사회통계연습
종속변수가 서열척도인 경우

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 종속변수가 서열척도인 경우
2. 종속변수가 범주형 변수인 경우
3. 질적 종속변수에 대한 선형회귀분석의 문제점

종속변수가 서열척도인 경우

종속변수가 서열척도인 경우

종속변수가 서열척도인 경우 그냥 선형회귀모형을 사용한다.

- 만일 종속변수 Y가 서열척도, 특히 리커트 척도인 경우 기존의 해석법을 대충 적용할 수 있다.

$$E(Y|X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- 가령 종속변수 Y가 주관적 건강(subjective health)이고 1점(=매우 건강하지 않음)에서 5점(=매우 건강함)까지의 서열척도라고 하자.
- 독립변수 X가 소득수준(10만원 단위)이라면, "소득이 10만원 증가할 때마다 주관적 건강은 $\widehat{\beta}$점 만큼 증가한다"고 일단 해석할 수 있다.

종속변수가 서열척도인 경우

- 종속변수는 1, 2, 3, 4, 5 사이에서 변화할 뿐이므로 완전한 양적변수는 아니고, 당연히 해석도 완전한 것은 아니다(Why?).
- 훨씬 더 추천하고 싶은 방법은, 높은 타당성과 높은 신뢰성(특히 내적 일관성)을 갖춘 여러 문항들을 더하여 합성지수를 만들고 이를 선형회귀모형으로 분석하는 것이다!
- 그러나 그런 문항을 도저히 찾을 수 없고, 분석하고자 하는 문항이 리커트 척도 딱 하나 뿐인데 만일 리커트 척도가 짝수 개의 항목을 가지고 있다면 이분화 (dichotomization)할 수도 있다.
- 4점 리커트 척도(1=매우 불행, 2=다소 불행, 3=다소 행복, 4=매우 불행)인 경우라면 선형확률모형을 사용하기 위해 0=불행, 1=행복으로 단순화할 수 있기 때문이다!
- 그러나 리커트 척도가 홀수 개의 항목을 가지고 있어 나누기도 애매하면 그냥 선형회귀모형을 사용하자.

종속변수가 서열척도인 경우

문항의 측도를 잘 살펴보고 함부로 서열성을 예단해서는 안된다.

- 어떤 문항이 주어졌을 때, 이것이 범주형 변수인가 아니면 서열척도인가를 잘 따져보고서 분석방법을 결정해야 한다.
- 서열척도가 아니라 사실 범주형 변수인데 함부로 선형회귀모형을 적용해서는 안된다.
- 가령 범주형 변수인 고용형태(1=일용직, 2=임시직, 3=상용직)는 미묘하게 서열척도인 것처럼 보이기도 한다(Why?). 그렇다고 이것을 그대로 선형회귀모형으로 분석하면 납득하기 어려운 해석에 부딪친다.

종속변수가 서열척도인 경우

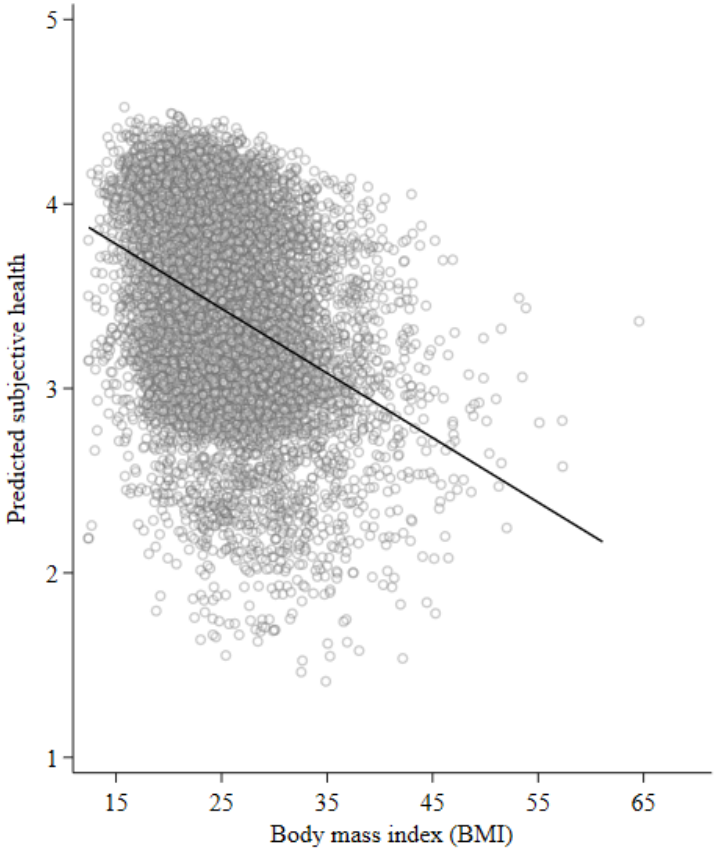
기존 방식에는 문제가 있고, 물론 더 나은 대안도 있다.

