계량분석
Dummy Variables

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 7, 2024

진행 순서

1. 이분변수
2. 다분변수
3. 범주형 독립변수 연습

이분변수

이분변수

사회학 연구에서는 범주형 변수를 사용해야 할 상황이 제법 많다.

- 우리는 지금까지 양적변수로 측정된 종속변수와 독립변수만을 사용했다. 하지만 생각해보면 양적변수로 표현하기 어려운 사회학적 연구대상도 존재한다.
- 먼저 질적변수인 종속변수는 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.
- 그것은 이른바 범주형 자료분석(categorical data analysis)의 영역이며 이 수업을 모두 이수하고 난 뒤에야 공부할 수 있다.

이분변수

그러면 질적변수인 독립변수는 어떨까?

- 예를 들어 성별, 인종, 종교, 최종학력은 독립변수로서 사회학적으로 중요한 의미를 갖는다.
- 성별은 {남, 녀}, 인종은 {백인, 흑인, 아시아인, 기타}, 종교는 {개신교, 가톨릭, 불교, 기타, 종교없음}, 최종학력은 {고졸 이하, 고졸, 대졸, 대학원졸 이상} 등 다양한 범주를 상상해 볼 수 있다.
- 물론 최종학력에서 범주형 변수(categorical variable) 대신 양적변수인 교육연수를 사용할 수도 있다. 에스트로겐 분비량으로 성별을 측정한다면 어떨까? 피부의 명도(brightness)로 인종을 측정한다면 어떨까?
- 그러한 척도도 상상해 볼 수 있지만, 대체로 이런 경우 의미가 미묘하게 달라진다는 점에 주의해야 한다(여기에 관해서는 곧 다룬다).

이분변수

우리는 이미 t 검정이나 ANOVA로 범주형 독립변수를 분석할 수 있었다.

- 그건 그렇다. 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?).
- 다행히 회귀분석에서도 범주형 변수를 사용할 수 있다!
- 이를 위해 우리는 이분변수(dichotomous variable)와 다분변수(polytomous variable)를 먼저 학습한다.

이분변수

이분변수 또는 가변수(dummy variable)란 무엇인가?

- 처방, 조건, 또는 상황 등이 존재하면(present) 1로, 그것이 부재하면(absent) 0으로 가부호화(dummy coding)된 변수이다.
- 예를 들어, 처리(treatment)에 관한 가변수라면 받았다(1) 또는 안받았다(0) 중 하나의 값을 갖는다.
- "성별이 여성이다"에 관한 가변수라면 그렇다(1) 또는 아니다(0) 중 하나가 된다.
- 이분변수는 이른바 명목변수(nominal variable)의 가장 단순한 형태임에 주의하자.

이분변수

이분변수의 원리와 해석은 사실 매우 간단하다.

- 성별을 예로 들자. 이때 남자는 0, 여자는 1로 가부호화(dummy coding)하였다고 하자.
- 아래와 같이 양적변수 X가 독립변수인 이변량 회귀모형을 추정하였다면, 회귀계수 $\widehat{b}_1$와 상수 $\widehat{b}_0$를 어떻게 해석할까?

$$\widehat{Y} = \widehat{b}_0 + \widehat{b}_1 X$$

- 위 모형에서 만약 X가 가변수라도 해석은 크게 다르지 않다.

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

- 이때 남녀 간 $\widehat{Y}$의 격차는 $\widehat{Y}_{\text{여}} - \widehat{Y}_{\text{남}} = \widehat{b}_1$이다. 이것은 그냥 가변수의 회귀계수이다!

이분변수

write 점수를 female로 예측하는 회귀분석을 연습해보자.

- 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."
- 여자의 write 점수는 어떻게 알 수 있을까?
- Stata에서 margins 명령어를 연습해보자.

이분변수

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?).

다분변수

다분변수

그런데 범주형 변수는 가변수만 있는게 아니라 다분변수도 있다.

