1. **Lecture**

**Types of data**

1. *Cross-section data*

Data across different units in one time period. Example: One round (year) of the National Sample Survey (NSS) of India, which surveys different households at a point in time. The next round (year) does not follow the same set of households, but is a separate cross-section.

1. *Pooled cross section data*

Cross section data across multiple time periods pooled together. Example: Combine multiple rounds of the NSS.

1. *Time-series data*

Data for one unit over time. Example: A data-set comprising GDP and inequality of one country over a span of say, 100 quarters.

1. *Panel data*

Combination of cross-section and time-series data. Data across many units followed up over different points in time. Example: (1) A dataset comprising GDP, poverty and inequality of 150 countries across 10 years. (2) The Annual Survey of Industries, which follows manufacturing firms between 1998-99 and 2014-15 in India.

Having seen these types of data, we will now delver deeper into using some of these data sets and focus on achieving causal estimations.

1. **Models using pooling Independent Cross Sections across Time**

Many surveys of individuals, families, and firms are repeated at regular intervals each year, or once every year or couple of years. These are independent cross section data such as the National Sample Survey in India which randomly samples households in India once in 5 years, or *Current Population Survey* (or CPS), which randomly samples households in the USA each year.

These individual cross section data may be pooled together over time. Reasons why we may pool individual cross section data:

1. One reason for using independently pooled cross sections is to increase the sample size. By pooling random samples drawn from the same population, but at different points in time, we can get more precise estimators and test statistics with more power.
2. But we may also be specifically interested in what happens to an outcome variable over time. Example:

Open FERTIL1.dta (W, p 341). The data is from the National Opinion Research Center’s *General Social Survey* for the even years from 1972 to 1984, inclusively. We use these data to estimate a model explaining the total number of kids born to a woman (*kids*). An interesting question to ask here would be: after controlling for other observable factors, what happened to fertility rates over time? To do this, simply add a binary variable for each year.

refers to year dummy variables. The base year is 1972 and is dropped from the regression. are individual characteristics including years of education, age, age squared, race, region of the country when living at age 16, and living environment at age 16. Results indicate that the coefficients on the year dummy variables show a sharp drop in fertility in the early 1980s.

Interpretation: The coefficient on year82implies that, holding education, age, and other factors fixed, a woman had on average .52 less children, or about one-half a child, in 1982 than in 1972. This is a very large drop. Another interpretation: 100 women in 1982 are predicted to have about 52 fewer children than 100 comparable women in 1972. This drop is separate from the decline in fertility due to the increase in average education levels (since we control for education)

State and time fixed effects reduce omitted variables bias: If state-level characteristics that affect wages are unavailable, controlling for state fixed effects may reduce some of the omitted variable bias. Similarly, all changes that took place in the economy may not be captured in variables. Time fixed effects can control for these changes.

Example:

Suppose we have NSS data for two years – 2011-12 and 2009-10. We want to measure the effect of training on wages. Suppose hypothetically that unionization laws changed in India between 2009-10 and 2011-12, making it difficult to form worker unions. As a result, bargaining power of workers go down and wages may be lower in 2011-12 compared to 2009-10. How will you control (capture) this in a regression? To capture this differences in wages across the two years, we can add a year dummy for 2011-12.

refers to log wages for individual ‘i’ living in state ‘s’ and time ‘t’. takes a value 1 if the worker underwent a skills training. Year201112 is a binary variable taking the value 1 if year==2011-12 and 0 if year=2009-10. The coefficient on year201112,, captures the year effect.

We can also include state fixed effects – one binary variable representing each state to capture time-invariant systematic unobserved differences in wages across states. These differences could arise from institutional and economic differences across states. For example, wages is Kerala is always higher than Maharashtra which in turn may always be higher than Uttar Pradesh. With a dummy variable representing each state, we can use to represent state fixed effects. Suppose there are Ns states, the vector consists of a series of Ns – 1 coefficients.

1. **“Fixed-effect estimator” using panel-data**

State fixed effects or other group fixed effects can account for unobservables at the state or group level in cross section or pooled cross section data. But we can account for unobservables at the individual level if we have panel data.

-------- (2)

Unobservable characteristics which are time-invariant for each unit are captured in Adding as a variable for each ‘i’ and estimating them implies holding these fixed factors constant. The βs estimated this way are called “fixed effect estimators” or “within estimators”. Controlling for individual fixed effects is possible by the virtue of panel nature of the data – that is, by observing the same individual units over time.

