

인과추론과 실무 : Difference in Difference

가짜연구소 인과추론팀

발표자 : 김지연

1. 목차

패널 데이터

- 패널 데이터(panel data)
 - 여러 개체들을 복수의 시간에 걸쳐 관측하여 얻는 데이터
- 횡단면 데이터(cross-sectional data)
 - 수집 기간과 무관하게 복수의 개체들에 관한 데이터를 수집
- 시계열 데이터 (time-series data)
 - 한 대상을 복수의 시간에 걸쳐 관측하는 데이터
- 반복된 횡단면 데이터 (repeated cross sections data / cross-sectional time-series data)
 - 민주주의 국가들의 30년에 걸친 경제 성장을 데이터를 수집
 - 각 해의 민주주의 국가들은 이전의 민주주의 국가들과 동일한 개체라고 볼 수 없음
 - 인구 구성 / 선거를 통한 입법부, 행정부의 교체 등등

2019년 7월 A 프로모션의 효과를 평가해보자

	프로모션 A	프로모션 B
2018년 7월	150억 (A1)	200억 (B1)
2019년 7월	110억 (A2)	100억 (B2)

1. 2019년 07월 프로모션 없는 그룹 대비 A가 10억 높은 매출을 기록, 프로모션 효과는 긍정적!
2. 작년 대비 40억 감소, 프로모션 효과 부정적!

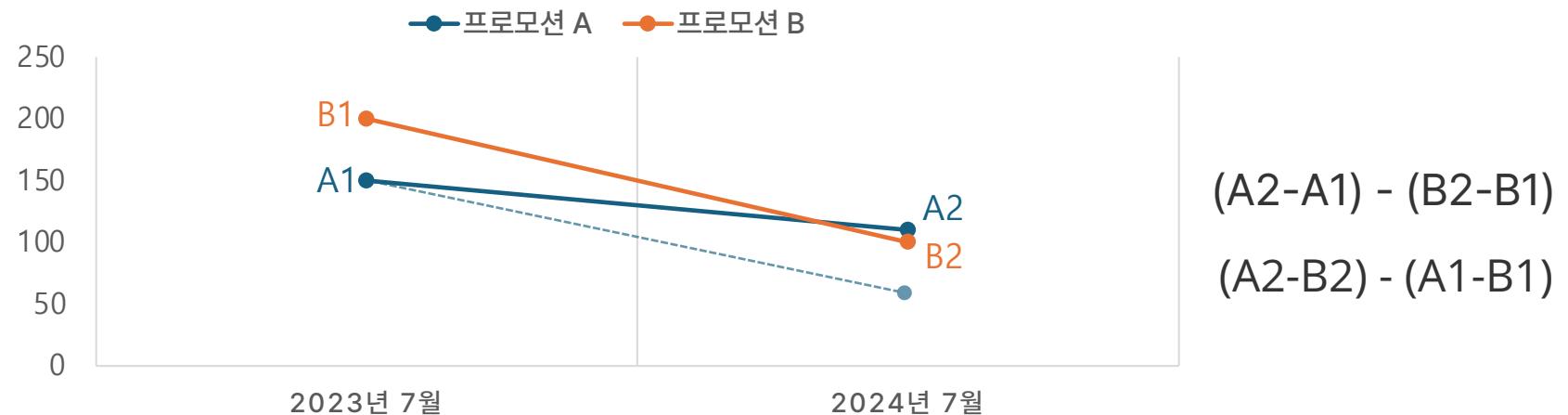
2019년 7월 A 프로모션의 효과를 평가해보자

	프로모션 A	프로모션 B
2018년 7월	150억 (A1)	200억 (B1)
2019년 7월	110억 (A2)	100억 (B2)

1. 2019년의 A,B의 횡단면만 비교하는 것은 각 특성에 의한 차이를 포함
2. A 프로모션의 2019년과 2018년만 비교하는 것은 시기별 특징에 따른 차이를 포함

2024년 7월 A 프로모션의 효과를 평가해보자

	프로모션 A	프로모션 B
2023년 7월	150억 (A1)	200억 (B1)
2024년 7월	110억 (A2)	100억 (B2)



이중차분법

$$ATT = E[Y_{it}(1) - Y_{it}(0)|D = 1, t > T_{pre}]$$

$$E[Y(0)|D = 1, Post = 1] \quad (\text{counterfactual})$$

$$= \underbrace{E[Y|D = 1, Post = 0]}_{(\text{실험군 기준값})} + \underbrace{E[Y|D = 0, Post = 1]}_{(\text{대조군 결과 추세})} - E[Y|D = 0, Post = 0]$$

$$ATT = (E[Y|D = 1, Post = 1] - E[Y|D = 1, Post = 0]) \\ - (E[Y|D = 0, Post = 1] - E[Y|D = 0, Post = 0])$$

t : 시간

D : 처치변수

처치 이후 $Post = 1$

처치 이전 $Post = 0$

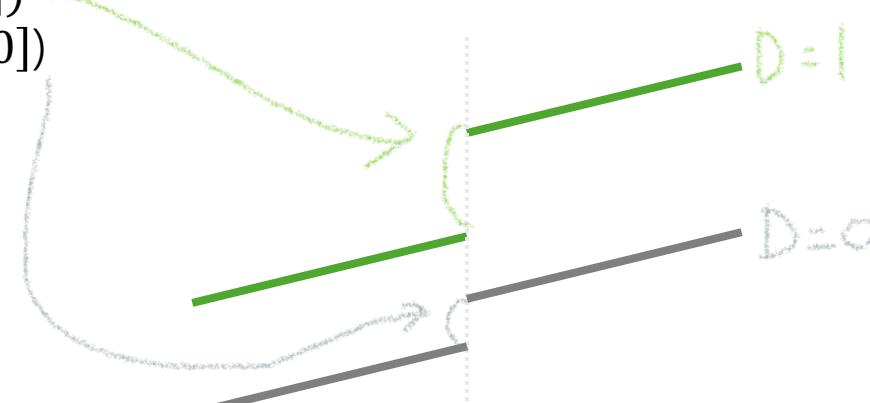
이중차분법

$$ATT = E[Y_{it}(1) - Y_{it}(0)|D = 1, t > T_{pre}]$$

$E[Y(0)|D = 1, Post = 1]$ (counterfactual)

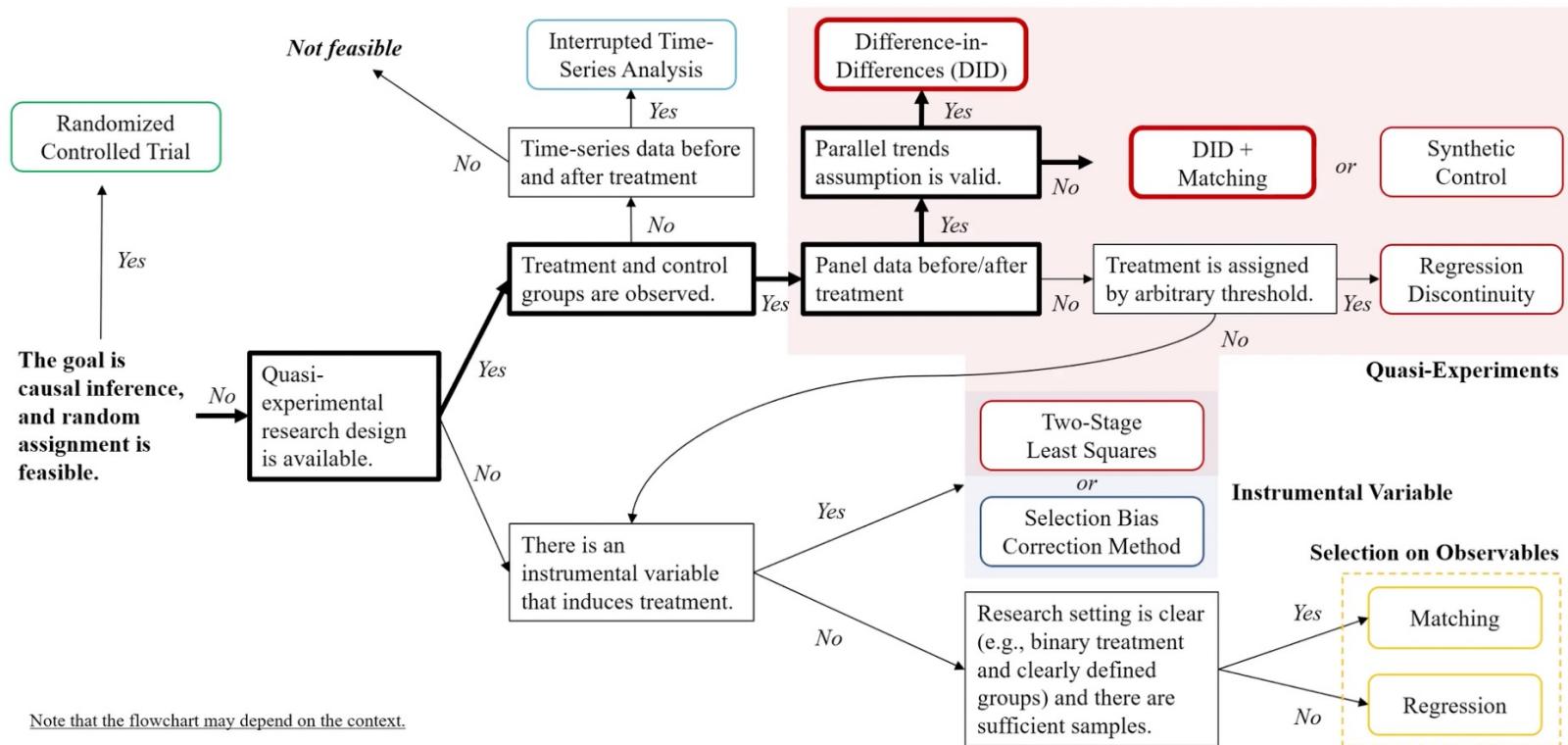
$$ATT = (E[Y|D = 1, Post = 1] - E[Y|D = 1, Post = 0]) \\ - (E[Y|D = 0, Post = 1] - E[Y|D = 0, Post = 0])$$

