계량분석
Classical Regression Assumptions

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 14, 2024

진행 순서

1. 회귀분석의 가정
2. 고전적 가정: 선형성
3. 고전적 가정: 완전 랭크
4. 고전적 가정: 비확률적 독립변수
5. 고전적 가정: 이상점 없음

회귀분석의 가정

회귀분석의 가정

보통최소자승(OLS)은 사실 몇 가지 가정에 입각해야만 성립한다.

- 기본적으로 (1) 고전적 가정(classical assumption), (2) 종속변수 Y에 대한 가정, (3) 오차항 ϵ에 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 가정이 많이 논의되는데, 교과서에 따라 가정의 목록이 조금씩 다르기 때문에 주의를 요한다.

회귀분석의 가정

- 이 가정들은 가장 기본적으로 (1) OLS 추정량(estimator) 자체를 도출하고 (2) OLS가 왜 BLUE (Best Linear Unbiased Estimator)인지 증명하는데 사용된다.
- 첫째, OLS 추정량 b와 se_b를 도출하는 과정에서 수학적으로 "계산이 가능하려면" 혹은 "계산이 용이하려면" 몇몇 가정이 필요하다.
- 둘째, OLS 추정량 b와 se_b이 다른 방식을 통한 추정량 b'와 $se_{b'}$보다 우월하다는 수학적 증명(Gauss-Markov Theorem)에 가정이 요구된다.
- 보다 구체적으로, 이 가정들이 성립할 때 OLS 추정량은 (1) 왜곡이 없고(unbiased) (2) (다른 추정량보다) 표준오차가 작다(efficient).

$$\text{불편성(unbiasedness)}: \; E(b) = \beta$$
$$\text{효율성(efficiency)}: \; se_b < se_{b'}$$

회귀분석의 가정

먼저 오차항에 대한 가정은 다음과 같다.

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

회귀분석의 가정

회귀모형에 대한 고전적 가정은 좀 더 광범위하다.

1. 선형성(linearity): Y와 X의 관계는 선형적으로 표현된다.
2. 완전 랭크(full rank): 모든 독립변수는 선형독립(linearly independent)이다. 사례수 n은 적어도 독립변수의 수 k보다 많다.
3. 비확률적 독립변수(non-stochastic Xs): 독립변수는 외생적(exogeneous)이다.
4. 어떤 교과서는 이상점 없음(no outliers)을 포함하기도 한다.
5. 그리고 오차항에 관한 가정들도 여기에 포함된다.

고전적 가정: 선형성

고전적 가정: 선형성

OLS는 선형모형의 오차를 최소화하는 상수와 회귀계수를 추정한다.

- 여기서 주목해야 할 부분은 선형모형(linear model)이라는 부분이다. 첫번째 고전적 가정은 바로 선형성(linearity)이다.
- 아래 회귀식에서 모든 X들과 Y의 관계는 선형적이다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

- 이 가정과 관련하여 형식적인 의미와 실질적인 의미 두 가지를 모두 충족해야 한다.

고전적 가정: 선형성

- 형식적으로 이 가정은 회귀식이 선형방정식으로 세워질 것을 전제로 하는데, 보통 변수를 데이터에서 골라 별 생각없이 모형에 집어넣으면 자연스럽게 충족된다.
- 엄밀한 수리모형(mathematical model)을 강조하면, 아래와 같은 (유명한) 비선형모형을 제시할 수도 있다.

$$Y = A \cdot L^{\alpha} \cdot K^{\beta} = \beta_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2}$$

- 그래도 걱정이 없다. 양변을 로그 변환(log transformation)하면 선형성을 확보할 수 있다.

$$\begin{aligned} ln(Y) &= ln(\beta_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2}) \\ &= ln\beta_0 + \beta_1 lnX_1 + \beta_2 lnX_2 \end{aligned}$$

고전적 가정: 선형성

- 물론 상호작용항(interaction term)이 존재하는 경우에도 비선형적이다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$$

- 이차항(quadratic term)이 존재하는 경우에도 비선형적이다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

- 보통 회귀모형을 아무 생각없이 선형모형으로 만들기 때문에 의식하지 않는 경우가 많다(Why?). 적절한 함수 변환으로 선형성을 확보하기도 쉬운 편이다.

