회귀분석

- 권수태 교수

JEONJU UNIVERSITY
The Place for Superstars

1. 회귀

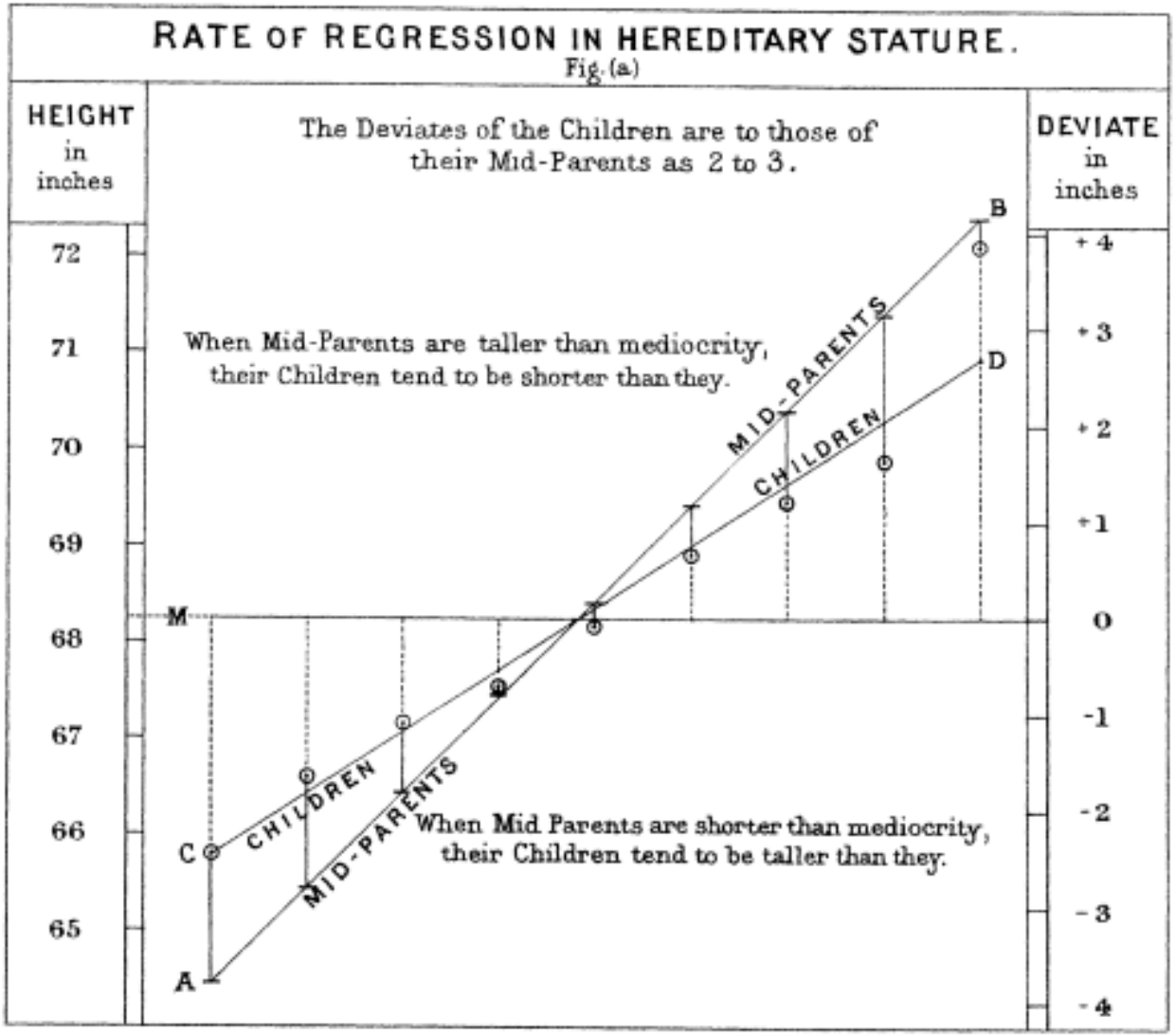
❖ 유래

- 19세기 영국의 과학자 프랜시스 갤턴(Francis Galton)이 1886년 발표한 ' Regression toward Meiocrity in Hereditary Stature ' 라는 논문에서 비롯됨
- 키 큰 선대 부모들이 낳은 자식들의 키가 점점 더 커지지 않고 다시 평균 키로 회귀하는 경향을 발견
- 사람의 키는 평균 키로 회귀하려는 경향을 가진다는 자연의 법칙이 있다는 것
  - ✓ 한 실험에서 205쌍의 부모와 성인 자녀에 대한 자료를 수집하여, 부모의 키의 평균(Mid parent height)를 구한 후 이를 몇몇 그룹으로 묶었음
  - ✓ 각각의 그룹에 대해 그 자녀들 키의 중위수(median height)를 구한 후 이를 그래프로 표시

