경영경제데이터분석

보충자료3

최 현 홍
(hongchoi@khu.ac.kr)

Contents

보충자료3

- PSM vs PSM+DID with lalonde 데이터

PSM vs PSM+DID with lalonde Data

Supplementary3

Recall) lalonde 데이터 설명

- **NSW 직업훈련 프로그램 이수가 미래 소득에 영향을 미치는가?**

DID를 배우고 해당 데이터를 다시 보니 어떤 생각이 드는가?

- **데이터셋 내 변수 목록**
  - **Treat**: NSW 참여자와 비참여자를 구분하는 더미 변수 (1: 참여)
  - Age: 참여자의 나이
  - Educ: 참여자의 교육 수준 (교육 년수)
  - Race: 참여자의 인종 (흑인, 백인, 히스패닉)
  - Married: 결혼 여부 (1: 결혼)
  - Nodegree: 고등학교 졸업 여부 (1: 학위 없음)
  - Re74: 프로그램 참여 전인 1974년 연 소득 (USD)
  - Re75: 프로그램 참여 전인 1975년 연 소득 (USD)
  - **Re78**: 프로그램 참여 후인 1978년 연 소득 (USD

Supplementary3

lalonde 데이터를 활용한 PSM

- 우리는 과거 lalonde 데이터를 활용, **PSM**을 이용하여 인과성 추론 분석을 수행하였음
  - 당시의 주요 분석 내용은 **처치군과 대조군의 78년 소득을 비교했더니 둘 사이의 차이가 유의미했다**는 것이었음 (p-value 0.084)
    - ✓ 처치군과 대조군의 매칭은 **나이, 교육, 결혼여부, 그리고 74년 소득**을 이용하여 이루어졌음 (다른 방식도 가능)
  - 즉, **NSW 참가 여부와 미래 높은 소득 사이에 인과성이 존재**한다고 주장할 수 있는 근거를 얻었음

- DID를 배우고 해당 데이터를 다시 보니 드는 생각:
  - 대상 지역이 **74년~78년 사이에 다양한 이유로 일자리가 늘었을 수도** 있지 않나?

Supplementary3

PSM의 분석 흐름

- **혼동변수를 적절히 고려하여 성향 점수 추정 및 매칭**
  - 예) Age, Educ, Married, Re74 → Treat
  - 다양한 성향점수 추정모형 및 매칭 방식 활용 가능
- **매칭 결과 검증**
  - 처치군과 대조군의 차이 (SMD, t-검정, 카이제곱검정 등)
  - 매칭된 처치군과 대조군의 성향점수 분포
- **분석 대상 인과관계(Treat → Re78) 검정**
  - 결과변수(Re78)가 연속변수이므로 t-검정 활용

Supplementary3

PSM + DID로 분석을 한다면?

- **PSM을 통해 매칭된 데이터**를 활용해보자
  - 병행추세 가정의 경우, **처치군과 대조군이 비슷한 특성을 가진 두 그룹(PSM으로 매칭됨)**이고 **해당 시차 동안 비슷한 외부 영향**을 받았을 것이니 성립한다고 가정하자

- 그런데, PSM을 통해 얻은 데이터를 **그대로 활용한 분석은 불가**
  - 처치 여부 Treat은 있지만, **시차에 따른 After 변수**가 없음
  - After 변수를 포함하려면 **하나의 응답자에 대해 2개의 행**을 만들어주어야 함 (처치 전/후)
  - 또한, 이제 **각 행이 특정 시점의 응답자**를 나타낸다면, **수입 변수 역시 새롭게 정의**될 필요가 있음
  - 예를 들어 **74년(처치 전) vs 78년(처치 후)**으로 데이터를 만든다면, After == 0인 경우 74년의 소득을, After == 1인 경우 78년의 소득을 활용하여 새로운 변수(revenue) 정의

Supplementary3

데이터 예시 (matched_data_lalonde_revised.csv)

