사회통계
회귀모형의 적합도와 가정

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 모형의 적합도
2. 회귀분석의 가정
3. 상관분석과 회귀분석
4. 마지막 코멘트

모형의 적합도

모형의 적합도

사용한 회귀모형이 현실 자료에 얼마나 적합한지 살펴보아야 한다.

- 우리가 모형을 세워 그것을 현실 자료에 맞추어(fit) 본 이상, 이것이 얼마나 잘 맞는가를 말할 수 있어야 한다. 이것이 모형의 적합도(goodness-of-fit)이다.
- 교재에 따라서는 (위에서 다룬) 회귀계수 b_1와 상수항 b_0의 유의성 검정 (significance test) 역시 적합도 지표 중 하나로 취급하기도 한다.
- 여러가지 적합도 지표 가운데 결정계수(coefficient of determination)와 일원분산분석만 살펴보자.

모형의 적합도

먼저 결정계수를 살펴보자.

- 주어진 Y의 전체 변량(total variation)은 (모형에 의해) 설명된 변량(explained variation)과 (그렇지 못하고) 남은 변량(residual variation)의 합으로 분해될 수 있다.

$$\sigma^2_{total} = \sigma^2_{explained} + \sigma^2_{residual}$$

- 그렇다면 설명된 변량 $\sigma^2_{explained}$와 전체 변량 σ^2_{total}의 비율은 모형의 높은 설명력을 의미한다고 볼 수 있다.

$$R^2 = \frac{\sigma^2_{explained}}{\sigma^2_{total}} = 1 - \frac{\sigma^2_{residual}}{\sigma^2_{total}}$$

- 이것이 바로 결정계수 R^2의 직관적 의미이다. 설명된 변량과 전체 변량의 비율이므로 0과 1사이에 놓인다. 1에 가까울수록 모형은 높은 적합도를 보인다고 할 수 있다.
- 이제 남은 문제는 세 변량들이 어떤 식으로 정의되는가를 파악하는 것이다.

모형의 적합도

독립변수 X가 없을 때 종속변수 Y를 가장 잘 예측하는 요소는 무엇일까?

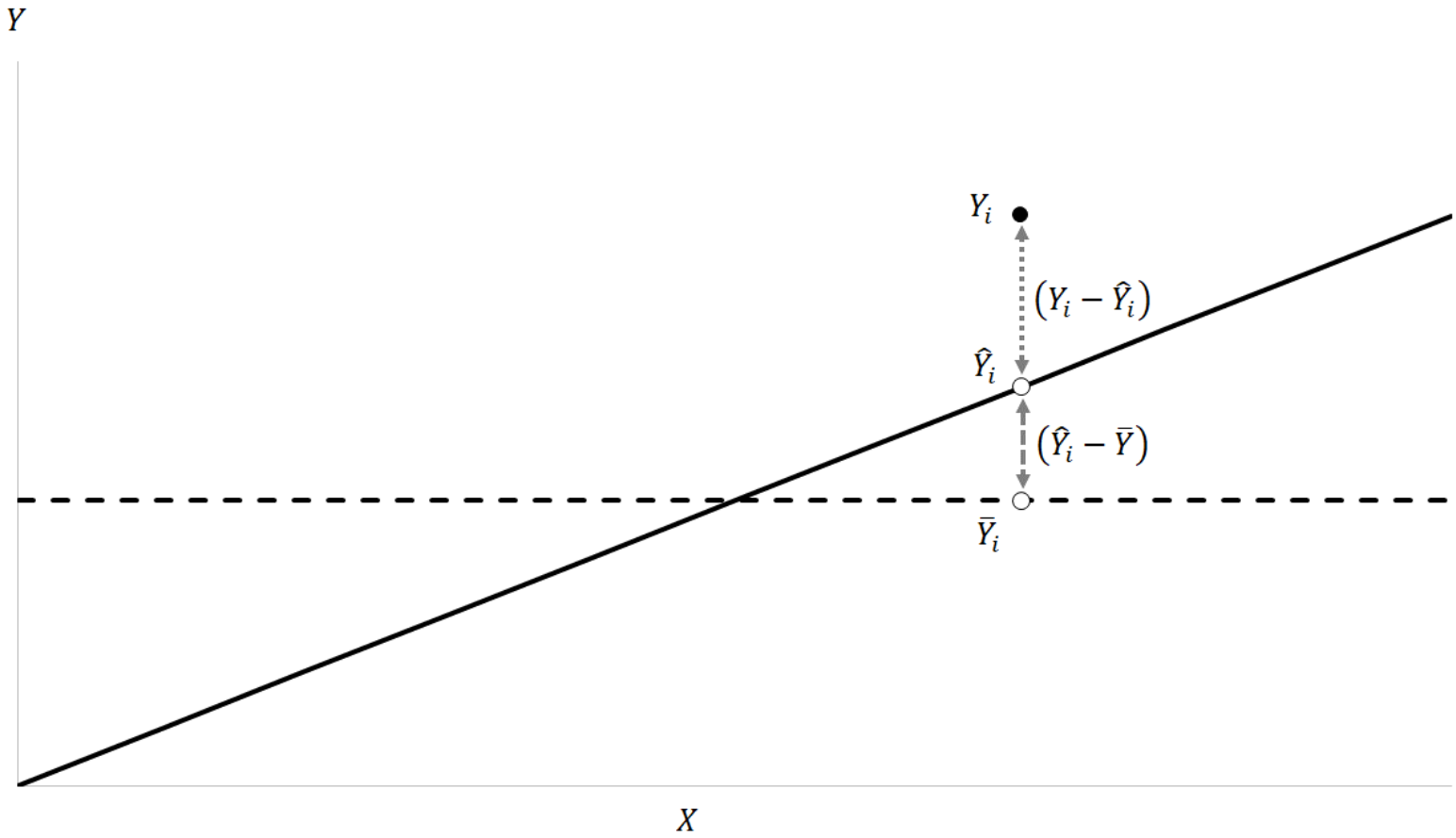
- 그것은 바로 종속변수의 평균 $\bar{Y}$이다(Why?). 우리는 사실 학기 초에 중심성향(central tendency)에 관해 공부하면서 이 사실을 은근슬쩍 배웠다.
- 이 원리는 우리에게 다음과 같은 중요한 원칙 하나를 제공한다:

  "어떤 모형을 세우던지 최소한 종속변수의 평균 $\bar{Y}$보다는 나은 설명력을 보여야 한다."
- 종속변수의 평균 $\bar{Y}$는 일종의 기준점이자 최소한의 바닥이다. 이것보다도 못한 모형은 아무런 가치도 없다.
- 반대로 종속변수의 실제 값 Y_i는 일종의 천장과도 같다. 이것은 모형이 추구할 수 있는 최고의 이상(ideal)이기 때문이다.
- (직관적으로 설명한다면) 당연히 모형의 예측값 $\hat{Y}_i$는 Y_i와 $\bar{Y}$ 사이 어딘가 놓이리라 생각할 수 있다(실제로는 꼭 그렇지 않을수도 있지만 너무 신경쓰지 말자).

모형의 적합도

- 그림을 통해서 우리는 다시 한 번 전체 변량이 잔여 변량과 설명된 변량의 합임을 직관적으로 이해할 수 있다.

$$(Y_i - \bar{Y}) = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y})$$



모형의 적합도

- 우리는 이제 제곱합(Sum of Squares; SS) 개념을 가지고 다음의 공식에 도달할 수 있다(증명 생략).

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

$$SS_{total} = SS_{residual} + SS_{explained}$$

- 이제 결정계수 R^2를 측정할 수 있다.

$$R^2 = \frac{SS_{explained}}{SS_{total}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$= 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

- 아까 그림을 통해 되새겨보면, Y_i에 더 가까운 위치에 $\hat{Y}_i$가 놓일수록 R^2 역시 높으리라고 추측할 수 있다(Why?).

모형의 적합도

이제 지난 수업의 예제 2의 회귀분석 결과에서 결정계수를 이해할 수 있다.

- "학자금 상환비율과 등록금 두 독립변수를 사용한 선형모형은 평균수입 변량 중 36.2%를 설명한다."
- 0.362라는 결정계수는 너무 낮을까? 그렇지만도 않다. 겨우 두 개의 변수만으로 이 정도 설명했다는 것은 나쁘지 않다. 많은 사회과학 연구를 돌이켜볼 때 이 정도면 제법 괜찮다.
- 결정계수는 보다 많은 독립변수를 집어넣을 때 계속 팽창하는 성향이 있다(Why?). 그런데 이것은 오캄의 면도날(Occam's razor)에 반하는 것이므로, 독립변수를 추가적으로 집어넣을 때마다 적절한 패널티를 가할 필요가 있다.
- 바로 아래 조정된 결정계수(adjusted R^2)는 바로 이렇게 패널티가 가해진 결정계수이다.

