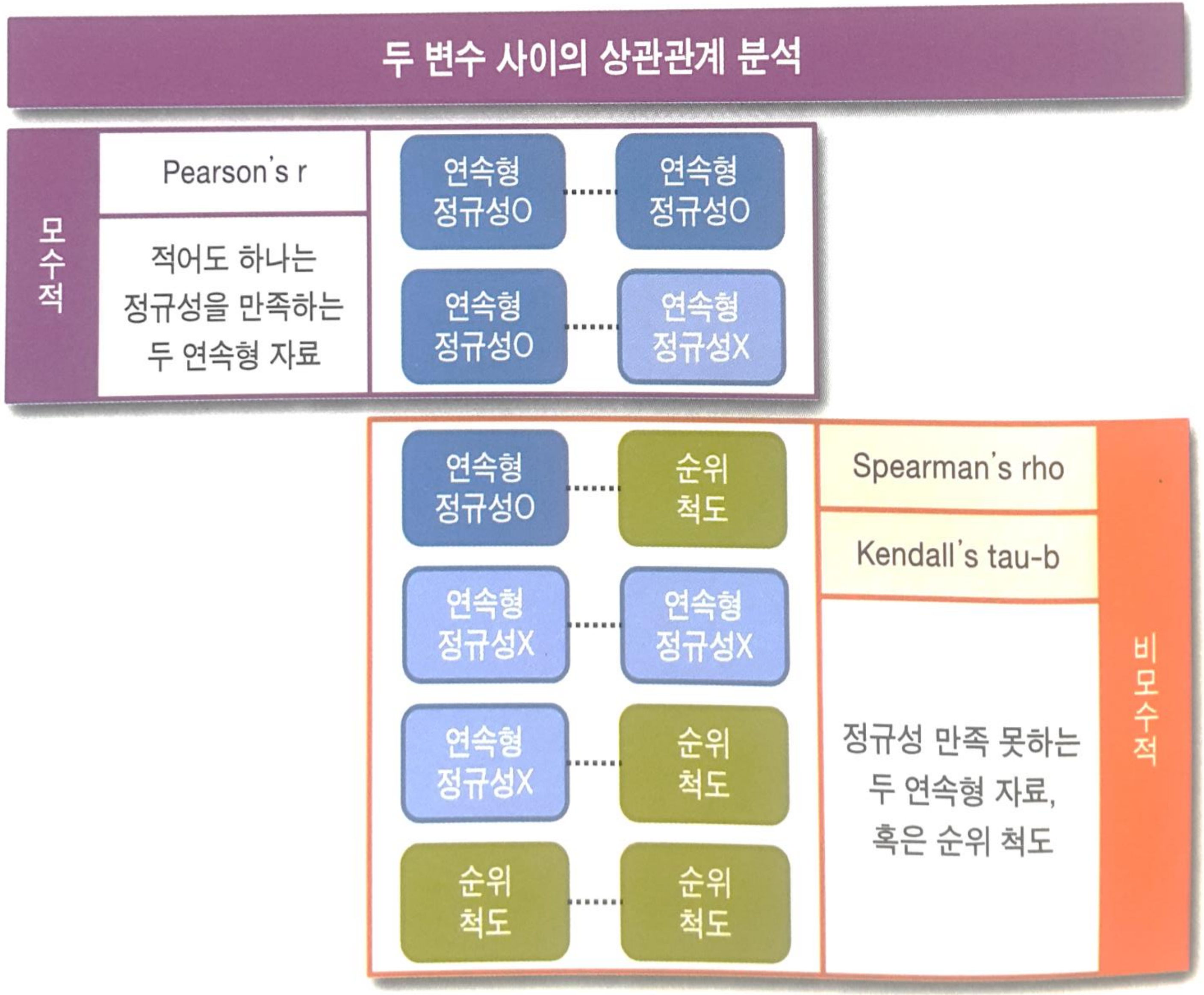
06 연속형 변수 사이의 선형관계 추정

20170502 모두의 바이오

박지혜

순서

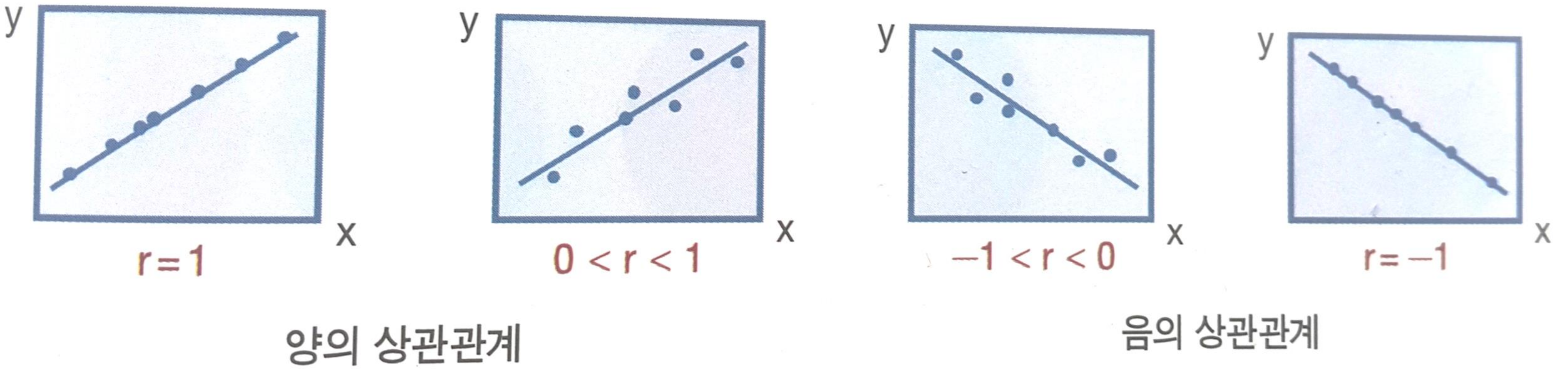
- 1. Pearson의 상관분석
- 2. Spearman의 순위상관분석
- 3. 단순회귀분석
- 4. 다중회귀분석