기존 데이터 대비 행 2배 (각 행은 응답자 X 시점을 나타냄)
after 및 revenue 변수 추가

	treat	after	revenue	age	educ	race	married	nodegree	re74	re75	re78	distance	weights	subclass
NSW1	1	0	0	37	11	black	1	1	0	0	9930.046	0.23098	1	1
NSW2	1	0	0	22	9	hispan	0	1	0	0	3595.894	0.447041	1	98
NSW3	1	0	0	30	12	black	0	0	0	0	24909.45	0.488269	1	109
NSW4	1	0	0	27	11	black	0	1	0	0	7506.146	0.473447	1	120
NSW5	1	0	0	33	8	black	0	1	0	0	289.7899	0.475522	1	131
NSW6	1	0	0	22	9	black	0	1	0	0	4056.494	0.447041	1	142
NSW7	1	0	0	23	12	black	0	0	0	0	0	0.466564	1	153
NSW8	1	0	0	32	11	black	0	1	0	0	8472.158	0.488962	1	164
NSW9	1	0	0	22	16	black	0	0	0	0	2164.022	0.485498	1	175
NSW10	1	0	0	33	12	white	1	0	0	0	12418.07	0.226109	1	2
NSW11	1	0	0	19	9	black	0	1	0	0	8173.908	0.437842	1	13
NSW12	1	0	0	21	13	black	0	0	0	0	17094.64	0.465874	1	24
NSW13	1	0	0	18	8	black	0	1	0	0	0	0.429366	1	35
NSW14	1	0	0	27	10	black	1	1	0	0	18739.93	0.206006	1	46
NSW15	1	0	0	17	7	black	0	1	0	0	3023.879	0.42093	1	57
NSW16	1	0	0	19	10	black	0	1	0	0	3228.503	0.443286	1	68
NSW17	1	0	0	27	13	black	0	0	0	0	14581.86	0.484471	1	79
NSW18	1	0	0	23	10	black	0	1	0	0	7693.4	0.455588	1	90
NSW19	1	0	0	40	12	black	0	0	0	0	10804.32	0.519332	1	97

Supplementary3

PSM + DID 분석 결과

```
newdata <- read.csv("matched_data_lalonde_revised.csv")

mynewmodel <- lm(revenue ~ treat + after + treat:after, data = newdata)

summary(mynewmodel)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1916.6      423.9   4.522 7.15e-06 ***
treat          179.0      599.4   0.299    0.765
after         3201.7      599.4   5.341 1.23e-07 ***
treat:after   1051.8      847.7   1.241    0.215
```

Conducting DID Analysis

PSM + DID 분석 결과 – 모수 추정치

- 모수 추정 결과 $Y = \beta_0 + \beta_1 Treat + \beta_2 After + \beta_3 Treat \times After$

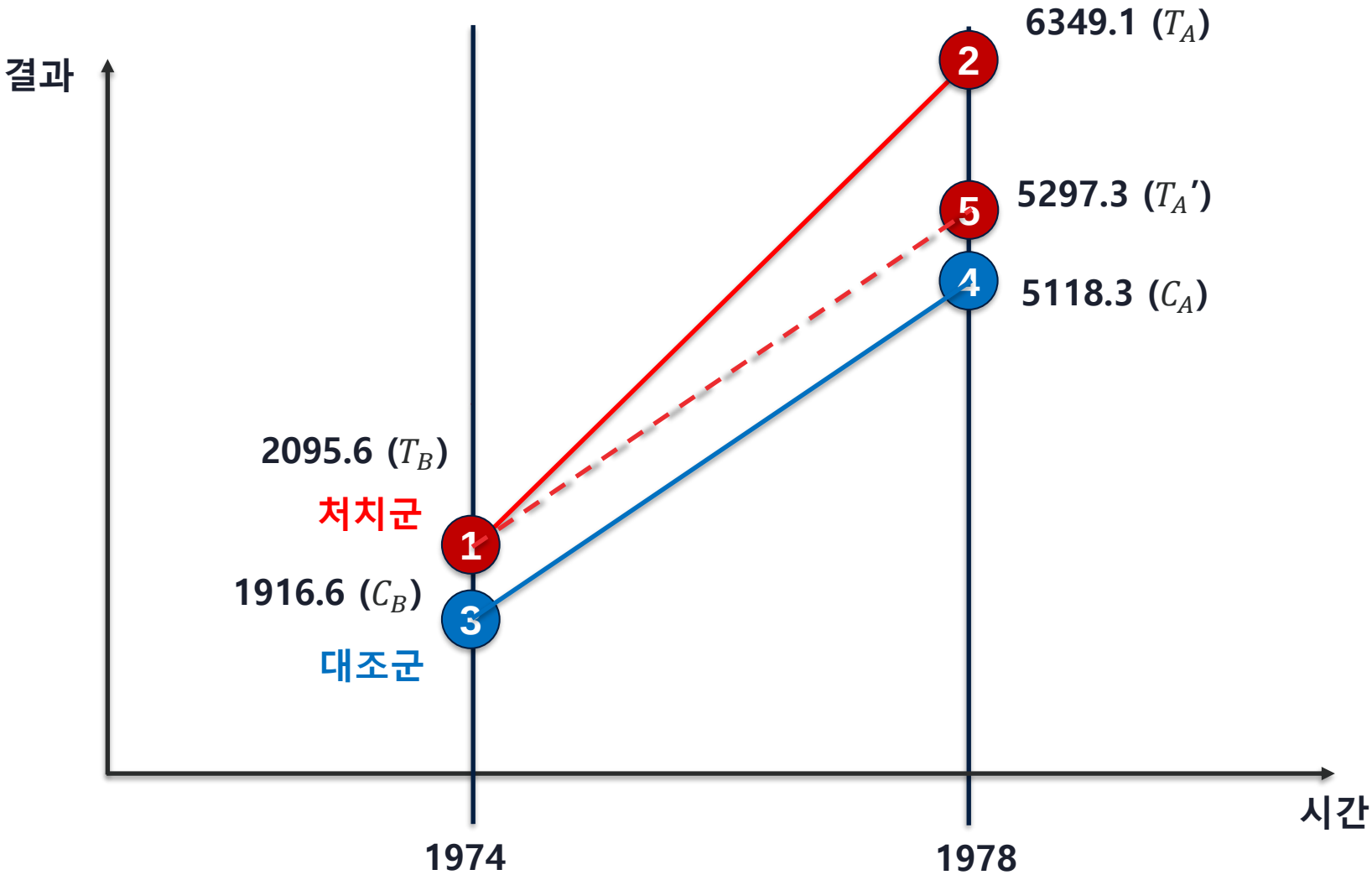
Variables	Coefficient	Std.Error	P-value	
Intercept (β_0)	1916.6	423.9	0.000	
Treat (β_1)	179.0	599.4	0.765	유의하지 않음
After (β_2)	3201.7	599.4	0.000	1% 유의수준에서 유의
Treat:After (β_3)	**1051.8**	**847.7**	**0.215**	유의하지 않음

