사회통계
다중회귀분석

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 회귀계수와 상수의 유의성 검정
2. 다중회귀모형

회귀계수와 상수의 유의성 검정

회귀계수와 상수의 유의성 검정

회귀분석에서도 표본을 넘어 모집단의 성격을 추론할 필요가 있다.

- 설령 우리가 미분 문제를 풀어 오차제곱합(SSE)을 최소화하는 회귀계수와 상수를 구했다고 하더라도 이것은 어디까지나 표본의 성격, 즉 통계량(statistic)일 뿐이다.
- 그렇기 때문에 모집단에서의 회귀모형과 표본에서의 회귀모형은 개념적으로 구별될 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad \text{(모집단 회귀모형)}$$
$$Y_i = b_0 + b_1 X_i + e_i \qquad \text{(표본 회귀모형)}$$

- ϵ_i가 오차항(error term)이라고 불리우는 반면, e_i는 잔차항(residual term)이라고 불리운다.

회귀계수와 상수의 유의성 검정

- 우리는 다음과 같은 가설 구조에 따라 모집단의 성격, 즉 모수(parameter)에 대해서도 추론해야 한다.

$$H_0 : \beta_0 = 0 \qquad H_a : \beta_0 \neq 0 \qquad \text{(상수의 경우)}$$
$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0 \qquad \text{(회귀계수의 경우)}$$

- 회귀계수가 0인가 여부를 테스트함에 주의하자.
- 추론하는 대상은 모집단의 상수와 회귀계수임에 주의하자.

회귀계수와 상수의 유의성 검정

- 우리는 모집단에서 수많은 표본을 뽑아 그로부터 상수 b_0와 회귀계수 b_1를 추정한 뒤, 이것들의 표집분포를 구축해 볼 수 있다.
- 우리는 (1) 표본회귀모형 $y = b_0 + b_1 X + e$에서 정규성(normality)을 가정하거나 (2) 중심극한정리(central limit theorem)에 힘입어 아래 다음을 알 수 있다.

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

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

회귀계수와 상수의 유의성 검정

- 이제 (상수항과 회귀계수 각각의) 표집분포의 표준편차를 표준오차(standard error) σ_{b_0} 또는 σ_{b_1} 라고 부를 수 있다.

$$\sigma_{\hat{b}} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2/(n-2)}{\sum(X_i - \bar{X})^2}} = \frac{\sigma_e}{\sqrt{\sum(X_i - \bar{X})^2}}$$

- 분자 σ_e는 오차의 표준편차이다. 이것은 오차평균제곱(mean squared error; MSE), 즉 오차제곱합(SSE)을 자유도로 나눈 값의 제곱근이다. 이 모형에서 회귀계수와 상수 두 개가 사용되었으므로 자유도는 $n-2$이다.
- 분모는 독립변수의 편차제곱이다.
- 일원분산분석(ANOVA)의 용어를 빌리면, 분자는 "모형의 MS_{error}"이고, 분모는 "독립변수의 SS_{total}"이라고 불릴 수 있다(Why?).

회귀계수와 상수의 유의성 검정

- 식 자체보다는 논리를 이해하자!
- 이것은 결국 상수항과 회귀계수의 표집분포의 표준편차이므로, 모집단에서 무한히 많은 표본을 뽑아 회귀계수를 구했는데 그것들이 어느 정도나 제각각인가를 나타낸다.
- 그러므로 b_1과 b_0의 표준오차 $\sigma_{\hat{b}}$는 각각의 추정값(estimates)에 대해 얼마나 확신하는가(confident)를 보여준다(Why?).
- (1) 모형의 오차가 클수록, (2) 독립변수 X_i의 분산이 작으면 회귀계수의 표준오차 σ_b가 커지므로, b는 믿을 수 없게(unreliable) 된다(Why?).

회귀계수와 상수의 유의성 검정

t 분포를 사용하여 회귀계수와 상수에 대한 유의성 검정을 수행한다.

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

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

- 이때 t 분포의 자유도는 (상수와 회귀계수를 포함하므로) $n-2$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.

