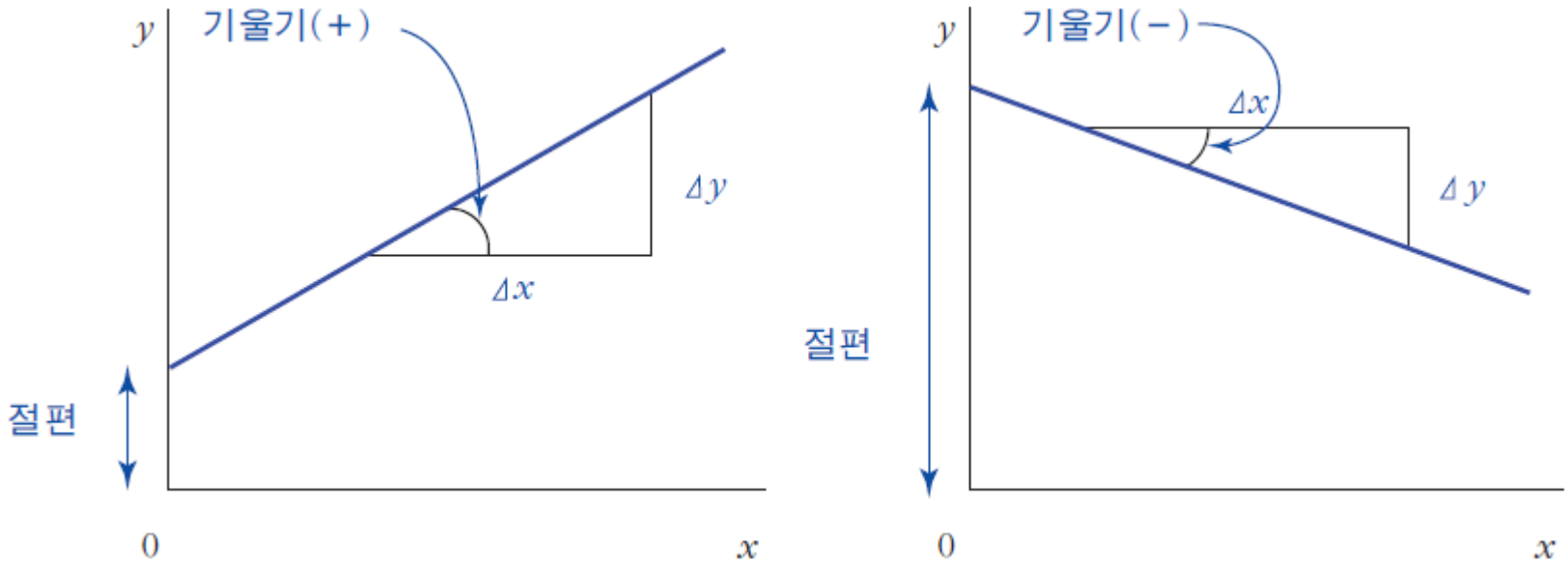
제 8 장 회귀직선

1. 기울기와 절편
2. 최소자승법
3. 회귀분석은 만병통치약이 아님
4. 중회귀분석
5. 총변동의 분해

1. 기울기와 절편

기울기와 절편

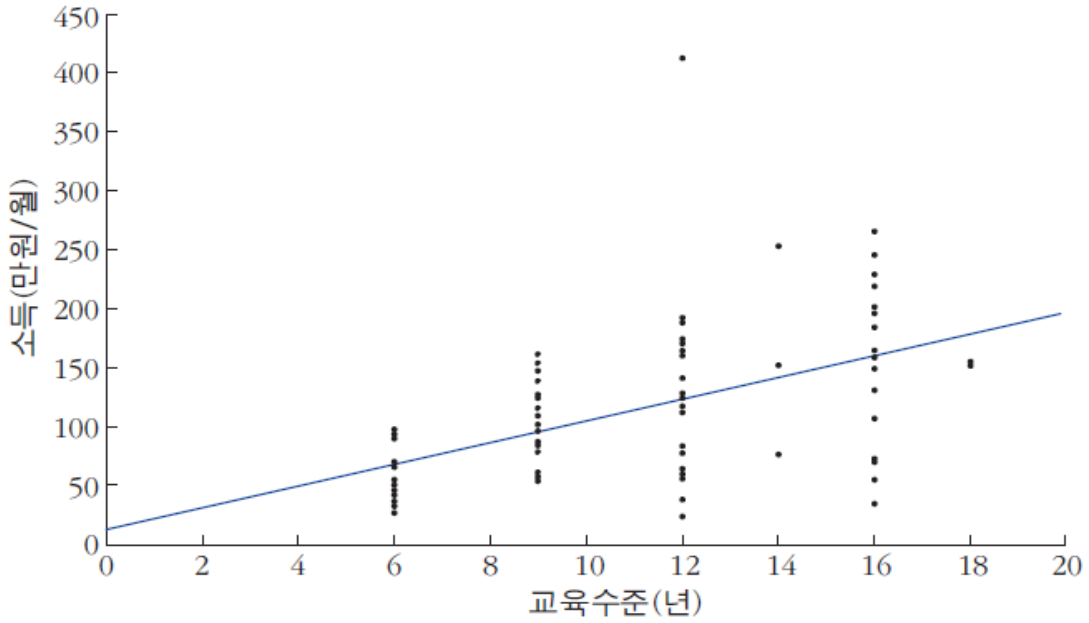
절편은 x가 0일 때 y값을 의미하며 기울기는 x가 1 만큼 증가할 때 y가 증가하는 정도를 의미한다.