$$= E[\Delta y | D = 1] - E[\Delta y | D = 0]$$



이중차분법을 언제 사용 할 수 있을까?

- 인과 효과 추정을 위해서는 RCT(Randomized Controlled Trials) 상황에서 A/B 테스트를 활용하는 것이 이상적
- 하지만 대부분 A/B테스트를 하기 어려운 환경 + 사후 분석이 대부분



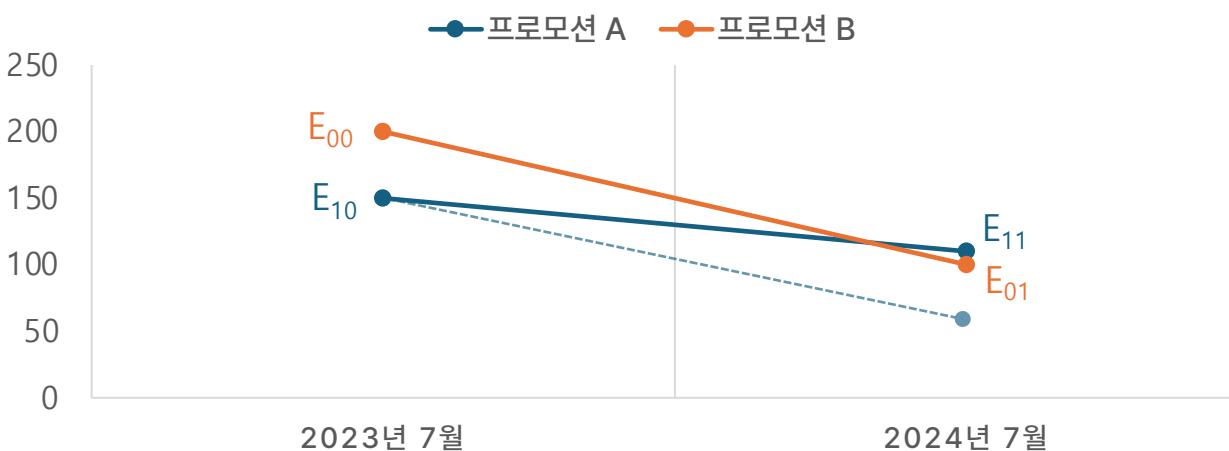
앨런 크루거 Alan Krueger의 최저임금 연구

- 최저임금 인상이 고용에 미치는 영향을 실증적으로 분석
- 연구 배경
 - 뉴저지 주는 1992년 4월 최저임금을 시간당 \$4.25에서 \$5.05로 인상
 - 이 연구는 이 최저임금 인상이 뉴저지와 인근 펜실베니아를 비교, 최저임금 영향력 확인 (펜실베니아의 경우 4월 전/후 시간당 최저임금 \$4.25 유지)
- 연구 방법
 - 최저임금 인상 전후의 고용 규모에 대한 데이터를 수집
 - 준실험 방법인 이중차분법을 활용

	뉴저지 (1)	펜실베니아 (2)	(1)-(2)
1992년 2~3월 (I)	20.44 (0.51)	23.33 (1.35)	-2.89 (1.44)
1992년 10~11월 (II)	21.03 (0.52)	21.17 (0.94)	-0.14 (1.07)
(II) - (I)	0.59 (0.54)	-2.16 (1.25)	2.76 (1.36)

이중차분법과 회귀

	처치 그룹	통제 그룹
2018년 7월	150억 (A1)	200억 (B1)
2019년 7월	110억 (A2)	100억 (B2)



$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Post_t + \beta_3 D_i Post_t + e_{it}$$

- $D = 0, Post = 0$ β_0 통제그룹의 기준값
- $D = 0, Post = 1$ $\beta_0 + \beta_2$ 시간에 따른 개입 전후 차이
- $D = 1, Post = 0$ $\beta_0 + \beta_1$ 개체간의 차이
- $D = 1, Post = 1$ $\beta_0 + \beta_1 + \beta_2 + \beta_3$ DID 추정량

이중차분법과 회귀

outcome	D	Post
18.076601	1	0
24.072336	1	0
27.165262	1	0
26.883542	1	0
30.273883	1	0
51.207363	1	1
52.592823	1	1
57.236814	1	1
65.162343	1	1
64.267490	1	1
9.531791	0	0
14.177175	0	0
19.934620	0	0
20.286638	0	0
22.906351	0	0
30.766237	0	1
32.826054	0	1
36.676317	0	1
43.247555	0	1
43.002449	0	1

 E_{10} E_{11} E_{00} E_{01}

통제 그룹	처치 전	처치 후
	처치 그룹	처치 그룹
	17.367315 (E_{00})	37.303723 (E_{01})
	25.294325 (E_{10})	58.093366 (E_{11})

$$(E_{11} - E_{10}) - (E_{01} - E_{00}) \\ = (58.093366 - 25.294325) - (37.303723 - 17.367315) = 12.86263$$

sm.ols('outcome ~ D + Post + D:Post', data).fit().summary().tables[1]

	coef	std err	t	P> t	[0.025	0.975]
Intercept	17.3673	2.498	6.952	0.000	12.071	22.663
T_d	7.9270	3.533	2.244	0.039	0.438	15.416
P_t	19.9364	3.533	5.643	0.000	12.447	27.426
T_d:P_t	12.8626	4.996	2.574	0.020	2.271	23.454

EITC의 고용 증가 효과

- 근로장려세제(EITC; Earned Income Tax Credit)란?
소득 기준으로 특정 계층 이하에 대해서 일을 했을 경우 소득에 부과되는 세금을 돌려주는 제도 (일종의 노동 장려책)
- 연구 배경
 - 그런데, EITC의 자격 조건 중에서 자녀 유무가 존재하는 경우가 많다.
그렇다면 미혼 여성 노동자의 경우 EITC가 노동 공급을 늘리게 될까?
- 연구 방법
 - EITC를 무작위 할당으로 주는 실험이 불가능, 차선책으로 DID(Difference in Differences)를 사용
 - 조건이 되는 자녀 유무에 따라서 처치 집단과 통제 집단을 나눌 수 있고, EITC가 실행되기 전과 후를 살펴봄으로써 DID 추정치를 구함

	variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
자녀의 숫자	children	0	13746	13746	1.19	1.38	0	0	1	2	9	
취업 여부	work	0	13746	13746	0.51	0.5	0	0	1	1	1	
연도	year	0	13746	13746	1993.35	1.7	1991	1992	1993	1995	1996	