1. 회귀

❖ 유래

어른키 평균
68.25 inch



JEONJU UNIVERSITY

1. 회귀

❖ 추정과 검정

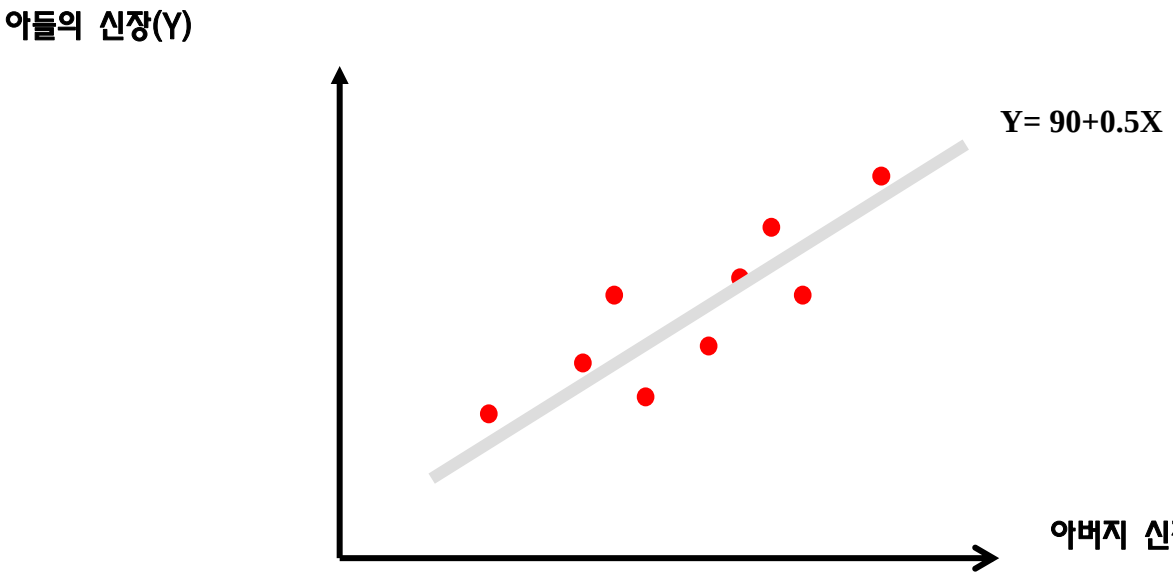
검정문제

❑ 아버지키가 크면 아들키도 큰가?

추정문제

❑ 아들키를 아버지키로 예측가능한가?

❑ 두 변수간에 관계식은?



JEONJU UNIVERSITY

1. 회귀

❖ 추정과 검정

- ➢ 통계학의 데이터분석은 추정과 검정의 단계로 이루어져 있음
- ➢ 독립변수(X)와 종속변수(Y)의 관계식에서
  - ✓ Y = a + b X
  - ✓ 추정 : 회귀식, 회귀계수
  - ✓ 검정 : 독립변수의 영향력(b=0?), 모형의 적합성 등

1. 회귀

❖ 회귀분석의 정의

- ➢ 독립변수와 종속변수의 관계를 규명
- ➢ 관계식(회귀식)을 추정하고
- ➢ 관계(영향력)의 유무를 검정

❖ 회귀분석의 종류

- ➢ 단순회귀분석: 독립변수의 수가 1
- ➢ 다중회귀분석: 독립변수의 수가 2이상

JEONJU UNIVERSITY

1. 회귀

❖ 회귀분석 적용 예

- ➢ 독립변수는 계량, 명목 가능
- ➢ 종속변수는 계량만 가능(명목일 때는 다른 분석 사용)

독립변수	종속변수
연령, 수입, 학력, 주거	구매성향
살균시간, 온도	균의 숫자
가격, 광고시간, 점포수	매출액
근무년수, 상여금, 교육	판매원 업적

1. 회귀

❖ 우리는 현재 (변수와 변수)관계에 관심

- ➢ 관계가 있다? 없다?(검정의 문제)
- ➢ 어느정도 관계가 있는가?(추정의 문제?)

