사회통계
단순회귀분석

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 단순선형회귀모형
2. 단순선형회귀분석 연습
3. 회귀계수와 상수의 도출

단순선형회귀모형

단순선형회귀모형

독립변수 X와 종속변수 Y 사이의 관계를 선으로 나타내보자.

- 이른바 선형회귀모형(linear regression model)은 아래와 같이 일차방정식(linear equation)으로 설정할 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- X와 Y, ϵ에는 하첨자 i가 붙어있으므로 관측치(observation)에 따라 값이 다르다.
- 이때 β_0를 상수(constant) 또는 절편(intercept)이라고 부르고, β_1를 회귀계수(regression coefficient) 또는 기울기(slope)라고 부른다.
- 한편 β_0와 β_1를 어떻게 설정하더라도 결국 자료를 완벽하게 설명할 수는 없다(Why?). 따라서 오차항(error term) ϵ_i을 선형회귀모형에 추가한다.

단순선형회귀모형

- 일단 다음의 두 가정이 필요하다.

$$E(\epsilon_i|X_i) = 0$$
$$E(\beta_k X_i|X_i) = \beta_k E(X_i|X_i) = \beta_k X_i$$

- 그러면 선형회귀모형의 조건부 기대값(conditional expectation)인 회귀식(regression equation)을 정의할 수 있다.

$$\begin{aligned} E(Y_i|X_i) &= E(\beta_0 + \beta_1 X_i + \epsilon_i|X_i) \\ &= E(\beta_0|X_i) + E(\beta_1 X_i|X_i) + E(\epsilon_i|X_i) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

- X가 한 단위 증가할 때, Y는 β_1 만큼 증가한다(Why?).
- $X = 0$일 때, $Y = \beta_0$이다(Why?).

단순선형회귀모형

우리는 초등수학에서 일차방정식을 그래프로 나타내기를 배웠다.

- desmos라는 웹사이트에서 $Y = \beta_0 + \beta_1 X$ 꼴의 일차방정식을 입력해보자.
- 물론 β_0와 β_1 자리에 어떤 숫자를 입력해야 한다.
- X가 수평축이고 Y가 수직축임에 주목하자.
- β_1를 이리저리 바꾸어서 이것이 기울기임을 확인하고, β_0를 이리저리 바꾸어서 이것이 절편임을 확인하자.
- 기울기는 X가 한 단위 변화할 때 Y가 변화하는 정도를 의미한다.
- 절편은 $X = 0$일 때 Y 값을 의미한다.

단순선형회귀모형

- 우리는 이미 상관분석을 배우면서 적합선(fitting line)을 이미 그려보았다.
- 마찬가지로 주어진 자료의 X와 Y 사이에 가장 잘 맞는 직선(best-fitting straight line)을 그어 그 관계를 나타내 보일 수 있다.
- 이렇게 자료를 관통하는 하나의 선을 찾는 것이 바로 회귀분석(regression analysis)의 핵심이다.

