LECTURE 9. DIFFERENCE-IN-DIFFERENCES AND FIXED EFFECTS ESTIMATION

Dr. Tom O'Grady

Note: Some of the first half of this lecture is based on a lecture given by Teppei Yamamoto at MIT. The diagrams are also his

For the last two weeks of Causal Analysis, we move beyond experiments and natural experiments to look at causal inference over time. Difference-in-differences, fixed effects and synthetic control all have in common the key feature that a treatment is applied over time in one or more units and not applied in one or more comparable units that serve as a counterfactual. These methods are frequently used used for policy evaluation. They can help us understand the impact of a policy that is introduced in some places (e.g. cities, regions or countries) but not others without any randomisation. We'll look at two examples today: minimum wage laws and tax laws.

Difference-in-differences and fixed effects estimation typically have lower internal validity than the previous examples we've seen in this module, because without genuine randomisation (or as-if randomisation in a natural experiment) it is harder to be sure that we have found a good counterfactual for the treated unit(s). But they may have greater external validity because we can potentially include a wider range of units in our analysis. In this lecture we first look at difference-in-differences (with two units and time periods) and then the more general method of fixed effects (with multiple units and time periods).

1. Two Groups and Two Time Periods: Difference-in-Differences Estimation

Difference-in-differences analysis involves the following key features:

1. Two *time periods*: $T = 0$ (pre-treatment) and $T = 1$ (post-treatment)

2. A *treatment* D that occurs between T=0 and T=1

3. Two *groups*: $G = 1$ (group that receives the treatment in second period) and $G = 0$ (group that never receives the treatment)

4. Multiple *units* within the groups, such as people: $i \in \{1, ..., N\}$

5. One of two *data setups*:
   - *Repeated Cross-Section*: a random sample of units in both groups in both periods (which makes data for each group comparable across both time periods)
   - *Longitudinal or Panel Data*: the exact same units are observed in both periods

The key idea behind difference-in-differences is that despite D being allocated to $G = 1$ non-randomly, change over time in the untreated unit serves as a counterfactual for what would have happened over time in the treated group in the absence of the treatment. Units can fall into one of four categories:

	Time Period	
Group	$T = 0$	$T = 1$
$G_i = 1$	$D_{i0} = 0$	$D_{i1} = 1$
(treatment group)	(untreated)	(treated)
$G_i = 0$	$D_{i0} = 0$	$D_{i0} = 0$
(control group)	(untreated)	(untreated)

1.1. Motivating example: New Jersey's Minimum Wage Law

One of the first difference-in-differences papers ever written – and perhaps the most famous example ever in the field of economics – is on this week's reading list and comes from the neighbouring American States of New Jersey and Pennsylvania. A common assumption in classical economics is that minimum wage laws reduce low-paid employment because they force firms to raise wages above the profit-maximising level that would occur without regulation. With labour more expensive, it becomes optimal for firms to cut back on the number of workers.

Many economists now think that this view is too simplistic, and possibly badly wrong. Card and Krueger's study is one of the most important pieces of evidence for this new, more sceptical approach to minimum wages. In April 1992, New Jersey introduced a 20% minimum wage increase from \$4.25 to \$5.05 per hour, whereas Pennsylvania did not. Card and Krueger carried out a survey of fast-food restaurants – one of the major employers of cheap labour – before and after the minimum wage increase in both New Jersey and in the region of Pennsylvania that borders it. The idea is that changes in employment in the Pennsylvania fast-food restaurants serve as a counterfactual for what would have happened to the New Jersey restaurants without the new law, because the Pennsylvania border area has a very similar economy to New Jersey. They found that, against economists' expectations, employment actually went up slightly in New Jersey compared to Pennsylvania after the law change.

Here, the groups G are New Jersey and Pennsylvania, the units i are individual fast-food restaurants, the treatment D is the minimum wage law change, and the time periods (0 and 1) are before and after the minimum wage law was enacted.

1.2. Difference-in-Differences Strategy

In difference-in-differences estimation, the causal effect that we would like to estimate is:

$$\tau \;=\; E[Y_{i1}(1) - Y_{i1}(0)|G_i = 1]$$

This is the ATE for the treatment group in the post-treatment period, because:

- $Y_{it}(0)$ is the potential outcome for unit i in period t when not treated
- $Y_{it}(1)$ is the potential outcome for unit i in period t when treated

Once again, though, we face the fundamental problem of causal inference. We don't observe period 1 outcomes for the treated group in the absence of the treatment $E[Y_{i1}(0)|G_i = 1]$.

