계량분석
Bivariate Ordinary Least Squares

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 24, 2024

진행 순서

선형회귀모형 입문

선형회귀모형 입문

독립변수 X와 종속변수 Y 사이의 관계를 선으로 나타내보자.

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

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

- 하첨자 i가 붙어있으므로 관찰값(observation)에 따라 상이한 X_i와 Y_i, ϵ_i를 담게 된다.
- 이때 β_0를 상수(constant) 또는 절편(intercept)이라고 부르고, β_1를 회귀계수(regression coefficient) 또는 기울기(slope)라고 부른다.
- ϵ_i는 오차항(error term)이라고 부른다.

선형회귀모형 입문

- 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?).

선형회귀모형 입문

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.

선형회귀모형 입문

회귀분석은 결국 자료를 관통하는 최적합선을 찾으려는 시도이다.

- 가장 잘 맞는 직선(best-fitting straight line)을 그어 자료를 요약하는 것으로 상상해보자.
- 회귀분석(regression analysis)은 바로 그 직선에 관한 방정식을 찾으려는 시도이다.
- 점은 주어진 자료이고, 적합선은 이를 설명하는 모형인 셈이다.
- 한편 β_0와 β_1를 어떻게 설정하더라도 결국 자료를 완벽하게 설명할 수는 없다(Why?).
- 따라서 우리는 추가적인 항(term)으로 오차항(ϵ_i)을 고려해야 한다.

선형회귀모형 입문

오차제곱합을 최소화하는 적합선이야말로 가장 잘 맞는 직선이다.

- 일단 적합선을 그린 다음에는 X가 주어질 때, Y를 예측(predict)할 수 있다.
- 실제(actual) Y_i와 예측된(predicted) $\hat{y}$ 간의 차이는 곧 오차(error)라고 볼 수 있다.

$$\epsilon_i = Y - \widehat{Y}_i = Y_i - (\widehat{\beta_0} + \widehat{\beta_1} X)$$

- 이 오차를 줄인다는 것은 이론적 예측과 현실 데이터 사이의 괴리를 줄인다는 의미와 일맥상통한다(Why?).

선형회귀모형 입문

- 단, 오차의 합을 그냥 최소화하지 않고 오차 제곱의 합(sum of squares)을 최소화한다(Why?).
- 오차의 제곱의 합을 최소화할 수 있는 β_0와 β_1을 찾음으로서 주어진 데이터를 가장 잘 설명할 수 있는 모형을 개발할 수 있게 된다.

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

- 이것이 보통최소자승(ordinary least squares; OLS)이라고 불리우는 회귀모형의 계산 원리이다.

선형회귀모형 입문

- OLS의 해(solutions) 찾기는 수학적으로 2차함수 최적화 문제(quadratic optimization problem)이다.
- 이 최적화 문제의 목적함수(objective function)는 오차의 제곱합(sum of squared error)이므로 2차항이다.
- 이 오차의 제곱을 β_0와 β_1에 대해 편미분(partial derivation)하되, 그 식을 0으로 놓고 풀면 "오차의 제곱을 최소화하는" β_0과 β_1의 값을 알아낼 수 있다.

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

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

- 이 식들을 풀어 정규방정식(normal equations)을 도출해 낼 수 있다.

선형회귀모형 입문

- 위와 같이 닫힌 형태의 해(closed-form solutions)를 분석적으로 구할 수 있는데, 만약 양적 방법론을 전공하려면 이 과정을 반드시 꼼꼼하게 이해해야 한다.
- 가우스-마코프 정리(Gauss-Markov Theorem)은 이와 같이 구한 해가 최선의 선형, 편향없는 추청치(best linear unbiased estimator; BLUE)임을 증명하고 있다.
- 처음에는 스케일러(scalar)로, 다음에는 행렬(matrix)로 이 풀이를 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. 적절한 경영경제 통계학 교과서나 계량경제학 (econometrics) 교과서를 참고하자.