회귀계수와 상수의 유의성 검정

예제 1. lungcancer.csv는 이것은 8개 북유럽 국가의 1인당 담배 소비량(smoke)과 인구 100만 명당 폐암 발병자수(cancer)를 나타낸다. 폐암 발병자수를 종속변수로, 담배 소비량을 독립변수로 하는 회귀분석을 수행하고 유의성 검정 결과를 해석하시오.

회귀계수와 상수의 유의성 검정

다시 폐암 자료를 가지고 회귀분석을 수행해보자.

- 이제 회귀계수 뿐 아니라, 표준오차(standard error), t 통계량, 유의확률, 95% 신뢰구간도 어떤 의미인지 이해할 수 있다.
- 흡연은 폐암 발생의 통계적으로 유의한(statistically significant) 변수인가?
- 회귀계수와 상수의 95% 신뢰구간은 각각 어떻게 되는가? 이것을 색칠공부하듯 그림으로 표시해보자.

회귀계수와 상수의 유의성 검정

통계적 유의성와 실질적 유의성은 다른 개념이다

- 통계적 유의성(statistical significance)을 해석할 때 가장 흔한 실수 중 하나는 유의확률(p-value)이 작은 것을 가지고 관계의 강도(strength)로 해석하는 것이다. 유의확률은 단지 $H_0 : \beta = 0$라는 옳은 귀무가설을 기각하는 가능성을 보여줄 뿐이다.
- 실질적 유의성(substantial significance)은 (통계적으로 유의한가와는 별개로) 그 강도가 실제로 얼마나 센가의 문제를 다룬다.
- 예컨대 1시간 게임을 더하게 되면 독서 시간이 2분 줄어든다는 발견(Cummings and Vandewater 2007)은 설령 통계적으로 유의하더라도 실질적으로는 그다지 유의하지 않다.
- 그러므로 통계적으로 유의한 결과를 얻었더라도, 그 관계를 "실질적인 의미가 담겨있는 언어로 해석하고 음미하여" 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.

다중회귀모형

다중회귀모형

현실에서는 아무도 단순회귀모형을 사용하지 않는다.

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

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

- 하지만 이것은 학습용 혹은 연습용에 불과하다.
- 반면 (종속변수는 여전히 1개지만) 모형 안에 k개의 독립변수가 투입된 경우를 다중회귀모형(multiple regression model)이라고 부른다.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

다중회귀모형

다중회귀분석은 특정 변수의 순효과를 살펴보는데 유용하다.

- 가령 결석 시수(skipped)가 올해 학점(termgpa)에 영향을 미친다는 회귀모형을 세우고 오차항을 최소화하는 b_0와 b_1을 다음과 같이 추정하였다고 하자.

$$\text{termgpa} = 3.043 - 0.076 \cdot \text{skipped}$$

- "결석 시수가 한 단위 증가할 때, 올해 학점은 0.076점만큼 감소한다."
- "결석 시수가 0일 때의 올해 학점은 3.043이다."
- 하지만 학점과 관련있는 변수는 결석 이외에도 많다!

다중회귀모형

- k개의 독립변수를 모형에 투입했다면 여러 영향력은 각각에 해당되는 변수 안으로 나뉘어 흡수된다(Why?).
- 그러므로 다중회귀분석은 다른 변수들의 효과를 통제한 상태에서 특정 변수의 순효과(net effect) 또는 부분효과(partial effect)를 살펴보는데 유리하다.
- 가령 결석 횟수(skipped) 외에, 숙제 제출(hwrte) 역시 올해 성적(termgpa)에 영향을 미칠 것이다.

$$\text{termgpa} = b_0 + b_1 \cdot \text{skipped} + b_2 \cdot \text{hwrte} + e$$

- 다중회귀분석을 통해 '숙제 제출의 효과를 통제했을 때(즉 숙제 제출의 정도가 모두 똑같을 때)' 결석 횟수가 한 단위 변화하면 올해 성적이 얼마만큼 변화하는지 살펴볼 수 있다.

다중회귀모형

다중회귀분석을 수행할 때는 변수의 사전 체크에 주의를 기울여야 한다.