두 변수 사이의 상관관계 분석
모수적
Pearson's r
적어도 하나는 정규성을 만족하는 두 연속형 자료
연속형 정규성O
연속형 정규성O
연속형 정규성O
연속형 정규성X
연속형 정규성O
순위 척도
연속형 정규성X
연속형 정규성X
연속형 정규성X
순위 척도
순위 척도
순위 척도
Spearman's rho
Kendall's tau-b
정규성 만족 못하는 두 연속형 자료, 혹은 순위 척도
비모수적

1. Pearson의 상관분석

상관계수



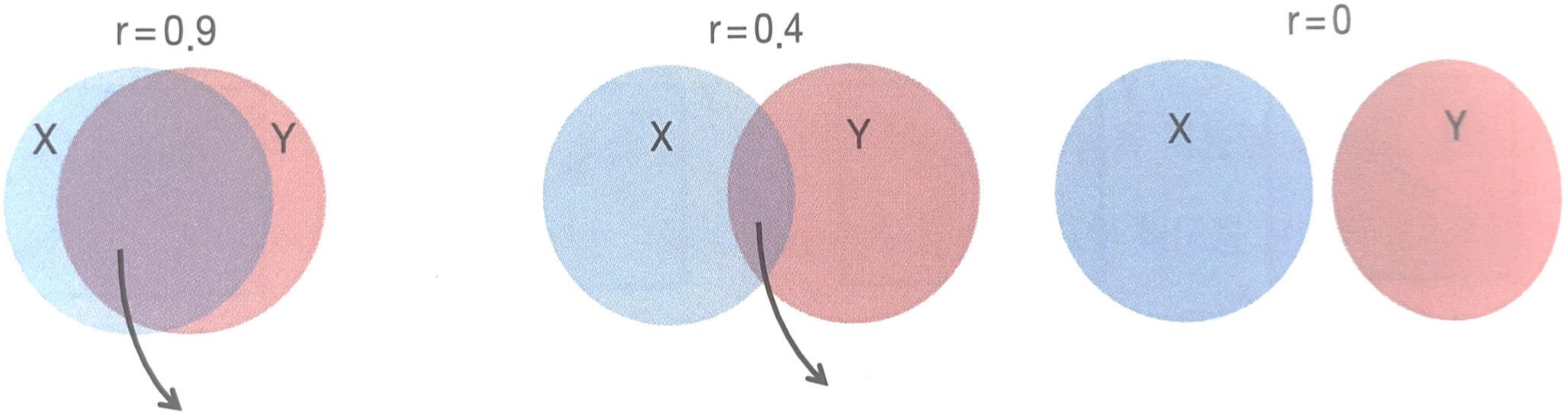
두 연속형 변수의 상관의 정도에 대해 알려줌 -1에서 1사이

상관계수는 직선의 기울기가 아니다

$$r_{xy} = \frac{Cov(X,Y)}{SD_x \times SD_y}$$

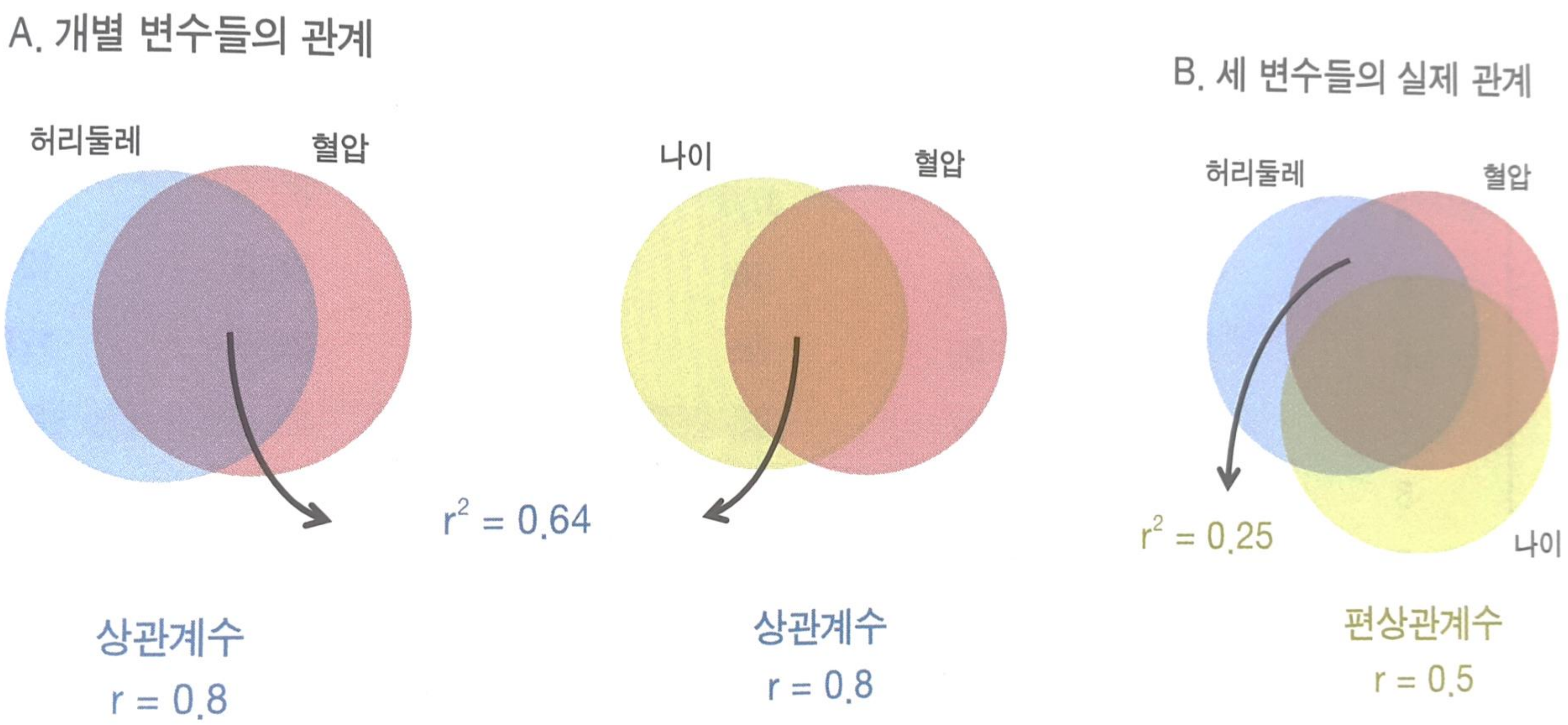
$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad SD_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \quad SD_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

