# Diff-in-Diff and Fixed Effects

Annie Chen

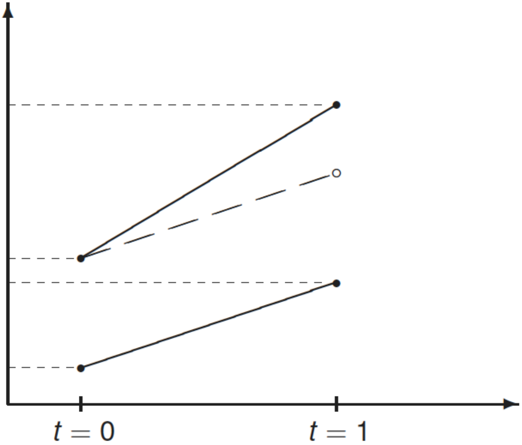
February 12, 2020

# Pset 5

- `R` code for plotting effect of observed covariates

# Review of D-i-D

Fill in the blanks:



# Review of D-i-D

- What is the causal estimand in a D-i-D?

# Review of D-i-D

- What is the causal estimand in a D-i-D?
  - The ATT in the post-treatment period: $\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|D_i = 1]$

# D-i-D Estimation with Regression

$$Y_{it} = \alpha + \beta_1 Treated_i + \beta_2 Period_t + \beta_3 (Treated_i \times Period_t) + u_{it}$$

- $\beta_1$ captures pre-existing differences between treated and control while $\beta_2$ is the change over time common to both groups.

# D-i-D Estimation with Regression

$$Y_{it} = \alpha + \beta_1 Treated_i + \beta_2 Period_t + \beta_3 (Treated_i \times Period_t) + u_{it}$$

- $\beta_1$ captures pre-existing differences between treated and control while $\beta_2$ is the change over time common to both groups.
- Same as two-way fixed effects? What happens in D-i-D model if I include both unit and time FE?

$$Y_{it} = \alpha + \delta_i + \tau_t + \beta_3 (Treated_i \times Period_t) + u_{it}$$

# Voter Graditude Example (Bectel and Hainmueller. 2011)

- Unit of analysis is electoral districts.

# Voter Graditude Example (Bectel and Hainmueller. 2011)

- Unit of analysis is electoral districts.
- Treatment is `Flooded`, a binary variable indicating whether an electoral district was directly affected by the 2002 flood.[1]

[1] Federal elections are set exogenously by the German Constitution.

# Voter Graditude Example (Bectel and Hainmueller. 2011)

- Unit of analysis is electoral districts.
- Treatment is `Flooded`, a binary variable indicating whether an electoral district was directly affected by the 2002 flood.[1]
- The dependent variable is the SPD PR vote share for a given electoral district.[2]

[1] Federal elections are set exogenously by the German Constitution.
[2] Models 4, 7, and 10 are the first-differenced DID regressions with the DV as the change in vote share between two election periods.

- Interested in $\mathbb{E}[Y_{i1,2002}|D_i = 1] - \mathbb{E}[Y_{i0,2002}|D_i = 1]$

- Interested in $\mathbb{E}[Y_{i1,2002}|D_i = 1] - \mathbb{E}[Y_{i0,2002}|D_i = 1]$
- Assume $E[Y_{0i,2002} - Y_{0i,1998}|D_i = 1] = E[Y_{0i,2002} - Y_{0i,1998}|D_i = 0]$

- Interested in $\mathbb{E}[Y_{i1,2002}|D_i = 1] - \mathbb{E}[Y_{i0,2002}|D_i = 1]$
- Assume $E[Y_{0i,2002} - Y_{0i,1998}|D_i = 1] = E[Y_{0i,2002} - Y_{0i,1998}|D_i = 0]$
  - Read: "In the absence of the flood, the average SPD vote share in the affected districts would have followed a similar trend as the average SPD vote share in unaffected districts."