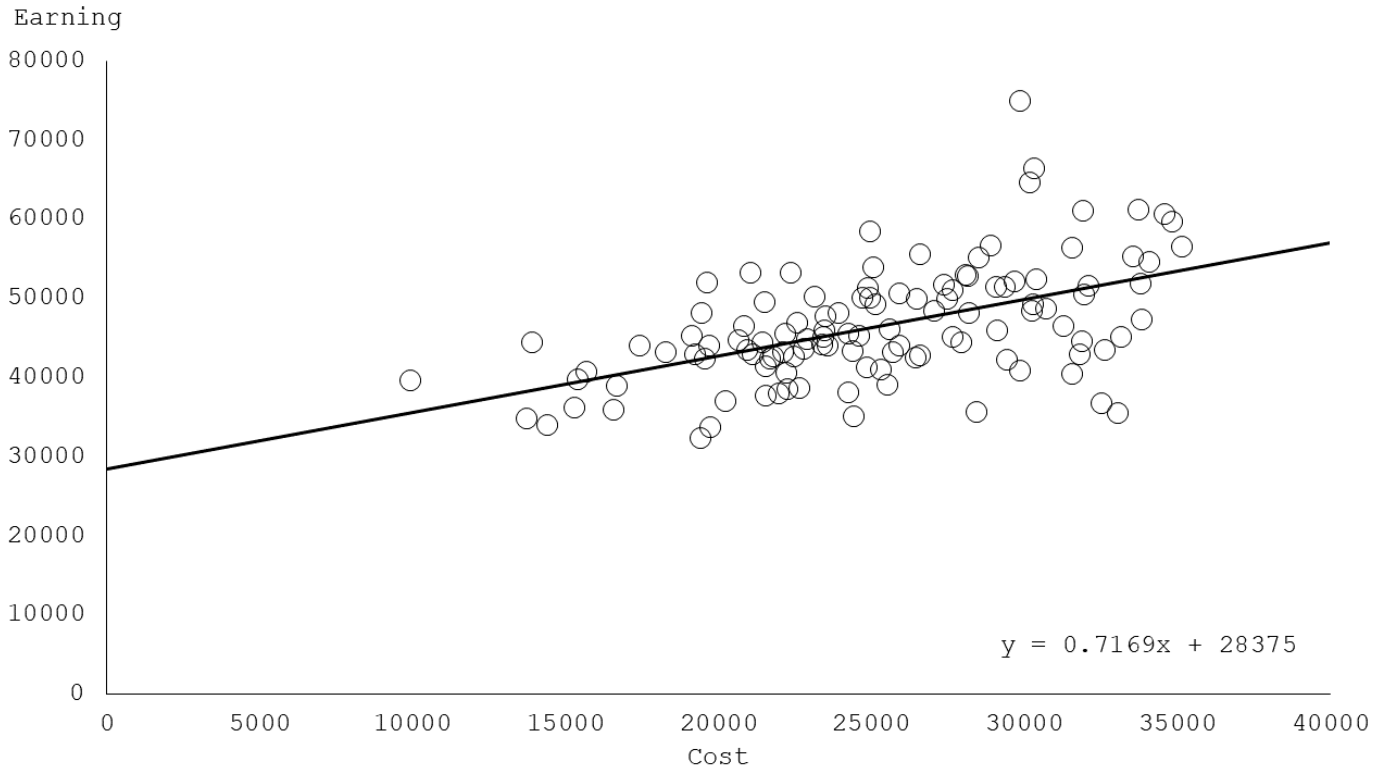
- 여러 개의 독립변수를 모형에 한꺼번에 투입하다보면 하나하나를 꼼꼼하게 살펴보지 않고 그냥 대충 집어넣는 경우가 많다. 이것은 매우 위험하다!
- 개별 변수의 척도가 어떻게 구성되어 있는지, 분포는 어떠한지, 결측치(missing values)가 있는지 등을 반드시 꼼꼼하게 살펴보아야 한다.
- 엑셀에서 필터링(filtering)이나 소팅(sorting) 등을 통해 꼼꼼히 자료를 살펴보아야 하고, [자료 분석]에서 "기술 통계법"이나 "히스토그램" 등을 활용해 변수의 분포도 살펴보아야 한다.
- 두 회귀분석에서 사용될 표본 크기가 똑같은지 확인해보자.
- 불편하지만 엑셀에서는 반드시 모든 독립변수들이 나란히 붙어있어야 한다.

