회귀분석

보통최소자승 회귀분석

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

July 5, 2023

진행 순서

안내사항

안내사항

수업에 관해 몇 가지 안내사항이 있습니다.

- 머지않아 양적 방법론을 사용해 논문을 쓰고자 하는 사람들을 위한 수업입니다.
- 펜과 노트를 준비하여 수식을 '손으로' 옮겨적는 연습을 하세요.
- 영어와 수학을 피하지 마세요.
- 강의안과 자료는 [여기]에서 무료로 다운받으세요. 공유 또는 판매하지 마세요. 오늘이 지나면 링크는 파기합니다.
- 메일로 질의응답하지 못하는 점 양해바랍니다.

안내사항

이 수업에 앞서 요구되는 기초지식이 있습니다.

- 특히 기술통계(descriptive statistics), 이론적 확률분포(theoretical probability distribution), 추정(estimation)과 가설검정(hypothesis test)의 논리 등 기초통계학은 알고 있어야 합니다.
- 이것들을 모르더라도 수업에는 따라올 수 있고 "어떻게(how)" 하는지는 배울 수 있지만 "왜(why)" 하는지는 모르게 됩니다.
- SPSS의 기초적인 사용방법은 숙지해야 합니다. 이것을 전혀 모르면 따라올 수 없습니다.
- 필요한 교재는 서점이나 도서관에서 자유롭게 선택할 수 있습니다.

선형회귀모형과 선형회귀식

선형회귀모형과 선형회귀식

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 기울기(slope)라고 부른다.
- 한편, β_0와 β_1를 어떻게 설정하더라도 결국 자료를 완벽하게 설명할 수는 없다. 따라서 오차항(error term) ϵ_i을 선형회귀모형에 추가한다.

선형회귀모형과 선형회귀식

- 선형회귀모형과 선형회귀식은 개념적으로 구분된다. 먼저 선형회귀모형은 다음과 같다.

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

- 선형회귀모형의 조건부 기대값(conditional expectation)인 선형회귀식(linear regression equation)은 다음과 같다(증명 생략). 더이상 오차항은 들어있지 않다.

$$E(Y|X) = \beta_0 + \beta_1 X$$

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

선형회귀모형과 선형회귀식

회귀분석의 핵심은 두 양적 변수 사이의 관계를 직선으로 요약하는 것이다.

