이중차분(Difference-in-Differences) 추정기법의 개념 및 응용

손호성

중앙대학교 · 공공인재학부

2018년 9월 14일

한국복지패널 학술대회 발표 자료

목차

서론

- 정책효과를 분석 할 때 “두 집단”에 대한 자료 이용해 대개 평가를 함:
  1. 정책의 수혜자: 처리집단(Y_i^1)
  2. 정책의 비수혜자: 통제집단(Y_i^0)

  $\Longrightarrow$ 위에서 Y는 결과변수(종속변수)의 값을 나타내며, i는 개인을 나타내는 기호, 그리고 숫자 1과 0은 정책의 수혜 여부를 나타냄. 즉, 1이면 정책의 수혜자이고 0이면 정책의 비수혜자

- 위 두 집단 간 비교를 해서 어떤 정책의 효과($\bar{\tau}_i$)를 정의하면 다음과 같은 식으로 정의할 수 있음:

$$\bar{\tau}_i = \bar{Y}_i^1 - \bar{Y}_i^0 \tag{1}$$

- 식 (1)에서 기호 위에 있는 작대기는 평균을 나타냄:
  1. $\bar{Y}_i^1$는 정책 수혜자 집단의 결과변수의 평균
  2. $\bar{Y}_i^0$는 정책 수혜자 집단의 결과변수의 평균

서론

- 정책의 효과를 추정할 때, 정책 수혜자 집단과 비수혜자 집단이 반드시 서로 **다른** 집단이어야 하는 것은 아님

  $\Longrightarrow$ 예를 들어, A라는 집단이 정책의 수혜자이고 t 시점부터 정책이 시행됐다고 한다면, 이 A라는 집단의 t 시점 **이후**가 정책 수혜자 집단이 되고, 같은 A라는 집단의 t 시점 **이전**이 정책 비수혜자 집단으로 간주될 수 있음

- 마찬가지로 두 집단이 반드시 서로 같은 집단일 필요도 없음

  $\Longrightarrow$ 예를 들어, A라는 집단이 정책의 수혜자, B라는 집단이 정책의 비수혜자이고, t 시점부터 정책이 시행됐다고 한다면, 정책의 효과는 A 집단과 B 집단 간의 t 시점 **이후의** 결과변수 값의 차이를 계산함으로써 구할 수 있음

서론

$$\bar{\tau}_i = \bar{Y}_i^1 - \bar{Y}_i^0 \tag{2}$$

- 실제 자료를 이용한 실증연구에서 식 (2)와 같은 결과변수 값의 차이를 계산하기 위해서 다음과 같은 선형 회귀식을 추정하게 됨:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \tag{3}$$

식 (3)에서 T_i는 정책의 수혜 여부를 나타내는 다음과 같은 이항변수 (dummy variable)임

$$T_i = \begin{cases} 1, & \text{정책의 수혜자} \\ 0, & \text{정책의 비수혜자} \end{cases}$$

$\Longrightarrow$ 즉 정책의 효과는 β_1의 추정값 $\hat{\beta}_1$임!

서론

- 추정한 $\hat{\beta}_1$이 정책의 **인과적** 효과를 반영한다고 주장하기 위해서는 다음과 같은 조건이 필요함

"정책의 수혜자와 비수혜자, 이 두 집단 간에는
정책의 수혜 여부만이 차이가 날뿐 다른 모든 측면에서 비슷함"

- 위 조건에서 "다른 모든 측면에서" 비슷하다라는 것의 의미는 다음과 같이 두 가지 측면에서 두 집단이 비슷하다는 것임:

① 관측이 가능하거나 측정이 가능한 특성(예: 교육수준, 소득수준, 나이, 체중)

② 관측이 불가능하거나 측정이 어려운 특성(예: 인내심, 끈기, 능력, 부지런함)

- 구체적인 정책 사례를 통해서 위 조건에 대해 설명하고자 함

사례

- 정책 사례: 서울시에서 t년도부터 둘째 자녀를 낳은 부부에게 1,000만원을 지급하는 정책을 시행했다고 하겠음. 즉, 정책은 1,000만원을 지급하는 것이고 정책의 효과는 둘째 자녀 출산 확률이라고 할 수 있음

  $\Longrightarrow$ 다른 도시에서는 이와 같은 정책이 시행되고 있지 않다고 가정하겠음

- 앞서 말한대로 이 정책의 효과를 엄밀하게 추정하기 위해서는 최소한 두 집단, 즉 정책의 수혜자와 비수혜자에 대한 자료가 필요함. 하지만 대개 실제로 부부 단위에서 이들 부부가 1,000만원을 받았는지 안 받았는지에 대한 자료가 부재하기 때문에 다음과 같은 전략으로 정책의 효과를 추정함:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \tag{4}$$

식 (4)에서 Y_i는 결과변수이고 T_i는 정책의 수혜 여부를 나타내는 처리변수임:

$$Y_i = \begin{cases} 1, & \text{둘째 자녀를 낳음} \\ 0, & \text{둘째 자녀를 낳지 않음} \end{cases} \qquad T_i = \begin{cases} 1, & \text{서울시 거주} \\ 0, & \text{다른 도시 거주} \end{cases}$$