자녀 우무 차이

$$\text{work} = \beta_0 + \beta_1 \text{post} + \beta_2 \text{anykids} + \delta(\text{anykids} \times \text{post}) + \varepsilon$$

EITC 시작 전/후 차이

DID 추정량

EITC의 고용 증가 효과

```
model = lm(work ~ anykids*post93, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = work ~ anykids * post93, data = data)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -0.5755 -0.4908  0.4245  0.5092  0.5540 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)              0.575460  0.008845 65.060 < 2e-16 ***
## anykidsTRUE             -0.129498  0.011676 -11.091 < 2e-16 ***
## post93TRUE              -0.002074  0.012931  -0.160  0.87261  
## anykidsTRUE:post93TRUE  0.046873  0.017158   2.732  0.00631 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4967 on 13742 degrees of freedom
## Multiple R-squared:  0.0126, Adjusted R-squared:  0.01238 
## F-statistic: 58.45 on 3 and 13742 DF,  p-value: < 2.2e-16
```

식별 가정

1. 평행 추세 가정
2. SUTVA
3. 강외생성
 1. 시간에 따라 변하지 않는 교란 요인
 2. 피드백 없음
 3. 이월 효과 없음

평행 추세 가정 (Parallel trend Assumption)

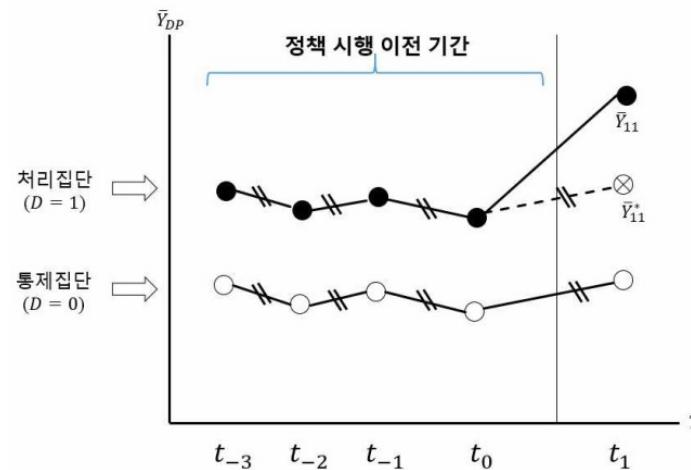
이중차분법은 처치가 없으면 평균적으로 실험군과 대조군의 결과 추세가 동일할 것이라는 가정이 필요

= 비교 그룹간 $Y(0)$ 처치 이전의 Pre-trend가 평행하다고 가정

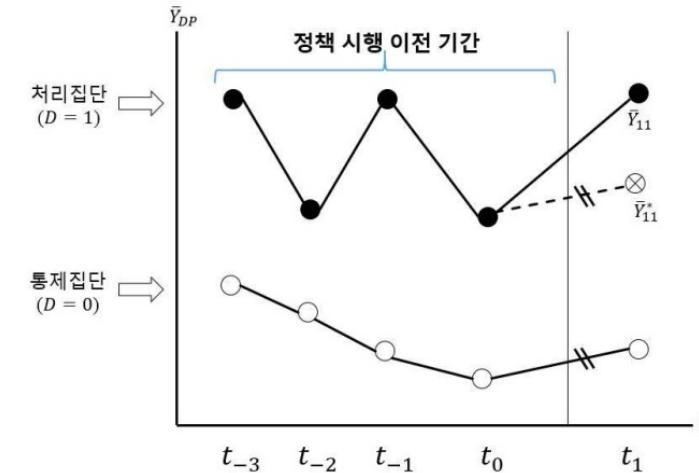
$$E[Y(0)_{it=1} - Y(0)_{it=0}|D=1] = E[Y(0)_{it=1} - Y(0)_{it=0}|D=0]$$

= 처치가 무작위 배정되지 않더라도 실험군과 대조군의 반사실 추세가 동일하다면, ATT 식별 가능

〈그림 7〉 DiD 추정량의 평행 추세 가정이 만족할만한 상황



〈그림 8〉 DiD 추정량의 평행 추세 가정이 만족하지 않을만한 상황



평행 추세 가정 (Parallel trend Assumption)

Event Study DID Model

Leads, lags 처치 전 1,2,3년 / 처치 이후 1,2,3년 변수를 고려

1. Treatment 이전 기간의 효과를 같이 측정
2. Treatment 이전 기간 leads term의 효과가 0이라고 한다면 두 집단이 평행추세를 따른다고 가정

Customer ID	Day	Treatment Group	Post	DID Terms		Relative-Time Terms			
				Push Notification (Treat×Post)	Relative Time (-2) Dummy	Relative Time (-1) Dummy	Relative Time (0) Dummy	Relative Time (1) Dummy	Relative Time (2) Dummy
1	1	1	0	0	1	0	0	0	0
1	2	1	0	0	0	1	0	0	0
1	3	1	1	1	0	0	1	0	0
1	4	1	1	1	0	0	0	1	0
2	1	1	0	0	0	1	0	0	0
2	2	1	1	1	0	0	1	0	0
2	3	1	1	1	0	0	0	1	0
2	4	1	1	1	0	0	0	0	1

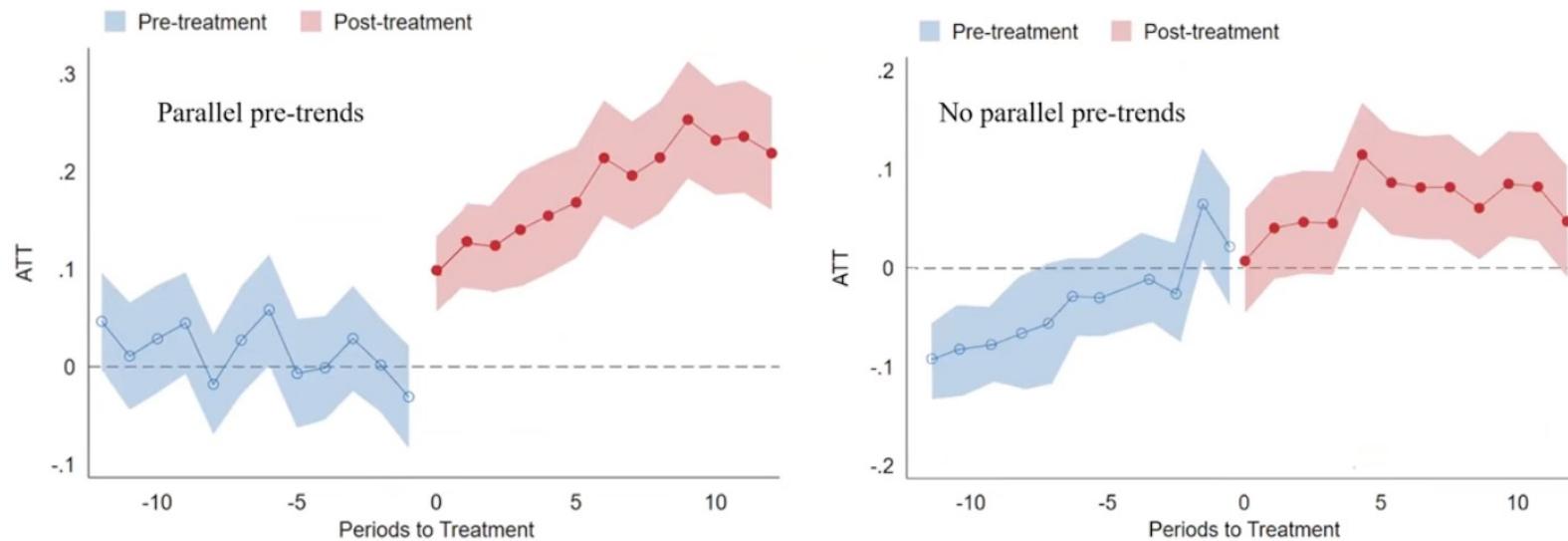
✓ Relative time dummies are perfectly collinear with treatment group or unit fixed effects.

✓ One group (typically, -1) should be omitted, which serves as a reference case.

