# 계량분석
## Dummy Variables

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 15, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 진행 순서

1. 이분변수
2. 다분변수
3. 이분변수 및 다분변수의 연습

# 이분변수

# 이분변수

경험적 사회과학 연구에서는 범주형 변수를 사용해야 할 상황이 많이 있다.

- 우리는 지금까지 숫자형 변수로 측정된 종속변수와 독립변수만을 사용했다. 하지만 생각해보면 숫자형 변수로는 표현하기 어려운 수많은 사회 현상이 존재한다.
- 가령 "숫자형 변수가 아닌 종속변수"는 어떨까? 예를 들어 (1) "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 경우나 (2) 1-5점 사이 리커트 척도로 측정된 "원자력발전에 대해 인지된 위험도"를 여러 요인들로 설명하는 경우가 그에 해당한다.
- 그것은 이른바 범주형 자료분석(categorical data analysis)의 영역이며 이 수업을 모두 이수하고 난 뒤에야 공부할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

그러면 "숫자형 변수가 아닌 독립변수"는 어떨까?

- 예를 들어 성별, 인종, 종교, 최종학력는 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.
- 다른 한편, 범주형 척도인 최종학력 대신 숫자형 척도인 교육연수(year of education)를 사용할 수도 있다. 데이터에 (두 변수가 모두 존재한다는 전제 아래) 학력을 범주형으로 측정하거나 숫자형으로 측정하거나 의미는 같을까? 대답은 "아니오" 이다(여기에 관해서는 곧 다룬다).
- 여기에 더하여, 아까 종속변수로 언급되었던 원자력발전에 대해 인지된 위험도나 성역할태도 등도 때로는 독립변수로 유용할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

우리는 이미 t 검정(t-test)이나 일원분산분석(ANOVA)으로 범주형 독립변수를 분석할 수 있었는데?

- 그건 그렇다! 하지만 (적어도 우리 배운 수준에서) t 검정과 ANOVA만으로는 "다른 변수의 영향력을 통제하기" 어려웠다.
- 우리가 배웠듯 회귀분석은 통제변수(control variables)를 여러 개 투입할 수 있는 탄력적인 도구라는 측면에서 큰 장점을 가지고 있다. 어떻게든 회귀분석 속에서 범주형 변수를 사용할 수는 없을까? 대답은 "그렇다"이다.
- 이를 위해 우리는 이분변수(dichotomous variable)와 다분변수(polytomous variable)를 학습한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

이분변수 또는 가변수(dummy variable)란 무엇인가?

- 처방, 조건, 또는 상황 등이 존재하면(present) 1로, 그것이 부재하면(absent) 0으로 더미 코딩(dummy coding)된 변수이다.
- 예를 들어, “처방 상태”에 관한 더미변수(treatment dummy)라면 받았다(1) 또는 안받았다(0) 중 하나의 값을 갖는다. “성별이 여성이다”에 관한 더미변수라면 그렇다(1) 또는 아니다(0) 중 하나가 된다.
- 이분변수는 이른바 명목변수(nominal variable)의 가장 단순한 형태(!)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

이분변수의 원리와 해석은 사실 매우 간단하다.

- 성별(변수명 FEMALE)을 예로 들자. 이때 남자는 0, 여자는 1로 더미 코딩되었다고 하자.
- 아래와 같이 숫자형 척도로 측정된 X를 포함한 이변량 회귀모형을 추정하였다면, "X가 한 단위 증가할 때 종속변수가 회귀계수($\hat{b}_1$)만큼 변화한다" 그리고 "X가 0일 때, 종속변수는 상수($b_0$)와 같다" 라고 해석하였다.

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

- 위 모형에서 만약 X가 숫자형 변수가 아니라 더미변수라도 해석은 마찬가지이다.

$$\hat{y}_{남} = \hat{b}_0 \qquad (\text{if } X_i = 0)$$
$$\hat{y}_{여} = \hat{b}_0 + \hat{b}_1 \qquad (\text{if } X_i = 1)$$

- 그러므로 남녀 간 $\hat{y}$의 격차는 $\hat{y}_{여} - \hat{y}_{남} = \hat{b}_1$ 이다. 이것은 그냥 더미변수의 회귀계수이다!