설명력

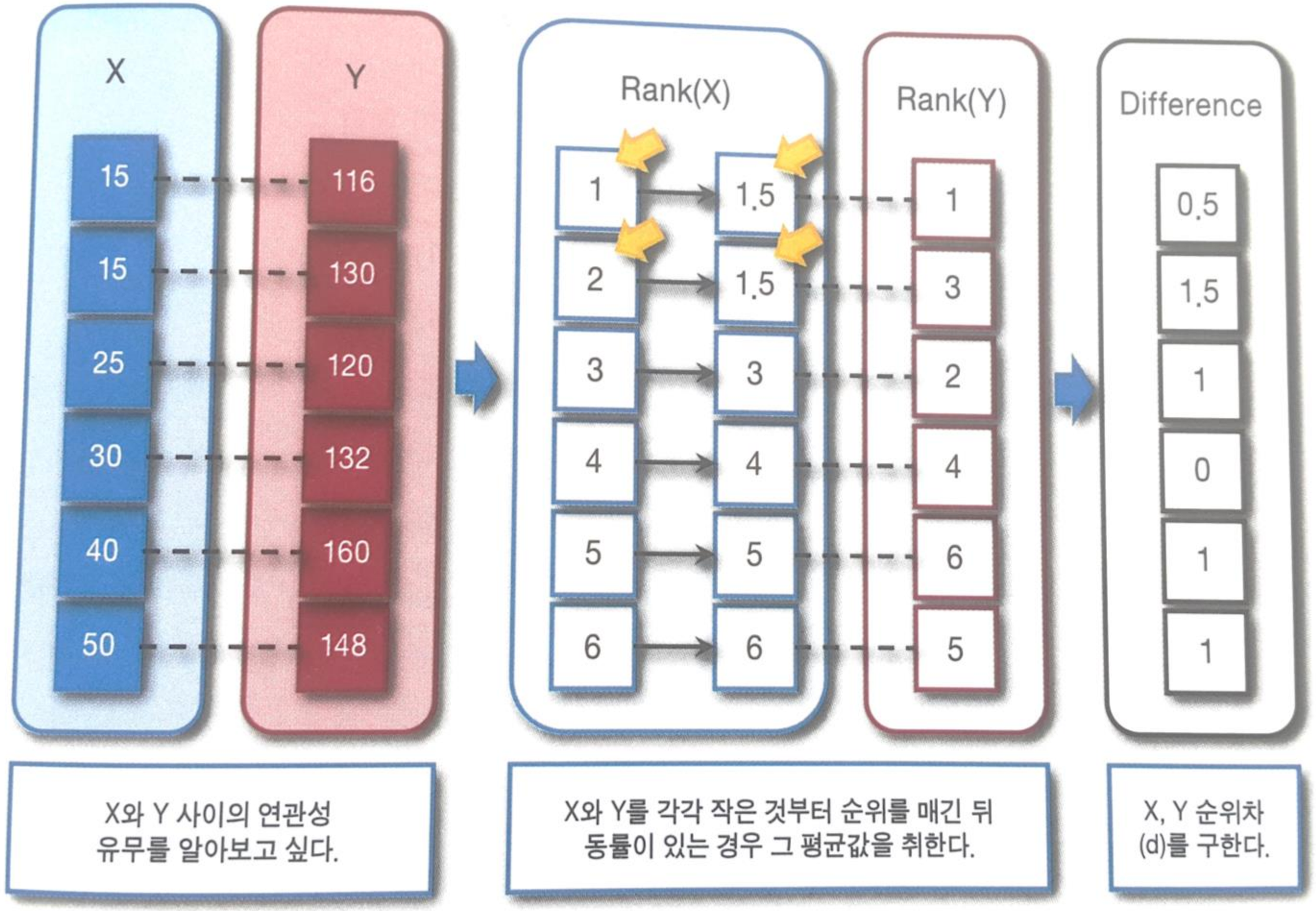


귀무가설 H_0	두 변수는 선형의 상관관계가 없다($r = 0$).
대립가설 H_1	두 변수는 선형의 상관관계가 있다($r \neq 0$).

편상관분석