회귀모형의 해석

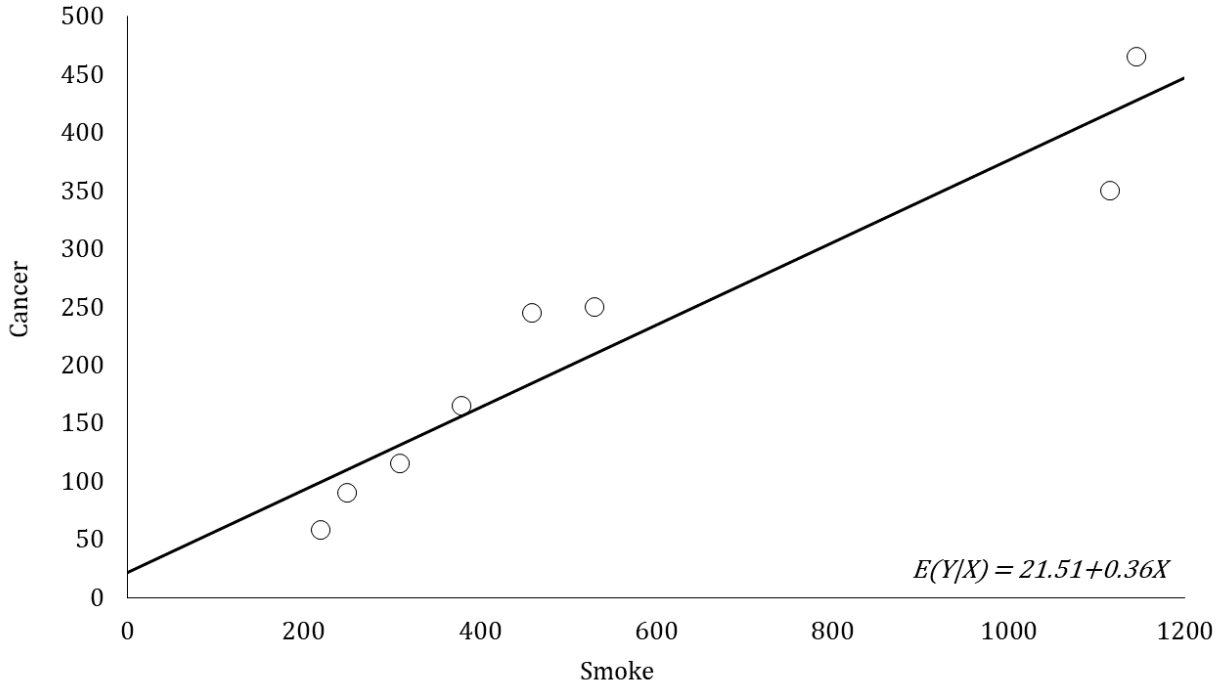
회귀모형의 해석

일단 Stata에서 실습부터 해보자.

- cancer.csv 자료를 사용하여 회귀분석을 수행한다면, 1인당 담배 소비량(smoke)이 100만 명당 폐암 발병자수(cancer)에 영향을 미친다고 보는 것이 타당하다.
- Stata에서 회귀분석 명령어는 regress이다. 명령어 뒤에는 종속변수가 먼저 옴에 주의할 것.
- 결과표는 크게 (1) 분산분석(ANOVA), (2) 요약정보, (3) 추정된 회귀모형의 세 부분으로 나뉜다.
- 이 결과표에 근거하여 추정된 회귀식을 작성해보자.

회귀모형의 해석

- 산점도를 그리고 적합선도 추가하자.
- 어떤 조건을 갖춘 적합선이 가장 잘 데이터를 나타낼 수 있을까? 만일 기울기와 절편이 달라지면 어떤 결과가 될까?



회귀모형의 해석

- 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!

회귀모형의 해석

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.
- predict 명령어를 사용해 $\hat{Y}$과 $\hat{\epsilon}$을 각각 추정해보자.

회귀모형의 해석

- 1인당 담배 소비량이 1000인 어떤 가상의 국가에서 인구 100만 명당 폐암 발병자수는 몇 명인지 예측해보자.
- 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.

회귀모형의 해석

해석상 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.

회귀식에서 유의성 검정

회귀식에서 유의성 검정

회귀분석에서도 표본을 넘어 모집단의 성격을 추리해야 한다.

