EC203: Applied Econometrics
Differences in differences

Dr. Tom Martin

University of Warwick

Illustrative reading:

- Wooldridge: Chapter 13
- Mostly Harmless Econometrics: Chapter 5

A cross section model

With cross-sectional data we typically specify models of the form:

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$$

- There are N individuals at a single point in time.
- With cross-sectional data we only have variation, in Y_i and X_{ji}, between individuals at a single point in time.
- Therefore, all estimators rely on between individual variation.

A pooled cross section model

A general pooled cross-section model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + \epsilon_{it}$$

- There are N_0 individuals in period $t = 0$, N_1 individuals in period $t = 1$, ... continuing, N_T individuals in period $t = T$.[1]
- In pooled cross-section individuals are different in each period. The number of observations in each period will typically differ.
- With pooled cross-sectional data we can exploit variation, in Y_{it} and X_{jit}, between individuals and across time.
- Therefore, estimators can exploit variation between individuals and variation across time.

[1] d_t is shorthand for a set of time dummies: $d_t = \sum_{s=2}^{T} \delta_s d_s$.

DD with pooled cross-section data

With pooled cross-section data a powerful estimation technique for estimating causal effects is difference in differences (DD).[2] Good DD techniques rely on good natural experiments, such as a change of government policy. In a typical set-up there is:

- A treatment group: a set of observations which recieves the new policy.
- A control group: a set of observations which does not recieve the new policy.
- Both the treatment and control groups are observed both before and after the policy change.
- The crucial assumption is the common trends assumption: in absence of the policy both the treatment and control groups would have followed the same trends. The greater the similarity between the treatment and control groups the more likely this assumption is to be satisfied.

[2] It is also implementable with panel data, but it is more common with pooled cross-sections.

DD with pooled cross-section data

Example: Does increasing the minimum wage decrease employment? In competitive labour markets, one would expect an increase in the minimum wage to cause employment to fall.

- In February 1992 both New Jersey and Pennsylvania (nieghbouring states in the US) both charged a state minimum wage of 4.25 dollars.
- On April 1^{st} 1992 New Jersey raised the state minimum wage from 4.25 to 5.05 dollars. The state minimum wage in Pennsylvania remained at 4.25.
- Card and Krueger (1994) collected data on number of people employed (among other things) in the same type of fast-food restaurants in New Jersey and Eastern Pennsylvania before and after the policy change.

DD with pooled cross-section data

In their set-up:

- The treatment group is New Jersey
- The control group is Pennsylvania
- This is known as a natural experiment: naturally occuring quasi-random variation in the minimum wage.
- Notes:
  - The data was collected both before and after the policy change.
  - There is no need for the restaurants to be the same in each period, as long as they are collected in the relevant geographic areas.
  - As such, we have a pooled cross-section of restaurants.

