

Identification Strategies in Applied Social Research

Regression Approach and Differences-in-Differences Models

Marco Francesconi
Essex

Spring 2012 — *EC944, Labour Economics – Lectures 1-2*

Introduction (1)

One of the most challenging empirical questions in the social sciences (and in empirical labour economics in particular) involves “what if” statements about *counterfactual* outcomes.

Examples:

- We would like to know if a specific *policy to reduce poverty* will improve the incomes and work opportunities of poor families and their children
- Young people would like to know what the *labour market consequences of getting a university degree* are for them

Introduction (2)

Examples (cont.):

- Workers would like to know whether they have (or can) gain *wage premia if they join a trade union*
- Women may want to know whether and how their labour market opportunities change *if they have a child*

All these cases require us to see what outcomes would have been observed if the poverty reduction policy was **not** implemented, or if the people who graduated **did not** go to university, or if the workers who joined a union **were not** union members, or if women **did not** have a child.

Differences, say, between graduates and non-graduates or union members and non-union members (in their outcomes) define the causal effects of interest, but it is not clear how to define a counterfactual world.

Introduction (3)

Two key difficulties are:

1. **definition** itself of counterfactual states (for example, in the case of union effects: should they be defined relative to a world where unionisation rates are equal to what they are now; or relative to a world where everyone is unionised; or relative to a world where nobody is unionised?)
2. **differences** in counterfactual outcomes are **not informative** when we observe **only one scenario for each individual** (i.e. either unionised or not; either graduate or not) [This is almost invariably the case when we consider real-world situations]

Introduction (4)

Solutions:

- in many disciplines (e.g., medical sciences) **randomised** trials, because randomization ensures that outcomes in the control group do capture the counterfactual for a treatment group
- in other disciplines where experiments in large scale cannot be done (e.g., social sciences) **observational studies**
- in economics, growing number of experimental research (e.g., field experiments, natural experiments, and lab experiments)

In all cases, but particularly with observational studies, it is essential to specify an identification strategy . This can be defined as:

“A combination of a clearly defined *source of (identifying) variation* in a *causal variable* and the use of a particular *econometric technique*.”

Introduction (5)

Social scientists have generally used either one of the following 6 different approaches:

- (a) **Control for confounding variables**
- (b) **Differences-in-differences**
- (c) **Fixed effects models (sibling models)**
- (d) **Instrumental variables**
- (e) **Regression discontinuity**
- (f) **Structural estimation of dynamic relationships**

We will be looking at each of these approaches with applications in the labour market and labour market policies.

Control for confounding variables (1)

(Source: Angrist and Krueger (1999) *Handbook of Labor Economics*)

Suppose we are interested in the following question: “Is the positive association between schooling and earnings a **causal** relationship?”

The basic idea of this approach is to use regression methods to control for variables that are confounded (i.e. related to) schooling. In particular, assume that

$$Y_i = \mathbf{X}'_i \beta + \rho S_i + e_i \quad (1)$$

where Y_i is the log wage (or earnings) of person i ; \mathbf{X}_i is a $(K \times 1)$ vector of control variables (which might include measures of ability and family background); S_i denotes years of education; and e_i is a regression error such that $E(e_i | X_i) = 0$ and $E(e_i | S_i) = 0$.

Control for confounding variables (2)

Given our question above, the key parameter of interest in equation (1) is ρ , the rate of return to schooling.

What conditions must be met for ρ to have a causal interpretation?

Causality can be based on an underlying relationship that describes what individual i would earn if he/she obtained different levels of education, i.e.:

$$Y_{s,i} [\equiv Y_i(S)] = f_i(S) \quad (2)$$

Notice that the function $f(\cdot)$ in (2) has a subscript i , while S does not. So, S is a variable, but *not* a random variable. The function $f_i(S)$ tells us what individual i would earn **for any value** of schooling S , and not just for the realised value S_i . That is, $f_i(S)$ answers “what if” questions.

Control for confounding variables (3)

Now, link the causal relationship (2) to the observed association between schooling and earnings:

$$f_i(S) = \beta_0 + \rho S + \eta_i. \quad (3)$$

Equation (3) is linear and is the same for all individuals (S does not have a subscript; and also ρ does not have a subscript). The only individual-specific random part of $f_i(S)$ is η_i , which captures unobserved factors that determine earnings or earning potentials.