- 예를 들어 5명의 사회경제적 지위(socioeconomic status; SES)를 를 세 범주 (1=low; 2=middle; 3=high)로 입력하였다고 하자.

id	ses
1	low
2	middle
3	high
4	high
5	middle

- 이 변수를 쪼개 다음과 같이 가부호화(dummy coding)할 수 있다:

  (1) "ses가 low이다"에 관한 첫번째 가변수(ses1)로 그렇다(1)/아니다(0).
  (2) "ses가 middle이다"에 관한 두번째 가변수(ses2)로 그렇다(1)/아니다(0).
  (3) "ses가 high이다"에 관한 세번째 가변수(ses3)로 그렇다(1)/아니다(0).

다분변수

- 사회경제적 지위 변수(ses) 하나를 다음과 같이 3개의 가변수로 재코딩(recoding)한 셈이다.

id	ses	ses1	ses2	ses3
1	low	1	0	0
2	middle	0	1	0
3	high	0	0	1
4	high	0	0	1
5	middle	0	1	0

- 잘 보면 (어디든지) 한 줄은 결국 필요가 없다. 나머지 두 줄에서 얼마든지 추측이 가능하기 때문이다.
- 아까 이분변수에서와 마찬가지로 바로 그 삭제된 범주가 기준집단이 된다. 나머지 모든 가변수가 0이면 자동적으로 이 범주를 의미하게 된다.

다분변수

다분변수에서 무엇을 기준집단으로 삼아야 할까?

- 기본적으로 상관없으므로 마음대로 정해도 된다. 하지만 이론적으로는 몇 가지 추천할 수 있는 기준이 있다:

  (1) 가장 사례수 N이 큰 (주류) 범주. 뒤집어 말하면 아주 작은 사례 수만 있는 범주는 피하는 것이 좋다(Why?).
  (2) 가장 동질성이 높은 범주. 뒤집어 말하면 "기타" 같은 범주는 피하는 것이 좋다 (Why?).

- 기준집단을 무엇으로 삼는가에 따라 통계적 유의성 구도는 당연히 다르게 나온다 (Why?). 실질적인 해석에서 손해를 보지 않으려면 여러 가지로 기준집단을 바꿔볼 필요가 있을 수 있다.
- 하지만 가장 중요한 것은 연구자의 관심이다. 관심이 가는 집단을 기준집단으로 삼아야 다른 집단과 자꾸 비교할 수 있다는 점을 기억하자(Why?).

다분변수

다분변수의 원리와 해석도 결국 이분변수와 같다.

- 사회경제적 지위(ses)의 범주는 3개가 있었으므로 여기서 하나를 뺀 2개의 가변수만 모델에 투입하게 된다.

$$Y = b_0 + b_1 \text{ses1} + b_2 \text{ses3} + e$$

- 위 회귀모형에서 ses2는 기준집단으로 빠져있는 점에 주목하자.
- 이제 가변수가 여러 개 있는 다중회귀모형을 해석하는 것과 마찬가지이다.

$$\widehat{Y}_{\text{low}} = \widehat{b}_0 + \widehat{b}_1 \qquad \text{(if ses=1)}$$

$$\widehat{Y}_{\text{mid}} = \widehat{b}_0 \qquad \text{(if ses=2)}$$

$$\widehat{Y}_{\text{high}} = \widehat{b}_0 + \widehat{b}_2 \qquad \text{(if ses=3)}$$

다분변수

write 점수를 ses로 예측하는 회귀분석을 연습해보자.

- 가장 먼저 다분변수인 ses의 범주를 빈도분포표를 통해 살펴보자. 레이블(label)과 함께 입력값에 대해서도 살펴보자.
- 다음으로는 기준집단이 될 범주를 정하자. 어느 쪽이 가장 많은 집단/범주인가? 또 어느 집단/범주가 가장 높은/낮은 종속변수의 평균값을 가지고 있나?
- Stata에서는 tabulate 명령어에 generate 옵션을 붙여 쉽게 다분변수를 직접 가부호화할 수 있다.
- 회귀분석을 할 때, 변수 이름 앞에 i.를 붙여 단항 연산자(unary operator)라고 선언하여 다분변수를 인식시킬 수 있다. 만약 b1.를 붙인다면 1번 범주가 기준집단이 된다.