- DID Estimator($\boldsymbol{\beta_3}$)의 추정치가 유의하지 않음
  - ✓ 즉, **순수한 처리 효과가 통계적으로 유의하지 않음**
  - ✓ 단 양수로 나타나긴 했음
  - ✓ (사실 PSM 분석 당시에도 p-value가 0.085였음)

- **물론 이러한 결과는 다른 매칭 결과를 활용했을 경우에 또 달라질 수 있음**

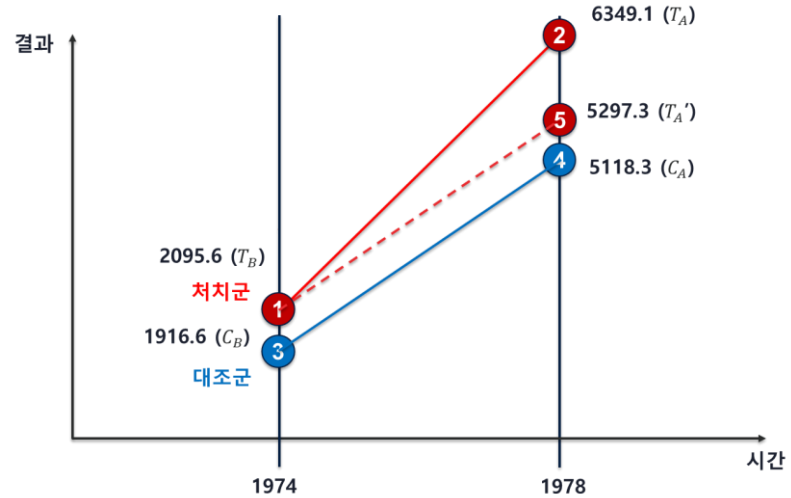
Conducting DID Analysis

PSM + DID 분석 결과 – 결과 시각화



Supplementary3

PSM과 DID 결과의 비교



PSM 당시 t-검정 결과

```
        Welch Two Sample t-test

data:  re78 by treat
t = -1.7292, df = 333.67, p-value = 0.08471
alternative hypothesis: true difference in
95 percent confidence interval:
 -2630.9570   169.3691
sample estimates:
mean in group 0 mean in group 1
       5118.350        6349.144
```

- PSM을 활용한 연구에서는 **매칭된 처치군과 대조군의 1978년 소득을 비교**했음. 즉, **②에서 ④를 뺀 값**이 유의한 값인지를 살핀 것
  - **대조군**과의 비교

- PSM+DID를 활용한 연구에서는 병행추세가정을 바탕으로 시간의 변화에 따른 효과를 제거했으므로, **②에서 ⑤를 뺀 값이 유의한 값**인지를 살핀 것
  - 대조군의 추세를 통해 구해진 **반사실**과의 비교

Supplementary3

어떤 결과가 맞는 걸까?