- ➢ 다른 변수값을 예측 또는 추정한다면
  - ✓ 수학 60점이니까 물리는 70점이겠다(계량=>계량) : 회귀분석
  - ✓ 영어가 550점이니까 불합격 하겠네(계량=>명목) : 로지스틱 회귀분석
  - ✓ 남자니까 검은색 좋아하겠네(명목=>명목) : 로그선형모형

❖ 회귀계수의 검정

➢ 모형 :

$$Y = \alpha + \beta X$$

➢ 가설 :

$$H_0 : \beta = 0 \; vs. \; H_1 : \beta \neq 0$$

$\Rightarrow$ 독립변수가 종속변수에 영향을 주는가?

(절편에 관한 검정은 중요하지 않음)

➢ Idea : 만약 β 가 0이라면 X의 변화가 Y에 전혀 영향을 주지 못함

1. 회귀

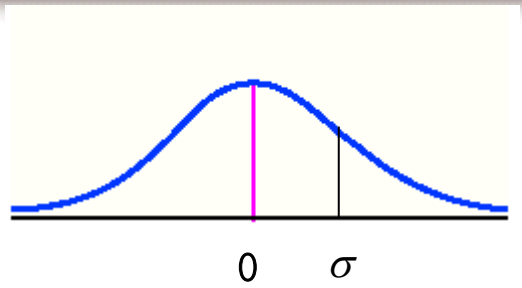
❖ 결정계수 (R 제곱)

- 유의성 검정에서 귀무가설이 기각되더라도 이는 기울기가 0이 아니라는 것뿐이지 추정된 회귀식이 전체자료를 잘 설명해 주고 있다고 판단하기는 어려움
- 그래서 표본자료로부터 추정된 회귀선이 그 측정자료에 어느 정도 적합한가를 측정하는 측도인 결정계수(coefficient of determination)가 필요
- 관측값 y의 총변동은 회귀선에 의해 설명되는 변동과 설명되지 않는 변동으로 나누어짐
  - ✓ SST=SSR+SSE
  - ✓ R^2=SSR/SST
    - R^2 =1이면 회귀선으로 y의 총변동이 완전히 설명된다는 것을 의미
    - R^2 =0이면 회귀선으로 x와 y의 관계를 전혀 설명하지 못한다는 의미