Relative-time Term 중 일부를 누락해서 baseline 생성

Event Study DID Model

Leads 변수들의 효과를 측정했을 때, 신뢰구간이 0 을 포함 하고 있을 시 통계적으로 유효하지 않음
= Parallel pre trend를 만족한다고 할 수 있음



평행 추세 가정 (Parallel trend Assumption)

평행 추세 가정 방법

1. 그래프를 통한 탐색
2. Stata 패키지 > estat ptrends
 - 처치그룹과 통제그룹의 처리 전 기간 동안의 기울기의 차이를 포착
 - 해당 값을 0으로 둔 회귀식을 세우고 Wald검정 수행, 이때 이 검정의 귀무가설은 선형 추세가 평행하다

SUTVA

실험군의 효과가 대조군으로 영향을 미치지 않아야 한다 (파급효과는 없어야 함)

예)

백신 접종과 집단 면역

- 한 학교에서 학생들에게 백신을 접종합니다.
- 한 학생이 백신을 맞으면 그 학생뿐만 아니라 주변 친구들도 감염 확률이 줄어듭니다(집단 면역 효과).

SUTVA 위배:

- 한 학생의 건강 상태는 다른 학생들이 백신을 맞았는지에 따라 달라집니다.
- 각 단위(학생)의 결과가 다른 단위(다른 학생들)의 처리(백신 접종)에 의해 영향을 받기 때문에 독립성이 깨집니다.

소셜 미디어 광고

- 어떤 회사가 두 그룹에게 소셜 미디어 광고를 보여줍니다. 그룹 A와 그룹 B가 있습니다.
- 그룹 A가 광고를 보면, 그룹 B도 이 광고에 대해 친구들과 이야기하면서 영향을 받을 수 있습니다.

SUTVA 위배:

- 그룹 B의 구매 행동은 그룹 A가 광고를 보았는지 여부에 따라 영향을 받습니다.
- 이는 각 단위(그룹)의 결과가 다른 단위(다른 그룹)의 처리(광고 노출)에 의해 영향을 받기 때문에 독립성이 깨집니다.

강외생성 - Post-treatment bias를 배제하기 위한 조건

- 패널 데이터는 시간이 개입되므로 현재의 변수가 과거의, 혹은 미래의 오차항과 상관되는지 문제에 대한 조건

시간에 따라 변하지 않는 교란 요인

- 교란요인이 시간에 따라 일정하다면, 교란요인이 존재해도 인과효과 식별 가능
- 시간 고정 공변량을 추가하여, 변수들을 통제 (예) 도시의 문화, 법률..

피드백 없음

- 과거의 결과인 Y_{it-1} 가 현재의 처치인 W_{it} 에 영향을 주지 않음
- 예를 들어 마케팅 프로모션의 일환으로 다운로드수가 1,000건에 도달할 때마다 오프라인 캠페인 진행

이월 효과와 시차종속변수 없음

- 과거의 처치가 현재 결과에 영향을 주지 않음
- 과거의 결과가 현재 결과의 직접적인 원인이 되지 않음 (필수X)

이중차분법의 변형

1. 시간에 따른 효과 변동
2. 이중차분법과 공변량

시간에 따른 효과 변동

이중차분법을 활용하여 처치효과를 측정할 때, 처치효과가 즉각적이지 않은 경우들도 존재

처치효과가 완전히 나타가지 전 기간을 포함하여, 처치 효과를 과소평가할 수 있음

=> 처치 이후 시간에 따른 ATT를 추정

```
def did_date(df, date):
    df_date = (df
        .query("date==@date | post==0")
        .query("date <= @date")
        .assign(post = lambda d: (d["date"]==date).astype(int)))

    m = smf.ols(
        'downloads ~ I(treated*post) + C(city) + C(date)', data=df_date
    ).fit(cov_type='cluster', cov_kwds={'groups': df_date['city']})

    att = m.params["I(treated * post)"]
    ci = m.conf_int().loc["I(treated * post)"]

    return pd.DataFrame({"att": att, "ci_low": ci[0], "ci_up": ci[1]}, index=[date])

post_dates = sorted(mkt_data["date"].unique())[1:] # 일자 리스트

atts = pd.concat([did_date(mkt_data, date)
    for date in post_dates])
```

	date	city	region	treated	tau	downloads	post
0	2021-05-01	5	S	0	0.000000	51.0	0
1	2021-05-02	5	S	0	0.000000	51.0	0
2	2021-05-03	5	S	0	0.000000	51.0	0
3	2021-05-04	5	S	0	0.000000	50.0	0
4	2021-05-05	5	S	0	0.000000	49.0	0
...
1627	2021-05-28	197	S	1	1.771233	53.0	1
1628	2021-05-29	197	S	1	1.771233	52.0	1
1629	2021-05-30	197	S	1	1.771233	54.0	1
1630	2021-05-31	197	S	1	1.771233	53.0	1
1631	2021-06-01	197	S	1	1.771233	55.0	1

시간에 따른 효과 변동

```

def did_date(df, date):
    df_date = (df
        .query("date==@date | post==0")
        .query("date <= @date")
        .assign(post = lambda d: (d["date"]==date).astype(int)))

    m = smf.ols(
        'downloads ~ I(treated*post) + C(city) + C(date)', data=df_date
    ).fit(cov_type='cluster', cov_kwds={'groups': df_date['city']})

    att = m.params["I(treated * post)"]
    ci = m.conf_int().loc["I(treated * post)"]

    return pd.DataFrame({"att": att, "ci_low": ci[0], "ci_up": ci[1]}, index=[date])

post_dates = sorted(mkt_data["date"].unique())[1:] # 일자 리스트 →
atts = pd.concat([did_date(mkt_data, date)
                  for date in post_dates]) ←

```

```

[Timestamp('2021-05-02 00:00:00'),
 Timestamp('2021-05-03 00:00:00'),
 Timestamp('2021-05-04 00:00:00'),
 Timestamp('2021-05-05 00:00:00'),
 Timestamp('2021-05-06 00:00:00'),
 Timestamp('2021-05-07 00:00:00'),
 Timestamp('2021-05-08 00:00:00'),
 Timestamp('2021-05-09 00:00:00'),
 Timestamp('2021-05-10 00:00:00'),
 Timestamp('2021-05-11 00:00:00'),
 Timestamp('2021-05-12 00:00:00'),
 Timestamp('2021-05-13 00:00:00'),
 Timestamp('2021-05-14 00:00:00'),
 Timestamp('2021-05-15 00:00:00'),
 Timestamp('2021-05-16 00:00:00'),
 Timestamp('2021-05-17 00:00:00'),
 Timestamp('2021-05-18 00:00:00'),
 Timestamp('2021-05-19 00:00:00'),
 Timestamp('2021-05-20 00:00:00'),
 Timestamp('2021-05-21 00:00:00'),
 Timestamp('2021-05-22 00:00:00'),
 Timestamp('2021-05-23 00:00:00'),
 Timestamp('2021-05-24 00:00:00'),
 Timestamp('2021-05-25 00:00:00'),
 Timestamp('2021-05-26 00:00:00'),
 ...
 Timestamp('2021-05-28 00:00:00'),
 Timestamp('2021-05-29 00:00:00'),
 Timestamp('2021-05-30 00:00:00'),
 Timestamp('2021-05-31 00:00:00'),
 Timestamp('2021-06-01 00:00:00')]

```

시간에 따른 효과 변동

```

def did_date(df, date):
    df_date = (df
        .query("date==@date | post==0")
        .query("date <= @date")
        .assign(post = lambda d: (d["date"]==date).astype(int)))

    m = smf.ols(
        'downloads ~ I(treated*post) + C(city) + C(date)', data=df_date
    ).fit(cov_type='cluster', cov_kwds={'groups': df_date['city']})

    att = m.params["I(treated * post)"]
    ci = m.conf_int().loc["I(treated * post)"]

    return pd.DataFrame({"att": att, "ci_low": ci[0], "ci_up": ci[1]}, index=[date])

post_dates = sorted(mkt_data["date"].unique())[1:] # 일자 리스트

atts = pd.concat([did_date(mkt_data, date)
                 for date in post_dates])