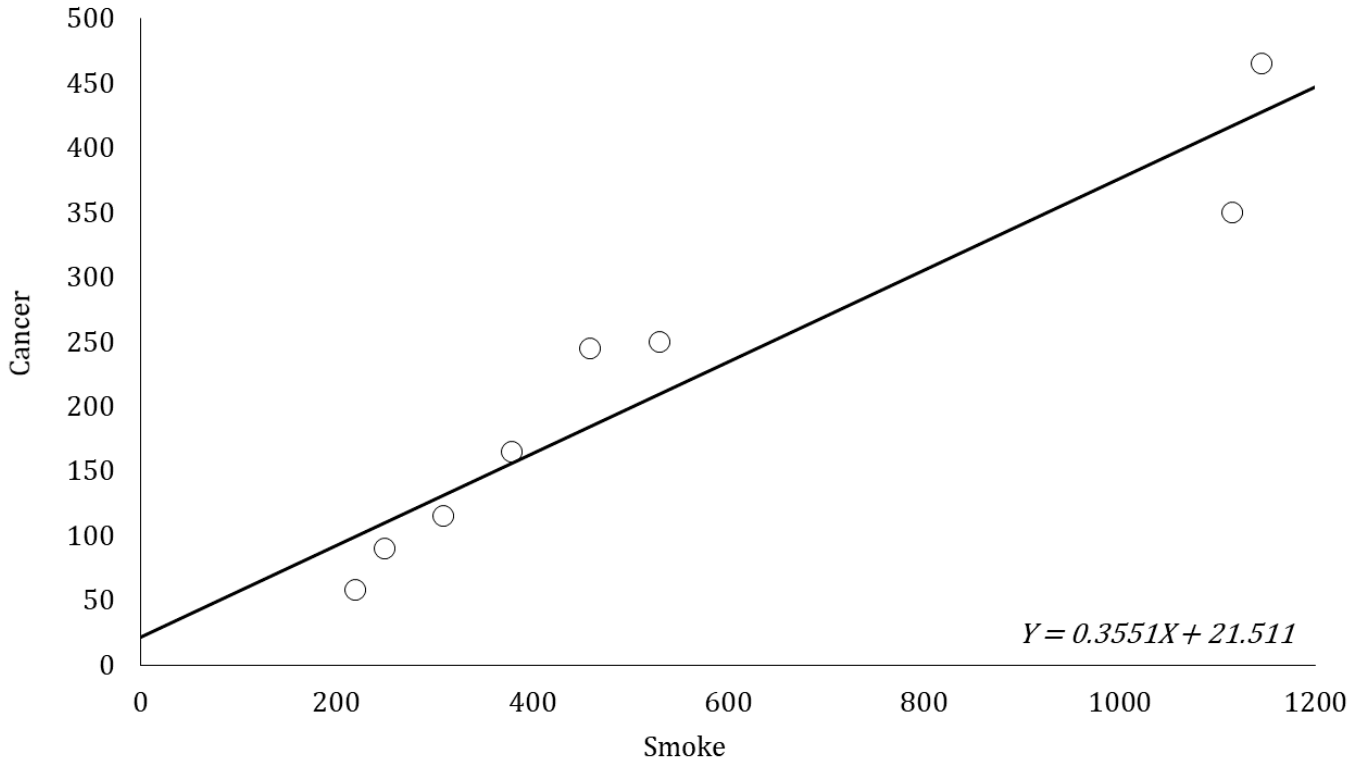
- 여러분이 이미 상관분석(correlation analysis)을 배웠다면 적합선(fitting line)에 친숙할 것이다.
- 마찬가지로 주어진 자료의 두 양적 변수(quantitative variable) X와 Y 사이에 가장 잘 맞는 직선(best-fitting straight line)을 그어 그 관계를 나타내 보일 수 있다.
- 이렇게 자료를 관통하는 하나의 선을 찾아내는 절차가 바로 회귀분석(regression analysis)의 핵심이다.

선형회귀모형과 선형회귀식

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

선형회귀모형과 선형회귀식

- 1인당 담배 소비량(smoke)이 100만 명당 폐암 발병자수(cancer)에 영향을 미친다고 보는 것이 타당하다.
- SPSS에서 [그래프]-[차트 작성기]에서 적합선이 있는 산점도(scatterplot)를 선택한다.
- 어떤 조건을 갖춘 적합선이 가장 잘 데이터를 나타낼 수 있을까? 만일 기울기와 절편이 달라지면 어떤 결과가 될까?



선형회귀모형과 선형회귀식

- SPSS에서 [분석]-[회귀분석]-[선형]을 선택하여 추정된 회귀식은 다음과 같다.

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

- 즉, $\widehat{\beta}_0 = 21.511$이고 $\widehat{\beta}_1 = 0.355$이다. 추정량(estimates)에 대해서는 이렇게 ˆ(hat)을 붙인다.
- "국가의 1인당 담배 소비량이 한 단위 증가할 때, 100만 명당 폐암 발병자수는 0.355만큼 증가한다."
- "아무도 흡연하지 않는 나라에서 100만 명당 폐암 발병자수는 21.511이다."
- 일반적으로 표현하자면, 독립변수 X의 값이 한 단위 변화하면 회귀계수(regression coefficient) β_1만큼 종속변수 Y에 영향을 미친다.
- 회귀계수 및 상수의 해석은 무척 단순하지만 연습을 필요로 한다!

회귀계수의 도출

단순선형회귀모형

오차를 전반적으로 최소화하는 적합선이야말로 가장 잘 맞는 직선이다.

