

# 계량분석

## OLS Assumptions (I)

김현우, PhD<sup>1</sup>

<sup>1</sup>충북대학교 사회학과 조교수

December 6, 2021

# 진행 순서

- 1 회귀분석의 가정
- 2 고전적 가정: 선형성
- 3 고전적 가정: Full Rank
- 4 고전적 가정: 비확률적 독립변수
- 5 고전적 가정: 이상점 없음

## 회귀분석의 가정

## 회귀분석의 가정

단순최소자승(OLS)은 사실 몇 가지 가정에 입각해야만 성립한다.

- 가장 기본적으로 (1) 고전적 가정(classical assumption), (2) 종속변수(y)에 대한 가정, (3) 오차항( $\epsilon$ )에 대한 가정을 구별해야 한다.
- (1)이 가장 넓은 범위에서 가정을 다루고 있고, (2)와 (3)은 궁극적으로 같은 내용이다.
- 교과서에 따라서는 어느 쪽인지 명시하지 않고 대충 넘어가다보니 학생들이 쉽게 혼동하곤 한다.
- 대체로 “오차항에 대한 가정”이 많이 논의되는데 재미있게도 교과서에 따라 가정의 목록이 조금씩 다르다.

# 회귀분석의 가정

- 이 가정들은 가장 기본적으로 (1) OLS 추정량(estimator) 자체를 도출하고 (2) OLS가 왜 BLUE (Best Linear Unbiased Estimator)인지 증명하는데 사용된다.
- 첫번째로 회귀계수와 그 표준오차(SE)를 도출하는 과정에서 수학적으로 “계산이 가능하려면” 혹은 “계산이 용이하려면” 반드시 필요하기 때문에 이 가정들이 요구된다.
- 두번째로 OLS 추정량( $b_k$ 과  $SE_{b_k}$ )이 다른 추정량( $b'_k$ 과  $SE_{b'_k}$ )보다 우월하다는 수학적 증명(Gauss-Markov Theorem)에 이 가정들이 요구된다. 보다 구체적으로, 이 가정들이 성립할 때 OLS 추정량은 (1) 왜곡이 없고(unbiased) (2) (다른 추정량보다) 작은 표준오차만을 가진다(efficient).

불편성(unbiasedness):  $E(b_k) = \beta_k$

효율성(efficiency):  $SE_{b_k} < SE_{b'_k}$

## 회귀분석의 가정

- 사실 이 내용은 상당히 복잡하기 때문에 사회과학 통계학 분야에서는 대학원에서도 다루어지는 문제로 여겨진다. 하지만 당황스럽게도 사회조사분석사 2급에서 몇 차례나 기출문제가 등장했다.
- 사회조사분석사 2급 수험서 레벨에서 오차항에 관해 흔히 세 가지 가정만 달랑 맥락없이 언급된다.
  - 정규성(normality):  $\epsilon_i \sim N(0, \sigma^2)$
  - 등분산성(homoscedasticity):  $\text{Var}(\epsilon_i|X) = \sigma^2$
  - 독립성(independence):  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$
- 이런 식으로 나열하는 것은 조금 논리적으로 엉성하지만 적어도 틀린 것은 아니니 받아들여도 된다.

# 회귀분석의 가정

좀 더 논리적으로 오차항에 대한 가정은 다음과 같다.

- ① 조건부 영평균(zero conditional mean):  $E(\epsilon_i|X_i) = 0$
- ② 등분산성(homoscedasticity):  $\text{Var}(\epsilon_i|X) = \text{Cov}(\epsilon_i, \epsilon_i) = \sigma^2$
- ③ 자기상관 없음(no autocorrelation):  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$
- ④ 정규성(normality):  $\epsilon_i \sim N(0, \sigma^2)$

# 회귀분석의 가정

“오차항에 대한 가정”을 포함하여 회귀모형에 대한 고전적 가정(classical assumptions)은 좀 더 광범위하다.

