

계량분석

이변량 단순최소자승

김현우, PhD¹

¹충북대학교 사회학과 조교수

October 25, 2021

진행 순서

- 1 선형모형 입문
- 2 회귀계수와 표준오차
- 3 모형의 적합도
- 4 회귀분석의 연습

선형모형 입문

선형모형 입문

데이터에서 독립변수(X)와 종속변수(y) 사이의 관계를 선형적으로 나타낼 수 있다.

- 이른바 선형모형(linear model)은 아래와 같이 설정할 수 있다.

$$y_i = \beta_0 + \beta_1 X_i + u_i$$

- 하첨자로 i가 붙어있으므로 관측치(observations)에 따라 상이한 X_i 와 y_i , u_i 를 담게 된다.
- 이때 β_0 를 상수(constant) 또는 절편(intercept)이라고 부르고, β_1 를 기울기(slope)라고 부른다.
- 만일 β_1 이 크다면 X가 한 단위 증가할 때, y는 더 크게 증가할 것이다(Why?).
- $X = 0$ 일 때, $y = \beta_0$ 이다(Why?).

우리는 초등수학에서 일차방정식(linear equation)의 그래프 그리기를 배웠다.

- 기억이 나지 않는다면 desmos라는 웹사이트에서 연습하자.
- β_1 를 이리저리 바꾸어서 이것이 기울기임을 확인하고, β_0 를 이리저리 바꾸어서 이것이 절편임을 확인하자.
- X가 수평축이고 y가 수직축임에 주목하자.

선형모형 입문

데이터를 관통하는 하나의 선이 바로 회귀분석(regression analysis)의 핵심이다.

- 데이터에 가장 잘 맞는 직선(best-fitting straight line)을 그어 그 의미를 상상할 수 있다.
- 그런데 우리는 이미 상관분석을 배우면서 적합선(fitting line)을 이미 그려보았다.

Stata에서 폐암 데이터를 활용해 가장 잘 맞는 직선을 연습해보자([Stata 코드] 참고).

- eCampus에서 lungcancer.dta를 사용하자. 이것은 8개 북유럽 국가의 일인당 담배 소비량과 인구 100만 명당 폐암 발병자수를 나타낸 것이다.
- 상관계수를 구해보고 또 산포도를 그린 뒤, 적합선도 그려보자.
- 어떤 조건을 갖춘 적합선이 가장 잘 데이터를 나타낼 수 있을까?

선형모형 입문

우리는 어떤 선형모델이 완벽하게 데이터를 설명할 수 없다는 현실을 받아야하여야 한다.

- b_0 와 b_1 를 어떻게 설정하더라도 결국 데이터를 완벽하게 설명할 수가 없다.
- 따라서 우리는 추가적인 항(term)으로 오차항(u_i)을 고려해야 한다.

뿐만 아니라, 우리는 대체로 모집단 대신 표본을 분석한다.

- 그렇기 때문에 모집단에서의 회귀모형과 표본에서의 회귀모형은 구별될 수 있다.

$$y_i = b_0 + b_1 X_i + e_i$$

- 여기서 우리는 더이상 β 를 사용하지 않고 있고, u_i 대신 e_i 를 사용하고 있다.
- u_i 가 오차항이라고 불리웠던 반면, e_i 는 잔차항(residual term)이라고 불리운다.

선형모형 입문

본격적으로 회귀모형을 분석하기에 앞서 일단 흡연과 폐암 데이터를 통해 실습해보자.

- 이 데이터에서 사용할 수 있는 독립변수와 종속변수는 각각 무엇인가?
- Stata 명령어는 **regress**이다. 명령어 뒤에는 종속변수가 먼저 오며 주의할 것.

회귀분석의 결과표는 크게 (1) **분산분석(ANOVA)**, (2) **요약정보**, (3) **추정된 회귀모형**의 세 부분으로 나뉜다.

