# 6. Fixed Effects, Difference-in-Difference, and Panel Data

- When we have longitudinal data we can potentially tackle OVB when the unobserved omitted factors are *stable over time*
- Setting:
  - We can measure the outcome variable for a set of objects (people, firms, ...) at several point in time
  - The key variable of interest (the „treatment“) changes over time
  - We study the association between the change in the treatment variable and the change in the outcome variable
- Two most important approaches
  - *Fixed Effects* (when we have panel data, that is we observe the same objects repeatedly)
  - *Difference-in-Difference* estimation (when we have repeated cross-sections and the treatment varies at an aggregate level)

# 6.1 Fixed Effects

- Consider again the potential outcome framework (time index $t = 1, \dots, T$)

$$Y_{C_{it}it} = \begin{cases} Y_{1it} & if \quad C_{it} = 1 \\ Y_{0it} & if \quad C_{it} = 0 \end{cases}$$

- Assume now that

$$E[Y_{0it}|A_i, X_{it}, t, C_{it}] = E[Y_{0it}|A_i, X_{it}, t]$$

  where
  - $X_{it}$ is a vector of observed (time varying) covariates and
  - $A_i$ is is a vector of *unobservable* factors that are fixed over time (no time index $t$! For instance, a person‘s ability or personality)
- The assumption states that $C_{it}$ is as good as randomly assigned conditional on $A_i$ and $X_{it}$
- This is a sensible identifying assumption whenever any unobserved determinants of the treatment (that also may affect the outcomes beyond the treatment) are time constant

- Consider now the following linear model

$$E[Y_{0it}|A_i, X_{it}, t] = \alpha + X_{it}'\beta + A_i'\gamma + \lambda_t$$

- And assume that the causal effect is a constant $\rho$

$$E[Y_{1it}|A_i, X_{it}, t] - E[Y_{0it}|A_i, X_{it}, t] = \rho$$

- Hence, we can write

$$Y_{it} = \alpha_i + \lambda_t + \rho C_{it} + X_{it}'\beta + \epsilon_{it}$$

where $\epsilon_{it} = Y_{0it} - E[Y_{0it}|A_i, X_{it}, t]$ and $\alpha_i = \alpha + A_i'\gamma$

- When we impose these assumptions, running a regression will estimate the causal effect $\rho$ of $C$ on $Y$

- This is a fixed effects model:
  - The $\alpha_i$ are parameters to be estimated (estimating a dummy for every person)
  - The $\gamma_i$ are time effects that are also estimated (estimating a dummy for every period)

**Study**

## Lazear's (2000) study on Performance Pay at Safelite

- Safelite is a large auto glass company in the US
- Business: replace broken windshields.
- New compensation scheme in January 1994: Piece rate scheme (PPP) replaced hourly-wage scheme in 1994
- The piece rate scheme was phased in over 19 months, starting from the headquarter town.
- The gradual implementation of piece rate allows for within-worker variation identifying the incentive effect of piece rate on effort.
- But: also high turnover rates; many workers also hired after the introduction of the PPP
- In the following:
  - Unit of observation = Worker in a given month;
  - Productivity measure: Average windshields installed by the worker on a given day.

# Safelite: Regression analysis

TABLE 3—REGRESSION RESULTS

| Regression number | Dummy for PPP person-month observation | Tenure | Time since PPP | New regime | $R^2$ | Description |
|---|---|---|---|---|---|---|
| 1 | 0.368<br>(0.013) | | | | 0.04 | Dummies for month and year included |
| 2 | 0.197<br>(0.009) | | | | 0.73 | Dummies for month and year; worker-specific dummies included (2,755 individual workers) |
| 3 | 0.313<br>(0.014) | 0.343<br>(0.017) | 0.107<br>(0.024) | | 0.05 | Dummies for month and year included |
| 4 | 0.202<br>(0.009) | 0.224<br>(0.058) | 0.273<br>(0.018) | | 0.76 | Dummies for month and year; worker-specific dummies included (2,755 individual workers) |
| 5 | 0.309<br>(0.014) | 0.424<br>(0.019) | 0.130<br>(0.024) | 0.243<br>(0.025) | 0.06 | Dummies for month and year included |

*Notes:* Standard errors are reported in parentheses below the coefficients.
Dependent variable: ln output-per-worker-per-day.
Number of observations: 29,837.