DD with pooled cross-section data

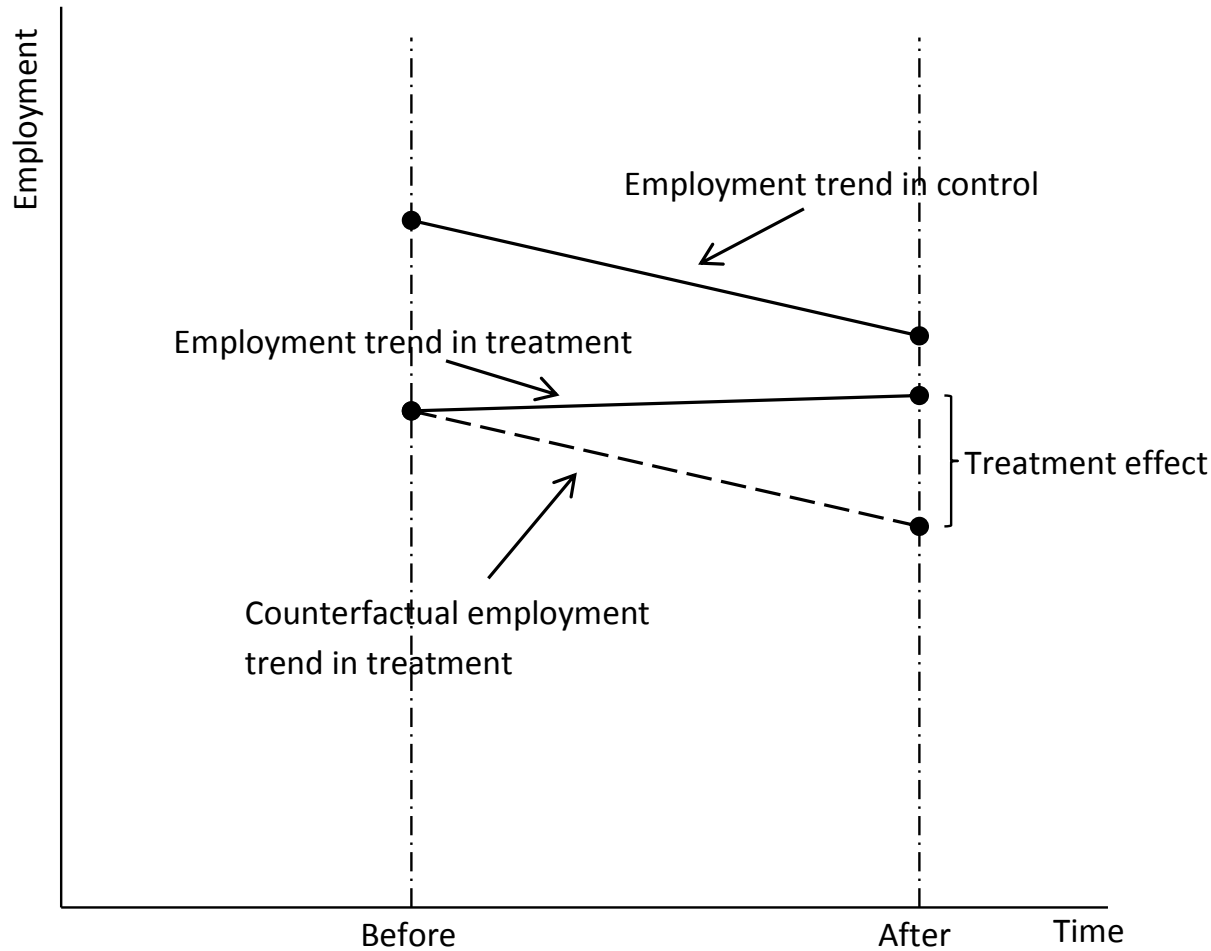
Difference in differences calculated using average employment in fast food restaurants:

	PA	NJ	Difference, NJ-PA
FTE employment before	23.33	20.44	−2.89
	(1.35)	(0.41)	(1.44)
FTE employment after	21.17	21.03	−0.14
	(0.94)	(0.52)	(1.07)
Change in FTE	−2.16	0.59	2.76
	(1.25)	(0.54)	(1.36)

The DD estimate is 2.76, that is, a rise of approximately three FTE employees, on average. The standard error is 1.36 and the t-statistic is approximately 2, implying we would reject at around the 5% level.[3]

[3] DD=$(F\bar{T}E_{PA,Aft} - F\bar{T}E_{PA,Bf}) - (F\bar{T}E_{NJ,Aft} - F\bar{T}E_{NJ,Bf}) = 2.76$

DD with pooled cross-section data



DD with pooled cross-section data

The important assumption, in all DD empirical strategies, is the common trends assumption.

- The common trends assumption: In the absence of the treatment, the treatment and control would have followed the same trend.
- Note this is not the same as stating the control and treatment groups have the same level of the outcome.

DD with pooled cross-section data

It is common to use regression to estimate DD estimates since:

- We can control for other variables.
- It is easy to calculate standard errors.
- We can extend the framework to include: i) multiple treatment/control states as well as time periods, ii) look at extended pre and post-treatment effects.

DD with pooled cross-section data

The required regression takes the form:

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s * d_t) + \epsilon_{ist}$$

Where:

- Y_{ist} is FTE employment in restaurant i, in state s, in period t. With $s \in (NJ, PA)$ and $t \in (feb, nov)$.
- $NJ_s = 1$ if the restaurant is in New Jersey (treatment state), and zero if in Pennsylvannia (control state).
- $d_t = 1$ for observations in November (post-treatment) and zero for observations in February (pre-treatment).
- $NJ_s * d_t$ is an interaction term between the state and time dummies.
- ϵ_{ist} is the error term for restaurant i, in state s, at time t.