- 지금은 무엇도 잘 이해가 가질 않는다. 당연한 일이다. 하지만 곧 모든 것을 다 이해할 수 있다.
- 적어도 하나는 이해할 수 있다. 표본수(number of observations)가 8개라는 점이다.
- 일단 “추정된 회귀모형” 부분을 살펴보자. 회귀계수(regression coefficient), 표준오차(standard error), t 값, 유의확률(p-value), 95% 신뢰구간이 보고되어 있다. 이 값들은 변수(X)와 상수(cons)에 대해 각각 따로 보고되어 있다.

회귀계수와 표준오차

회귀계수와 표준오차

- 위와 같이 **닫힌 형태의 해(closed-form solutions)**를 분석적으로 구할 수 있는데, 만약 양적 방법론을 전공하려면 이 과정을 반드시 꼼꼼하게 이해해야 한다. 처음에는 스칼라(scalar)로, 다음에는 행렬(matrix)로 이 과정을 거듭 이해해야 한다.
- 이 과정에서 **미적분(calculus)**과 **선형대수학(linear algebra)**에 대해 어느 정도의 지식이 요구된다. 수학의 기초를 꼼꼼히 다져두지 않으면 앞으로 배울 수 있는 내용의 수준에서 (공부한 사람과) 압도적인 격차가 벌어진다.
- 양적 방법론을 전공하지 않는다면 너무 여기에 집착하지 않고 Stata를 그냥 믿고 바른 해석과 폭넓은 응용에 집중하자.

모형의 적합도

모형의 적합도

이제 모형의 적합도(goodness-of-fit)를 살펴보자.

- 우리가 모형(model)을 세워 그것을 현실 데이터에 맞추어(fit) 본 이상, 이것이 얼마나 잘 맞는가를 말할 수 있어야 한다. 이것이 모형의 현실 적합도이다.
- 우리는 특히 세 가지 적합도 지표(goodness-of-fit indices)를 공부한다: 결정계수(R^2), 조정된(adjusted) 결정계수(R^2), 그리고 RMSE.
- 어느 한 가지 지표에만 맹목적으로 의존하지 않고 모든 지표들을 균형있게 살펴보면서 자신이 세운 모델이 얼마나 데이터에 잘 맞는가를 확인해야 한다.
- 물론 적합도 지표에 근거해 여러 가지 모델들을 비교 평가할 수도 있다.

모형의 적합도

R^2 는 종종 결정계수(coefficient of determination)라고도 불린다.

- 결정계수는 제법 흥미로운 수리적 아이디어를 통해 계산된다.
- 우리에게 데이터가 주어지고 여기에 총변량(total variation)이 있다면, 이것은 모형에 의해 설명되는 변량(explained variation)과 그렇지 못하고 남은 변량(residual variation)의 합이라고 할 수 있다.

$$\text{Variation}_{\text{Total}} = \text{Variation}_{\text{Explained}} + \text{Variation}_{\text{Residual}}$$

모형의 적합도

- 그렇다면 설명되는 변량의 비율(ratio)은 모형의 높은 설명력 내지 데이터 적합도를 의미한다고 볼 수 있다.

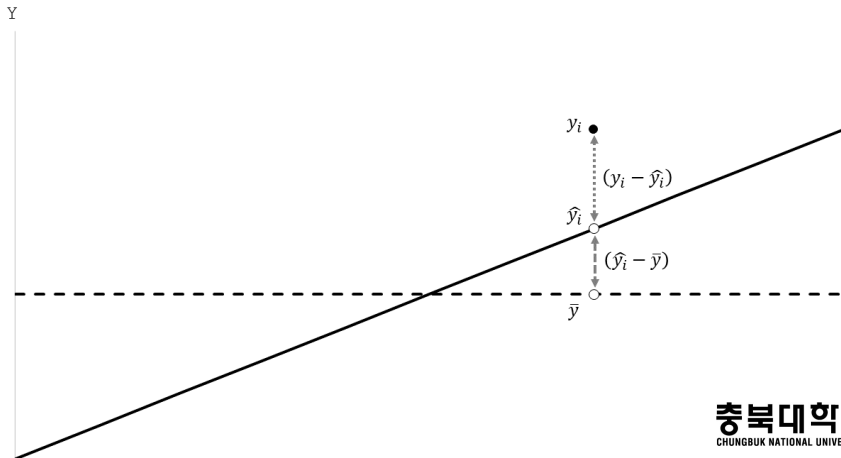
$$R^2 = \frac{\text{Variation}_{\text{Explained}}}{\text{Variation}_{\text{Total}}} = 1 - \frac{\text{Variation}_{\text{Residual}}}{\text{Variation}_{\text{Total}}}$$