다중회귀모형

예제 2. college.csv는 미국 대학 파일별 졸업생의 평균수입(earnings), 학자금 상환비율(debt), 등록금(cost), 졸업률(grad), 대학이 도시에 위치해 있는가 여부(city)에 관한 정보를 담고 있다. 고래는 미국 유학을 계획 중에 있으며, 특히 등록금에 비해 졸업 후에 얼마를 버는지 잘 따져보고 대학을 결정하고자 한다. 학자금 상환비율의 영향력을 통제하였을 때와 그렇지 않을 때, 등록금과 졸업생 평균수입의 연관성이 어떻게 달라지는지 살펴보시오.

다중회귀모형

- 분석에 투입할 자료만을 남겨두고 나머지는 편의상 삭제할 수 있다. 각 변수별로 정렬해보고 맨 끄트머리를 살펴보아 혹시 극단치(outliers)가 있지 않는지 살펴보자.
- 반드시 산점도와 적합선을 그려 선형관계가 존재하는지도 살펴보아야 한다. 등록금(cost)와 평균수입(earnings)의 관계는 산점도와 적합선으로 표현될 수 있다.



다중회귀모형

- 이제 우선 등록금이 졸업후 평균수입에 미치는 영향을 파악하기 위해 회귀분석을 수행할 수 있다.
- 한편 학자금 상환비율의 영향력을 배제하고 등록금만의 순효과를 파악하기 위해서는 다중회귀분석이 적절할 것이다.
- 그러므로 먼저 (1) 등록금과 졸업후 평균수입 사이의 단순회귀분석을 먼저 수행하고, 그 다음에 (2) 학자금 상환비율을 추가한 다중회귀분석을 수행하여 그 차이를 살펴보면 된다.
- 엑셀에서 [자료]-[자료 분석]을 통해 “회귀 분석”을 선택하자. 다중회귀분석에서는 모든 독립변수를 한번에 선택해야 하는 점에 주의하자.

다중회귀모형

- 다음은 단순회귀분석에서 추정된 등록금의 효과로, 학자금 상환비율의 영향력은 제대로 반영하지 못하고 있다.

$$\text{earnings} = 28375.405 + 0.717 \cdot \text{cost}$$

- 반면 다중회귀분석의 추정 결과는 다음과 같다.

$$\text{earnings} = 2526.991 + .570 \cdot \text{cost} + 334.338 \cdot \text{debt}$$

- 잠시 차이점을 살펴보자.

다중회귀모형

- “학자금 상환비율의 효과를 무시하였을 때, 등록금이 1000달러 증가하면 평균수입은 717달러만큼 증가한다.”
- “학자금 상환비율의 효과를 통제하였을 때, 등록금이 1000달러 증가하면 평균수입은 570달러만큼 증가한다.”
- “등록금과 학자금 상환비율의 중앙값(median)은 각각 24957.5달러와 90%인데, 그 경우 예상되는 평균수입은 약 46,843 달러이다.”
- “고래가 만일 등록금 30,000달러 정도에 학자금 상환비율이 최소 90%인 학교에 간다면, 고래의 예상되는 평균수입은 약 49,717달러이다.”

다중회귀모형

예제 3. WAGE2.CSV을 사용하여 월급(LWAGE)을 예측하기 위해 본인 교육수준(educ), 아버지 교욕수준(feduc), 직무 경력(exper), 근무 시간(hours), 지능지수(IQ), 형제자매의 수(sibs)를 독립변수로 하는 다중회귀모형을 설정하고 이를 수행한 뒤, 적절히 해석하시오. 하나의 독립변수를 골라 종속변수와의 연관성을 시각화하고 해석하시오.

다중회귀모형

모든 요소를 빼먹지 않아야 한다.

- 두 변수간 연관성을 살펴보기 위해 적어도 산점도와 적합선을 꼭 그릴 것.
- 회귀계수와 상수에 대한 귀무가설을 제대로 설정할 것.
- 회귀식을 정확히 서술한다. 통계적으로 유의하지 않더라도 빼지 않는다!
- (통계적으로 유의하면) 회귀계수를 정확하게 해석한다. 특히 해석의 단위에 신경쓴다.
- 유의성 검정 뿐만 아니라 평범한 언어로 해석해야 한다!