- SPSS에서 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(predicted) Y_i (혹은 $\widehat{Y}$) 간의 차이는 곧 오차(error)라고 볼 수 있다.

$$\widehat{e_i} = Y_i - \widehat{Y_i} = Y_i - (\widehat{\beta_0} + \widehat{\beta_1} X_i)$$

- 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.

회귀계수의 도출

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 $\widehat{\beta}_0 = 21.511$, $\widehat{\beta}_1 = 0.355$이므로 독립변수인 $smoke_i$에 원하는 값을 대입하면 종속변수인 $cancer$를 예측할 수 있다.

$$\widehat{cancer} = 21.511 + 0.355 \cdot smoke$$

- 만일 1인당 담배 소비량이 1000인 가상의 국가에서는 인구 100만 명당 폐암 발병자수가 몇 명인지 예측해보자.
- 1인당 담배 소비량이 평균이라면 인구 100만 명당 폐암 발병자수는 몇 명일까?
- 예측은 회귀분석의 대단히 유용한 기능 중 하나이다. 이것으로 주가(stock price)나 집값 등에 관한 모형을 세우고 회귀계수 및 상수를 추정한 뒤, 조건별로 가격을 예측해 볼 수 있다.

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.

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

- 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.

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

- 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 $y = b_0 + b_1 X + e$에 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.

회귀계수와 상수의 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.

회귀계수의 유의성 검정

유의확률에 관한 정보를 요약하기 위해 이제부터 *을 붙이기로 한다.

- 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).

회귀계수의 유의성 검정

연습 2. 당신은 슬슬 학점에 대해 걱정이 들기 시작했다. 아르바이트를 열심히 하다보니 학교 수업을 4번 정도 빼먹었던 것이다. 당신은 자신의 시험 성적을 예측해보고자 과거 수강생 자료를 확보하였고 회귀분석을 활용하여 자신의 학점(GPA)을 예측해보고자 한다.

(1) attend.sav를 사용하여 독립변수를 skipped로 종속변수를 termgpa로 하여 회귀분석을 수행하시오.

(2) 추정된 회귀식을 제시하고 유의성 검정을 수행한 뒤, 이를 해석하시오.

(3) 당신의 예상 학점은 몇 점인지 예측하시오.

회귀계수의 유의성 검정

- 결석 시수(skipped)가 올해 학점(termgpa)에 영향을 미친다는 회귀모형을 세우고 (오차제곱합을 최소화하는) β_0와 β_1을 다음과 같이 추정할 수 있다.

$$Y = 3.043 - 0.076X$$

- "결석 시수가 한 단위 증가할 때, 올해 학점은 0.076점만큼 감소한다."
- "결석 시수가 0일 때의 올해 학점은 3.043점이다."

회귀계수의 유의성 검정

- SPSS 분석 결과에 따르면 $b_1 = 0.625$이고 $SE_{b_1} = 0.115$이다. 그러므로 t 값은 5.443 (=0.625/0.115)이다.
- 양측검정(two-tailed test)에서 이러한 t 값보다 극단적인 값이 나올 확률은 0.001 보다 작다.
- 그러므로 $\beta_1 = 0$이라는 귀무가설을 0.1% 유의수준에서도 자신있게 기각하고 이렇게 결론내릴 수 있다:
  "결석 시수는 99.9% 신뢰수준에서 통계적으로 유의하게 올해 학점과 연관되어 있다."
- 보통 통계적으로 유의하지 않으면 (찻잔 속의 태풍이므로) 해석을 아예 하지 않는 경우가 많다.

회귀계수의 유의성 검정

- 마지막으로 추정된 상수 $\widehat{\beta}_0$와 회귀계수 $\widehat{\beta}_1$을 가지고 $\widehat{Y}$을 구할 수 있다.
- 앞서 추정된 결과에 따르면 $\widehat{\beta}_0 = 3.043$ 이고 $\widehat{\beta}_1 = -0.076$이므로 독립변수 X, 즉 skipped에 원하는 값을 대입하면 종속변수 예측값 $\hat{Y}$을 얻을 수 있다.