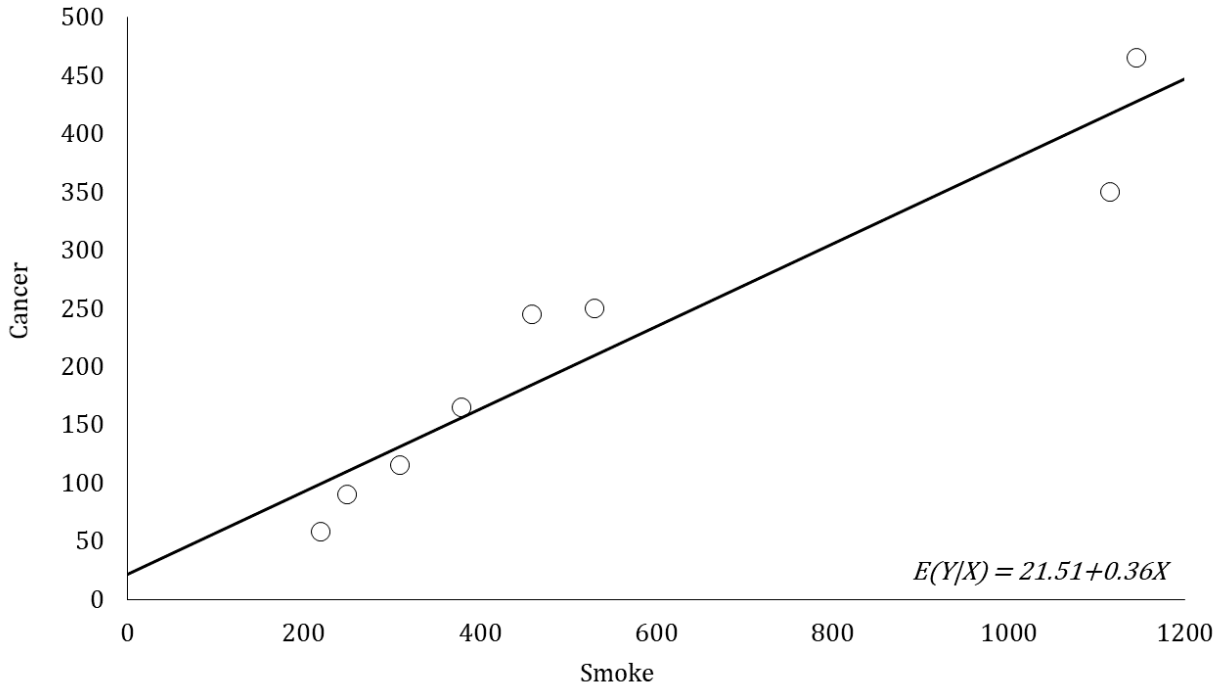
단순선형회귀분석 연습

단순선형회귀분석 연습

예제 1. lungcancer.csv는 8개 북유럽 국가의 1인당 담배 소비량(smoke)과 인구 100만 명당 폐암 발병자수(cancer)를 나타내고 있다. 독립변수와 종속변수를 선택한 뒤, 둘 사이의 관계를 나타내기 위해 가장 잘 맞는 직선을 그리시오. 회귀식을 도출하고 관계를 해석하시오.

단순선형회귀분석 연습

- 1인당 담배 소비량(smoke)이 100만 명당 폐암 발병자수(cancer)에 영향을 미친다고 보는 것이 타당하다.
- 먼저 산점도를 그리고 적합선을 추가하자. 이때, "수식을 차트에 표시"한다.
- 어떤 조건을 갖춘 적합선이 가장 잘 데이터를 나타낼 수 있을까? 만일 기울기와 절편이 달라지면 어떤 결과가 될까?



단순선형회귀분석 연습

- 회귀식은 다음과 같이 추정되었다.

$$E(Y_i|X_i) = 21.511 + 0.355 \cdot X_i$$

- "국가의 1인당 담배 소비량이 한 단위 증가할 때, 100만 명당 폐암 발병자수는 0.355명 만큼 증가한다."
- "아무도 흡연하지 않은 국가에서 100만 명당 폐암 발병자수는 21.511명이다."
- 일반적으로 표현하자면, 독립변수 X의 값이 한 단위 변화(unit change)하면 회귀계수 b_1 만큼 종속변수 Y에 영향을 미친다.
- 회귀계수 및 상수의 해석은 무척 단순하지만 연습을 필요로 한다!

단순선형회귀분석 연습

회귀식을 일단 추정했다면 이제 마음껏 예측에 사용할 수 있다!

- 추정된 상수 $\widehat{\beta}_0$와 회귀계수 $\widehat{\beta}_1$를 통해 예측된(predicted) Y, 즉 $\widehat{Y}$을 얻을 수 있다.
- 앞서 추정한 회귀식에 따르면 $\widehat{\beta}_0 = 21.511$, $\widehat{\beta}_1 = 0.355$이다.
- 이 값들이 적용된 회귀식 안 X_i에 원하는 값을 대입하면 Y_i를 예측(prediction) 할 수 있다.

$$\widehat{Y}_i = 21.511 + 0.355 \cdot X_i$$

- 추정량(estimates)에 대해서는 이렇게 ˆ(hat)을 붙인다.

단순선형회귀분석 연습

예제 2. 앞에 사용한 자료에 따르면, 1인당 담배 소비량이 1000인 어떤 가상의 국가에서 인구 100만 명당 폐암 발병자수는 몇 명인지 예측하시오.