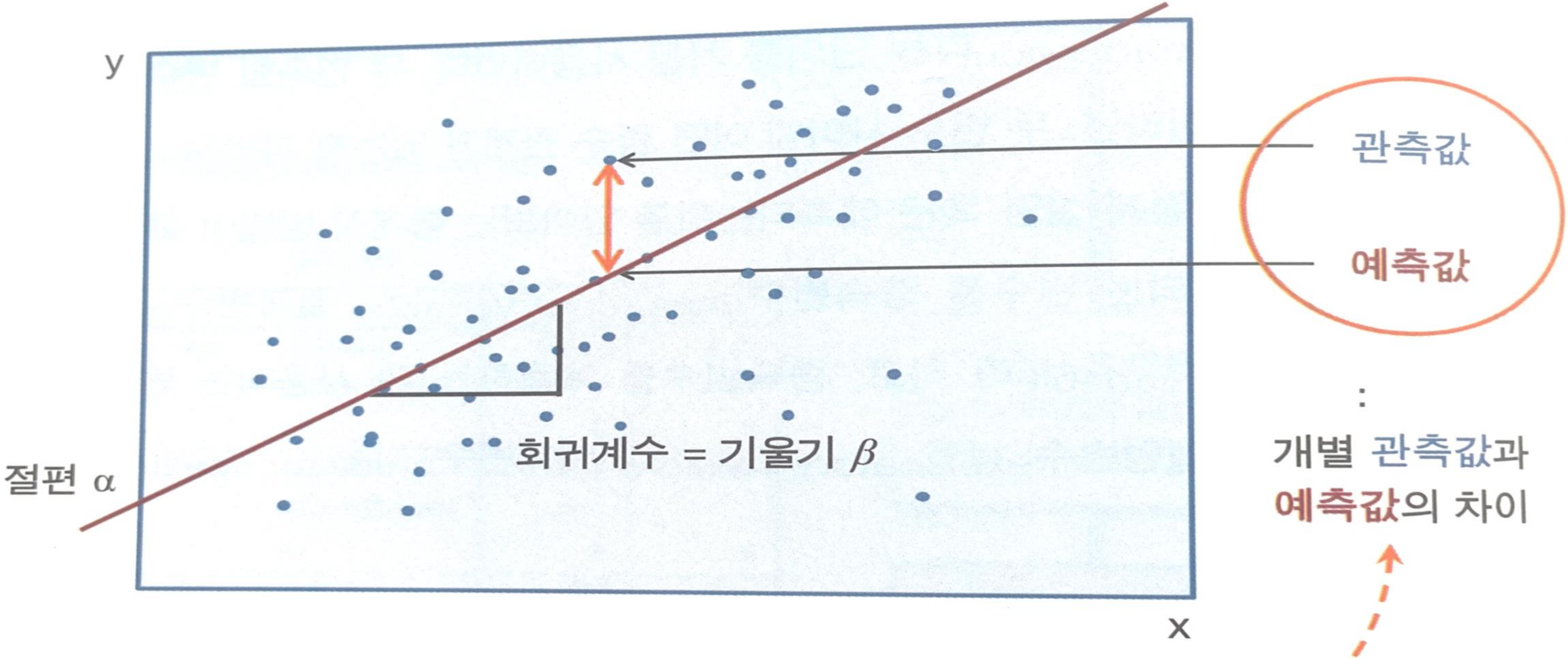

A. 개별 변수들의 관계
허리둘레
혈압
나이
혈압
$r^2 = 0.64$
상관계수
$r = 0.8$
상관계수
$r = 0.8$
B. 세 변수들의 실제 관계
허리둘레
혈압
나이
$r^2 = 0.25$
편상관계수
$r = 0.5$

