회귀분석
회귀분석의 활용

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

July 5, 2023

진행 순서

1. 가변수와 범주형 변수
2. 상호작용 효과와 비선형적 관계
3. 회귀분석의 가정 및 주의사항

가변수와 범주형 변수

가변수와 범주형 변수

회귀분석에서도 질적 차이를 드러낼 수 있다.

- 지금까지 회귀분석에서는 등간(interval) 또는 비율(ratio) 척도로 이루어진 양적 변수만 고려하였다. 그 결과는 "X가 한 단위 증가할 때, Y는 β만큼 증가한다"와 같이 해석하였다.
- 그러나 명목(nominal) 또는 서열(ordinal) 척도와 같은 질적 변수(qualitative variable)를 분석할 수는 없는 것일까?
- 가령 (1) 종속변수가 질적 변수라면 어떨까? 단순최소자승(OLS)으로는 그런 경우를 분석할 수 없고 범주형 자료분석(categorical data analysis)을 공부해야 한다.
- 그렇다면 (2) 독립변수가 질적 변수라면 어떨까? 이때는 OLS 알고리즘을 사용할 수 있다! 그러나 회귀식에 질적 변수인 채로 그냥 투입해서는 안된다(Why?).
- 질적 변수의 회귀계수를 위와 같이 해석해보라(왜 불가능할까?).

가변수와 범주형 변수

회귀분석에 앞서 질적 변수는 가변수로 바꾸어 사용해야 한다.

- 인종(1=백인; 2=흑인; ...), 종교(0=없음; 1=기독교; 2=불교; ...), 리커트(Likert) 척도로 측정된 만족도(1=매우 불만족; 2=다소 불만족; 3=다소 만족; 4=매우 만족) 등은 모두 질적 변수로 분류될 수 있다.
- 이런 변수들은 그대로 회귀분석에서 독립변수로 사용할 수 없다. 다만 경우에 따라 리커트 척도는 그대로 사용하기도 한다(Why?).
- 질적 변수들은 일단 가변수(dummy variable)로 바꾸어 독립변수로 활용할 수 있다.

가변수와 범주형 변수

회귀분석에서 가변수의 작동 원리와 해석은 매우 간단하다.

- Y는 종교성(religiosity) 점수인 단순회귀식을 다음과 같이 상정하자.

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

- X가 양적변수인 경우 해석은 다음과 같다:

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

- X가 가변수(e.g., 남자=0; 여자=1)인 경우라도 해석은 마찬가지이다.

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

- 남녀 간 $\widehat{Y}$의 종교성 격차는 $\widehat{Y}_{여} - \widehat{Y}_{남} = \widehat{\beta}_1$이다. 그런데 $\widehat{\beta}_1$는 바로 가변수의 회귀계수 자체다!

가변수와 범주형 변수

가변수의 회귀계수는 곧장 남녀 간 차이를 말해준다.

- 남자를 지칭하는 가변수와 여자를 지칭하는 가변수를 따로 회귀모형에 투입할 필요는 없다. 그 정보는 중복된 것이기 때문이다(Why?).
- 중요한 것은 '회귀모형에 넣지 않은 가변수 쪽을 비교 대상으로 삼는다'는 사실이다 (Why?).
- 따라서 아래와 같이 회귀식이 추정된 경우에 가변수 해석은 다음과 같다.

$$Religiosity = 1.81 + 0.75 Female$$

- "여자의 종교성 점수는 남자보다 0.75점 높다."
- "남자의 종교성 점수는 평균적으로 1.81점이다."

가변수와 범주형 변수

연습 6. 당신은 유학에 앞서 시골에 있는 대학에 가면 근처에 취업할 곳이 마땅치 않아서 평균수입이 작을 것이라는 가설을 세웠다. college.sav에서 찾을 수 있는 등록금(cost), 졸업율(grad), 학자금 상환비율(debt)의 영향력을 통제하였을 때, 도시에 캠퍼스가 있는지 여부(city)가 졸업생의 평균수입(earnings)에 어떤 미치는가 회귀식으로 추정하고 해석하시오.

가변수와 범주형 변수

- city는 가변수이다. 만약 {1, 2}로 입력되어 있었다면 이를 {0, 1}로 재부호화(recoding)하여 바꾸어야 한다.

$$Earnings = 10004.966 + 0.435Cost + 178.099Grad + 141.478Debt + 2526.789City$$

- "도시 대학을 졸업한 사람은 비도시 대학을 졸업한 사람보다 약 2,527달러 평균수입이 높다."
- city는 95% 신뢰수준에서 통계적으로 유의하다. 그러므로 위 해석은 표본을 넘어 모집단에 대해서도 적용될 수 있다.