Substitute the observed S_i for S in (3) and plug it into (2) to obtain:

$$Y_i(S) = \beta_0 + \rho S_i + \eta_i \quad (4)$$

Control for confounding variables (4)

Equation (4) looks like (1), without covariates \mathbf{X}_i , except that because of (3) the coefficients in (4) can be given a causal interpretation. The OLS estimate of ρ in (4) is:

$$\rho_{OLS} = \frac{\text{Cov}(Y_i, S_i)}{\text{Var}(S_i)} = \rho + \frac{\text{Cov}(S_i, \eta_i)}{\text{Var}(S_i)}. \quad (5)$$

The term $\text{Cov}(S_i, \eta_i)/\text{Var}(S_i)$ in (5) is the coefficient from a regression of η_i on S_i , and reflects any correlation between the realised S_i and unobserved individual earnings potential (η_i in this case). Therefore, **if** educational attainment were **randomly assigned**, as in a controlled experiment, then $\text{Cov}(S_i, \eta_i) = 0$.

Control for confounding variables (5)

In practice, however, schooling is a consequence of individual and family decisions as well as institutional and other forces (e.g., schools and teachers, peer groups, and government) that can generate correlation between η_i and schooling.

Therefore, the OLS estimate of ρ is likely to be **inconsistent**.

The (control for confounding variables) regression strategy tries to overcome this problem by decomposing the random part of individual earnings potential, η_i , into a linear function of the observable characteristics \mathbf{X}_i :

$$\eta_i = \mathbf{X}_i' \beta + \epsilon_i, \quad (6)$$

where ϵ_i is the error term uncorrelated with \mathbf{X}_i' by construction, and β is a vector of population regression coefficients.

Control for confounding variables (6)

Plug (6) into (4) to get:

$$Y_i = \beta_0 + \mathbf{X}'_i\beta + \rho S_i + \epsilon_i. \quad (7)$$

This is the equation we would ultimately estimate using “real-world” data, either cross-sectional or panel data (such as the Labour Force Survey in the UK and Europe, the Current Population Survey in the US, or British Household Panel Survey, National Longitudinal Survey of Youth, etc.).

Control for confounding variables (7)

The key identifying assumption is that the observable characteristics \mathbf{X}_i are the **only** reason why η_i and S_i (or, equivalently, $f_i(S)$ and S_i) are correlated. In other words:

$$E(S_i \epsilon_i) = 0. \tag{8}$$

Expression (8) is known as the **“selection on observables” assumption**, where the regressor of interest (S) is assumed to be determined independently of potential outcomes (ϵ) after accounting for a set of observable characteristics (\mathbf{X}_i).

Control for confounding variables (8)

With (7) and under assumption (8) we have:

$$\rho_{OLS} = \frac{\text{Cov}(Y_i, S_i)}{\text{Var}(S_i)} = \rho + \underbrace{\Gamma'_{SX}\beta}_{\text{"omitted variables bias"}}, \quad (9)$$

where Γ_{SX} is a $(K \times 1)$ vector of coefficients from a regression of each element of X_i on S_i . Equation (9) is the **“omitted variables bias” formula**, which relates a bivariate regression coefficient to the coefficient on S_i in a regression that includes additional covariates.

Control for confounding variables (9)

Using (9) we can say that if the omitted variables are positively correlated with earnings ($\beta > 0$) and positively correlated with schooling ($\Gamma_{SX} > 0$), then

$$\frac{\text{Cov}(Y_i, S_i)}{\text{Var}(S_i)} > \rho,$$

or the OLS estimate of ρ is larger than the true causal effect of schooling on earnings.

In sum, the OLS estimate of ρ in equation (1) provides a consistent estimate of causal parameter of interest provided that (8) holds.

Control for confounding variables (10)

Note that (8) is an assumption about ϵ_i and S_i , whereas $E(X_i\epsilon_i) = 0$ is a statement about covariates and is true by definition/construction. In all regression analyses, it is important to distinguish between:

- error terms that represent the *random part* of models of potential outcomes, from
- *mechanical decompositions* (where the relationship between errors and regressors has no behavioural content).