One possibility would be to fill in the missing potential outcome with observations from the treated units before they were treated, in period 0. That is, we ignore the control units altogether and just do a simple before-and-after comparison of the treated units:

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$E[Y_{i0}(0)\|G_i = 1]$	$E[Y_{i1}(1)\|G_i = 1]$
Control Group ($G_i = 0$)	$E[Y_{i0}(0)\|G_i = 0]$	$E[Y_{i1}(0)\|G_i = 0]$

However, this involves a very strong assumption: that the treated units' potential outcomes did not change over time, so that the treated units at time $t = 0$ provide a valid counterfactual for what would have happened without the treatment at time $t = 1$. This is very often likely to be violated. Suppose, for instance, that there was economic growth over the year 1992 in New Jersey. Then even in the absence of the minimum wage law change, unemployment might have fallen over time between $t = 0$ and $t = 1$. The $t = 0$ figure would be too low. More generally, all people, towns etc. tend to change a lot over time, and we cannot just assume that their past behaviour is a perfect counterfactual for the present.

A second possibility would be to fill in the missing potential outcome with observations from the untreated units at $t = 1$. That is, we ignore the first-period data altogether and just compare the treated and untreated units directly:

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$E[Y_{i0}(0)\|G_i = 1]$	$E[Y_{i1}(1)\|G_i = 1]$
Control Group ($G_i = 0$)	$E[Y_{i0}(0)\|G_i = 0]$	$E[Y_{i1}(0)\|G_i = 0]$

However, this also involves another very strong assumption: that the control units provide a good counterfactual to the treated units, even though treatment was not randomly assigned. There is no reason to believe that this would be the case. There are many reasons why employment patterns in Pennsylvania restaurants might be different in general to New Jersey, since there are many other legal and economic difference between states.

Difference-in-Differences instead involves a combination of these two strategies. We **take the difference over time in the $G = 1$ group and subtract from it the difference over time in the $G = 0$ group:**

$$\tau_{dd} \;=\; \Big\{E[Y_{i1}|G_i = 1] - E[Y_{i0}|G_i = 1]\Big\} - \Big\{E[Y_{i1}|G_i = 0] - E[Y_{i0}|G_i = 0]\Big\}$$

This assumes that changes over time in the control group provide a valid counterfactual for change over time in the treatment group in the absence of the treatment. This is known as the **parallel trends assumption**. While it is by no means guaranteed to hold, this assumption is much more reasonable than either of the two previous assumptions. It does not require the border areas of New Jersey and Pennsylvania to be similar to each other in a cross-sectional sense, nor does it require New Jersey in $t = 0$ to be similar to New Jersey in $t = 1$. Instead it simply requires that changes in the border of Pennsylvania provide a good guide to the changes that would have occurred in New Jersey. For example, it requires that economic growth be the same across both regions. This is probably a reasonable assumption, especially for the subset of restaurants close to the state borders where economic trends are unlilkely to differ strongly.

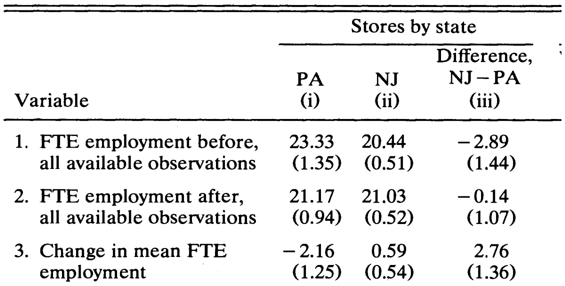
1.3. Estimation

To find the difference-in-differences treatment effect, we once again replace the potential outcomes with observed outcomes. We simply take the mean difference in observed outcomes over time in the control group and subtract it from the same mean difference in the treated group:

$$\begin{aligned}\hat{\tau} \;=\; & \Big\{\frac{1}{N}\sum_{i=1}^{m}(Y_i \mid G_i = 1, T_i = 1) - \frac{1}{N}\sum_{i=m+1}^{N}(Y_i \mid G_i = 1, T_i = 0)]\Big\} - \\ & \Big\{\frac{1}{M}\sum_{i=1}^{m}(Y_i \mid G_i = 0, T_i = 1) - \frac{1}{M}\sum_{i=m+1}^{N}(Y_i \mid G_i = 0, T_i = 0)]\Big\}\end{aligned}$$

