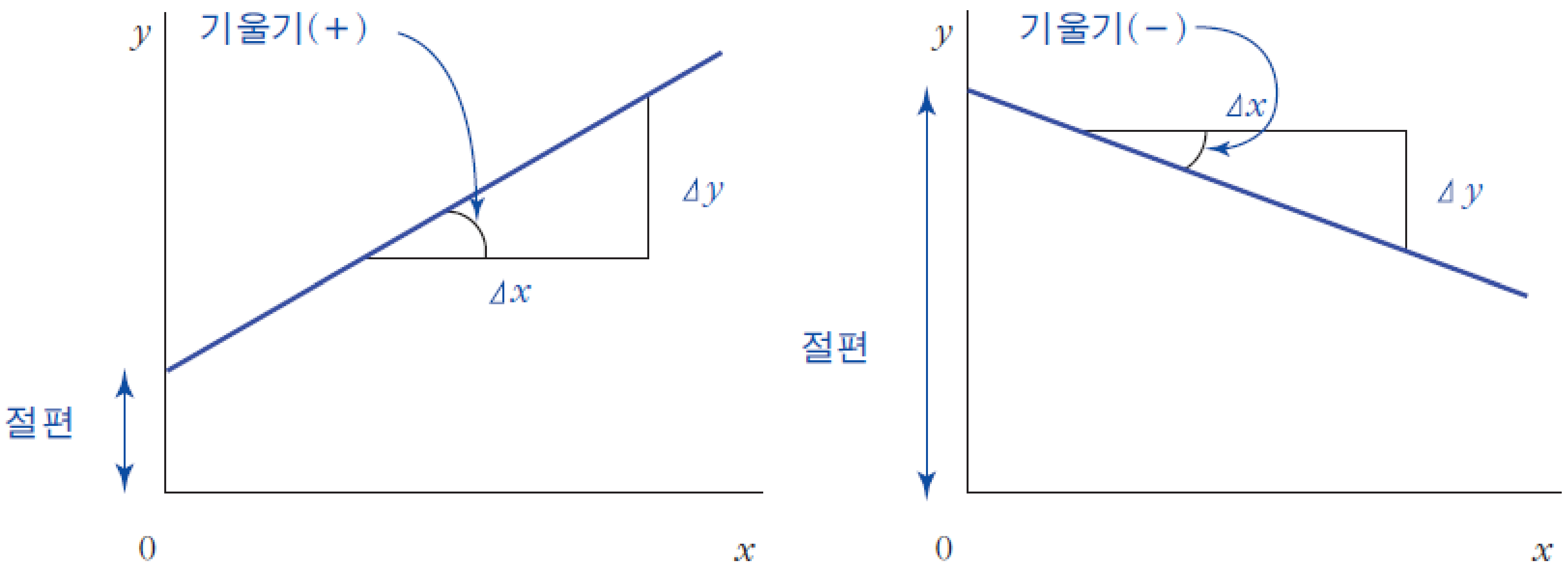
4.2 회귀직선

1) 회귀분석과 최소자승법
2) 중회귀분석의 직관적 이해

1. 회귀분석과 최소자승법

기울기와 절편

- 절편은 x가 0일 때 y값을 의미하며 기울기는 x가 1 만큼 증가할 때 y가 증가하는 정도를 의미한다.