**Safelite (continued): What do the worker fixed effects do here?**

- Regression without worker fixed effects (row 1)
  - this gives us an estimate of the causal effect of the treatment on the *average performance* of all workers working at a given point in time
  - (when believe that the treatment is as good as randomly assigned conditional on the time period which seems very plausible here)

- However: if we are interested in the causal effect of the treatment on the performance of an *average given worker* this is a „biased“ estimate
  - This is the case when the ability of workers depends on the treatment
  - For instance, when the PPP allows to hire better workers
  - Then $E[Y_{0it}|t, C_{it}=1] > E[Y_{0it}|t, C_{it}=0]$

    i.e. workers hired under the PPP would be better even without the PPP
  - In this respect the conditional independence assumption is violated
  - There is a classical selection bias and the PPP dummy should give a too high estimate for the causal effect of the PPP on a *given* worker

## What do the worker fixed effects do here?

- The worker fixed effects model (row 2) takes this problem into account
- It imposes the weaker assumption that

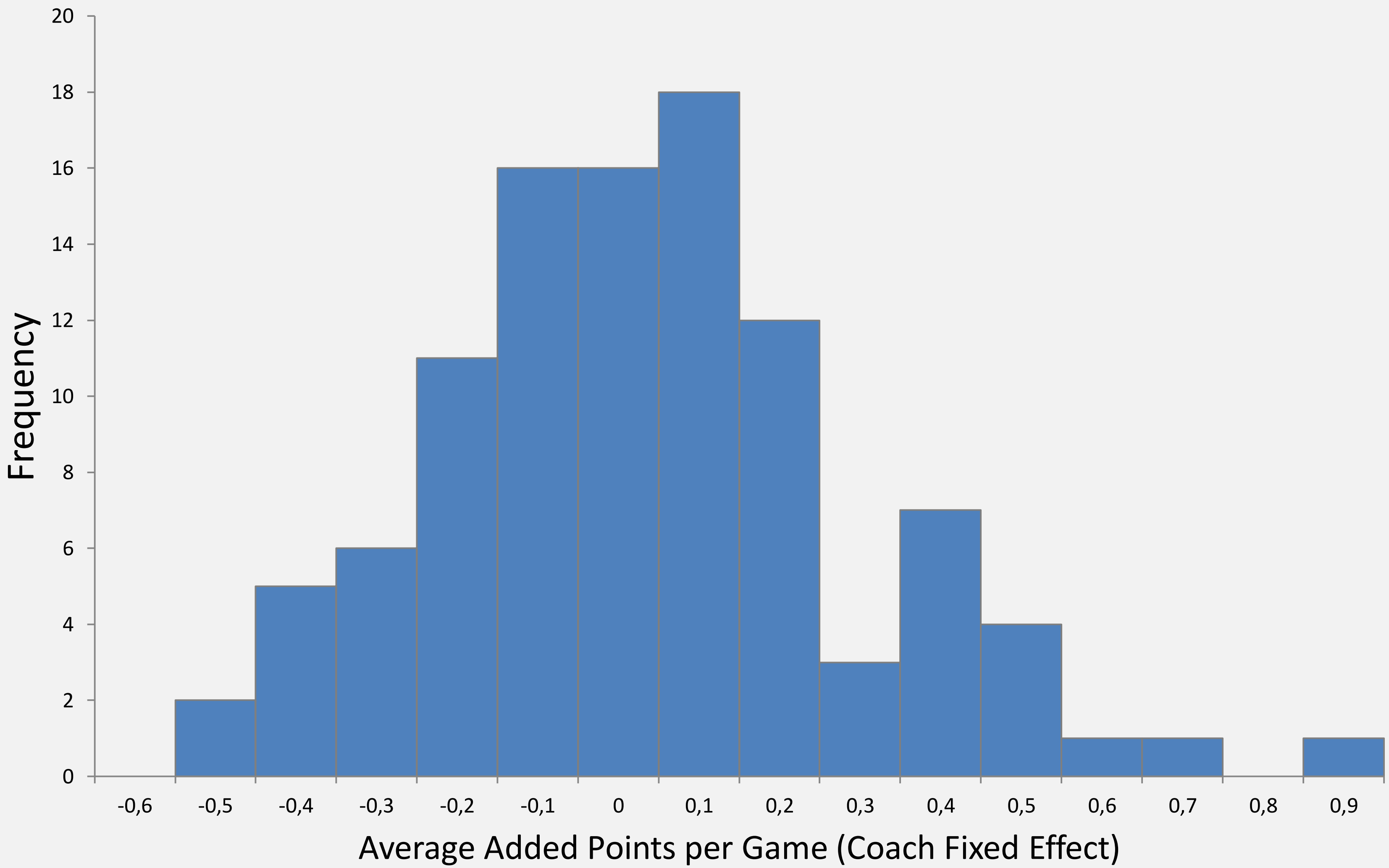
$$E[Y_{0it}|A_i, t, C_{it}] = E[Y_{0it}|A_i, t]$$