2. Spearman의 순위상관분석



$$\text{Rho} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

3. 단순회귀분석



예측값 $\hat{y} = \alpha + \beta \times x$

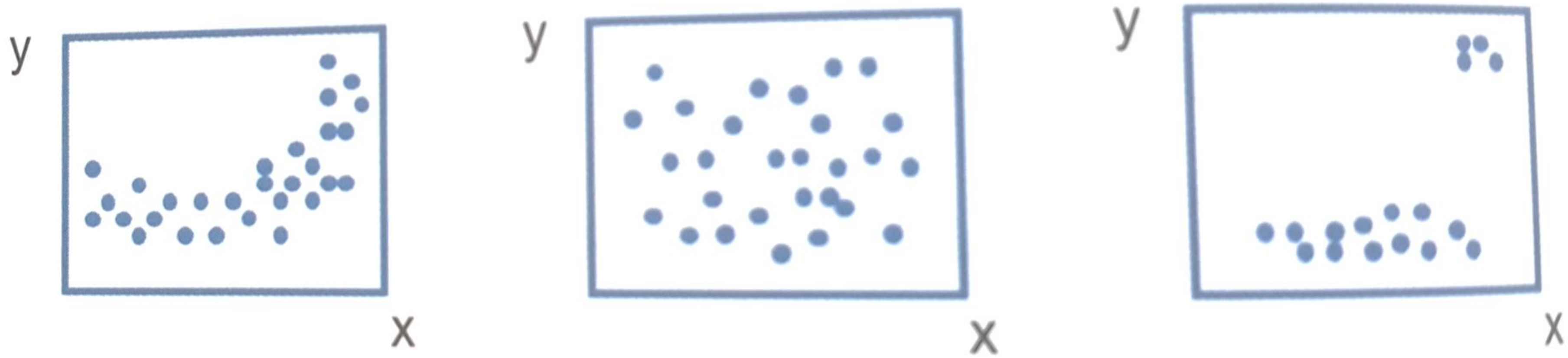
관측값 $y = \alpha + \beta \times x + \varepsilon$

회귀모형에 대한 기본 가정

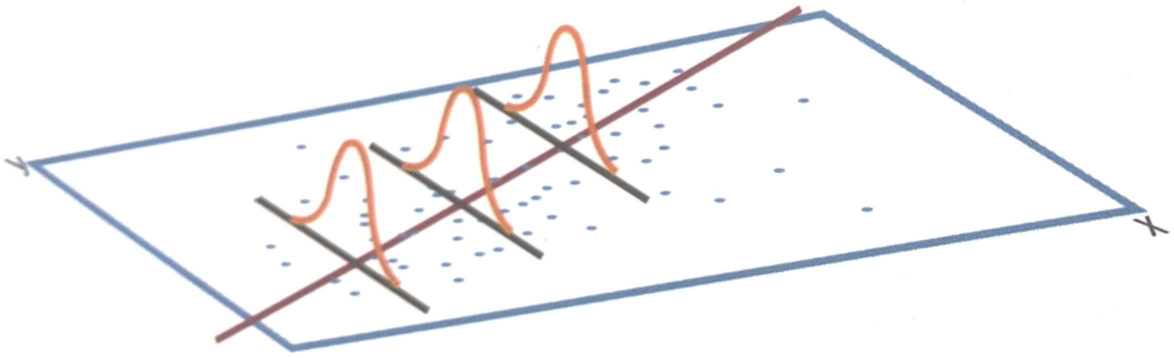
1	선형성	• 독립변수(x)와 종속변수(y)의 관계는 선형 관계에 있다. • 산점도로 확인
2	오차항의 정규성	• 모든 독립변수(x)의 값에서 종속변수(y)는 정규 분포를 이룬다. • 정규 P-P 곡선으로 확인
3	오차항의 독립성	• 개별 잔차들은 서로 독립이다. • 잔차산점도, Durbin-Watson 통계량으로 확인
4	오차항의 등분산성	• 모든 독립변수(x)의 값에서 종속변수(y)의 분산은 같다. • 잔차산점도로 확인

독립변수와 종속변수가 선형관계에 있을 것!!

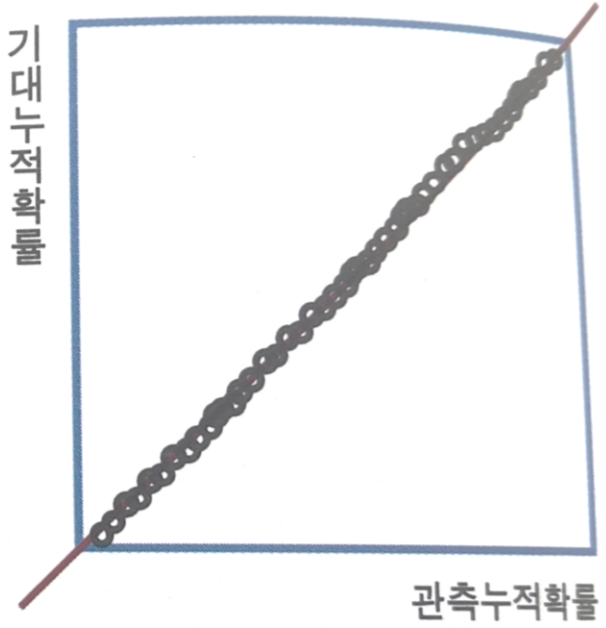
산점도–선형성의 확인



모든 독립변수 값에서 종속변수는 정규분포를 이룰 것!!



