통계학 강의노트

상관관계 vs. 인과관계

상관관계가 있지만

어느 것이 원인이고 어느 것이 결과인지 명백하지 않을 때가 있다.

원인과 결과가 시간에 따라 뒤바뀌기도 하고,

양쪽이 동시에 원인이면서 결과일 수도 있는 것이다.

상관관계 vs. 인과관계

상관관계

❖ 상관관계는 어떤 변수가 증가할 때 다른 변수가 함께 증가 또는 감소하는지를 관찰해 파악

Ex) 체중과 신장 : 양의 상관관계 존재

→ 키가 크면 대체적으로 체중이 증가

❖ 상관계수:

➢ 상관계수의 범위 : −1 ~ 1

✓ 상관계수가 음수이면 음의 상관을, 양수이면 양의 상관을 갖음

➢ 상관계수가 0 일 경우 : 서로 관계가 전혀 없음을 의미

✓ 상관계수는 선형(linear) 상관만을 측정하므로 상관계수가 0이라는 말은 선형관계가 존재하지 않는다는 의미임

상관관계 vs. 인과관계

상관관계의 예시

❖ 광고와 매출액

➢ 광고와 매출액은 상호작용을 하여 원인도 되고 결과도 됨

→ 광고를 늘리면 상품 매출액이 증가해 광고비를 더 지출할 수 있는 여유가 생겨 광고를 더 하게 됨

☞ 초기에는 광고가 매출액 증가의 원인일 수 있지만 나중에는 매출액 증가가 광고 증가의 원인이 됨

❖개인소득과 보유 주식

➢ 개인소득과 보유 주식은 서로 원인과 결과가 상호작용하는 관계임

→ 소득이 많을수록 주식을 많이 사게 되지만 주식을 많이 사면 다시 배당 등으로 소득이 증가

상관관계 vs. 인과관계

두 변수 사이의 상관관계가 존재하더라도 원인이 다른 곳에 있는 경우

❖ 교회수가 늘면 범죄 발생률이 증가한다??(교회가 범죄 증가의 원인???)

→ 진짜 원인: 인구증가

☞ 인구가 늘면 교회도 많아지고 범죄도 증가하는 것임

❖ 프랑스 도시 스타라스부르그에서는 황새의 둥지수와 출생률 사이의 상관관계가 높은 것으로 나타났다.

→ 진짜 원인: 인구증가

☞ 인구가 증가하면 출생률도 높아지고, 또 주택이 증가하므로 황새가 둥지 틀 곳이 많아지는 것임

❖ 미국 매사추세츠 주의 장로교 목사 월급과 쿠바 하바나의 럼(rum)주 가격 사이의 높은 상관관계가 있다.(목사들이 술 무역을 통해서 돈을 번다???)

→ 진짜 원인: 시간 흐름에 따른 물가상승

☞ 세월의 흐름에 따라 거의 모든 물가와 월급이 올라가게 마련임

상관관계 vs. 인과관계

잘못된 인과관계의 추정

남태평양 뉴 헤브리디스 섬 주민들은 몸의 이(벌레)가 건강의 원인이라고 믿고, 건강하려면 이를 몸에 많이 지녀야 한다고 생각했다. 즉, 건강한 사람에게는 이가 있지만 환자에게는 이가 없는 경우가 많다는 원주민의 과거 수세기에 걸친 경험과 관찰을 토대로 이런 결론을 내린 것이다.

- 실제로는 이 섬에는 이가 득실거려 대부분의 사람들 몸에 이가 있었는데, 이가 옮기는 열병에 걸리게 되면 체온이 올라가서 이가 살기 어려운 조건이 되므로 환자의 몸에서 이가 달아나는 것이었음

→ 건강하면 이가 꼬이고, 이는 열병을 옮기고, 열병은 이를 쫓아내고, 이가 없어지면 열병이 낫고, 건강해지면 다시 이가 꼬이는 순환을 반복

☞ 원인과 결과가 뒤죽박죽으로 뒤엉킴

상관관계 vs. 인과관계

인과관계가 성립할 수 있는 조건(by John S Mill)

❖ 두 개의 변수들은 상관관계를 갖지만 그것은 우연일 뿐 서로 인과관계가 없는 경우가 많음

❖ 인과관계가 있더라도 다른 변수들이 그 사이에 존재할 수도 있음

❖ 인과관계가 성립할 수 있는 3대 조건

첫째, 원인은 결과보다 시간적으로 앞서야 한다

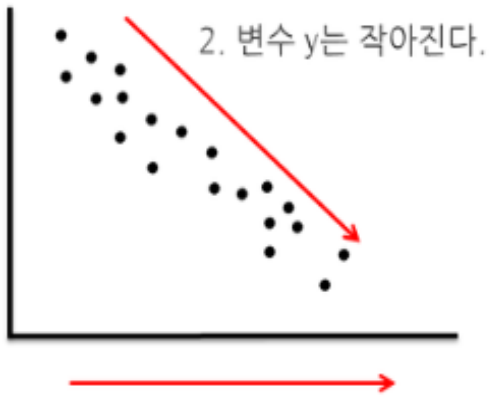
둘째, 원인과 결과는 서로 관련 있어야 한다

셋째, 결과는 원인이 되는 변수만으로 설명되어야 하고,
다른 변수에 의한 설명은 제거되어야 한다.