- When $A_i$ captures the workers unobserved ability this assumption states that for workers *of the same ability* the counterfactual performance is independent of the treatment
- The fixed effects model in a sense estimates the unoberved abilities of the workers (using that a worker‘s performance is observed over many months)
- It thus estimates the causal effect of the PPP conditional on worker‘s abilities
- Note: The model without fixed effects is here not wrong, it estimates something different
  - Without worker fixed effects it estimates the total effect on performance which includes a *selection* and an *incentive effect*
  - With worker fixed effects it estimates the pure incentive effect

**Study** **Estimating the Value Added of Managers**

- How much do individual managers matter for firm behavior and performance?
- Bertrand and Schoar (2003) use
  - Manager-firm matched panel dataset, where they can track individual top managers across different firms over time.
  - Estimate how much of the unexplained variation in firm practices can be attributed to manager fixed effects (controlling for firm fixed effects and time-varying firm characteristics)
- Mühlheusser/Schneemann/Sliwka/Wallmeier (2017) consider the case of soccer coaches
  - Coaches frequently move between teams
  - Allows to disentangle the effect of the coach from the strength of the club by estimating models with manager and team fixed effects
  - Data on 20 seasons of the German Bundesliga

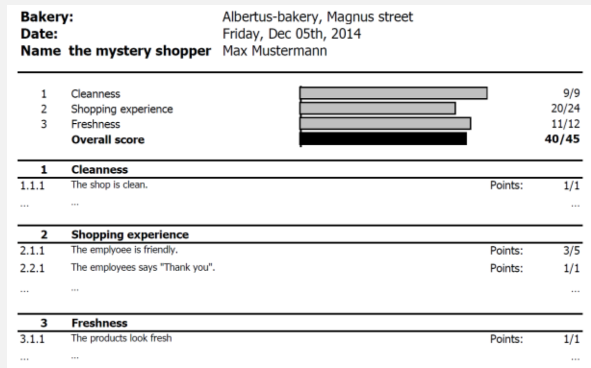
# Estimated Soccer Coach Fixed Effects for the Bundesliga



**Study**

## Mystery shopping in a bakery chain (Friebel et al, mimeo)

- Each month, each shop of a bakery chain is visited by a mystery shopper.
- The mystery shopper buys some bread and evaluates the performance of the employees in a bakery ("mystery shopping score") using the following sheet:



**Bakery:** Albertus-bakery, Magnus street
**Date:** Friday, Dec 05th, 2014
**Name the mystery shopper** Max Mustermann

| | | |
|---|---|---|
| 1 | Cleanness | 9/9 |
| 2 | Shopping experience | 20/24 |
| 3 | Freshness | 11/12 |
| | **Overall score** | **40/45** |

| | | | |
|---|---|---|---|
| **1** | **Cleanness** | | |
| 1.1.1 | The shop is clean. | Points: | 1/1 |
| ... | ... | | ... |
| **2** | **Shopping experience** | | |
| 2.1.1 | The emplyoee is friendly. | Points: | 3/5 |
| 2.2.1 | The employees says "Thank you". | Points: | 1/1 |
| ... | ... | | ... |
| **3** | **Freshness** | | |
| 3.1.1 | The products look fresh | Points: | 1/1 |
| ... | ... | | ... |

- Goal of the mystery shopping: Measure the effort of employees. Shop-managers receive a bonus based on the mystery shopping scores
- But Friebel et al find:
  - Regressing shop performance on mystery shopping scores shows that they have no predictive power
  - But the $R^2$ increases substantially (from 2% to 23%) when including mystery-shopper-fixed effects in the regressions.