가변수와 범주형 변수

- cost, grad, debt의 중앙값(median)은 각각 24957.5, 67, 90이다. 이를 아래 회귀식에 대입하면 아래와 같다.

$$\begin{aligned} Earnings &= 10004.966 + 0.435Cost + 178.099Grad + 141.478Debt + 2526.789City \\ &= 10004.966 + 0.435 \cdot 24957.5 + 178.099 \cdot 67 + 141.478 \cdot 90 + 2526.789City \\ &= 45527.132 + 2526.789City \end{aligned}$$

- "다른 모든 변수들이 중앙값에 고정되었을 때, 비도시 대학 졸업자(city=0)의 예상되는 평균수입($\widehat{earnings}$)은 약 45,527달러이다."
- "다른 모든 변수들이 중앙값에 고정되었을 때, 도시 대학 졸업자(city=1)의 예상되는 평균수입($\widehat{earnings}$)은 약 48,054달러이다."
- "다른 모든 변수들이 중앙값에 고정되었을 때, 도시 대학 졸업자(city=1)는 비도시 대학 졸업자(city=0)보다 약 2,527달러 더 많이 번다."

가변수와 범주형 변수

물론 범주형 변수를 가변수들로 변환할 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).

가변수와 범주형 변수

- 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!

가변수와 범주형 변수

- 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 변수를 기준집단(reference group) 또는 기저범주(baseline category)라고 부른다.
- 위의 연습 6에서 도시(city) 가변수 외에 비도시 가변수를 따로 넣지 않았던 이유도 여기에 있다.
- 모든 가변수의 해석은 기준집단에 비교하여 이루어진다. 그러므로 위 연습 6에서 도시(city)와 같은 가변수를 해석할 때도 모형에 투입하지 않은 쪽을 기준으로 비교하여 해석하였다.

가변수와 범주형 변수

예제 7. 인구학자인 민주는 미국의 센서스 지역(region)에 따라 연령 중앙값(medage)이 다를 것이라고 예상하고 있다. census.csv를 활용하여 이에 관한 가설을 세워 회귀분석을 수행하고, 그 결과를 해석하시오.

가변수와 범주형 변수

- SPSS에서 [변환]-[가변수 작성]을 활용해 범주형 변수를 가변수로 간단히 바꿀 수 있다.
- 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. 여기서는 남부(South)를 기준집단으로 삼기로 한다.

가변수와 범주형 변수

- 추정된 회귀식은 다음과 같다. 해석은 이에 근거하여 이루어진다.

$$Medage = 29.619 - 0.094NCntrl + 1.615NE - 1.334West$$

- "북부중앙(NCntrl)의 연령 중앙값은 남부와 통계적으로 유의한 차이가 없다."
- "북동부(NE)의 연령 중앙값은 남부에 비해 1.615세 높다."
- "서부(West)의 연령 중앙값은 남부에 비해 1.334세 낮다."
- "남부(South)의 연령 중앙값은 29.619세이다."

가변수와 범주형 변수

- 사실 이 회귀분석은 일원분산분석과 수학적으로 완전히 똑같은 추정 결과를 보여준다.
- 먼저 센서스 지역(region)를 문자에서 숫자로 부호화(encoding)한다. [변환]-[자동 코딩변경]을 통해 할 수 있다.
- 그런 다음, [분석]-[평균 비교]-[일원배치 분산분석]을 따라 일원분산분석을 수행하면 (아까 회귀분석과) 완전히 똑같은 F 값과 유의확률(p-value)을 얻는다.

회귀분석의 가정 및 주의사항

예제 8. servpc.sav는 2010년 우리나라 군 별 인구 1천 명당 공무원 수(servpc)와 재정자립도(finind), 그리고 광역자치단체별 가변수들을 담고 있다. 다음의 물음에 답하시오.

(1) 종속변수는 재정자립도, 독립변수는 광역자치단체별 가변수들인 선형모형을 세우고 회귀분석하시오. 회귀계수, 유의성 검정 및 적합도 검정 결과를 모두 충실히 해석하시오.