상관분석

상관분석

❖ 상관분석: 두 변수가 어떠한 관계에 있는지를 파악하는 분석

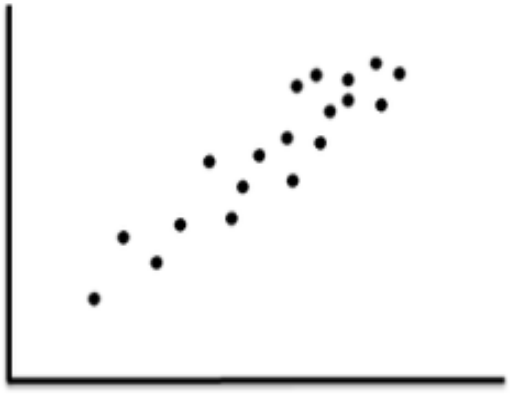


상관분석

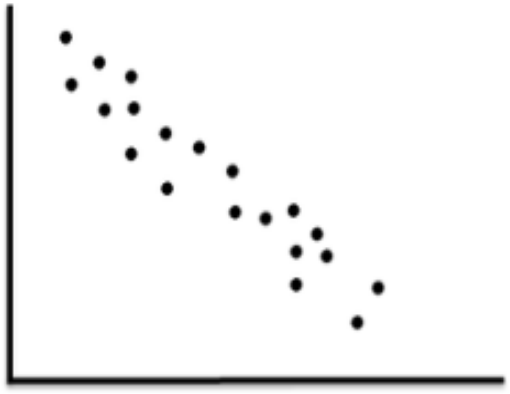
상관분석

❖ 양(+)의 상관관계: x가 증가할수록 y도 증가

❖ 음(−)의 상관관계: x가 증가할수록 y는 감소



양(+)의 상관관계



음(−)의 상관관계

상관계수

상관계수의 필요성

➢ 두 변수 사이의 관계가 어느 정도 강한가?

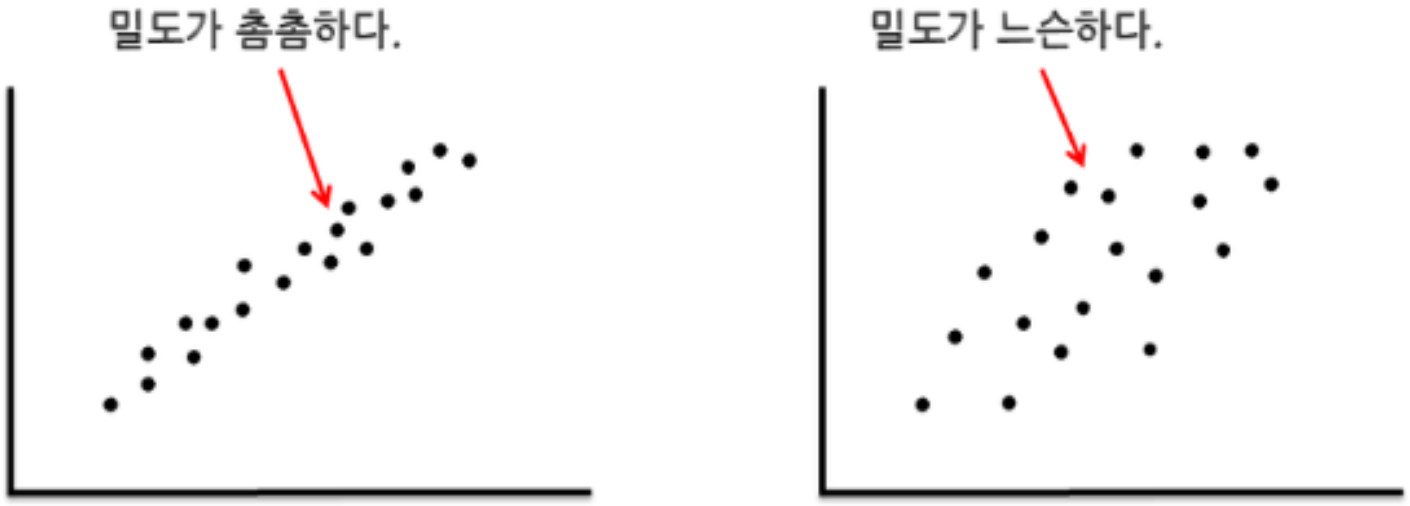
평균과 표준편차만으로는 알 수 없다.

'상관계수'라는 새로운 통계량 필요

상관계수
(correlation coefficient)

상관계수

상관계수의 개념



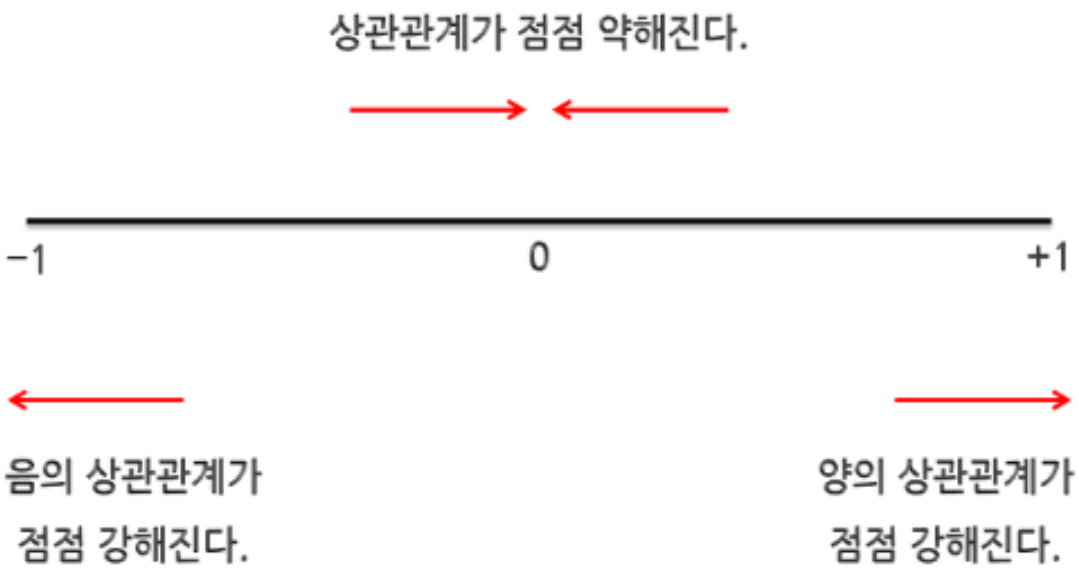
- 둘 다 양(+)의 상관관계이지만 같다고 하기에는 "밀도"의 차이가 존재

