사회통계연습
범주형 변수의 활용

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 범주형 변수의 분석
2. 범주형 변수의 실제 사용
3. 회귀분석 체크리스트
4. 연구가설

범주형 변수의 분석

범주형 변수의 분석

셋 이상의 범주가 담긴 변수를 독립변수로 분석할 수도 있다

- 가령 5명의 사회경제적 지위(socioeconomic status; SES)를 아래와 같이 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 더미 코딩할 수 있다:

  (1) "ses가 low이다"의 가변수(low)로 그렇다(1)/아니다(0).
  (2) "ses가 middle이다"의 가변수(middle)로 그렇다(1)/아니다(0).
  (3) "ses가 high이다"의 가변수(high)로 그렇다(1)/아니다(0).

범주형 변수의 분석

- 사회경제적 지위(ses) 변수 하나를 3개의 가변수로 재부호화한 셈이다.

id	ses	low	middle	high
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.
- 그러므로 위 범주형 변수로부터 3개의 가변수를 만들었더라도 모두 다 독립변수로 회귀모형에 집어넣을 수는 없고 반드시 하나를 빼고 집어넣어야만 한다!

범주형 변수의 분석

- 하나를 빼야 하는 이유는 (1) 나머지 5개의 가변수로부터 마지막 하나의 가변수 내용을 완벽하게 추측할 수 있고, (2) 계산상의 이유로 똑같은 독립변수를 둘 이상 집어넣어서는 안되기 때문이다.
- 이렇게 빠진 변수가 기준집단 또는 근거범주(base category)가 된다.
- 모든 가변수의 해석은 기준집단에 비교하여 이루어진다.

범주형 변수의 분석

- 만일 기준집단을 high로 한다면, 회귀식은 다음과 같이 세울 수 있다.

$$Y = \beta_0 + \beta_1 \text{low} + \beta_2 \text{middle} + \epsilon$$

- 범주형 변수를 가변수로 바꾸어 회귀식에 투입하였다면 다음과 같이 해석할 수 있다 (Why?).

$$\widehat{Y}_{\text{low}} = \widehat{\beta}_0 + \widehat{\beta}_1$$
$$\widehat{Y}_{\text{middle}} = \widehat{\beta}_0 + \widehat{\beta}_2$$
$$\widehat{Y}_{\text{high}} = \widehat{\beta}_0$$

- low 집단과 high 집단의 차이는 $\widehat{Y}_{\text{low}} - \widehat{Y}_{\text{high}} = \widehat{\beta}_1$ 이다.
- middle 집단과 high 집단의 차이는 $\widehat{Y}_{\text{middle}} - \widehat{Y}_{\text{high}} = \widehat{\beta}_2$ 이다.
- 즉 각 집단을 나타내는 가변수의 회귀계수는 곧장 기준집단과의 차이를 드러낸다!

범주형 변수의 분석

연습 1. HSB2.SAV에서 쓰기 점수(write)를 종속변수로, 인종변수(race)를 독립변수로 하는 회귀분석을 실시하고 그 결과를 해석하시오.

범주형 변수의 분석

- SPSS에서 [변환]-[가변수 작성]을 활용해 범주형 변수를 가변수로 간단히 바꿀 수 있다.
- Jamovi에서는 '요인(factor)'에 범주형 변수를 그대로 집어넣거나(요인에 대입하는 경우 기준집단을 고를 수 없다), OneHotEncoding 같은 모듈을 설치하여 대응한다.
- 회귀분석에서 기준집단을 하나 빼놓는 것을 잊지 않아야 한다. 이때 투입하는 변수보다 오히려 빼놓는 변수를 현명하게 선택해야 한다!
- 회귀분석 결과표에 어느 쪽이 기준집단인지 명확히 서술해야 한다.
- 가변수가 여러 개인 경우(e.g., 사회경제적 지위)는 말할 필요도 없고, 하나인 경우(e.g., 여성)도 마찬가지이다.

범주형 변수의 분석

연습 2. 2023년 방송매체 이용형태조사 자료(media.sav)에 따르면, 우리나라 주중 일 평균 OTT 이용 시간(Q559) 평균은 어떻게 되는가? 지역별(DM3)로 이 평균값은 차이가 있는가? 적절한 회귀분석을 수행하고 그 결과를 해석하시오.