단순선형회귀분석 연습

- 답은 376.511명이다.

$$\begin{aligned}\widehat{Y}_i =& 21.511 + 0.355 \cdot X_i \\ =& 21.511 + 0.355 \cdot 1000 \\ =& 376.511\end{aligned}$$

- 예측은 회귀분석의 대단히 유용한 기능(e.g., 돈벌이) 중 하나이다.
- 이것으로 주가(stock price)나 집값 등에 관한 모형을 세우고 회귀계수 및 상수를 추정한 뒤, 조건별로 가격을 예측해 볼 수 있다.

단순선형회귀분석 연습

예제 3. 중고차 딜러 새우는 자사에서 취급하고 있는 K모델 승용차 10대의 연식과 가격을 조사하였다(usedauto.csv). 연식(AGE)이 가격(PRICE)에 어떠한 영향을 미치는지 (1) 산점도를 그리고 (2) 회귀식을 추정하시오. 만일 연식이 8년 된 차량이 입고되었다면 얼마에 내놓는 것이 합리적인지 예측하시오.

회귀계수와 상수의 도출

회귀계수와 상수의 도출

오차를 전반적으로 최소화하는 적합선이야말로 가장 잘 맞는 직선이다.

- 엑셀에서 그려진 적합선과 회귀식은 도대체 어떻게 추정된 것일까?
- 일단 회귀식이 추정되었다면 새로운 X_i 값이 주어질 때, 그에 대응하는 Y_i를 예측(predict)할 수 있다.
- 실제 자료 Y_i와 예측된 Y_i (혹은 $\widehat{Y}$) 간의 차이는 곧 오차(error)라고 볼 수 있다.

$$\epsilon_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$$

회귀계수와 상수의 도출

- 예상되는 오차 ϵ_i를 줄이는 것은 현실 자료와 이론적 예측 사이의 괴리를 줄이는 것과 일맥상통한다(Why?).
- 물론 오차 하나(e.g., ϵ_1 또는 ϵ_4)만 줄이는 것이 아니라 전체적인 오차를 줄이는 것이 중요하다.
- 이때 오차의 합을 그냥 최소화하지 않고 오차 제곱의 합(sum of squared error; SSE)을 최소화한다(Why?).
- 오차 제곱의 합을 최소화하는 β_0와 β_1을 찾음으로써, 주어진 자료를 가장 잘 설명할 수 있는 모형을 만들 수 있게 된다.

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_i^n \epsilon_i^2$$

- 이것이 바로 보통최소제곱(ordinary least squares; OLS)이다.

회귀계수와 상수의 도출

lungcancer.csv를 가지고 보통최소제곱을 엑셀에서 차근차근 계산해보자.

- (1) 추정된 y 또는 $\widehat{Y}_i$
- (2) 오차 또는 $\epsilon_i = Y_i - \widehat{Y}_i$
- (3) 오차제곱 또는 $\epsilon_i^2 = (Y_i - \widehat{Y}_i)^2$
- (4) 오차제곱합 또는 $\sum \epsilon_i^2 = (Y_i - \widehat{Y}_i)^2$