(2) 독립변수로 공무원 수도 추가한 선형모형을 다시 한 번 세우고 회귀분석하시오. 회귀계수, 유의성 검정 및 적합도 검정 결과를 모두 충실히 해석하시오.

상호작용 효과와 비선형적 관계

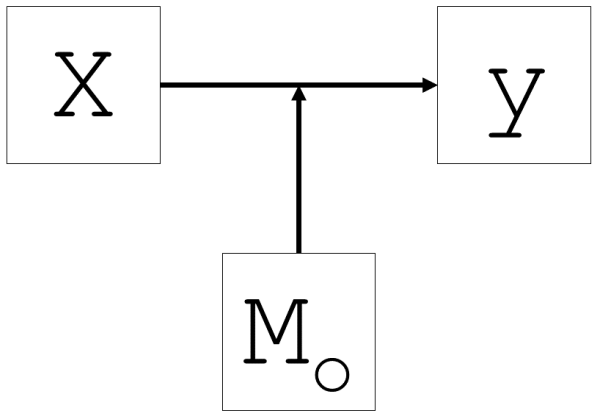
상호작용 효과와 비선형적 관계

독립변수의 효과는 때로 조건부일 수도 있다.

- "경력(career)이 축적됨에 따라 임금이 증가할 것이다" 라는 가설은 단순선형회귀모형에서 쉽게 테스트될 수 있다.
- "대졸자의 임금은 비대졸자보다 높을 것이다" 라는 가설도 가변수를 활용한 단순선형회귀모형에서 쉽게 테스트될 수 있다.
- 그런데 이런 단순한 선형관계에 만족하지 않고, 한층 더 복합적인 분기(junction)를 고려한다면 더욱 세련된 형태로 가설과 이론을 발전시킬 수 있다.
- 가령 "경력 축적을 통한 임금 증가폭은 대졸자보다 비대졸자가 높을 것이다"와 같은 연구 가설은 (앞서 제시한 가설보다) 훨씬 세련되고 흥미롭다!

상호작용 효과와 비선형적 관계

- 어떤 독립변수 X의 종속변수 Y에 대한 효과는 조절변수(moderating variables) M_O에 따라 상이할 수도 있다.
- 이를 파악하기 위해서는 (1) 독립변수 X와 조절변수 M_o 각각의 주효과(main effects)를 먼저 살펴보고, 그 위에 (2) 두 변수가 얽혀 만들어내는 상호작용 효과(interaction effect)를 추가적으로 살펴보아야 한다.
- 조절변수와 상호작용 효과를 표현하는 경로 도식은 다음과 같다.



상호작용 효과와 비선형적 관계

회귀분석 안에서 상호작용 효과는 비교적 쉽게 파악할 수 있다.

- 위 사례를 토대로 두 개의 주효과에 관한 연구 가설을 세울 수 있다.

  (1) "대졸자(collgrad)는 비대졸자보다 더 많은 임금(ln_wage)을 받을 것이다."
  (2) "재직경력(tenure)이 증가하면 임금(ln_wage)은 더 커질 것이다."

- 다음으로 상호작용 효과에 관해 다음의 연구 가설을 세울 수 있다.

  (3) "대졸자의 재직경력 증가에 따른 임금 증가폭보다 비대졸자의 재직경력 증가에 따른 임금 증가폭이 더 클 것이다."

상호작용 효과와 비선형적 관계

- 조절변수는 대졸 여부(collgrad)이며 다음과 같이 회귀식을 구성한다.

$$\ln\_\text{wage} = b_0 + b_1 \text{collgrad} + b_2 \text{tenure}$$

- 대졸 여부(collgrad)와 재직경력(tenure)이 포함되어 있지만, 이것들로는 개별적인 효과(=주효과)만을 볼 수 있을 뿐이다.
- 이때, 재직경력의 임금 상승효과(=회귀계수)는 다음과 같이 두 부분(주효과와 상호작용 효과)으로 분해될 수 있다!

$$b_2 = \gamma_0 + \gamma_1 \text{collgrad}$$

- 위 식도 (오차항없는) 일종의 회귀식처럼 볼 수 있다(Why?).
- 두 식을 결합하면 아래와 같다.

$$\ln\_\text{wage} = b_0 + b_1 \text{collgrad} + \gamma_0 \text{tenure} + \gamma_1 (\text{tenure} \cdot \text{collgrad})$$