- 통계에서는 숫자를 사용해서 밀도를 표현하는데 이 밀도를 표현한 숫자를 상관관계라고 부름(***r***로 표시)

상관계수

상관계수의 범위

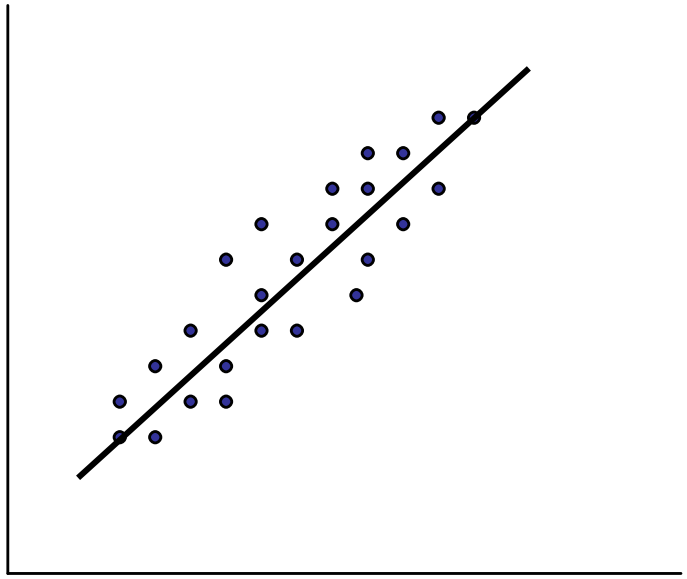
- **$-1 \leq r \leq 1$**
  - 상관계수 = 1 또는 –1 → 완전상관(perfect correlation), 모든 점들 정확히 하나의 선 위에 위치
  - 양의 상관관계 → 점의 분포가 우상향
  - 음의 상관관계 → 점의 분포가 우하향



상관계수

산포도와 상관계수

(a)



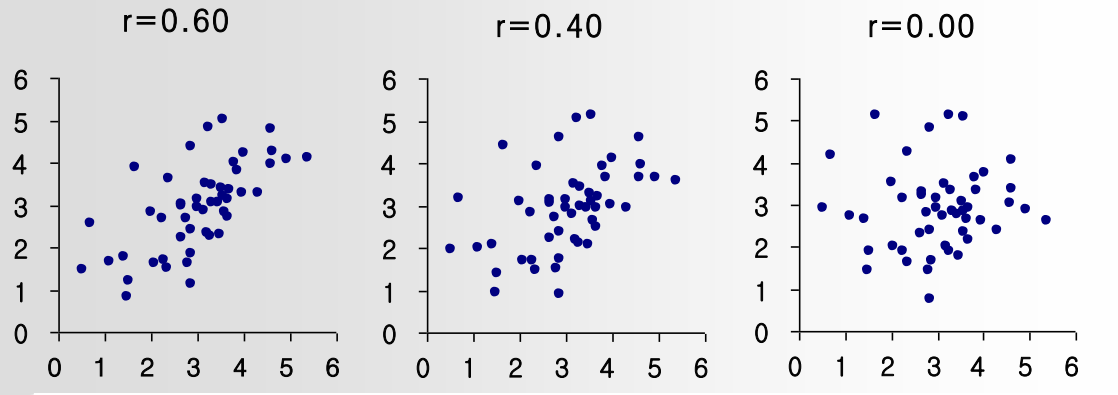
(b)

1에 가까운 상관계수는 점들이
선 주위에 **몰려** 있음을 의미

0에 가까운 상관계수는 점들이
선 주위에 **퍼져** 있음을 의미

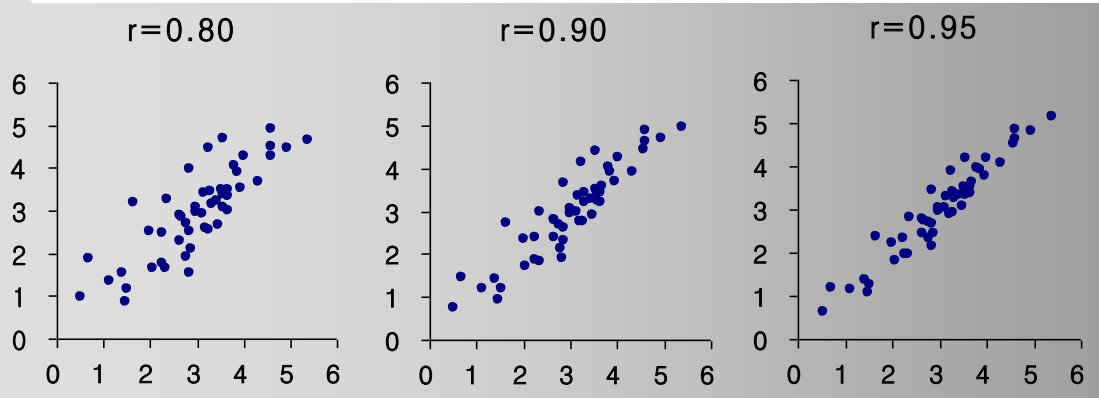
상관계수

상관계수 변화에 따른 산포도 모양의 변화 (양의 상관관계)