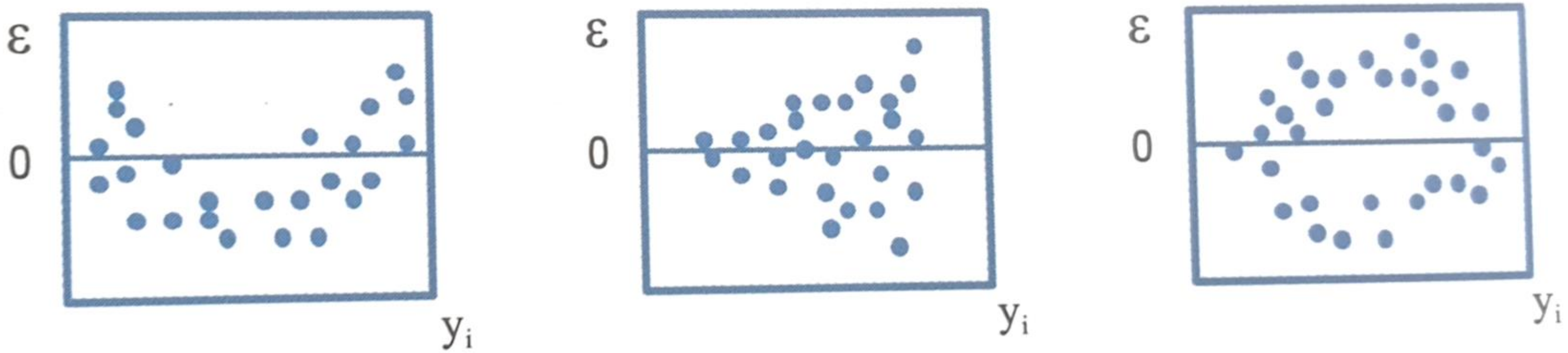
오차항의 정규성에 대한 3차원 모식도



정규 p-p 도표

잔차들은 서로 영향을 받지 않고 독립적일 것!!
모든 독립변수의 값에서 잔차의 분산은 같을 것!!

전차산점도—오차항의 독립성과 등분산성 확인



단순회귀분석을 시행하는 전체 과정

step 1 산점도로 두 변수의 선형관계 확인

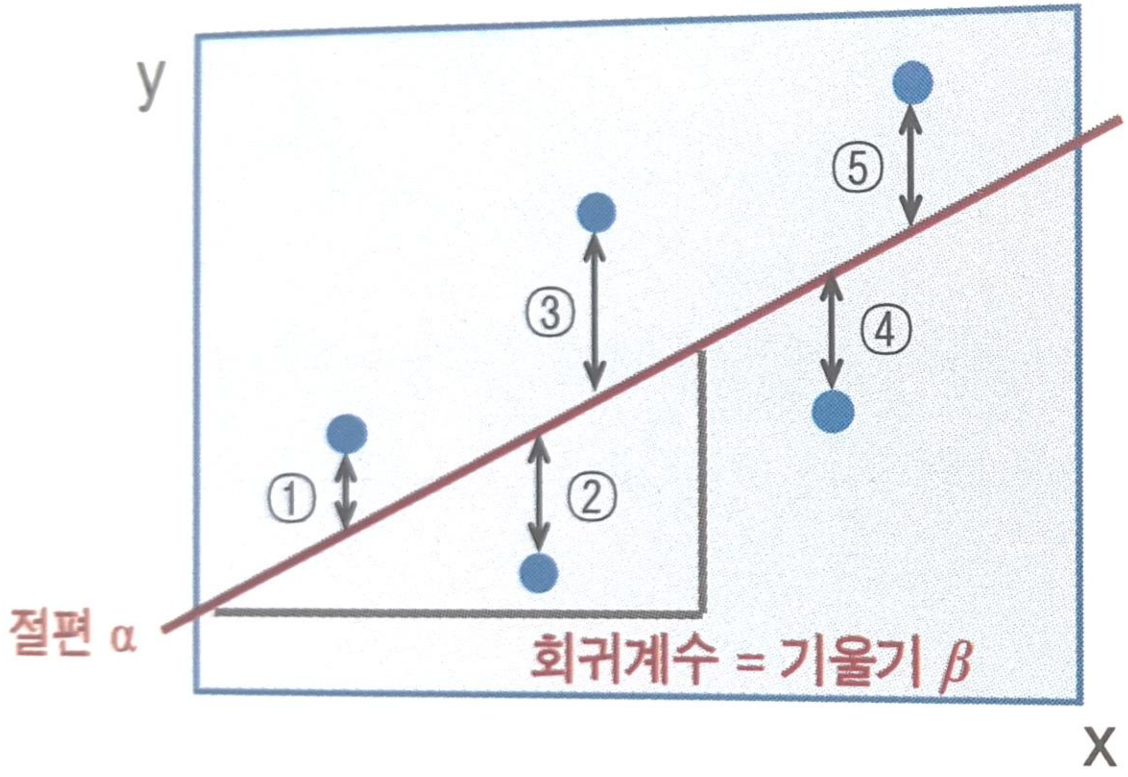
step 2 $y = \alpha + \beta \times x$ 회귀식의 추정 및 결정계수 R^2의 산출

step 3 회귀식($y = \alpha + \beta \times x$)의 유의성 검정($p < 0.05$)

step 4 회귀계수(β)의 유의성 검정($p < 0.05$)

step 5 정규 P-P 곡선, 잔차산점도로 회귀분석의 기본 가정 점검

회귀식 추정



잔차의 제곱합을 최소화하는

(①^2 + ②^2 + ③^2 + ④^2 + ⑤^2)