- 두 결과 중 어떤 결과가 맞는 걸까?
  - 그래서 인과관계가 있는 건데 없는 건데?

- **둘 중 무엇이 맞는 결과라고 혹은 인과관계가 있거나 없다고 단정지어 말 할 수 없음**
  - 자료의 수집, 변수 측정, 모형의 적합성 등에서 다양한 오류와 불확실성이 존재함
  - 우리가 판단의 기준으로 삼는 p-value 역시 확률적인 지표
  - **단정지어 말하는 사람을 조심**하자

Supplementary3

그럼 어떻게 설득할 수 있을까?

- 취할 수 있는 접근법 중 하나: **다양한 분석에서 주장하고 싶은 결과가 안정적**으로 나타난다는 것을 보임
  - 예) 기본모형 + 다양한 변수 포함/제외 모형 비교

선행연구 예)
런던 혼잡구역 진입 제한 → 교통사고 감소

기본 모형

Table 3
Basic DID models for car casualties.

Basic DID models	Model 4 car all casualties		Model 5 car KSI		Model 6 car slight injured	
	Coef. (std. err.)		Coef. (std. err.)		Coef. (std. err.)	
CCZoon	−3.81E−01	(4.92E−10)*	7.20E−01	(2.49E−09)*	−4.65E−01	(4.87E−10)*
CCYear	−2.22E−01	(2.18E−11)*	−2.01E−01	(2.18E−14)*	−2.23E−01	(2.92E−11)*
CCYear × CCZone	−3.17E−02	(5.43E−11)*	−1.10E−01	(4.34E−10)*	−1.19E−02	(1.12E−10)*
Constant	3.94E+00	(4.92E−10)*	6.93E−01	(3.47E−14)*	3.90E+00	(4.87E−10)*
Obs	244		244		244	
BIC	489.39		463.03		487.79	
Likelihood-ratio test of alpha = 0:	chibar2(01) = 2245.44		chibar2(01) = 95.46		chibar2(01) = 2052.78	
	Prob ≥ chibar2 = 0.000		Prob ≥ chibar2 = 0.000		Prob ≥ chibar2 = 0.000	

* Figures are significant at: 99%.

결과가 3개인 이유:
종속변수(사고)를 다르게 설정
-모든 사고, 사망 및 중상, 경상

변수 포함 모형 (변수 조합을 다르게 할 수도 있음)

Table 4
Full DID models for cycle casualties.