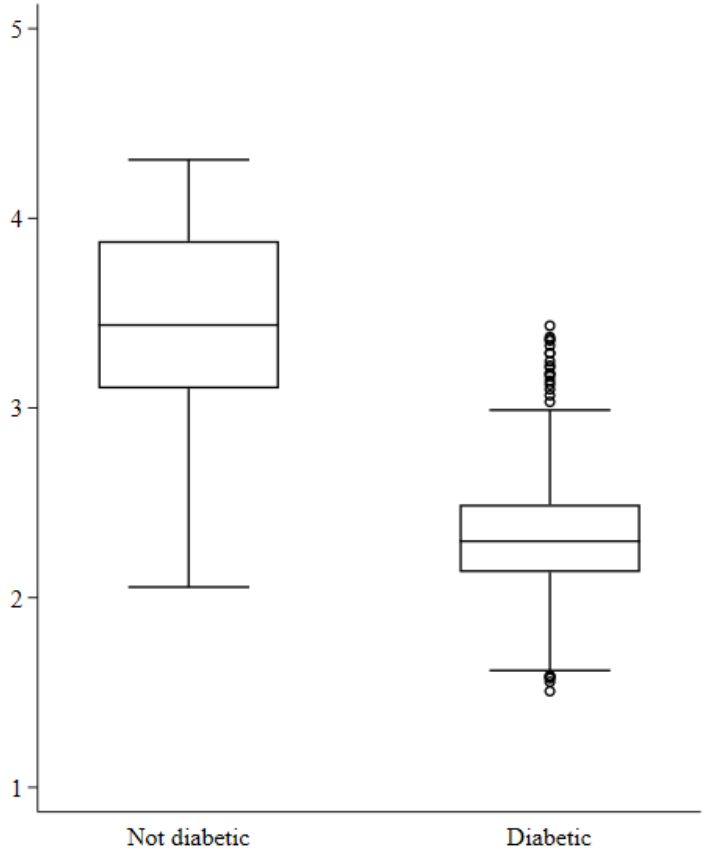
- 선형회귀모형으로 해석하려다보면 이상한 예측값(가령 1보다 작거나 5보다 큰 값)이 도출되거나 하는 등의 한계를 갖는다.
- 지난 수업 연습 3의 시각화 자료를 다시 돌이켜보자. y축이 이상하다!
- 훨씬 더 바람직한 접근방법 중 하나는 서열 로지스틱 회귀분석(ordinal logistic regression analysis)이다.
- 그러나 이런 모형은 (OLS 알고리즘이 아니라) MLE 알고리즘을 채택하고 작동 원리와 해석법을 당장 배우기엔 다소 어렵다.
- 이런 한계와 대안이 있음을 기억하고 그냥 선형회귀모형을 사용하자!

종속변수가 서열척도인 경우

연습 1. nhanes2.sav에서 주관적 건강(health)을 종속변수로, 성별(female), 흑인 여부(black), 연령(age), 고혈압 유병 여부(highbp), 당뇨병 유병 여부(diabetes), 체질량지수(BMI)을 독립변수로 하는 회귀식을 추정하고 그 결과를 해석하시오. 이 결과를 토대로 (1) 체질량지수와 (예측된) 주관적 건강 그리고 (2) 당뇨병 유병 여부와 (예측된) 주관적 건강 간의 관계를 시각화하시오.