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

- "수업 시수를 4번 빼먹은 당신의 예상 학점은 2.739이다."

다중회귀모형

다중회귀모형

현실에서는 아무도 단순회귀모형을 사용하지 않는다.

- 단순회귀모형(simple regression model)에서는 오로지 하나의 독립변수만 고려하였다.

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

- k개의 독립변수가 투입된 모형을 다중회귀모형(multiple regression model)이라고 부른다.

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

다중회귀모형

다중회귀분석은 (다른 변수들의 효과를 통제한 상태에서) 특정 변수의 순효과를 살펴보는데 유용하다.

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

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

- 회귀계수와 상수의 해석은 오로지 주어진 두 변수 간 사이의 1:1 관계로만 제한된다.

  (1) "결석 시수가 한 단위 증가할 때, 올해 학점은 0.076점만큼 감소한다."
  (2) "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.
- 그러므로 다중회귀분석은 (다른 변수의 영향력으로부터 독립된) 특정 변수의 순효과(net effect) 또는 부분효과(partial effect)를 살펴보는데 유리하다.
- 가령 결석 횟수(skipped) 외에도 숙제 제출 백분율(hwrte)도 올해 성적(termgpa)에 영향을 미칠 것이다.

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

- 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 (여러 변수를 사용하는 만큼) 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 척도(scale)가 어떻게 구성되어 있는지, 분포(distribution)는 어떠한지, 극단치(outlier)는 없는지, 결측치(missing values)가 있는지 등을 반드시 꼼꼼하게 살펴보아야 한다.
- SPSS에서도 정렬(sorting) 등을 통해 꼼꼼히 자료를 살펴보아야 하고, [그래프]-[차트 작성기]에서 '히스토그램' 등을 꼼꼼히 체크해야 한다.

다중회귀모형

연습 3. college.sav는 미국 대학 파일별 졸업생의 평균수입(earnings), 학자금 상환비율(debt), 등록금(cost), 졸업률(grad), 대학이 도시에 위치해 있는가 여부(city)에 관한 정보를 담고 있다. 당신은 미국 유학을 계획 중에 있으며, 특히 등록금에 비해 졸업 후에 얼마를 버는지 잘 따져보고 대학을 결정하고자 한다. 학자금 상환비율(debt)의 영향력을 통제하였을 때와 그렇지 않을 때, 등록금(cost)과 졸업생 평균수입(earnings)의 연관성이 어떻게 달라지는지 살펴보시오.

다중회귀모형

- 별도의 회귀분석을 두 번 수행해야 한다는 점에 주의해야 한다!