DD with pooled cross-section data

The required regression takes the form:

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s * d_t) + \epsilon_{ist}$$

With associated assumptions:

- $E[\epsilon] = 0$
- $E[\epsilon|X] = 0$
  - Where breaking down this assumption: $E[\epsilon|NJ] = 0$, $E[\epsilon|d] = 0$, and importantly, $E[\epsilon|NJ * d] = 0$ is the common trends assumption.
- $V[\epsilon|X] = \sigma^2$
- $cov[\epsilon_{ist}, \epsilon_{jst}|X] = 0$
- $\epsilon|X \sim N(0, \sigma^2)$

DD with pooled cross-section data

The required regression takes the form:

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s * d_t) + \epsilon_{ist}$$

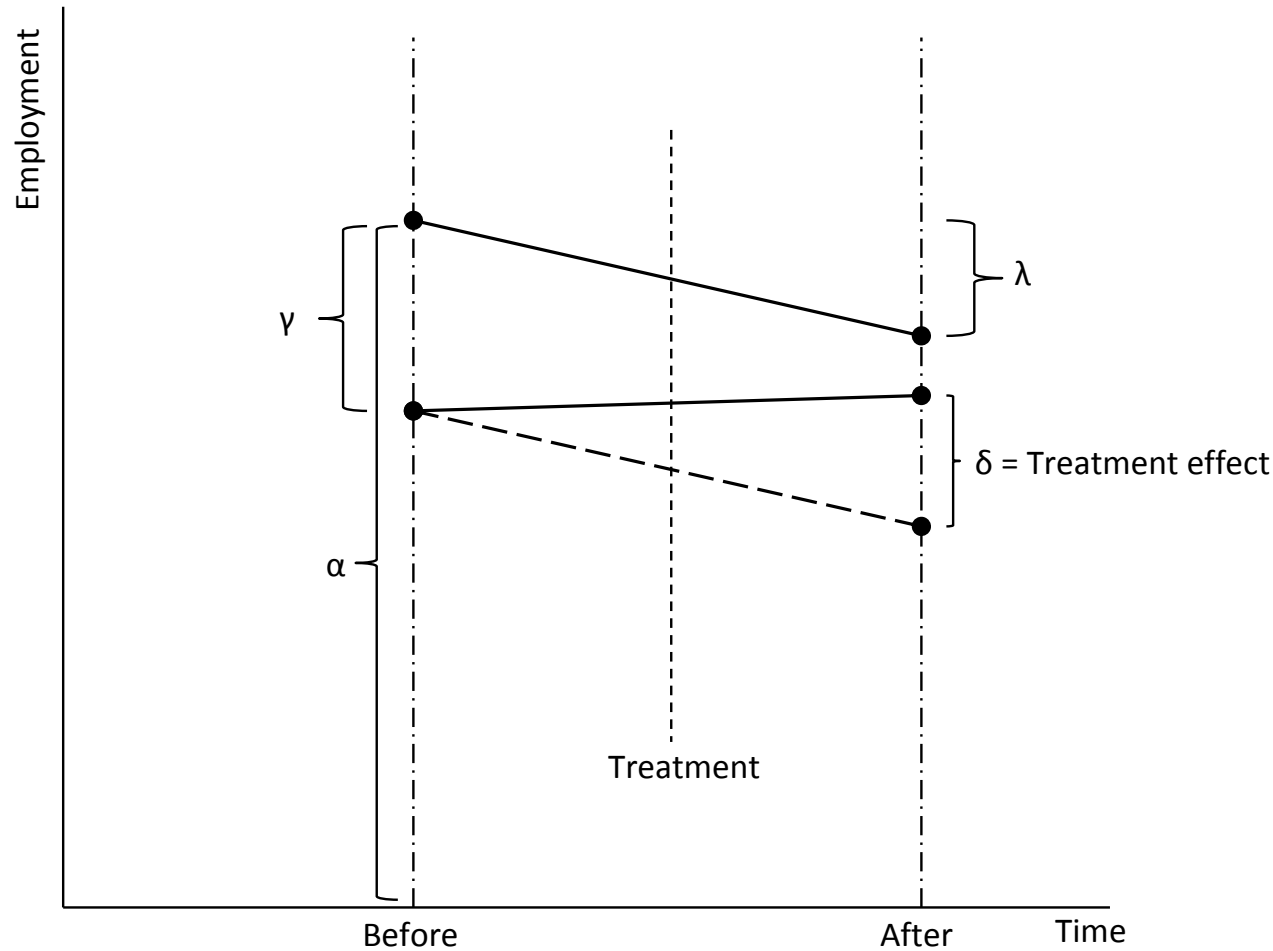
Turning dummy variables on and off we get:

- PA pre treatment: $E[Y|NJ = 0, d = 0] = \alpha$
- PA post treatment: $E[Y|NJ = 0, d = 1] = \alpha + \lambda$
- NJ pre treatment: $E[Y|NJ = 1, d = 0] = \alpha + \gamma$
- NJ post treatment: $E[Y|NJ = 1, d = 1] = \alpha + \gamma + \lambda + \delta$

Such that the DD estimate is:

$$DD = (NJ_{post} - NJ_{pre}) - (PA_{post} - PA_{pre}) = \delta$$

DD with pooled cross-section data



DD with pooled cross-section data

The estimated regression in the current example would be:

$$\hat{Y}_{ist} = 23.33 - 2.89NJ_s - 2.16d_t + 2.76NJ_s * d_t$$

Turning dummy variables on and off we get:

- PA pre treatment: $E[Y|NJ = 0, d = 0] = 23.33$
- PA post treatment: $E[Y|NJ = 0, d = 1] = 21.17$
- NJ pre treatment: $E[Y|NJ = 1, d = 0] = 20.44$
- NJ post treatment: $E[Y|NJ = 1, d = 1] = 21.03$

Such that the DD estimate is:

$$DD = (21.03 - 20.44) - (21.17 - 23.33) = 2.76$$

The increase in the minimum wage increased FTE by 2.76 on average. (The coefficient is also significant at the 5% level.)