다분변수

- Stata에서 margins 명령어로 조건부 평균값을 예측할 수 있다.

  (1) “ses가 low인 집단(ses1==1)의 작문 성적은 평균적으로 50.62점이다.”
  (2) “ses가 middle인 집단(ses2==1)의 작문 성적은 평균적으로 51.93점이다.”
  (3) “ses가 high인 집단(ses3==1)의 작문 성적은 평균적으로 55.91점이다.”

- 회귀계수와 상수의 해석을 통해 집단 간 차이를 해석할 수 있다.

  (4) “ses가 high인 집단은 middle인 집단보다 작문 성적이 평균적으로 3.99점 높다.”
  (5) “ses가 low인 집단은 middle인 집단보다 작문 성적이 평균적으로 1.31점 낮다.”

다분변수

- 기준집단이 되는 범주를 하나를 빼고 나머지 가변수를 “반드시” 모두 투입해야 한다.
- 예를 들어 low/middle/high로 사회경제적 지위를 분류했을 때, (middle은 기준집단이라서 뺐지만) low 마저도 빼고 high만 모델에 투입한다면, low와 middle 두 집단이 사실상 함께 기준집단이 된다(Why?).
- 이렇게 회귀모형을 만들었다면, high 집단의 작문 점수를 해석할 때 non-high 집단 (즉 low 집단 + middle 집단)과 대조하는 방식으로 이루어져야만 한다. 가변수와 같은 셈이다!
- 다시 말해, 마음대로 하나를 빼거나 하면 그 뺀 범주가 기준집단과 통합되는 효과가 있음을 염두에 두어야 한다.

범주형 독립변수 연습

범주형 독립변수 연습

가변수 이외의 여러 변수들을 통제한 상태에서 가변수의 해석을 연습해보자.

- nlswork.dta에서 ln_wage를 종속변수로, union, race, grade, ttl_exp 독립변수로 하는 회귀모형을 만들자.
- 다중회귀분석에서는 언제나 모든 변수들을 하나하나 꼼꼼하게 살펴보고 모형에 투입해야 한다. 종속변수와 독립변수를 살펴보자. 결측치는 목록별 삭제(listwise deletion)로 대응하자.
- 추정된 회귀모형에서 이분변수인 노조 가입여부 union의 회귀계수를 해석해보자. 그 다음에는 다분변수인 인종 race을 해석해보자.
- 위계적 회귀모형(hierarchical regression models)을 구축하고 어떻게 변화하는지 살펴보자. 이때 적합도 지표로 R^2를 반드시 보고하자.

범주형 독립변수 연습

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.

범주형 독립변수 연습

범주형 변수는 때로 문턱효과를 살펴볼 때 꼭 필요할 수 있다.

- 교육연수(양적변수) 변수와는 달리 최종학력(범주형 변수)은 문턱효과(threshold effects)를 살펴보기에 유리하다.
- 교육연수 grade는 양적변수이므로 해석상 단위 변화(unit change)에 따라 회귀계수의 영향력이 일정하게 작동한다.
- 즉, "고2 → 고3 변화"는 "고3 → 대1 변화"와 동질적인(homogeneous) 것으로 가정된다.
- 이 가정은 타당한가? 사회적으로 구성된 최종학력의 의미는 다르다. 고등학교 중퇴와 고졸 사이에는 질적인 차이, 즉 문턱효과가 있다.

범주형 독립변수 연습

- 따라서 (고졸 미만과 고졸을 질적으로 구분하는) 다분변수인 최종학력을 회귀모형에 투입하는 것이 나을수도 있다(이것은 연구자가 판단할 문제이다).
- 이 자료에서도 대졸 여부 collgrad가 있으므로 이를 회귀모형에 넣고 비교해보자.
- (가령 출산자녀 수처럼 응답 범주의 숫자가 너무 적다면) 일부러 양적변수를 범주형 변수처럼 취급하여 회귀계수를 살펴보아야 한다.
- 사후검정(post-estimation)의 일종인 Wald 검정(Wald test)를 통해 쟁점이 되는 회귀계수가 정말 같은가 여부도 테스트해 볼 수 있다.