모형의 적합도

일원분산분석 결과표 역시 모형 전체의 적합도를 보여준다.

- 아래와 같이 선형모형을 설정하였을 때 최악의 추정 결과는 무엇일까?

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_{k-1} X_{k-1} + \beta_k X_k + \epsilon_i$$

- 그건 (상수 빼고) 모든 회귀계수가 0이 되는 상황이다. 보다 구체적으로 말해서 "모집단에서 (상수 빼고) 모든 β가 0인 상황"이다.
- 제대로 된 모형이라면 최소한 이런 경우만큼은 부정할 수 있어야 한다.
- 만일 이런 상황을 부정할 수 없다면 "이 모형은 완전히 쓸모없다" 라는 말을 부정할 수 없는 것이나 마찬가지이기 때문이다.

모형의 적합도

- 그러므로 다음과 같은 가설구조를 검증해 볼 수 있다.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_a : (\beta_1 \neq 0) \text{ or } (\beta_2 \neq 0) \text{ or } ... \text{ or } (\beta_k \neq 0)$$

- 이 가설구조를 테스트할 수 있는 방법은 일원분산분석(one-way ANOVA)이다.
- 기억이 나지 않으면 제12주차 강의안을 꼭 복습하자!

모형의 적합도

- 제11주차 때 배운 일원분산분석의 검정통계량 F 값은 다음과 같았다.

$$F_{(k-1,n-k)} = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)}$$

- 다중회귀분석의 맥락에서 검정통계량 F 값은 다음과 같다. 단 k는 독립변수의 수를 의미한다.

$$F_{(k,n-k-1)} = \frac{\sigma^2_{explained}}{\sigma^2_{residual}} = \frac{MS_{explained}}{MS_{residual}} = \frac{SS_{explained}/k}{SS_{residual}/(n-k-1)}$$

- F 값이 클수록 유의확률(p-value)이 작아지므로 위 귀무가설을 더 자신있게 기각할 수 있다.

모형의 적합도

- 분산분석표 안에 채워진 숫자는 다음과 같다.

Variance	SS	df	MS	F
Explained	$SS_{explained} = \sum(\hat{Y}_i - \bar{Y})^2$	k	$MS_{explained}$	F
Residual	$SS_{residual} = \sum(Y_i - \hat{Y}_i)^2$	$n-k-1$	$MS_{residual}$	
Total	$SS_{total} = \sum(Y_i - \bar{Y})^2$	$n-1$	MS_{total}	

- 계산에 집착하기 보다는 그 논리 전개에 주목하여 공부하자.

모형의 적합도

이제 예제 2의 회귀분석 결과에서 분산분석을 이해할 수 있다.

- "유의한 F"라고 괴상하게 이름 붙여진 항목을 보자. 이것은 일원분산분석의 유의확률(p-value)을 뜻한다.
- 이것이 0.05보다 작다는 사실은 95% 신뢰수준에서 다음의 귀무가설을 기각할 수 있다는 의미이다(Why?).

$$H_0 : \beta_1 = \beta_2 = 0$$

- 우리는 "모든 회귀계수가 0이다"라는 귀무가설을 기각하고 "최소한 하나의 회귀계수는 0이 아니다"라는 대립가설을 채택할 수 있었다. 다행이다~

모형의 적합도

예제 2. HPRICE2.CSV는 미국 어느 지역에서 주택가격의 중앙값과 기타 특징에 관한 자료이다. 주택가격의 중앙값(price)을 종속변수로 하고, 다른 모든 것을 독립변수로 하는 다중회귀모형을 설정하고 이를 수행한 뒤, 적절히 해석하시오. 이 회귀모형은 현실 주택가격을 얼마나 잘 설명하는지 평가하시오.

회귀분석의 가정

회귀분석의 가정

단순최소자승은 사실 몇 가지 가정에 입각하고 있다.