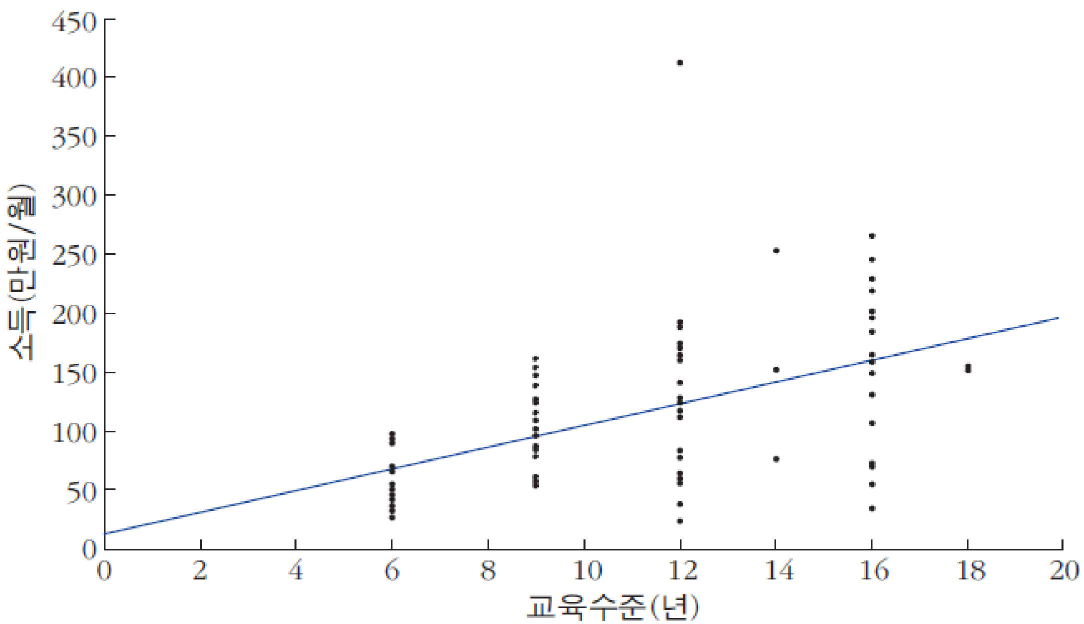


류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 161

1. 회귀분석과 최소자승법

교육수준과 월소득

- 교육과 월소득의 관계: 만 30-40 세의 도시 남성 198명을 대상으로 조사



평균 교육년수 = 12.5년

교육년수의 표준편차 = 2년

평균 소득 = 163만원

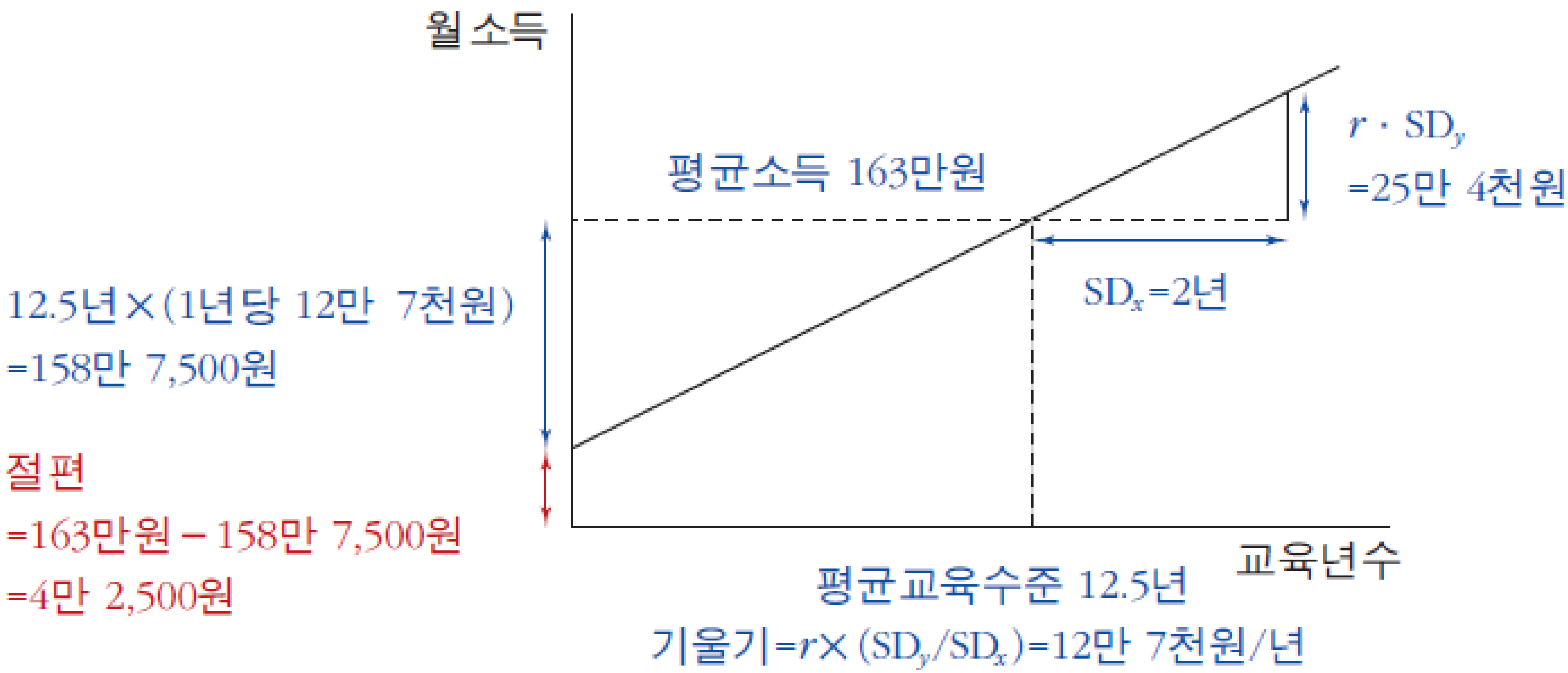
소득의 표준편차 = 77만원

상관계수 = 0.33

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 161

1. 회귀분석과 최소자승법

교육수준과 월소득

