사회통계연습

모형의 적합도와 결과표 작성

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 모형의 적합도
2. 결과표 작성

모형의 적합도

모형의 적합도

이제 자신의 회귀모형을 평가해보자.

- 우리가 모형(model)을 세워 그것을 현실 데이터에 맞추어(fit) 본 이상, 이것이 얼마나 잘 맞는가를 말할 수 있어야 한다. 이것이 모형의 자료 적합도(goodness-of-fit)이다.
- 크게 세 가지 적합도 지표(goodness-of-fit indices)를 공부한다:

  (1) 결정계수(R^2) 또는 조정된(adjusted) 결정계수(R^2)
  (2) 일원분산분석(one-way ANOVA)
  (3) 제곱근평균제곱오차(root mean square error; RMSE).
- 어느 한 가지 지표에만 맹목적으로 의존하지 않고 모든 지표들을 균형있게 살펴보면서 자신이 세운 모형이 얼마나 자료에 잘 맞는가를 확인해야 한다.
- 물론 적합도 지표에 근거해 여러 가지 모형들을 비교 평가할 수도 있다.

모형의 적합도

먼저 결정계수를 살펴보자.

- 주어진 자료에 전체 변량(total variation)이 있다면, 이것은 (모형에 의해) 설명된 변량(explained variation)과 (그렇지 못하고) 남은 변량(residual variation)의 합이라고 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.

모형의 적합도

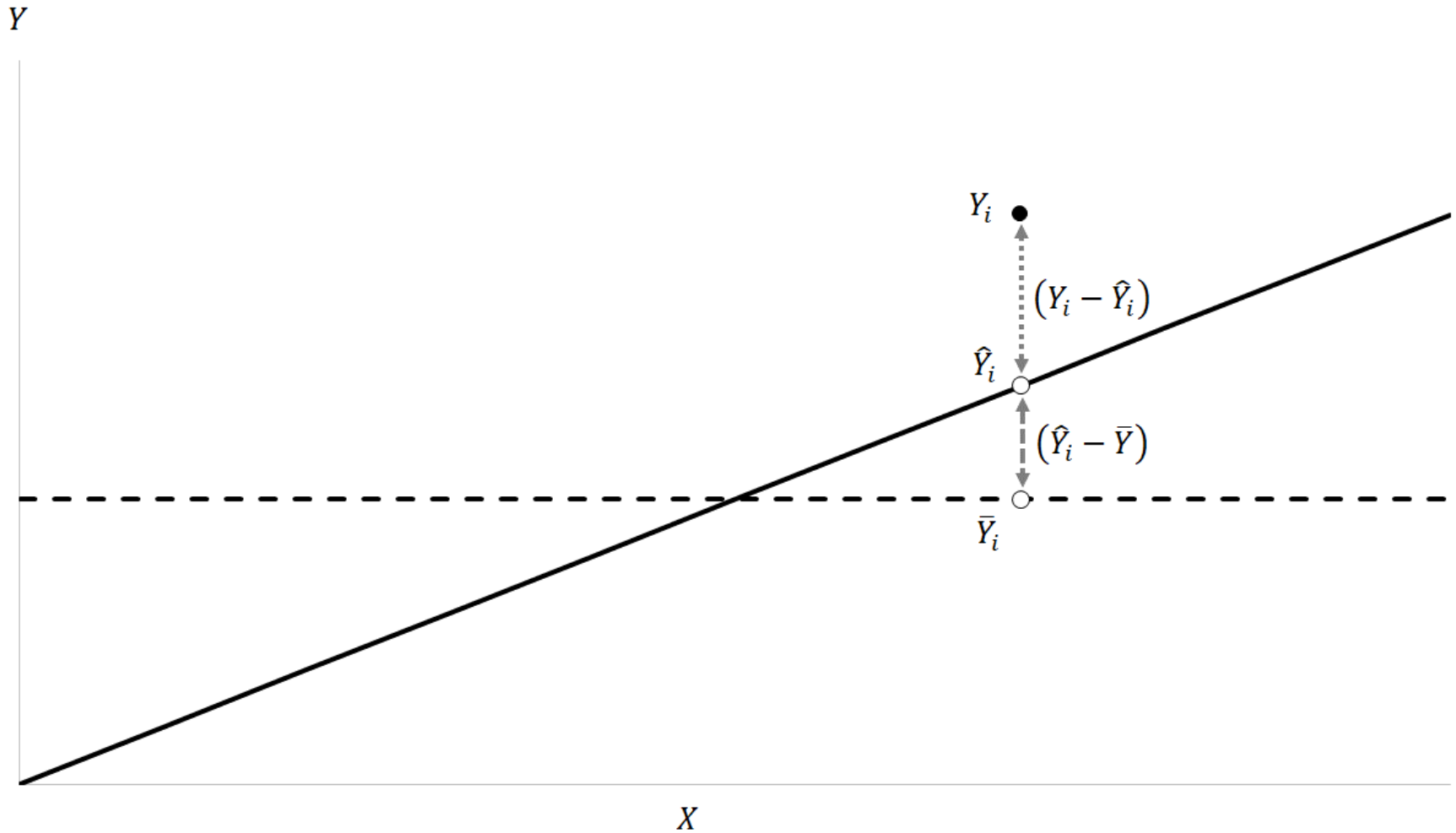
독립변수 X가 없을 때 종속변수 Y를 가장 잘 예측하는 요소는 무엇일까?