사례

- 서울시의 1,000만원 지급 정책이 둘째 출산 확률 여부에 미친 효과를 분석하기 위해 전국을 대표하는 표본 중에서 첫째 자녀만 낳은 부부를 분석 표본으로 한 다음, 식 (4)와 같은 선형 회귀식을 선형확률모형(Linear Probability Model)을 이용해서 그 효과를 추정했을 때 그 결과 값이 다음과 같다고 하겠음

$$\begin{aligned} Y_i &= \hat{\beta}_0 + \hat{\beta}_1 T_i \\ &= 0.33 + 0.23 T_i \end{aligned} \tag{5}$$

  $\Longrightarrow$ 정책의 효과를 반영하는 β_1이 0.23으로 추정되었으므로 이 정책은 둘째 자녀를 낳을 확률을 23%p 증가시키는 효과가 있다고 결론 내릴 수 있음

- 식 (5)를 좀 더 구체적으로 해석을 하면 절편 값이 0.33이므로 이것이 나타내는 것은 다른 도시에 거주하고 있는 첫째를 낳은 부부 중에서 둘째를 낳은 부부의 비율이 33%이고 $\hat{\beta}_1$이 0.23이므로 서울시에 거주하고 있는 첫째를 낳은 부부 중에서 둘째를 낳은 부부의 비율은 56%$(= 0.33 + 0.23)$임을 의미함

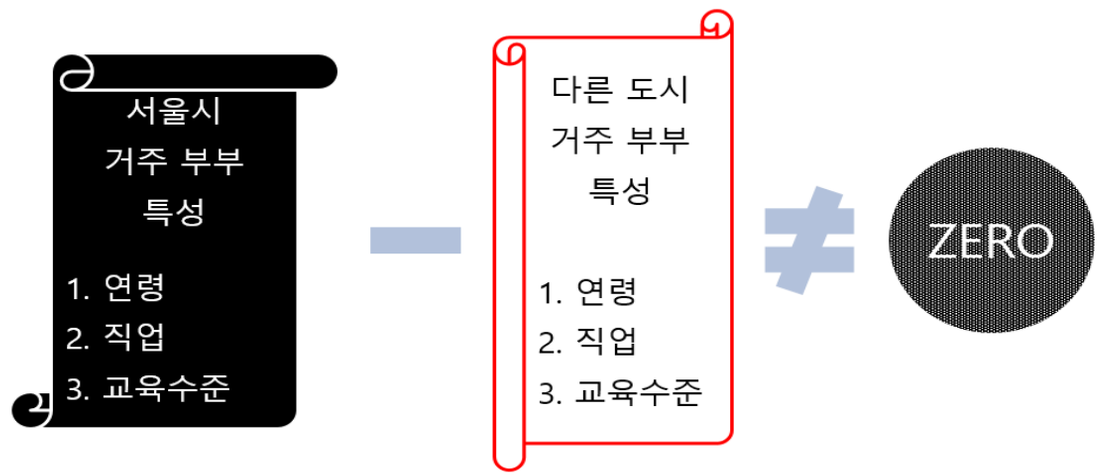
  $\Longrightarrow$ 따라서 정책의 효과는 23%p$(= 56\% - 33\%)$임

단순회귀분석의 한계

$$Y_i = 0.33 + 0.23T_i$$

- **질문: 위의 0.23이 과연 정책의 인과적 효과를 반영 할까?**
- $\hat{\beta}_1$이 정책의 인과적 효과를 반영한다고 주장하기 위해서는 다음과 같은 조건 필요:

  $\Longrightarrow$ 두 집단은 정책의 수혜 여부 외에는 모든 측면에서 비슷!
- 위 예에서 두 집단이 정책의 수혜 여부를 제외하고는 모든 측면에서 비슷하다고 할 수 있을까? 당연히 그렇지 않을 것임



단순회귀분석의 한계

$$\hat{\beta}_1 = 0.23 = \left\{ \begin{array}{c} \text{정책의 효과?} \\ \text{연령의 차이?} \\ \text{직업 관련 차이?} \\ \text{소득수준의 차이?} \\ \text{교육수준의 차이?} \\ \vdots \end{array} \right.$$

- 0.23이라는 추정값이 과연 정책의 효과를 반영하는지 혹은 두 집단의 다른 특성 차이로 인해 발생한 효과를 반영하는지 식별할 수 없다는 문제점이 발생

  $\Longrightarrow$ 극단적으로 0.23이라는 추정값이 모두 이러한 특성 차이로 인해 야기되었을 수도 있음. 그렇다고 한다면 실질적으로 정책의 효과는 없는 것임!

다중회귀분석의 한계

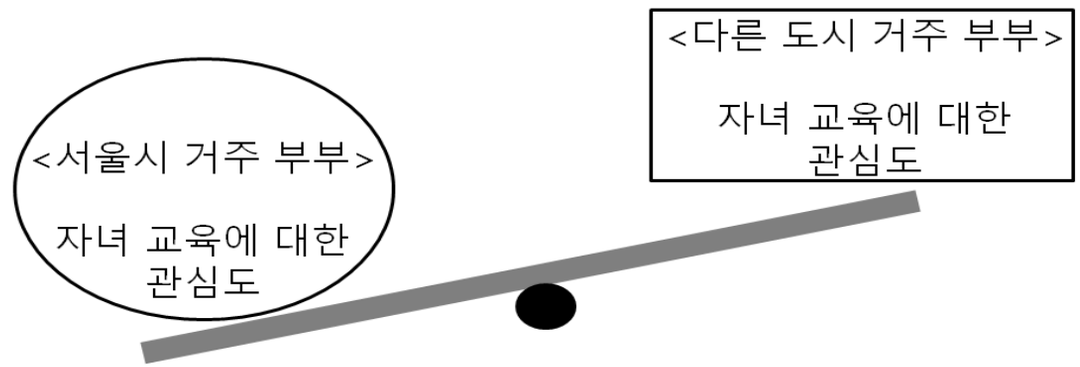
- 이러한 문제를 극복하기 위해 대개 연구자는 관측이 가능한 변수(X, Z 등)를 통제해서 식 (6)과 같은 방식으로 정책의 효과를 추정함

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 Z_i + \cdots + \varepsilon_i \tag{6}$$

  $\Longrightarrow$ 그럼 위 식을 토대로 추정한 $\hat{\beta}_1$은 정책의 인과적 효과를 반영할까?