- Interested in $\mathbb{E}[Y_{i1,2002}|D_i = 1] - \mathbb{E}[Y_{i0,2002}|D_i = 1]$
- Assume $E[Y_{0i,2002} - Y_{0i,1998}|D_i = 1] = E[Y_{0i,2002} - Y_{0i,1998}|D_i = 0]$
  - Read: "In the absence of the flood, the average SPD vote share in the affected districts would have followed a similar trend as the average SPD vote share in unaffected districts."
- $\hat{Y}_{it} = \hat{\nu}_i + \hat{\delta}_t + \hat{\beta} D_{it} + X_{it}^T \hat{\alpha} + \hat{\epsilon}_{it}$

- Interested in $\mathbb{E}[Y_{i1,2002}|D_i = 1] - \mathbb{E}[Y_{i0,2002}|D_i = 1]$
- Assume $E[Y_{0i,2002} - Y_{0i,1998}|D_i = 1] = E[Y_{0i,2002} - Y_{0i,1998}|D_i = 0]$
  - Read: "In the absence of the flood, the average SPD vote share in the affected districts would have followed a similar trend as the average SPD vote share in unaffected districts."
- $\hat{Y}_{it} = \hat{\nu}_i + \hat{\delta}_t + \hat{\beta} D_{it} + X_{it}^T \hat{\alpha} + \hat{\epsilon}_{it}$
  - where $\hat{\nu}_i$ and $\hat{\delta}_t$ are district and time fixed effects respectively.

- `Flooded` and `PostPeriod` are equivalent to the interaction between Treatment and Period (Treatment = 1 are flooded districts and Period = 1 is post-treatment)

```
all_data %>%
  filter(data_period == "1998-2002") %>%
  dplyr::select(c("wkr", "year", "spd_z_vs",
                  "Flooded", "PostPeriod")) %>%
  head()
```

```
## # A tibble: 6 x 5
##     wkr  year spd_z_vs Flooded PostPeriod
##   <dbl> <dbl>    <dbl>   <dbl>      <dbl>
## 1     1  1998     47.2       0          0
## 2     1  2002     44.0       0          1
## 3     2  1998     43.5       0          0
## 4     2  2002     42.4       0          1
## 5     3  1998     44.6       0          0
## 6     3  2002     41.2       0          1
```

# First-Difference Estimator

$\Delta Y_{it} = \beta \Delta D_{it} + \Delta u_{it}$ (first, without covariates)

```
library(plm)
fd_mod <- all_data %>%
  filter(data_period == "1998-2002") %>%
  plm(data = ., index = c("wkr"),
      model = "fd", # first difference estimator
      formula = spd_z_vs ~ Flooded * PostPeriod)
```

# First-Difference Estimator

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = spd_z_vs ~ Flooded * PostPeriod, data = ., model = "fd",
##     index = c("wkr"))
##
## Balanced Panel: n = 299, T = 2, N = 598
## Observations used in estimation: 299
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -11.97010  -1.50274  -0.12258   1.24436  11.77026
##
## Coefficients: (1 dropped because of singularities)
##             Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) -2.88037    0.22058 -13.058 < 2.2e-16 ***
## Flooded      7.14401    0.70827  10.087 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5238.1
## Residual Sum of Squares: 3901.6
## R-Squared:      0.25515
## Adj. R-Squared: 0.25265
## F-statistic: 101.74 on 1 and 297 DF, p-value: < 2.22e-16
```

# Correcting Standard Errors

```r
# this does not support twoway clustering...
coeftest(fd_mod, vcov. = vcovHC(fd_mod, type = "HC2",
                                cluster = "group"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.88037    0.22758 -12.656 < 2.2e-16 ***
## Flooded      7.14401    0.47313  15.099 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

