

Week 1. Statistical Models and Social Science

Fox., John. 2016. *Applied Regression Analysis and Generalized Linear Models*, 3rd Ed.

Ch. 1. Statistical Models and Social Science

Sanghoon Park

Department of Political Science
University of South Carolina

2020-09-13

Statistical Models and Social Reality

“모든 모델은 틀렸다(wrong), 다만 몇몇 모델만이 유용할 뿐이다” (Box 1979: 202).

- 복잡한 현실을 특정한 모형으로 모두 설명하는 것은 불가능
- 사회 구조 혹은 다른 체계적 요인들이 개인의 선택을 조건짓고, 제약
- 특정한 개인을 예측하는 것이 아니라 일반적 경향성을 포착하고자 함.

사회과학에는 수많은 ‘추상적인’ 이론들이 존재

- 추상적인 사회이론들이 만약 복잡한 현실세계를 이해하는 데 유용한가?
- 만약 유용하다면, 이와 같은 이론(논리)과 현실(경험)을 어떻게 연결할 수 있을까?

Statistical Models and Social Reality

이론은 특정 현상들을 '정의'(define)하고, 그 정의된 현상들 간의 관계를 서술, 설명, 및 예측

- 우리는 이 추상적인 이론을 구성하는 각 부분들이 무엇인지 설명하고, 평가하거나 측정하는 과정을 통해 이론과 현실을 연결(bridging)

- 예) 교육 수준 \uparrow \rightarrow 소득 수준 \uparrow

- ▶ 교육 수준과 소득 수준을 어떻게 정의할 것인가?
- ▶ 교육 수준과 소득 수준의 관계를 어떻게 설명할 것인가?
- ▶ 교육 수준과 소득 수준을 실제로 어떻게 측정할 것인가?
- ▶ 이 관계에 영향을 미치는 다른 요인은 존재하는가? 존재한다면 무엇인가?

Statistical Models and Social Reality

이론은 특정 현상들을 '정의'(define)하고, 그 정의된 현상들 간의 관계를 서술, 설명, 및 예측

- 우리는 이 추상적인 이론을 구성하는 각 부분들이 무엇인지 설명하고, 평가하거나 측정하는 과정을 통해 이론과 현실을 연결(bridging)
 - 예) 교육 수준 \uparrow \rightarrow 소득 수준 \uparrow
 - ▶ 교육 수준과 소득 수준을 어떻게 정의할 것인가?
 - ▶ 교육 수준과 소득 수준의 관계를 어떻게 설명할 것인가?
 - ▶ 교육 수준과 소득 수준을 실제로 어떻게 측정할 것인가?
 - ▶ 이 관계에 영향을 미치는 다른 요인은 존재하는가? 존재한다면 무엇인가?

Statistical Models and Social Reality

이론은 특정 현상들을 '정의'(define)하고, 그 정의된 현상들 간의 관계를 서술, 설명, 및 예측

- 우리는 이 추상적인 이론을 구성하는 각 부분들이 무엇인지 설명하고, 평가하거나 측정하는 과정을 통해 이론과 현실을 연결(bridging)
 - 예) 교육 수준 \uparrow \rightarrow 소득 수준 \uparrow
 - ▶ 교육 수준과 소득 수준을 어떻게 정의할 것인가?
 - ▶ 교육 수준과 소득 수준의 관계를 어떻게 설명할 것인가?
 - ▶ 교육 수준과 소득 수준을 실제로 어떻게 측정할 것인가?
 - ▶ 이 관계에 영향을 미치는 다른 요인은 존재하는가? 존재한다면 무엇인가?

Statistical Models and Social Reality

통계모델: 복잡한 현실을 단순하게(simplified) 서술한 것 (Fox, 2016, 2)

- 고용노동자들에 대한 Large-N 표본의 설문조사 자료
- 교육 수준, 성별, 인종, 거주지역 등과 같은 특성들로 개인의 소득에 대한 회귀분석을 시행하였다고 가정

과연 위의 모델로 개개인의 소득 수준을 완벽하게 설명할 수 있는가? 불가능함.

- 효과(effect): 관측하여 변수로 만들어 모델에 포함한 것들이 종속변수(소득)의 '체계적'(systematic) 변화에 미치는 영향
- 잔차(residuals): 관측한 것들로 설명하지 못하는 불확실성, 변수들의 효과로 설명하지 못하는 종속변수의 변화

Statistical Models and Social Reality

통계모델: 복잡한 현실을 단순하게(simplified) 서술한 것 (Fox, 2016, 2)

- 고용노동자들에 대한 Large-N 표본의 설문조사 자료
- 교육 수준, 성별, 인종, 거주지역 등과 같은 특성들로 개인의 소득에 대한 회귀분석을 시행하였다고 가정

과연 위의 모델로 개개인의 소득 수준을 완벽하게 설명할 수 있는가? 불가능함.