## Estimating Fixed Effects Models

- Estimating the coefficients of individual dummy variables seems demanding in large panels (1000 employees = 1000 fixed effects)
- However, if we are not interested in knowing the specific values of the individual fixed effects, we can estimate the model in a simpler manner
- Consider

$$Y_{it} = \alpha_i + \lambda_t + \rho C_{it} + X_{it}'\beta + \epsilon_{it}$$

- Now take the average across all time periods $\overline{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \rho\overline{C}_i + \overline{X}_i'\beta + \overline{\epsilon}_i$$

and substract this from $Y_{it}$

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(C_{it} - \overline{C}_i\right) + (X_{it}' - \overline{X}_i')\beta + \epsilon_{it} - \overline{\epsilon}_i$$

→ The $\alpha_i$ are eliminated!

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(C_{it} - \overline{C}_i\right) + (X'_{it} - \overline{X}'_i)\beta + \epsilon_{it} - \overline{\epsilon}_i$$

- Hence,
  - replace the outcome variable by its deviation from the mean over time
  - replace the explanatory variables by their deviations from their means over time
  - Regress the „de-meaned“ outcome on the „de-meaned“ explanatory variables

  → This gives us an estimate of $\rho$

  → We can estimate $\rho$ and $\beta$ without having to estimate the $\alpha_i$

- This model is sometimes also called the *within-estimator:*

  It estimates the effect of $\rho$ on $Y$ from the within person variation in $C$

Python

# Fixed Effects Regressions in Python

- Panel regressions in Python can be done with library linearmodels
- Install by pip install linearmodels
- Import by from linearmodels import PanelOLS
- In order to run a panel regression use a MultiIndex DataFrame that is a DataFrame that uses two indices
  - one index for the entity variable (the omitted time constant variable)
  - one index for the time variable

```
df=df.set_index([‘entity', 'year’])
```

- Then fit the model by

```
reg = PanelOLS.from_formula('y ~ x + EntityEffects + TimeEffects', data=df).fit()
```

- Then print the output with print(reg)
  (Note the different notation to statsmodels: can directly print the results)

## Your Task **Fixed Effects**

- Open the notebook in which you estimated the association between Management Practices and ROCE
- For a part of the observaations the data set contains panel data
- The paper by Bloom et al. (2012) contains the following table, where the third colums shows the result of a fixed effects regression
- Please replicate this regression using PanelOLS
- Note:
  - The variable account_id contains an identifier for each firm
  - The variable emp contains the number of employees and ppent the capital (fixed assets)
  - You can generate logs by using np.log(x) directly in the formula

| | (1) | (2) | (3) |
|---|---|---|---|
| Sector | Manufact. | Manufact. | Manufact. |
| Dependent variable | Log (Sales) | Log (Sales) | Log (Sales) |
| Management | 0.523*** (0.030) | 0.233*** (0.024) | 0.048** (0.022) |
| Ln(Employees) | 0.915*** (0.019) | 0.659*** (0.026) | 0.364*** (0.109) |
| Ln(Capital) | | 0.289*** (0.020) | 0.244*** (0.087) |
| Country controls | No | Yes | NA |
| Industry controls | No | Yes | NA |
| General controls | No | Yes | NA |
| Firm fixed effects | No | No | Yes |
| Organizations | 2,927 | 2,927 | 1,453 |
| Observations | 7,094 | 7,094 | 5,561 |

## Your Task — Fixed Effects (Simulated Sales Training Evaluation VII)

Generate the following notebook

```
n=2000
df1=pd.DataFrame(index=range(n))
df1['ability']=np.random.normal(100,15,n)
df1['year']=1
df1['persnr']=df1.index
df1['training']=0
## Now copy the DataFrame (i.e. generate observations for second year)
df2=df1.copy()
df2['year']=2
## Training only in year 2:
df2['training']=((df2['ability']+np.random.normal(0,10,n)>=100))
## Generate DataFrame that spans both years by appending the two data frames
df=pd.concat([df1,df2], sort=False)
df['sales']= 10000 + df.training*5000 + df.ability*100 + df.year*2000
                + np.random.normal(0,4000,2*n)
```

**Note:**

- The script generated a data frame simulating two years of data in which
  - Sales of each subject are observed in each year
  - training is affected by ability
  - subjects are only trained in year 2