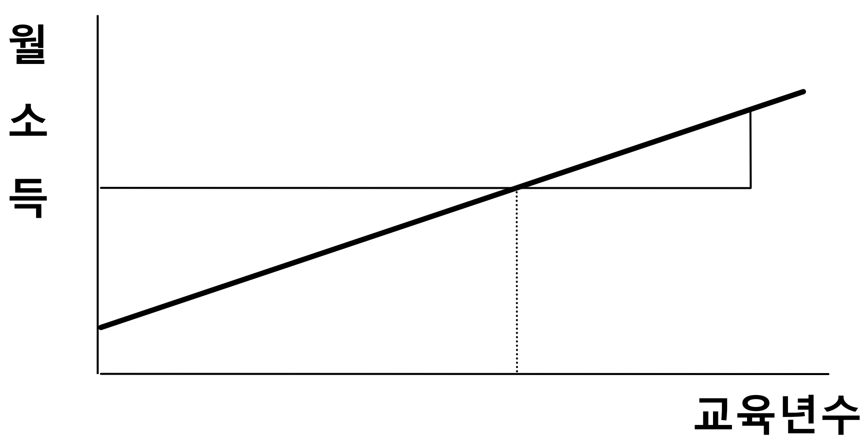


류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 163

1. 회귀분석과 최소자승법

교육수준과 월소득

- 교육과 월소득의 관계: 도시 여성 108명을 대상으로 분석한 조사
  - (추정된 월소득) = 98,000원 + (72,000원/년) × (교육년수)
  - $\text{기울기} = \mathrm{r} \times \frac{SD_y}{SD_x} = 0.43 \times \frac{60\text{ 만원}}{3.6\text{ 년}} \cong 72{,}000\text{ 원/년}$
  - y-절편 = 890,000원 – (72,000원/년)x11년 = 98,000원