2. 회귀분석

❖ 단순 회귀분석

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$



➢ 모형

$$Y_i = \alpha + \beta X_i + \epsilon_i, \qquad i = 1, \ldots, n$$

종속 변수 | 상수 항 | 회귀 계수 | 독립 변수 | 오차 항

$$E(\epsilon_i) = 0, \qquad Var(\epsilon_i) = \sigma^2$$

입력변수

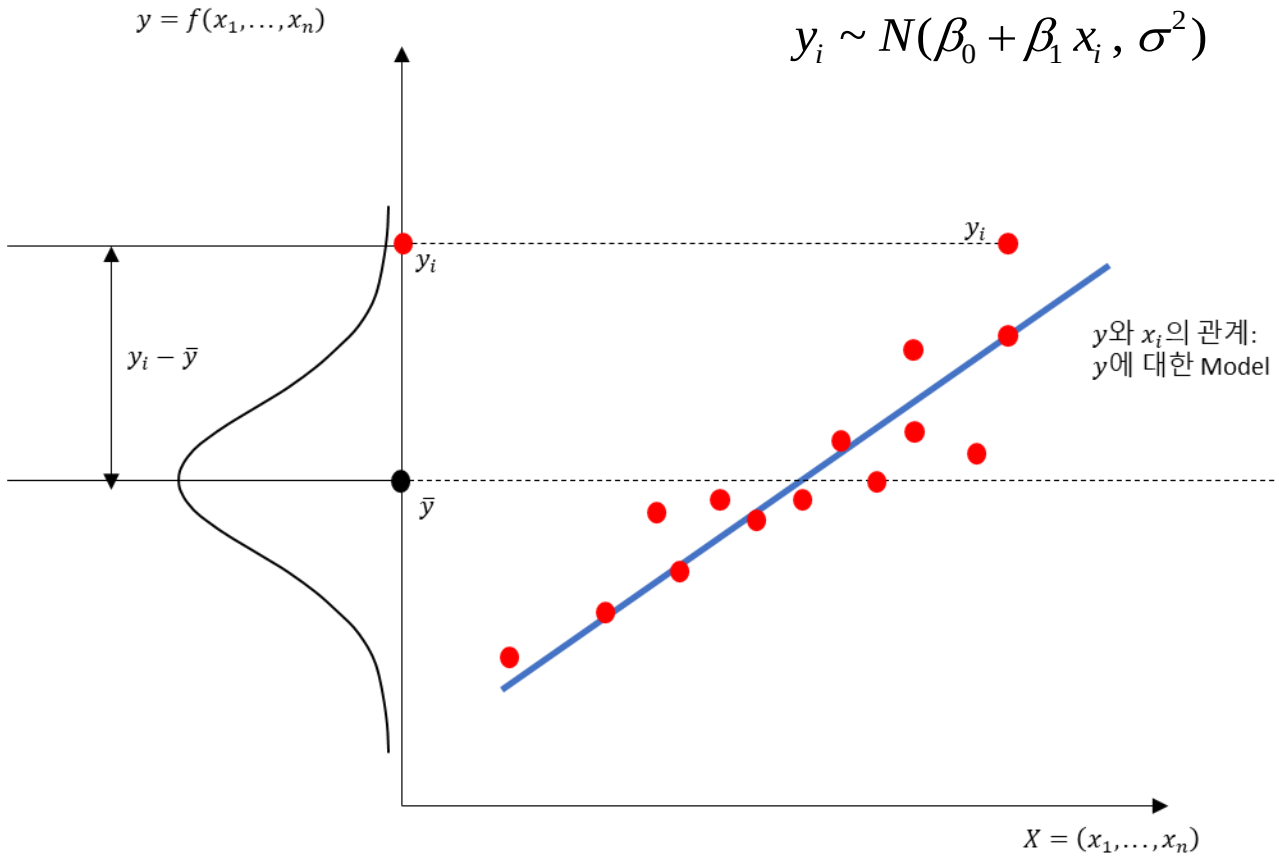
모수: 추정할 값

확률변수: 추정 못함

x_i

2. 회귀분석

❖ 단순 회귀분석



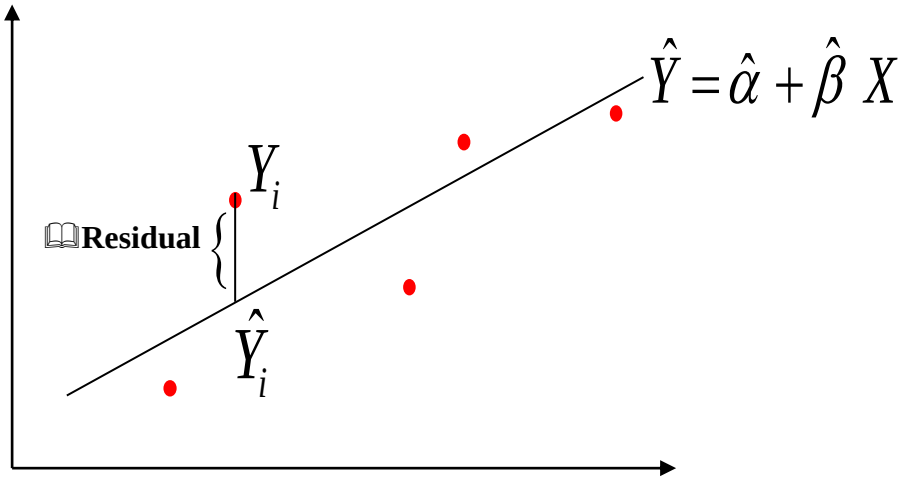
2. 회귀분석

❖ 단순 회귀분석

➢ α, β 추정

✓ 잔차(residual)를 최소화 하는 회귀직선식을 구함

$$\min \sum (Y_i - \hat{Y}_i)^2$$



2. 회귀분석

❖ 단순 회귀분석

➢ α, β 추정

✓ 최소자승법(Ordinary Least Squares :OLS)에 의한 최소자승추정량(Least Square Estimates)

$$\min \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

$$\frac{\partial}{\partial \hat{\alpha}} \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 = 0$$

$$\frac{\partial}{\partial \hat{\beta}} \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 = 0$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