1. 기울기와 절편

교육수준과 월소득

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



평균 교육년수 = 12.5년

교육년수의 표준편차 = 2년

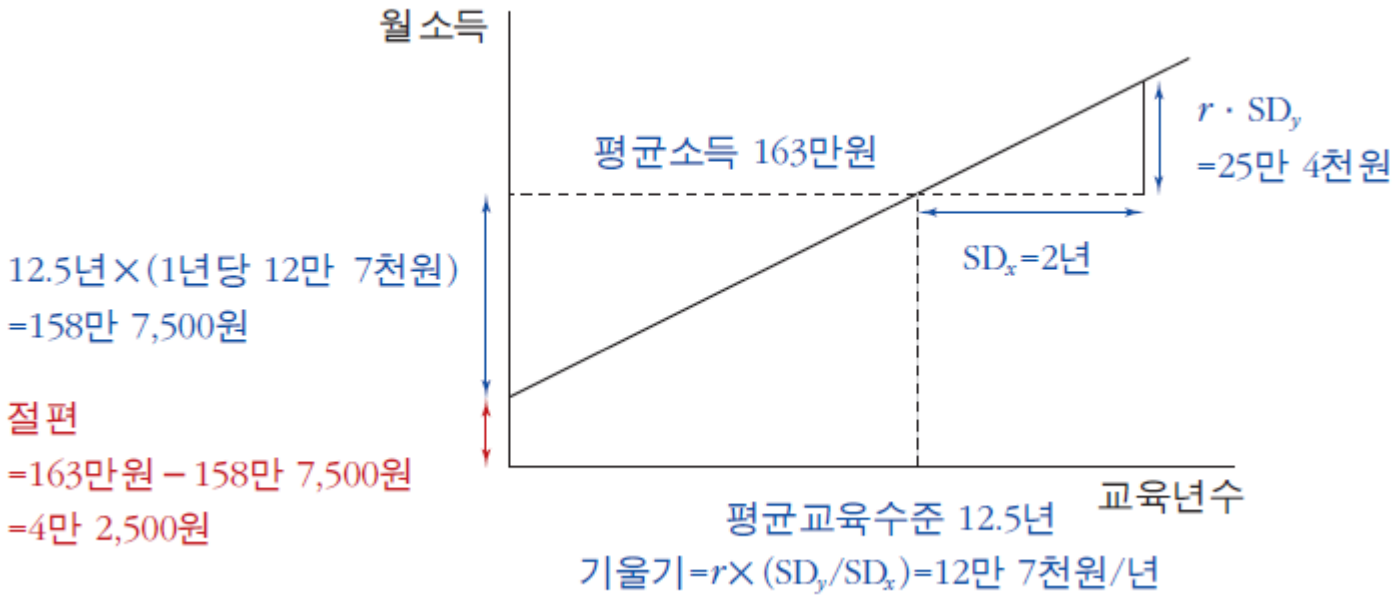
평균 소득 = 163만원

소득의 표준편차 = 77만원

상관계수 = 0.33

1. 기울기와 절편

교육수준과 월소득



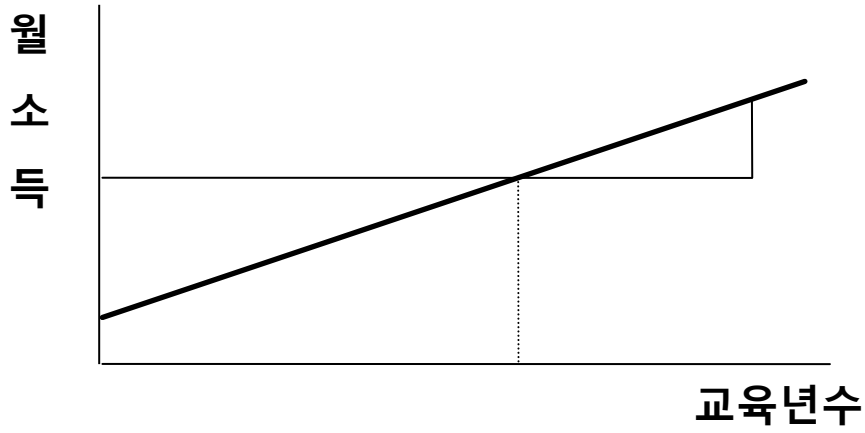
1. 기울기와 절편

교육수준과 월소득

교육과 월소득의 관계: 도시 여성 108명을 대상으로 분석한 조사

- (추정된 월소득) = 98,000원 + (72,000원/년) × (교육년수)
- $\text{기울기} = r \times \frac{\text{SD}_y}{\text{SD}_x} = 0.43 \times \frac{60\text{만원}}{3.6\text{년}} \approx 72{,}000\text{원/년}$