- 관측이 가능한 변수를 통제해도 두 집단 간에는 관측이 불가능한 변수 측면에서도 다를 소지가 매우 크고 또한 관측이 가능한 변수를 통제한다고 해서 관측이 불가능한 특성이 통제가 되는 것은 아님

  $\Longrightarrow$ 정책의 수혜자와 비수혜자 간에는 대개 여러가지 면에서 다를 소지가 매우 큼



다중회귀분석의 한계

$$\hat{\beta}_1 = 0.23 = \left\{ \begin{array}{c} \text{정책의 효과?} \\ \sout{\text{연령의 차이?}} \\ \sout{\text{직업 관련 차이?}} \\ \sout{\text{소득수준의 차이?}} \\ \sout{\text{교육수준의 차이?}} \\ \text{자녀 교육에 대한 관심 차이?} \\ \text{자녀에 대한 가치관 차이?} \\ \text{결혼관의 차이} \\ \vdots \end{array} \right\}$$

- 따라서 어떤 정책의 **효과만을** 식별하기 위해서는 이러한 두 집단 간에 존재할 수 있는 관측 **불가능한** 특성에 대한 통제 또한 이루어져야 함

  $\Longrightarrow$ 즉, 식 (7)에서처럼 연구자가 관측할 수 없는 특성 δ_i에 대한 통제가 이루어져야 $\hat{\beta}_1$의 추정값이 정책의 효과만을 반영한다고 주장할 수 있음

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 Z_i + \cdots + \delta_i + \varepsilon_i \tag{7}$$

- 문제는 δ_i를 **관측할 수 없기 때문에** 회귀식에 이 변수를 통제할 수 없다는 점
- 그럼 어떻게 이 문제를 해결할 수 있을까?

  $\Longrightarrow$ 최근에 정책평가와 관련해서 많이 활용되고 있는 여러 준실험 설계 기법은 이와 같은 상황을 해결하기 위해 개발되었다고 해도 과언이 아님. 그 중에 특히 많이 활용되고 있는 것이 이중차분(DiD) 추정기법임

이중차분 추정기법 적용을 위해 필요한 정보

- 이중차분 추정기법으로 어떤 정책의 효과성을 분석하기 위해서는 다음과 같이 총 네 개 집단에 대한 자료를 필요로 함:
  1. 정책 **수혜자** 집단의 정책 시행 **이후** 자료
  2. 정책 **수혜자** 집단의 정책 시행 **이전** 자료
  3. 정책 **비수혜자** 집단의 정책 시행 **이후** 자료
  4. 정책 **비수혜자** 집단의 정책 시행 **이전** 자료
- 위에서 언급한 서울시의 1,000만원 지급 정책의 효과성을 이중차분 추정기법으로 분석하기 위해서는 다음과 같은 정보를 필요로 한다는 것을 의미함:
  1. 서울시에 거주하는 부부의 정책 시행 **이후** 자료
  2. 서울시에 거주하는 부부의 정책 시행 **이전** 자료
  3. 다른 도시에 거주하는 부부의 정책 시행 **이후** 자료
  4. 다른 도시에 거주하는 부부의 정책 시행 **이전** 자료

이중차분 추정기법 적용을 위해 필요한 정보

- 이중차분 추정기법이 어떻게 두 집단 간에 존재하는 관측 가능한 특성 및 관측 불가능한 특성에 대한 통제가 이루어지는지를 간단한 수학적 모델로 살펴보겠음:
  1. 처리집단의 정책 시행 이후: $Y_{it}^1 = D_{it} + \delta_i + \gamma_{it}^1$
  2. 처리집단의 정책 시행 이전: $Y_{it}^0 = \quad\quad \delta_i + \gamma_{it}^0$

     $\Longrightarrow$ 첫 번째 차분: $1 - 2 = Y_{it}^1 - Y_{it}^0 = D_{it} + \gamma_{it}^1 - \gamma_{it}^0$
  3. 통제집단의 정책 시행 이후: $Y_{it}^1 = \lambda_i + \xi_{it}^1$
  4. 통제집단의 정책 시행 이전: $Y_{it}^0 = \lambda_i + \xi_{it}^0$

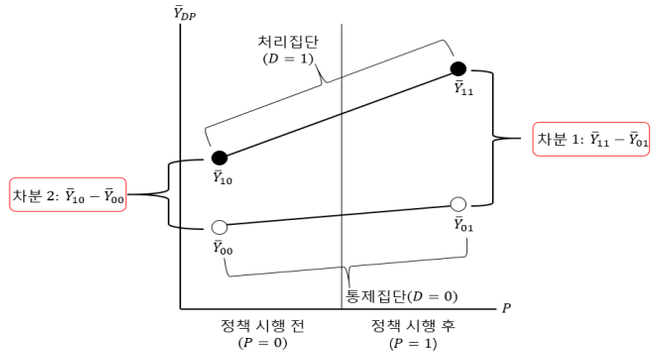
     $\Longrightarrow$ 두 번째 차분: $3 - 4 = Y_{it}^1 - Y_{it}^0 = \xi_{it}^1 - \xi_{it}^0$
- 첫 번째 차분 $-$ 두 번째 차분:

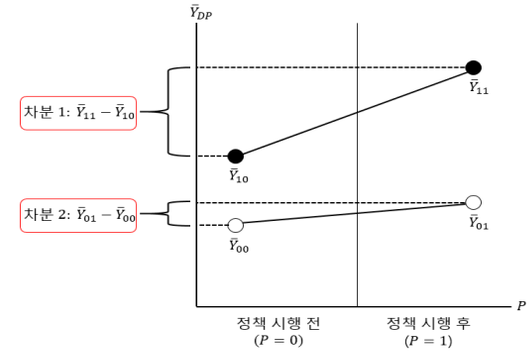
$$D_{it} + (\gamma_{it}^1 - \gamma_{it}^0) - (\xi_{it}^1 - \xi_{it}^0)$$

$\Longrightarrow$ 이중차분 전략으로 정책의 효과(D_{it})만을 식별하기 위해서는 다음과 같은 조건을 만족하면 됨:

$$(\gamma_{it}^1 - \gamma_{it}^0) - (\xi_{it}^1 - \xi_{it}^0) = 0$$
$$\Longrightarrow \gamma_{it}^1 - \gamma_{it}^0 = \xi_{it}^1 - \xi_{it}^0$$

이중차분(DiD) 추정기법의 원리: 그림을 이용해서





- DiD 추정량:

$$(\bar{Y}_{11}-\bar{Y}_{01})-(\bar{Y}_{10}-\bar{Y}_{00}) \qquad \text{혹은} \qquad (\bar{Y}_{11}-\bar{Y}_{10})-(\bar{Y}_{01}-\bar{Y}_{00})$$

이중차분(DiD) 추정기법의 원리: 회귀분석을 이용해서

- DiD 추정량은 대개 다음과 같은 회귀모형을 이용해서 추정하게 됨:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 P_t + \hat{\beta}_2 D_i + \hat{\beta}_3 (P_t \times D_i)$$

여기서

$$P_t = \begin{cases} 1, & t\text{가 정책 시행 이후} \\ 0, & t\text{가 정책 시행 이전} \end{cases} \qquad D_i = \begin{cases} 1, & i\text{가 처리집단} \\ 0, & i\text{가 통제집단} \end{cases}$$

그리고 $(P_t \times D_i)$ T_t와 D_i의 교호작용 변수임

- 그러면,

$$\begin{aligned}
\bar{Y}_{11}(D=1, P=1) &= \hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 1 + \hat{\beta}_3 \times 1 = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 \\
\bar{Y}_{10}(D=1, P=0) &= \hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 1 + \hat{\beta}_3 \times 0 = \hat{\beta}_0 + \hat{\beta}_2 \\
\bar{Y}_{01}(D=0, P=1) &= \hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 0 + \hat{\beta}_3 \times 0 = \hat{\beta}_0 + \hat{\beta}_1 \\
\bar{Y}_{00}(D=0, P=0) &= \hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 + \hat{\beta}_3 \times 0 = \hat{\beta}_0
\end{aligned}$$

즉, DiD 추정량은:

$$\begin{aligned}
(\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00}) &= \left[\left(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3\right) - \left(\hat{\beta}_0 + \hat{\beta}_2\right)\right] \\
&\quad - \left[\left(\hat{\beta}_0 + \hat{\beta}_1\right) - \left(\hat{\beta}_0\right)\right] \\
&= \left(\hat{\beta}_1 + \hat{\beta}_3\right) - \left(\hat{\beta}_1\right) = \hat{\beta}_3
\end{aligned}$$

이중차분(DiD) 추정기법의 원리: 표를 이용해서

- DiD 회귀식이 무엇을 추정하는지는 다음과 같은 표를 보면 명확해짐:

	통제집단 $(D=0)$	처리집단 $(D=1)$	차이
정책 시행 전 $(P=0)$	$\bar{Y}_{00}$	$\bar{Y}_{10}$	$\bar{Y}_{10}-\bar{Y}_{00}$
정책 시행 후 $(P=1)$	$\bar{Y}_{01}$	$\bar{Y}_{11}$	$\bar{Y}_{11}-\bar{Y}_{01}$
차이	$\bar{Y}_{01}-\bar{Y}_{00}$	$\bar{Y}_{11}-\bar{Y}_{10}$	$(\bar{Y}_{11}-\bar{Y}_{10})-(\bar{Y}_{01}-\bar{Y}_{00})$

	통제집단 $(D=0)$	처리집단 $(D=1)$	차이
정책 시행 전 $(P=0)$	$\hat{\beta}_0$	$\hat{\beta}_0+\hat{\beta}_2$	$\hat{\beta}_2$
정책 시행 후 $(P=1)$	$\hat{\beta}_0+\hat{\beta}_1$	$\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3$	$\hat{\beta}_2+\hat{\beta}_3$
차이	$\hat{\beta}_1$	$\hat{\beta}_1+\hat{\beta}_3$	$\hat{\beta}_3$