평균 : 3

표준편차 : 1



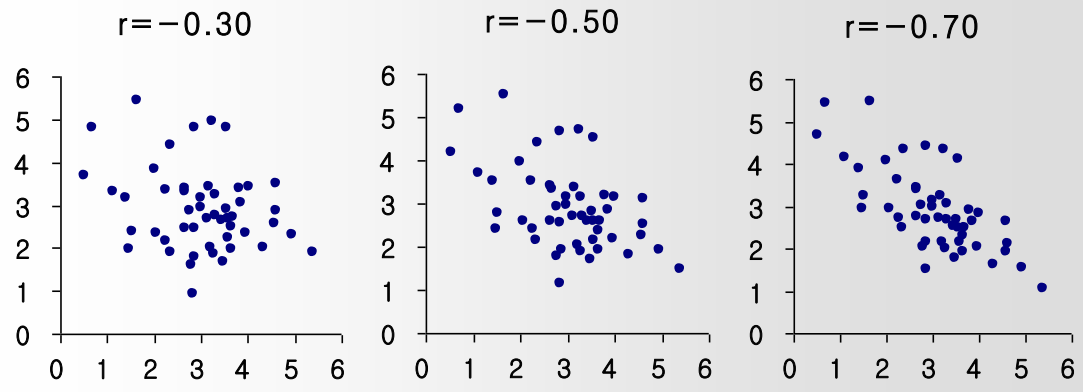
선형 관계 점차 뚜렷

상관계수

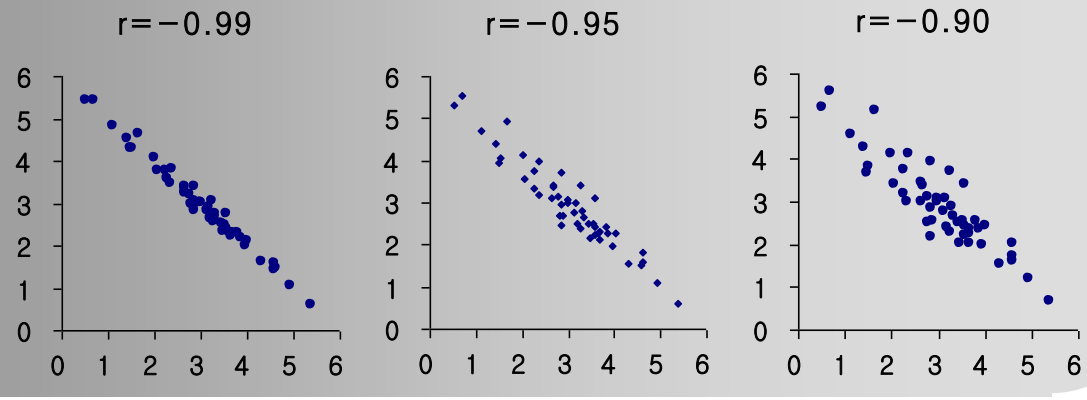
상관계수 변화에 따른 산포도 모양의 변화 (음의 상관관계)