회귀계수와 상수의 도출

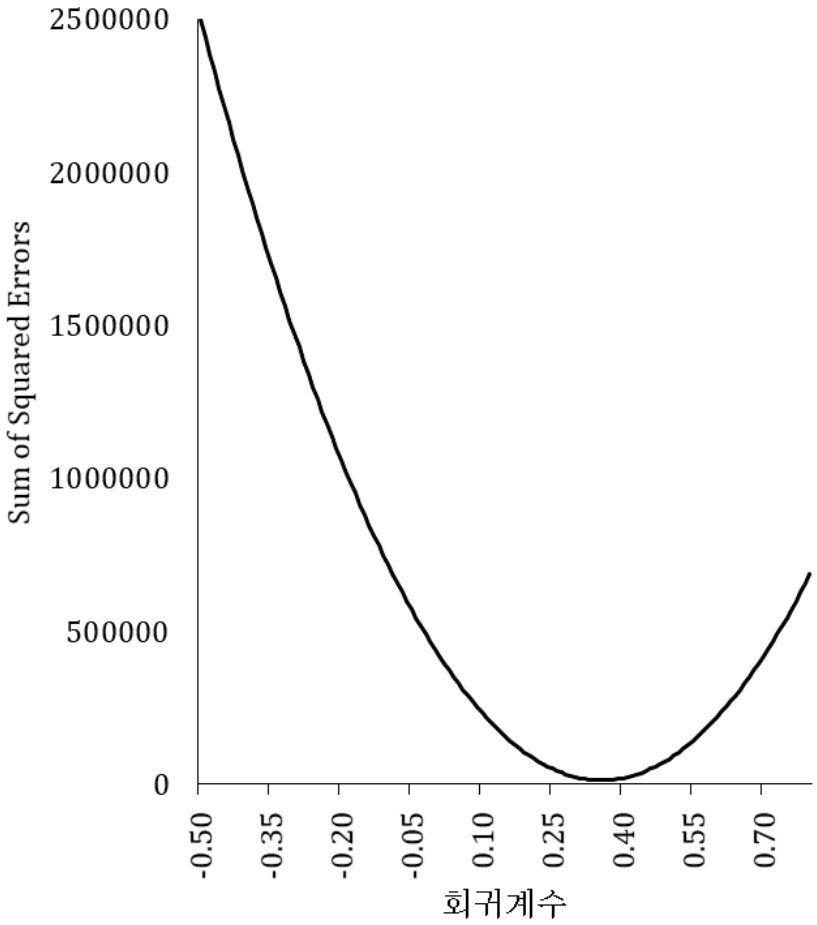
- 앞서 구한 회귀식을 사용하여 $\hat{Y}_i$을 계산하는 것이 출발점이다!