- 이 질문은 생각보다 매우 중요하다! 그것은 바로 종속변수의 평균($\bar{Y}$)이다(Why?).
- 이 원리는 우리에게 중요한 원칙 하나를 제공한다: "어떤 모형을 세우던지 최소한 종속변수의 평균($\bar{Y}$)보다는 나은 설명력을 보여야 한다."
- 종속변수의 평균 $\bar{Y}$은 일종의 기준점이자 최소한의 바닥이다. 이것보다도 못한 모형은 아무런 가치도 없다.
- 한편 종속변수의 실제 값(Y_i)은 일종의 이상점이자 최고의 천장이다. 이것은 모형이 추구할 수 있는 최고의 이상이다.
- (직관적으로 설명한다면) 당연히 모형의 예측값 $\hat{Y}_i$은 Y_i와 $\bar{Y}$ 사이 어딘가 놓이리라 생각할 수 있다(실제로는 꼭 그렇지 않을수도 있다).

모형의 적합도

- (아주 엄밀하지는 않지만) 그림을 통해서 우리는 총변량이 “설명된 변량”과 “잔여 변량”의 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.

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

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

모형의 적합도

연습 1. CAR_PRICES.SAV에서 중고차 가격(Price)을 차량의 연식(Age)으로 설명하시오. 이후 추가적으로 주행거리(Miles)를 독립변수로 추가하였을 때 모형 적합도의 개선 여부를 판단하시오.

모형의 적합도

모형 적합도 판정은 상당히 정형화되어 있다.

- "차량의 연식만을 고려한 회귀모형으로도 중고차 가격 변량의 83%를 설명한다."
- "차량 연식과 주행거리를 추가하였을 때, 이 회귀모형은 중고차 가격 변량의 84.6%를 설명한다."
- 변수를 하나 추가하였지만 통계적으로 유의하지도 않고 R^2는 겨우 1.6% 증가했을 뿐이다.
- 바로 아래 조정된 결정계수(adjusted R^2)는 무엇인가? 기본적으로 결정계수는 보다 많은 독립변수를 집어넣을 때 무제한적으로 팽창하는 성향이 있다. 그런데 이것은 오캄의 면도날 원칙(Occam's Razor)에 반하는 것이므로, 독립변수를 추가적으로 집어넣을 때마다 적절한 패널티를 가할 필요가 있다.
- 조정된 결정계수는 바로 이렇게 패널티가 가해진 결정계수이다.