평균 : 3

표준편차 : 1

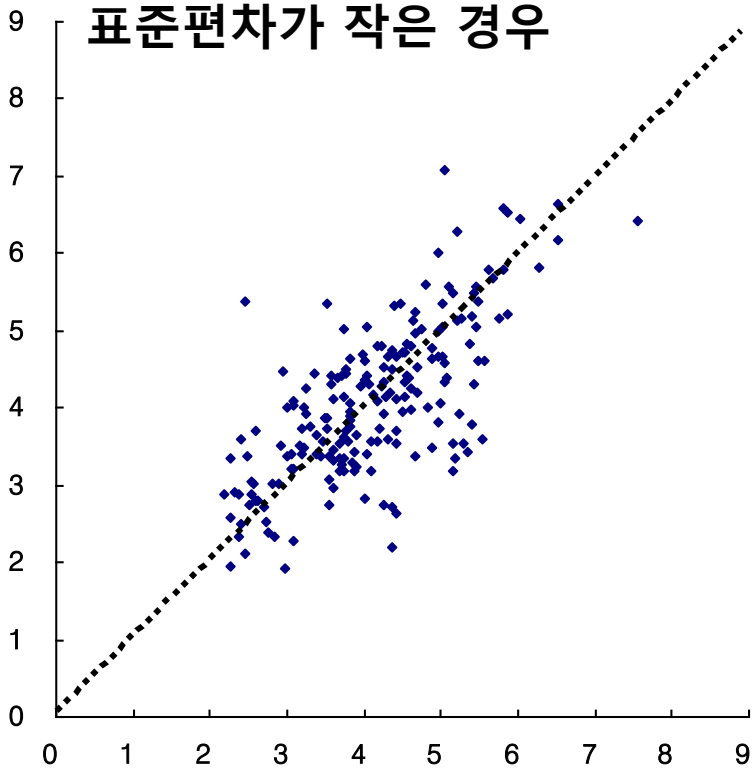


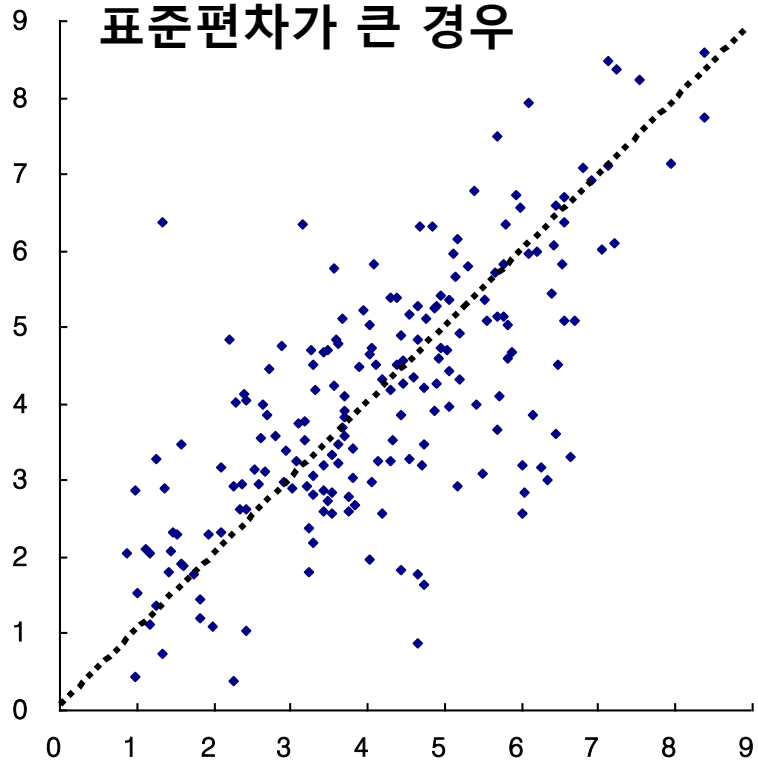
선형 관계 점차 뚜렷



상관계수의 특징 및 유의사항

표준편차를 변화시킬 때의 시각적 효과

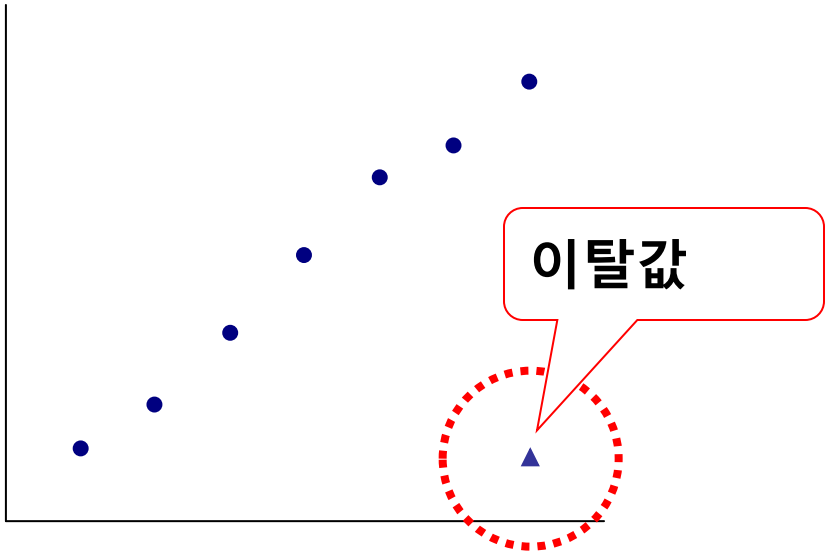


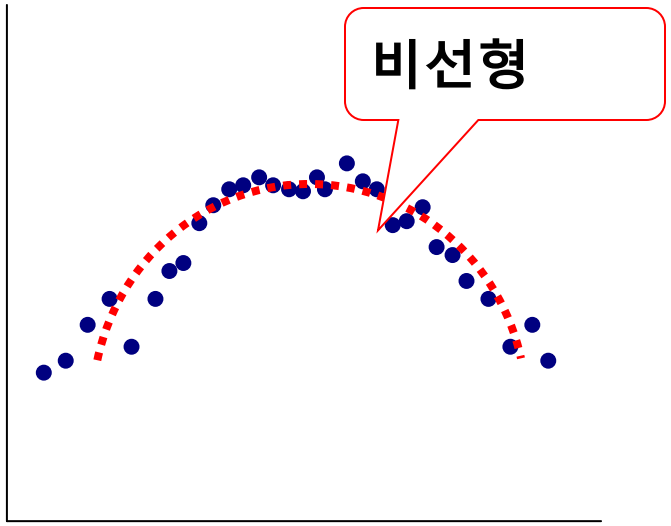


상관계수가 같아도 표준편차가 작으면
산포도가 더 빽빽하게 밀집된 것처럼 보인다.