고전적 가정: 선형성

선형모형이 자료에 실제로 잘 적합하는가를 살펴보자.

- 설령 형식적으로 선형성 가정을 충족하더라도, 그 선형모형이 "실질적으로" 데이터에 잘 맞지 않을 수 있다!
- transit.csv에서 map은 버스지도를 무료 배부 부수(단위 1,000부)를 의미하고 rider는 증가한 버스 승객의 수(단위 1,000명)이다. 관찰단위는 시이다.
- 첫번째로 rider를 Y 축으로, map을 X 축으로 산포도와 적합선을 그려보면 선형모형의 적합도가 낮음을 확인할 수 있다.
- 두번째로 (더 세련된 방법은) 이른바 RVF 도표(Residual-Versus-Fitted plot)를 그려보는 것이다. 이것은 오차항(residuals)을 Y 축으로, 예측된 Y (fitted values)를 x축으로 하여 그린 산포도(와 적합선)를 의미한다.
- 직선 적합선이 적합하지 않음을 두번째 산포도에서 선명하게 확인할 수 있다.

고전적 가정: 선형성

- 그림을 그려보아 선형모형이 자료에 적합한지 살펴보자.
- 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 다른 가정 위배 여부의 식별에도 좋은 출발점이 된다.
- 선형모형이 부적합한 것 같다면 로그 변환, 이차항(quadratic term) 등을 사용하고 모형 적합도의 개선 여부를 판단한다.

고전적 가정: 완전 랭크

고전적 가정: 완전 랭크

이것은 오로지 수학적인 계산에 관한 가정이다.

- 완전 랭크(full rank)는 우리말 번역이 마땅치 않은 선형대수학(linear algebra)의 개념이다. 선형대수학을 이수하지 않으면 이 개념을 이해하기 어렵다.
- 연립방정식(simultaneous equations)이 주어졌을 때, 미지수(unknowns)의 고유한 해(solution)를 구할 수 있는 어떤 조건들이 충족되지 않으면 해가 무한히 많거나 아예 구할 수 없다(Why?).
- 완전 랭크는 바로 이렇게 고유한 미지수, 즉 상수와 회귀계수를 구할 수 있는 최소한의 조건이 된다.
- 독립변수들이 선형종속적(linear dependent)인 경우나 변수의 수보다 관측치의 수가 작은 경우 등에서 완전 랭크 가정이 위배된다.

고전적 가정: 완전 랭크

독립변수의 수 k에 비해 표본 크기 n가 작으면 문제가 된다.

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

$$n \geq k + 2$$

- 예컨대 독립변수가 2개라면 표본 수는 최소 5개가 되어야 한다.
- 오늘날 사회과학 통계학은 대부분 대규모 표본을 사용한다. 독립변수가 다소 많아지더라도 표본의 크기만큼이나 많아지는 경우는 사실상 없다.
- "독립변수 당 최소한 20개 정도 표본이 있어야 한다"는 이야기를 할 때가 있다. 이것은 회귀분석의 가정과는 별개로 믿을 수 있는 안정적인 추정이 가능하기 위한 조건으로 받아들여야 한다.

고전적 가정: 완전 랭크

선형종속적 변수가 들어가면 통계분석 패키지가 보통 알아서 제거해준다.

- 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?)
- 선형종속적(linearly dependent)이면 안된다. 예컨대 두 변수 X_1와 X_2가 있을 때, 아래와 같은 선형식이 성립하면 X_2는 X_1에 대해 선형종속적이다.

$$X_2 = \gamma_0 + \gamma_1 X_1$$

- 이 경우 아래 역시 성립하므로 X_1도 X_2에 대해 공히 선형의존적이다.

$$X_1 = \frac{\gamma_0}{\gamma_1} + \frac{1}{\gamma_1} X_2$$

고전적 가정: 완전 랭크

- 두 독립변수가 선형종속적인 것은 결국 똑같은 두 독립변수를 집어넣은 것과 같다 (Why?).
- 가령 나이와 출생연도 사이에도 “태어난 해 = 올해 연도 - 나이” 라는 선형식이 성립한다.
- 이렇게 완전히 똑같거나 선형종속적인 독립변수가 존재하는 상황을 완전공선성 (perfect collinearity)이라고 부른다.

고전적 가정: 완전 랭크

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.