2. 회귀분석

일차함수

$$y = a + bx$$

절편 기울기

단순회귀

$$y = w_0 + w_1 x$$

상수항 회귀계수

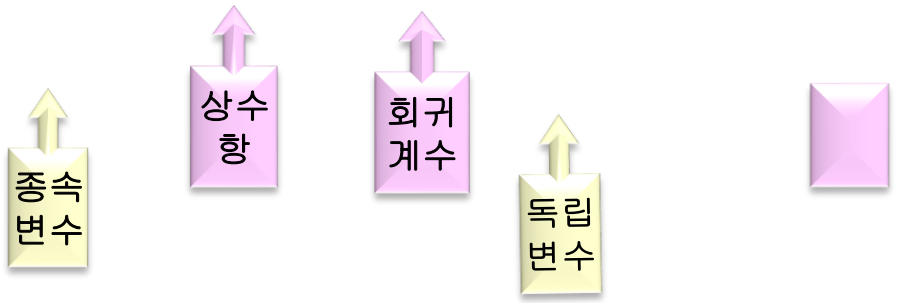
종속변수 독립변수

미지수, 매개변수

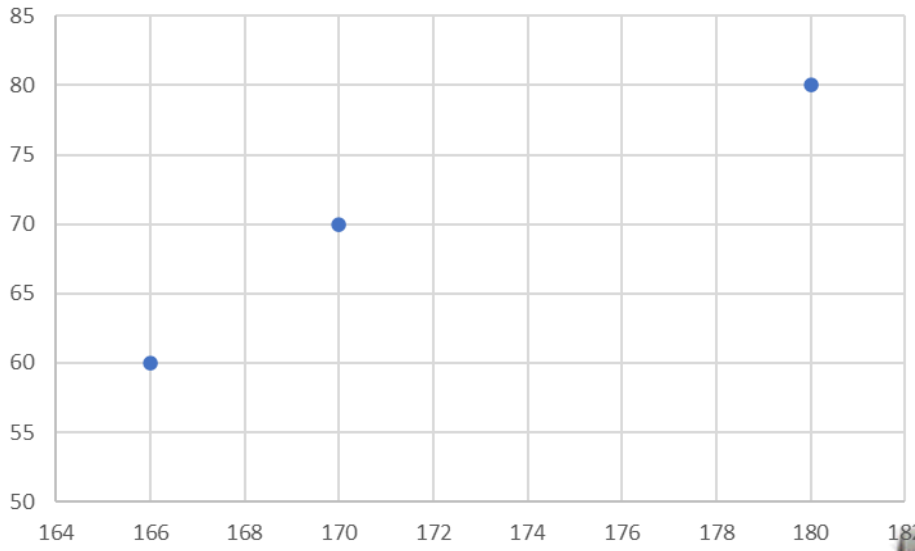