where there are N units in the treated group and M in the control group (note that for simplicity, we are assuming no attrition between the two periods, so that the same number of units are observed in each period). This is simple to calculate in the Card and Krueger data, as shown in Figure 1. Employment fell on the Pennsylvania side of the border by an average of 2.16 employees per restaurant, which provides a counterfactual for New Jersey. In New Jersey, employment rose by 0.59 employees. The difference-in-differences effect of the minimum wage rise was therefore a *rise* in average employment of 2.75 workers (0.59 - -2.16), the opposite effect to the prediction of classical economic theory.[1]

Figure 1: **Means by State and Time Period in Card-Krueger Study**



	Stores by state		
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

As always, this can also be calculated using a regression. The simple difference in means is numerically identical to the following regression:

$$Y_i = \mu + \gamma G_i + \delta T_i + \tau G_i T_i$$

[1] Economists explain this result in several ways. One theory – although not endorsed by Card and Krueger – is that fast-food restaurants enjoy "monopsony" power as the sole employer in some areas, and therefore use their market power to hold down wages below the competitive market rate implied by economic theory. In this situation, a rise in the minimum wage serves only to raise the wage toward that competitive level, allowing firms to maintain employment while taking a hit to their profits.

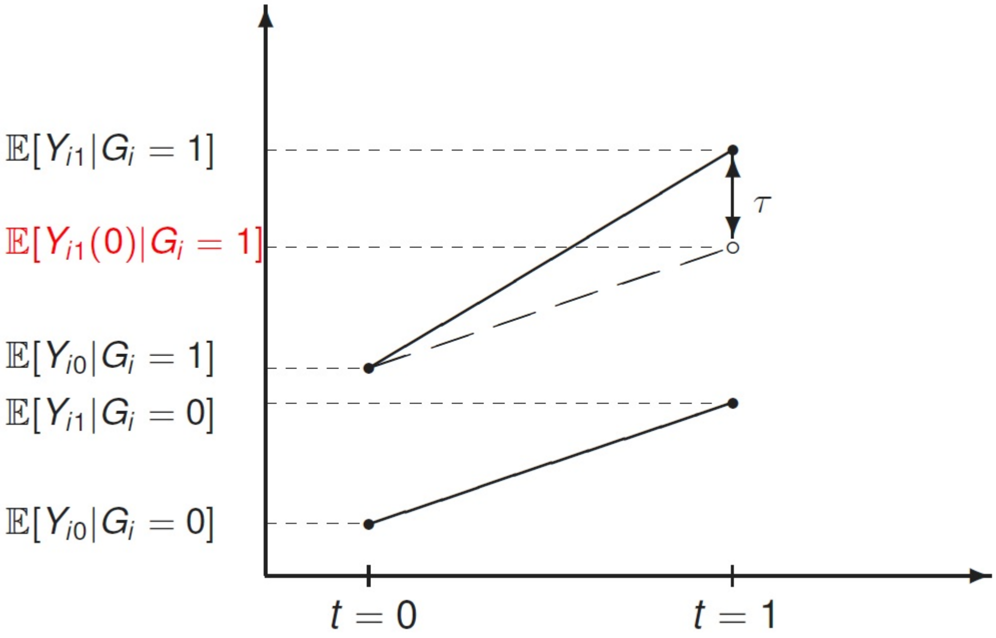
where $T_i \in \{0,1\}$ is a dummy variable indicating whether unit i (e.g. a restaurant) is observed in the pre- or post-period and $G_i \in \{0,1\}$ is a dummy variable indicating whether unit i is in the treatment or control group (e.g. New Jersey or Pennsylvania). We can easily see that $\hat{\tau} = \tau_{dd}$ by calculating predicted values from the regression for our four groups, corresponding directly to the cells in Figure 1:

	Control $G_i = 0$	Treated $G_i = 1$	Treated - Control
Before $(T_i = 0)$	$\hat{\mu}$	$\hat{\mu} + \hat{\gamma}$	$\hat{\gamma}$
After $(T_i = 1)$	$\hat{\mu} + \hat{\delta}$	$\hat{\mu} + \hat{\gamma} + \hat{\delta} + \hat{\tau}$	$\hat{\gamma} + \hat{\tau}$
After - Before	$\hat{\delta}$	$\hat{\delta} + \hat{\tau}$	$\hat{\tau}$

2. The Parallel Trends Assumption

The parallel trends assumption for the outcome variable is critical for difference-in-differences estimation. Figure 2 illustrates it graphically: we assume that in the absence of the treatment, the treated group would have followed the same trend for the outcome variable as the control group. This is used to fill in the missing potential outcome $E[Y_{i1}(0)|G_i = 1]$, and τ is the difference between this counterfactual and the actual outcome in the treated group.

Figure 2: **Graphical Illustration of Difference-in-Difference Estimation**


$\mathbb{E}[Y_{i1}|G_i = 1]$
$\mathbb{E}[Y_{i1}(0)|G_i = 1]$
$\mathbb{E}[Y_{i0}|G_i = 1]$
$\mathbb{E}[Y_{i1}|G_i = 0]$
$\mathbb{E}[Y_{i0}|G_i = 0]$
τ
$t = 0$
$t = 1$

2.1. Violation: Time-Varying Confounders

Time-constant differences (imbalances) between the treated and control groups are perfectly permissible here. Places with similar initial characteristics might be more likely to follow similar trends over time, but this doesn't have to be the case for difference-in-differences to work. Instead, the parallel trends assumption rules out the possibility of *time-varying confounders.* It rules out the possibility that something changed in the treated group but not the control group between $t = 0$ and $t = 1$ in a way that affected the outcome. In such a case, the control group could not serve as an over-time counterfactual for the treated group. For example, it could be problematic if New Jersey experienced much higher economic growth than Pennsylvania over the period, and economic growth also caused restaurants to hire more workers.

This essentially means that we need to worry about balance in terms of *changing* attributes. Because difference-in-differences is often used for policy evaluation, one particularly common violation of common trends is selection into, or targeting of, policies in ways that are correlated with time-varying confounders:

- **Selection** problems occur when units select into the treatment based on changes in their attributes. For instance, states whose economic growth is rising may be more likely to introduce rises in minimum wages

- **Targeting** problems occur when policymakers target programs toward groups whose position is currently getting worse (or better)

These violations are far less likely if the treatment was applied in a way that appears to have been random. Therefore differences-in-differences estimation is particularly powerful when combined with an actual or natural experiment that led to the treatment being randomly assigned, or something close to it. If there is less confounding in terms of time-constant attributes, there is probably also less confounding in terms of time-variable attributes. Also note that one advantage of the regression setup is that we can include time-varying group-level covariates X that might help bolster our confidence in the parallel trends assumption. It is a good idea to include covariates that that are likely to have changed at the same time as the treatment you are interested in. For example, we could control for state GDP:

$$Y_i = \mu + \gamma G_i + \delta T_i + \tau G_i T_i + \beta GDP$$

2.2. Violation: Attrition or Sampling Differences

A further potential violation arises in 'unbalanced' panel data or repeated cross-sectional data: **differential attrition** or **differences in sampling**. Panel data is sometimes referred to as either 'balanced' or 'unbalanced':

- **Balanced** panel data means that the same units are observed in *both* time periods.

- **Unbalanced** panel data means that some units dropped out over time. Therefore the second period features only a subset of units that began the study.

In unbalanced panel data, differential attrition occurs when the units that drop out have different potential outcomes compared to those who remain, *and* attrition differs between the treatment and control groups. Suppose that in the initial samples taken in period 1, both groups were near-identical on average and would have followed a parallel trend in the absence of treatment had they been followed fully into period 2. It is possible that with differential attrition, the remaining subsamples in period 2 would no longer have followed a parallel trend due to the distinctiveness of those who left the sample in one of the two groups.

If there is attrition, it is always sensible to conduct balance tests for the period 1 and 2 samples in the treated and control group and then compare attrition patterns across the two groups. It is common to conduct analysis using only the balanced subset of units (those who remained in both periods) when attrition is similar across the two groups, although differential attrition based on unobserved characteristics remains possible. Note that this is just like finding the ATE for always-reporters in experiments. It is probably quite rare to encounter situations where attrition behaviour will be the same across treated and control groups.