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

작문(write) 성적을 성별(female)로 예측하는 회귀분석을 연습해보자 (**[Stata 코드]** 참고).

```
. reg write female

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =     13.94
       Model |  1176.21384         1  1176.21384   Prob > F        =    0.0002
    Residual |  16702.6612       198  84.3568745   R-squared       =    0.0658
-------------+----------------------------------   Adj R-squared   =    0.0611
       Total |  17878.875        199  89.843593    Root MSE        =    9.1846

------------------------------------------------------------------------------
       write | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   4.869947   1.304191     3.73   0.000     2.298059    7.441835
       _cons |   50.12088   .9628077    52.06   0.000     48.22221    52.01955
------------------------------------------------------------------------------
```

- "남자의 작문 성적은 평균적으로 50.12점이다."
- "여자의 작문 성적은 평균적으로 54.99 (=50.12+4.87)점이다."
- "여자의 작문 성적은 남자의 작문 성적보다 평균적으로 4.87점($b_1$=54.99-50.12) 높다."

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

- 기본적인 원칙 중 하나는 우리가 남자 변수와 여자 변수를 동시에 집어넣지 않았다는 점이다. 우리의 데이터에서 "여자가 아니면 곧바로 남자이기 때문에" 두 변수를 동시에 집어넣는 것은 아무런 의미도 없다.
- 이때 집어넣지 않은 쪽은 기준집단(reference group) 또는 근거범주(base category)가 된다. 우리의 예제에서는 남자가 기준집단이다.
- 우리는 기준집단이 되는 성별 범주를 0으로 더미 코딩하였다. 남자를 0으로, 여자를 1로 했으므로 상수는 곧바로 기준집단인 남자의 작문 점수를 보여준다.
- 이와는 별개로, 여기서 성별 변수의 이름을 보자! female로 되어있으므로 남자가 0, 여자가 1로 더미 코딩되어 있으리라고 쉽게 예측할 수 있다. 변수의 이름을 통해 코딩 방식(coding scheme)이 어떤지 직관적으로 알 수 있도록 하자. 그러므로 변수의 이름을 sex나 gender로 짓는 것은 사실 별로 좋지 않다(Why?).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 다분변수

# 다분변수

그런데 범주형 변수는 더미변수만 있는게 아니라 다분변수도 있다.

- 예를 들어 5명의 사회경제적 지위(socioeconomic status)를 를 세 범주의 코딩 방식(1=low; 2=middle; 3=high)으로 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"에 관한 첫번째 더미변수(status1)로 그렇다(1)/아니다(0).
  (2) "ses가 middle이다"에 관한 두번째 더미변수(status2)로 그렇다(1)/아니다(0).
  (3) "ses가 high이다"에 관한 세번째 더미변수(status3)로 그렇다(1)/아니다(0).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 다분변수

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

| id | ses | status1 | status2 | status3 |
|---|---|---|---|---|
| 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 집단/범주가 기준집단이 된다. 나머지 모든 더미변수가 0이면 자동적으로 이 그룹/범주를 의미하게 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 다분변수

다분변수에서 무엇을 기준집단으로 삼아야 할까?

- 기본적으로 상관없으므로 마음대로 정해도 된다. 하지만 몇 가지 추천되는 기준이 있다.
  (1) 가장 사례 수(n)가 가장 많은 (주류) 범주. 뒤집어 말하면, 아주 작은 사례 수만 있는 범주는 피하는 것이 좋다.
  (2) 종속변수의 평균이 제일 높거나 반대로 제일 낮은 범주. 뒤집어 말하면, 중간 정도 되는 범주는 피하는 것이 좋다.
- 기준집단을 무엇으로 삼는가에 따라 통계적 유의성이 조금씩 다르게 나올 수 있다. 실질적인 해석에서 손해를 보지 않으려면 여러 가지로 기준집단을 바꿔볼 필요가 있을 수 있다.
- 하지만 궁극적으로 중요한 것은 연구자의 관심이다. 관심이 가는 집단을 기준집단으로 삼아야 다른 집단과 자꾸 비교할 수 있다는 점을 기억하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수

다분변수의 원리와 해석도 결국 이분변수와 같다.

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

$$y_i = b_0 + b_1 SES_i^{[low]} + b_2 SES_i^{[high]} + e_i$$

- 위 회귀모형에서 $SES_i^{[middle]}$ 은 기준집단으로서 빠져있는 점에 주목하자.
- 이제 더미변수가 여러 개 있는 다변량 회귀모형을 해석하는 것과 마찬가지이다.

$$\begin{aligned}
\hat{y}_{low} &= \hat{b}_0 + \hat{b}_1 && (\text{if } SES_i^{[low]} = 1) \\
\hat{y}_{middle} &= \hat{b}_0 && (\text{if } SES_i^{[middle]} = 1) \\
\hat{y}_{high} &= \hat{b}_0 + \hat{b}_2 && (\text{if } SES_i^{[high]} = 1)
\end{aligned}$$

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

# 다분변수

작문(write) 성적을 사회경제적 지위(ses)로 예측하는 회귀분석을 연습해보자(**[Stata 코드]** 참고).