상관계수의 특징 및 유의사항

상관계수가 유용하지 않은 경우

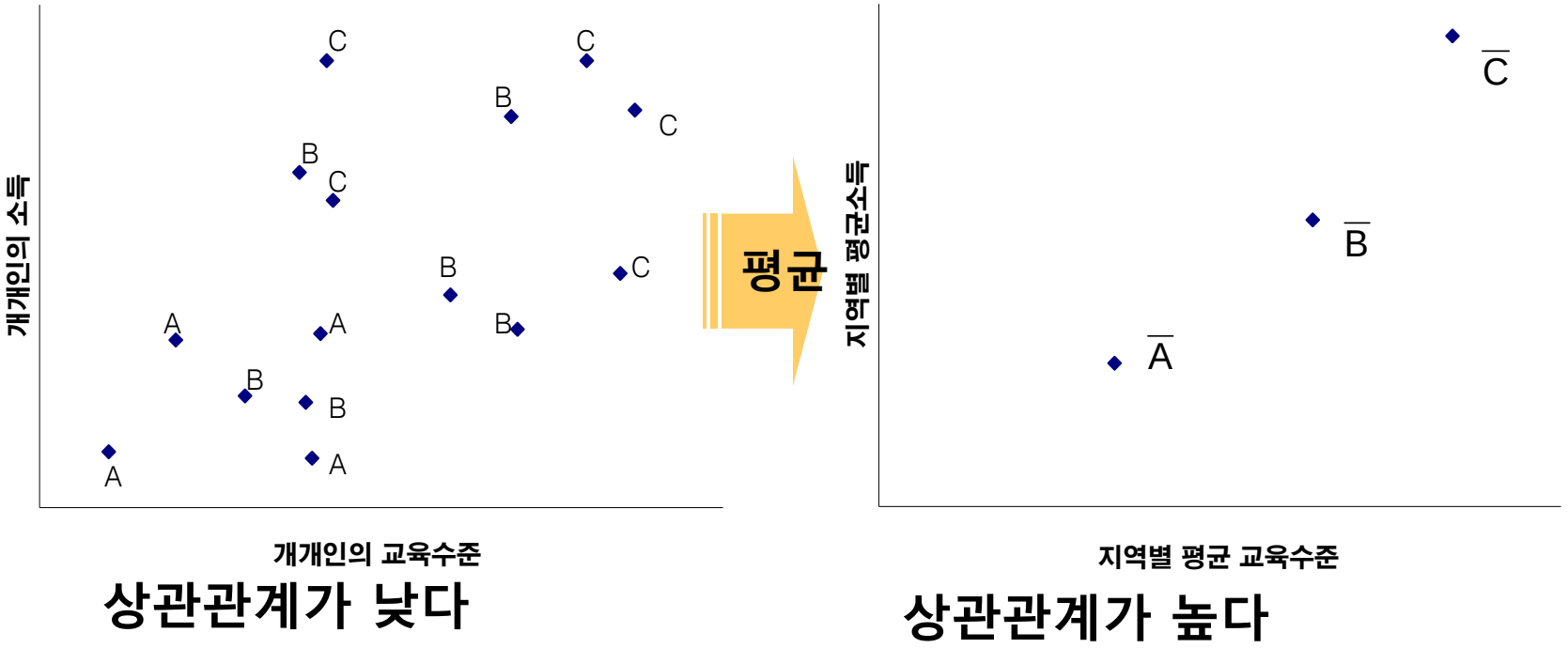




상관계수는 이탈값이 존재하거나 분포가 비선형일 때 유용성이 떨어진다.

상관계수의 특징 및 유의사항

상관계수가 실제의 관계를 과장하는 경우



상관관계가 낮다

상관관계가 높다

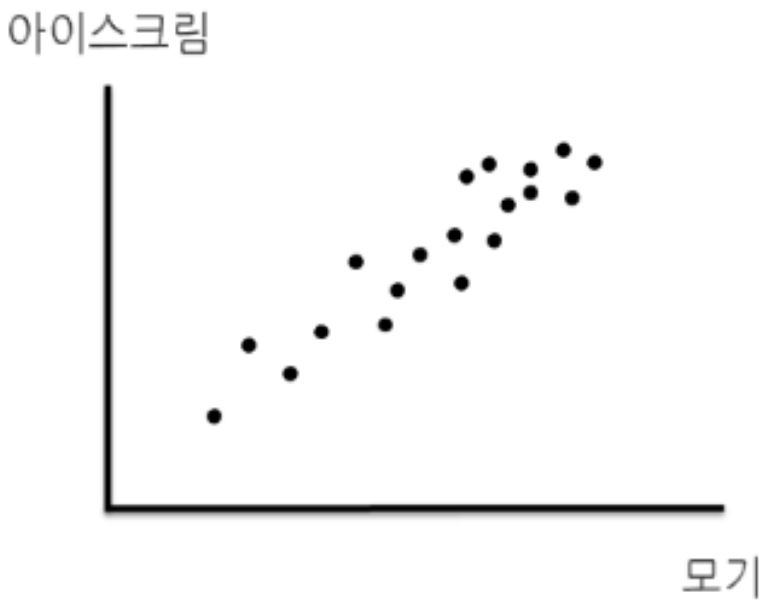
집단별 비율이나 평균에 기초하여 구한 상관계수는 실제의 관계를 과장한다.

*참고 : 생태학적 오류: 집단 -> 개인 / 개체주의적 오류: 개인 -> 집단

상관분석

상관계수가 곧바로 인과관계를 의미하지는 않는다.

❖ 상관분석은 두 변수가 어떠한 관계에 있는지는 파악할 수 있지만, 서로의 인과관계는 파악하기 어려움



상관분석과 회귀분석

➢ 상관분석(correlation analysis)

- 두 변수 사이의 선형관계의 강도와 방향을 요약하는 수치를 구하는데 사용됨.
- 상관분석은 두 변수가 얼마나 밀접하게 연관되어 있는가 하는 정도를 나타낸다.

➢ 회귀분석(regression analysis)

- 두 변수(종속변수, 독립변수) 사이의 함수적 관계를 기술하는 수학적 방정식을 구하는데 사용됨
- 이 식은 독립변수의 값이 주어질 때 종속변수의 값을 추정하거나 예측하는데 사용된다.