회귀식을 추정

즉, **절편 α**와 **회귀계수 β**를 추정

결정계수 R^2

- 총변동
- = 이 회귀식으로 설명할 수 있는 변동 + 그 외 설명되지 않는 변동
- 회귀식의 기여율 = 결정계수 R^2 = $\frac{\text{회귀식에 의해 설명되는 변동}}{\text{총변동}}$
- 두 변수 사이의 상관관계 = 상관계수 r = $\pm\sqrt{R^2}$

회귀식의 유의성 검정

귀무가설 H_0	회귀모형($y = \alpha + \beta \times x$)으로 설명할 수 없다.
대립가설 H_1	회귀모형($y = \alpha + \beta \times x$)으로 설명할 수 있다.

회귀식의 유의성 검정

$$F = \frac{\text{회귀식에 의해 설명되는 변동의 평균}}{\text{그 외 설명되지 않는 변동의 평균}}$$

분산분석표

k(독립변수의 수), n(총개체 수)

요인	제곱합(변동)	자유도	제곱합의 평균	F 통계량	유의확률
군	SSR(회귀식으로 설명되는 변동)	k	MSR=SSR/k	F=MSR/MSE	p value
오차	SSE(그 외 설명되지 않는 변동)	n−k−1	MSE=SSE/(n−k−1)		
전체	TSS(총변동)	n−1			

* SSR = sum of squares for regression
SSE = sum of squares for error
TSS = total sum of squares

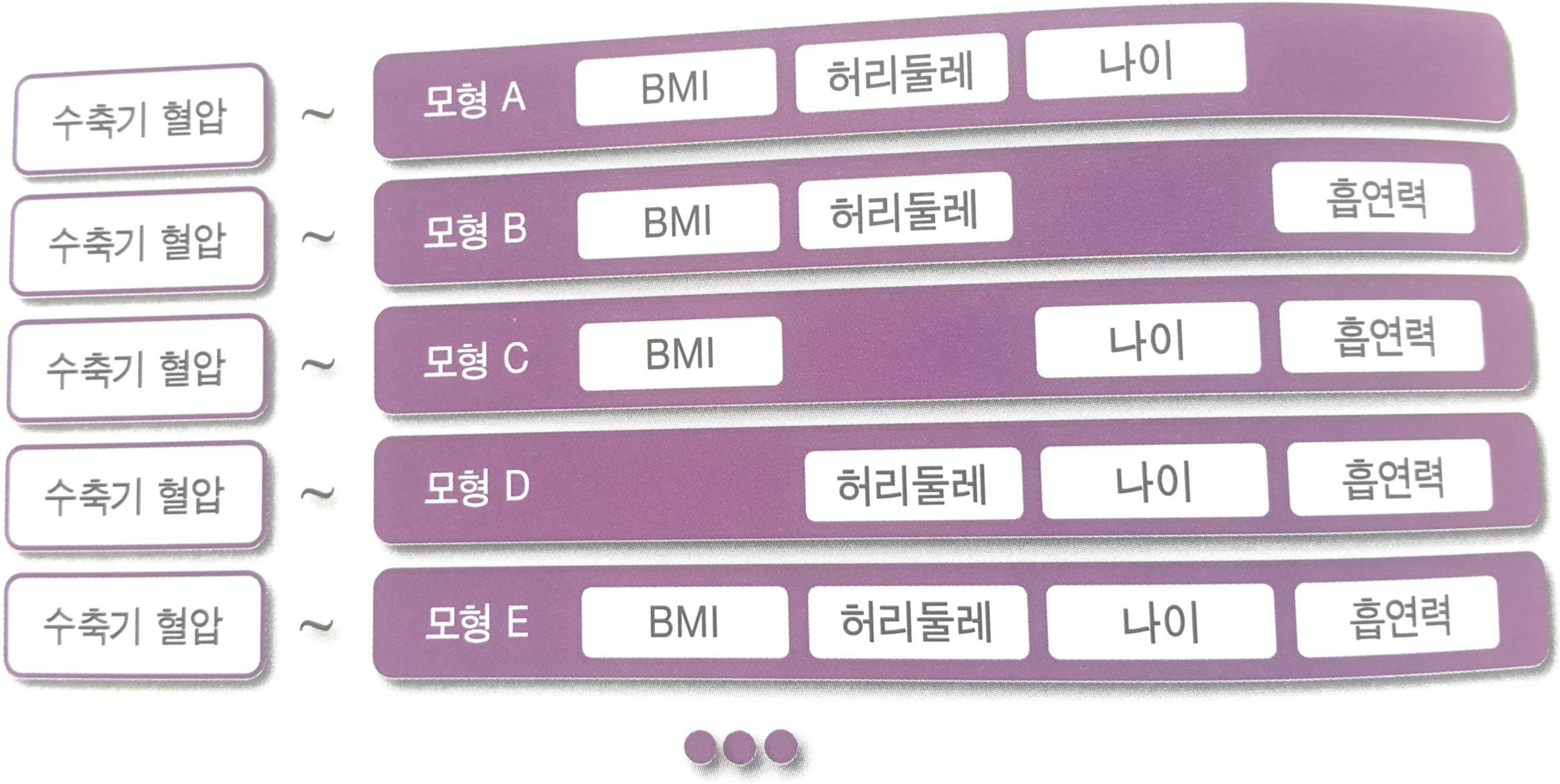
MSR = mean squares for regression
MSE = mean squares for error

