사회통계연습
가변수의 활용

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 질적 변수의 분석
2. 회귀분석에서 가변수의 사용
3. 가변수 사용에서 주의사항

질적 변수의 분석

질적 변수의 분석

회귀분석에서 질적 변수는 다룰 수 없는 걸까?

- 우리는 이미 t 검정이나 ANOVA로 범주형 독립변수를 분석할 수 있었다.
- 하지만 통제변수(control variables)를 여러 개 투입할 수 있는 회귀분석과는 달리, (적어도 우리 배운 수준에서) t 검정과 ANOVA만으로는 다른 변수의 영향력을 통제하기 어려웠다(Why?).
- 지금까지 회귀분석에서는 등간(interval) 또는 비율(ratio) 척도로 이루어진 양적 변수만 고려하였다.
- 그 해석은 "X가 한 단위 증가할 때, Y는 β만큼 증가(또는 감소)한다"와 같았다.
- 그러나 명목(nominal) 또는 서열(ordinal) 척도와 같은 질적 변수(qualitative variable)를 분석할 수는 없는 것일까?

질적 변수의 분석

- 이 질문에는 사실 두 가지 측면이 담겨져 있으므로 나누어 살펴보아야 한다.
- 첫째, 종속변수가 질적 변수라면 어떨까? 예를 들어, 성역할 태도(gender role attitude) 중 하나로 "엄마는 직장보다 자녀를 우선시해야 한다" 라는 진술에 대해 동의/부동의를 묻고, 어떤 요인에 의해 그러한 태도를 갖게 되는지 살펴보는 문제가 있을 수 있다.
- 대답은 "할 수 있다" 이다. 다만 학부 수준에서 배우는 보통최소제곱(OLS) 알고리즘으로는 그런 경우를 분석하는데 한계가 있고, 원칙적으로는 대학원 수준의 범주형 자료분석(categorical data analysis)을 공부해야 한다(그러나 우리는 나중에 한다!).

질적 변수의 분석

- 둘째, 독립변수가 질적 변수라면 어떨까? 예를 들어 성별, 인종, 종교, 최종학력은 독립변수로서 사회학적으로 중요한 의미를 갖는다.
- 게다가 독립변수가 질적 변수라면 회귀분석에서도 질적 차이를 부각시켜 해석할 수 있다! 가령 최종학력에서 범주형 변수(categorical variable) 대신 양적변수인 교육연수를 사용하면 어떨까? 에스트로겐 분비량으로 성별을 측정한다면 어떨까? 피부의 명도(brightness)로 인종을 측정한다면 어떨까?
- 이때도 대답은 "할 수 있다" 이다. 게다가 보통최소제곱(OLS) 알고리즘까지도 평범하게 사용할 수 있다!

질적 변수의 분석

다만 회귀식에 질적 변수를 그대로 투입해서는 안된다.

- 왜 그런지 질적 변수의 회귀계수를 한 번 해석해보자.
- 인종(1=백인; 2=흑인; ...), 종교(0=없음; 1=기독교; 2=불교; ...), 리커트(Likert) 척도(1=매우 동의; 2=다소 동의; 3=다소 부동의; 4=매우 부동의) 등은 원칙적으로 모두 질적 변수로 분류될 수 있다.
- 이런 변수들은 그대로 회귀분석에서 독립변수로 사용할 수 없다. 해석이 안되기 때문이다.

회귀분석에서 가변수의 사용

회귀분석에서 가변수의 사용

질적 변수들은 일단 가변수로 바꾸어야 해석할 수 있게 된다.

- 가변수(dummy variable)란 처방, 조건, 또는 상황 등이 존재하면(present) 1로, 그것이 부재하면(absent) 0으로 더미 코딩(dummy coding)된 변수이다.
- 예를 들어, 처치(treatment)에 관한 가변수라면 '받았다(1)' 또는 '안받았다(0)' 중 하나의 값을 갖는다.
- "성별이 여성이다"에 관한 가변수라면 '여성이다(1)' 또는 '여성이 아니다(0)' 중 하나가 된다.
- 더미 코딩할 때 반드시 상호배타적(mutually-exclusive)이고 전체포괄적(all-inclusive)인 범주를 준비하고 적용해야만 한다.

회귀분석에서 가변수의 사용

회귀분석에서 가변수의 작동 원리와 해석은 매우 간단하다.

- 종속변수 Y가 쓰기 점수(write)인 단순회귀식을 다음과 같이 상정하자.

$$Y = \beta_0 + \beta_1 X$$

- X가 양적 변수인 경우 해석은 다음과 같다:

  (1) "X가 한 단위 증가할 때 Y가 β_1 만큼 증가(또는 감소)한다."
  (2) "X가 0일 때, Y는 β_0와 같다."

- 독립변수 X가 가변수인 성별(1=여자; 0=남자)인 경우에 해석은 다음과 같다.

$$\widehat{Y}_{남} = \widehat{\beta}_0 \qquad (\text{if } X = 0)$$
$$\widehat{Y}_{여} = \widehat{\beta}_0 + \widehat{\beta}_1 \qquad (\text{if } X = 1)$$

- 남녀 간 $\widehat{Y}$의 쓰기 점수 격차는 $\widehat{Y}_{여} - \widehat{Y}_{남} = \widehat{\beta}_1$ 이다.
- 그런데 가변수의 회귀계수가 바로 $\widehat{\beta}_1$ 다!