✓ 상관분석은 두 변수 사이의 인과관계를 밝히는 것이 아니라, 두 변수가 서로 어느 정도 관련되어 있는가를 나타냄

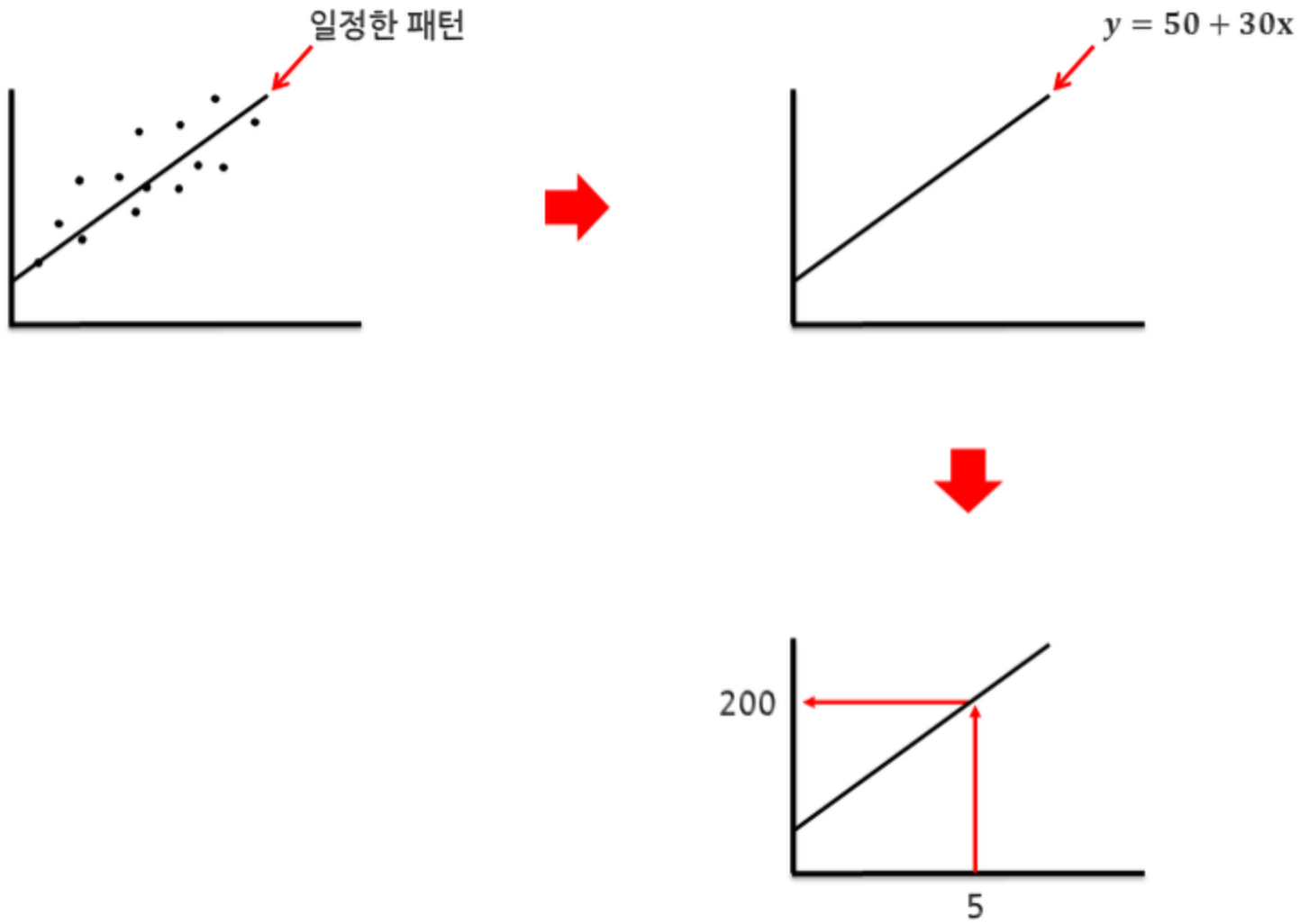
회귀분석

회귀분석의 개념 및 특징

- 회귀분석: 독립변인이 종속변인에 영향을 미치는지 알아보고자 할 때 실시하는 분석방법
- 두 변수의 관계에 분명한 방향(direction)이 있을 때
  ex) 인구증가 – 범죄율 증가
- 반응변수와 설명변수(종속변수와 독립변수)
  - 변수의 역할을 구분
- 두 변수의 직선적 관련성을 구체적인 수식으로 표현
- 예측(prediction)을 위한 목적에도 사용

회귀분석

회귀분석의 개념 및 특징



회귀분석

회귀분석의 종류

- 단순회귀분석: 영향을 주는 변수가 1개
  ex) 광고 -> 판매량

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

- 다중회귀분석: 영향을 주는 변수가 2개 이상
  ex) 광고, 교육훈련, 근로자의 사기, 기업 이미지 등 -> 판매량

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

* β_0 : y절편, β_1 : 기울기, ε : 오차항

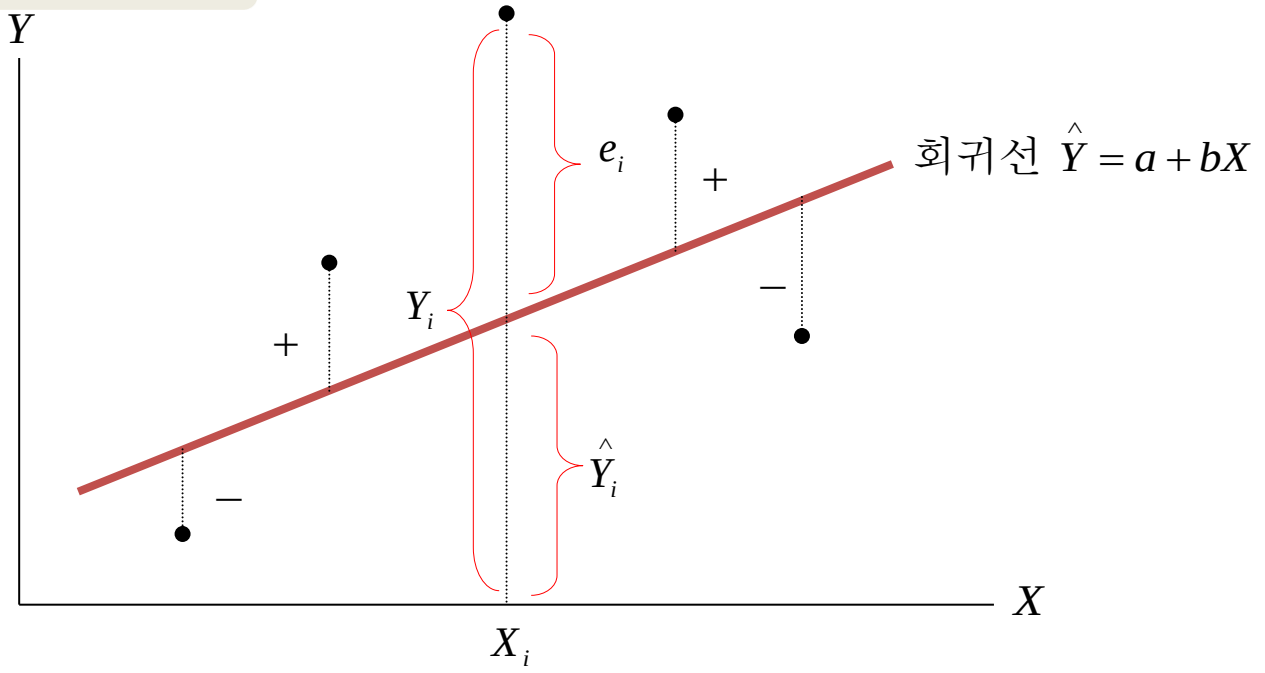
회귀분석

최소제곱법(Least Squares Method)