상호작용 효과와 비선형적 관계

- 결국, 독립변수인 재직경력(tenure)와 조절변수인 대졸 여부(collgrad) 외에도 두 변수를 곱한 새로운 변수가 포함되어 있다. 이 변수를 특별히 상호작용항(interaction term)이라고 한다.
- 곱하기 이전의 두 변수를 주요항(main terms)이라고 한다.
- 상호작용 효과를 살펴보기 위해 상호작용항 tenure · collgrad을 넣을 때는 두 주요항 tenure와 collgrad을 반드시 함께 식 안에 넣어야 한다(Why?)!
- 위 회귀식의 상호작용항 회귀계수 $\hat{\gamma}_1$에 대해 유의성 검정($H_0 : \Gamma_1 = 0$)을 하여, 이것이 만일 통계적으로 유의하다면 상호작용 효과(interaction effect)는 모집단에서도 존재한다고 말할 수 있다.

상호작용 효과와 비선형적 관계

- 상호작용항을 포함한 회귀식은 아래와 같았다.

$$\text{ln\_wage} = b_0 + b_1\text{tenure} + b_2\text{collgrad} + b_3(\text{tenure} \cdot \text{collgrad})$$

- 상호작용항을 포함한 회귀모형의 해석은 가변수의 해석과 크게 다르지 않다.

$$\begin{aligned}
\text{ln_wage}_{\text{비대졸}} &= b_0 + b_1\text{tenure} && \text{(if collgrad==0)} \\
\text{ln_wage}_{\text{대졸}} &= b_0 + b_1\text{tenure} + b_2 + b_3\text{tenure} && \text{(if collgrad==1)} \\
&= (b_0 + b_2) + (b_1 + b_3)\text{tenure}
\end{aligned}$$

- 대졸이 아닌 경우(collgrad==0), 상수 b_0와 회귀계수 b_1로만 해석한다.
- 대졸인 경우(collgrad==1), 상수 $(b_0 + b_2)$와 회귀계수 $(b_1 + b_3)$에 따라 해석한다.
- 여기서 주목할 부분은 collgrad에 따라 tenure의 상수와 회귀계수에 조금씩 부스터가 더해졌다는 점이다.

가변수와 범주형 변수

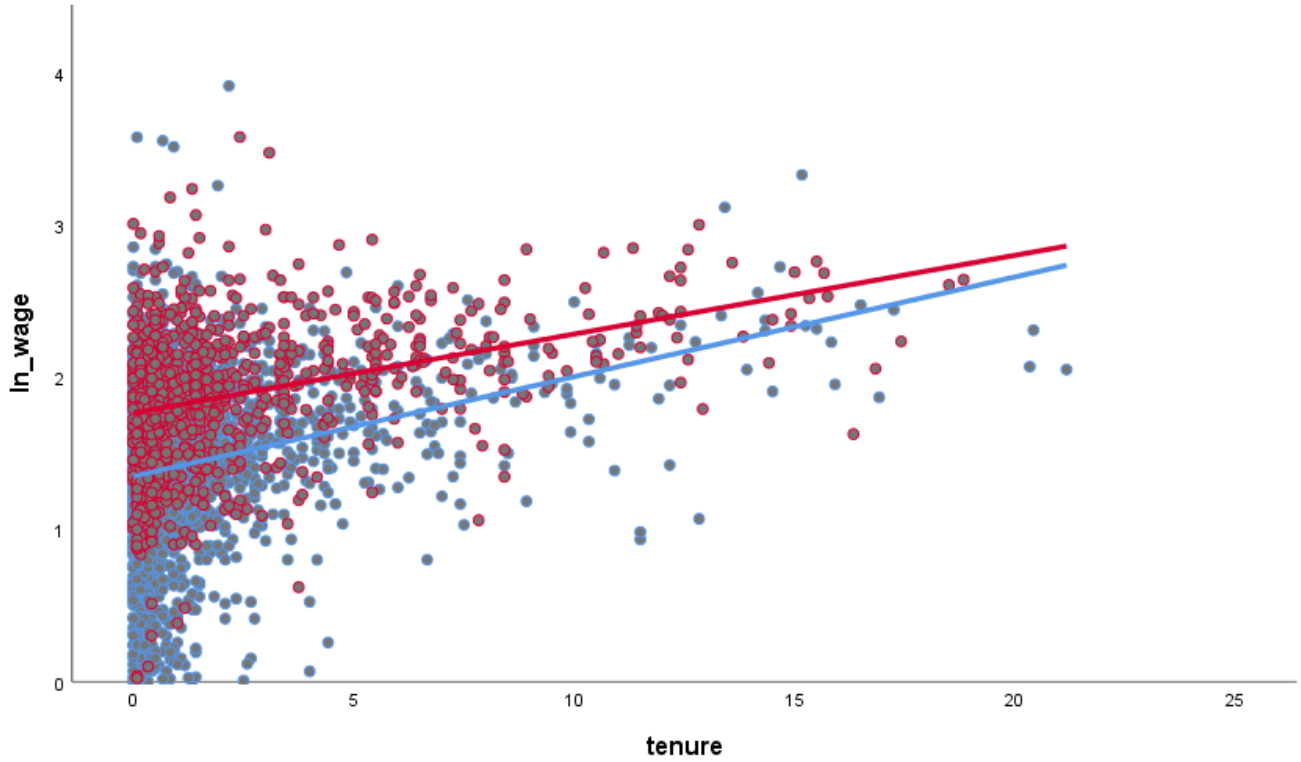
예제 9. nlswork.sav를 사용하여 임금(ln_wage)을 종속변수로, 재직경력(tenure)과 대졸 여부(collgrad)를 독립변수로 하는 회귀식을 추정하시오. 이때 추가적으로 두 변수의 상호작용 효과를 고려하였을 때 결과가 어떻게 달라지는지 해석하시오.