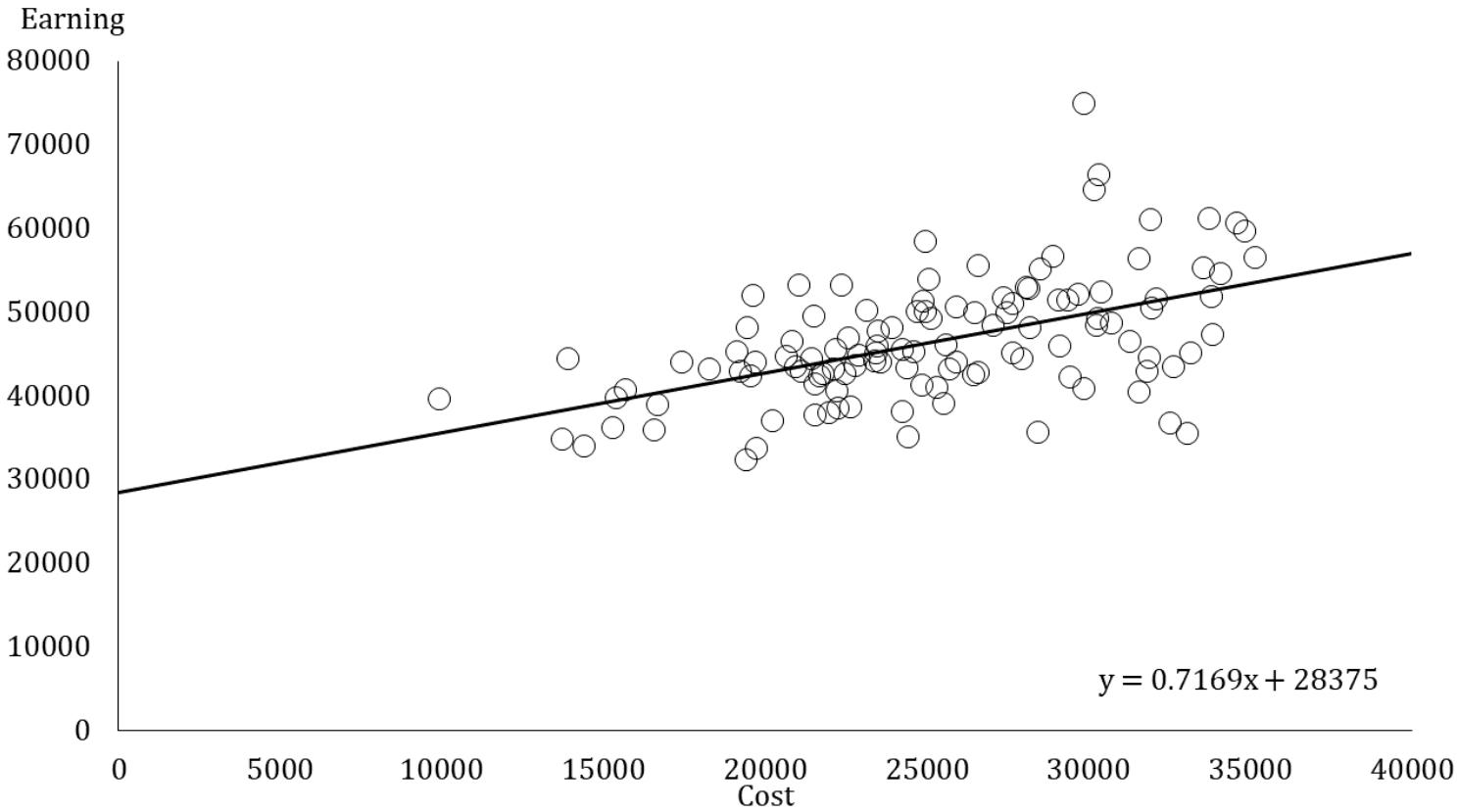
$$\text{earnings} = b_0 + b_1 \text{cost} + \epsilon_1 \qquad \text{(Model 1)}$$

$$\text{earnings} = b_0 + b_1 \text{cost} + b_2 \text{debt} + \epsilon_2 \qquad \text{(Model 2)}$$

- 즉 (1) 등록금과 졸업후 평균수입 사이의 단순회귀분석을 먼저 수행하고, 그 다음에 (2) 학자금 상환비율을 추가한 다중회귀분석을 수행한다. 각각을 따로 해석하고 나중에 두 모형 사이의 차이를 살펴본다.
- 두 모형에서 사용하는 관찰값(=사례)이 똑같도록 주의하자! 그러므로 분석에 사용되는 세 변수 중 하나라도 결측치나 극단치가 있으면 일관하여 삭제 또는 대체한다.

다중회귀모형

- 회귀분석은 어디까지나 직선으로 두 변수의 관계를 나타낸다. 그러므로 산점도와 적합선을 그려보고 직선으로 나타낼 수 없는 관계는 혹시 아닌지 체크해야 한다.
- 가령 등록금(cost)과 평균수입(earnings)의 관계는 다음과 같다.



다중회귀모형

- 첫번째 모형은 단순회귀분석이며, 여기서는 (학자금 상환비율은 고려하지 않은 채) 등록금과 졸업후 평균수입의 연관성만을 살펴보고 있다.