2. 회귀분석

단순회귀

$$y_i = w_0 + w_1 x_i$$



키	몸무게
166	60
170	70
180	80

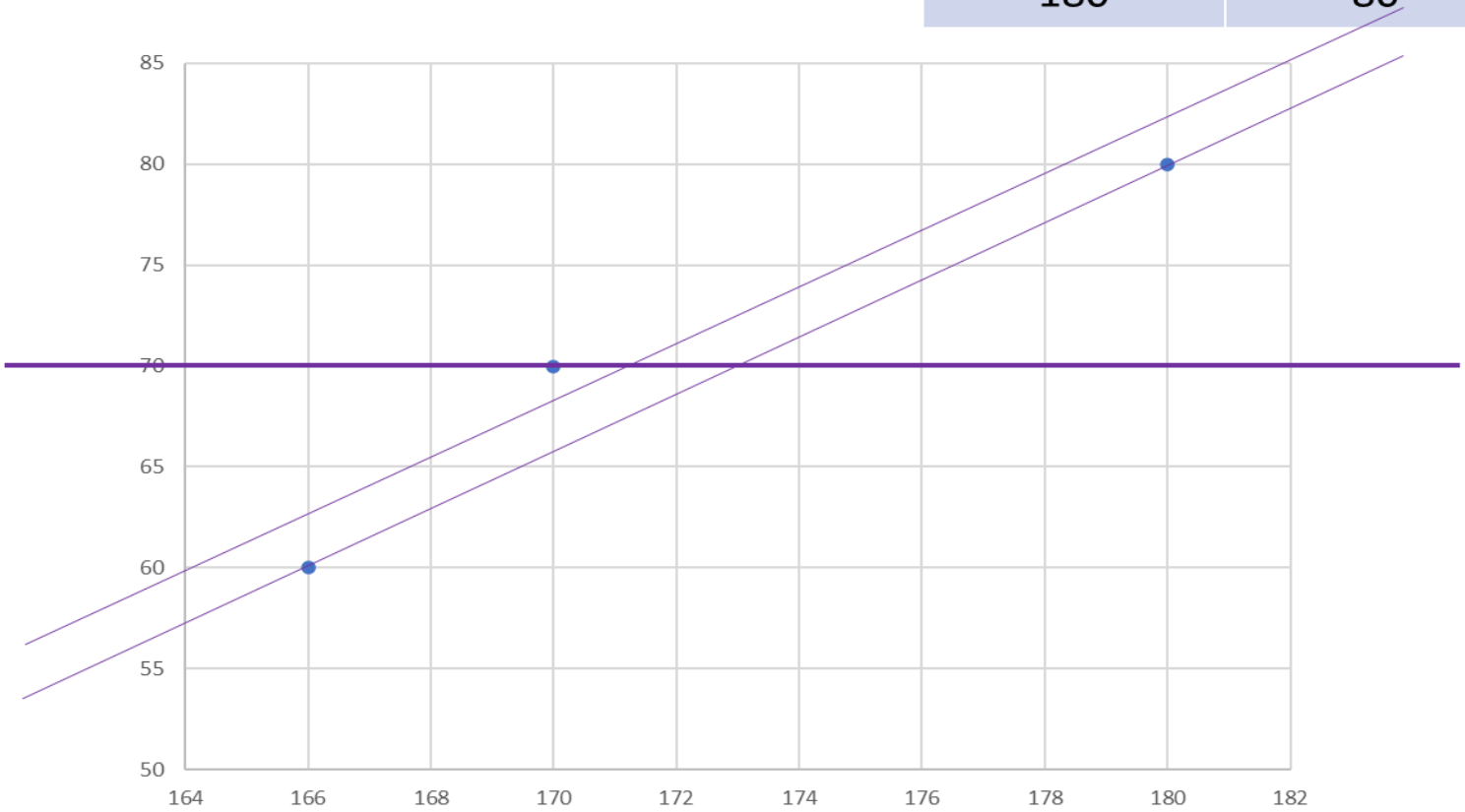


JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