상호작용 효과와 비선형적 관계

- 상호작용항을 포함하지 않은 회귀식과 포함한 회귀식을 각각 따로 추정하여 그 결과를 결과표에 나란히 정리한다.
- 상호작용항을 만들기 위해서는 SPSS에서 [변환]-[변수 계산]을 선택하여 그저 두 변수를 곱하면 된다.
- 상호작용항은 95% 신뢰수준에서 통계적으로 유의하다. 이는 상호작용 효과가 (표본뿐 아니라) 모집단에서도 존재함을 시사한다.
- 상호작용항의 회귀계수는 약 -0.014이므로 대졸자는 재직경력이 길어질수록 임금상승에서 (비대졸자보다) 상대적으로 조금씩 손해를 본다.
- 다만 이 값은 경력효과 0.065보다는 작으므로 (경력에 따른 임금상승의) 주효과를 뒤집을만큼 큰 것은 아니다!

상호작용 효과와 비선형적 관계

- 회귀계수만 봐서는 정확히 어떤 차이가 나타나는지 식별하기 어려우므로 상호작용 효과는 보통 그래프를 그려가면서 파악한다.
- 한 가지 방법은 [그래프]-[회귀 변수 도표]에서 '색상 기준'에 상호작용항을 집어넣은 다음, '옵션'에서 '각 범주형 색상 집단에 대한 회귀선 적합'을 선택하는 것이다.
- SPSS보다 R이나 Stata 등에서 보다 세련된 기법을 사용할 수 있다.



(빨간색은 대졸자, 파란색은 비대졸자임)

상호작용 효과와 비선형적 관계

상호작용항의 유의성 검정 절차는 조금 주의해야 한다.

- 상호작용항이 통계적으로 유의하다면 주요항의 유의성 여부에 대해서는 별도로 신경쓸 필요가 없다(Why?).
- 만약 상호작용항에 이론적으로 주목할 생각이 아니라면, 상호작용항을 회귀식에 넣어서는 안된다(Why?).
- 달리 말해, 상호작용항이 있다면 이것을 무시한 채 주요항만 해석해서는 안된다. 만일 상호작용항이 통계적으로 유의하지 않다면 주요항의 유의성을 판단하기 위해 상호작용항을 제거하고 다시 모형을 추정해야 한다.
- 그러므로 상호작용 효과를 검토할 생각이라면, 상호작용항을 넣지 않은 모형과 넣은 모형을 나란히 보고하는 편이 바람직하다.

상호작용 효과와 비선형적 관계

분석에 사용하는 주요항의 척도는 조금씩 다를 수 있다.

- 위 예제에서 tenure는 양적변수였고, collgrad는 가변수였다.
- 그러므로 주요항의 척도가 어떻게 구성되어 있는가에 따라 아래 세 가지 유형이 나올 수 있다(Why?):
  (1) 양적변수 × 가변수
  (2) 양적변수 × 양적변수
  (3) 가변수 × 가변수.
- 위 예제는 (1) 유형이였다.