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

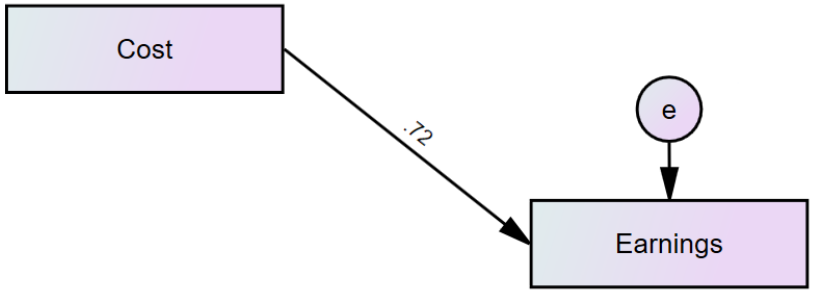
- "학자금 상환비율을 고려하지 않는다면, 등록금이 1000달러 증가할 때 평균수입은 717달러만큼 증가한다."
- 두번째 모형은 다중회귀분석이며, 여기서는 학자금 상환비율의 영향력을 통제하면서 등록금과 졸업후 평균수입의 연관성을 살펴보고 있다.

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

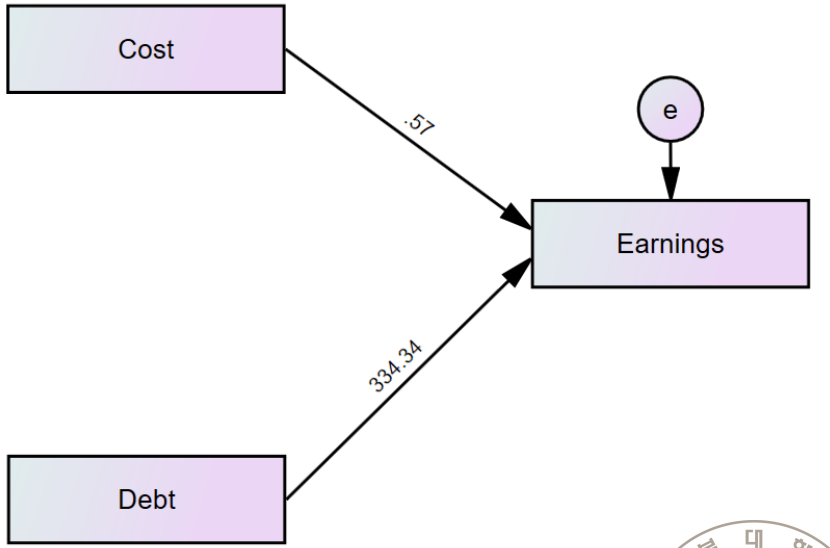
- "학자금 상환비율의 효과를 통제하였을 때, 등록금이 1000달러 증가하면 평균수입은 570달러만큼 증가한다."

다중회귀모형

- 물론 회귀계수를 해석하기 전에 유의성 여부를 먼저 판독해 두는 편이 좋다. 다행히 두 변수 모두 다 99.9% 신뢰수준에서 통계적으로 유의하다.
- 모형을 구축할 때나 그 결과물을 해독할 때 경로 도식(path diagram)을 그려보면 때때로 유용하다.



(Model 1)



(Model 2)

다중회귀모형

연습 4. attend.sav를 다시 사용하여 termgpa를 종속변수로, 기말시험(final), 예전 학점(priGPA), 과제(hwrte), 결석(skipped)을 독립변수로 하는 회귀모형을 구성하시오. 결과를 해석하고 어떤 변수가 termgpa를 가장 잘 설명하는지 서술하시오.

다중회귀모형

회귀분석의 결과표를 엑셀로 복사하여 가지런하게 꾸며야 한다.

- 모형 간에 똑같은 관찰값(=사례)을 유지하면서 변수만 추가해 나갈 수 있다.
- 이러한 분석 전략을 위계적 회귀분석(hierarchical regression analysis)이라고 부른다.
- 회귀분석 결과표에는 SPSS에서 사용하던 변수명을 그대로 남겨두지 말고 (독자가 알아볼 수 있도록) 똑바로 고쳐쓸 것!