**Now analyze the generated data:**

- Run an OLS regression of sales on training and year
- Define the time and entity indices
- Run a fixed effects regression

**But note three important caveats:**

1. When the treatment $C_{it}$ hardly varies over time it is hard to evaluate the causal effect effect $\rho$
   - In the extreme when $C_{it}$ is completely stable then $C_{it} = \overline{C}_{it}$
   - Not identifying a significant effect in the data then does not necessarily imply that there is no such effect
2. When on top of that there is measurement error in $C_{it}$ this can generate most of the variation in $C_{it}$ over time
   - this will lead to attenuation bias
   - thus there can be a tendency to underestimate the true effect
3. Fixed effects can only eliminate *time constant* omitted variables
   - If the treatment is correlated with time varying unobserved variables omitted variable issues remain

# 6.2 Difference-in-Difference Estimation

- Sometimes the regressor of interest varies only at a more aggregate level (say a state, or a firm) which we index by $s$
- Moreover, sometimes we do not observe the same individuals repeatedly but have different samples at different points in time $t$
- We can then use a so-called difference-in-difference estimation strategy
- This is like a fixed effects model with a fixed effect at a more aggregate level
- The underlying idea is again that there is an additive structure in the potential outcomes that is

$$E[Y_{0ist}|s,t] = \gamma_s + \lambda_t$$

- Assume that the causal effect of the treatment is a constant

$$E[Y_{1ist} - Y_{0ist}|s,t] = \delta$$

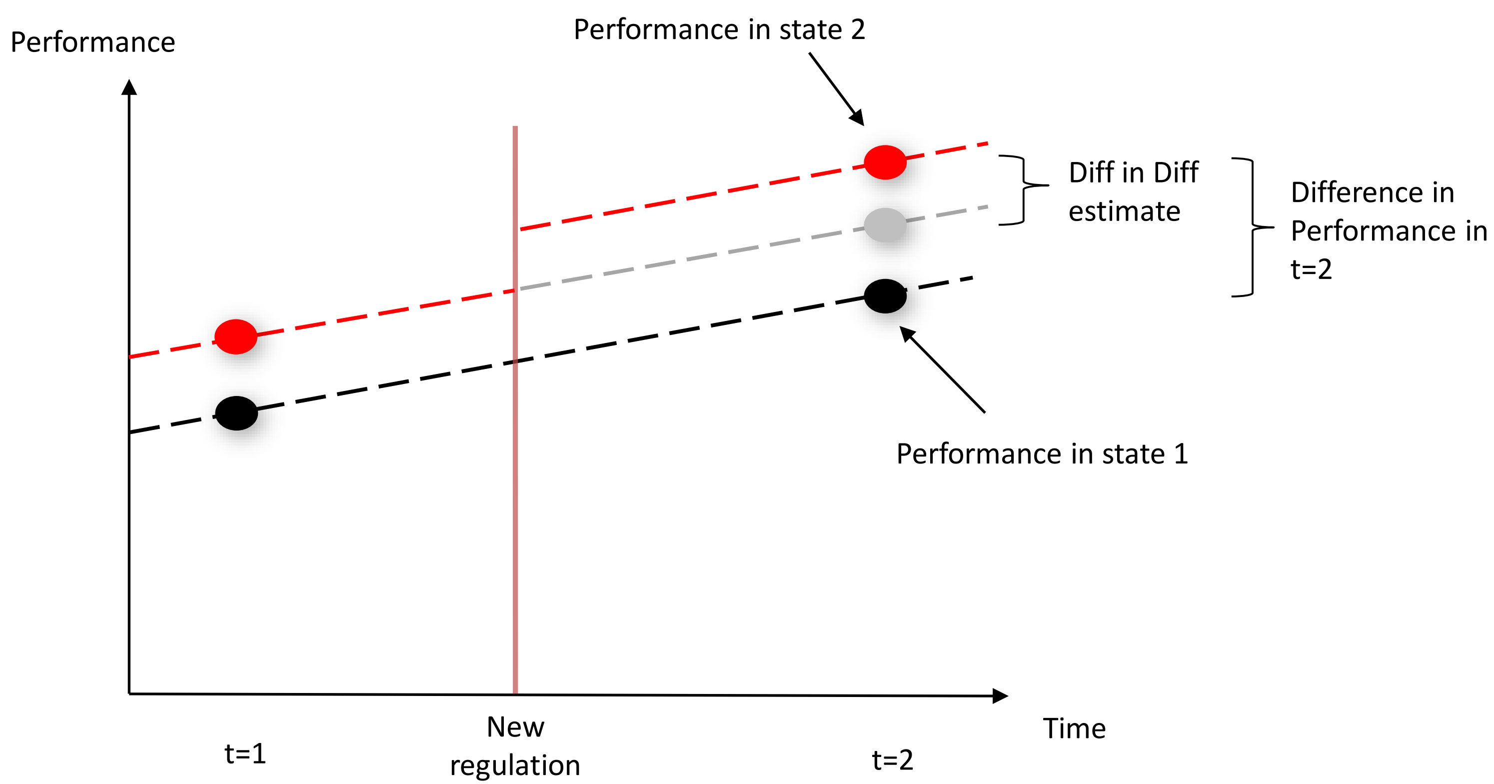
- Then

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

where $E[\epsilon_{ist}|s,t] = 0$

## Diff-in-Diff: Graphical illustration

- Suppose that there are two states (regions, firms, departments,…)
- A new regulation is introduced in state 2 but not in state 1
- We want to study the effect on firm performance



**Consider**

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

- Say we have two periods $t = 1,2$ and two states $s = 1,2$
- A new policy is adopted in state 2 in period 2 such that