고전적 가정: 완전 랭크

- 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(Why?).
- 한 변수가 다른 여러 변수들과 조금씩 공선성을 가져 결국 종속변수의 변량(variation)을 설명할만큼 충분히 독자적인 변량을 갖지 못할 수도 있다.

고전적 가정: 완전 랭크

분산팽창인자는 좀 더 세련된 다중공선성 진단법으로 알려져 있다.

- 직관적으로 다중공선성은 독립변수 사이의 지나치게 밀접한 관계 때문에 발생하는 문제이다. 그러므로 특정 독립변수가 얼마나 다른 독립변수들에 의해 지나치게 잘 설명된다면 다중공선성 문제의 원인이 될 것이다(Why?).
- 바로 이 논리를 반영하여 다중공선성 문제의 심각도를 측정할 수 있다.
- 먼저 특정 독립변수를 새로운 종속변수로, 나머지 독립변수를 그대로 독립변수로 하여 회귀분석하고 그 결정계수 R^2를 구한다.

고전적 가정: 완전 랭크

- 1에서 결정계수를 뺀 값 $1 - R^2$을 공차(tolerance)라고 부른다.
- 곰곰히 생각해보면 공차는 특정 독립변수가 다른 독립변수들에 의해 설명되지 않은 정도를 보여준다(Why?).
- 이것은 특정 독립변수의 (다른 독립변수로부터의) 상대적 독립성을 보여준다.
- 그런데 우리는 문제의 심각성을 알고 싶으므로 이 공차의 역(inverse)을 취해야 한다. 이 값이 바로 k번째 독립변수의 분산팽창인자(VIF)이다.

$$VIF_k = \frac{1}{1 - R_k^2}$$

고전적 가정: 완전 랭크

- 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 그렇게 까지 크지 않았을 두 변수(예컨대 소득과 최종학력) 사이의 상관계수도 국가를 분석단위로 삼으면 매우 커진다(Why?).
- 상호작용항이나 다항함수를 사용할때도 분산팽창인자가 커진다. 그러나 이 경우에는 너무 고민하지 않아도 된다.
- 개별 변수의 분산팽창인자 뿐 아니라 평균 분산팽창인자(Mean VIF)에도 주목하자. 1보다 훨씬 크면 대응책을 고민해야 한다(Hamilton 1992).

고전적 가정: 완전 랭크

공선성 문제에 대한 몇 가지 정형화된 대응책이 이미 준비되어 있다.

- 완전공선성이 나타나는 가장 흔한 이유는 연구자가 실수로 똑같은 변수를 두 번 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 랏소회귀(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 있고 확률적으로 변하지 않는다.

- 통계 용어 stochastic은 적절한 번역어가 없다. 연관 용어인 probability나 random과 함께 확률이라는 단어로 대충 뭉뚱그려진다.
- 평범하게 생각하면 독립변수 X가 자료에서 이미 주어져 있는 것이 당연하게 들릴 수도 있다. 그러나 이 가정은 $E(Y|X)$의 계산에서 필수적이다.
- 독립변수 X_i가 실험설계에서처럼 무작위 처리(random treatment)로 주어지는 것을 상상하자.
- 게다가 설령 "독립변수가 확률적(stochastic)이다"라고 가정하더라도, 다시 오차항에 대한 가정 $Cov(\epsilon, X) = 0$이 성립하면 아무런 문제도 없다(나중에 다시 다룬다).