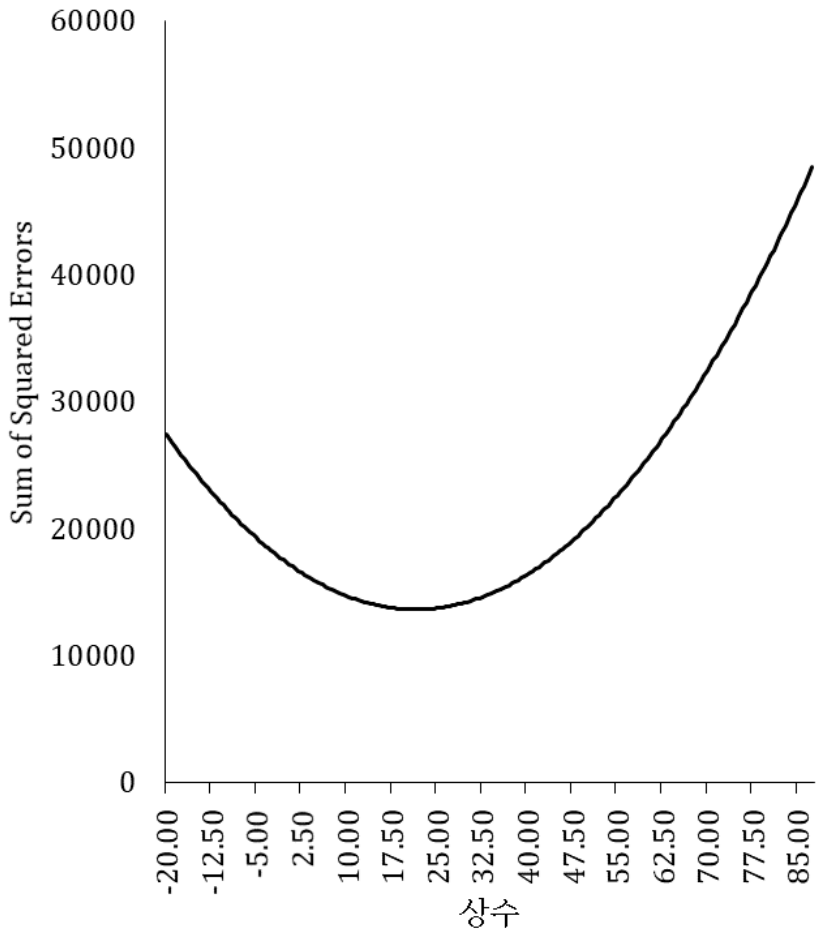
D2 =21.511 + 0.355*B2

	A	B	C	D	E	F	G
1	country	smoke	cancer	y_hat	e	e^2	
2	Iceland	220	58	99.611	-41.611	1731.475	
3	Norway	250	90	110.261	-20.261	410.5081	
4	Sweden	310	115	131.561	-16.561	274.2667	
5	Denmark	380	165	156.411	8.589	73.77092	
6	Holland	460	245	184.811	60.189	3622.716	
7	Switzerlan	530	250	209.661	40.339	1627.235	
8	Finland	1115	350	417.336	-67.336	4534.137	
9	Great Brita	1145	465	427.986	37.014	1370.036	
10						**13644.1**	**SSE**

회귀계수와 상수의 도출

- 어떤 상수 $\widehat{\beta}_0$와 회귀계수 $\widehat{\beta}_1$의 조합이 오차제곱합(SSE)을 가장 작게 만들 수 있을까?





회귀계수와 상수의 도출

수학적으로 어떻게 회귀계수와 상수를 도출되었을까?

- 그 구체적인 원리는 미적분학(calculus)에서 2차함수 최적화 문제(quadratic optimization problem)라고 불린다.
- 이 최적화 문제의 목적함수(objective function)는 오차제곱합 $\sum \epsilon_i^2$ 이므로 2차항이다.
- 미적분학에서 미분의 의미는 기하학에서 기울기(slope)를 말한다(Why?). 아까 그래프에서 기울기가 0이 되는 지점은 어디이고 무엇을 의미할까?
- 그러므로 오차제곱합을 β_0와 β_1에 대해 편미분(partial derivation)한 식을 0으로 놓고 풀면 (오차의 제곱을 최소화하는) β_0과 β_1 값을 알아낼 수 있다.

$$\frac{\partial \sum \epsilon_i^2}{\partial \beta_0} = 0$$

$$\frac{\partial \sum \epsilon_i^2}{\partial \beta_1} = 0$$

회귀계수와 상수의 도출

엑셀로도 물론 쉽게 회귀식을 추정할 수 있다.

- 엑셀에서는 [데이터]-[데이터 분석]을 통해 "회귀 분석"을 선택한다.
- 회귀분석의 결과표는 크게 (1) 회귀분석 통계량, (2) 분산분석, (3) 추정된 회귀모형의 세 부분으로 나뉜다.
- 지금은 무엇도 잘 이해가 가질 않는다. 당연한 일이다. 하지만 적어도 하나는 이해할 수 있다. 관측수(number of observations)가 8개라는 점이다.
- 일단 "추정된 회귀모형" 부분을 살펴보자. 회귀계수와 상수가 있다!

회귀계수와 상수의 도출

예제 4. 고래는 아르바이트에 몰두하다가 사회통계 수업을 그만 4번 정도 빼먹었다. 고래는 자신의 최종 학점을 예측하기 위해 과거 수강생 자료를 확보하였다(attend.csv). 독립변수를 skipped, 종속변수를 termgpa로 하는 회귀분석을 수행하고, 그 결과를 해석하시오. 고래의 예상 학점은 몇 점인지 예측하시오.

회귀계수와 상수의 도출

- 결석 시수(skipped)가 학점(termgpa)에 영향을 미친다는 아이디어에 따라 회귀모형을 세우고, "오차항을 최소화하는 β_0와 β_1"을 다음과 같이 추정할 수 있다.

$$Y_i = 3.043 - 0.076 \cdot X_i$$

- "결석 시수가 한 단위 증가할 때, 올해 학점은 0.076점만큼 감소한다."
- "결석 시수가 0일 때의 올해 학점은 3.043이다."
- 추정된 모형에서 $\widehat{\beta}_0 = 3.043$ 이고 $\widehat{\beta}_1 = -0.076$이므로 독립변수 X_i에 원하는 값을 대입하면 종속변수 예측값 $\hat{Y}_i$를 얻을 수 있다.

$$\widehat{Y}_i = 3.043 - 0.076 \cdot 4 = 2.739$$

- "수업을 4번 빼먹은 고래의 예상 학점은 2.739점이다."