- 이것이 바로 결정계수(R^2)의 직관적 의미이다. 변량의 비율이므로 0과 1 사이에 놓인다. 1에 가까울수록 모델은 높은 적합도를 보인다고 할 수 있고, 0에 가까울수록 모델은 형편없는 적합도를 보인다고 할 수 있다.
- 이제 남은 문제는 세 변량들이 어떤 식으로 정의되는가를 파악하는 것이다.

모형의 적합도

- (아주 엄밀하지는 않지만) 그림을 통해서 우리는 총변량이 “설명된 변량”과 “잔여 변량”의 합임을 직관적으로 이해할 수 있다.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$



모형의 적합도

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

$$SS_{\text{Total}} = SS_{\text{Explained}} + SS_{\text{Residual}}$$
$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

- 당연히 R^2 에 관해서도 정의할 수 있다.

$$R^2 = \frac{SS_{\text{Explained}}}{SS_{\text{Total}}} = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SS_{\text{Residual}}}{SS_{\text{Total}}} = 1 - \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- 아까 그림을 통해 되새겨보면, y_i 에 더 가까운 위치에 \hat{y}_i 가 놓일수록 R^2 역시 높으리라고 추측할 수 있다.

모형의 적합도

잠깐 다시 폐암 데이터의 회귀분석 결과를 되돌아보자.

- 이제 결정계수(R-squared)가 어떤 의미인지 이해할 수 있다.
- 0.2965라는 결정계수는 너무 낮을까? 그렇지만도 않다. 많은 사회과학 연구를 돌아해보면 이 정도면 괜찮다.
- 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 결정계수이고, 보다 구체적인 내용은 다음 주에 배우게 된다.
- 그 밑에 RMSE를 보자. 7.8778이라는 값은 모형을 통한 예측이 데이터와 평균적으로 7.8778만큼의 오차를 낼을 시사한다.

회귀분석의 연습

회귀분석의 연습

Stata에서 미세먼지 해결을 위한 국민여론조사(2019) 자료를 활용해 연습해보자([Stata 코드] 참고).

- 코드북을 참조해 “제1차 국민정책제안 각 과제”의 의견 총합과 “중장기 정책과제”의 의견 총합 사이의 관계를 보자.
- 많은 사회조사에서는 이른바 숫자형 척도가 좀처럼 조사되지 않고 대체로 범주형인 경우가 많다. 이 수업 말고 범주형 자료분석(categorical data analysis)을 배울 다른 기회가 있을 것이다. 일단 상관분석이나 회귀분석이나 숫자형 변수를 다루므로 우리는 유사한 범주형 변수를 모두 더하여 하나의 연속형 변수를 만들기로 한다.
- 상관계수를 구해보고 또 산포도를 그린 뒤, 피팅 라인도 그려보자.
- 회귀분석을 수행하고 회귀계수와 표준오차, t 값과 유의확률을 해석해보자.
- 결정계수와 RMSE를 해석해보자.

회귀분석의 연습

Stata에서 성별-연령별 유동인구 자료를 활용해 연습해보자([Stata 코드] 참고).

- 여자 50대를 종속변수로 남자 50대를 독립변수로 회귀분석을 수행한다.
- 그 이전에 산포도를 그려보고 극단치를 제거하자.
- 회귀분석을 수행하고 회귀계수와 표준오차, t 값과 유의확률을 해석해보자.
- 결정계수와 RMSE를 해석해보자.

회귀분석의 연습

오늘 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 인과관계로 거짓보고만 하지 않으면 된다.