평균 교육년수 = 11년

교육년수의 표준편차 = 3.6년

평균 소득 = 89만원

소득의 표준편차 = 60만원

상관계수 = 0.43

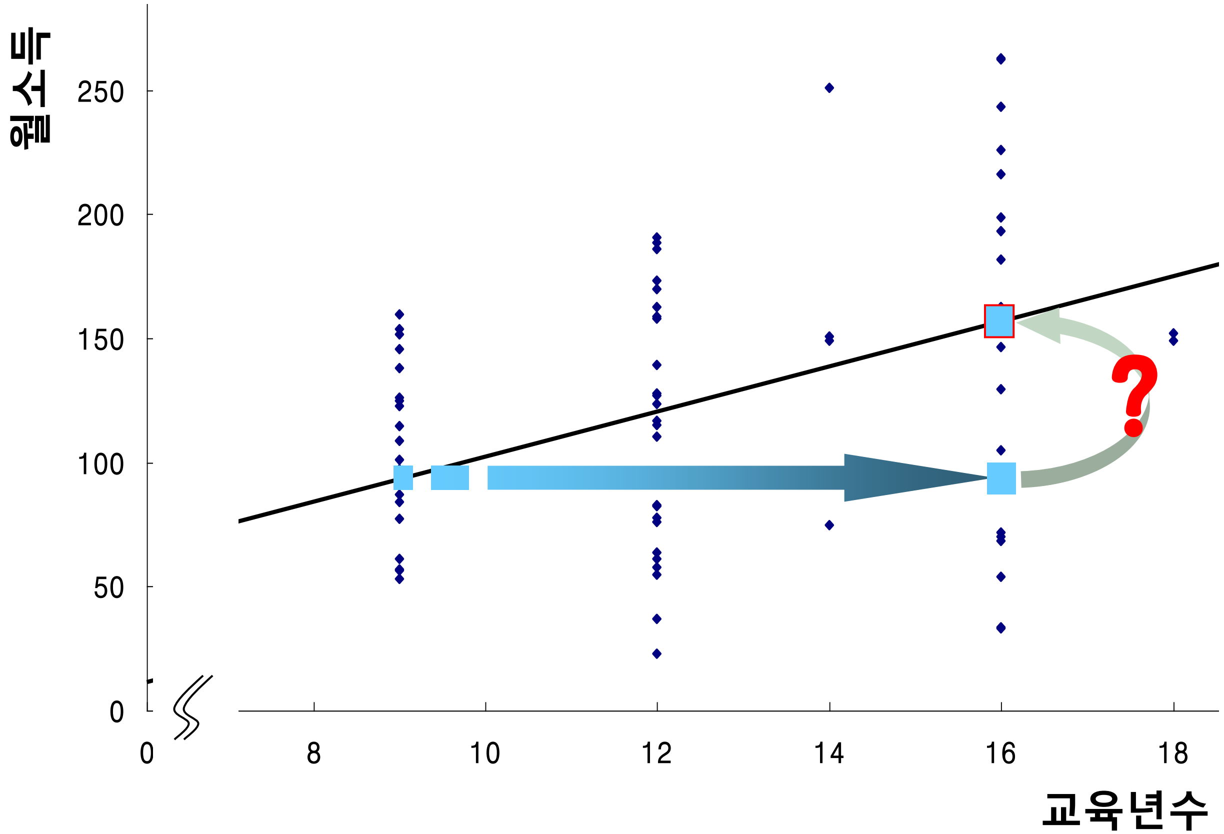
1. 회귀분석과 최소자승법

교육수준과 월소득

- 회귀직선 이용, 교육년수가 12년인 고졸여성과 16년인 대졸여성의 소득 추정
  - **교육년수가 12년인 여성의 추정소득**
    - 98,000원 + (72,000원/년) × 12년 = 962,000원
  - **교육년수가 16년인 여성의 추정소득**
    - 98,000원 + (72,000원/년) × 16년 = 1,250,000원

1. 회귀분석과 최소자승법

기울기 추정치에 대한 해석상의 주의



추정된 기울기를 '외부개입-내부반응'의 정도로 해석할 수 있을까?

- 통제된 실험으로부터 얻은 자료를 가지고 구한 기울기면 그런 해석이 가능하나 경험적 연구로 얻은 기울기면 꼭 그런 것은 아니다.

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 161

1. 회귀분석과 최소자승법

최소자승법

- **최소자승직선:** 모든 직선 중에서 x를 통해 y를 추정할 때 발생하는 추정오차들의 "제곱의 합"으로 측정한 전반적 크기를 가장 작게 만들어주는 직선
  - 산포도상의 각각의 점으로부터 하나의 직선까지의 수직거리를 정의
  - 수직거리의 "제곱 합"이 최소화 되는 직선을 회귀직선으로 선택
  - 수직거리의 제곱합을 최소화하는 것이나 RMS로 측정한 수직거리의 전반적 크기를 최소화하는 것이나 수학적으로 동일한 최적화 문제임
  - 즉, 최소자승법(method of least squares)은 모든 직선 가운데 수직거리의 전반적 크기를 최소화 해주는 직선을 구하는 방법임