One key question we ought to ask is:

Is the selection-on-observables assumption (condition (8)) plausible?

This assumption makes total sense if there is random assignment *conditional* on X . But even without random assignment, it may be plausible if we know a lot about the process generating the regressor of interest. For example, colleges/universities may screen students using certain characteristics (e.g., test scores, exam grades, and family income). Conditional on these characteristics, all applicants may be acceptable and will be chosen on a first-come/first-served basis.

Regression method problems

1. *Choice of covariates is crucial*

- Schooling is *not* random. Obvious covariates to include in our regressions would be variables that are correlated with both schooling and earnings. One of such variables, which has been widely used, is *test scores*. But test scores may not be perfect controls for all the differences in earnings potential between more and less educated individuals.
- Plus, we need to see whether the estimates are sensitive to the inclusion of *additional control variables*. If results are sensitive to changes in the set of control variables, then there are reasons to wonder whether there might be *unobserved* covariates that would change the estimates even further.

2. *Measurement error*

- This biases the OLS estimates towards zero (if variables are measured with additive linear error that is uncorrelated with correctly-measured values).
- The inclusion of control variables that are correlated with actual values of the variable on interest and uncorrelated with measurement error tends to aggravate this attenuation bias.

3. *Variables used to control for ability may be endogenous*

Consider (4) and assume that $Cov(S_i, \eta_i) = 0$. Thus, a simple bivariate regression would generate a consistent estimate of ρ .

Suppose now schooling affects test scores, A_i , in the following way:

Control for confounding variables (14)

$$A_i = \gamma_0 + \gamma_1 S_i + \eta_{1i}. \quad (10)$$

Assume $\gamma_1 > 0$ and $\text{Cov}(S_i, \eta_{1i}) = 0$. Therefore, the OLS estimates of (10) would be **consistent** for γ_1 .

What happens if we include A_i in equation (7)?

The endogeneity of A_i means that $\text{Cov}(\eta_i, \eta_{1i}) \neq 0$ and, possibly, this is positive. Thus:

Control for confounding variables (15)

$$\rho_{OLS|A} = \frac{\text{Cov}(Y_i, S_{.A_i})}{\text{Var}(S_{.A_i})} = \rho + \gamma_1 \varphi_{01},$$

where $S_{.A_i}$ denotes the residuals from a regression of S_i on A_i , and φ_{01} is the coefficient from a regression of η_i on η_{1i} .

If $\gamma_1 > 0$ and $\varphi_{01} > 0$, then $\rho_{OLS|A}$ is *less* than the true value of ρ . This is not because of omitted-variable bias, but because of the bias induced by **conditioning on an outcome variable** (A_i).

Differences-in-Differences (DD) Models (1)

The DD approach is a simple method applied to sets of group means in cases when certain groups of agents are **exposed to the causing variable** of interest and other groups are not. It can be applied to *panel data* as well as to (repeated) *cross sectional* data.

Example: “Mariel Boatlift” effect of migration on the employment of natives (*Source: Card (1990) Industrial and Labor Relations Review*).

- Between May and September 1980 the Miami labour force increased by about 7% as a consequence of the ‘Mariel’ immigration.
- Question: “Was this immigration **harmful** to the employment prospects of residents?”

Differences-in-Differences (DD) Models (2)

- We need an identification strategy. The DD approach will require us to select comparison cities that can be used to estimate what would have happened in the Miami labour market in the absence of the Mariel immigration (we need to construct “counterfactuals”).
- Card (1990) chose Atlanta, Los Angeles, Houston and Tampa-St Petersburg, because, **like Miami**, they had large Black and Hispanic populations.