모형의 적합도

분산분석 결과표 역시 모형 전체의 적합도를 보여준다.

- 아래와 같이 선형모형을 설정하였을 때 최악의 추정 결과는 무엇일까?

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$$

- 그건 (상수 빼고) 모든 회귀계수가 0이 되는 상황이다. 보다 구체적으로 말해서 "모집단에서 (상수 빼고) 모든 β가 0인 상황"이다.
- 제대로 된 모형이라면 최소한 이런 경우는 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.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_a : (\beta_1 \neq 0) \text{ or } (\beta_2 \neq 0) \text{ or } (\beta_3 \neq 0)$$

- 이 가설구조를 테스트할 수 있는 방법은 일원분산분석(one-way ANOVA)이다.

모형의 적합도

- 제6주차 때 연습한 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 기각할 수 있다.

모형의 적합도

- 분산분석표 안에 채워진 숫자는 다음과 같은 의미이다(교재 309).

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}	

- SPSS에서는 Explained 대신에 Regression으로 표기하고 있다.

모형의 적합도

이제 회귀분석 결과에서 ANOVA표를 이해할 수 있다!

- CAR_PRICES.SAV에서 F 값은 46.835이고 유의확률(Sig.)은 0.000이므로 0.05보다 당연히 작다. 그러므로 99.9% 신뢰수준에서 다음의 귀무가설을 기각할 수 있다.

$$H_0 : \beta_{\text{Age}} = \beta_{\text{Miles}} = 0$$

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

결과표 작성

결과표 작성

회귀분석을 수행하고나서 그 결과물을 표로 정리하여 보고해야 한다.

- 김수한·이명진(2014)을 참고로 사회과학연구에서 어떤 종류의 표와 그림이 사용되는지 살펴보자.
- 분석에 사용한 변수의 목록을 표로 정리하고 있음에 유념하자.
- 여기 사용된 모든 종류의 표를 직접 만들 수 있어야 한다. 특히 회귀분석의 결과표가 어떤 식으로 이루어졌는지 잘 흉내내어 만들어보아야 한다.

결과표 작성

- 아래는 훨씬 더 큰 다중회귀분석 결과표의 일부를 담고 있다.

〈표 4〉 한국인의 반기업정서: 대기업의 과거, 현재, 미래에 대한 평가와 지원에 대한 회귀분석

	모형 1 과거_성장	모형2 현재_성과	모형3 미래_기대	모형4 미래_지원	모형5 전체평가
사회제도 비신뢰	0.007**	0.012***	0.018***	0.025***	0.093***
	(0.003)	(0.002)	(0.003)	(0.004)	(0.008)
연령	0.007***	−0.006***	−0.005**	0.001	−0.014**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)
교육수준	0.072***	−0.016	0.051*	0.172***	0.275***
	(0.022)	(0.020)	(0.022)	(0.030)	(0.063)
가계소득 수준	0.041*	−0.050**	−0.040*	−0.043	−0.108
	(0.019)	(0.018)	(0.020)	(0.027)	(0.056)
성장지역_대도시	−0.056	−0.118**	−0.090*	−0.062	−0.339**
	(0.044)	(0.042)	(0.046)	(0.061)	(0.129)
상수	3.461***	2.111***	1.813***	1.635***	3.507***
	(0.141)	(0.133)	(0.145)	(0.194)	(0.411)
관찰수	1,296	1,296	1,296	1,296	1,296
R-squared	0.043	0.062	0.063	0.093	0.156
Adj. R-squared	0.033	0.052	0.053	0.083	0.147

주: *** p<0.001, ** p<0.01, * p<0.05; 괄호의 값은 표준편차

김수한·이명진. 2014. "한국사회의 반기업정서." 『한국사회학』 48(1): 39-70.