  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. 기울기와 절편

교육수준과 월소득

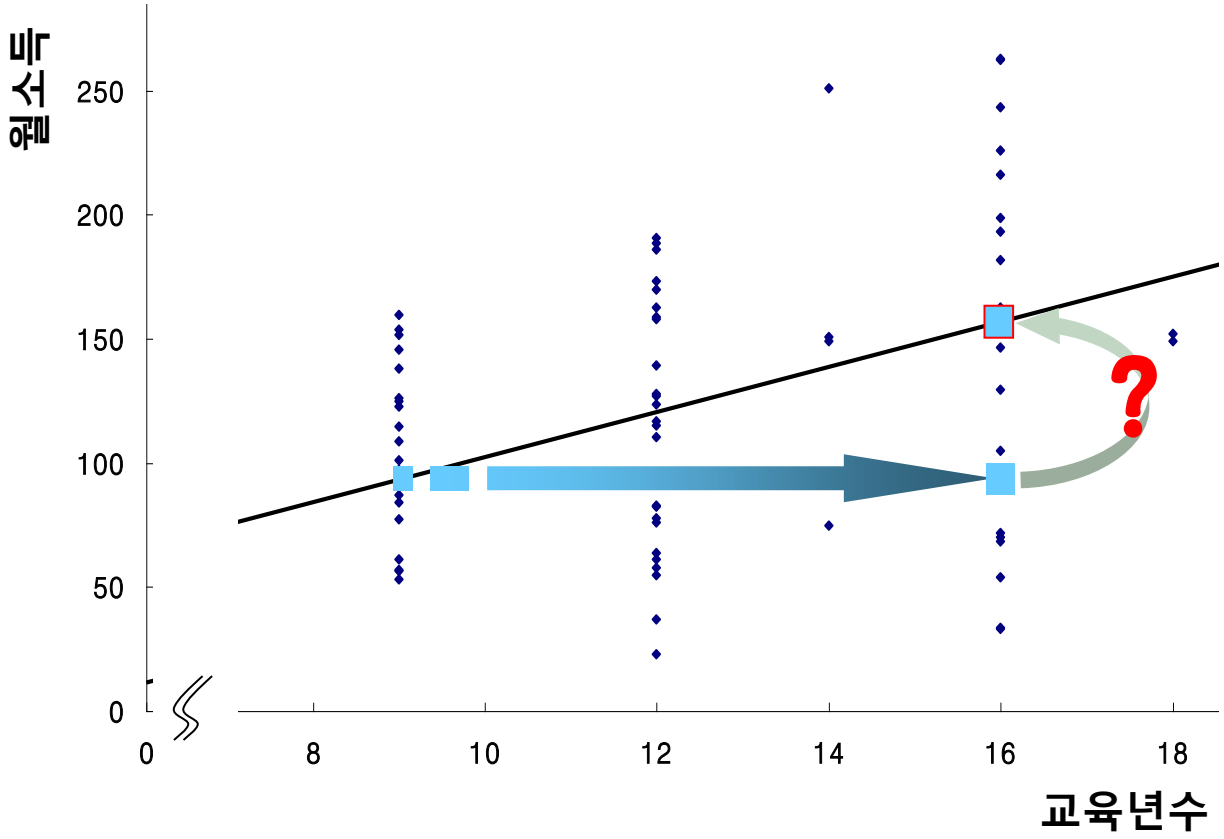
회귀직선 이용, 교육년수가 12년인 고졸여성과 16년인 대졸여성의 소득 추정

- **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. 기울기와 절편

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.

2. 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

2. 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

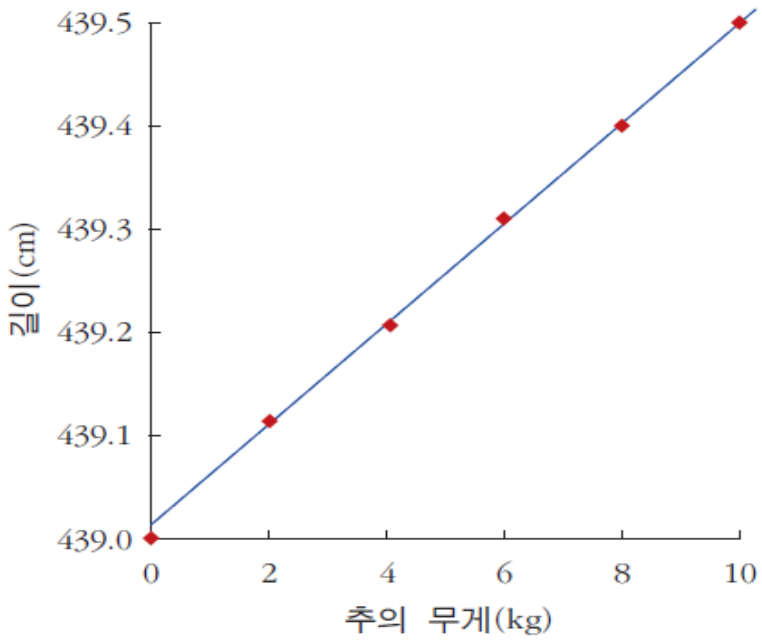
- x: 매단 추의 무게 , y: 용수철의 길이

2. 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

2. 최소자승법

빅맥지수

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

빅맥지수와 환율의 수준이 같지 않다.