- 설령 우리가 오차제곱항을 최소화하는 b_0과 b_1를 구했다고 하더라도 이것은 어디까지나 표본의 성격, 즉 통계량(statistic)일 뿐이다.
- 우리는 (다른 추리통계학에서와 마찬가지로) 다음과 같은 가설 구조에 따라 모집단의 성격, 즉 모수(parameter)에 대해서도 추리해야 한다.

$$\text{상수: } H_0 : \beta_0 = 0, \quad H_a : \beta_0 \neq 0$$
$$\text{회귀계수: } H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0$$

- 모집단에서 수많은 표본을 뽑아 그로부터 b_0와 b_1를 구한 뒤, 이것들의 표집분포를 그린다고 상상해보자.
- 그 가상적인 표집분포의 표준편차를 (상수 및 회귀계수의) 표준오차(standard error)라고 부를 수 있다.

회귀식에서 유의성 검정

- 보통 우리는 표본을 분석하므로 모회귀모형(population regression model)과 표본회귀모형(sample regression model)은 개념상 구별된다.

$$\text{모집단: } Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\text{표본: } Y_i = b_0 + b_1 X_i + e_i$$

- 여기서 우리는 β 대신 b를, ϵ_i 대신 e_i를 사용하고 있다.
- ϵ_i가 오차항이라고 불리웠던 반면, e_i는 잔차항(residual term)이라고 불리운다.

회귀식에서 유의성 검정

t 분포를 사용하여 회귀계수와 상수에 대한 유의성 검정을 수행한다.

- 주어진 표본에서 회귀분석으로 추정된 회귀계수 $\widehat{b}_1$ 의 t 값은 아래와 같다.

$$t = \frac{\widehat{b}_1 - \beta_1}{SE_{b_1}} = \frac{\widehat{b}_1}{SE_{b_1}}$$

- 이때 t 분포의 자유도는 $n-1$이다.
- 귀무가설이 옳다는 전제 아래 그린 표집분포는 t 분포한다. 표본에서 추정된 검정통계량 t 값의 위치를 확인해보고 그보다 극단적인 t 값을 얻게 될 확률, 즉 유의확률(p-value)을 계산할 수 있다.
- 만일 유의확률이 0.05보다 작다면 우리는 95% 신뢰수준에서 귀무가설($H_0 : \beta_1 = 0$)을 기각하고 대립가설($H_a : \beta_1 \neq 0$)을 채택할 수 있다.

회귀식에서 유의성 검정

Stata에서 유의성 검정 결과를 확인해보자.

- mroz 자료를 사용하여 여성의 소득 lwage을 아버지 학력수준 fatheduc으로 설명해보자.
- 유의성 검정을 수행해보고 의미를 해석해보자.
- 추정 결과(ereturn)는 행렬과 스케일러로 재사용할 수 있다.
- 검정통계량 t 값과 유의확률(p-value)를 직접 계산해보자.
- 유용한 기능 가운데 하나는 e(sample)인데 회귀모형에 사용된 표본을 알려준다.

회귀식에서 유의성 검정

통계적 유의성이 말해주는 의미를 명확히 이해하자.

- 통계적 유의성을 해석할 때 가장 흔한 실수 중 하나는 유의확률(p-value)이 작은 것을 가지고 선형적 관계의 강도(strength of linear relationship)로 해석하는 일이다.
- 유의확률은 단지 $H_0 : \beta = 0$라는 옳은 귀무가설을 기각하는 가능성을 보여줄 뿐이다.
- 실질적 유의성(substantial significance)은 통계적으로 유의한가 여부와 상관없이 실제로 얼마나 그 강도가 센가의 문제를 다룬다.
- 예컨대 한 시간 게임을 더하게 되면 독서시간이 2분 줄어든다는 발견(Cummings and Vandewater 2007)은 설령 통계적으로 유의하더라도 실질적으로는 그다지 유의하지 않다.
- 그러므로 (통계적으로 유의한 결과를 얻었더라도) 그 관계의 그래프를 그려보고 실제로 해석해보아 실질적 유의성이 얼마나 높은지 판단할 필요가 있다.

Cummings, Hope M. and Elizabeth A. Vandewater. 2007. "Relation of Adolescent Video Game Play to Time Spent in Other Activities." Archives of Pediatrics Adolescent Medicine 161(7): 684-689.

회귀분석의 결과표

회귀분석의 결과표

Stata에서는 esttab 또는 estout 등으로 통계표를 꾸밀 수 있다.

