범주형 자료분석
Introduction

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

March 10, 2025

진행 순서

범주형 자료분석의 필요성

범주형 자료분석의 필요성

사회과학 및 보건의료 분야의 연구에는 종종 범주형 자료가 사용된다.

- 어떤 변수들, 예컨대 소득액, 체질량지수(body mass index; BMI), 지역별 투표율, 사업체 근로자수 등은 그나마 양적 변수로 측정될 수 있다.
- 그러나 주관적 사회계층, 비만 여부, 지난 대선 투표 여부, 고용형태(정규직/비정규직 등) 등은 종종 가변수(dummy variable) 등 범주형 척도에 따라 측정된다.
- 보통최소제곱(ordinary least square; OLS) 알고리즘을 활용하는 선형회귀모형으로는 이러한 범주형 종속변수를 분석하기 곤란하다.
- 시간을 들여 종속변수로서 흥미로운 가변수(e.g., 취업 여부, 고혈압 상태, 노조가입 여부, A정당 지지, 푸드스탬프 수급 등)에 관해 상상해보자!

범주형 자료분석의 필요성

보통 셋 중 하나의 대응책을 활용한다.

- 첫째, 교차표(cross-tabulation)와 χ^2 분석 등 기초통계분석만을 활용하여 범주형 자료를 분석한다. 예전에는 이런 방식으로 접근하는 연구자들이 많았다.
- 둘째, 리커트 척도(Likert scale)로 측정된 여러 문항들을 합성지수(composite index)로 변환하여 양적 변수처럼 분석한다. 이 과정에서 요인분석(factor analysis) 등을 사용한다(다음 주 수업의 주제가 된다).
- 셋째, 범주형 자료를 분석하기 위해 특화된 회귀분석의 도구들을 사용한다(오늘 수업의 주제가 된다).

범주형 자료분석의 필요성

범주형 자료분석은 매우 다양한 기법들을 포괄하고 있다.

- 첫째, 종속변수가 가변수인 경우(e.g., 지난 대선 투표 여부), 이항 로지스틱(binary logistic) 회귀모형 등을 활용한다.
- 둘째, 종속변수가 명목 척도(nominal scale)를 따르는 변수인 경우(e.g., 지지하는 정당), 다항 로지스틱(multinomial logistic) 회귀모형 등을 활용한다.
- 셋째, 종속변수가 서열 척도(ordinal scale)를 따르는 변수인 경우(e.g., 주관적 사회계층), 서열 로지스틱(ordinal logistic) 회귀모형 등을 활용한다.
- 넷째, 종속변수가 비음정수(non-negative integer)인 가산 자료(count data)인 경우(e.g., 지난 5년 간 이사 횟수), 포와송(Poisson) 회귀모형 등을 활용한다.

범주형 자료분석의 필요성

- 그 밖에도 제로팽창(zero-inflated) 회귀분석, 이산선택모형(discrete choice model), 2단계(two stage) 회귀분석, 토빗(tobit) 회귀분석, 생존분석(survival analysis) 매우 다양한 상황에 대응한 범주형 자료분석 기법이 있으며 지금도 계속해서 개발되고 있다.
- 이 수업에서는 가장 대표적이고 폭넓게 사용되는 범주형 자료분석 기법만을 몇 가지 다루기로 한다.