범주형 독립변수 연습

가변수는 예외처리에도 유용하게 쓰인다.

- 이론적인 목적에 따라, 도시에 살지 않는(c_city==0), 남부의(south==1), 흑인(race==2)을 특별한 소수자 집단으로 부각시키거나 그 영향력을 통제하고 싶다고 하자(두 표현은 회귀분석에서는 같은 의미이다).
- 그러면 이들 집단을 지시하는 가변수를 따로 만들어 하나의 변수처럼 회귀모형에 투입할 수 있다.
- 무엇이 예외인가에 대해서는 물론 이론적으로 결정되는 측면이 강하지만, 자료를 꼼꼼하게 살펴보아야만 비로소 알 수 있는 부분도 크다.

범주형 독립변수 연습

- 2000년에서 2010년 사이의 자살률을 검토하는 시계열 분석(time-series analysis)을 수행한다면 어떤 시기가 예외적일까? 아마도 2008년 금융위기는 예외적인 사건이므로 통제하는 것이 바람직할 것이다.
- 서울시의 420여개 행정동을 분석단위(unit of analysis)로 삼아 행정만족도를 조사한다고 할 때, 어떤 지역이 예외적일까? 아마도 강남4구(강남구, 서초구, 송파구, 강동구)에 속하는 행정동은 특히 부유한 지역들이므로 이들을 통제하는 것이 바람직할 것이다.

범주형 독립변수 연습

한승용(2008)을 보고 가변수가 활용되는 방식을 살펴보자.

- 연구가설은 무엇인가(172)? 사용한 원자료는 무엇인가(173)?
- 분석단위는 무엇인가(182)? 데이터가 어떻게 구성되어 있을지 상상해 보라.
- 가설을 테스트하기 위해 〈표 6〉 에서 〈표 9〉 까지 제시된 회귀분석 결과표를 보라. 각각 어떻게 가변수를 구성했는지 추론해 보라(183).
- 표를 보고 가변수의 회귀계수를 하나 해석해 보라(183).

한승용. 2008. “사회적 통합과 자살: 연휴가 자살자수 감소에 미치는 영향.”『한국인구학』31(1): 169-198.

범주형 독립변수 연습

<표 7> 혼인상태별 자살양상에 대한 회귀분석, 연휴통합

구분	유배우		무배우	
	B	S.E	B	S.E
-3	-0.43	1.64	-2.34	1.62
-2	-0.53	0.83	-1.73**	0.82
-1	-1.37*	0.83	-1.97**	0.82
연휴	-1.85**	0.59	-1.17**	0.58
+1	-0.20	0.83	-1.00	0.82
+2	-0.58	0.83	-0.47	0.82
+3	-0.37	0.83	1.17	0.82
수목금	-0.66**	0.21	-0.57**	0.20
토일	-1.58**	0.23	-1.35**	0.22
여름	-0.56**	0.24	-0.54**	0.24
가을	-2.03**	0.24	-1.93**	0.24
겨울	-4.35**	0.25	-4.08**	0.25
2001	0.41	0.30	0.87**	0.29
2002	2.65**	0.30	3.28**	0.29
2003	5.91**	0.30	6.24**	0.29
2004	7.13**	0.30	6.40**	0.29
2005	7.23**	0.30	7.72**	0.29
상수	11.58**	0.29	10.90**	0.28
F 비	102.13**		97.07**	
R2	0.44		0.43	

주: 1) * p<.10, ** p<.05.
2) 기준은, 연휴기간 vs. 나머지 기간 / 월화 vs. 수목금-토일 / 봄 vs. 각 계절 / 2000년 vs. 각 연도.