범주형 변수의 분석

- 평균적인 OTT 이용 시간은 약 79.6분이다(Why?).
- 회귀분석에서 지역은 범주형 변수이므로 가변수로 먼저 변환하거나 요인으로 투입해야 한다.
- 거듭 강조하지만 가변수의 해석은 기준집단을 비교로 해야 한다(여기서는 서울로 하자).
- "인천/경기는 서울보다 약 11.6분 OTT 이용 시간이 길다.."
- "서울에 비해 부울경은 OTT 이용 시간이 약 9.6분 더 길다."
- "서울의 평균 OTT 이용 시간은 약 75분이다."
- R^2와 F 값도 적절히 해석하자.

범주형 변수의 분석

- 잠깐! 이런 연구질문은 일원분산분석을 통해 답하던 것이 아니었나?
- 여러 개의 가변수를 넣은 회귀분석은 일원분산분석과 완전히 똑같은 결과를 보여준다!
- 이번에는 일원분산분석을 수행하고 그 결과를 앞서 얻은 결과와 비교해보자.
- 일원분산분석 결과에 비해 회귀분석의 장점은 무엇인가?

범주형 변수의 분석

연습 3. 같은 자료를 사용하여, 주중 일 평균 TV 시청시간(Q240)은 어떻게 되는지 살펴보자. 주거형태(dm7)에 따라 TV 시청시간은 어떻게 다른지 살펴보시오.

범주형 변수의 분석

기준집단 하나를 빼고 나머지 가변수는 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 변수의 실제 사용

범주형 변수의 실제 사용

(1) 독립변수 : 우울

이 연구의 독립변수인 우울 정도를 측정하기 위하여 CES-D(Center for Epidemiological Studies - Depression Scale, Radloff, 1977; Andresen, Malmgren, Carter and Patrickl, 1994; Chou K. L., Chi I. and Chou N.W.S, 2004) 단축형을 사용하였다. CES-D는 원래 20문항으로 구성되어 있으나 노인의 응답부담을 줄이기 위해 10문항으로 구성된 단축형이 개발되었으며 높은 수준의 신뢰도를 지니고 있는 것으로 검증되었다(Andresen, Malmgren, Carter and Patrickl, 1994).

우울의 문항 내용은 '무슨 일을 하던 정신을 집중하기가 어려웠다', '우울했다', '하는 일마다 힘들게 느껴졌다', '잠을 설쳤다, 잠을 잘 이루지 못했다', '세상에 홀로 있는 듯한 외로움을 느꼈다', '사람들이 나에게 차갑게 대하는 것 같았다', '생활이 즐거웠다', '슬픔을 느꼈다', '사람들이 나를 싫어하는 것 같았다', '도무지 무엇을 시작할 기운이 나지 않았다'로 구성되어 있으며, 본 연구의 신뢰도 Cronbach's a 계수는 .874로 나타났다.

범주형 변수의 실제 사용

(2) 종속변수 : 자살생각

본 연구의 종속변수인 자살생각은 Augustine Osman, Courtney L. Bagge, Peter M. Gutierrez, Lisa C. Konick, Beverly A. Kopper and Francisco X. Barrious(2001)이 고안한 The suicidal Behaviors Questionnaire-Revised(SRQ - R) : Validation with clinical and Nonclinical samples. Assessment. 8.4를 사용하였다.

자살생각의 문항 내용은 '자살을 생각하거나 시도 경험', '지난 일년동안 자살 생각 여부', '자살관련 얘기를 타인에게 한 경험', '자살을 시도할 가능성'으로 구성되어 있으며, 본 연구의 신뢰도 Cronbach's a 계수는 .805로 나타났다.

(3) 매개변수 : 가족연대감

이 연구의 매개변수인 가족연대감은 Edelstein B. A., Heisel M. J. Mckee D. R. et. al(2009)가 개발한 Reasons for Living Scale-older Adult version (RFL-OA)척도를 한국어 번역하여 사용하였다.

가족 연대감의 문항의 내용은 '자살은 가족에게 너무 큰 상처를 줄 것이다, 나는 그들을 고통스럽게 하고 싶지 않다', '나는 가족을 매우 사랑하고 그들과 즐거운 시간을 누린다. 그들을 떠날 수 없을 것이다', '내겐 힘든 시기에 나를 지지해 주는 사랑스러운 가족이 있다', '가족이 내게 의지하고 나를 필요로 한다'로 구성되어 있으며, 본 연구의 신뢰도 Cronbach's a 계수는 .889로 나타났다.