- *Least Squares Demo* - University of South Carolina

1. 회귀분석과 최소자승법

후크의 실험

- 후크의 실험: 매단 추의 무게와 용수철 길이의 관계

표 8-1 후크의 법칙

추의 무게(kg)	길이(cm)
0	439.00
2	439.12
4	439.21
6	439.31
8	439.40
10	439.50

- x: 매단 추의 무게 , y: 용수철의 길이

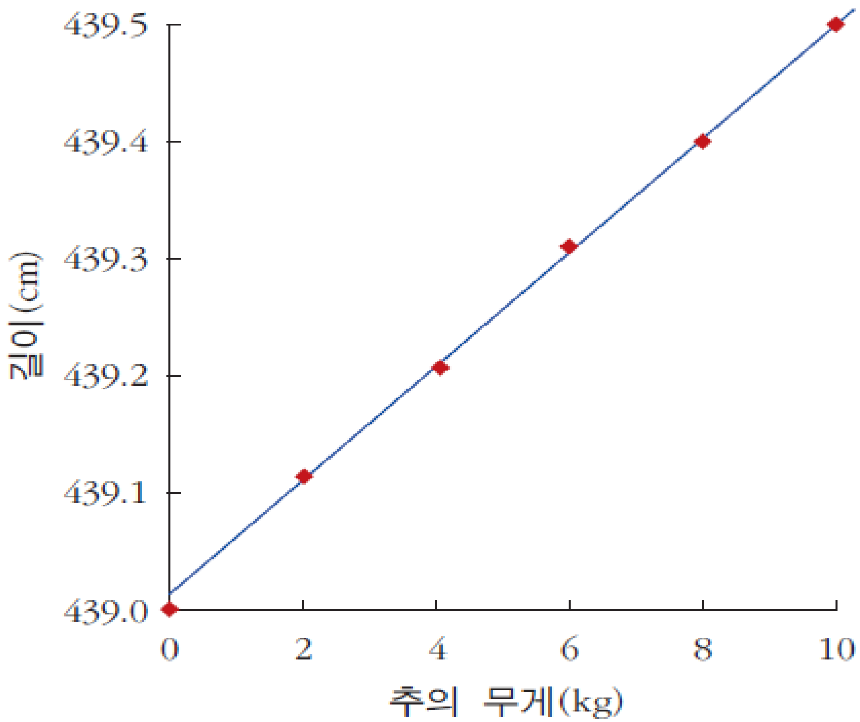
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 168

1. 회귀분석과 최소자승법

후크의 실험

- 추정된 용수철 길이(cm)=439.01cm+0.05cm/kg(매단 추의 무게(kg))

산포도



최소자승추정치

절편 = 439.01cm

기울기 = 0.05cm/kg

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 168

1. 회귀분석과 최소자승법

빅맥지수

$$(\text{환율}) = -57.31 + 1.81\ (\text{빅맥지수})$$

- 빅맥지수와 환율의 수준이 같지 않다.
- Absolute PPP 성립한다고 보기 어렵다.

1. 회귀분석과 최소자승법

빅맥지수

$$\ln(\text{환율}) = 0.29 + 1.01 \times \ln(\text{빅맥지수})$$

- 빅맥지수가 1% 변화할 때 환율도 대체로 1% 남짓 (1.01%) 변화하는 것으로 판단됨
- Relative PPP가 성립하지 않는다고 볼 통계적 근거가 없음

1. 회귀분석과 최소자승법

자산가격결정모형(CAPM)