```

2021-05-02 00:00:00								
	date	city	region	treated	tau	downloads	post	
0	2021-05-01	5	S	0	0.0	51.0	0	
1	2021-05-02	5	S	0	0.0	51.0	1	
32	2021-05-01	15	S	0	0.0	50.0	0	
33	2021-05-02	15	S	0	0.0	49.0	1	
64	2021-05-01	20	S	0	0.0	48.0	0	
...
1537	2021-05-02	195	S	0	0.0	55.0	1	
1568	2021-05-01	196	S	0	0.0	52.0	0	
1569	2021-05-02	196	S	0	0.0	53.0	1	
1600	2021-05-01	197	S	1	0.0	51.0	0	
1601	2021-05-02	197	S	1	0.0	52.0	1	

2021-06-01 00:00:00								
	date	city	region	treated	tau	downloads	post	
0	2021-05-01	5	S	0	0.000000	51.0	0	
1	2021-05-02	5	S	0	0.000000	51.0	0	
2	2021-05-03	5	S	0	0.000000	51.0	0	
3	2021-05-04	5	S	0	0.000000	50.0	0	
4	2021-05-05	5	S	0	0.000000	49.0	0	
...
1610	2021-05-11	197	S	1	0.000000	52.0	0	
1611	2021-05-12	197	S	1	0.000000	51.0	0	
1612	2021-05-13	197	S	1	0.000000	50.0	0	
1613	2021-05-14	197	S	1	0.000000	50.0	0	
1631	2021-06-01	197	S	1	1.771233	55.0	1	

시간에 따른 효과 변동

```

def did_date(df, date):
    df_date = (df
        .query("date==@date | post==0")
        .query("date <= @date")
        .assign(post = lambda d: (d["date"]==date).astype(int)))

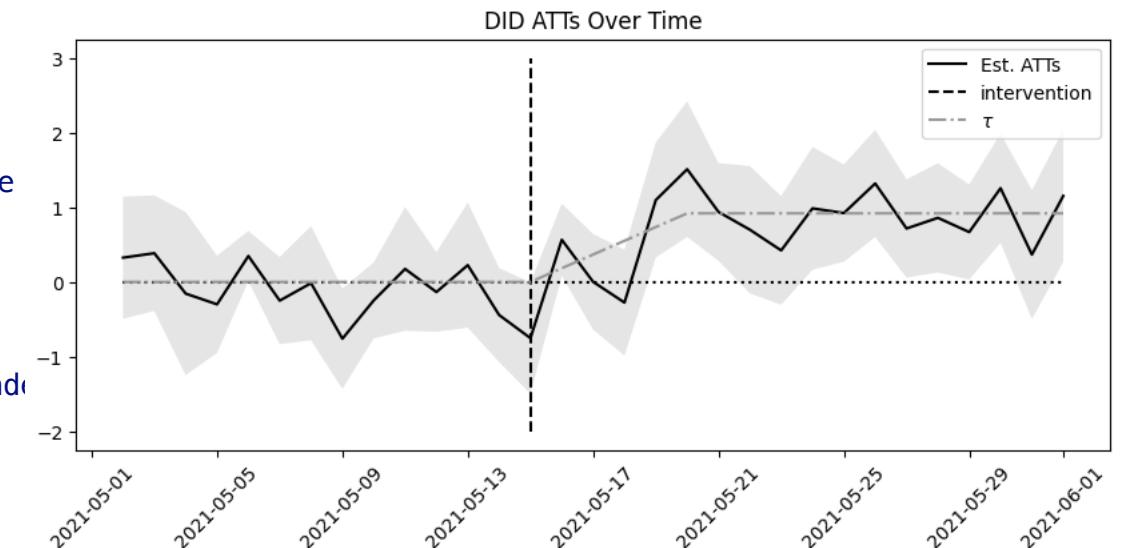
    m = smf.ols(
        'downloads ~ I(treated*post) + C(city) + C(date)', data=df_date
    ).fit(cov_type='cluster', cov_kwds={'groups': df_date['city']})

    att = m.params["I(treated * post)"]
    ci = m.conf_int().loc["I(treated * post)"]

    return pd.DataFrame({"att": att, "ci_low": ci[0], "ci_up": ci[1]}, index=[date])

post_dates = sorted(mkt_data["date"].unique())[1:] # 일자 리스트
atts = pd.concat([did_date(mkt_data, date)
                  for date in post_dates])

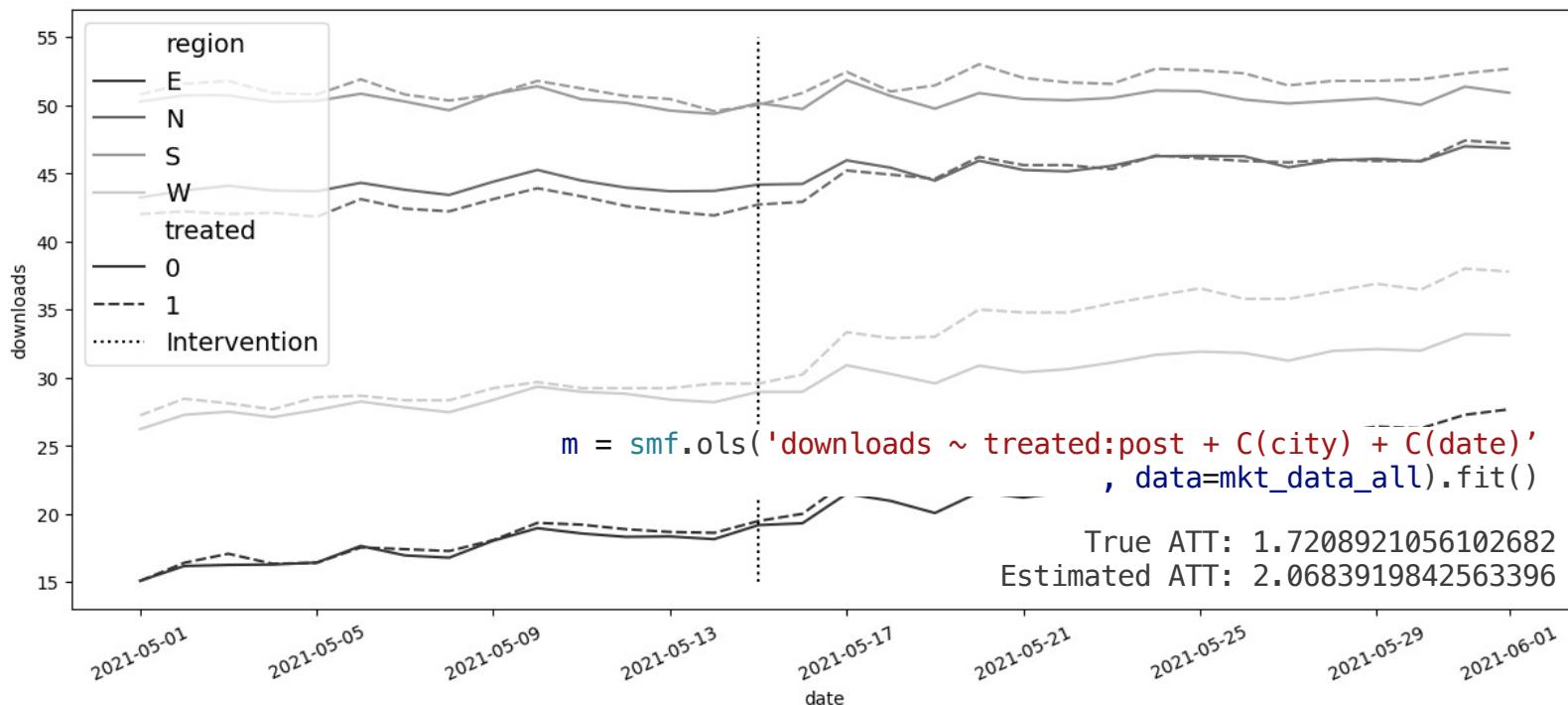
```



이중차분법과 공변량

평행추세 가정을 만족하지 않지만, 공변량을 조건부로 두었을 때 해당 가정을 만족하는 경우 사용

$$E[Y(0)_{it=1} - Y(0)_{it=0}|D=1, X] = E[Y(0)_{it=1} - Y(0)_{it=0}|D=0, X]$$



이중차분법과 공변량

평행추세 가정을 만족하지 않지만, 공변량을 조건부로 두었을 때 해당 가정을 만족하는 경우 사용

```
m = smf.ols('downloads ~ treated:post + C(city) + C(date)',  
            data=mkt_data_all).fit()
```

treated:post : treated와 post 변수 간의 상호작용 항. DID 추정량

C(city) : 범주형 변수 city에 대한 더미 변수

C(date) : 범주형 변수 date에 대한 더미 변수

```
m_saturated = smf.ols('downloads ~ (post*treated)*C(region)',  
                      data=mkt_data_all).fit()
```

- **post*treated**: post와 treated의 상호작용 항

- post + treated + post:treated

- **(post*treated)*C(region)** : post와 treated의 상호작용 항이 각 region 내에서 다를 수 있음을 나타냅니다.