- 효과(effect): 관측하여 변수로 만들어 모델에 포함한 것들이 종속변수(소득)의 '체계적'(systematic) 변화에 미치는 영향
- 잔차(residuals): 관측한 것들로 설명하지 못하는 불확실성, 변수들의 효과로 설명하지 못하는 종속변수의 변화

Statistical Models and Social Reality

통계모델: 복잡한 현실을 단순하게(simplified) 서술한 것 (Fox, 2016, 2)

- 고용노동자들에 대한 Large-N 표본의 설문조사 자료
- 교육 수준, 성별, 인종, 거주지역 등과 같은 특성들로 개인의 소득에 대한 회귀분석을 시행하였다고 가정

과연 위의 모델로 개개인의 소득 수준을 완벽하게 설명할 수 있는가? 불가능함.

- 효과(effect): 관측하여 변수로 만들어 모델에 포함한 것들이 종속변수(소득)의 '체계적'(systematic) 변화에 미치는 영향
- 잔차(residuals): 관측한 것들로 설명하지 못하는 불확실성, 변수들의 효과로 설명하지 못하는 종속변수의 변화

Statistical Models and Social Reality

통계모델: 복잡한 현실을 단순하게(simplified) 서술한 것 (Fox, 2016, 2)

- 고용노동자들에 대한 Large-N 표본의 설문조사 자료
- 교육 수준, 성별, 인종, 거주지역 등과 같은 특성들로 개인의 소득에 대한 회귀분석을 시행하였다고 가정

과연 위의 모델로 개개인의 소득 수준을 완벽하게 설명할 수 있는가? 불가능함.

- 효과(effect): 관측하여 변수로 만들어 모델에 포함한 것들이 종속변수(소득)의 '체계적'(systematic) 변화에 미치는 영향
- 잔차(residuals): 관측한 것들로 설명하지 못하는 불확실성, 변수들의 효과로 설명하지 못하는 종속변수의 변화

Statistical Models and Social Reality

통계모델의 잔차는 클 수도, 작을 수도 있음.

- 하지만 잔차가 아무리 작다라고 하더라도 모델이 완벽하게 현실을 설명·예측할 수 있다고 할 수 없음.
- 통계모델은 근본적으로 서술적(descriptive) (Fox, 2016, 3)
 - 어디까지나 논리적인 기대(이론)와 일치하는 방향으로 경험적인 관측(데이터)가 배열되어 있는지를 보여줄 뿐임.
 - 따라서 통계모델은 항상 불확실성을 내재하지만 그럼에도 변수들의 체계적 효과, 유의미한 측면을 보여줌.

Statistical Models and Social Reality

정리하자면, 통계모델은 그 자체로 현실을 직접적으로 보여주는 것이 아니라 간단하게 유의미한 특성들만을 구조화하여 보여줄 뿐

- 구조적 조응성을 가지지만 기능적 조응성을 가지지는 않음.
- 예) 모형 비행기를 통해 실제 비행기의 구조를 이해하는 데 도움을 얻을 수 있지만, 그렇다고 모형 비행기가 실제로 날 수 있는 것은 아님.

통계모델의 실용성(practice)과 정확성(accuracy)

- 데이터(현실)을 정확하게 보여주지 못하는 모델은 쓸모없지만
- 정확하게 현실을 보여주는 모델이라고 하더라도 모두 쓸모 있는 것은 아님.
 - 통계모델 자체는 굉장히 정교하고 정확하더라도 실제 현실을 설명하는 데 아무 도움이 되지 않을 수 있음.
 - 모델 이전에 이론 수립(theory-building)이 중요한 이유

Observation and Experiment

통계모델의 구분

- 실험(experimental) 자료
 - 연구자가 설명변수들을 직접적으로 통제 (randoly assigned)
 - 무작위(randomized) 실험을 통한 인과추론
 - 처치(treatment)를 제외한 나머지 요인들을 모두 무작위로 통제
- 관측(observational) 자료
 - 설명변수들의 값은 연구자에 의해 관측된 결과
 - 설명변수들의 값에 따른 종속변수의 '평균'의 변화 (differences in mean)
 - 따라서 인과추론은 정당화되지 않음