Full DID models	Model 7 bicycle all casualties (Poisson)		Model 8 bicycle KSI (Poisson)		Model 9 bicycle slight injured (Poisson)	
	Coef. (std. err.)		Coef. (std. err.)		Coef. (std. err.)	
CCZoon	1.45E+00	(3.71E−01)*	5.06E−01	(6.38E−02)*	1.63E+00	(4.33E−01)*
CCYear	−1.14E−01	(4.77E−04)*	−1.73E−01	(1.71E−02)*	−1.01E−01	(7.11E−03)*
CCYear × CCZone	1.25E−01	(4.76E−03)*	2.67E−02	(2.41E−02)	1.27E−01	(1.20E−03)*
Resident population	1.14E−03	(7.83E−04)	1.29E−04	(5.42E−04)	1.34E−03	(8.51E−04)
Resident population aged 0–15	−1.93E−04	(1.71E−03)	−1.53E−03	(8.09E−04)***	2.91E−06	(1.92E−03)
Resident population aged 16–59	2.39E−04	(4.26E−04)	−5.02E−04	(1.39E−03)	3.77E−04	(3.46E−04)
Percentage of resident population aged 0–15	6.77E+00	(2.27E+00)*	4.27E+00	(1.36E+00)*	7.27E+00	(2.31E+00)*
Percentage of resident population aged 16–59	5.52E+00	(2.53E+00)**	4.95E+00	(1.59E+00)*	5.71E+00	(2.64E+00)**
Employee population	5.70E−06	(1.72E−06)*	2.55E−06	(1.78E−06)	6.24E−06	(1.63E−06)*
Land area	3.02E−02		9.50E−04	(2.18E−02)	3.85E−02	(6.16E−03)*
Employee population density	−3.41E−07	(8.63E−07)	−3.84E−06	(3.60E−07)*	8.11E−08	(1.08E−06)
Resident population density	−5.79E−05	(1.30E−05)*	−7.08E−05	(3.28E−05)	−5.42E−05	(1.91E−05)*
Length of minor road	−3.34E−05	(1.63E−05)**	−3.21E−06	(2.01E−05)	−4.08E−05	(1.40E−05)*
Length of motor road	−5.16E−05	(3.92E−05)	−4.63E−06	(5.31E−06)	−6.21E−05	(4.20E−05)
Length of A-road	1.15E−06	(3.78E−05)	5.36E−05	(3.76E−05)	−9.09E−06	(3.56E−05)
Length of B-road	−1.15E−04	(2.26E−05)*	−2.37E−04	(8.69E−05)*	−8.94E−05	(7.91E−06)*
Density of minor road	3.39E−05	(3.55E−06)*	−1.86E−05	(6.72E−05)	4.30E−05	(1.26E−05)*
Density of motor road	4.28E−04	(1.68E−03)	1.36E−03	(1.29E−04)*	−7.81E−05	(1.85E−03)
Density of A-road	−1.21E−03	(1.50E−02)	−2.72E−02	(9.21E−03)*	3.04E−03	(1.41E−02)
Density of B-road	2.52E+00	(1.03E+00)**	7.65E+00	(1.43E+00)*	1.53E+00	(1.42E+00)
Count of junctions	3.65E−03	(1.04E−03)*	2.42E−03	(1.36E−04)*	4.06E−03	(9.77E−04)*
Count of roundabout	2.11E−02	(9.76E−02)	−1.51E−01	(1.02E−01)	5.65E−02	(9.55E−02)
IMDscore	1.78E−02	(1.58E−03)**	2.17E−02	(1.69E−02)	1.69E−02	(1.58E−03)*
PPE	1.12E−03	(6.03E−05)*	1.62E−03	(5.84E−04)*	1.02E−03	(2.50E−05)*
Constant	−6.22E+00	(2.88E+00)*	−6.26E+00	(2.43E+00)*	−6.78E+00	(2.98E+00)*
Obs	244		244		244	
BIC	307.52		339.29		305.37	
Likelihood-ratio test of alpha = 0:	chibar2(01) = 0.0E+00		chibar2(01) = 0.0E+00		chibar2(01) = 0.0E+00	
	Prob ≥ chibar2 = 0.500		Prob ≥ chibar2 = 0.500		Prob ≥ chibar2 = 0.500	

* Figures are significant at: 99%.
** Figures are significant at: 95%.
*** Figures are significant at: 90%.

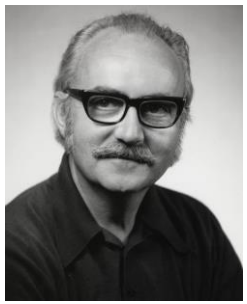
Li, H., Graham, D. J., & Majumdar, A. (2012). The effects of congestion charging on road traffic casualties: A causal analysis using difference-in-difference estimation. *Accident Analysis & Prevention*, *49*, 366-377.

Supplementary3

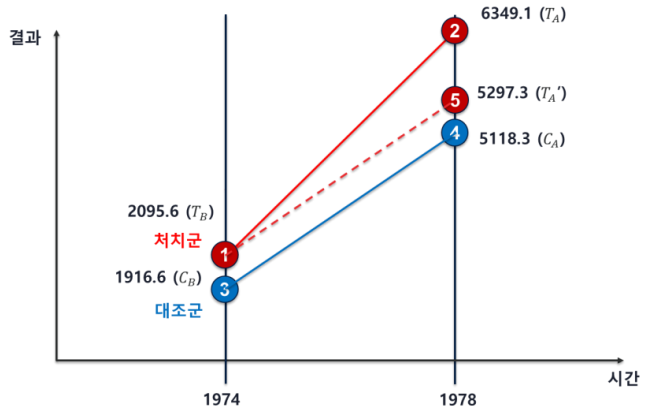
어떤 결과가 맞는 걸까?

- 사회과학 분야의 많은 문제에서 **"참"은 존재하지만 알기 어려움**
  - **Only possible to argue**

- **All Models are Wrong, but Some are Useful**
  - All models are **approximations**
  - The practical question is: **how wrong do they have to be not to be useful**



George E. P. Box



Supplementary3

https://scholar.google.com/

Google 학술검색

모든 언어 한국어 웹

거인의 어깨에 올라서서 더 넓은 세상을 바라보라 - 아이작 뉴턴