Absolute PPP 성립한다고 보기 어렵다.

2. 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

2. 최소자승법

자산가격결정모형(CAPM)

CAPM : 수익률에 대한 단일요인 모형(Single factor model of return)

시계열 회귀분석 방정식

$$r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it}$$

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

2. 최소자승법

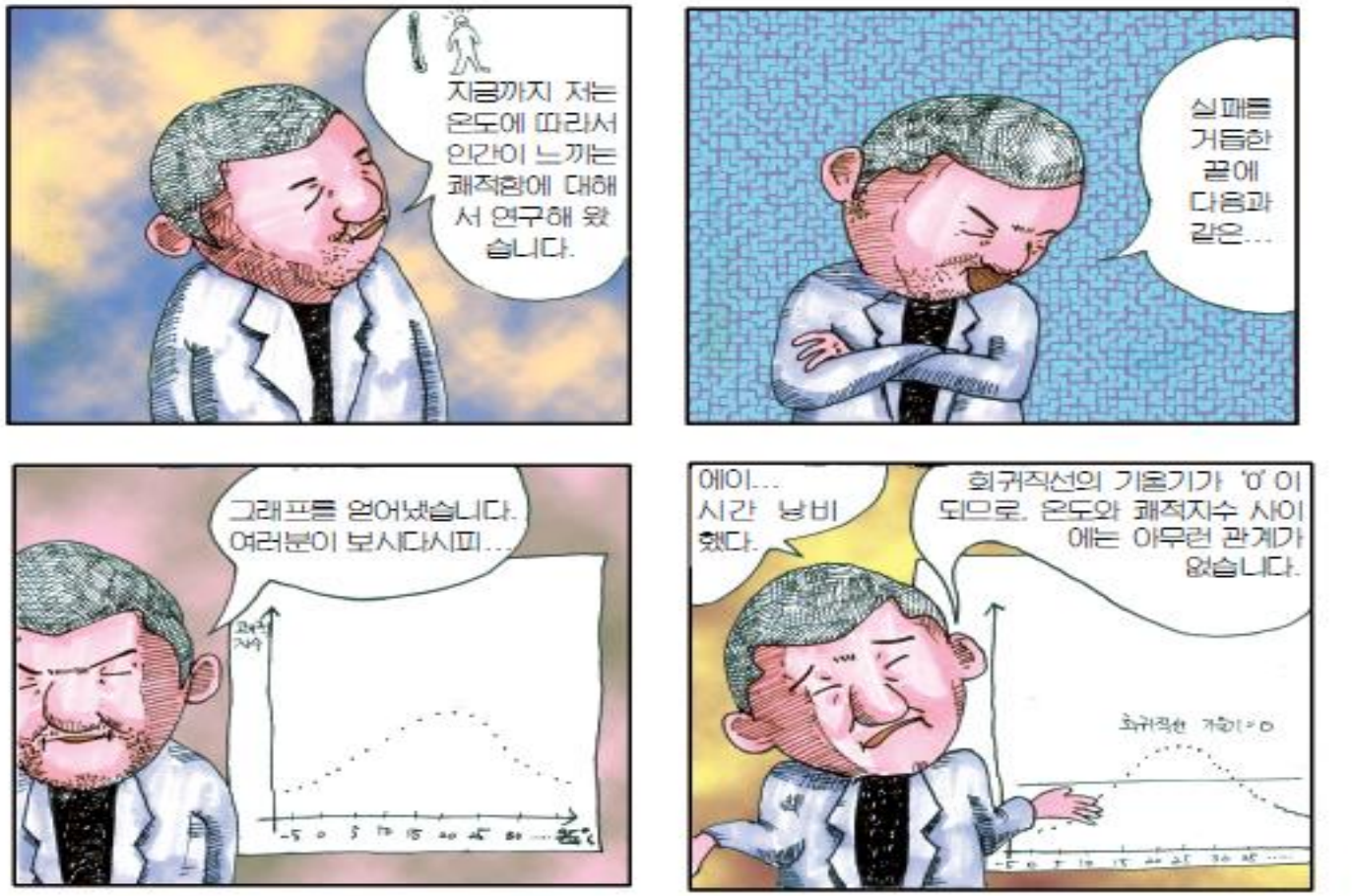
CAPM β 의 추정

- 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

3. 회귀분석은 만병통치약이 아님

회귀직선과 비선형관계



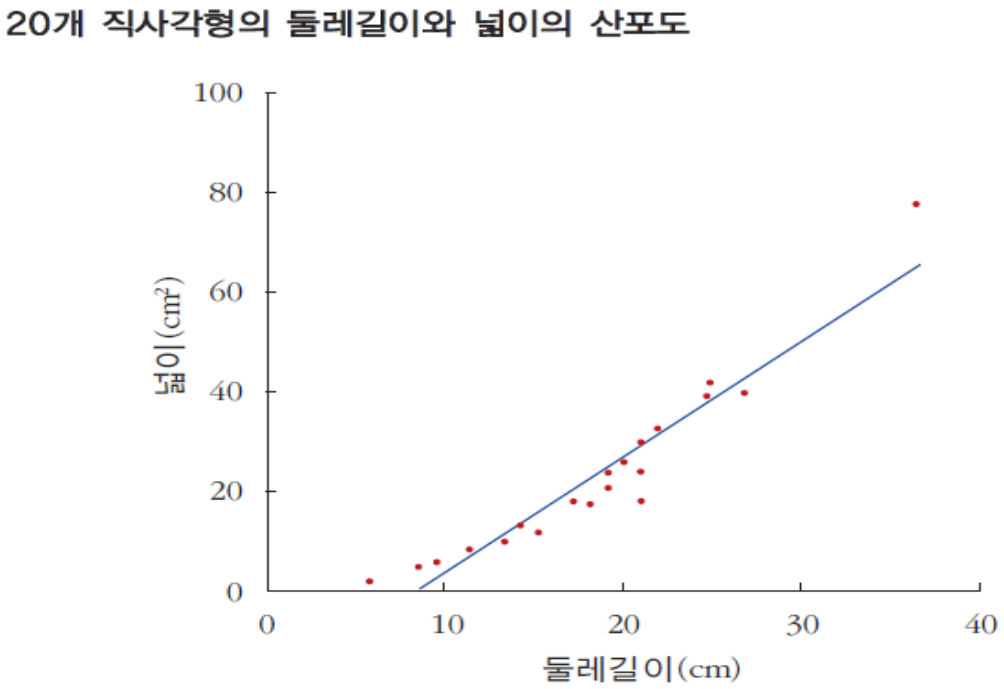
회귀직선은 선형관계만을 측정

3. 회귀분석은 만병통치약이 아님

직사각형의 둘레길이와 넓이의 산포도

넓이와 둘레길이간 상관계수=0.98: 이는 넓이와 높이라는 제3의 요인들 영향

넓이와 둘레길이간 관계는 선형 관계가 아니라 비선형의 관계임



4. 중회귀분석

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

4. 중회귀분석

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

$$y = a + b \times E + c \times S + (\text{오차})$$

4. 중회귀분석

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 질적변수이다. 일반적으로 더미변수는 하나의 질적인 설명변수가 종속변수에 미치는 영향을 파악하고자 할 때 이용한다.

4. 중회귀분석

남녀 노동자간 임금격차

남녀간의 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)

4. 중회귀분석: 기술진보와 물가지수

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

4. 중회귀분석

2000년 서울 강남지역의 아파트 가격