Algebraically, coefficients from a first-difference regression below is algebraically the same as (2):

--------- (3)

Notice that the fixed effect being constant over time, disappears in (3).

*Example:* Freeman (1984) uses four datasets to estimate the effects of being a union member on wages. He controls for individual fixed effects assuming that selection into union status is based on unobserved but fixed individual characteristics. Controlling for individual fixed effects address the missing variable bias from variables constant for an individual across time. Note that the cross section estimates are always higher than fixed effects estimator. What does this say about the nature of bias, and why? --- Answer: Positive selection bias in the cross-section estimate, because those who have higher motivation may join the union, and higher motivation also leads to better wages.

**Table 1: Effects if union status on wages**

|  |  |  |
| --- | --- | --- |
| **Survey** | **Cross-section estimate** | **Fixed effects estimator** |
| May Current Population Survey | .19 | .09 |
| National Longitudinal Survey | .28 | .19 |
| Michigan PSID | .23 | .14 |
| QES | .14 | .16 |

1. **Difference-in-Difference estimator (Policy Analysis with Pooled Cross Sections)**

Example (W, p345) Kiel and McClain (1995) study the effects of building of an incinerator on housing prices in North Andover, Massachusetts. The rumour that a new incinerator would be built in North Andover began after 1978, and construction began in 1981. The incinerator was expected to be in operation soon after the start of construction; the incinerator actually began operating in 1985. We will use data on prices of houses that sold in 1978 and another sample on those that sold in 1981. The hypothesis is that the price of houses located near the incinerator would fall relative to the price of more distant houses. A naive analyst would use only the 1981 data and estimate a very simple model:

rprice is the price of house sold in real terms. Nearinc is a binary variable which takes the value 1 if the house is near the incinerator, 0 otherwise. Estimating this equation using the data in KIELMC.dta for the year 1981 tells us that the average selling house price in the areas near the incinerator was $30,688.27 less than for the latter group. Unfortunately, this does not mean that being near an incinerator causes lower housing prices because if we run the same regression for the year 1978 we get a similar result (-18,824.37 is the coefficient on nearinc).

Then, how can we tell whether being near the incinerator reduced housing price? The key is to look at how the coefficient on *nearinc* changed between 1978 and 1981, which is:

= 30,688.27 - (-18,824.37) = - 11,863.9.

This above estimator is known as the difference in difference estimator because it is a difference of two differences - the difference in prices of those near the incinerator and those far from the incinerator after the incinerator’s building began, and the difference in prices of those near the incinerator and those far from the incinerator before the incinerator’s building was announced/rumoured. Mathematically, this is:

Above is simple mean difference. But if we want to include other characteristics of the house as controls, we need to arrive at this through a regression. How? See the regression below:

is the difference in difference estimator. How?

Verify that is indeed the estimator.

More generally, suppose you want to measure the effect of event x on y after the event x happened. You have data before x happened (time=1 or pre-treatment period), and after x happened (time=2 or post-treatment period). The data may be panel or pooled cross-section. Let post=0 and post=1 denote the time periods before and after event x happened, respectively. Treatment=1 denotes the units (households or firms, etc.) which received the treatment, or for whom x happened (treatment group). Treatment=0 denote units for whom x did not happen (control group). The difference-in-difference estimator is in the regression below.

The fundamental assumption of the Diff-in-Diff estimator

The key identifying assumption is that if there is no treatment, the change in y is the same in both control and treatment group. Mathematically, should not be statistically significantly different from 0 when there is no treatment. In other words, trend in the outcome *variable* (the effect of post) would be the same in control and treatment groups in the absence of treatment. Show W, figure 5.2.1. In our house price and incinerator example, house prices should be growing the same way, in the absence of incinerator, in localities near and far from the incinerator.

1. **Difference in difference across multiple time periods**

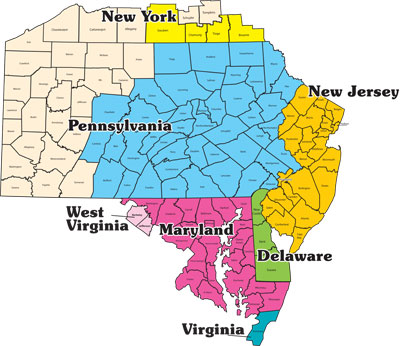
Suppose there is a policy change in different states implement at different time points. This is typical in many countries which have a federal set up. There are many jurisdictions wherein states have the power and autonomy to implement laws or implement their own version of a central law. In such cases, the regression design below can be used to estimate the effects of these policy changes.