$\Longrightarrow$ DiD 추정량 $=(\bar{Y}_{11}-\bar{Y}_{10})-(\bar{Y}_{01}-\bar{Y}_{00})=\hat{\beta}_3$

DiD 추정기법의 식별조건(Identifying Assumption)

- DiD로 도출한 정책의 효과가 순수하게 정책의 효과만을 반영한다고 주장하기 위해서는 한 가지 가정을 만족해야 하는데, 다음과 같은 모집단 회귀식, 즉 오차항을 포함해서 DiD 추정기법의 식별조건을 알아보도록 하겠음:

$$Y_{it} = \beta_0 + \beta_1 P_t + \beta_2 D_i + \beta_3 (P_t \times D_i) + \varepsilon_{it}$$

- DiD 추정량을 보면:

$$\begin{aligned}
\left(\bar{Y}_{11} - \bar{Y}_{10}\right) - \left(\bar{Y}_{01} - \bar{Y}_{00}\right) &= [(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \varepsilon_{11}) - (\beta_0 + \beta_2 + \varepsilon_{10})] \\
&\quad - [(\beta_0 + \beta_1 + \varepsilon_{01}) - (\beta_0 + \varepsilon_{00})] \\
&= (\beta_1 + \beta_3 + \varepsilon_{11} - \varepsilon_{10}) - (\beta_1 + \varepsilon_{01} - \varepsilon_{00}) \\
&= \beta_3 + (\varepsilon_{11} - \varepsilon_{10}) - (\varepsilon_{01} - \varepsilon_{00})
\end{aligned}$$

$\Longrightarrow$ 여기서 $\varepsilon_{11} = E(\varepsilon_{it}|D_i = 1, P_t = 1)$, $\varepsilon_{10} = E(\varepsilon_{it}|D_i = 1, P_t = 0)$ 등을 말함

- 즉 DiD 전략 하에서의 식별조건은 다음과 같음:

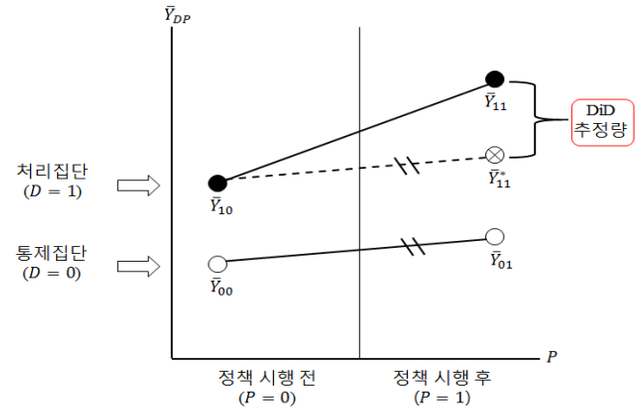
$$\begin{aligned}
\varepsilon_{11} - \varepsilon_{10} - \varepsilon_{01} - \varepsilon_{00} &= 0 \\
\Longrightarrow \varepsilon_{11} - \varepsilon_{10} &= \varepsilon_{01} - \varepsilon_{00}
\end{aligned}$$

DiD 추정기법의 식별조건(Identifying Assumption)

식별조건: $\varepsilon_{11} - \varepsilon_{10} = \varepsilon_{01} - \varepsilon_{00}$

- 위 식별조건을 평행 추세(parallel trends) 가정이라고 함. 이 가정이 의미하는 것은:

  $\Longrightarrow$ 만약 정책이 시행되지 않았더라면 결과변수 Y값이 처리집단과 통제집단 간에 비슷한 추세를 보였을 것이라는 것을 의미!