- 이 가정들은 (1) 단순최소자승을 통해 추정량(estimator)을 도출하고 (2) OLS 추정량이 왜 왜곡이 없고(unbiased) 다른 추정량보다 작은 표준오차만을 가지는지, 즉 효율적(efficient)인지 증명하는데 조용히 쓰인다.
- 재미있게도 교과서에 따라 가정의 목록이 조금씩 다르다. 보통 다음 중 어느 하나만을 설명한다.

  (1) 고전적 가정(classical assumption)
  (2) 종속변수 Y에 대한 가정
  (3) 오차항 ϵ에 대한 가정
- 이것은 상당히 복잡하고 대학원에서나 다루어지는 문제로 여겨진다.

회귀분석의 가정

- 그런데 어이없게도 사회조사분석사 2급에서 몇 차례나 기출문제가 등장했다. 수험서 레벨에서 오차항 ϵ에 관해 흔히 세 가지 가정을 한다.

  1. 정규성(normality)

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

  2. 등분산성(homoscedasticty)

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

  3. 독립성(independence)

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

- 약간 논리적으로 엉성하지만 적어도 틀린 것은 아니니 받아들여도 된다.

회귀분석의 가정

- 논리적으로 좀 더 엄격하게는 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

- 오차항 ϵ에 대한 가정 뿐 아니라, 좀 더 광범위하게 회귀모형에 대한 고전적 가정 (classical assumptions)도 필요하다.

  1. 선형성(linearity)

     "Y와 X의 관계는 선형적으로(linearly) 구성된다."

  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.

회귀분석의 가정

회귀분석의 가정은 고급통계학의 관문과도 같다.

- 회귀모형에서 가정이 깨지면 더이상 단순최소자승(OLS)이 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. 그렇기 때문에 학부 사회통계학에서 가정은 더이상 다루어지지 않는다.

상관분석과 회귀분석

상관분석과 회귀분석

근데 결정계수는 왜 하필 R^2 일까? 우리가 배운 r 은 상관계수였는데 혹시?

- 사실 상관계수 r 의 제곱은 단순회귀모형의 R^2 와 일치한다!
- 학부 사회통계학 레벨에서는 수학적으로 엄밀하게 이를 증명하는 대신 실습을 한 번 해보자.
- Earnings를 종속변수로 하고 Cost를 독립변수로 하였을 때, 단순회귀모형의 결정계수는 얼마인가? 또 CORREL$(\cdot)$ 함수로 구한 상관계수 r 의 제곱은 얼마인가?

상관분석과 회귀분석

회귀분석을 활용하여 비선형적 관계도 추정할 수 있다.

- 상관계수는 변수 간 선형적 관계의 강도(strength of the linear relationship)를 알려줄 뿐이다.
- 그러나 회귀분석은 매우 탄력적이라서 두 변수 사이에 비선형적 관계(nonlinear relationship)가 있더라도 어렵지 않게 살펴볼 수 있다. 구체적인 방법은 사실 아래처럼 매우 쉽지만 학부 과정에서는 다루지 않는다.

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

상관분석과 회귀분석

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)라고 불리우는 일반적인 사회과학 데이터를 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.

마지막 코멘트

마지막 코멘트

단순최소자승 회귀분석은 기초사회통계의 마지막 관문이다.

- 지금까지 걸어온 길을 반추해보자. 단순최소자승 회귀분석은 가설검정의 논리, t 검정, 일원분산분석(one-way ANOVA)에 이르기까지 모든 기법을 총동원하고 있다.
- 물론 우리는 중간시험 전까지 가설검정의 논리를 이해하기 위해서 확률변수, 확률분포, 표집분포, 표준오차, 누적분포함수, 표준정규분포, 임계값, 유의수준을 배웠다.
- 이제 여러분은 기초사회통계에 필요한 모든 지식을 학습하였다.
- 우리는 (SPSS 등) 전문적인 통계분석 패키지 대신 엑셀을 배웠다.

마지막 코멘트

그럼 다음 학기 전공필수 사회통계연습에서 우리는 무엇을 배울까?

- 기초통계분석의 원리를 학습하였으니, 우리는 그보다 수준높은 기법을 실제로 연습한다.
- 이때부터 우리는 전문적인 통계분석 패키지인 SPSS을 사용한다.
- 그때도 숙제는 매주 있다! 당연하지~
- 기초사회통계는 당연히 다음 수업에서 계속 쓰인다. 잊지 않기 위해서라도 방학 때 사회조사분석사 2급 준비를 해두자(어차피 졸업 요건이다).