다중회귀모형

- 회귀분석 결과표 안에 회귀계수 뿐 아니라, 표준오차(standard error), t 값, 또는 신뢰구간(confidence interval) 셋 중 하나는 함께 보고해야 한다.
- 통계적으로 유의하지 않은 변수를 멋대로 빼서는 안된다(물론 상수도 마찬가지이다).
- 표가 너무 길어지면 몇몇 변수들을 생략할 수 있으나 이 경우에도 반드시 보고해야 한다.
- 세부적인 것들은 출판된 논문을 최대한 흉내내가면서 배운다(김성훈2020.pdf를 살펴보자).

다중회귀모형

여러 독립변수를 고려하다보면 어떤 변수가 가장 중요한지 알고 싶어진다.

- 회귀계수 값의 크기로 변수 간 영향력의 차이를 비교할 수 없다(Why?).
- 이를 비교하기 위한 한가지 방법은 표준화 회귀계수(standardized regression coefficient)를 사용하는 것이다.
- 표준화에도 몇 가지 방법이 있는데 대체로 다음을 사용한다.

$$\beta_k^* = \frac{s_X}{s_{Y_k}} \beta_k$$

- 표준화 회귀계수는 값의 크기로 어떤 변수가 얼마나 중요한지 알려준다.

다중회귀모형

상수 해석을 위해 평균중심화가 필요할 수도 있다.

- 여기서 평균중심화(mean centering)란 독립변수 원점수 대신 이를 평균(mean)으로 빼준 편차(deviance)를 사용하는 방식이다.

$$Y_i = \beta_0 + \beta_1(X_i - \bar{X}) + \epsilon_i$$

- 예컨대 독립변수 X가 연령이고 종속변수 Y가 연봉인 회귀모형에서 상수의 의미 해석은 다소 모호하다(Why?).
- 이때 연령 원점수 X 대신 연령 편차 $(X - \bar{X})$를 대신 사용한다면 상수는 다음과 해석할 수 있다: "연령이 평균일 때, 연봉은 b_0이다."

다중회귀모형

- SPSS에서 변수를 우클릭하여 [기술통계량]을 구하고 평균값을 복사한다.
- 다시 [변환]-[변수 계산]에서 복사한 평균값을 빼주어 평균중심화한 새로운 변수를 만든다.
- 물론 회귀분석을 수행할 때는 새롭게 만든 변수를 사용하는 것을 잊지 말 것.
- 평균중심화 여부와 상관없이 회귀계수는 완전히 똑같아야 한다(달라지면 무언가 잘못한 것이다).
- 오로지 상수값만 달라지며 그 상수 b_0는 해당 독립변수가 평균일 때($X = \bar{X}$), 예측되는 Y 값이라고 해석된다.

다중회귀모형

연습 5. nlswork.sav를 사용하여 로그 임금 $(\ln_w age)$, (tenure) (south) . .

모형의 적합도

모형의 적합도

개발한 회귀모형이 현실 자료에 얼마나 적합한지 살펴보아야 한다.

- 우리가 모형을 세워 그것을 현실 자료에 맞추어(fit) 본 이상, 이것이 얼마나 잘 맞는가를 말할 수 있어야 한다. 이것이 모형의 적합도(goodness-of-fit)이다.
- 우리는 여러가지 적합도 지표 가운데 결정계수(coefficient of determination)와 분산분석만 공부한다.

모형의 적합도

첫번째로 결정계수의 기본 원리를 살펴보자.