범주형 변수의 실제 사용

Table 4. 위계적 회귀분석 : 우울, 가족연대감, 자살생각에 미치는 영향(n=2,034)

	가족연대감 (Model 1)			자살생각 (Model 2)			자살생각 (Model 3)			자살생각 (Model 4)		
	b	s.e.	p	b	s.e.	p	b	s.e.	p	b	s.e.	p
우울	-.700	.041	***	.413	.023	***				.322	.024	***
가족연대감							-.188	.012	***	-.130	.012	***
여성(남성)	.180	.042	***	-.082	.024	***	-.060	.024	*	-.059	.023	*
거주지역 도시(농촌)	.028	.042		.032	.023		.039	.024		.035	.023	
독거아님 (독거)	.219	.045	***	-.003	.025		.002	.026		.025	.026	
주관적 건강상태(보통)												
건강	.158	.049	***	.005	.027		.014	.028		.025	.026	
건강하지 않음	.011	.047		.061	.026	*	.133	.026	***	.062	.025	*
월평균가구소득(0-99만원)												
100-199만원	.076	.046		-.071	.026		-.080	.026	**	-.061	.025	*
200-299만원	.148	.058	*	-.093	.032	**	-.099	.033	**	-.073	.032	*
300-399만원	.127	.077		-.135	.043	**	-.129	.044	**	-.118	.042	**
400만원 이상	.300	.108	**	-.112	.060		-.090	.061		-.073	.059	
연령(65세 이상)												
70세 이상	.148	.056	**	-.115	.031	***	-.100	.032	**	-.096	.030	**
75세 이상	.116	.052	*	-.127	.029	***	-.113	.029	***	-.112	.028	***
교육수준 (초등학교 중퇴~졸업)												
무학	-.154	.049	**	.015	.027		-.012	.028		-.005	.027	
중학교(중퇴~졸업)	.043	.055		.017	.031		.024	.031		.022	.030	
고등학교(중퇴~졸업)	.099	.058		.001	.032		-.012	.033		.014	.032	
대학교(중퇴~졸업)	.257	.084	**	.072	.047		.104	.048	*	.105	.046	*
대학원 이상	-.072	.170		.129	.095		.084	.096		.120	.092	
상수	4.615	.085	***	.908	.047	***	1.933	.068	***	1.509	.073	***
R^2	.232			.192			.167			.235		

* p<.05, ** p<.01, *** p<.001

회귀분석 체크리스트

회귀분석 체크리스트

회귀분석의 수행에는 기본적인 원칙들이 있으며 대체로 꼭 지켜야 한다.

1. 연구가설의 제시
2. 변수 설명(정의, 측도, 결측치)
3. 합성지수라면 구성 방식과 내적 일관성 평가
4. 회귀분석을 결정한 이유
5. 회귀식 및 회귀분석 결과표 제시
6. 시각화 제시 및 해석
7. 회귀계수 및 상수 해석
8. 유의성 검정 결과 해석
9. 모형 적합성(ANOVA 및 결정계수) 해석

회귀분석 체크리스트

- 연구가설(research hypothesis)에 대해서는 조금 있다가 좀 더 자세히 살펴보기로 한다.
- 변수가 측정하고자 하는 개념, 정의, 측도 등 기본적인 요소는 잘 요약하여 표로 나타내기도 한다(솔직히 강력하게 추천한다).
- 역부호화 또는 재부호화를 했다면 그 사실을 밝혀야 한다.
- 필요에 따라 변수를 변환했을 때도 이유와 개선 여부를 밝혀야 한다.
- 결측치는 반드시 처리해야 하며 그 사실을 밝혀야 한다("잘 모르겠다"는 결측치다!). 만일 결측치를 평균으로 대체했다면 그 사실을 밝혀야 한다.
- 합성지수를 만들 때 (1) 왜 만들어야 하는가, (2) 내적 일관성이 있었는가? (3) 어떻게 만들었는가를 밝혀야 한다.
- 크론바흐 알파 값은 별도의 표로 만들지 않고 글 속에서 자연스럽게 언급해도 충분하다.