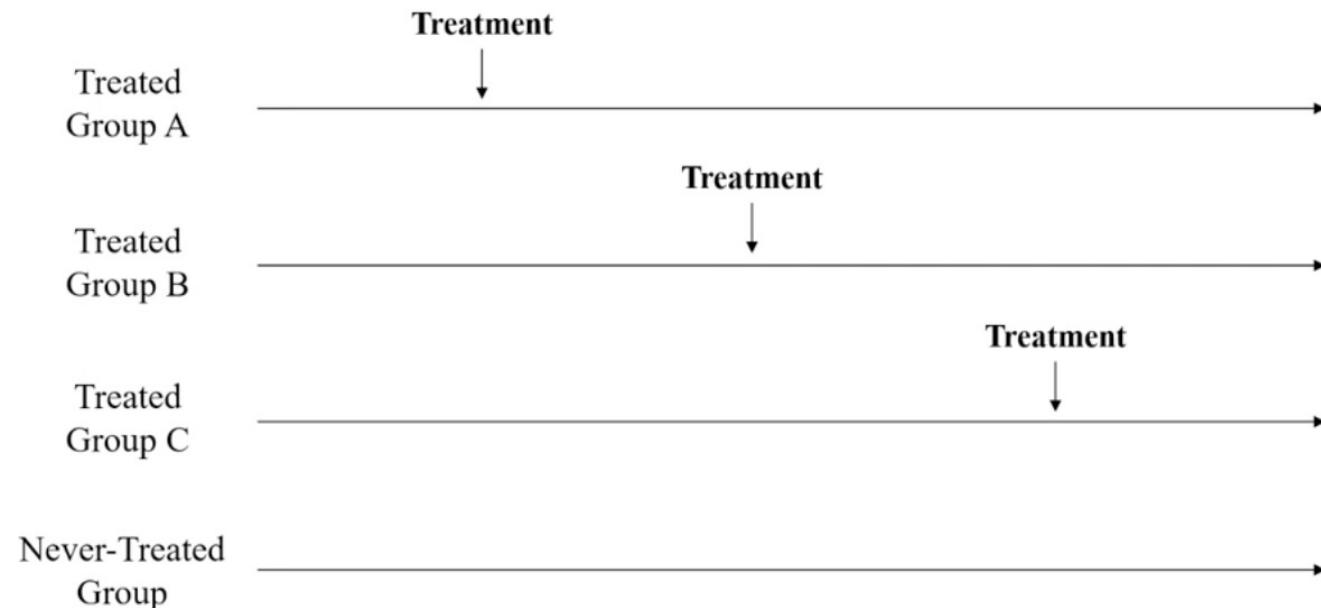
- post
 - treated
 - post:treated
 - C(region)
 - post:C(region)
 - treated:C(region)
 - post:treated:C(region)

처치의 시차 도입 (DID with Staggered Treatment)

기존의 이중차분법은 블록 디자인 기반으로, 동일 시점에 처치 전/후 데이터를 활용

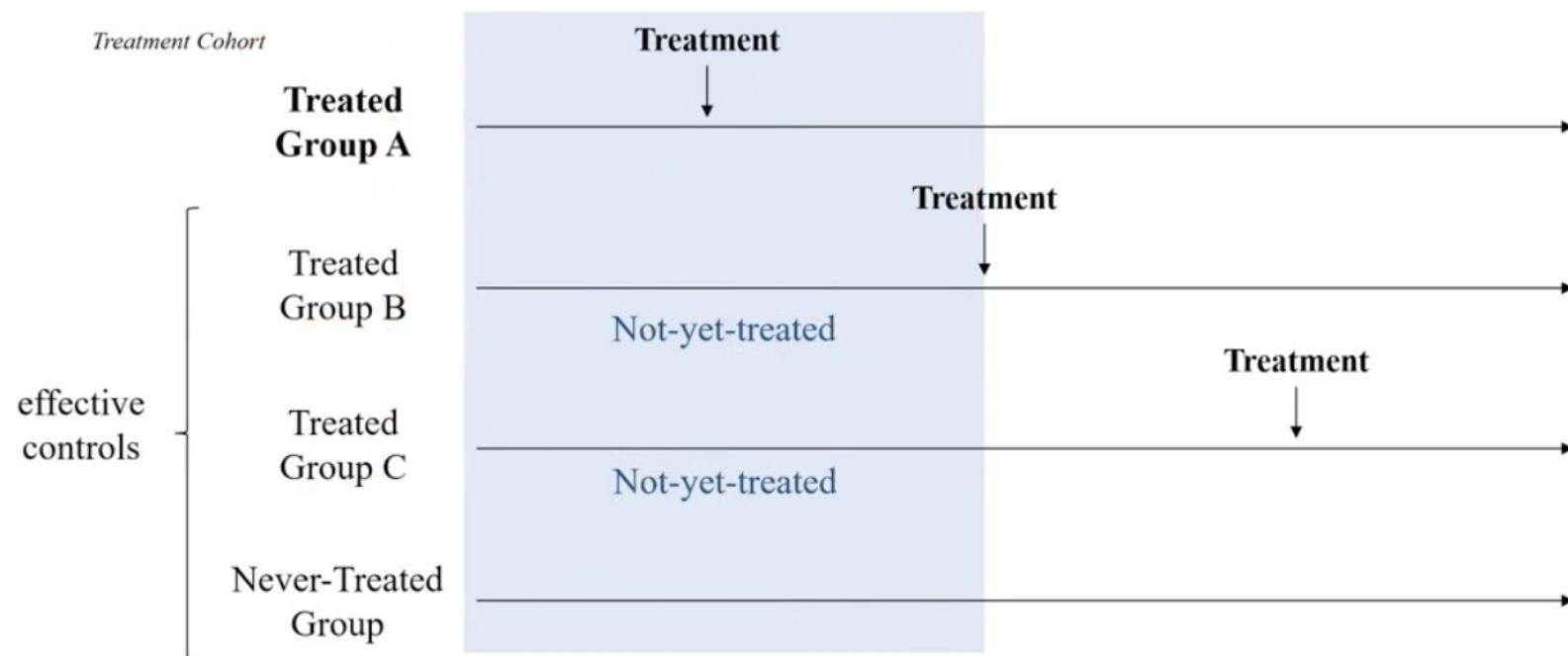
그룹별 처치를 받는 시점이 다른 경우에는 어떻게 접근하면 될까?

만약 코로나처럼 모든 사람이 처치받는 시점이 다르다면..?