종속변수가 서열척도인 경우





종속변수가 서열척도인 경우

연습 2. 한국인의 의식·가치관 조사 자료(korvals.sav)에서 성별(SQ2), 연령(SQ3), 월평균 가구소득(DQ6), 거주지역 규모(SQ1_1), 한국사회로부터 받는 대우의 공정성 인식(Q24_*), 정치적 이데올로기(DQ8)를 독립변수로 하고, 대기업–중소기업 간 갈등에 대한 인식(Q20_7)과 정규직–비정규직 간 갈등에 대한 인식(Q20_8)을 종속변수로 하는 회귀분석을 각각 수행하시오. 회귀분석의 결과로 확인된 공정성 인식과 성별과 결혼에 대한 인식의 연관성을 시각화하시오.

종속변수가 범주형 변수인 경우

종속변수가 범주형 변수인 경우

범주형 종속변수인 경우 가변수로 바꾸어 선형확률모형을 사용한다.

- 독립변수 X가 범주형 변수일 때, 어떻게 분석하였는지 다시 떠올려보자.
- 이때는 기준집단(reference group)을 하나 정해 그것만 빼고 나머지를 가변수로 만들어 독립변수로 회귀모형에 투입하였다.
- 종속변수인 경우도 마찬가지로 생각하면 된다. 하나 기준집단을 정해 그것은 뺀다.
- 나머지는 가변수를 종속변수로 하는 선형확률모형과 완전히 똑같다.

종속변수가 범주형 변수인 경우

연습 3. empltype.sav에서 고용 유형(WGSTAT)을 종속변수로, 성별(SEX), 도시 거주 여부(URBAN), 교육수준(EDUC), 연령(AGE)을 독립변수로 하는 회귀식을 추정하고 그 결과를 해석하시오.

종속변수가 범주형 변수인 경우

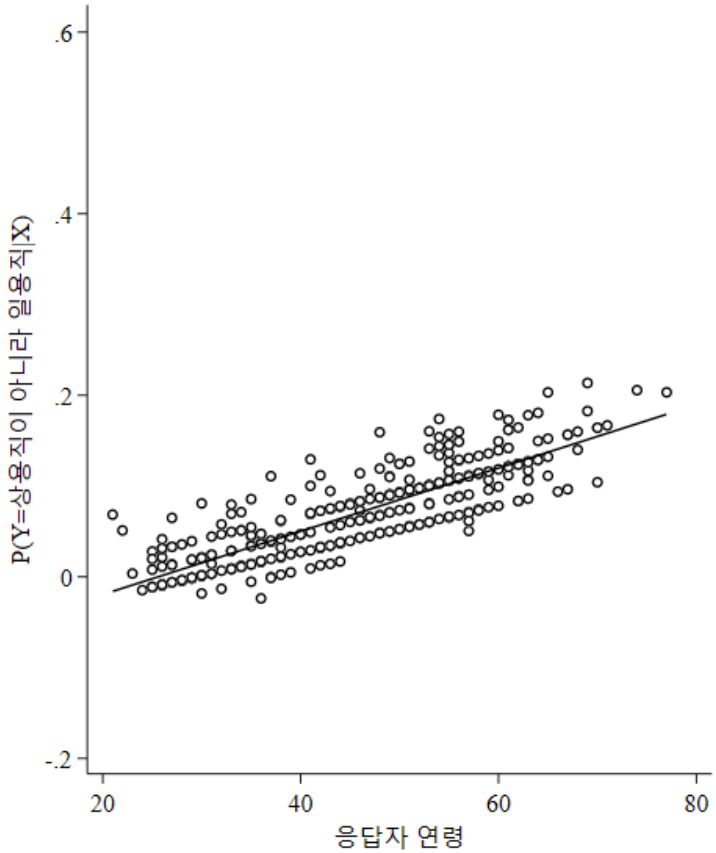
- WGSTAT을 더미변수로 바꿀 때는 측도에 조심해야 한다.
- (기준집단으로 상용직을 삼는다면) 응답자가 상용직이 아니라 임시직일 확률을 분석하기 위해, 일용직인 응답자를 분석에서 잠깐 제외해야 한다(Why?)!
- 이렇게 해야만 분석할 때 (일용직은 아예 생각하지 않고) 상용직이냐 임시직이냐에만 분석의 초점을 두게 된다.
- 마찬가지로 상용직이 아니라 일용직일 확률을 분석하려면 임시직인 응답자를 분석에서 제외해야 한다.
- SPSS에서는 [데이터]-[케이스 선택]에서 일용직 혹은 상용직을 분석에 앞서 배제하자. 이때는 배제를 위해 ~= 기호를 사용한다(회귀분석 단계에서 "선택변수" 기능을 사용할 수도 있다).
- Jamovi에서는 [데이터]-[필터]를 사용한다. 이때는 배제를 위해 ! = 기호를 사용한다 (회귀분석 단계에서 "선택변수" 기능을 사용할 수도 있다).