- ① 선형성(linearity):  $y$ 와  $X$ 의 관계는 선형적으로 표현된다.
- ② Full Rank: 사례 수( $n$ )는 적어도 독립변수의 수( $k$ )보다 두 개 많고, 똑같은 독립변수가 두 개 이상 존재하지 않는다.
- ③ 비확률적 독립변수(non-stochastic  $X_s$ ): 독립변수는 외생적이다.
- ④ 어떤 교과서는 극단치 없음(no outliers)을 포함하기도 한다. 이해는 하지만 가정이라고 하기엔 좀 무리가 있다.



## 고전적 가정: 선형성





## 고전적 가정: 선형성

이 가정의 실질적인 타당성 여부는 데이터에 그 선형모형이 얼마나 잘 적합 (fit)하는가에 달려있다([Stata 코드] 참고).

- 설령 형식적으로 선형성 가정을 충족하더라도 그 선형모형이 “실질적으로” 데이터에 잘 맞지 않을 수 있다.
- eCampus에서 transit.csv를 다운받아 Stata에서 열어보자. map은 버스지도를 무료 배부 부수(단위 1,000부)를 의미하고 rider는 증가한 버스 승객의 수(단위 1,000명)이다. 관찰단위는 시이다.
- 첫번째로 rider를 y축으로, map을 x축으로 산포도와 적합선(fitting line)을 그려보면 선형모형의 적합도가 낮음을 확인할 수 있다.
- 두번째로 (더 세련된 방법은) 이른바 **RVF (Residual-Versus-Fitted) 도표(plot)**를 그려보는 것이다. 이것은 오차항(residuals)을 y축으로, 예측된 y (fitted values)를 x축으로 하여 그린 산포도(와 적합선)를 의미한다.
- 직선의 적합선이 매우 부적합함을 두번째 산포도에서 선명하게 확인할 수 있다.

## 고전적 가정: 선형성

- 한마디로 **그림을 그려보아 선형모형이 데이터에 적합한지 살펴보는 것**이 핵심이다. 엄밀하게 말해 여기까지 오면 더이상 형식적 가정 문제라기보다 실질적인 선형모형의 타당성 문제라고 보아야 한다.
- 일단 모든 X에 대해 y와의 산포도를 그려보는 것이 첫 출발점이다. 모든 변수에 대해 그리기 귀찮기 때문에 Stata의 **graph matrix** 명령어를 사용할 수도 있다.
- 단 하나의 산포도로 살펴볼 수 있는 가장 세련된 방법은 위에서 설명한 RVF 도표이다. 이것은 “예측된 y (fitted values)”를 사용한다는 점에서 여러 그래프를 하나로 압축하고 있는 셈이다(Why?).
- RVF 도표는 선형성 뿐 아니라 이상점(outliers)이나 이분산성(heteroscedasticity) 식별 등에도 좋은 출발점이 된다!
- 선형모형이 부적합한 것 같다면 로그 변환이나 이차항을 사용하고 모형 적합도의 개선 여부를 판단한다.

## 고전적 가정: Full Rank

## 고전적 가정: Full Rank

이것은 완전히 수학적인 계산 과정에 관한 가정이다.

- Full rank는 우리말 번역이 마땅치 않은 선형대수학(linear algebra)의 개념이다. 만약 선형대수학을 한 학기 정도 이수하지 않으면 이 개념을 이해하기 어렵다.
- 하지만 우리 모두는 적어도 “연립방정식(simultaneous equations)이 주어졌을 때, **미지수(unknowns)의 고유한 해**를 구할 수 있는 어떤 조건들”이 있고, 그 조건들이 충족되지 않으면 미지수가 무한히 많거나 아예 구할 수 없거나 함을 알고 있다. Full Rank는 바로 이렇게 고유한 미지수, 즉 상수와 회귀계수를 구할 수 있는 조건들을 의미한다.
- Full Rank가 위배되는 상황은 크게 다음 두 가지를 지적할 수 있다:
  - (1) 변수의 수에 비해 표본 크기가 작은 경우
  - (2) 둘 이상의 완전히 똑같거나 **선형의존적(linear dependent)**인 변수가 있는 경우

# 고전적 가정: Full Rank

독립변수의 수( $k$ )에 비해 표본 크기( $n$ )가 작은 경우는 문제가 된다([Stata 코드] 참고).