고전적 가정: 이상점 없음

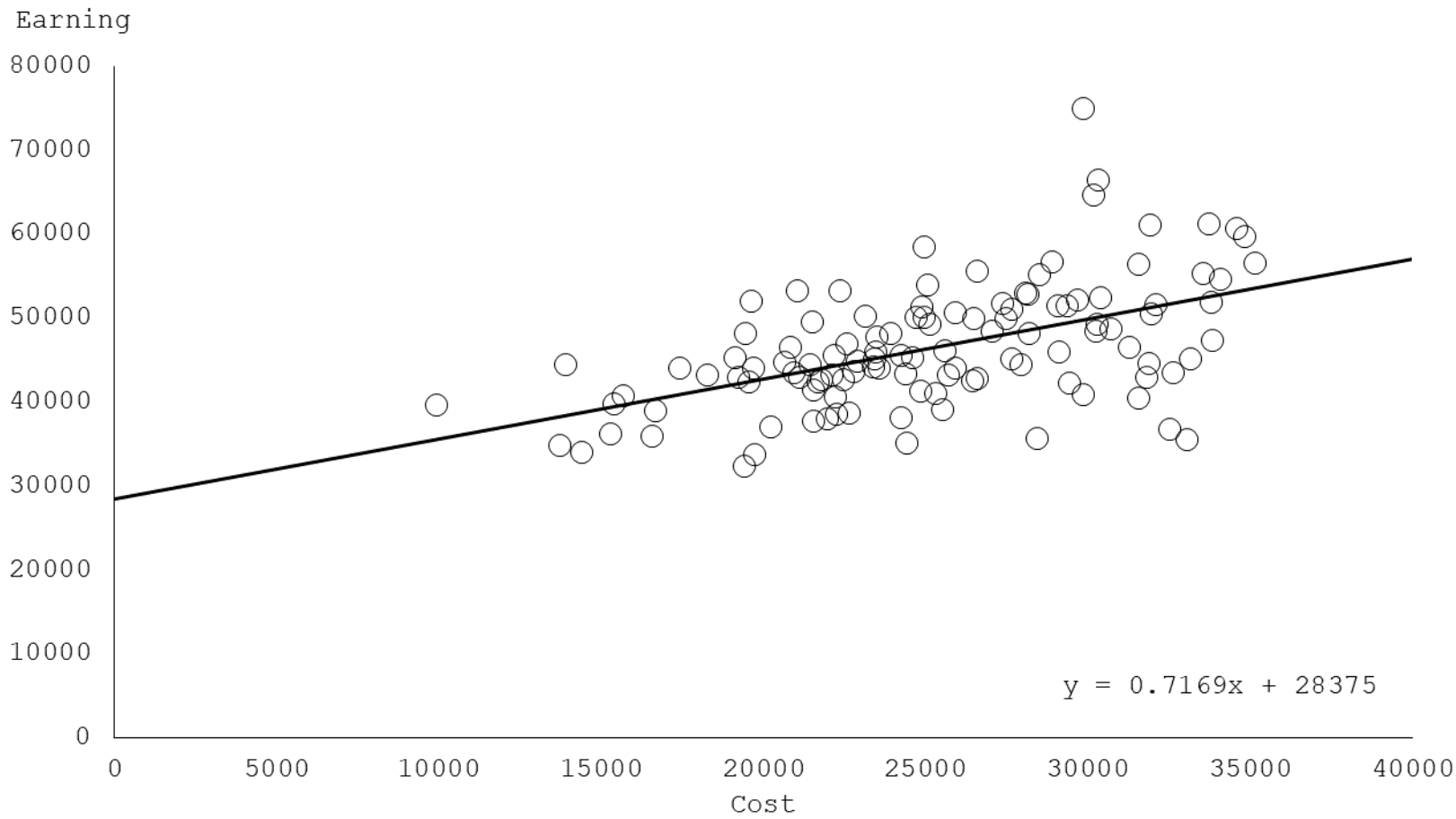
고전적 가정: 이상점 없음

이상점이 존재하면 추정량은 심각하게 왜곡된다. 그것도 아주 심각하게!