회귀분석 체크리스트

- 회귀분석을 선택한 이유를 밝혀야 한다. 특히 분석 목적과 변수들의 측도를 언급해야 한다.
- 회귀분석 결과표는 최대한 간결하면서도 필요한 정보는 모두 보여주도록 만든다. 쓸떼없는 중복이나 공백은 피해야 한다.
- 유의도 검정 결과는 별(*)로 요약하며, 별의 갯수와 유의확률의 관계 역시 노트로 표 밑에 달아야 한다.
- 결과표에 회귀계수만 달랑 보여주는 것은 추천하지 않고, 표준오차 또는 t 값을 함께 보고하는 것을 추천한다. 그런데 표준오차와 t 값을 둘 다 보고할 필요는 없다(Why?).
- 시각화는 반드시 필수인 것은 아니다. 그러나 보통 글과 수식만 보면 사람들이 지루해 하기 쉬우므로 종종 유용하다!

회귀분석 체크리스트

- 회귀계수 및 상수 해석에는 정답이 있다! 이 정답에서 빗나가면 감점이다.
- 회귀계수의 해석과 사회이론적인 설명을 서로 연결할 수 있어야 한다. 그러나 과장은 금물!
- ANOVA 결과표는 종종 표로 만들기도 하지만 (상대적으로 소극적인 모형 적합성 검정이기 때문에) 표 안에 적당히 F 값을 보고하는 것으로 대체하기도 한다.
- 결정계수 R^2 의 해석에도 정답이 있다! 이 정답에서 빗나가면 감점이다.

연구가설

연구가설

사회이론이 진짜 설명이자 주장이고, 사회통계학은 도구일 뿐이다.

- 사회학도는 이론으로 자신의 주장을 가다듬는다. 이론을 만드는 구체적인 방식과 절차는 다음 학기 사회조사방법론에서 배운다.
- 사회이론에 입각하여 자신의 연구 질문(research question)을 먼저 세워야 한다.
- e.g., "A대학에서는 1학년 학생들의 학업 스트레스 대처 교육 프로그램을 개발하였는데, 과연 이 프로그램은 학업 스트레스 감소 효과가 있을 것인가?"
- 그리고 연구 질문에 대한 잠정적인 대답인 연구가설이 이어진다.
- e.g., "교육 프로그램에 참여한 학생들은 그렇지 않은 학생보다 학업 스트레스 수준이 낮을 것이다."

연구가설

- 그런데 통계분석 단계에서도 가설을 세운다는 점을 떠올리자.
- 다름아닌 귀무가설(null hypothesis) H_0과 대립가설(alternative hypothesis) H_a이다.
- 학업 스트레스 검사 결과, 평균 점수를 μ라고 했을 때, 가령 다음과 같은 t 검정을 실시할 수 있다.

$$H_0 : \mu_T - \mu_C = 0$$
$$H_a : \mu_T - \mu_C \neq 0$$

- 회귀분석($Y = \beta_0 + \beta_1 X$)에서도 마찬가지로 다음과 같은 가설을 세울 수 있다.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

연구가설

- 다시 말해, 우리가 흔히 말하는 가설(hypothesis)은 적어도 두 가지 의미에서 구분해야 한다.
- (1) 사회이론의 단계에서는 연구질문에 대한 잠정적 답변인 연구가설(research hypothesis)을 의미하고, (2) 통계분석 단계에서는 귀무가설(null hypothesis) 또는 대립가설(alternative hypothesis)을 말한다.
- 이 점을 꼭 주의깊게 이해하자: 연구가설과 대립가설은 결국 같은 것이다(Why?)!
- 우리가 연구보고서와 논문을 작성할 때는 보통 (1) 부분만 보고한다! 그러나 (2) 부분도 정확히 이해하고 있어야 통계분석 결과를 올바르게 해석할 수 있다.
- 시험에서는 모두 기술하는 편을 추천한다. 다만 여러 가설들을 뒤죽박죽 섞어 언급하면 좋지 않다.

연구가설

이론의 제시와 통계분석은 '이론상' 구분된다. 하지만...

- 현실적으로 이론의 제시와 통계분석을 기름과 물처럼 완벽히 구별하기란 어렵다.
- 그 이유 중 하나는, 이론을 제시하고서 그것을 제대로 테스트할 수 있는 좋은 통계 원자료(raw data)를 구하기 어렵다는 것이다.
- 좋은 원자료를 생산하려면 그 자체로 상당한 투자를 필요로 한다. 우리는 이론에 딱 맞는 원자료를 스스로 생산하기보다는, 종종 이미 공개된 원자료를 무료로 사용한다.
- 이런 경우 (그 원자료가 우리의 이론에 맞추어 진 것이 아니므로) 이론 제시의 단계에서 우리가 가진 원자료를 통해 검증가능한가를 염두에 둘 수 밖에 없다!