(추정된 아파트 가격)= -20.394+1,549(평수)+0.76(연령)　　단위: 만 원

- (i) 아파트 평수와 (ii) 아파트 연령의 2개 설명변수 갖는 중회귀분석 모형임
- 하지만 여전히 중요한 제3의 변수가 모형에서 누락된 결과 "아파트가 red wine 처럼 오래될수록 비싼" 것으로 잘못 추정됨
- 아파트 단지규모는 아파트 연령과도 관련되어 있고(강남 개발 초기에 지어진 아파트가 대단지로 들어섬) 동시에 아파트 가격과도 밀접하게 연관되어(대단지는 편의시설이 발달되어 있어 아파트 값이 비쌈) 있는 혼동요인(confounding factor)으로 작용하고 있음

4. 중회귀분석

2000년 서울 강남지역의 아파트 가격

(추정된 아파트 가격)=-20.291+1,538(평수)-137(연령)+2(단지규모)

- 위 식은 (i) 아파트 평수, (ii) 아파트 연령 등 기존 2개의 설명변수에 (iii) 아파트 단지규모라는 설명변수를 추가하여 확장한 중회귀분석 모형을 추정한 식임
- 제3의 요인인 아파트 단지규모를 통제한 결과 아파트 연령과 가격간의 관계가 보다 상식에 부합되게 얻어짐→동일한 (평수, 단지규모)에 속하는 아파트를 비교해보면 오래된 아파트가 1년당 평균 137만원 꼴로 값이 낮음
- 물론 오래된 아파트의 재건축 가능성을 고려하면 아파트 연령과 아파트 가격간의 관계는 비선형일 수도 있을 것으로 예상됨

4. 중회귀분석

회귀분석: 홈런 on BB

규정타석을 채운 타자들로만 이루어진 "동질적" 자료 이용

홈런이 많은 타자는 큰 것 '한 방'을 노리기 때문에 삼진도 많이 당함

- Babe Ruth: "I hit big, I miss big. I would like to live as big as I could."

$$(\text{사사구수}) = a + 0.51(\text{홈런수}) + e$$

4. 중회귀분석

회귀분석: 홈런 on BB

모든 타자들로 이루어진 "이질적" 자료 이용 (타석수 통제되지 못함)

홈런이 많은 타자는 큰 것 '한 방'을 노리기 때문에 삼진도 많이 당함

$$(\text{삼진수}) = \alpha + \beta(\text{홈런수}) + \varepsilon$$

$$b = 2.40 (>> 0.51)$$

4. 중회귀분석

중회귀분석: 홈런 on (BB and 타석수)