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

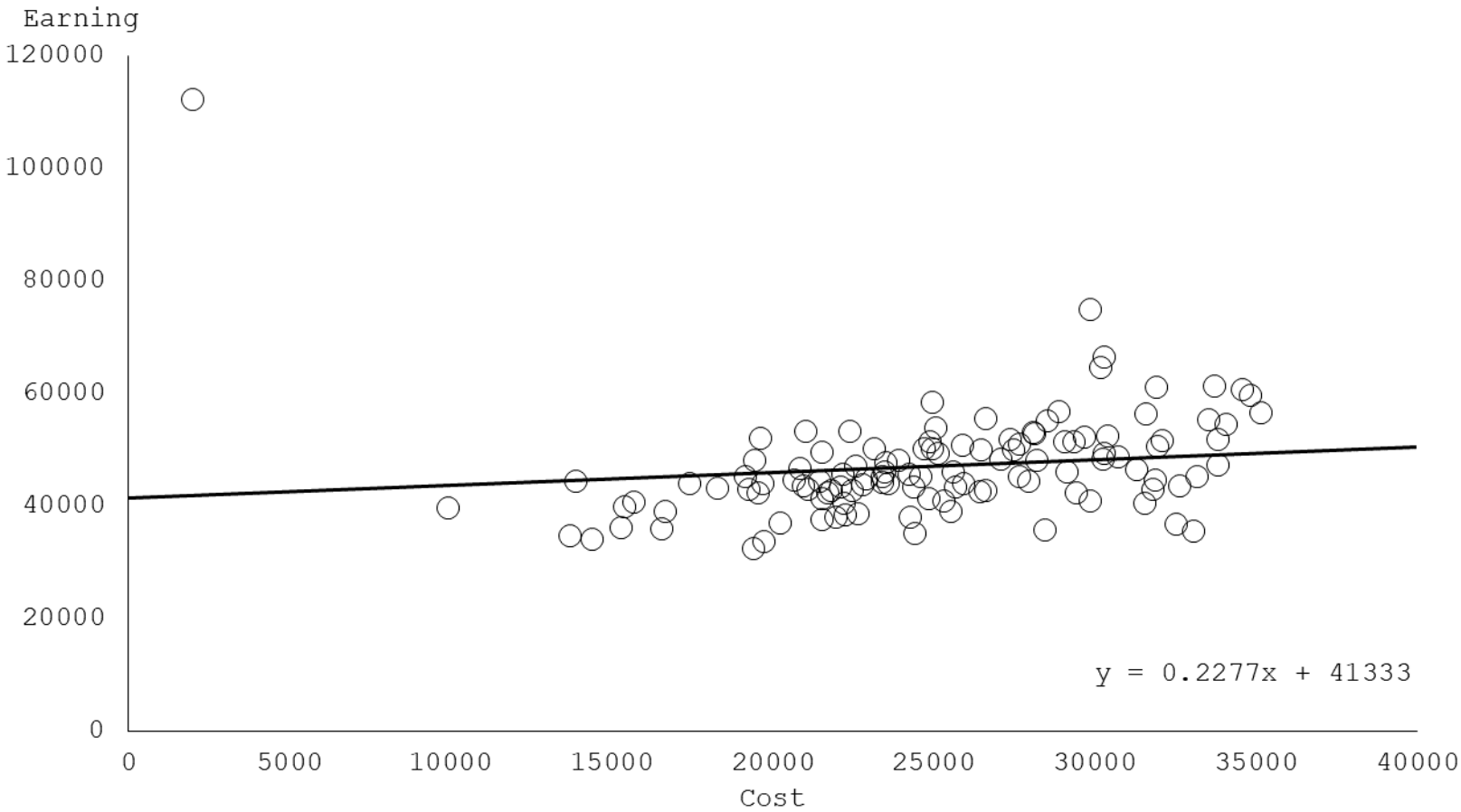
고전적 가정: 이상점 없음

- 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 식별은 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), 허위 인플루언서 만들기, 게임 어뷰징(game abusing) 식별 등이 있다. 이것들을 이해하려면 머신러닝 분야에서 깊이있는 공부가 요구된다.
- 여기서 우리는 사회과학 통계학에서 기본적으로 잘 알려진 이상점 식별(outlier detection) 또는 영향력있는 사례 식별(influential case detection)의 두 가지 도구만을 공부한다:

  (1) 스튜던트화된 잔차(Studentized residuals)
  (2) Cook's distance

고전적 가정: 이상점 없음

이상점은 적합선과 크게 동떨어져 있으므로 잔차가 유독 크다.

- 앞의 그래프를 자세히 관찰해보면 모형에서 추정된 적합선과 이상점 사이의 거리 $(\hat{Y} - Y)$, 즉 잔차(residual)가 유달리 크다는 것을 알 수 있다.
- 그러므로 모든 관측치의 잔차를 계산해 본 뒤, 이를 표준화하여 유독 큰 잔차가 나타나는 관측치가 있다면 이것을 이상점으로 식별할 수 있다.
- 즉 표준화된 잔차(standardized residuals)로 이상점을 식별할 수 있다.

$$z_i = \frac{e_i}{\sqrt{MSE(1 - h_i)}}$$

- MSE는 평균제곱오차(mean squared error), h_i는 이른바 레버리지(leverage)로 불리우며 투영 행렬(projection matrix) 또는 모자 장수(hat maker)라고 불리우는 행렬의 대각행렬이다. 분모 부분을 더 깊이 이해하려면 많은 수학적 설명이 필요하므로 그냥 표준화가 목적인 것으로 일단 받아들이자.