- 얼마나 표본 크기가 작으면 이런 문제가 생길까? 최소한 다음이 성립해야 한다.

$$n \geq k + 2$$

- 예컨대 독립변수가 2개라면 샘플은 최소 5개가 되어야 회귀분석이 작동한다.
- 오늘날 사회과학 통계학은 대부분 대규모 표본을 사용한다. 변수가 다소 많아지더라도 표본의 크기만큼이나 많아지는 경우는 사실상 없다.



## 고전적 가정: Full Rank

- 얼마나 표본 크기가 작으면 이런 문제가 생길까? 최소한 다음이 성립해야 문제가 없다.

$$n > k + 2$$

- 예컨대 독립변수가 2개라면 샘플은 최소 5개가 되어야 회귀분석이 작동한다.
- 오늘날 사회과학 통계학은 대부분 대규모 표본을 사용한다. 독립변수가 다소 많아지더라도 표본의 크기만큼이나 많아지는 경우는 사실상 없다.
- 그것과는 별개로 “독립변수 당 최소한 20개 정도 표본이 있어야 한다”는 이야기를 할 때가 있다. 예를 들어 5개 독립변수를 사용하고 싶다면 표본 수가 100개는 되어야 한다는 식이다. 이것은 회귀분석의 가정과는 별개로 믿을 수 있는 추정이 가능하기 위한 필요조건으로 받아들여야 한다.

## 고전적 가정: Full Rank

둘 이상의 완전히 똑같거나 선형의존적 변수가 들어가는 경우는 흔하게 일어나지만 통계분석 패키지가 알아서 제거해준다([Stata 코드] 참고).

- 어떤 변수들은 종종 (거의) 실질적인 의미에서 차이가 없다. 예를 들어, 횡단면 분석(cross-sectional analysis)의 맥락에서 연령(age)과 태어난 해(birth year)를 동시에 독립변수로 고려하는 것은 아무런 의미도 없다(Why?)
- 선형의존적이라는 말은 선형독립적(linear independent)이 아니라는 의미이다. 예컨대 두 변수  $X_1$  와  $X_2$  가 있을 때,  $X_2 = a + bX_1$  와 같은 선형식이 성립하면  $X_2$  는  $X_1$  에 대해 선형의존적이다. 이 경우  $X_1 = (a/b) + (1/b)X_2$  역시 성립하므로  $X_1$  도  $X_2$  에 대해 공히 선형의존적이다.
- 이렇게 완전히 똑같거나 선형의존적인 변수가 존재하는 상황을 완전공선성(perfect collinearity)이라고 부른다.
- $X_2$  가  $X_1$  에 의해 완벽하게 설명되므로 이런 경우 역시 사실상 똑같은 변수를 여러 개 집어넣은 것과 같다.

## 고전적 가정: Full Rank

완전공선성의 가정 성립과 위반 사이에는 약간의 회색지대가 있다.

- 완전공선성까지는 아니지만 공선성(collinearity)의 정도가 매우 높아 모형의 추정 결과가 불안정해지는 현상을 **다중공선성(multicollinearity)**라고 부른다.
- 완전공선성은 가정 위배이지만 다중공선성은 그 자체로 가정 위배는 아니다.
- 다중공선성이 존재하는가를 식별하는 가장 기본적인 방법 두 가지는 (1) **상관계수행렬(correlation coefficient matrix)**을 살펴보는 것과 (2) **분산팽창인자(Variance Inflation Factors; VIF)**를 살펴보는 것이다.
- 상관계수행렬을 살펴보면 구체적으로 어떤 두 변수 사이의 상관계수가 지나치게 높은지 파악할 수 있다. 많은 연구논문이나 보고서에서 상관계수행렬을 보고하는 것은 이 때문이다.
- 그러나 이 방식은 오로지 두 변수 사이에서 나타나는 공선성 문제만 볼 수 있기 때문에 제한점이 있다. **한 변수가 다른 여러 변수들과 조금씩 공선성을 가져 결국 종속변수의 변량(variation)을 설명할만큼 충분히 독자적인 변량을 갖지 못할 수도 있다.**

## 고전적 가정: Full Rank

분산팽창인자는 좀 더 세련된 다중공선성 진단법으로 알려져 있다([Stata 코드] 참고).