In the minimum wage example, a few restaurants did not respond to the second wave of the sample (after the change) and some closed down. If these restaurants were distinctive and this attrition occurred differently in each state, then the remaining restaurants might have followed very different trends in the treated and control groups. One possibility is that worse-performing restaurants in New Jersey were more likely to close down after the minimum wage change. That would leave only restaurants with higher numbers of employees in the second period, inflating the treatment effect. In reality though, attrition occurred only on a very small scale in the Card and Krueger study. They also report findings for both the full sample and the balanced sub-sample, and the results are virtually identical.

The issues are very similar in repeated cross-sectional data. Here we might worry that the sampling of units was carried out differently in both periods and across groups, so that the period-1 and period-2 samples would not necessarily have followed parallel trends. Again, balance tests may help detect sampling differences.

2.3. Checking Pre-Treatment Trends

The parallel trends assumption is a strong one, and is often likely to be violated in the real world where policy changes occur alongside many other changes. Although we can't test for it completely (there can always be unobserved time-varying confounders), we can gather circumstantial evidence. One key test is whether or not **the treated and untreated groups were following parallel trends before the treatment occurred.** This is like a balance test. If the units were trending in the same way before the treatment changed, then it is more likely that they would have continued to do so in the absence of any change.

It is standard practice in difference-in-differences estimation to collect data on trends that occurred before the treatment. Unfortunately Card and Krueger did not collect this data; as a result, their paper probably wouldn't be publishable today in a good journal. Figures 3 and 4 show what this data would look like in a good scenario (no violation of parallel trends) and a bad scenario. As we saw above, Pennsylvania saw a slight decline in employment in its restaurants following the policy change (at time=0) whereas New Jersey saw a rise. In the ideal scenario (Figure 3), both would have been following a slight downward trend before the policy change. But in a bad scenario (Figure 4), the two states have non-parallel trends before the change: New Jersey's employment was already growing before the policy change, whereas in Pennsylvania it was declining. In Figure 4, what looks like a positive effect of the treatment in New Jersey between 0 and 1 is nothing more than a continuation of trends that began before the policy change, perhaps due to state differences in trends in demand for fast food, etc.

Figure 3: **Card and Krueger: Parallel Trends Assumption Satisfied**

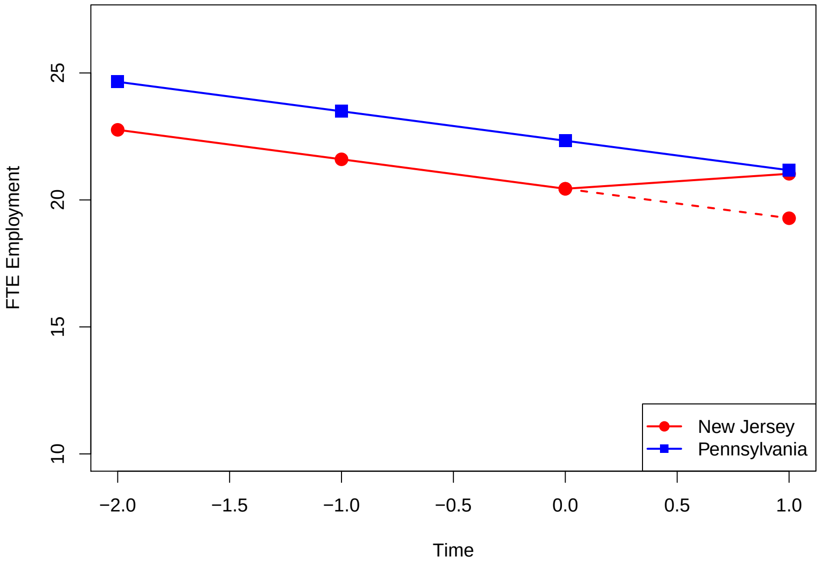
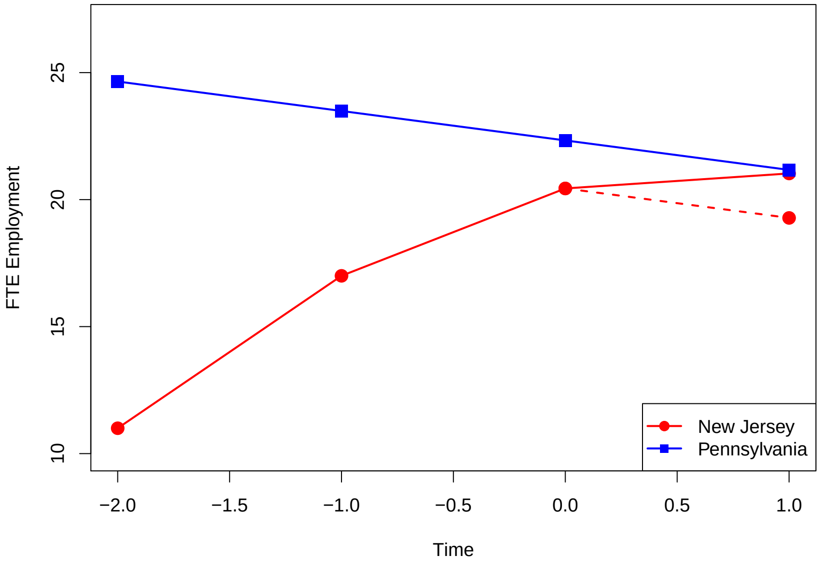


