

[Data analysis] 시간의 흐름에 따라 패널 데이터 분석 (데이터 분석의 힘 chapter.5) —
나무늘보의 개발 블로그

노트북: 첫 번째 노트북

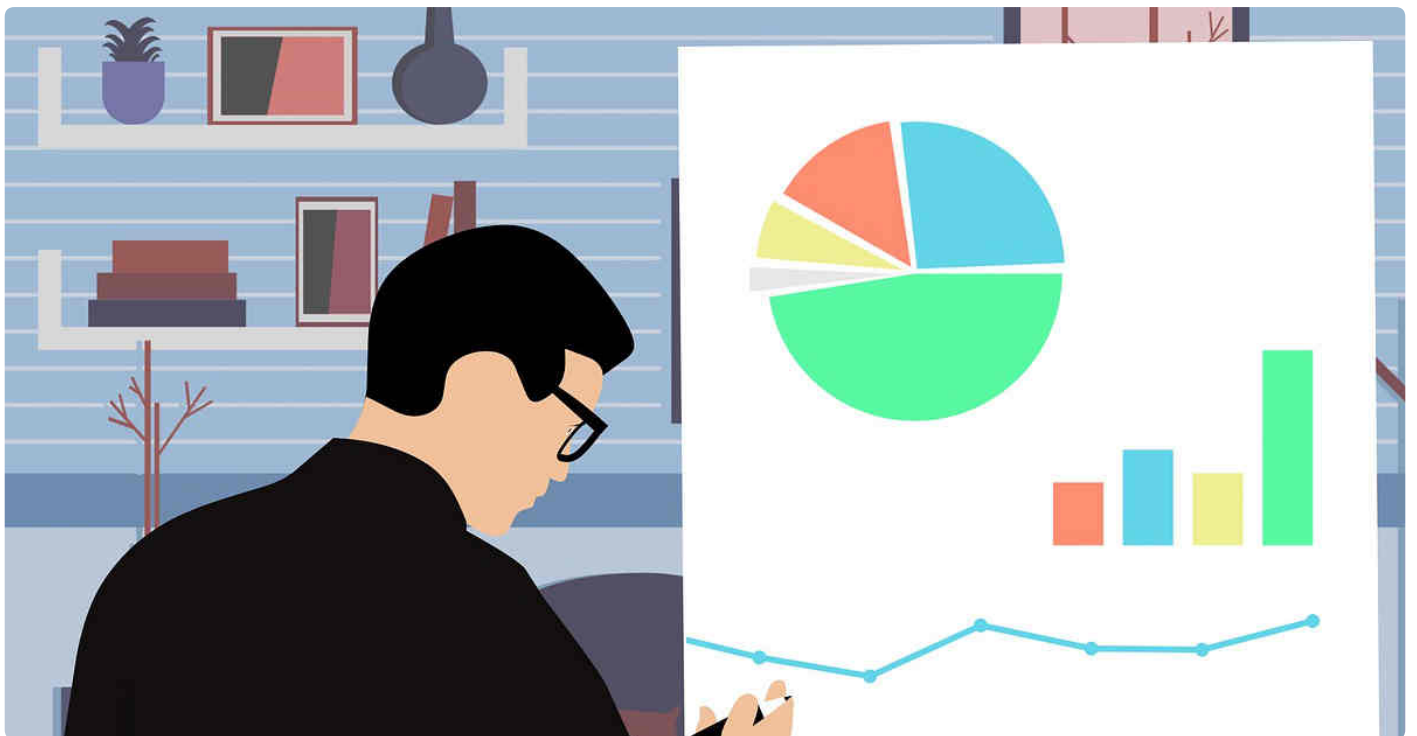
만든 날짜: 2021-03-27 오후 6:40

URL: <https://continuous-development.tistory.com/234>

Data scientist/Data Science

[Data analysis] 시간의 흐름에 따라 패널 데이터 분석 (데이터 분석의 힘 chapter.5)

2021. 3. 27. 18:40 수정 삭제 공개



패널데이터란?

복수의 집단에 대해 복수의 기간에 걸쳐 수집한 데이터를 가리킨다.

ex) 여러 기업의 재무 데이터를 복수의 기간동안 관측하거나, 지역별 부동산 데이터를 몇 년간 관측하는 것

덴마크의 개인 납세 데이터를 이용한 연구

덴마크의 경우 1991년 세제가 개혁되면서 연간소득이 1억 2000만원 이상인 외국인 노동자의 소득세가 대폭 줄어들었다.

클레벤 교수팀은 외국인 노동자를 개인 집단으로, 영향을 받지 않은 외국인 노동자를 비교집단으로 나누고 정책의 효과를 분석하기로 했다.