모든 타자들로 이루어진 “이질적” 자료 이용하되 중회귀분석 이용하여 통계적 방법으로 타석수 통제

홈런이 많은 타자는 큰 것 ‘한 방’을 노리기 때문에 삼진도 많이 당함

$$(\text{삼진수}) = \alpha_1 + \beta_1(\text{홈런수}) + \beta_2(\text{타석수}) + \varepsilon$$

$$b_1 = 0.63 \qquad b_2 = 0.14$$

4. 중회귀분석

결혼시장 분석

- 심리학, 사회학, 인류학 등에서 행하는 설문조사를 통한 남녀 배우자 선호 비교 연구: 정직한 진술 (truth telling)이 보장되지 않는 맹점이 있음
- 실제 선택한 결과를 관측한 현시선호(revealed preference)의 데이터를 연구할 필요가 있음.
- 국내 모 결혼정보회사의 상세한 개인 프로필 및 선택에 대한 현시 선호 데이터를 사용하여, 우리 나라 중매결혼시장에서 남녀의 배우자 선호의 차이를 비교함. **사회 경제적인 조건과 외모 조건에 대한 선호에 있어서 남녀의 차이가 어떻게 드러나는가?**

4. 중회귀분석

결혼시장 분석:사회경제적 조건 및 신체적 조건들

	남			여		
특성	중간값	평균	표준편차	중간값	평균	표준편차
나이	32.2	33.4	3.4	29.6	30.3	2.7
신장(cm)	173	173.7	4.3	162	162.7	3.9
체중(kg)	68	69.3	6.8	50	50.2	4.1
인상등급(0~5)	3	3	1.2	3	3	1.1
연봉(원)	3500	4833	2575	2200	3504	1860

학력	남	여
대학원 이상(%)	22	28
대졸 이상(%)	89	83
고졸(%)	11	17

결혼 적령기의 전체 인구집단과 비교할 때, 결혼정보회사의 회원들은 나이는 다소 많고, 신장은 다소 크고 연봉은 많으며, 학력이 높음.

4. 중회귀분석

결혼시장 분석:'결혼정보회사'의 배우자지수

- 사회경제적 위세 지수 (SESI: Socio Economic Status Index)

  : 학력, 학벌, 직업, 소득 등을 포괄하는 지수 (회사 측에서 작성).

- 신체적 매력 지수(PAI: Physical Attractiveness Index)

  : 키, 체중, 인상등급 등을 포괄하는 지수 (회사 측에서 작성).

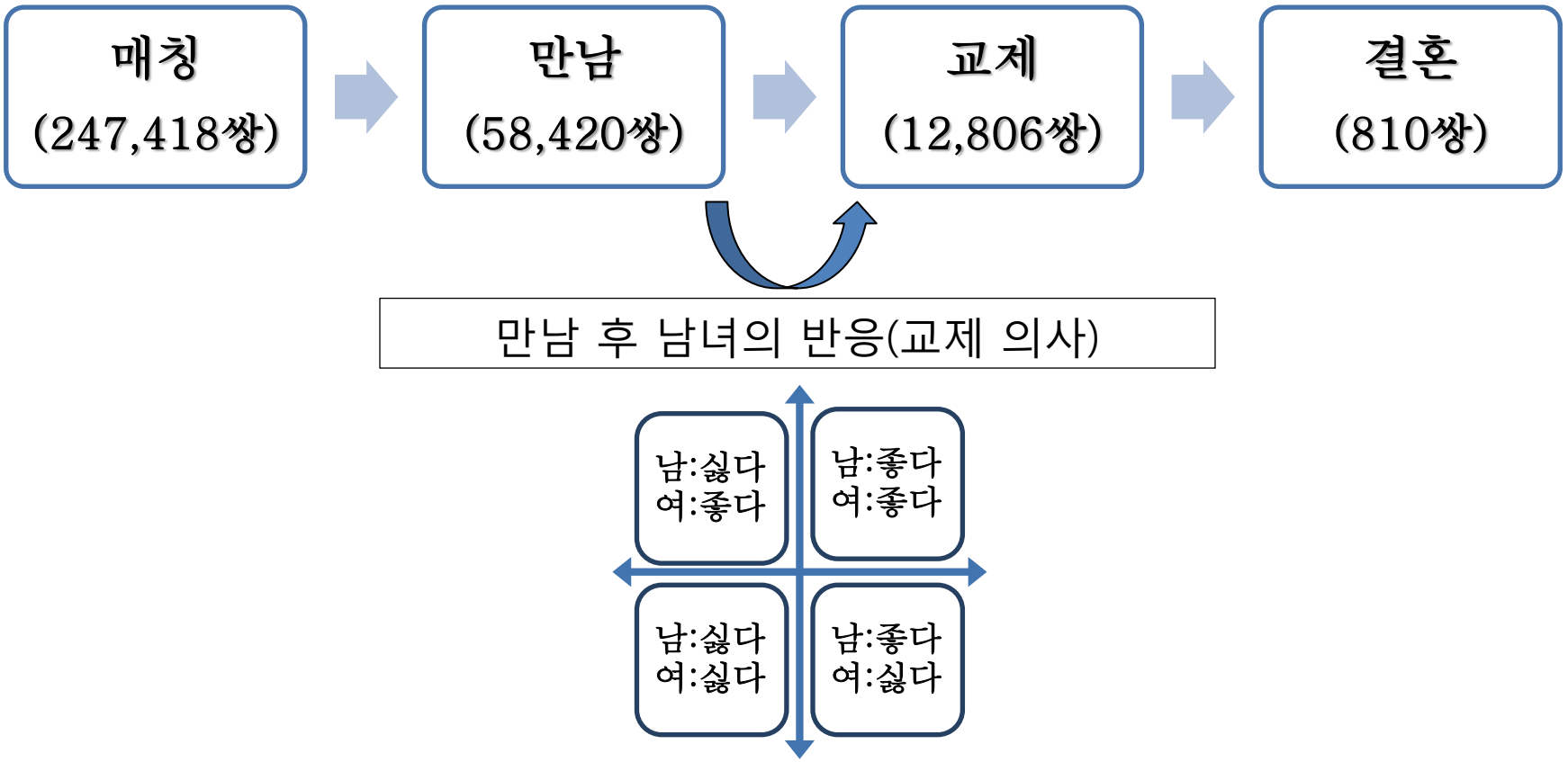
- 가정환경 지수 (FBI: Family Background Index)