고전적 가정: 이상점 없음

표준화된 잔차로 이상점을 식별하려는 접근법에는 치명적인 한계가 있다.

- 적합선 자체가 이미 이상점에 의해 왜곡된 뒤에 잔차가 계산되기 때문이다!
- 그러므로 개별 사례를 일단 하나씩 빼놓고 적합선을 추정한 뒤, 그로부터 해당 사례의 오차를 구하고 표준화하는 방식이 보다 바람직하다. 이것을 스튜던트화된 잔차 (Studentized residuals)라고 부른다.

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_i)}}$$

- 이때 스튜던트(Student)란 William S. Gosset의 가명이다.

고전적 가정: 이상점 없음

- 교과서와 통계분석 패키지에 따라 통일되지 않은 용어가 다양하게 쓰인다.
- 어떤 교과서/소프트웨어는 이를 스튜던트화 삭제된 잔차(Studentized deleted residuals) 또는 외재적으로 스튜던트화된 잔차(externally Studentized residuals)라고 부른다(Kutner, Nachtsheim, and Neter 2004).
- 다른 곳에서는 스튜던트화 잔차를 좀 다른 의미로 사용하기도 한다(그러므로 맥락을 잘 살펴보아야 한다).
- 위의 그림을 다시 예로 든다면, “해당 이상점 관측치를 포함하지 않고” 회귀분석을 수행하여 적합선을 그린 뒤, 그로부터 “해당 이상점 관측치에 대한 잔차”를 계산하는 셈이다.

고전적 가정: 이상점 없음

영향력있는 사례와 이상점은 좀 다른 개념이다.

- 이상점이 있더라도 실질적으로 거의 영향을 미치지 못할 수도 있다(Why?).
- 예를 들어 이상점이 고르게 분포하거나 적합선 상에 놓여있어 적합선의 기울기에 영향을 미치지 못하는 상황을 상상해보자!
- 그러므로 이상점 유무와 영향력 유무는 좀 별개의 문제로 접근해야 한다.
- 영향력있는 사례를 식별하는 방식으로 Cook's distance, DBFITS, DFBETAS 등이 특히 널리 알려져 있다.

고전적 가정: 이상점 없음

- Cook's distance는 특정 사례가 예측된 Y (fitted values) 전체에 미치는 영향력을 추정한다.

$$D_i = \frac{\sum(\hat{Y} - \hat{Y}_{(i)})^2}{k \cdot MSE}$$

- 분모는 오로지 표준화하는 역할을 수행한다. k는 추정된 회귀계수의 수를 지칭한다.
- 특정 사례를 데이터에서 일시적으로 제거한 뒤 추정하여 그것이 예측된 Y (fitted values)에 미치는 영향을 실제로 평가하는 셈이다.
- (DFBETA는 어떤 사례가 개별 회귀계수에 미치는 영향력을 평가하는 반면), D는 어떤 사례가 예측된 Y 전체에 미치는 영향력을 평가한다는 점에서 다르다.

고전적 가정: 이상점 없음

이상점/영향력있는 사례의 제거는 신중하게 단계적으로 수행한다.

- 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. 하지만 일단 히스토그램을 그려보고 너무 나간 사례들을 단계적으로 제거하면서 영향을 살펴보자.
- 이상점/영향력있는 사례를 제거한 결과만 보고하기보다는, 포함하기도 하고 제거하기도 하여 각각 따로 모형을 추정하고 결과를 함께 보고할 수도 있다. 특히 온라인 부록(online appendix)을 활용하자.

고전적 가정: 이상점 없음

- 이상점 식별과 영향력있는 사례 식별은 이론적으로 구분된다.
- 하지만 실증분석에서는 그냥 두 기준을 함께 적용하여 서로 상이한 두 측면을 보완하기도 한다.
- 이를 그림으로 나타낸 것이 이른바 LVR 도표(Leverage-Versus-Residual-squared plot)로 “두 값 모두가 크면” 삭제를 고려하게 된다.