회귀분석에서 가변수의 사용

연습 1. HSB2.SAV에서 쓰기 점수(write)를 종속변수로, 성별(female)를 독립변수로 하는 단순회귀분석을 실시하고 그 결과를 해석하시오.

회귀분석에서 가변수의 사용

- b_0는 어떻게 해석할까? "남자의 write 점수는 평균적으로 50.12점이다."
- b_1는 어떻게 해석할까? "여자의 write 점수는 남자의 write 점수보다 평균적으로 4.87점(b_1=54.99-50.12) 높다."
- 다시 말해, 가변수의 회귀계수는 곧장 남녀 간 차이를 말해준다.
- 한편 여자의 write 점수는 어떻게 알 수 있을까?

회귀분석에서 가변수의 사용

우리는 남자 변수와 여자 변수를 동시에 집어넣지 않았다.

- 자료에서 여자가 아니면 곧바로 남자이기 때문에 두 변수를 동시에 집어넣는 것은 아무런 의미도 없다.
- 이때 집어넣지 않은 쪽을 기준집단(reference group) 또는 근거범주(base category)라고 부른다. 우리의 예제에서는 남자가 기준집단이다.
- 우리는 기준집단이 되는 성별 범주를 0으로 가부호화하였다. 남자를 0으로, 여자를 1로 했으므로 상수는 곧바로 기준집단인 남자의 write 점수를 보여준다.
- (표현 그대로) 기준집단을 기준으로 해석하게 된다. 따라서 여러가지 의미에서 기준이 될 만한 집단을 기준집단으로 삼는 편이 좋다(Why?).

회귀분석에서 가변수의 사용

연습 2. galton.csv에서 아버지와 어머니의 키로 본인의 키를 추정하는 회귀식을 추정하시오. 그 뒤, 본인의 성별을 추가하였을 때 결과를 어떻게 달라지는지 해석하고 모형 적합도가 개선되는지 여부 또한 평가하시오.

가변수 사용에서 주의사항

가변수 사용에서 주의사항

가변수는 꼭 0과 1로만 가부호화해야 할까?

- 만약 노조가입 여부 union에서 {0, 1}이 아니라 {1, 2}, 또는 {-1, 1}로 가부호화하면 어떨까?
- 사실 그래도 된다. 회귀계수와 상수는 변하지만 적합도 지표(R^2나 F 등)은 변하지 않는다(Why?).
- 그러면 왜 관습처럼 {0, 1}로 주로 가부호화할까? 단지 그렇게 할 때 계수 해석이 편리하고 절편 해석도 편리하기 때문이다.
- 보다 구체적으로, (1) 0을 사용하면 기준집단의 평균값이 다른 모든 집단/범주들이 0일 때 자연스럽게 상수와 일치하기 때문이고, (2) 1을 사용하면 회귀계수가 정확히 해당 집단/범주의 평균값을 보여주기 때문이다.

가변수 사용에서 주의사항

연습 3. nlswork.sav에서 임금 ln_wage를 종속변수로 하되 교육연수 grade, 직무경험 ttl_exp, 노조원 여부를 union을 독립변수로 하여 회귀식을 추정하고 해석하시오. 그 뒤, 가변수인 노조원 여부를 노조원이 아닐 때 -1로 재부호화(recoding)하여 회귀식을 다시 추정하고 해석하시오.

가변수 사용에서 주의사항

- 회귀계수와 상수가 달라졌으므로 그 해석도 달라짐에 주의하자.
- 독립변수 X가 독특한 가변수인 노조원 여부(1=노조원; -1=비노조원)인 경우에 해석은 다음과 같다(Why?).

$$\widehat{Y}_{\text{노조}} = \widehat{\beta}_0 + \widehat{\beta}_1 \qquad (\text{if } X = 1)$$
$$\widehat{Y}_{\text{비노조}} = \widehat{\beta}_0 - \widehat{\beta}_1 \qquad (\text{if } X = -1)$$

- 노조원 지위에 따른 임금 $\widehat{Y}$의 격차는 $\widehat{Y}_{\text{노}} - \widehat{Y}_{\text{비노}} = 2\widehat{\beta}_1$ 이다(Why?).
- 해석이 기묘해졌다.

가변수 사용에서 주의사항

그런데 두 범주에 따라 양적변수가 달라지는 상황은 이미 배우지 않았나?

- 사실 우리는 두 개의 범주가 주어졌을 때, 양적변수가 각각 어떻게 달라지는가를 분석하기 위해 t 검정을 배웠다.
- 회귀분석은 뭐가 다른 것일까?
- 쓰기 점수를 여성 독립변수로 회귀분석한 결과와 쓰기 점수를 성별에 따라 t 검정한 결과를 비교해보자.
- 회귀계수의 유의성 검정에서 t 값을 사용한 만큼 두 결과는 같다(Why?).