  : 부의 학력, 부의 직업, 부의 재산, 양친 생존여부, 부모 이혼여부, 형제관계 등을 포괄하는 지수로 회사에서 작성.

지수	남					여				
	최저값	중간값	최고값	평균	표준편차	최저값	중간값	최고값	평균	표준편차
SESI	34.1	73	100	72	11	33.2	71	99.7	70	13
PAI	21	80	100	76	12	14	82	100	79	10
FBI	7.6	54	99.3	54	19	7.6	62	98.1	61	28

4. 중회귀분석

결혼시장 분석:결혼에 이르는 단계와 만남 후 남녀의 반응



4. 중회귀분석

결혼시장 분석:분석모형 및 추정 결과

- (반응) = α + β_1(상대의SESI) + β_2(상대의PAI) + β_3 (상대의FBI) + ε
- 반응: 좋다=1, 싫다=0

	남자의 반응		여자의 반응	
	추정치	표준오차	추정치	표준오차
SESI	0.0123	0.0130	0.0191*	0.0014
PAI	0.0319*	0.0034	0.0118*	0.0027
FBI	0.0139*	0.0053	0.0029	0.0087

4. 중회귀분석

결혼시장 분석:해석과 함의

사회경제적 조건과 외모의 교환 관계(Trade off between SESI & PAI)

- 남자가 여자를 평가할 때는 사회경제적 조건(SESI)에 비해 외모(PAI) 중시
- 여자가 남자를 평가할 때는 외모(PAI)에 비해 사회경제적 조건(SESI) 중시

5. 총변동의 분해

통화증가율과 인플레이션율

표 8-2 우리나라의 연간 통화증가율과 인플레이션율, 1986-2008년도 (단위: %)

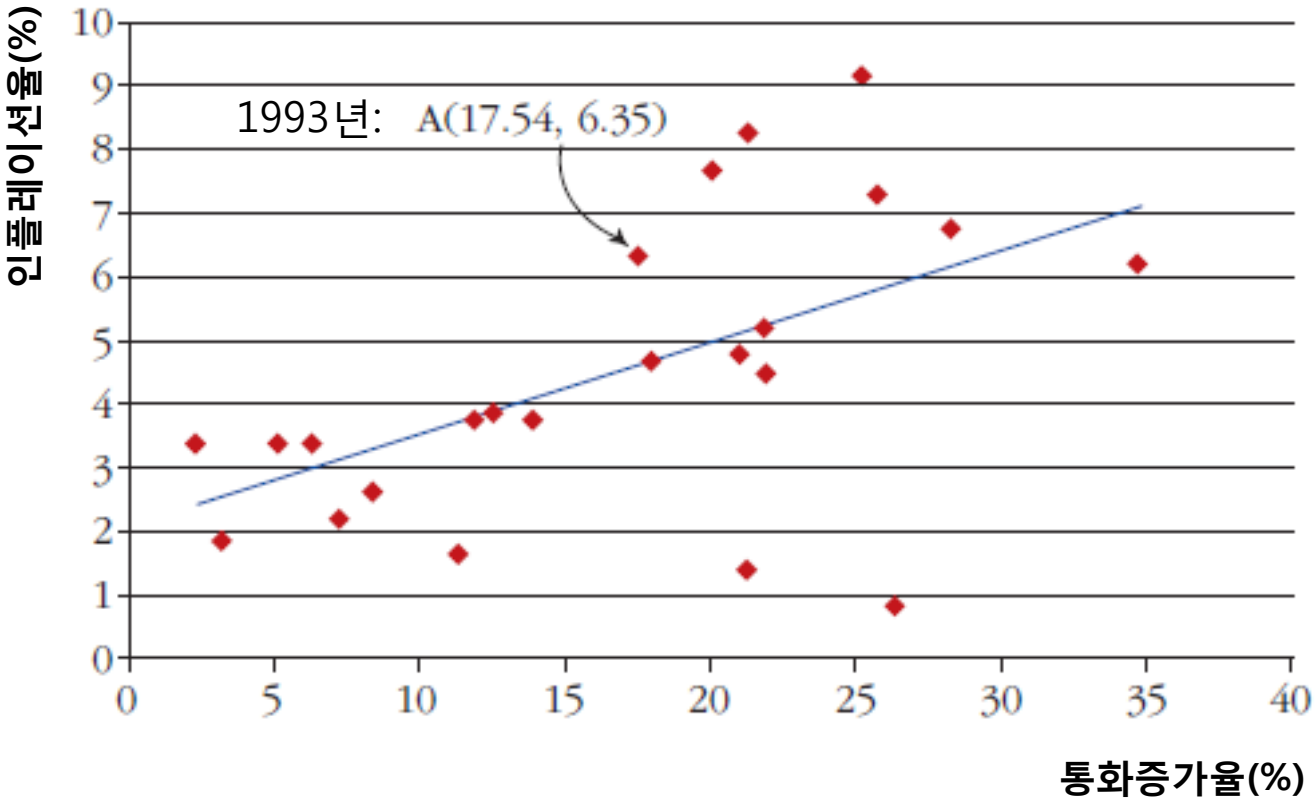
연도	통화증가율	인플레이션율	연도	통화증가율	인플레이션율	연도	통화증가율	인플레이션율
1986	26.4	0.88	1994	21.85	5.2	2002	13.93	3.81
1987	34.79	6.22	1995	21.1	4.81	2003	2.36	3.39
1988	28.3	6.78	1996	18.02	4.68	2004	6.33	3.4
1989	25.78	7.32	1997	21.38	8.29	2005	7.27	2.22
1990	25.31	9.16	1998	21.27	1.46	2006	11.3	1.68
1991	20.13	7.72	1999	3.19	1.89	2007	12.47	3.89
1992	21.95	4.51	2000	5.16	3.44	2008	11.96	3.75
1993	17.54	6.35	2001	8.53	2.64			