  - $C_{11} = C_{21} = C_{21} = 0$ and $C_{22} = 1$
- Then
  - $E[Y_{ist}|s = 1, t = 2] - E[Y_{ist}|s = 1, t = 1] = \lambda_2 - \lambda_1$ and
  - $E[Y_{ist}|s = 2, t = 2] - E[Y_{ist}|s = 2, t = 1] = \lambda_2 - \lambda_1 + \delta$
- The difference-in-difference is

$$(E[Y_{ist}|s = 2, t = 2] - E[Y_{ist}|s = 2, t = 1])$$
$$- \quad (E[Y_{ist}|s = 1, t = 2] - E[Y_{ist}|s = 1, t = 1]) \quad = \quad \delta$$

which is the causal effect of interest

## Regression Diff-in-Diff

- Note: we could estimated the causal effect $\delta$ from just working with the differences and replace the expectations with the respective averages
- Typically it is more convenient to simply run a regression
  - Let $TREAT_i$ be a dummy indicating whether an observation comes from the treated region
  - Let $POST_t$ be a dummy indicating whether an observation comes from a period after the treatment has been implemented
- Then we can regress

$$Y_{it} = \alpha + \beta \cdot TREAT_i + \gamma \cdot POST_t + \delta \cdot (TREAT_i \cdot POST_t) + \epsilon_{it}$$

- The coefficient $\delta$ of the interaction term $TREAT_i \cdot POST_t$ yields an estimate of the causal effect
- Note:
  - Regression DiD also provides statistical tests
  - And it can be applied if there are more than two periods

**Study**

## Dismissal Protection and Hiring

- Bauernschuster (2013) studies the effect of a firm-size threshold of the German dismissal protection law in 2004 on the hiring behavior of small firms
- From 1999 until the end of 2003 the dismissal protection law applied only to establishments with more than five (full time equivalent) workers
- In 2004, this threshold was shifted up to ten full-time equivalent workers
- Dismissal protection regulation was abandoned for workers hired after December 31, 2003 by establishments with 6 to 9 FTE employees
- Bauernschuster studies the effect on hiring applying a diff-in-diff strategy
  - „Treatment group“: establishments with 6-10 employees
  - „Control group“: establishments with 11-20 employees
- Uses data from the IAB establishment panel (on which the LPP is based)
- Dependent variables are hiring rates and total number of hirings per establishment in the first half of the year