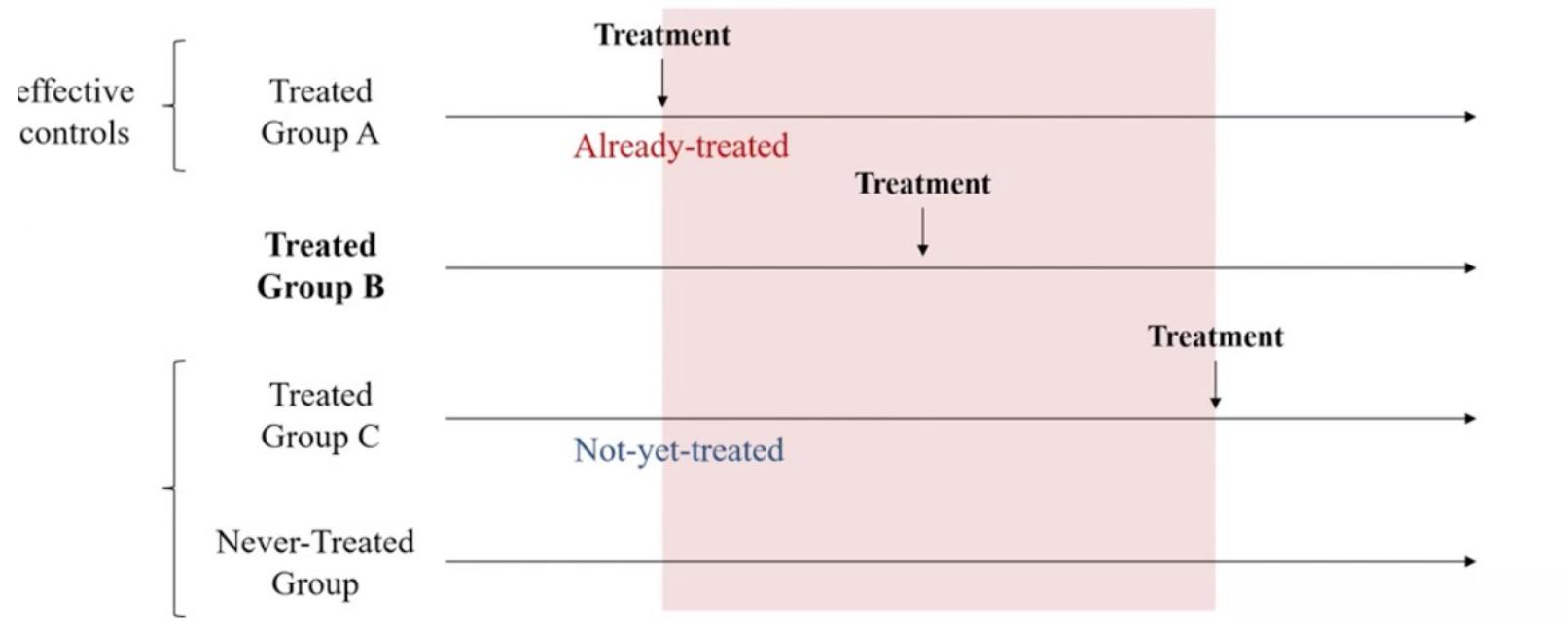
처치의 시차 도입 (DID with Staggered Treatment)

이중차분법의 기본 개념은 처치를 받은 그룹과 처치를 받지 않은 그룹의 처치 전후 차이를 보는 것이기 때문에 Never-Treated Group 뿐만 아니라, Treated B와 C도 아직 처치를 받지 않았기 때문에 좋은 대조군으로 활용할 수 있음



처치의 시차 도입 (DID with Staggered Treatment)

이중차분법의 기본 개념은 처치를 받은 그룹과 처치를 받지 않은 그룹의 처치 전후 차이를 보는 것이기 때문에
Never-Treated Group 뿐만 아니라, Treated B와 C도 아직 처치를 받지 않았기 때문에 좋은 대조군으로 활용할 수 있음



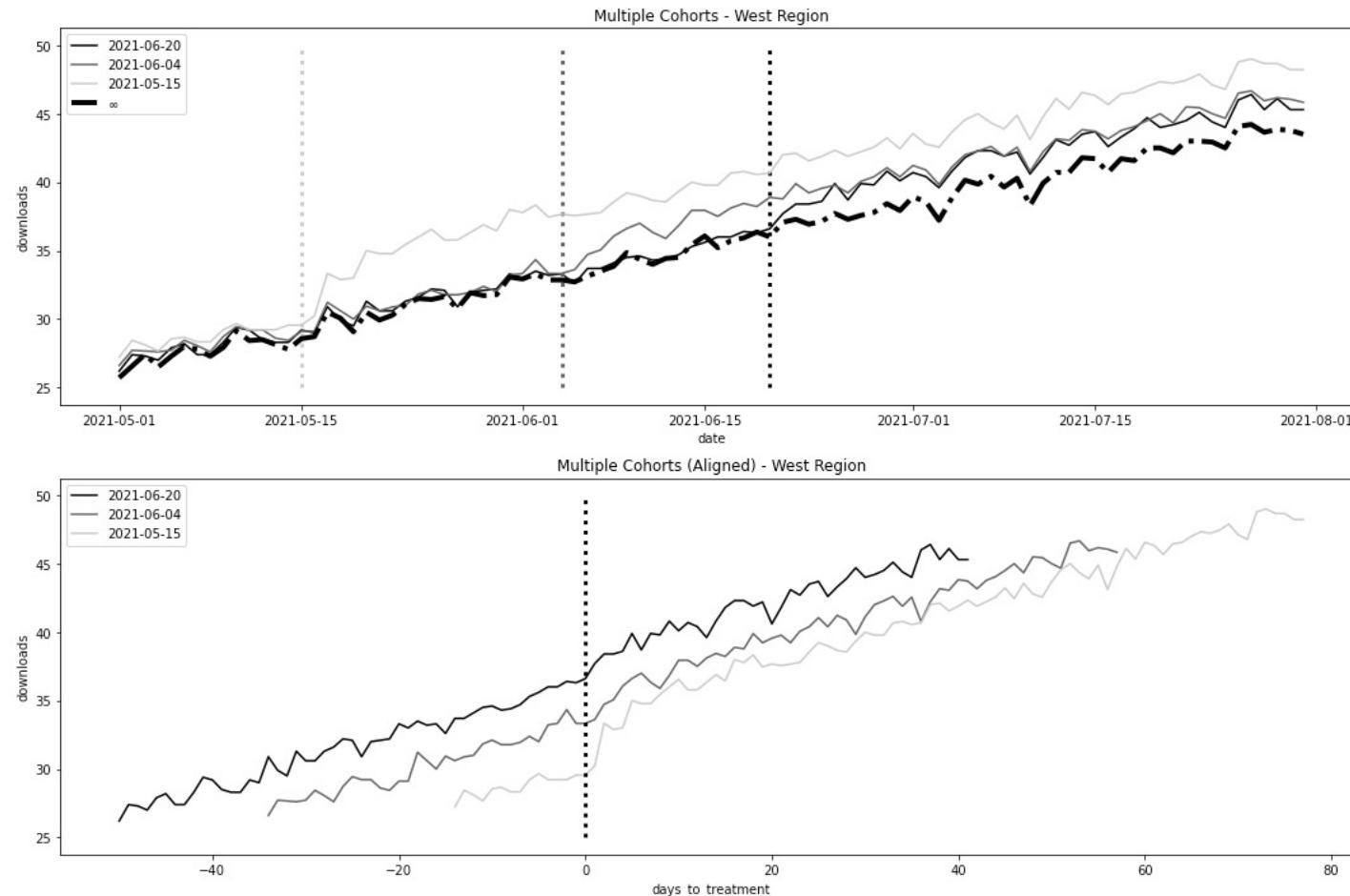
처치의 시차 도입 (DID with Staggered Treatment)

1. 처치를 받는 시점이 다른 그룹마다 코호트 그룹으로 분류



처치의 시차 도입 (DID with Staggered Treatment)