주: 통화증가율은 M2증가율이고 인플레이션율은 소비자물가지수 상승률임.

- 1986~2008년간의 평균 인플레이션율은 4.50%였는데 반해 1993년의 인플레이션율은 6.35%로 표본 내 23년간의 평균보다 1.85% 높음
- 이중 얼마만큼의 차이를 통화증가율의 차이로 설명할 수 있을까?

5. 총변동의 분해

통화증가율과 인플레이션율



5. 총변동의 분해

통화증가율과 인플레이션율

- 1993년: (x_i, y_i)=(17.54%, 6.35%)

$$y_i - \bar{y} = [(a+bx_i) - \bar{y}] + [y_i - (a+bx_i)]$$
$$T = R + E$$

- T=1993년의 인플레이션율(y_i)이 지난 23년의 평균과 차이나는 부분 전체
- R=그 전체 가운데 1993년의 통화증가율(x_i)로 설명되는 부분
- E=그 전체 가운데 1993년의 통화증가율로는 설명되지 않는 부분

5. 총변동의 분해

총변동의 분해

$$\sum (y_i - \bar{y})^2 = \sum [(a + b\,x_i) - \bar{y}]^2 + \sum [y_i - (a + b\,x_i)]^2$$

$$\boldsymbol{SST} \quad = \quad \boldsymbol{SSR} \quad + \quad \boldsymbol{SSE}$$

(회귀식으로 설명됨) (회귀식으로 설명 안 됨)

- **SST**[총제곱합 (total sum of squares)]: y의 평균 주위로의 총변동
- **SSR**[회귀제곱합 (regression sum of squares)]: 회귀직선에 의해 설명되는 변동분
- **SSE[**잔차제곱합 (residual sum of squares) 또는 오차제곱합 (error sum of squares)]: 회귀직선에 의해 설명되지 않는 변동분

5. 총변동의 분해

결정계수(R^2)

결정계수(R^2) = 총변동에서 차지하는 설명되는 변동분의 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (0 \le R^2 \le 1)$$

결정계수의 값이 1에 가까울수록 회귀직선의 설명력은 높다.

단순회귀분석의 경우 결정계수인 R^2 값은 두 변수간 상관계수인 r 의 제곱과 같게 된다. (단순회귀분석의 경우에는 R^2 = r 제곱)

5. 총변동의 분해

조정된 결정계수(adjusted R^2)

설명변수를 추가하면 추가할수록 R^2는 언제나 증가함

- R^2=1-SSE/SST 인데 SST는 고정된 반면 SSE는 설명변수 추가될수록 감소

이 문제를 해결하기 위해 아래의 "조정된 결정계수"를 정의함

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \quad (n = \text{표본크기},\ k = \text{설명변수의 개수})$$

- SSE와 SST가 각각의 자유도로 나누어진 형태로 등장
- SST의 자유도=(n-1): 표준편차 구할 때의 자유도와 동일
- SSE의 자유도=(n-k-1): n개 자료 이용 총 (k+1)개의 계수 추정한 결과
- 조정된 결정계수는 설명변수가 추가된다고 해서 반드시 늘지는 않음