Figure 4: **Card and Krueger: Parallel Trends Assumption Violated**



2.4. Placebo Differences-in-Differences

A more formal test of the parallel trends assumption can be carried out using a placebo test. Recall that this involves testing for the absence of an effect that we expect to be absent. In this case, we can **test for the absence of a difference-in-differences effect in the periods preceding the treatment.** In Figure 3, treatment occurred between periods 0 and 1 and a positive treatment effect is clearly visible over that period as a difference in the trends of the treatment and control group. But if parallel trends is satisfied, we should not expect to find any evidence of such differences between the groups *before* the treatment occurred. So all we need to is carry out the exact same differences-in-differences regression as before, but with T recoded so that the placebo treatment instead occurs between periods -1 and 0 or between -2 and -1. Clearly, in Figure 3 these placebo tests would find no effect. But in Figure 4, these placebo tests would find very clear effects, which provides a formal statistical argument that parallel trends is violated. In this week's problem set, we see an example that falls somewhere between Figures 3 and 4. Trends before the treatment are slightly non-parallel, but the formal placebo test suggests that in reality there was no statistically significant difference between the groups.

3. Fixed Effects Estimation

Using the method of Fixed Effects, Difference-in-differences estimation can be extended to:

- multiple units
- multiple time periods
- a continuous rather than binary treatment variable
- a treatment variable that changes at different rates in different places at different times

This makes the method extremely adaptable, and hence it is very widely used. It can be applied to get causal estimates in any situation where we have multiple observations of the same "groups" over time, *and* the treatment changes by different amounts in different places at each time. We might have panel data on multiple individuals over many years (commonly known as a 'longitudinal study'), or on multiple countries or regions. Suppose, for instance, that we wanted to find the causal effect of minimum wage changes on employment in US states in general. Minimum wages change at different times in different places, and by different amounts. Fixed effects allows us to find the causal effect of a one-unit change in the minimum wage. To do this, we need to account for the fact that:

- Each time period is unique. A minimum wage rise at some time t may just happen to coincide with high employment simply because employment was booming in general at t

- Each state is unique. A minimum wage rise in some state i may just happen to coincide with high employment simply because g is wealthy and has high employment rates.

To account for this, we use a **two-way fixed effects model with T time periods and N groups:**

$$Y_it \ = \ \mu + \sum_{i=2}^{N} \gamma_i G_i + \sum_{t=2}^{T} \delta_t T_t + \tau D_{it} + \beta X_{it}$$

where the G_i are dummy variables equalling 1 if the observation is part of group i, the T_i are dummy variables equalling 1 if the observation is in time period t, D_{it} is the value of the treatment variable in group i at time t and X_{it} are time-varying covariates, if we choose to include them. This equation expresses our prediction for a single group i at time period t. It generalises the previous regression formula for two-period difference-in-differences. Each state i has its own dummy variable (with the first state serving as the omitted reference category) with its own unique coefficient. Likewise, each time period t has its own dummy variable (with the first period serving as the omitted reference category) with its own unique coefficient. Suppose we wanted to predict full time employment in fast-food restaurants in New Jersey in 1993. Then, the equation becomes:

$$Employment_{NJ.1993} \;=\; \hat{\mu} + \hat{\gamma}_{NJ} + \hat{\delta}_{1993} + \hat{\tau} Min.Wage_{NJ.1993} + \hat{\beta} X_{NJ.1993}$$