- CAPM : 수익률에 대한 단일요인 모형(Single factor model of return)
- 시계열 회귀분석 방정식

$$\mathrm{r_{it}} = \alpha_i + \beta_i r_{mt} + \epsilon_{it}$$

- Fama & French(1992) : ‘기업규모’ 및 ‘장부가치/시장가치 비율’ 등 두 요인을 추가하여 수익률의 종목간 변동을 추가로 설명

1. 회귀분석과 최소자승법

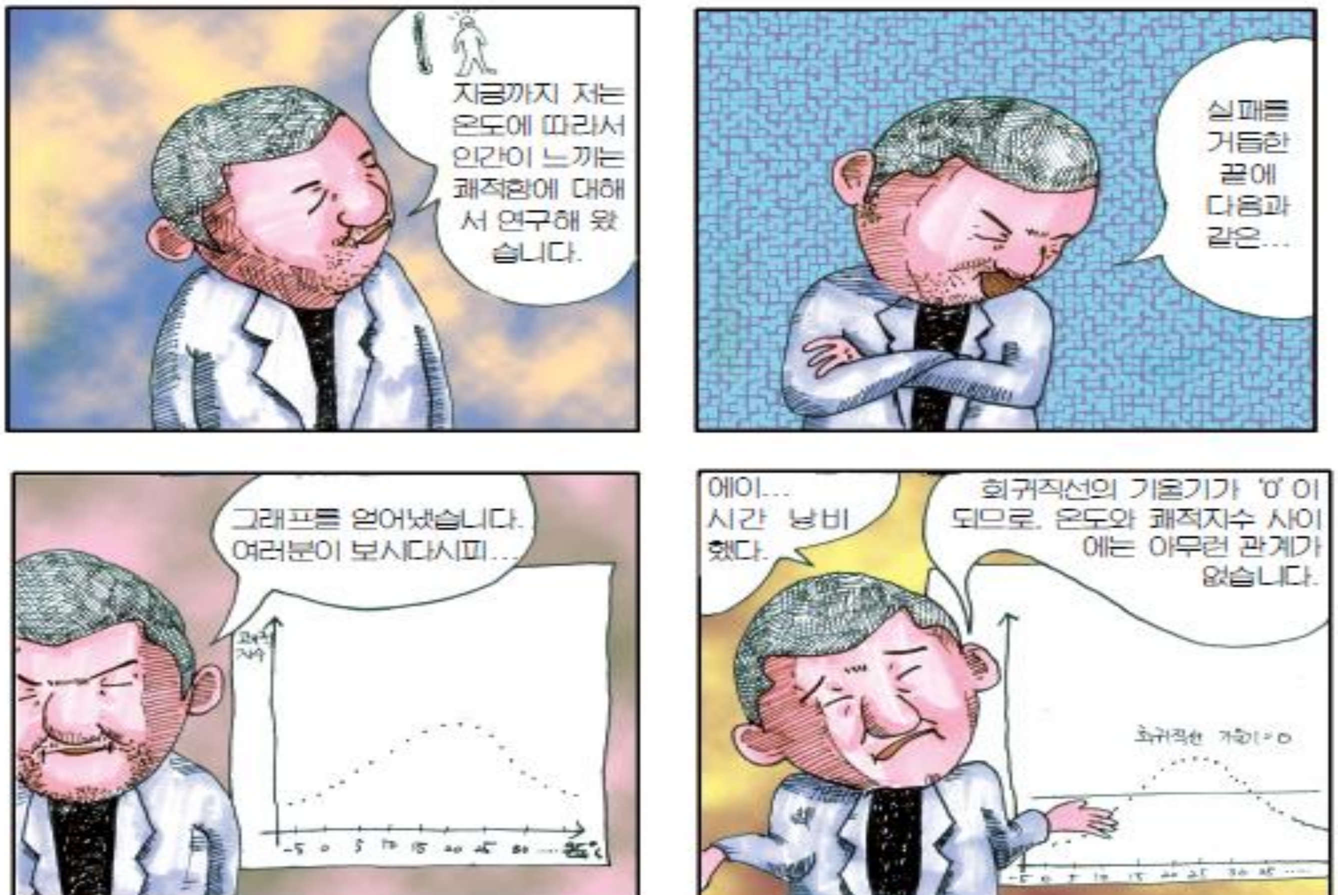
CAPM β의 추정

- 종목별로 월별 주가수익률 데이터를 이용하여 개별주식의 수익률을 시장포트폴리오(KOSPI)의 수익률에 대해 회귀분석한 결과
- 데이터: 월별 주가 자료(1999. 2.-2001. 12.)

기업명	β	t-value
Samsung Electronics	1.24	6.39
SK Telecom	0.95	3.71
KT	1.09	5.97
KEPCO	0.71	4.79
POSCO	1.00	8.12