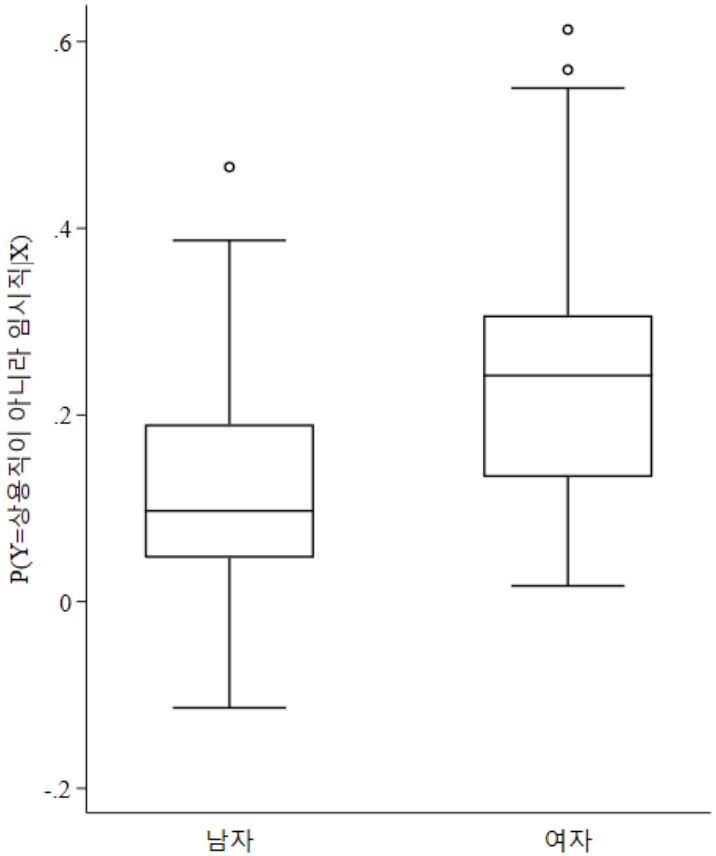
종속변수가 범주형 변수인 경우

- 만약 선택변수에서 그것들을 배제하지 않고 분석하면 어떻게 될까?
- 그러면 상용직이 아니라 임시직일 확률을 분석한 것이 아니게 된다. 일용직인 사람들이 자료에 포함되어 있으므로 상용직 또는 일용직이 아니라 임시직일 확률을 분석한 셈이 된다(Why?).
- 가변수 자료를 곰곰히 살펴보고 그 이유를 따져보자!

종속변수가 범주형 변수인 경우

- 가령 (1) 연령에 따른 상용직이 아니라 일용직이 될 확률 변화와, (2) 성별에 따른 상용직이 아니라 임시직이 될 확률 변화를 각각 시각화할 수 있다.





질적 종속변수에 대한 선형회귀분석의 문제점

그런데 이래도 괜찮은 것일까?

- 종속변수가 가변수 혹은 범주형 변수일 때, OLS 알고리즘을 사용한 선형회귀모형을 그대로 사용하면 몇 가지 문제가 있는 것으로 알려져 있다.
- 첫번째 문제는 선형회귀모형의 가정, 특히 등분산성(homoscedasticity)에 위배된다는 점이다.
- 두번째 문제는 말도 안되는 예측확률이 나온다는 점이다(e.g., 확률이 0보다 작거나 1보다 큼).
- 그러나 그러한 문제에도 불구하고 (1) 그 심각성이 상대적으로 적다는 것과 (2) 이 문제를 극복할 수 있는 더 좋은 분석 기법은 좀 어렵다는 것을 고려할 수 밖에 없다.

더 나은 대안은 무엇일까?

- 보다 근래에는 종속변수가 가변수라면 로지스틱 회귀분석 등이 보다 널리 사용된다.
- 종속변수가 서열척도인 경우에는 서열 로지스틱 회귀분석이, 종속변수가 범주형 변수인 경우에는 다항(multinomial) 로지스틱 회귀분석이 적합하다.
- 이런 한계와 대안이 있음을 기억하고 일단은 그냥 선형확률모형을 사용하자!