사회통계연습
회귀계수의 해석과 유의성 검정

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 단순선형회귀모형
2. 회귀분석에서 유의성 검정
3. 다중회귀모형

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.

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

- 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.

단순선형회귀모형

우리는 초등수학에서 일차방정식을 그래프로 나타내기를 배웠다.

- 그 내용이 기억나지 않는다면 desmos라는 웹사이트에서 $Y = \beta_0 + \beta_1 X$를 그려보자.
- β_1를 이리저리 바꾸어서 이것이 기울기임을 확인하고, β_0를 이리저리 바꾸어서 이것이 절편임을 확인하자.
- X가 수평축이고 Y가 수직축임에 주목하자.
- 기울기는 X가 한 단위 변화할 때 Y가 변화하는 정도를 의미한다.
- 절편은 $X = 0$일 때 Y 값을 의미한다.

단순선형회귀모형

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.

단순선형회귀모형

오차를 전반적으로 최소화하는 적합선이야말로 가장 잘 맞는 직선이다.

- 적합선과 회귀식은 도대체 어떻게 추정하는 것일까?
- 일단 회귀식이 추정되었다면 새로운 X 값이 주어질 때, (그에 대응하는) Y를 예측(predict)할 수 있다.
- 실제 자료 Y와 예측된(predicted) Y (혹은 $\widehat{Y}$) 간의 차이는 곧 오차(error)라고 볼 수 있다.

$$\widehat{e} = Y - \widehat{Y} = Y - (\widehat{\beta}_0 + \widehat{\beta}_1 X)$$

- 모집단에서는 오차, 표본에서는 잔차(residuals) e라고 구분된다(하지만 종종 잔차도 그냥 오차라고 부른다).
- 회귀식에서 추정량(estimates)에 대해서는 이렇게 ˆ(hat)을 붙인다.

단순선형회귀모형

- 예상되는 오차 $\widehat{\epsilon_i}$를 줄인다는 것은 현실 자료와 이론적 예측 사이의 괴리를 줄인다는 의미와 일맥상통한다.
- 물론 (오차 하나만 줄이는 것이 아니라) 오차를 전체적으로 줄이는 것이 중요하다.
- 단, 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.

단순선형회귀모형

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

단순선형회귀모형

- 1인당 담배 소비량(smoke)이 100만 명당 폐암 발병자수(cancer)에 영향을 미친다고 보는 것이 타당하다.
- 회귀식은 다음과 같이 추정되었다.

$$E(Y|X) = 0.355 \cdot X + 21.511$$

- "국가의 1인당 담배 소비량이 한 단위 증가할 때, 100만 명당 폐암 발병자수는 0.355 만큼 증가한다."
- "아무도 흡연하지 않은 국가에서 100만 명당 폐암 발병자수는 21.511이다."
- 일반적으로 표현하자면, 독립변수 X의 값이 한 단위 변화(unit change)하면 회귀계수 b_1 만큼 종속변수 Y에 영향을 미친다.

단순선형회귀모형

- Jamovi에서는 [분석]-[회귀]-[선형 회귀분석]을 선택한다. 왜 선형일까?
- SPSS에서는 [분석]-[회귀분석]-[선형]을 선택한다.
- 회귀분석의 결과표는 크게 (1) 회귀분석 통계량, (2) 분산분석, (3) 추정된 회귀식의 세 부분으로 나뉜다.
- 사실 산점도와 적합선을 먼저 그려보아 두 변수 간 관계를 미리 짐작해보고 극단치 (outliers) 존재 유무 등을 미리 살펴보아야 한다.
- 이때 적합선은 사실 회귀식을 시각화한 것이다.

회귀분석에서 유의성 검정

회귀분석에서 유의성 검정

회귀분석에서도 표본을 넘어 모집단의 성격을 추리해야 한다.

- 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.

회귀분석에서 유의성 검정

- 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.

회귀분석에서 유의성 검정

- 우리는 모집단에서 수많은 표본을 뽑아 그로부터 상수 b_0와 회귀계수 b_1를 추정한 뒤, 이것들의 표집분포를 구축해 볼 수 있다.
- 우리는 (1) 모회귀계수 또는 모상수 β에 대해 정규성(normality)을 가정하거나 (2) 중심극한정리(central limit theorem)에 힘입어 아래 다음을 알 수 있다.

$$\beta_0 = E(\widehat{b_0})$$
$$\beta_1 = E(\widehat{b_1})$$

- 이제 (상수항와 회귀계수의) 표집분포의 표준편차를 표준오차(standard error)라고 부를 수 있다.

회귀분석에서 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 (상수와 회귀계수를 포함하여) $n-2$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 5% 유의수준 또는 95% 신뢰수준에서 귀무가설($H_0 : \beta_1 = 0$)을 기각하고 대립가설($H_a : \beta_1 \neq 0$)을 채택할 수 있다.

회귀분석에서 유의성 검정

연습 2. CONSUMPTION.CSV는 2000년-2016년 사이 분기별 미국인의 경제자료를 담고 있다. 이 자료에서 가처분소득(income)이 한 단위 증가할 때, 연 평균 소비(consumption)가 얼마만큼 변화하는지를 나타내는 한계소비성향(marginal propensity to consume)을 추정하시오(두 변수 모두 단위는 달러). 또한 가처분소득이 35,000달러일 때, 연 평균 소비지출액이 얼마인지 예측하시오.

