theta_{ij1}$ (B)
- At follow-up 2 (T1=0 and T2=1): $Y_{ij2} = \beta_0+\beta_1+0 + \beta_3 + 0 + \beta_5+ \mu_{ij} + \mu_j + \theta_{ij2}$ (G)

Taking the difference:

Fu1-Base:	$E(Y_{ij1} - Y_{ij0} \mid P_{ij}=1) = \beta_2 + \beta_4$	(B-A)
Fu2-Base:	$E(Y_{ij2} - Y_{ij0} \mid P_{ij}=1) = \beta_3 + \beta_5$	(G-A)
Fu2-Fu1:	$E(Y_{ij2} - Y_{ij1} \mid P_{ij}=1) = (\beta_3 + \beta_5) - (\beta_2 + \beta_4)$	(G-B)

assuming:
$E(\theta_{ij1} - \theta_{ij0} \mid P_{ij}=1) = 0$
$E(\theta_{ij2} - \theta_{ij0} \mid P_{ij}=1) = 0$
$E(\theta_{ij2} - \theta_{ij1} \mid P_{ij}=1) = 0$

II. <u>In Non-Program Areas ($P_{ij}=0$)</u>

- At baseline (T1=0 and T2=0): $Y_{ij0} = \beta_0 + 0 + 0 + 0 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij0}$ (C)
- At follow-up 1 (T1=1 and T2=0): $Y_{ij1} = \beta_0 + 0 + \beta_2 + 0 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij1}$ (D)
- At follow-up 2 (T1=0 and T2=1): $Y_{ij2} = \beta_0 + 0 + 0 + \beta_3 + 0 + 0 + \mu_{ij} + \mu_j + \theta_{ij2}$ (H)

Taking the difference:

Fu1-Base:	$E(Y_{ij1} - Y_{ij0} \mid P_{ij}=0) = \beta_2$	(D-C)
Fu2-Base:	$E(Y_{ij2} - Y_{ij0} \mid P_{ij}=0) = \beta_3$	(H-C)
Fu2-Fu1:	$E(Y_{ij2} - Y_{ij1} \mid P_{ij}=0) = \beta_3 - \beta_2$	(H-D)

assuming:
$E(\theta_{ij1} - \theta_{ij0} \mid P_{ij}=0) = 0$
$E(\theta_{ij2} - \theta_{ij0} \mid P_{ij}=0) = 0$
$E(\theta_{ij2} - \theta_{ij1} \mid P_{ij}=0) = 0$

Then, the estimated program impacts are:

Impact 1 (Base to Fu1) = $(\beta_2 + \beta_4) - \beta_2 = \beta_4$
Impact 2 (Base to Fu2) = $(\beta_3 + \beta_5) - \beta_3 = \beta_5$
Impact 3 (Fu1 to Fu2) = $(\beta_3 + \beta_5) - (\beta_2 + \beta_4) - (\beta_3 - \beta_2) = (\beta_5 - \beta_4)$

Notice that μ_{ij} and μ_j were differenced out, again.

The validity of this model rests in the assumption that time-varying unobservables do not vary over time in a systematic way related to program status. That is, time-varying unobservables are not a source of endogeneity of the program variable.

Graphically (see ppt material).

<u>What if you only have a panel of communities, not individuals?</u>
You can only control for community-level fixed unobservables. You cannot control for individual-level unobservables. You will have to assess whether that is enough to control for the potential sources of endogeneity in your model.

Random Effects Model (RE)

Model: $Y_{ti} = X_{ti}\beta + \alpha P_{ti} + Z_i\delta + \mu_i + \varepsilon_{ti}$

Error terms assumptions:

1. ε_{ti} and μ_i are random variables normally distributed with zero means and variances σ^2_{ε} and σ^2_{μ} respectively. They are not correlated with each other.
2. The error terms are not correlated with explanatory variables.

Note: FE assumes μ_i are fixed coefficients that affect intercept
RE assumes μ_i are random and not correlated with regressors

- RE resolves potential serial correlation and heteroscedasticity problems
- RE does not solve endogeneity problem

Given the assumptions, an additional parameter of interest:

$$\rho = \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + \sigma^2_{\varepsilon}}$$

It is a measure of the degree of correlation across time.
- If $\rho = 0$, all observations are independent ($\sigma_{\mu} = 0$)
- If $\rho = 1$, means perfect correlation for time observations so there are only N observations.

Important: Similar to the "design effect" of multiple stage survey sample selection

Given the assumptions the Generalized Least Squares (GLS) can be used to obtain optimal estimates (STATA `xtreg`).

However, OLS with corrected standard errors is an option if one is not interested in obtaining an estimate of ρ (STATA).

Choosing between FE and RE

1. If there is no endogeneity, the estimated parameters obtained with FE should be identical to those obtained with RE. But, the RE model has the advantage of being more efficient;

2.If there is endogeneity, then the estimated parameters obtained with FE should be different to those obtained with RE, and FE is preferred as it generates unbiased and consistent estimates.

Haussman test:

It is based on the difference between $\hat{\beta}_{FE}$ y $\hat{\beta}_{RE}$

- If $\hat{\beta}_{FE} = \hat{\beta}_{RE}$, there is no endogeneity, use RE

- If $\hat{\beta}_{FE} \neq \hat{\beta}_{RE}$, there is endogeneity, use FE