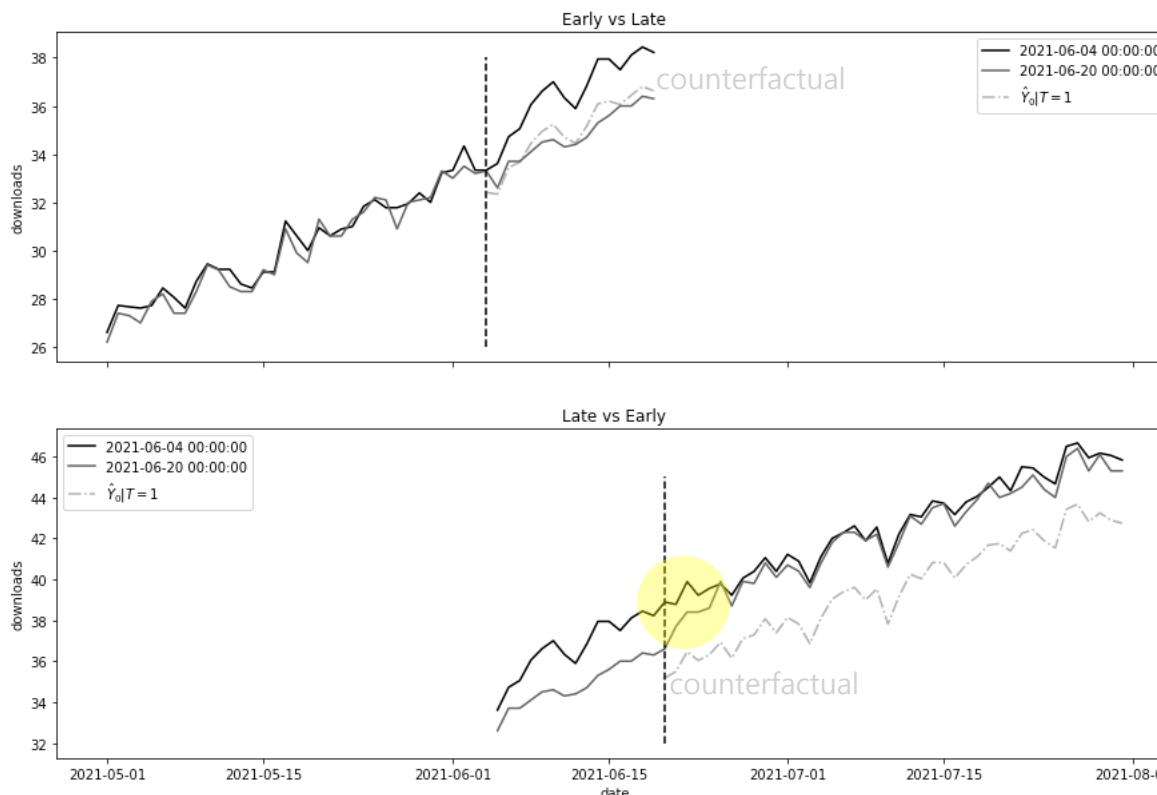
2. 처치를 받는 시점을 기준으로 Cohort 정렬



처치의 시차 도입 (DID with Staggered Treatment)

3. ATE 효과의 하향 편향 발생

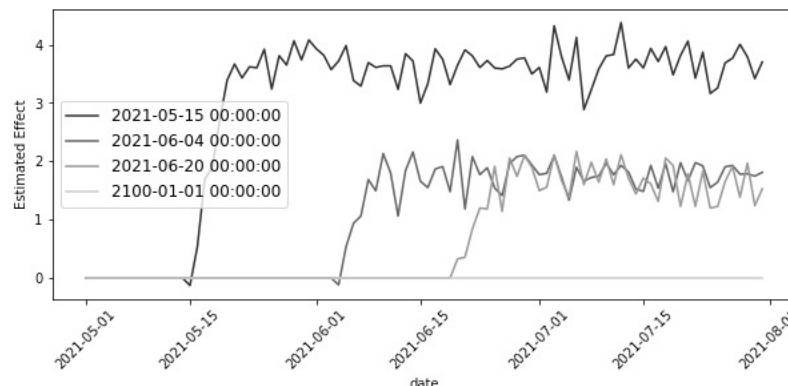
- 처치를 받고난 직후부터 처치 효과를 완전히 관찰 할 수 있을때 까지 시간이 걸림
- Cohort 0604의 경우 이미 처치를 받은 이후 일정 시간의 경과가 흘러 처치를 완전히 받고 있는 추세이지만, Cohort 0620은 처치를 받은 직후로 처치를 완전히 받기까지 일정 시간이 필요
- 이 과정에서 대조군에 대한 추세가 과대 추정되어 ATT 추정값은 하향 편향



처치의 시차 도입 (DID with Staggered Treatment)

4. 시간에 따른 이질적 효과의 처리

- 각 시간과 실험 대상에 대해 다른 weight를 적용 $Y_{it} = \tau_{it} W_{it} + \alpha_i + \gamma_t + eit$
- 실험 대상마다 처치효과를 가지게 되면 너무 많은 매개변수가 생성 → cohort별로 대상을 그룹화 $Y_{it} = \tau_{gt} W_{it} + \alpha_i + \gamma_t + eit$



- 각 DID comparison에 대해 시각적으로 표현 > Goodman-Bacon's decomposition

처치의 시차 도입 (DID with Staggered Treatment)

Local DID Model

- 처치를 전혀 받지 않은 그룹을 대조군으로 사용하여, 각 Cohort에 대해 하나의 DID 모델을 추정
결과를 가중평균으로 합산
- Callaway and Sant' Anna (2021)
 - Never treated group 과 Not yet treated group와의 비교를 사용
- Sun and Abraham (2021)
 - Never treated group과의 비교만을 통해서 효과를 추정

Effect of offline showrooms on demand generation and operational efficiency

연구 배경

- 온라인 기반의 온라인 쇼핑 업체인 Warby Parker의 오프라인 쇼룸 개장에 대한 효과를 분석하기 위한 연구
- 준실험 데이터를 사용하여 오프라인 쇼룸이 전체 수요와 온라인 채널 수요를 증가시키고, 운영 효율성을 개선하며, 고객의 쇼핑 행동에 긍정적인 영향을 미친다는 것을 밝혀내고자 함

처리 그룹: 쇼룸에서 반경 30마일 이내의 우편번호

통제 그룹: 이 반경 밖의 우편번호

DDD

Case 1: WWDC에 참석한 개발자들이 더 창의적인 결과물에 영향을 줄까?

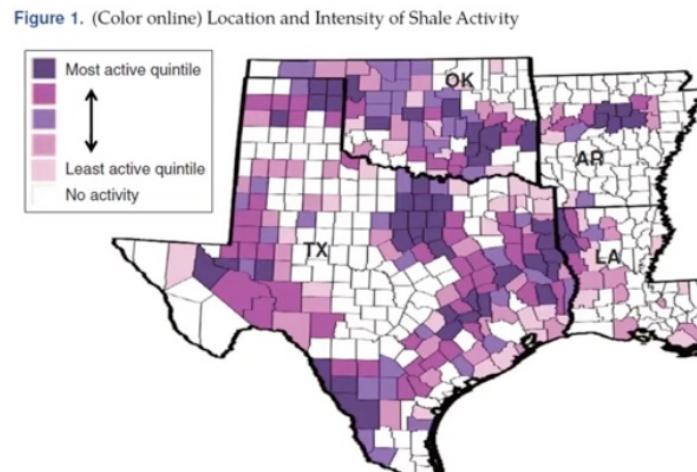
- WWDC를 참석한 개발자와 참석하지 않은 개발자 차이의 효과를 측정
- 이때 회사의 환경에 따라 참석 여부가 달라질 수 있다! > 이 부분을 Effect Modifier로 두로 이중차분의 차분을 진행!

		Change After Treatment			
		Pre-Treatment	Post-Treatment	Effect Modifier = 0	Effect Modifier = 1
Treatment Group	T_{Pre}	T_{Post}	$T_{0,Post} - T_{0,Pre}$	$T_{1,Post} - T_{1,Pre}$	
Control Group	C_{Pre}	C_{Post}	$C_{0,Post} - C_{0,Pre}$	$C_{1,Post} - C_{1,Pre}$	
		$DID = (T_{0,Post} - T_{0,Pre}) - (C_{0,Post} - C_{0,Pre})$		$DID = (T_{1,Post} - T_{1,Pre}) - (C_{1,Post} - C_{1,Pre})$	
		$DDD = [(T_{1,Post} - T_{1,Pre}) - (C_{1,Post} - C_{1,Pre})] - [(T_{0,Post} - T_{0,Pre}) - (C_{0,Post} - C_{0,Pre})]$			

DDD

Case 2: Shale Boom으로 인한 지역 경제의 활성화에 대한 영향력 분석

- Shale이 발견된 지역의 전후 비교를 진행, 지역별로 > Staggered treatment DID
- 초기자금이 많이 필요한 철강 등의 사업에서는 외부 투자가 더 많이 필요
 - > 외부 투자가 많이 필요한 산업의 전/후 차이 (DID)
 - > 외부 투자가 많이 필요하지 않은 산업의 전/후 차이 (DID)



Shale Boom **(Natural Shock)**

More deposit in local banks by oil firms

Access to Local Finance **(Treatment)**

- Industries with high external financing requirements
- Industries with low external financing requirements

감사합니다

Q&A