Example: Let be the minimum wage in state s at time t in India, and could be the (1) average daily wage in state s at time t or (2) employment in state s and time t. Note that each states set a different minimum wage, which is revised every year or every couple of years. The above regression design helps us estimate the effects of minimum wage on wage or employment. are time and state fixed effects respectively (compare them to treatedit and postit in the previous equation. The parameter of interest is , which provides the impact of minimum wage on wage/employment provided there is no endogeneity bias.

1. **Hands on**

New Jersey raised its minimum wage from $4.25 to $5.05 in April 1992. In the neighbouring state of Pennsylvania the minimum wage remained at the previous value of $4.25. This is a natural experiment, the effects of which were investigated by Card and Krueger whose results can be found in the following publication:

*Card, D. and A.B. Krueger (1994) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. American Economic Review. Vol. 84 (4), pp. 772-793.*



Open minwage.dta (You can also download this dataset from [David Card’s UC Berkley website](http://davidcard.berkeley.edu/data_sets.html), but the one I shared is processed and readily useable). Type describe to see a list of variables and their labels. Note that there should be 820 total observations, corresponding to two observations (before and after the minimum wage legislation) for each of 410 restaurants.

1. This raise in the minimum wage is a natural experiment. A natural experiment occurs when some exogenous event— often a change in government policy — changes the environment in which individuals, families, firms, or cities operate. A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change. [[1]](#footnote-1) In this above setting, identify the control and treatment groups, and post- and pre-treatment period.
2. The main outcomes of interest in this study are starting wage and full-time equivalent (FTE) employment. Generate the FTE variable for each observation as the number of full-time employees (including managers) plus 0.5 times the number of part-time workers.
3. Show the distribution of wages in both states, before and after the reform. Is the minimum wage increase visible in the histogram?
4. Create a table similar to subset of Table 2 in the Card and Krueger study. Your table should show the means of FTE employment, starting wages, the percent of workers earning exactly $4.25, and number of hours open for restaurants by wave (the variable `post') and by (NJ and PA). Your table should also show the t-statistic for a test of the equality of means between NJ and PA (using “ttest” command in stata). What do you conclude from these t-statistics?
5. One way to evaluate the effect of the minimum wage on employment would be to compute the difference in FTE employment for NJ and PA in the second time period, when NJ has a higher minimum than PA. Compute the effect of the minimum wage in this way using OLS regression. Do you think this is an unbiased estimate of the effect of minimum wages on employment (what conditions would need to be true for it to be an unbiased estimate)? How does Table 2 help you think about this?
6. Using only the averages in your version of Table 2 from above, compute a difference-in-differences estimate of the effect of minimum wages on FTE employment.
7. Using a regression of starting wages on a constant term, a dummy variable for post, a dummy variable for NJ, and NJ\*post, and compute the effect of raising the minimum wage on starting wages. Note that there are four parameters estimated in this regression: provide an interpretation for each one (hint: compare your coefficient estimates to the data in your Table 2) in terms of the implied conditional expectation function.
8. Similarly compute the DD estimate for the effect of minimum wages on employment. Compare the coefficients to the FTE averages in the paper’s table 2.
9. Under what assumptions do estimates in 7 and 8 provide an unbiased estimate of the causal impact of minimum wages on employment? Invoking the assumptions you outlined in the previous question, estimate what New Jersey employment (in terms of average FTEs) would have been, had the minimum wage policy not passed.
10. Do you think the evidence on the effect of minimum wages on employment is persuasive? Are there major threats to validity that remain, or are you convinced?
11. In a subsequent study, Card and Krueger produced a figure showing fast food restaurant employment in NJ and PA over a longer time period. How does this data strengthen or weaken the credibility of your findings about the causal impact of minimum wages? See figure 2 in: *Card, David and Alan B. Krueger. 2000.* *"Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply."* *American Economic Review*, *90(5): 1397-1420.*

1. Unlike a true experiment, in which treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. [↑](#footnote-ref-1)