- 직관적으로 다중공선성은 독립변수 사이의 지나치게 밀접한 관계 때문에 발생하는 문제이다. 그러므로 “특정 독립변수”가 얼마나 다른 독립변수들에 의해 지나치게 잘 설명된다면 다중공선성 문제의 원인이 될 것이다.
- 그러므로 특정 독립변수를 새로운 종속변수로, 나머지 독립변수를 그대로 독립변수로 하여 새로운 회귀분석을 수행하고 그 결정계수( $R^2$ )를 구한다.
- 1에서 결정계수를 뺀 값( $1 - R^2$ )을 **공차(tolerance)**라고 부른다. 곰곰히 생각해보면 공차는 “특정 독립변수”가 다른 독립변수들에 의해 설명되지 않은 정도를 보여준다 (Why?). 이것은 “특정 독립변수”의 (다른 독립변수로부터의) 상대적 독립성을 보여준다.
- 그런데 우리는 문제의 심각성을 알고 싶으므로 이 공차의 역(inverse)을 취해야 한다. 이 값이 바로 해당 “특정 변수”의 분산팽창인자(VIF)이다.

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

## 고전적 가정: Full Rank

- 분산팽창인자가 얼마나 크면 문제인지 대략적인 지표(rule-of-thumb)가 교과서에 따라 제각각이다. 엄격하게 5 이상은 문제라고 말하거나 15 까지도 괜찮다고 말하기도 한다.
- 분산팽창인자는 특히 집계자료(aggregate data)를 사용할 때 커지는 경향이 있다. 개인을 분석단위로 삼으면 그렇게까지 크지 않았을 두 변수(예컨대 소득과 최종학력) 사이의 상관계수도 국가를 분석단위로 삼으면 매우 커지기 때문이다.
- 상호작용항이나 다항함수를 사용할때도 분산팽창인자가 커진다. 이로 인한 문제는 너무 고민하지 않아도 된다. 정 신경쓰이면 평균중심화(mean centering)를 하자.
- 개별 변수의 분산팽창인자 뿐 아니라 평균 분산팽창인자(Mean VIF)에도 주목하자. 1보다 훨씬 크면 대응책을 고민해야 한다(Hamilton 1992).

## 고전적 가정: Full Rank

공선성 문제에 대한 몇 가지 정형화된 대응책이 이미 준비되어 있다.

- 완전공선성이 나타나는 가장 흔한 이유는 연구자가 실수로 똑같은 변수를 두 번 집어넣거나 범주형 변수를 더미 코딩하고 모두 다 집어넣은 경우이다. 그러므로 첫번째 대응책은 똑같은 변수 중 하나를 제거하는 것이다. 정 원한다면 변수를 따로따로 넣은 모형을 여러 개 추정하고 나란히 비교하여 보고할 수도 있다.
- 두번째 대응책은 좀 더 복잡한데, 합성지수(composite index) 같은 잠재변수(latent variable)를 만들어 이를 사용하는 것이다. 둘 이상의 변수가 “그렇게나 유사하다면” 아예 하나로 합친 새로운 변수를 만들어 분석에 사용할 수 있기 때문이다.
- 세번째 대응책은 능형회귀(ridge regression)나 라쏘(LASSO regression)처럼 우도함수(likelihood function)에 패널티 항(penalty term)을 넣는 특수한 알고리즘을 사용하는 것이다. 머신러닝(machine learning) 분야에서는 제법 많이 사용된다.

## 고전적 가정: 비확률적 독립변수

고전적 가정: 비확률적 독립변수

독립변수는 외생적으로 주어져 고정되어있고(fixed) 확률적으로 변화하지 않아야 한다(non-stochastic).