# Dismissal Protection and Hiring (continued)

## The Dynamics of Hiring

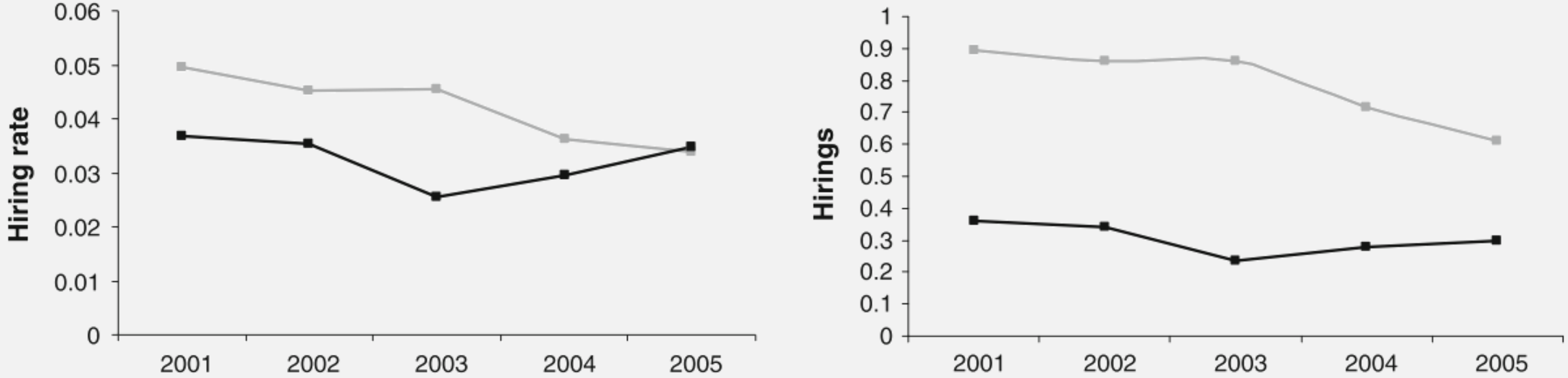



**Fig. 2** Dynamics of hirings in treatment and control group. The *left figure* shows average hiring rates for treatment (*black line*) and control groups (*grey line*) over time. The *right figure* shows average absolute hirings for treatment (*black line*) and control groups (*grey line*) over time. The treatment group comprises establishments with more than five and up to ten full-time equivalent workers while the control group consists of establishments with more than ten and up to 20 full-time equivalent employees. Data source: IAB Establishment Panel

Bauernschuster (2013)

# Dismissal Protection and Hiring (continued)

**Table 1** DiD estimates on hirings

| Parameter | Hiring rate (1) | | Hiring rate (2) | | Hiring rate (3) | |
|---|---|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error |
| DiD 2004 | 0.013** | 0.007 | 0.016** | 0.007 | 0.020** | 0.009 |
| DiD 2005 | 0.021** | 0.009 | 0.021** | 0.009 | 0.020** | 0.009 |
| Treatment group | −0.020*** | 0.006 | −0.019*** | 0.007 | −0.024*** | 0.006 |
| Year 2004 | −0.009* | 0.004 | −0.013*** | 0.004 | −0.014*** | 0.006 |
| Year 2005 | −0.012** | 0.005 | −0.013*** | 0.005 | −0.013* | 0.006 |
| Control set 1 | No | | Yes | | Yes | |
| Control set 2 | No | | No | | Yes | |
| *N* | 1,749 | | 1,658 | | 1,285 | |
| $R^2$ | 0.0059 | | 0.0725 | | 0.1197 | |

The table reports the results of OLS difference-in-differences regressions with hiring rates as the dependent variable. The treatment group comprises establishments with more than five and up to ten full-time equivalent workers while the control group consists of establishments with more than ten and up to 20 full-time equivalent employees. The baseline year is 2003. Specification (1) includes no further controls. In specification (2), we additionally control for capital stock, works council, collective labor agreement, age, and industry (control set 1). In specification (3), we add the ratio of female workers, ratio of unqualified workers, ratio of apprentices, wage per worker in the previous year, value added per worker in the previous year as well as net hirings in the previous year as further controls (control set 2). Standard errors are clustered at the establishment level. ***, **, * denote significance at the 1, 5, and 10% levels, respectively. Data source: IAB Establishment Panel

Bauernschuster (2013)

## The Common Trend Assumption

- Crucial for both a Diff-in-Diff and a Fixed Effects estimation strategy:
  Treatment and control group follow the same underlying time trend!
- If this is violated, then both approaches yield biased estimates
- It is called the *common trend assumption*
- Again this is an identifying assumption: We can claim that we identify a causal effect when the common trend assumption holds
- It can be very useful to check the trends in several periods
  - before the change occurs
  - after the change occurs
- If trends differ already before the intervention both strategies are problematic to identify causal effects