Observation and Experiment

관측 자료를 통한 통계모델이 인과관계(causality)을 담보하지 않는 이유

- 적절한 변수를 모델에서 누락했을 경우
 - 관측되지 않은 혼재변수(confounding variables)가 존재
 - 만약 종속변수에 영향을 미치는 변수인데 모델에 들어가지 않았다면 통제가 충분하지 못했다는 것
 - 만약 누락된 변수가 다른 설명변수들에 영향을 미치는 변수라면 간접적인 방식으로 종속변수에 영향을 줄 수 있으므로 마찬가지로 통제가 충분히 이루어지지 않았다는 것
- 적절하지 않은 변수를 모델에 포함했을 경우
- 통제가 제대로 이루어졌다는 확신을 할 수 없기 때문에 관측자료를 바탕으로 한 통계분석을 인과적으로 해석하는 데 한계가 존재
 - 상대적으로 실험자료는 관측자료에 비해 통제가 용이
 - 그러나 설명변수의 식별이라는 문제에 있어서는 마찬가지로 불확실성이 내재
 - 어떠한 방법으로도 모든 적절한 변수들이 통제되었다고 확신할 수는 없음.
 - 연구 대상인 사회현상과 인간 행태가 가지는 본연적 불확실성

Observation and Experiment

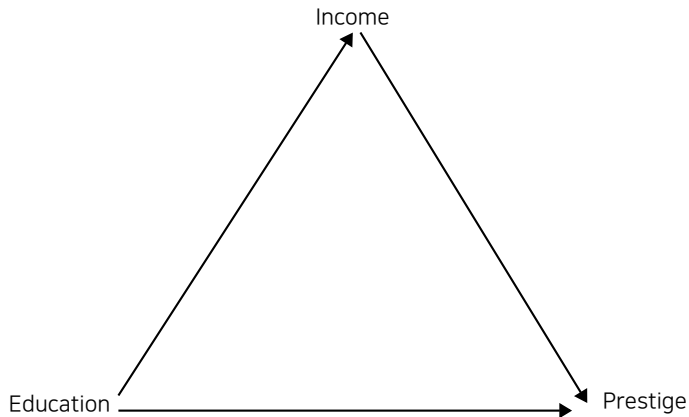


Figure 1: 교육, 소득, 직업에 관한 인과모델

Observation and Experiment

Figure 1의 이해

- 소득 수준은 직업 수준에만 영향을 미치지만, 교육 수준은 소득 수준과 직업 수준 모두에게 영향을 미침.
- 직업 수준과 소득 수준의 관계는 어느 정도 허위적
 - 교육 수준을 통제했을 때, 소득 \rightarrow 직업의 효과는 감소할 것: 허위적 요소의 제거
- 반대로 교육 수준과 직업 수준의 관계는 소득 수준에 의해 어느 정도 매개됨 (mediated).
 - 소득 수준을 통제했을 때, 교육 \rightarrow 직업의 효과가 감소하였다면, 그 감소한 만큼이 교육이 소득을 경유해 직업 수준에 미치는 '간접적 효과'

관측자료를 이용할 경우, 종속변수에 가장 우선적인 원인이 되는 설명변수와 그 설명변수-종속변수 관계에 끼어드는 변수를 구별하는 것이 중요
사회과학에서 “원인”(cause)이라는 개념은 조금 느슨하게 사용됨.

Populations and Samples

연구 문제: 고등학교 3학년 수험생이 공부를 오래할수록 모의고사 성적을 더 잘 받을까?

- 모집단: 전체 고등학교 3학년 수험생
- 표본: 모의고사 시험을 친 고등학교 3학년 수험생 집단

이때, 표본은 전체 고등학교 3학년 수험생에 대한 대표성을 지닌, 그보다 작은 규모의 집단을 의미

- 우리가 관심을 가지고 있는 것은 '일반화할 수 있는' 모집단
- 하지만 현실에서 모집단을 관측하기란 불가능
- 따라서 우리는 모집단을 대표할 수 있는 표본의 특성을 통해 모집단 또한 그러할 것이라는 추론(inference)을 하게 됨.
 - 기술추론(descriptive inference): 표본의 생김새를 보고 모집단도 어떻게 생겼을 것이라는 추론
 - 인과추론(causal inference): 표본에서 어떠한 변수들 간의 관계가 모집단 수준에서도 그러할 것이라는 추론

Populations and Samples

통계 추론은 어느 정도 불확실성을 내재한 채로 수집된 자료들이 나타내는 경향성(pattern)의 안정성에 관한 것

- 실험 자료를 통해서 우리는 그 관계가 안정적일 때, 인과성(causation)을 얘기할 수 있음.
- 마찬가지로 불확실성을 내재한 자료들을 통해 모집단에 관한 일반화(generalization)를 할 수 있지만, 어디까지나 가치판단의 문제와 직결

Fox, John Jr. 2016. Applied Regression Analysis and Generalized Linear Models.
3 ed. CA: Thousand Oaks, SAGE Publications.

Table of Contents

Statistical Models and Social Reality

Observation and Experiment

Populations and Samples