- 통계 용어 stochastic은 적절한 번역어가 없다. probability, random, stochastic은 확률이라는 단어로 대충 뭉뚱그려진다.
- 독립변수에는 샘플링에 따른 변동(sampling variation)가 나타나지 않아야 하는데, 평범하게 생각하면 독립변수 X가 데이터의 형태로 주어져 있는 것이 당연하게 들릴 수도 있다.
- 이 가정에 따르면 독립변수의 값은 “모형 외부로부터” 결정된다고 전제된다. 그러므로 이론상 “모형 내부에서” 일종의 방정식 시스템(equation system)이 존재한다면 이 가정을 반드시 완화해야 한다. 경제학에서는 일찌감치 방정식 시스템을 도입하는 편이고, 그 밖의 사회과학에서는 조금 나중에 다층회귀모형(multilevel regression modeling)에서 방정식 시스템을 도입한다.





고전적 가정: 이상점 없음

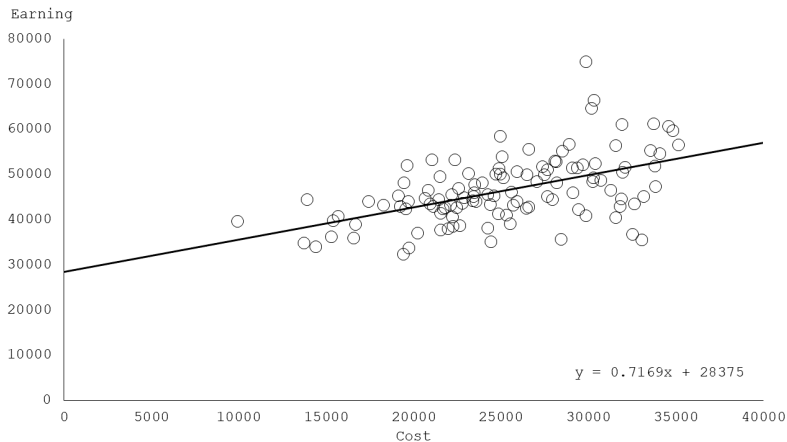
고전적 가정: 이상점 없음

이상점(outliers)이 존재하면 추정량은 심각하게 왜곡된다. 그것도 아주 심각하게!

- 이것은 사실 가정이라고 하기엔 좀 무리가 있다. 하지만 회귀분석 결과의 실질적인 의미를 해치는 굉장히 심각하고 중요한 문제이다. 개인적으로 모형을 점검할 때 가장 중요한 요소라고 본다.
- 이상점의 가장 초보적이고 흔한 원인은 사람의 실수이다. 자료를 입력한 사람의 실수 등으로 인해 특정 변수 특정 케이스에 너무 크거나 너무 작은 값이 들어갈 수 있다.
- 그러므로 데이터를 꼼꼼히 훑어보고 소팅 또는 필터링해가며 살펴보는 것이 중요하다.

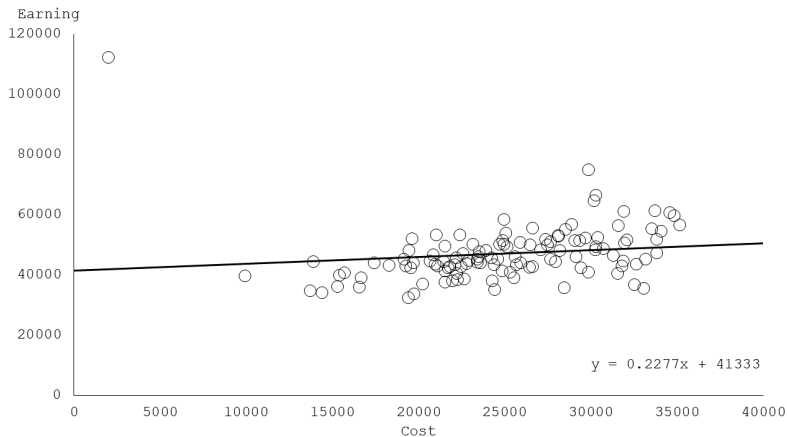
# 고전적 가정: 이상점 없음

- 아래 그래프는 등록금(cost)과 졸업생 평균수입(earnings)의 관계의 예제이다.