Differences-in-Differences (DD) Models (3)

Rationale for DD method

Let:

- Y_{0i} = unemployment status of individual i in absence of migration in the city where i lives
- Y_{1i} = unemployment status of individual i if Mariel immigrants come to the city where i lives

We define the **average unemployment rate** in city c in year t as:

- $E(Y_{0i}|c, t)$ if no immigration occurs
- $E(Y_{1i}|c, t)$ if immigration occurs

Differences-in-Differences (DD) Models (4)

We know that:

- Mariel immigration happened in Miami ($c = \text{Miami}$) in 1980 ($t = 1980$). So, $E(Y_{1i}|c, t)$ can only be computed for Miami, after 1980
- The Mariel Boatlift study uses comparison cities to measure the counterfactual average $E(Y_{0i}|c = \text{Miami}, t > 1980)$, i.e. what the unemployment rate in Miami would have been if the Mariel immigrants had not come

The DD method identifies causal effects by restricting the conditional mean function $E(Y_{0i}|c, t)$ in a special way. In particular, suppose:

$$E(Y_{0i}|c, t) = \beta_t + \gamma_c. \quad (11)$$

Differences-in-Differences (DD) Models (5)

Therefore, without immigration, the unemployment rate can be expressed as the sum of a year effect (common to all cities), given by β_t , and a city effect that is fixed over time (given by γ_c). Suppose, for simplicity, that:

$$E(Y_{1i}|c, t) = E(Y_{0i}|c, t) + \delta, \quad (12)$$

where δ is the effect of the Mariel immigration (**treatment effect**). This is our parameter of interest here.

The unemployment status of people living in Miami and the comparison cities in 1979 and 1981 can then be written as:

$$Y_i = \beta_t + \gamma_c + \delta M_i + \epsilon_i, \quad (13)$$

where $E(\epsilon_i|c, t) = 0$, and

Differences-in-Differences (DD) Models (6)

$$M_i = \begin{cases} 1 & \text{if } i \text{ lived in Miami after 1980} \\ 0 & \text{otherwise.} \end{cases}$$

Now, differencing unemployment rates across cities and years gives:

$$\begin{aligned} \text{DD} &= \{E(Y_i|c = \text{MIAMI}, t = 1981) - E(Y_i|c = \text{COMP}, t = 1981)\} \\ &\quad - \{E(Y_i|c = \text{MIAMI}, t = 1979) - E(Y_i|c = \text{COMP}, t = 1979)\} \\ &= \{\beta_{1981} + \gamma_{\text{MIAMI}} + \delta - \beta_{1981} + \gamma_{\text{COMP}}\} \\ &\quad - \{\beta_{1979} + \gamma_{\text{MIAMI}} - \beta_{1979} + \gamma_{\text{COMP}}\} && (14) \\ &= \delta. && (15) \end{aligned}$$

Hence, the DD method identifies the causal impact of the Mariel immigration.

Differences-in-Differences (DD) Models (7)

Equation (13) could be estimated in a regression of micro data for cities and years. As proved in (14), that regression will allow us to identify our **treatment effect**.

At this point, the only regressors in (13) are years, cities and M_i (the interaction between living in Miami and time after 1980). But we could estimate a variant of the model in which we adjust for individual specific characteristics (e.g., race, age, education, etc.), X_i . That is:

$$Y_i = \mathbf{X}'_i \beta_0 + \beta_t + \gamma_c + \delta M_i + \epsilon_i, \quad (16)$$

where β_0 includes a constant.

Differences-in-Differences (DD) Models (8)

DD problems

The key identifying assumption in the previous models (and of the DD method in general) is that the interaction terms are zero **in the absence of the intervention**. In Card's model, this means that there is only one trend captured by β_t common to both the *treatment group* (Miami) and the *comparison group* (or *control group*, the other cities).

But if unemployment rates evolve differently across cities because of different shocks affecting cities (as well as individuals in those cities) differently, then DD method will not identify δ .

As an illustration, suppose that equation (13) is now replaced by:

Differences-in-Differences (DD) Models (9)

$$Y_i = \beta_{t,c} + \gamma_c + \delta M_i + \epsilon_i. \quad (17)$$

Notice $\beta_{t,c}$ now captures time *trends that differ across cities*. The DD estimator will be:

$$DD = \delta + (\beta_{1981,MIAMI} - \beta_{1979,MIAMI}) - (\beta_{1981,COMP} - \beta_{1979,COMP}). \quad (18)$$

Therefore, the DD method does **not** identify the treatment effect of interest as long as the *trends in unemployment rates* in Miami and the comparison cities between 1979 and 1981 *differ*.