# Fixed Effects via Demeaning

$\tilde{Y}_i = \beta \tilde{D}_i + \tilde{u}_i$

```
all_data %>%
  filter(data_period == "1998-2002") %>%
  plm(data = ., index = c("wkr", "PostPeriod"),
      model = "within", # within estimator
      formula = spd_z_vs ~ Flooded * PostPeriod) %>%
  coeftest(., vcov. = vcovHC(., type="HC2"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## Flooded       7.14401    0.46983  15.206 < 2.2e-16 ***
## PostPeriod1 -2.88037    0.22737 -12.668 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Including covariates

- Exercise: Carefully select covariates ($\mathbf{X}_{it}$) and reestimate the new model including these covariates: $\hat{Y}_{it} = \hat{\nu}_i + \hat{\delta}_t + \hat{\beta} D_{it} + X_{it}^T \hat{\alpha} + \hat{\epsilon}_{it}$.

```
all_data %>%
  dplyr::select(starts_with("xx")) %>%
  colnames()
```

```
##  [1] "xxsinc_SPD"         "xxpopdensity_"      "xxshpop_o60_'
##  [4] "xxpopnetinp1000_"   "xxue_"              "xxshagric_"
##  [7] "xxshmanu_"          "xxshtradservice_"   "xxshotherserv
## [10] "xxshforeign_"
```

# Including covariates

- Exercise: Carefully select covariates ($\mathbf{X}_{it}$) and reestimate the new model including these covariates: $\hat{Y}_{it} = \hat{\nu}_i + \hat{\delta}_t + \hat{\beta} D_{it} + X_{it}^T \hat{\alpha} + \hat{\epsilon}_{it}$.
- Assume that these are all _ _ _ - _ _ _ _ _ _ _ _ _ covariates.

```
all_data %>%
  dplyr::select(starts_with("xx")) %>%
  colnames()
```

```
##  [1] "xxsinc_SPD"          "xxpopdensity_"      "xxshpop_o60_'
##  [4] "xxpopnetinp1000_"  "xxue_"                "xxshagric_"
##  [7] "xxshmanu_"           "xxshtradservice_"  "xxshotherserv
## [10] "xxshforeign_"
```

# Including covariates

```
form_covars <- spd_z_vs ~ Flooded * PostPeriod +
                  # controls
                  xxpopdensity_ + # share pop density
                  xxshpop_o60_ + # share pop over 60
                  xxpopnetinp1000_ + # share pop outflow
                  xxue_ + # unemployment
                  xxshagric_ + # employment share (agriculture)
                  xxshmanu_ + # employment share (manufacturing)
                  xxshtradservice_ + # employment share (trade)
                  xxshotherservice_ + # employment share (other ser
                  xxshforeign_ +# share of foreigners
                  xxsinc_SPD # SPD incumbent in Land
```

```r
full_mod <- all_data %>%
  filter(data_period == "1998-2002") %>%
  plm(data = ., index = c("wkr", "PostPeriod"),
      model = "within", form_covars)
```

| | Model 1 |
|---|---|
| Flooded | 6.91*** |
| | (0.76) |
| PostPeriod1 | −3.98*** |
| | (0.69) |
| xxpopdensity_ | −0.06 |
| | (1.43) |
| xxshpop_o60_ | 0.40 |
| | (0.22) |
| xxpopnetinp1000_ | −0.04 |
| | (0.03) |
| xxue_ | −0.13 |
| | (0.16) |
| xxshagric_ | −1.58 |
| | (4.05) |
| xxshmanu_ | −1.20 |
| | (4.03) |
| xxshtradservice_ | −1.31 |
| | (4.02) |
| xxshotherservice_ | −1.12 |
| | (4.02) |
| xxshforeign_ | 20.09 |
| | (21.33) |
| xxsinc_SPD | −1.12 |
| | (0.63) |
| $R^2$ | 0.44 |

# Parallel Trends Assumption

- $\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0]$

```
all_data %>%
  filter(data_period == "1994-1998") %>%
  plm(data = ., index = c("wkr"),
    model = "fd",
    formula = spd_z_vs ~ Flooded * PostPeriod)
```