상호작용 효과와 비선형적 관계

양적변수 $\times$ 양적변수인 경우에도 크게 다르지 않다.

- 만약 재직경력(tenure)과 연령(age)의 상호작용 효과를 살펴본다면 다음의 회귀식을 세울 수 있다.

$$\text{ln\_wage} = b_0 + b_1 \text{tenure} + b_2 \text{age} + b_3 (\text{tenure} \cdot \text{age})$$

- 회귀식의 추정 절차와 해석은 완전히 똑같다. 그러나 그래프를 그릴 때는 '특정한 연령 대표값'을 부여하여 해석한다.
- 가령 20세를 기준으로 상호작용 효과를 다음과 같이 계산할 수 있다.

$$\begin{aligned} \text{ln\_wage}_{20} &= b_0 + b_1 \text{tenure} + b_2 \cdot 20 + b_3 (\text{tenure} \cdot 20) \\ &= (b_0 + b_2 \cdot 20) + (b_1 + 20 \cdot b_3) \cdot \text{tenure} \end{aligned}$$

상호작용 효과와 비선형적 관계

- 유사하게 연령이 30세인 경우와 40세인 경우의 회귀식을 계산할 수 있다.
- 엑셀에서 tenure에 걸맞는 값을 순차적으로 대입하여 $\widehat{ln_wage}$를 예측한 뒤, 산점도(scatterplot)와 적합선(fitting line)을 그릴 수 있다. 적합선의 '기울기'로 상호작용 효과를 해석한다.
- 조절변수를 숫자형 변수로 그대로 두면 해석이 까다로우므로 대표적인 범주 몇 개만을 골라 거기에 초점을 두고 해석하는 것이 요령이다!
- 사실 양적변수 $\times$ 양적변수의 상호작용 효과 해석은 제법 까다로우므로 조절변수를 아예 처음부터 범주화하는 경우도 제법 많다.
- 가령 연속변수로서의 연령이 아니라 20대, 30대, 40대 가변수로 재부호화하여 사용할 수도 있다.

상호작용 효과와 비선형적 관계

마지막 유형으로 가변수 × 가변수인 경우를 살펴보자.

- 가령 남부(south)와 대졸 여부(collgrad)의 상호작용 효과를 살펴보기 위해 다음의 회귀식을 추정할 수 있다.

$$\text{ln\_wage} = b_0 + b_1 \text{south} + b_2 \text{collgrad} + b_3 (\text{south} \cdot \text{collgrad})$$

- {0, 1}로 더미 코딩하는 이상 이 유형도 쉽게 해석할 수 있다.

$$\begin{aligned}
\text{ln_wage}_{\text{비대졸\&비남부}} &= b_0 && \text{(if collgrad==0 \& south==0)} \\
\text{ln_wage}_{\text{비대졸\&남부}} &= b_0 + b_1 && \text{(if collgrad==0 \& south==1)} \\
\text{ln_wage}_{\text{대졸\&비남부}} &= b_0 + b_2 && \text{(if collgrad==1 \& south==0)} \\
\text{ln_wage}_{\text{대졸\&남부}} &= b_0 + b_1 + b_2 + b_3 && \text{(if collgrad==1 \& south==1)}
\end{aligned}$$

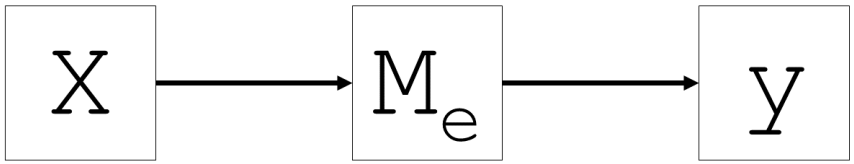
상호작용 효과와 비선형적 관계

- 회귀계수는 각각의 조건에 따른 부스터 크기를 나타낸다. 가령 남부는 비남부에 비해 임금에 b_1 만큼의 차이가 있다.
- 가변수 $\times$ 가변수의 경우에는 그래프를 그릴 때 산점도와 적합선이 아니라 히스토그램(histogram)으로 나타내는 것이 보통이다.
- 예측된 Y, 즉 (1) $Y_{\text{비대졸\&비남부}}$, (2) $Y_{\text{비대졸\&남부}}$, (3) $Y_{\text{대졸\&비남부}}$, (4) $Y_{\text{대졸\&남부}}$를 엑셀에서 각각 계산하여 히스토그램으로 나타낸다.