This controls for all the things that are unique about New Jersey in general via $\hat{\gamma}$ (e.g. New Jersey may always tend to have higher employment in restaurants than other states). It also controls for all the things that are unique about the year 1993 (e.g. employment in restaurants may have been high in general across the USA that year). It therefore isolates the effect of changes in New Jersey's minimum wage. Because of this, the fixed effects method is sometimes referred to as a **within estimator**, because it measures only the impact of changes within each group, as opposed to measuring differences *between* groups or time periods as in random/mixed-effects modeling (often subsumed under the heading "multilevel modeling"). This makes fixed effects estimation ideal for uncovering the causal effect of one-unit changes in treatments, but much less useful for description or prediction, where multilevel approaches are more typically used. In R, two-way fixed-effects is implemented easily via the plm package:

```
library(plm)
model.name <- plm(outcome ~ treatment + covariates, data = ,
 index = c("group.name", "time"), effect = "twoways")
```

The fixed effects estimator relies on the same assumptions as difference-in-differences. The groups need not have the same potential outcomes in terms of fixed attributes. All fixed attributes are controlled for in the group dummy variables. All of the groups need to be following parallel trends, so that τ gives the effect of changes in only the treatment, rather than other confounders that might be trending at the same time. Including time-varying covariates can help bolster this assumption. In addition, it is also possible to include **state-specific time trends** that control for all time-varying factors that change smoothly over time within states. This is the more common approach in contemporary fixed effects papers, and is covered very nicely in Angrist and Pischke. Rather than simply re-producing their material here, I'll leave you to read it. Make sure you understand what they say.

3.1. Standard Errors

A final complication that arises in fixed effects estimation, as in any statistical approach where units share some geographical dependence, is serial correlation in the error term. Fixed effects models can be implemented with a standard regression, including with R's `lm()` function. But the usual calculations of standard errors embedded in these functions assume that every unit in our analysis represents a unique, independent observation. Another way of putting this is that it assumes that all units ended up in our study via a random sampling process.

With fixed effects models, we cannot assume this, because the error terms within any geographical area (which we'll call a **cluster**) are clearly correlated. If New Jersey has an unexpectedly high level of employment in 1993, it's also likely to have an unexpectedly high level of employment in 1994 as well. We cannot treat our sample as if every datapoint provides a unique piece of information. Intuitively, the usual standard errors over-state how precise our analysis really is. To correct for this, we need to use different, **cluster-robust standard errors**, which re-weight the regression errors to account for correlation within clusters. While this may seem like an academic issue, it is often very important in practice. Heteroskedasticity-robust standard errors are rarely very different from the usual standard errors, but it is by no means uncommon for cluster-robust standard errors to be ten times larger, entirely changing the statistical significance of our results.

In R, cluster-robust standard errors can be calculated via two different routes. A general approach, that works for almost all statistical models that involve some type of regression, uses two different packages:

```
library(lmtest)
library(multiwayvcov)
coeftest(model.name, cluster.vcov(model.name, dataset$group.name))
```

where `model.name` is the name you have given to your model (e.g. a linear regression). In the `plm` package, cluster-robust standard errors for two-way fixed effects models can easily be calculated:

```
library(plm)
coeftest(model.name, vcov=vcovHC(model, cluster="group.name", type="HC1"))
```

Two caveats are important here:

1. These robust methods assume that we have a large sample of clusters. When the number of clusters falls below about 20, they will give the wrong answer. In those instances, bootstrapping should be used instead, which can be done in R with the `clusterSEs` package

2. It is important to carefully define what a 'cluster' is. **A cluster is the unit at which the treatment is applied**. For example, minimum wages are applied by state. In cluster-robust standard errors in two-way fixed-effects models, the clusters are therefore almost always the same as the groups. But in difference-in-differences with two periods and two groups (treatment and control), the treatment and control groups might be made up of various clusters where the treatment is actually applied. For instance, in the Problem Set this week, the dataset consists of units nested within 65 districts (clusters), where the treatment was applied. The standard errors still need to be clustered by district, even though ultimately there are only two groups, treatment and control. The bottom line is to always think about how the treatment is applied in a given application.