```
##
## Model Formula: spd_z_vs ~ Flooded * PostPeriod
## <environment: 0x7fed44820078>
##
## Coefficients:
## (Intercept)     Flooded
##  4.61036210 -0.00040711
```

# Parallel Trends Assumption

- $\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0]$
- Check leads: 1994 - 1998

```
all_data %>%
  filter(data_period == "1994-1998") %>%
  plm(data = ., index = c("wkr"),
    model = "fd",
    formula = spd_z_vs ~ Flooded * PostPeriod)
```

```
##
## Model Formula: spd_z_vs ~ Flooded * PostPeriod
## <environment: 0x7fed44820078>
##
## Coefficients:
## (Intercept)     Flooded
##  4.61036210 -0.00040711
```

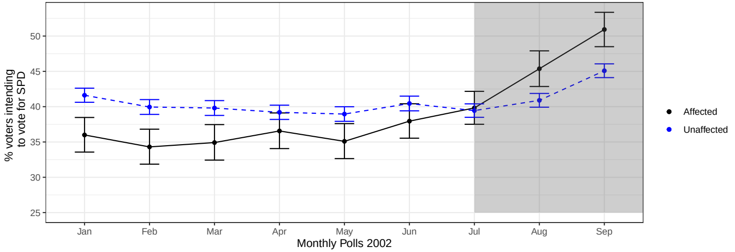
# Parallel Trends Assumption

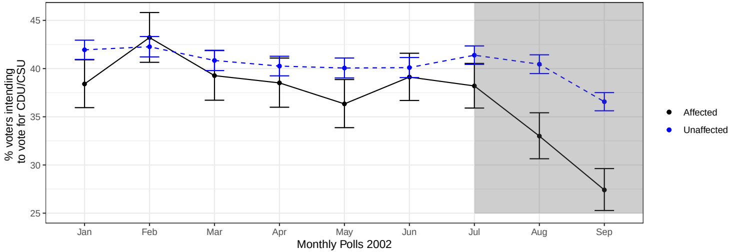
- $\mathbb{E}[Y_{i1}(0) - Y_{i0}(0) | D_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0) | D_i = 0]$
- Check leads: 1994 - 1998

```
all_data %>%
  filter(data_period == "1994-1998") %>%
  plm(data = ., index = c("wkr"),
    model = "fd",
    formula = spd_z_vs ~ Flooded * PostPeriod)
```

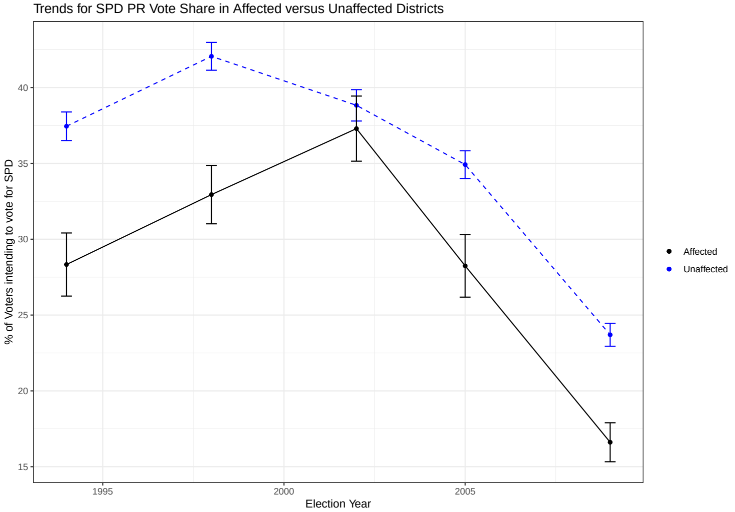
```
##
## Model Formula: spd_z_vs ~ Flooded * PostPeriod
## <environment: 0x7fed44820078>
##
## Coefficients:
## (Intercept)     Flooded
##  4.61036210 -0.00040711
```

# Visual Inspection of Parallel Trends Assumption



# More Diagnostics



# Competing Explanations

- A competing hypothesis involves the possible confounding salience of the Iraq war issue during the 2009 campaign.