조절변수의 연습

조절효과와 매개효과는 다른 개념이며 구별되어야 한다.

- (1) 조절효과(moderation effect)는 상호작용 효과를 통해 살펴보고, (2) 매개효과(mediation effect)는 위계적 회귀분석 그리고/또는 경로분석(path analysis)에서 살펴본다.
- 매개효과는 다음과 같은 경로 도식으로 나타낸다.



- 이 특강에서는 시간 부족으로 매개효과를 자세히 다루지 못하지만, (1) Baron-Kenny 매개분석에 관한 여러 교과서와 논문을 참고하고, (2) 경로분석, 더 나아가 구조방정식모형(structural equation modeling)을 공부하는 것을 추천한다.

회귀분석의 가정 및 주의사항

회귀분석을 활용한다면 비선형적 관계도 쉽게 추정할 수 있다.

- 앞서 회귀분석은 두 변수의 관계를 요약하는 직선을 찾아내는 문제라고 설명하였다.
- 그렇긴 하지만, 회귀분석은 매우 탄력적이라서 두 변수 사이에 비선형적 관계 (nonlinear relationship)도 아래처럼 살펴볼 수 있다.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- desmos에서 (위 회귀식과 같은) 2차방정식을 그려보자.
- 눈치빠른 학생들은 비선형적 관계와 상호작용항이 사실 동일한 개념임을 알 수 있을 것이다.

가변수와 범주형 변수

예제 11. HPRICE2.SAV를 사용하여 주택가격(lprice)을 종속변수로 하위계층 비율(lowstat)을 독립변수로 하는 회귀식을 추정하시오. 두 변수 사이의 관계를 직선과 곡선으로 각각 나타내고 어느 쪽이 좀 더 타당한지 추측해보시오. 하위계층 비율과 주택가격의 비선형적 관계를 전제로 한 회귀식으로 추정하고 해석하시오.

회귀분석의 가정 및 주의사항

- X^2을 2차항(quadratic term)이라고 부른다. 단순히 제곱하여 만들면 된다.
- 회귀식 안에는 2차항 뿐 아니라 X도 빼놓지 않고 투입해야 한다(Why?).
- 2차항을 해석할 때는 반드시 그래프를 그리는 편이 좋다(Why?). 무엇보다 x 축의 끄트머리가 40% 정도에 불과하므로 최저점이 어느 위치인지 잘 판단해야 한다.

회귀분석의 가정 및 주의사항

회귀분석의 가정 및 주의사항

단순최소자승 알고리즘은 사실 몇 가지 가정에 입각하고 있다.

- 이 가정들은 (1) 단순최소자승의 해(solution)가 존재하고, (2) OLS 추정량이 어째서 왜곡이 없으며(unbiased), (3) 다른 추정량보다 작은 표준오차만을 가지는지, 즉 효율적(efficient)인지 증명하는데 조용히 쓰인다.
- 재미있게도 교과서에 따라 가정의 목록이 조금씩 다르지만, 사실 다음의 두 형태로 구분된다.

  (1) 고전적 가정(classical assumption)
  (2) 오차항 ϵ에 대한 가정

회귀분석의 가정 및 주의사항

- 먼저 회귀모형에 대한 고전적 가정은 다음과 같다.

  (1) 선형성(linearity): Y와 X의 관계는 선형적으로(linearly) 표현된다.
  (2) 완전공선성 없음(no perfect collinearity): 똑같은 독립변수 두 개 이상 넣지 않는다.
  (3) 비확률적 독립변수(non-stochastic Xs): 독립변수는 외생적이다.
- 어떤 교과서는 (4) 극단치 없음(no outliers)을 포함하기도 한다. (분명히 필요하긴 하지만) 가정이라고 하기엔 좀 무리가 있다.
- 물론 위에 더하여 오차항에 대한 가정도 필요하다!

회귀분석의 가정 및 주의사항

- 다음과 같이 오차항 ϵ에 대한 가정을 나열할 수 있다.

  (1) 외생성(exogeneity)/조건부 영평균(zero conditional mean)

  $$E(\epsilon_i|X_i) = 0$$

  (2a) 등분산성(homoscedasticty)

  $$Var(\epsilon_i|X) = Cov(\epsilon_i, \epsilon_i) = \sigma^2$$

  (2b) 자기상관 없음(no autocorrelation)

  $$Cov(\epsilon_i, \epsilon_j) = 0$$

- 그리고 위 가정들보다 좀 더 강한 가정 하나를 "편의상" 추가한다.