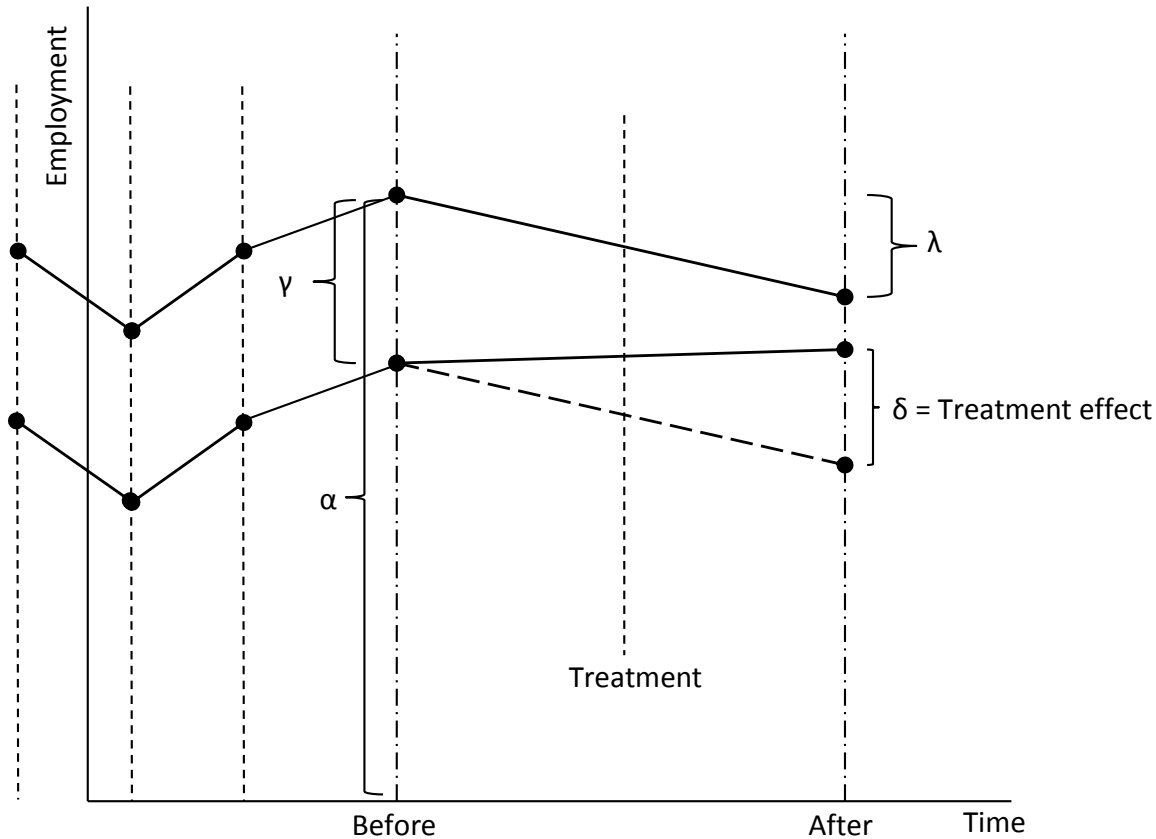
DD with pooled cross-section data

How convincing is the above story?

- First note, the treatment and control state can differ in terms of average employment rates, this is captured by the state fixed effect (the NJ dummy).
- The DD estimate will have a causal interpretation if the employment trends would be the same in both states in the absence of the treatment. (The common trends assumption.)
- Treatment induces deviation from the trend as seen in the previous figures.
- This is a difficult assumption to test, as by definition you don't know what would have happened to the treatment group had they not entered the treatment.

DD with pooled cross-section data

Even if pre-treatment trends were the same (as below), the concern is if anything else occurred between pre and post treatment in the treatment state but not the control. For instance, other policy changes which asymmetrically shifted average employment.



DD with pooled cross-section data

A more convenient representation of the DD model is:

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s * d_t) + \epsilon_{ist}$$
$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \epsilon_{ist}$$

- $\gamma_s = \alpha + \gamma NJ_s$: represents the state fixed effects (intercept and state dummy).
- $\lambda_t = \lambda d_t$: represents the time fixed effects (time period dummy).
- $D_{st} = NJ_s * d_t = 1$ if state s at time t is in the treated state and zero otherwise. Such that δ is the DD estimator.[4]
- ϵ_{ist} is the error.

[4]Note the intercept α is typically assumed to be included in the state (or time) fixed effects for convenience. In the current case $D_{st} = 1$ if $NJ = 1$ and $T = 1$, i.e. in New Jersey in the post treatment time period. $D = 0$ in Pensylvannia in both periods and in New Jersey in the pre-treatment period.

DD with pooled cross-section data

Extending the regression to more than two states and periods:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \epsilon_{ist}$$

- $\gamma_s = \sum_S \gamma_j S_j$: represents the state fixed effects.
- $\lambda_t = \sum_T \lambda_j d_j$: represents the time fixed effects.[5]
- $D_{st} = 1$ if state s at time t is in the treated state and zero otherwise. Such that δ is the DD estimator.
- ϵ_{ist} is the error.

[5]Note the intercept is assumed included in the state or time fixed effects for convenience.

DD with pooled cross-section data

Extending the regression to more than two states and periods:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \beta_1 X_{1ist} + ... + \beta_k X_{kist} + \epsilon_{ist}$$

- $\gamma_s = \sum_S \gamma_j S_j$: represents the state fixed effects.
- $\lambda_t = \sum_T \lambda_j d_j$: represents the time fixed effects.
- $D_{st} = 1$ if state s at time t is in the treated state and zero otherwise. Such that δ is the DD estimator.
- $\beta_1 X_{1ist} + ... + \beta_k X_{kist}$ are a set of control variables.
- ϵ_{ist} is the error.

DD with pooled cross-section data

Notes on control group and control variables:

1. Selection of control group: the DD set-up can be made much more general. For instance, instead of states one could use demographic groups, some of which are affected by a policy change and some are not.
2. Selection of control variables: it is important to only include exogenous controls (i.e. not affected by the policy), otherwise you will remove variation caused by the policy. That is, you will remove the effect you are trying to estimate.