- 이것들은 사용자들이 만들어 배포한 명령어(user-written commands)이지만 굉장히 폭넓게 쓰인다.
- 두 명령어 모두 같은 사람이 만들었는데 접근 방식이 살짝 다르다. 상황에 따라 편리한 쪽을 사용하면 된다.
- 일단 한 번 만들어두면 앞으로 계속해서 쓸 수 있으므로 억지로 외우거나 하지 않아도 된다.
- outreg2 명령어도 있고 이쪽도 제법 유명하다. 다만 사용법이 esttab 또는 estout 과는 다르다.

회귀분석의 결과표

회귀분석의 결과표를 엑셀로 복사해 붙여놓고 가지런하게 꾸며야 한다.

- 회귀분석 결과표에는 Stata에서 사용하던 변수명을 그대로 남겨두지 말고 (독자가 알아볼 수 있도록) 똑바로 고쳐써야 한다.
- 결과표 안에 회귀계수 뿐 아니라, 표준오차(standard error), t 값, 또는 신뢰구간(confidence interval) 셋 중 하나는 함께 보고해야 한다(Why?).
- 통계적으로 유의하지 않았다고 포함했던 변수를 멋대로 표에서 빼서는 안된다(물론 상수도 마찬가지이다).
- 표가 너무 길어지면 (연구에서 핵심이 아닌) 변수들은 생략하는 관행도 있다. 그러나 이 경우에도 논문에서 반드시 보고해야 한다.

회귀분석의 결과표

유의확률에 관한 정보를 요약하기 위해 *을 붙인다.

- 유의확률이 0.001보다 작으면 상관계수 옆에 별 3개(***), 0.01보다 작으면 별 2개(**), 0.05보다 작으면 별 1개(*), 0.1보다 작으면 대거(dagger) 하나(†)를 붙일 수 있다.
- 이런 표식은 통계적으로 유의하게(statistically significantly) 귀무가설을 기각할 수 있음을 의미한다.
- 유의확률에 따른 별 붙이기는 관습의 문제이고 연구자마다 다르다(어떤 이들은 아예 붙이지 않는다).

회귀분석의 결과표

〈표 4〉 고3 학업성취도에 대한 회귀분석의 결과 1

	모형 1		모형 2		모형 3	
	계수	표준오차	계수	표준오차	계수	표준오차
비일반고[1]	.062**	.179	.017	.155	.019	.154
중3 학업성취도			.509***	.008		
남성[2]					-.009	.107
자아존중감					.165***	.014
학습시간					.091***	.008
학습태도					.355***	.013
교사-학생 관계					.046*	.019
가구소득					.056**	.114
부모 관리/전문직[3]					.034	.116
부모 대졸 이상[4]					.055**	.123
부모 감독					-.005	.035
부모 애정					.009	.029
4년제 대학교 졸업 부모 기대[5]					.082**	.194
대학원 졸업 부모 기대[5]					.099***	.253
아직 정해지지 않은 부모 기대[5]					.028	.243
상수	8.952***	.065	6.498***	.126	-.819	.985
사례수	1,361		1,361		1,361	
결정계수	.004		.261		.285	

***$p < .01$, **$p < .05$, *$p < .1$.
계수는 표준화 회귀계수. 단, 상수의 계수는 비표준화 회귀계수.
기준집단: 1 = 일반고; 2 = 여성; 3 = 부모 비전문/관리직; 4 = 부모 고졸 이하; 5 = 2-3년제 전문대학 졸업 이하 부모 기대.

연습문제

연습문제

연습 1. 2021년 청소년건강행태조사 자료를 활용하여 다음에 모두 답하시오.

(1) 최근 7일 동안 아침식사 빈도, 과일 섭취빈도, 산음료 섭취빈도, 단맛 나는 음료 섭취빈도, 패스트푸드 섭취빈도, 물 섭취빈도 간의 연관성을 살펴보기 위해 상관분석을 수행하시오.

(2) 주중 스마트폰 사용 평균 시간이 주중 수면시간과 어떠한 연관성을 갖고 있는지 가설을 세우고 회귀분석을 수행하시오. 시각화와 함께 적절히 해석하시오.

(2) 스마트폰 사용 평균 시간이 범불안장애 지표(GAD-7)에 어떠한 연향을 미치는지 가설을 세우고 회귀분석을 수행하시오. 시각화와 함께 적절히 해석하시오.