- 가장 먼저 다분변수인 사회경제적 지위(ses)의 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 더미코딩할 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.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 다분변수

```
. reg write status1 status3
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 858.715441 | 2 | 429.35772 | Number of obs | = 200 |
| Residual | 17020.1596 | 197 | 86.396749 | F(2, 197) | = 4.97 |
| | | | | Prob > F | = 0.0078 |
| | | | | R-squared | = 0.0480 |
| | | | | Adj R-squared | = 0.0384 |
| Total | 17878.875 | 199 | 89.843593 | Root MSE | = 9.295 |

| write | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| status1 | -1.309295 | 1.657609 | -0.79 | 0.431 | -4.578231 | 1.959642 |
| status3 | 3.987477 | 1.548883 | 2.57 | 0.011 | .9329576 | 7.041997 |
| _cons | 51.92632 | .9536454 | 54.45 | 0.000 | 50.04565 | 53.80698 |

- "ses가 low인 집단(status1==1)의 작문 성적은 평균적으로 50.62점이다."
- "ses가 middle인 집단(status2==1)의 작문 성적은 평균적으로 51.93점이다."
- "ses가 high인 집단(status3==1)의 작문 성적은 평균적으로 55.91점이다."
- "ses가 high인 집단은 middle인 집단보다 작문 성적이 평균적으로 3.99점 높다."
- "ses가 low인 집단은 middle인 집단보다 작문 성적이 평균적으로 1.31점 낮다."

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 다분변수

- 기준집단의 평균은 회귀모형의 상수가 되므로 그림에서는 y축 절편(interecept)에 그대로 반영된다.
- 기준집단이 되는 범주를 하나를 빼고 나머지 더미변수를 “반드시” 모두 투입해야 한다. 예를 들어 (1) low, (2) middle, (3) high로 사회경제적 지위를 분류했을 때, (middle은 기준집단이라서 뺐지만) low 마저도 빼고 high만 모델에 투입한다면, low와 middle 두 집단이 사실상 함께 기준집단이 된다. 이렇게 모델을 만들었다면, high 집단의 작문 점수를 해석할 때 non-high 집단(즉 low 집단 + middle 집단)과 대조하는 방식으로 이루어져야만 한다. 더미변수와 같은 셈이다!
- 다시 말해, 마음대로 하나를 빼거나 하면 그 뺀 범주가 기준집단과 통합되는 효과가 있음을 염두에 둘 것.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수 및 다분변수의 연습

# 이분변수 및 다분변수의 연습

더미변수 이외의 여러 변수들을 통제한 상태에서 더미변수의 해석을 연습해보자(**[Stata 코드]** 참고).

- eCampus에서 nlswork.dta를 다운받자. 종속변수로 로그임금(ln_wage)를 사용하자. 독립변수로 union, race, grade, ttl_exp를 사용하자.
- 다변량 회귀분석에서는 언제나 모든 변수들을 하나하나 꼼꼼하게 살펴보고 모델에 투입해야 한다. 종속변수와 독립변수를 살펴보자. 결측치는 listwise deletion으로 제거하자.
- 추정된 회귀모형에서 이분변수인 노조 가입여부(union)의 회귀계수를 해석해보자. 그 다음에는 다분변수인 인종(race)을 해석해보자.
- 위계적 회귀모형(hierarchical regression models)을 구축하고 어떻게 변화하는지 살펴보자. 이때 적합도 지표로 $R^2$를 반드시 보고하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수 및 다분변수의 연습

만약 노조가입 여부(union)에서 {0, 1}이 아니라 {1, 2}, 또는 {-1, 1}로 더미 코딩하면 어떨까?

- 사실 그래도 된다. 나중에 짬을 내서 직접 해보자. 회귀계수와 상수는 변할 수 있지만 적합도 지표($R^2$나 F 등)은 변하지 않는다(Why?).
- 그러면 왜 관습처럼 {0, 1}로 주로 더미 코딩할까? 단지 그렇게 할 때 계수 해석이 편리하고 절편 해석도 편리하기 때문이다.
- 보다 구체적으로, 0을 사용하는 이유는 특히 기준집단의 평균값이 다른 모든 집단/범주들이 0일 때 자연스럽게 상수와 일치하기 때문이고, 1을 사용하는 이유는 회귀계수가 정확히 해당 집단/범주의 평균값을 보여주기 때문이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수 및 다분변수의 연습

교육연수(숫자형) 변수와는 달리 최종학력(범주형) 변수는 문턱효과(threshold effects)를 살펴보기에 유리하다(**[Stata 코드]** 참고).

- 학력의 경우 교육연수(grade)는 숫자형 변수이므로 해석상 단위 변화(unit change)에 따라 회귀계수의 영향력이 일정하게 작동한다. 즉, "고2 → 고3 변화"는 "고3 → 대1 변화"와 동등하다(homogeneous)고 가정된다. 이 가정은 타당한가?
- 사회적으로 구성된 최종학력의 의미는 다르다. 고등학교 중퇴와 고졸 사이에는 질적인 차이, 즉 문턱효과가 있기 때문이다.
- 따라서 고졸 이하와 고졸을 질적으로 구분하여 다분변수인 최종학력을 모델에 투입하는 것이 나을수도 있다(이것은 연구자가 판단할 문제이다). 이 데이터에서는 대졸 여부(collgrad)가 있으므로 이를 모델에 넣고 비교해보자.
- (숫자가 너무 많지 않다면) 일부러 숫자형 변수를 범주형 변수처럼 한 번 쯤 취급하여 회귀계수를 살펴볼 수 있다. 그 뒤, 사후검정(post-estimation)을 통해 쟁점이 되는 회귀계수가 정말 같은가 여부도 테스트해 볼 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수 및 다분변수의 연습

더미변수는 예외처리에도 유용하게 쓰인다.

- 만일 이론상 도시에 살지 않는(c_city==0), 남부의(south==1), 흑인(race==2)이 불이익을 받는 소수자집단이므로 특별한 예외로 부각시키거나 그 영향력을 통제하고 싶다면(두 표현은 회귀분석에서는 같은 의미이다), 그들을 지칭하는 더미 변수를 만들고 이를 다시 회귀모형에 더미변수로 투입할 수 있다.