# 고전적 가정: 이상점 없음

- 단 하나의 이상치만으로도 그 관계는 심각하게 왜곡된다.



# 고전적 가정: 이상점 없음

이상점 식별(outlier detection)은 사실 그 자체로 굉장히 큰 주제이다.

- 이상점의 존재를 식별하는 수리적 알고리즘도 이미 상당히 발달해있고 기법도 다양하게 개발되었다. 대표적인 영역들로 선거부정(election frauds), 신용카드 사기(credit frauds), 게임 어뷰징(game abusing) 식별 등이 있다. 이것들을 이해하려면 좀 깊이있는 공부가 요구된다.
- 여기서 우리는 사회과학 통계학에서 기본적으로 잘 알려진 이상점 식별(또는 영향력있는 사례 식별)의 두 가지 도구만을 공부한다:
  - (1) 스튜던트화된 잔차(Studentized residuals)
  - (3) Cook's Distance







고전적 가정: 이상점 없음

영향력있는(influential) 사례의 유무는 이상점의 유무와 별개 문제이다  
([Stata 코드] 참고).

- 이상점이 있더라도 실질적으로 거의 영향을 미치지 못할 수도 있다. 예를 들어 “이상점이 고르게 분포하거나 적합선 상에 놓여있어” 적합선의 기울기에 영향을 미치지 못하는 상황을 상상해보자!
- 그러니 이상점 유무와 영향력 유무는 좀 별개의 문제로 접근해야 한다.
- 영향력있는 사례를 식별하는 방식으로 Cook's Distance, DBFITS, DFBETAS가 특히 널리 알려져 있다.

# 고전적 가정: 이상점 없음

- Cook's Distance는 특정 사례가 “예측된  $y$  (fitted values) 전체”에 미치는 영향력을 추정한다.

$$D_i = \frac{\sum (\hat{y} - \hat{y}_{(i)})^2}{kMSE}$$

- $k$ 는 추정된 회귀계수의 수를 지칭한다.
- “특정 사례”를 데이터에서 일시적으로 제거한 뒤 추정하여 그것이 예측된  $y$  (fitted values)에 미치는 영향을 평가한다.

고전적 가정: 이상점 없음

이상점/영향력있는 사례의 제거는 서두르지 말고 천천히 단계적으로 수행한다([Stata 코드] 참고).

- 스튜던트화된 잔치를 계산한 뒤에는 대략적인 기준(rule-of-thumb)으로 절대값이 2보다 큰 사례를 삭제할 수 있다(Neter et al 2004).
- Cook's Distance를 계산한 뒤에는 대략적인 기준으로 1보다 크거나  $4/n$ 보다 큰 사례를 삭제할 수 있다(Neter et al 2004).
- 기계적으로 위와 같이 적용할 수도 있다. 일단 나의 의견은 일단 히스토그램을 그려보고 너무 나간 사례들을 “단계적으로 제거하면서” 영향을 살펴보라는 것이다.
- 또한 이상점/영향력있는 사례를 제거한 결과만 보고하기보다는, 포함하기도 하고 제거하기도 하여 각각 따로 모형을 추정하고 결과를 함께 보고할 수도 있다. 특히 온라인 부록(online appendix)으로 이와 같은 결과를 보고할 수도 있다.

고전적 가정: 이상점 없음

이상점과 영향력있는 사례는 개념적으로 분명히 구분된다([Stata 코드] 참고).

- 우리는 스튜던트화된 잔차를 살펴보아 “오차항의 크기에 기반하여” 이상점을 식별하였다.
- 반면 Cook's distance는 “예측된  $y$ 의 크기에 기반하여” 영향력있는 사례를 식별하였다.
- 서로 상이한 두 측면을 동시에 고려하는 전략도 있다. 이를 그림으로 나타낸 것이 이른바 **LVR (Leverage-Versus-Residual-Squared) 도표(plot)**로 “두 값 모두가 크면” 삭제를 고려하게 된다.