- 회귀분석에서 연구자의 관심은 모든 관측치들로부터의 오차를 최소화시키는 회귀선을 찾아내 독립변수의 변화에 따라 종속변수가 어떻게 변하는지를 예측하는 것

- 모든 표본에 대하여 오차의 제곱 합이 최소화되는 계수를 추정하는 방법

- 회귀식은 각각의 경우에 대한 거리가 가까울수록 오차가 줄어들며 오차를 최소화하는 직선식이 표본자료를 가장 잘 설명



$$최소 \sum e_i^2 = 최소 \sum (Y_i - \hat{Y}_i)^2 = 최소 \sum (Y_i - a - bX_i)^2$$

$$\therefore \quad \sum Y_i = na + b \sum X_i$$

$$\sum X_i Y_i = a \sum X_i + b \sum X_i^2$$

회귀분석

유의성검정(significance test)

- 각 독립변수와 종속변수의 관련도가 유의한지 또는 종속변수에 대한 설명력을 가지고 있는가를 밝히는 것임
- 유의확률(p값)이 유의수준(0.1/0.05/0.01)보다 작으면 통계적으로 유의미한 결과

결정계수(coefficient of determination)

- 표본회귀선이 표본자료를 얼마나 잘 설명하는가를 평가하는 기준의 하나가 결정계수 (R^2)
  즉 결정계수는 종속변수의 변화가 독립변수에 의하여 어느 정도 설명되는가를 나타냄
- 결정계수의 범위는 0과 1사이

  ex) 결정계수 1: 종속변수의 변화는 독립변수에 의하여 100% 설명됨

  결정계수 0.25: 종속변수의 변화는 독립변수에 의하여 25% 설명됨

  나머지 75%는 다른 독립변수들과 오차에 의하여 영향을 받게 됨

단순회귀분석 예시

- **신생아 배둘레와 몸무게 증가에 대한 회귀분석**

신생아	1	2	3	4	5	6	7	8
배둘레	35.0	32.0	30.0	31.5	32.7	30.0	36.0	30.5
몸무게	3.45	3.20	3.00	3.20	3.30	3.20	3.85	3.15

신생아	9	10	11	12	13	14	15
배둘레	34.7	30.5	33.0	35.0	31.8	38.0	33.0
몸무게	3.65	3.40	3.50	4.00	3.10	4.20	3.45

단순회귀분석 예시

몸무게 = -0.85+0.13 *배둘레

↓

Regression coefficients, 회귀계수

-0.85 : Intercept, Y 축의절편

→ -0.85 kg 관심의 대상이 아니다

0.13 : Slope, 기울기

배둘레가 1 증가할 때 몸무게의 변화량

→ 배둘레가 1 cm 증가하면, 몸무게가 0.13 kg 증가함

결정계수가 0.8045라면,

→ 몸무게의 차이를 배둘레로 80.45% 만큼 설명할 수 있다.

다중회귀분석 예시

➢ **SNS 활용이 당선경쟁력에 미치는 영향(국가정책연구 제26권 제4호: p165-191)**

- 연구목적: 후보자들의 SNS의 활용이 당선경쟁력에 얼마나 영향을 미치는지 분석
- 연구방법: 다중회귀분석

<표 1> 변수의 조작적 정의

구분	변수명		내용
종속변수	당선경쟁력		당선경쟁력 지수(CI)
독립변수	SNS 요인	SNS 이용	후보자들이 이용하는 SNS의 수(0~5)[6]
		SNS 관계형성	후보자들 트위터의 팔로워 수 자연로그
통제변수[7]	공약 요인	경제	경제와 관련한 공약이 있으면 1, 없으면 0
		복지	복지와 관련한 공약이 있으면 1, 없으면 0
		교육	교육과 관련한 공약이 있으면 1, 없으면 0
		문화	문화와 관련한 공약이 있으면 1, 없으면 0
		교통	교통과 관련한 공약이 있으면 1, 없으면 0
		환경	환경과 관련한 공약이 있으면 1, 없으면 0
		안전	안전과 관련한 공약이 있으면 1, 없으면 0
	후보자 요인	정당	여당이면 1, 야당이면 0
		성별	남자면 1, 여자면 0
		연령	연령(만)
		경력	의정횟수
		직업	국회의원0, 정당인1, 변호사2, 교수3, 기타4

6) 트위터, 페이스북, 블로그, 싸이월드, 미투데이

다중회귀분석 예시

<표 5> 회귀분석결과

변수			Coeff.	Std. Err.	t	표준화계수
SNS 요인	SNS 이용 수		−0.02656**	0.00488	−2.75	−0.30301
	트위터 팔로워 수		0.00953*	0.00964	1.95	0.22165
공약 요인	경제		−0.00949	0.01579	−0.60	−0.06154
	복지		−0.04396	0.03178	−1.38	−0.15969
	교육		0.01585	0.02188	0.72	0.07861
	문화		0.02621*	0.01571	1.67	0.16978
	교통		0.00121	0.01419	0.09	0.00901
	환경		−0.00424	0.01317	−0.32	−0.03148
	안전		0.00052	0.01736	0.03	0.00305
후보자 요인[9]	성별		0.02256	0.01964	1.15	0.11936
	정당		−0.03203**	0.01533	−2.09	−0.23791
	나이		0.00112	0.00095	1.19	0.12238
	경력		0.01351*	0.00767	1.76	0.18901
	직업	정당인	0.01582	0.01755	0.90	0.11488
		변호사	−0.03305	0.02768	−1.19	−0.12893
		교수	−0.01700	0.02385	−0.71	−0.08119
		기타	−0.01629	0.02322	−0.70	−0.07780
상수			0.90238	0.08071	11.18	

N	96	R-squared	0.3377

**p<0.05, *p<0.1