```
. reg ln_wage i.union i.race grade ttl_exp disadv
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| | | | | Number of obs | = 19,233 |
| | | | | F(6, 19226) | = 1622.20 |
| Model | 1414.78134 | 6 | 235.796891 | Prob > F | = 0.0000 |
| Residual | 2794.62157 | 19,226 | .14535637 | R-squared | = 0.3361 |
| | | | | Adj R-squared | = 0.3359 |
| Total | 4209.40291 | 19,232 | .218874943 | Root MSE | = .38126 |

| ln_wage | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| 1.union | .1774441 | .0065837 | 26.95 | 0.000 | .1645394 | .1903488 |
| race | | | | | | |
| 2 | -.0295298 | .0072659 | -4.06 | 0.000 | -.0437716 | -.0152881 |
| 3 | .0629695 | .0264522 | 2.38 | 0.017 | .0111209 | .1148181 |
| grade | .0694622 | .0012025 | 57.77 | 0.000 | .0671053 | .0718191 |
| ttl_exp | .0307723 | .0006129 | 50.21 | 0.000 | .029571 | .0319737 |
| disadv | -.219412 | .0110152 | -19.92 | 0.000 | -.2410028 | -.1978211 |
| _cons | .6173548 | .015766 | 39.16 | 0.000 | .586452 | .6482575 |

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

# 이분변수 및 다분변수의 연습

- 무엇이 예외인가에 대해서는 물론 이론적으로 결정되는 측면이 강하지만, 자료를 꼼꼼하게 살펴보아야만 비로소 알 수 있는 부분도 크다.
- 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 강남4구에 속하는 행정동은 특히 부유한 지역들이므로 통제하는 것이 바람직할 것이다.
- 극단적으로 오로지 하나의 사례(observation)에 대해서만 더미변수로 1을 부여한다면 OLS의 회귀계수와 분산 추정 단계에서 그 사례를 제거하는 것과 똑같은 효과를 같는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

# 이분변수 및 다분변수의 연습

한승용(2008)을 보고 더미변수가 활용되는 방식을 살펴보자.

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

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

충북대학교
CHUNGBUK NATIONAL UNIVERSITY