- 주어진 자료에 전체 변량(total variation)이 있다면, 이것은 (모형에 의해) 설명된 변량(explained variation)과 (그렇지 못하고) 남은 변량(residual variation)의 합이라고 분해될 수 있다.

$$\sigma^2_{total} = \sigma^2_{explained} + \sigma^2_{residual}$$

- 그렇다면 설명된 변량 $\sigma^2_{explained}$와 전체 변량 σ^2_{total}의 비율은 모형의 높은 설명력을 의미한다고 볼 수 있다.

$$R^2 = \frac{\sigma^2_{explained}}{\sigma^2_{total}} = 1 - \frac{\sigma^2_{residual}}{\sigma^2_{total}}$$

- 이것이 바로 결정계수 R^2의 직관적 의미이다. 설명된 변량과 전체 변량의 비율이므로 0과 1사이에 놓인다. 1에 가까울수록 모형은 높은 적합도를 보인다고 할 수 있다.
- 만약 R^2가 0.45라면 "이 회귀모형은 주어진 자료의 전체 변량의 45% 를 설명한다"고 해석한다.

모형의 적합도

두번째로 분산분석 결과표 역시 모형 전체의 적합도를 보여준다.

- 아래와 같이 선형회귀모형을 설정하였을 때 최악의 추정 결과는 무엇일까?

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_{k-1} X_{k-1} + \beta_k X_k + \epsilon$$

- 그건 (상수 빼고) 모든 회귀계수가 0이 되는 상황이다(Why?).

$$\beta_1 = \beta_2 = \cdots = \beta_{k-1} = \beta_k$$

- 제대로 된 모형이라면 최소한 이런 경우는 부정할 수 있어야 한다. 만일 이런 경우가 나타난다면 "이 모형은 완전히 쓸모없다"는 말이나 마찬가지다.

모형의 적합도

- 그러므로 다음과 같은 가설구조를 검증해 볼 수 있다.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0 = \beta_k = 0$$
$$H_a : (\beta_1 \neq 0) \text{ or } (\beta_2 \neq 0) \text{ or } \cdots \text{ or } (\beta_{k-1} \neq 0) \text{ or } (\beta_k \neq 0)$$

- 이 가설구조를 테스트할 수 있는 방법은 일원분산분석(one-way ANOVA)이다.
- 다중회귀분석의 맥락에서 F 값은 다음과 같다. 단 k는 독립변수의 수를 뜻한다.

$$F_{(k,n-k-1)} = \frac{\sigma^2_{explained}}{\sigma^2_{residual}} = \frac{MS_{explained}}{MS_{residual}} = \frac{SS_{explained}/k}{SS_{residual}/(n-k-1)}$$

- F 값이 충분히 크면 유의확률(p-value)이 작아지므로 위 귀무가설을 기각할 수 있다.

다중회귀모형

연습 5. college.sav를 다시 사용하여 졸업생 평균수입(earnings)을 종속변수로, 학자금 상환비율(debt)과 등록금(cost)을 독립변수로 한 다중회귀식을 추정하시오. 그 적합도를 살펴보고 충실히 해석하시오.

모형의 적합도

- “학자금 상환비율과 등록금 두 독립변수를 사용한 선형모형은 자료의 전체 변량 중 36.2%를 설명한다.”
- 0.362라는 결정계수는 너무 낮을까? 그렇지만도 않다. 겨우 두 개의 변수만으로 이 정도 설명했다는 것은 나름 훌륭하다(분과에 따라 해석이 조금씩 다르다).
- 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.
- 바로 아래 조정된 결정계수(adjusted R^2)가 바로 이렇게 패널티가 가해진 결정계수이다(여기서는 별 차이가 없었다).

모형의 적합도

- ANOVA 표의 "Sig." 항목을 보자. 이것은 유의확률(p-value)을 뜻한다. 이것이 0.05 보다 작다는 사실은 95% 신뢰수준에서 다음의 귀무가설을 기각할 수 있다는 의미이다(Why?).

$$H_0 : \beta_1 = \beta_2 = 0$$

- 우리는 "모든 회귀계수가 0이다"라는 귀무가설을 기각하고 "최소한 하나의 회귀계수는 0이 아니다"라는 대립가설을 채택할 수 있었다. 다행이다~