회귀계수(β)의 유의성 검정

4) 회귀계수(β)의 유의성 검정

귀무가설 H_0	독립변수는 종속변수와 관계가 없다($\beta = 0$).
대립가설 H_1	독립변수는 종속변수와 관계가 있다($\beta \neq 0$).

4. 다중회귀분석


수축기 혈압 ~ 모형 A BMI 허리둘레 나이
수축기 혈압 ~ 모형 B BMI 허리둘레 흡연력
수축기 혈압 ~ 모형 C BMI 나이 흡연력
수축기 혈압 ~ 모형 D 허리둘레 나이 흡연력
수축기 혈압 ~ 모형 E BMI 허리둘레 나이 흡연력

목적은…

1. 주요 독립변수를 파악하려고
2. 교란변수를 통제하려고
3. 종속변수를 예측하려고

	단순회귀분석	다중회귀분석
회귀식 추정	$y = \alpha + \beta \times x$	$y = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3$
결정계수	결정계수 R^2	수정결정계수 R^2_{adj}
회귀식의 유의성	H_0: 회귀식으로 설명할 수 없다. ($\beta = 0$)	H_0: 회귀식으로 설명할 수 없다. ($\beta_1 = \beta_2 = \beta_3 = 0$)
회귀계수의 유의성	H_0: $\beta = 0$	H_0: $\beta_1 = 0$ H_0: $\beta_2 = 0$ H_0: $\beta_3 = 0$

독립변수를 선택하는 방법

- 전진선택법 - 회귀계수의 유의확률이 가장 낮은 변수를 먼저 넣고, 유의확률이 가장 낮은 변수를 순서대로 하나씩 추가
- 한번 선택되면 절대 제거안됨

- 후진선택법 – 모든 변수를 회귀모형에 넣은 상태에서 유의확률이 높은 변수부터 순서대로 하나씩 제거하여 유의한 변수들만 남을 때 까지 계속 제거
- 한번 제외되면 다시 선택되지 못함

단계선택법의 예

STEP 1　회귀계수(β)의 유의확률(p)이 가장 작은 독립변수인 BMI가 모형에 처음 **진입**되었다.

수축기 혈압 ~ 모형 1 | BMI

STEP 2　남은 변수 중, 모형에 추가했을 때 유의확률이 가장 작은 허리둘레가 모형에 **진입**되었다.
두 변수 보정상태에서 두 회귀계수(β)가 모두 유의하면 두 변수를 모두 남긴다.

수축기 혈압 ~ 모형 2 | BMI | 허리둘레

STEP 3　남은 변수 중, 모형에 추가했을 때 유의확률이 가장 작은 나이가 모형에 **진입**되었다.

수축기 혈압 ~ 모형 3 | BMI | 허리둘레 | 나이

STEP 4　나이를 보정하고 나니 허리둘레의 회귀계수(β)가 유의하지 않아 허리둘레가 **제거**되었다.

수축기 혈압 ~ 모형 4 | BMI | 나이

STEP 5　더 이상 새로 추가할 조건에 맞는 독립변수가 없을 때까지 이 과정을 반복한다.

⋮

다중공선성(multicollinearity)

- 독립변수들 간에
- 완전한 또는 거의 완전한 선형의 종속관계가 존재하는 것

- 이런 경우 임상적으로 의미가 더 크고 또 종속변수와 관련성이 높은 변수 하나만 선택하여 모형에 투입해야 함.

- 공차한계와 분산팽창요인으로 다중공선성을 파악할 수 있음.

공차한계 (Tolerance)와 분산팽창요인(VIF)

- R_k^2 : 회귀식의 독립변수 $X_1, X_2, \ldots\ldots, X_n$ 중 X_k를 종속변수로 하고, 나머지 변수들을 독립변수로 회귀분석을 수행했을 때의 결정계수
- R_k^2이 크다는 것은 이미 다른 변수들에 의해 X_k가 거의 설명되고 있음을 의미

- 공차한계 Tolerance = 1- R_k^2
- 분산팽창요인 Variance Inflation Factor,
- VIF = $\frac{1}{Tolerance} = \frac{1}{1- R_k^2}$
- VIF가 클수록 회귀식의 신뢰도를 떨어뜨림.
- 보통 10이상이면 다중공선성이 존재하는 것으로 간주