1. 회귀분석과 최소자승법

회귀직선과 비선형관계



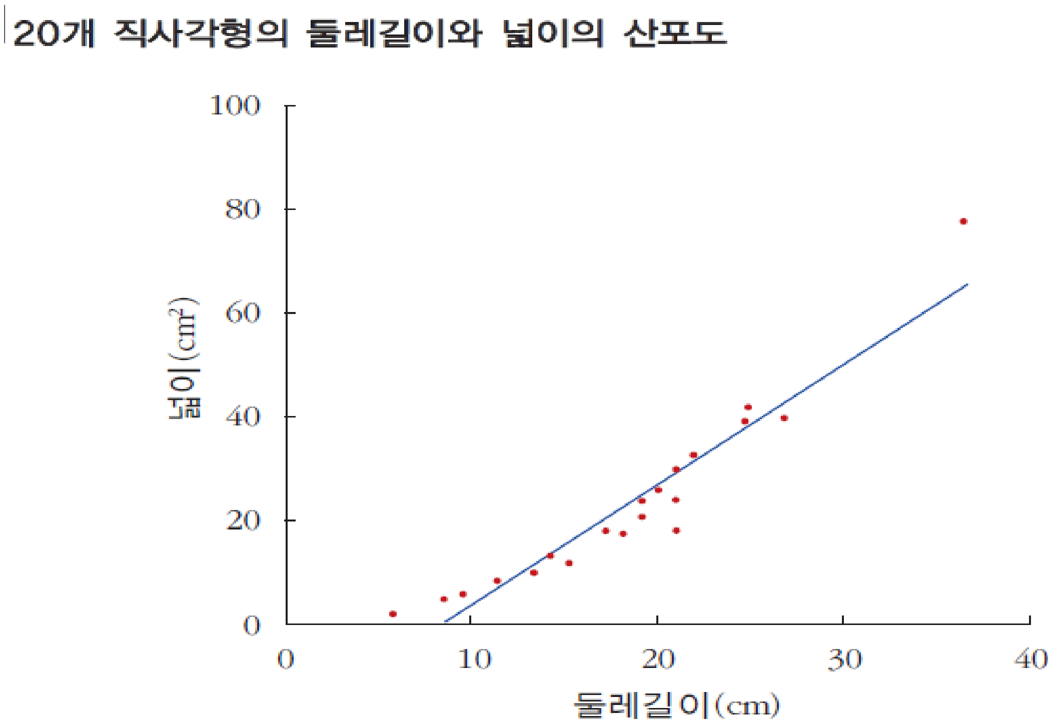
회귀직선은 선형관계만을 측정

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 160

1. 회귀분석과 최소자승법

직사각형의 둘레길이와 넓이의 산포도

- 넓이와 둘레길이간 상관계수=0.98: 이는 넓이와 높이라는 제3의 요인들 영향
- 넓이와 둘레 … 임



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 170

2. 중회귀분석의 직관적 이해

중회귀분석

- 종종 제3의 변수가 두 변수 x와 y 각각에 영향을 미쳐, 관심의 대상인 두 변수 상호간의 순수한 관계를 왜곡시키게 됨. 제3의 변수를 통제할 필요성 대두

1) 실험 (experiment)
2) 통계적 통제 1: 자료를 제3의 변수값에 따라 분류, 집단 별로 따로따로 분석
3) 통계적 통제 2: 중회귀분석

2. 중회귀분석의 직관적 이해

중회귀분석

- 부모의 사회적 지위(S)가 자녀의 소득(y)에 미치는 효과를 통제했을 때 자녀 본인의 교육수준(E)이 본인의 소득(y)에 미치는 순수 효과(b). 여기서 “순수”란 S를 통제했다는 의미임

$$y = a + b*E + c * S + (\text{오차})$$

2. 중회귀분석의 직관적 이해

남녀 노동자간 임금격차

- 남녀 간에 임금격차가 존재하는지 보기 위하여 다음의 단순회귀분석 모형을 추정하려고 한다.

(임금) = a + b(남성 더미변수) + (오차)

- 여기서 남성 더미변수는 남성에게 1을, 여성에게는 0의 값을 부여하는 질적 변수이다. 일반적으로 더미변수는 하나의 질적인 설명변수가 종속변수에 미치는 영향을 파악하고자 할 때 이용한다.

2. 중회귀분석의 직관적 이해

남녀 노동자간 임금격차

- 남녀 간의 임금격차는 남녀간 교육수준의 차이 등이 통제되지 않는 한 성차별의 증거로 보기 어렵다.

- 교육수준을 설명변수로 추가, 단순회귀분석 모형을 중회귀분석 모형으로 확장

(임금) = a + b(남성 더미변수) + c(교육연수) + (오차)

2. 중회귀분석의 직관적 이해

기술진보 반영한 물가지수 작성법

예) 컴퓨터 기술진보를 감안한 상태에서 지난 10년간의 컴퓨터 가격지수 작성

- 지난 10년간 판매된 데스크탑 컴퓨터에 대해 CPU, 메모리 등 각종 스펙(x), 판매연도(첫 해를 기준으로 $d_2,\ldots,d_{10}$ 등 총 9개의 연도더미변수들), 판매가격(y) 정보 수집
- log (y) 를 상수항, x, $d_2,\ldots,d_{10}$ 에 중회귀분석하여 $d_2,\ldots,d_{10}$ 의 계수 추정치인 $b_2,\ldots,b_{10}$ 얻음
- 첫 해의 가격지수를 100으로 두면, 둘째 연도, ..., 10번째 연도의 가격지수는 각각 100*exp(b_2), ..., 100*exp(b_{10}) 등으로 추정됨. 이를 시계열 그림으로 표현