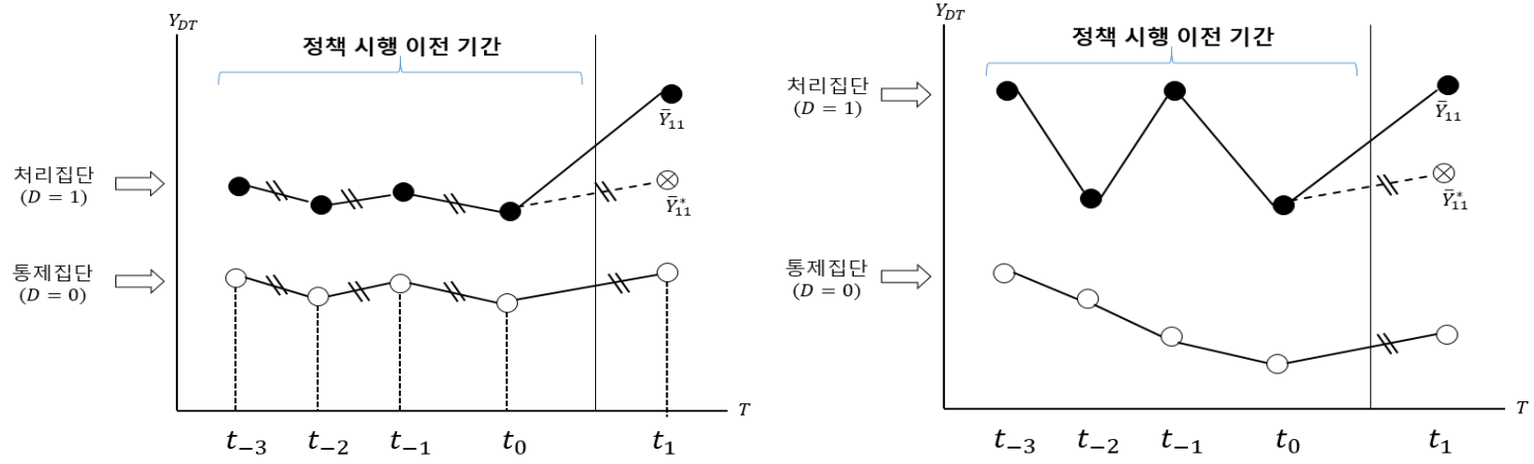


- 즉 정책이 시행되지 않았다면 실제 Y값은 $\bar{Y}_{11}$가 아닌 $\bar{Y}_{11}^{*}$이었을 것이라는 것을 의미!
- 위 가정을 만족해야 DiD 하에 도출된 효과 값이 정책의 인과적 효과를 반영하는 것인데 과연 이 가정을 test 할 수 있을까?

  $\Longrightarrow$ 불행하게도 이 가정을 test 할 수 없음! 이유는?

DiD 추정기법의 식별조건(Identifying Assumption)

- 정책은 실제로 시행이 됐기 때문에 우리는 이 평행 추세 가정을 “직접적으로” test 할 수는 없지만 “간접적으로” test 할 수 있음!
- How? 정책 시행 **이전에** 두 집단의 Y값이 평행한 추세를 보였는지를 확인!



$\Longrightarrow$ 어떤 시나리오가 평행 추세 가정이 만족한다는데 좀 더 강력한 근거가 될까? 왼쪽!

DiD 추정기법의 식별조건(Identifying Assumption)

- 따라서, DiD 추정기법을 사용해서 어떤 변수의 인과적 효과를 추정하기 위해서 연구자가 반드시 해야 할 사안은 정책 시행 **이전에** 두 집단의 결과변수 값이 비슷한 추세를 보였다는 것을 test하는 것임

  $\Longrightarrow$ 만약 두 집단의 결과변수 값이 정책 시행 이전에 비슷한 추세를 보이지 않았다면 정책 시행 이후에 관측된 결과변수 값의 차이가 정책에 기인한 것인지 혹은 다른 요인에 기인한 것인지 판단할 수 없음!

- 정책 시행 이전의 평행 추세 정도를 test 하는 것의 중요성이 현재 많은 연구에서 간과되고 있음

  $\Longrightarrow$ 좋은 저널의 편집장이나 리뷰어들은 평행 추세 가정이 만족하는지 test할 것을 반드시 요구함

- 그렇기 때문에 DiD 추정기법을 적용하기 위해서는 두 집단의 정책 도입 이전 시점 (예를 들어, t_{-3}까지)에 관측된 결과변수 값에 대한 자료가 필수적

회귀식을 활용한 추정 방식

- 회귀분석을 이용해서 DiD 추정을 할 때 다음과 같은 모형을 추정해서 추정값을 도출함:

$$Y_{it} = \beta_0 + \beta_1 P_t + \beta_2 D_i + \beta_3 (P_t \times D_i) + \varepsilon_{it}$$

- 위 방식보다 좀 더 나은 방식은 고정효과 모형을 이용한 추정 방식임:

$$Y_{it} = \alpha_0 + \alpha_1 (P_t \times D_i) + \gamma_i + \delta_t + \varepsilon_{it}$$

$\Longrightarrow$ 후자 방식의 장점: 표준오차 감소할 확률 높고 좀 더 모형의 fit이 좋아짐

- 자료를 이용해 $\hat{\beta}_3$ 혹은 $\hat{\alpha}_1$을 도출해서 이에 대한 통계적 검정을 하면 됨!

평행 추세 가정 Test 하는 방법

- 평행 추세 가정 test는 다음과 같은 회귀식을 추정해서 할 수 있음:

$$Y_{it} = \beta_0 + \beta_1 P_t + \beta_2 D_i + \sum_{t=2}^{T} \alpha_t (D_i \times Z_t) + \varepsilon_{it}$$

위에서 D_i와 P_t는 앞서 내린 정의와 같고 Z_t는 다음과 같은 이항변수를 나타냄:

$$Z_t = \begin{cases} 1, & \text{연도} = t \\ 0, & \text{연도} \neq t \end{cases}$$

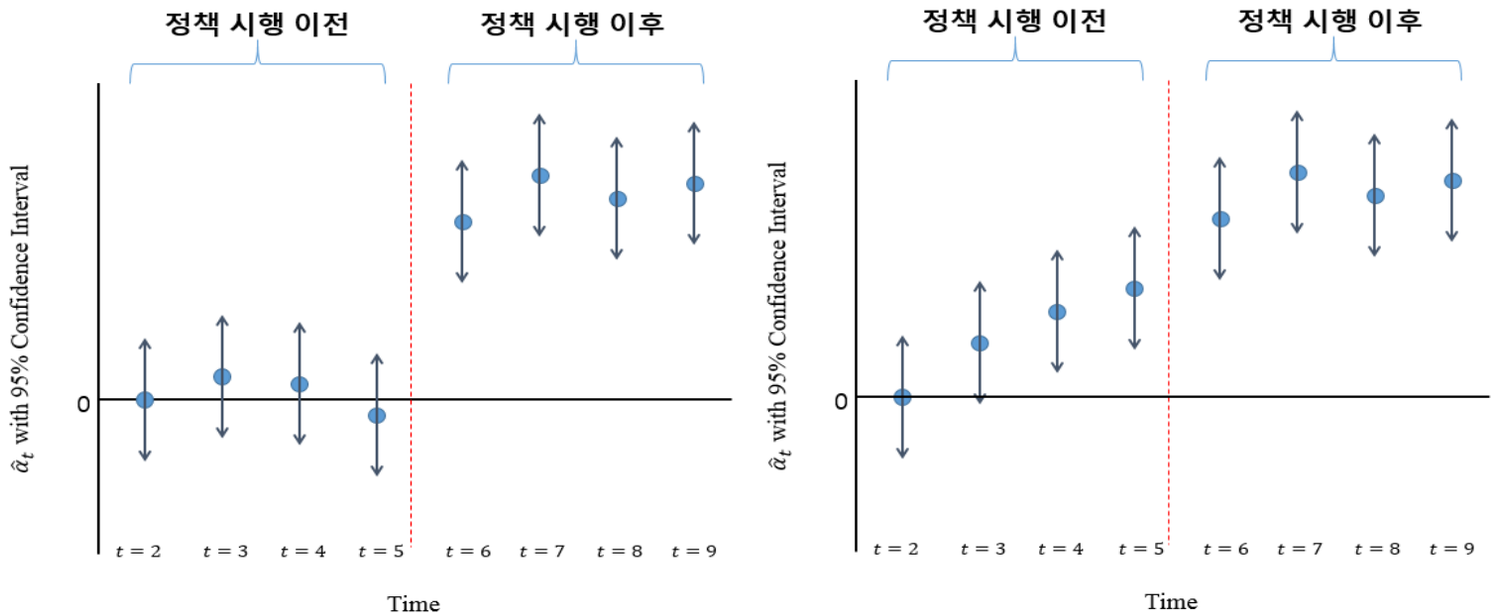
$\Longrightarrow$ 다시 말해, Z_t는 연도 더미변수임. 만약 5개 년도 자료이면, 네 개의 연도 더미를 식에 포함한다는 것을 의미함(base 연도 더미는 제외하고)

- 위 식에서 α_2는 연도 $t = 2$ 시점에 처리집단과 통제집단 간의 결과변수 값의 차이가 연도 $t = 1$ 시점(누락한 base 연도 더미)에서의 처리집단과 통제집단 간의 결과변수 값의 차이에 비해 얼마나 차이가 나는지를 나타냄

$\Longrightarrow$ 위 식을 추정한 후에 $\hat{\alpha}_t$ 추정값과 95% 신뢰구간을 그래프 등으로 보여주면 됨!

평행 추세 가정 Test 하는 방법

- 만약 평행 추세 가정이 만족하다고 볼 근거가 있기 위해서는 왼쪽과 같은 결과가 나와야 함



DiD를 활용한 연구 사례: Say Yes 정책의 평가

- Say Yes to Education 정책의 궁극적인 목적

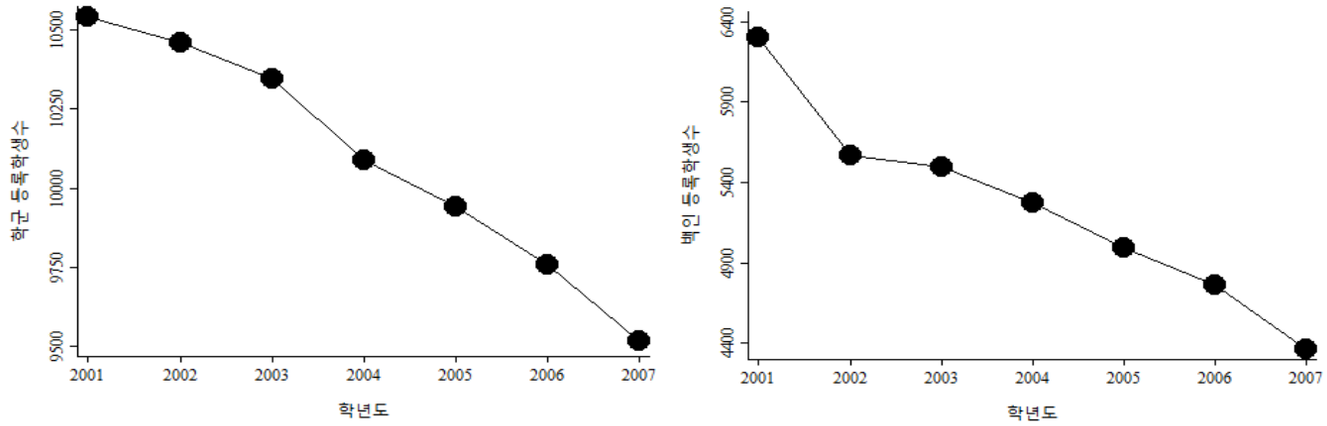
  $\Longrightarrow$ 교육 인센티브를 제공해서 Syracuse 시의 경제활성화

- 정책의 간략 소개

  1. 방과 후 및 계절학기 프로그램, 보건 및 법률 상담소, 대학진학 카운슬링 서비스 지원
  2. 대학 학자금 제공. 자격 요건 :
     - Syracuse 시에 거주해야 함
     - Syracuse 시에 속해 있는 공립학교에 3년 이상 등록을 해야 함
     - Syracuse 시에 속해 있는 공립 고등학교에서 졸업을 해야 함

DiD를 활용한 연구 사례: Say Yes 정책의 평가

- Say Yes to Education 정책의 도입 계기:



- 위와 같이 도시로부터 인구가 빠져나가면
  1. 뉴욕 주로부터 받는 학생 지원 감소 ⇒ 재정 수입의 감소
  2. 인구 감소 ⇒ 조세 수입의 감소 및 불경기 초래
  3. 중산층 및 고소득층 비율의 감소 ⇒ 범죄율 상승 등 초래

대조집단 선정: Rochester 학군

- 정책의 효과 증명하기 위해서는 Syracuse 시와 굉장히 비슷한 학군과의 비교 필요
- Syracuse 시는 뉴욕 주에서 5번 째로 큰 도시인데 뉴욕 주에서 Syracuse 시와 비슷한 도시로 Rochester 시를 들 수 있음

표 1. Syracuse 시 vs. Rochester 시

	정책 시행 전		정책 시행 후	
변수	Syracuse	Rochester	Syracuse	Rochester
영어 비능통자 학생 비율	11%	12%	11%	11%
저소득층 학생 비율	90%	87%	90%	85%
특별교육 대상자 비율	17%	19%	18%	19%
중식 지원 받는 학생 비율	72%	76%	83%	80%
흑인 학생 비율	53%	51%	62%	61%
남미 학생 비율	16%	17%	24%	25%
아시아 학생 비율	6%	6%	3%	3%
백인 학생 비율	22%	22%	10%	10%

대조집단 선정 : Rochester 시

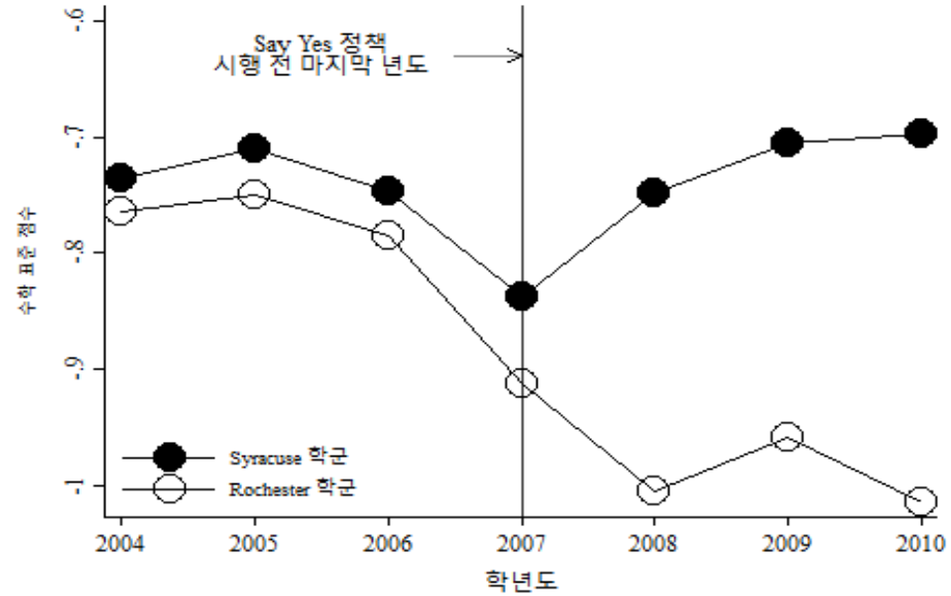
- 본 연구에서는 Rochester 시가 Syracuse 시와 매우 비슷하다는 가정 하에 두 집단의 정책 '전' 결과 변수의 차이와 두 집단의 정책 '후' 결과 변수의 차이를 비교해서 정책의 효과 추정

  $\Longrightarrow$ Difference-in-Differences(DiD) 전략!

- DiD를 적용하기 위해서는 우선 두 집단이 비슷하다는 걸 보여줘야 함 :

  1. 인구통계학적 성격 비슷했음
  2. 정책 시행 전까지 두 도시의 결과 변수의 추세(trend)가 비슷?

결과



Syracuse 시 성적 vs. Rochester 시 성적

$\Longrightarrow$ 전형적인 DiD 세팅!

DiD 관련 기타 이슈

- DiD는 두 개의 차분을 토대로 효과 값 추정하는데, 때로는 세 개의 차분을 토대로 효과 값 추정하는 게 유용할 수도 있음

  $\Longrightarrow$ Triple-Differences 모델

- DiD의 한계점 1: 대개 어떤 처리집단과 비교하기 위한 통제집단이 많은 경우가 많음

  ① 뉴욕 주에는 1,000개가 넘는 도시가 있음. 본 연구에서는 Rochester와 비교를 했지만 실제로 Syracuse 시와 비슷한 도시가 **여러 개** 있을 수도 있음

  $\Longrightarrow$ 그럼 어떤 도시를 택할 것인가? 사람마다 의견이 다를 수 있음(예, 서울과 비슷한 도시는? 부산? 분당?)

  ② 만약 평행 추세 가정을 간접적으로 만족하는 통제집단이 2개 이상이면 연구자 입장에서 가장 결과값이 좋게 나오는 통제집단을 임의로 선택해서 연구 결과를 보여줄 소지가 매우 큼!

  $\Longrightarrow$ 이를 어느 정도 극복한 연구설계기법: Synthetic Control Methods!

- DiD의 한계점 2: 일반적으로 회귀분석을 토대로 DiD 추정량을 추정하면 가설검정 시에 사용되는 표준오차 값이 틀린 경우가 많음

  $\Longrightarrow$ Clustered standard errors 등 사용 필요!