1991년 개혁 이후로 집단의 이민자 수가 비교집단에 비해 크게 늘었다. 1980년에 비해 2배 늘었고 2005년에는 4배 가까이 늘었다.

패널 데이터 분석의 사고 방식

패널 데이터 분석은 RCT가 불가능할 경우 사용 할 수 있는 자연 실험 방법 중 하나이다.

1991년 전후 복수 기간의 데이터를 수집했다고 하자. 개입 집단과 비교 집단이 있을때

두 집단은 1991년 이전에는 같은 움직임을 보이다가 1991년 이후에는 매우 다른 움직임을 보인다. 이를 근거로 1991년의

세제 개혁이 이민자 수에 영향을 주었다는 인과관계를 주장할 수 있다는 것이 패널 데이터 분석의 기본적인 사고 방식이다.

개입효과는 개입 후의 차와 개입전의 차를 빼서 개입 효과를 구한다.

이 분석 방법은 이중차분법이라고도 불리운다. 또는 시간에 따라 변하지 않는 고정적인 효과를 생략한다는 의미로 고정효과 추정법이라고 불리운다.

패널 데이터 분석에 필요한 평행 트렌드 가정

패널 데이터 분석의 가정 - 만약 개입이 일어나지 않았다면 개입집단의 평균값과 비교 집단의 평균값은 평행한 추이를 보인다.

평행 트렌드 가정을 데이터로 입증할 수가 없으므로 데이터 분석자는 이 가정이 성립하리라는 증거를 최대한 열거해야 한다.

데이터 분석자가 제공할 수 있는 정보

개입 이전의 데이터를 모아 개입집단과 비교집단 사이에 평행 트렌드 가정이 성립하는지 조사해야 한다.

개입 이후 개입집단에만 영향을 미친 다른 사건이나 변수가 없었는지 확인해야 한다.

평행 트렌드 가정은 어떤 때 무너질까

개입 요소 이외에는 다른 개입이 없어야 한다.

패널데이터 분석의 강점과 약점

강점

x가 y에 미친 인과관계를 분석할 수 있다.

평행 트렌드 가정이 지켜지는 한, 개입집단과 비교집단 간에 원래 차이가 있었어도 문제가 되지 않는다는 점

개입을 받은 모든 대상의 개입 효과에 대해 인과관계를 측정할 수 있다는 점

약점

많은 상황에서 성립하지 않는다.

복수 기간에 걸쳐 개입집단과 비교집단의 데이터를 수집해야 한다.

요약

패널 데이터 분석을 이용하려면 복수의 집단에 대해 복수의 기간에 걸쳐 데이터를 수집할 수 있어야 한다.

패널 데이터 분석의 원칙

개입을 전후에서 개입집단과 비교집단 양쪽의 데이터를 입수할 수 있는지 확인한다.

평행 트렌드 가정이 성립하는지 검증한다.

평행 트렌드 가정이 성립할 가능성이 높다면 두 집단의 평균값 추이를 그래프로 그림으로써 개입효과의 평균값을 측정한다.

패널 데이터 분석의 강점

필요한 데이터만 확보된다면 RD 디자인이나 집군분석 이상으로 광범위하게 이용 할 수 있다.

결과를 그래프로 보여줄수 있어 쉽고 투명한 분석이 가능하다.

개입집단 전체에 대한 개입 효과를 분석할 수 있다. 분석 대상이 제한된 RD 디자인이나 집군분석에 비해 강점이다.

패널 데이터 분석의 약점

분석에 필요한 가정이 성립할 것이라는 근거를 제시할 수는 있지만 입증할 수는 없다. 이는 RCT와 비교했을 때 큰 약점이다.

RD 디자인 이나 집군분석에 필요한 가정에 비해 평행 트렌드 가정은 매우 까다로운 가정이며 실제로는 성립하지 않는 경우도 많다.

'Data scientist > Data Science' 카테고리의 다른 글

[Data analysis] 시간의 흐름에 따라 패널 데이터 분석 (데이터 분석의 힘 chap...

[Data Science] 데이터 사이언스 개념 - 10.딥러닝

[Data Science] 데이터 사이언스 개념 - 9.신경망이 기초

[Data Science] 데이터 사이언스 개념 - 8.토픽 모델 / 네트워크 분석□

[Data Science] 데이터 사이언스 개념 - 7.비지도 학습□

[Data Science] 데이터 사이언스 개념 - 6.분류문제□

데이터 분석의 힘

패널 데이터 분석

패널데이터 분석



나아무늘보

혼자 끄적끄적하는 블로그 입니다.