  (3) 정규성(normality)

  $$\epsilon_i \sim N(0, \sigma^2)$$

회귀분석의 가정 및 주의사항

회귀분석의 가정은 고급통계학의 관문과도 같다.

- 회귀모형에서 가정이 깨지면 더이상 OLS가 최적이라고 보장할 수 없다.
- 그러나 가정이 깨져도 이에 대응하는 별도의 기법이 존재한다. 이 고급 기법들은 각각의 가정 위배에 대응하여 다루어진다.
- 그런 의미에서 OLS의 가정에 관해 심도있게 학습하는 것은 고급통계학으로의 관문(gateway)으로 이어진다고 할 수 있다.
- $1 + 1 = 2$ 를 배우는 것은 금방이지만 왜 이 식이 성립하는지 공리(axioms)나 가정을 공부하는 것은 어렵다.
- 더 본격적으로 회귀분석을 공부한다면 가정에 대해 꼼꼼히 학습하고 그 진단과 대응책 역시 다루게 된다(이 특강에서는 더이상 다루지 않는다).

회귀분석의 가정 및 주의사항

우리가 배운 회귀분석은 인과분석이 아니다.

- "상관분석은 상관관계를 살펴보고, 회귀분석은 인과관계를 살펴본다"는 식의 헛소문은 전혀 사실이 아니다.
- 회귀분석을 통해 계산된 회귀계수는 사실 수학적으로 꼼꼼히 따져보면 표준화된 상관계수에 불과하다.
- 비실험적 자료(non-experimental data) 혹은 관찰자료(observational data)를 가지고 평범하게 회귀분석하였다면 단지 상관관계만을 파악한 것이다.
- 물론 인과관계가 아니라고 연구로서 무가치해지는 것은 아니다. 다만 회귀분석의 결과를 해석할 때 (실험설계가 아닌 이상) 마치 인과관계인 것처럼 말하지 않아야 한다.

회귀분석의 가정 및 주의사항

- 보다 엄격한 인과분석을 위한 설험계획(design of experiment; DoE)이 있으며 이것은 앞으로도 계속해서 중요한 분야가 될 것이다.
- 일정 기간에 걸쳐 반복적으로 추적 조사한 자료를 종단분석(longitudinal analysis)하는 기법이 있으며 이 또한 중요한 방법론이다.
- 횡단면 자료(cross-sectional data)인 경우에도 도구변수(instrumental variable)나 경향점수(propensity score) 등을 활용하여 보다 엄격하게 인과적 관계를 추론하려는 접근방법도 있다.
- 이 모든 방법론은 향후 그 중요성을 점점 더할 것으로 전망된다(Why?).

회귀분석의 가정 및 주의사항

통계적 유의성와 실질적 유의성은 다른 개념이다

- 통계적 유의성(statistical significance)을 해석할 때 가장 흔한 실수 중 하나는 유의확률(p-value)이 작은 것을 가지고 선형적 관계의 강도(strength of linear relationship)로 해석하는 것이다.
- 곰곰히 생각해보면, 유의확률은 단지 $H_0 : \beta = 0$라는 옳은 귀무가설을 기각하는 가능성을 보여줄 뿐이다.

회귀분석의 가정 및 주의사항

- 실질적 유의성(substantial significance)은 통계적으로 유의한가 여부와 상관없이 실제로 얼마나 그 강도가 센가의 문제를 다룬다.
- 예컨대 한 시간 게임을 더하게 되면 독서시간이 2분 줄어든다는 발견(Cummings and Vandewater 2007)은 설령 통계적으로 유의하더라도 실질적으로는 그다지 유의하지 않다.
- 그러므로 통계적으로 유의한 결과를 얻었더라도 그 관계를 "실질적인 의미가 담겨있는 언어로 해석하고 음미하여" 실질적 유의성이 얼마나 높은지 판단할 필요가 있다.
- 두 변수 사이의 관계를 보여주는 그래프를 그리라고 자꾸 강조하는 이유도 여기에 있다.

Cummings, Hope M. and Elizabeth A. Vandewater. 2007. "Relation of Adolescent Video Game Play to Time Spent in Other Activities." Archives of Pediatrics Adolescent Medicine 161(7): 684-689.