다중회귀모형

다중회귀모형

실제 업무나 연구에서는 다중회귀모형을 사용한다.

- 단순회귀모형(simple regression model)에서는 오로지 독립변수 X와 종속변수 Y, 딱 두 개의 변수만 고려하였다.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- 반면, (종속변수는 여전히 1개지만) 모형 안에 k개의 독립변수가 투입된 경우를 다중회귀모형(multiple regression model)이라고 부른다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

- k개의 독립변수를 모형에 투입했다면 여러 영향력은 각각에 해당되는 변수 안으로 나뉘어 흡수된다.
- 그러므로 다중회귀분석은 (다른 변수의 영향력으로부터 독립된) 특정 변수의 순효과(net effect) 또는 부분효과(partial effect)를 살펴보는데 유리하다.

다중회귀모형

- 가령 임금(ln_wage)을 종속변수로, 나이(age)와 교육연수(educ)를 독립변수로 하는 회귀모형을 세웠다고 하자.

$$ln\_wage = b_0 + b_1 age + b_2 educ + e$$

- 다중회귀분석을 통해 "나이(age)의 효과를 통제했을 때(즉 같은 나이일 때)" 교육연수(educ)가 한 단위 변화하면 임금(ln_wage)이 얼마만큼 변화하는지 살펴볼 수 있다.

다중회귀모형

연습 3. MCAS.SAV에서 (메사추세츠 주의 224개 학군별로 수집된) 학생 대 교사 비율(str), 교사의 평균 연봉(tsal), 가계소득 중앙값(inc), 한부모가계 비율(sgl)이 학업성취도 평균점수(score)와 어떤 연관성을 맺고 있는지 회귀분석을 수행하고 회귀계수 및 유의성 검정 결과를 해석하시오.

다중회귀모형

다중회귀분석에서는 변수 체크에 주의를 기울여야 한다.

- 여러 개의 독립변수를 모형에 한꺼번에 투입하다보면 하나하나를 꼼꼼하게 살펴보지 않고 그냥 대충 집어넣는 경우가 많다. 이것은 매우 위험하다!
- 개별 변수의 척도가 어떻게 구성되어 있는지, 분포는 어떠한지, 결측치가 있는지 등을 반드시 꼼꼼하게 살펴보아야 한다.
- 이러한 자료 전처리(data pre-processing) 과정은 대체로 지루하지만 꼭 필요하므로 산업인력 수요가 크다.

다중회귀모형

다중회귀모형의 해석에 친숙해져야 한다.

- 다른 변수들의 영향력을 통제하였을 때, 특정 독립변수 X의 값이 한 단위 변화(a unit change)하면 회귀계수 b만큼 종속변수 Y가 변화한다.
- 가령 str, tsal, inc, sgl이 score에 영향을 미친다는 회귀모형을 세우고 보통최소자승법(OLS)에 따라 상수와 회귀계수를 다음과 같이 추정하였다고 하자.

$$\text{score} = -.496\text{str} - .023\text{tsal} + .293\text{inc} - .878\text{sgl} + 231.894$$

- "교사의 평균 연봉(tsal), 가계소득 중앙값(inc), 한부모가계 비율(sgl)의 효과를 통제하였을 때, 학생 대 교사 비율(str)이 10% 증가할 때마다 학업성취도 평균점수는 4.96점 씩 감소한다."
- 통계적으로 유의하지 않은 회귀계수는 해석하지 않는다(Why?).

다중회귀모형

다중회귀분석에서는 보통 변수들의 역할을 구분한다.

- 관심변수(variables of interest)와 통제변수(control variables)를 구분하는 것이 대표적이다.
- 사회학 연구의 차원에서 볼 때, 통계분석은 결국 근거(증거)를 마련하는 과정이다. 우리는 핵심 주장을 가지고 있기 마련이다!
- 그 핵심 주장을 검증하기 위해 꼭 필요한 변수들이 바로 관심변수 역할을 수행하고, 나머지는 단지 통제변수에 지나지 않는다.
- 연구문제에 따라 "교육수준이 임금에 미치는 영향"에 관심이 있다면 교육수준이 관심변수가 되고, 나머지 (잠재적으로 임금에 영향을 미칠 수 있는) 다른 모든 독립변수들은 이른바 통제변수로 취급된다.
- 하지만 이것은 어디까지나 주관적으로 부여하는 의미일 뿐이고, 회귀모형 안에서 변수들의 지위는 평등하다(Why?).

다중회귀모형

연습 4. MCAS.SAV에서 교사의 평균 연봉(tsal)이 학업성취도 평균점수(score)과 어떤 연관성을 갖고 있는지 살펴보고자 한다. 이에 관한 상관분석을 수행하고 산점도를 그리시오. 단순회귀분석을 수행하고 결과를 함께 해석하시오. 그 다음, 가계소득 중앙값(inc)을 통제변수로 하는 다중회귀분석을 수행하고 그 결과를 해석하시오. 왜 두 분석의 결과가 다른지 이유에 관해 논하시오.