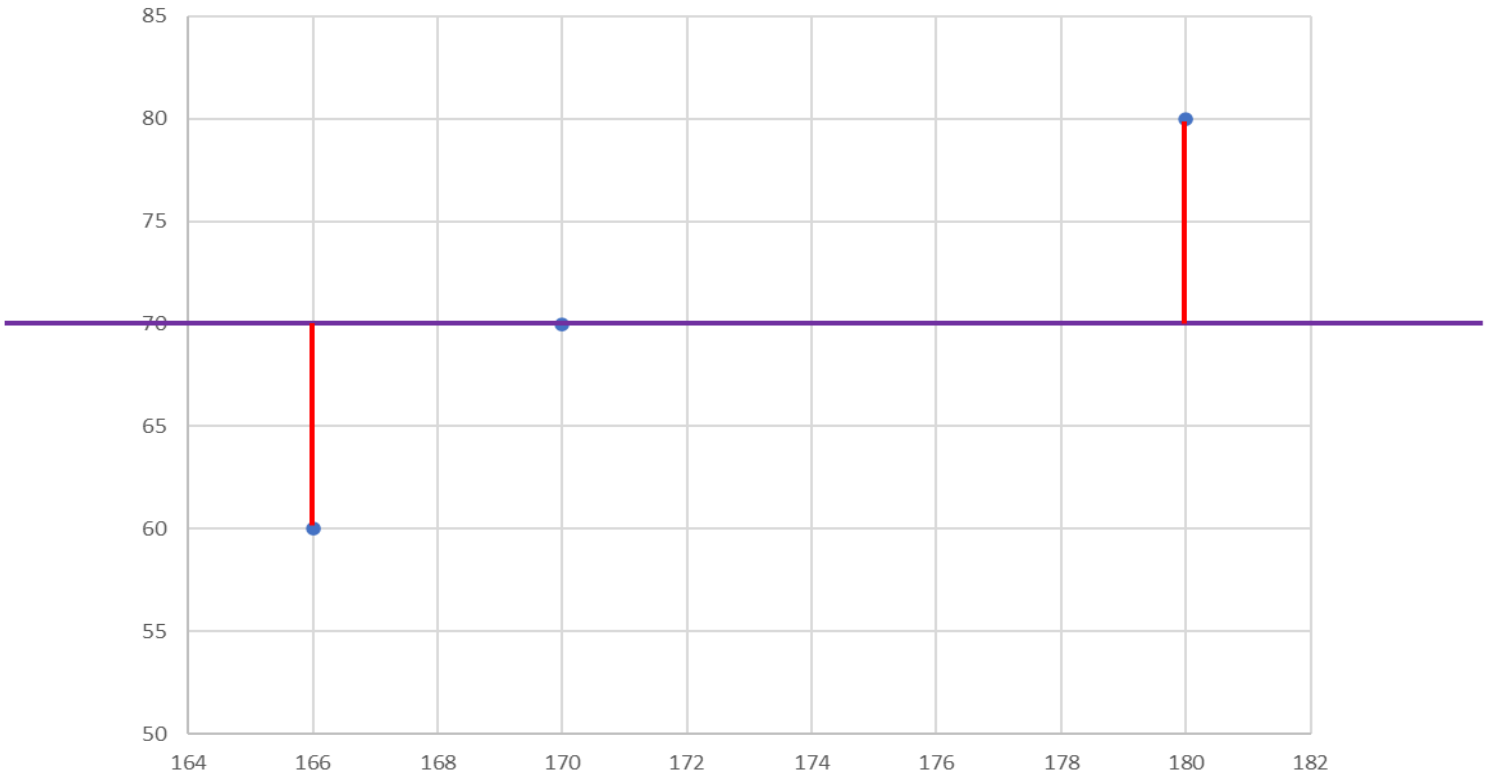


JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

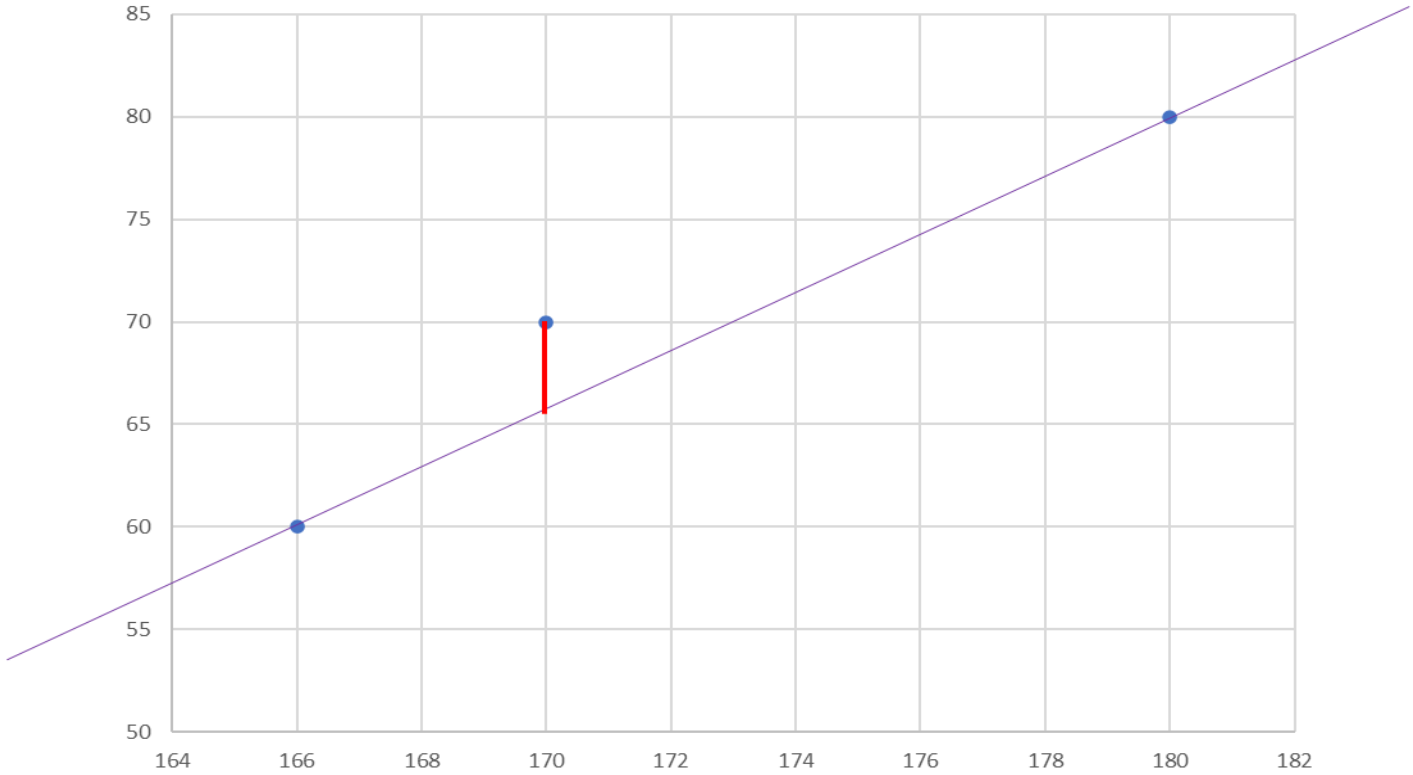


JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

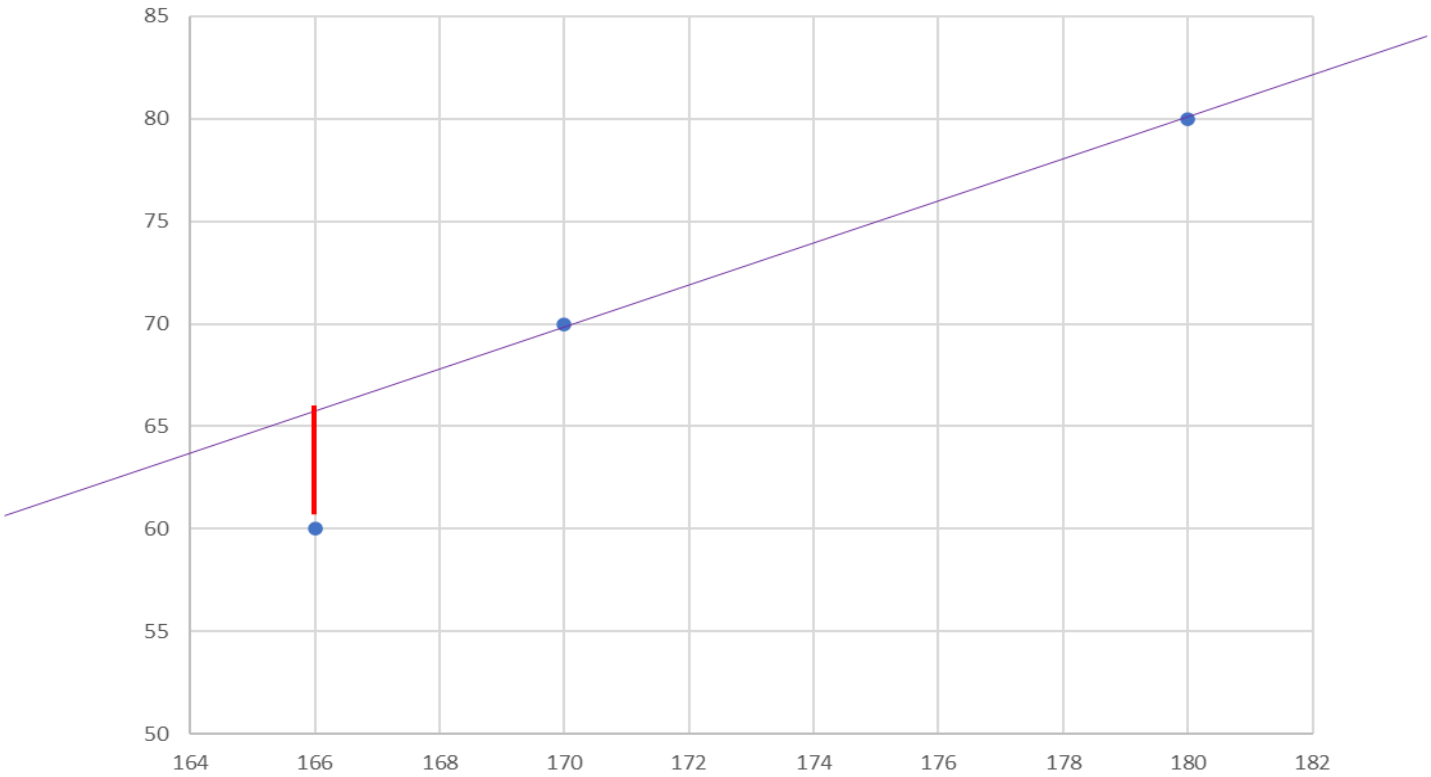


JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

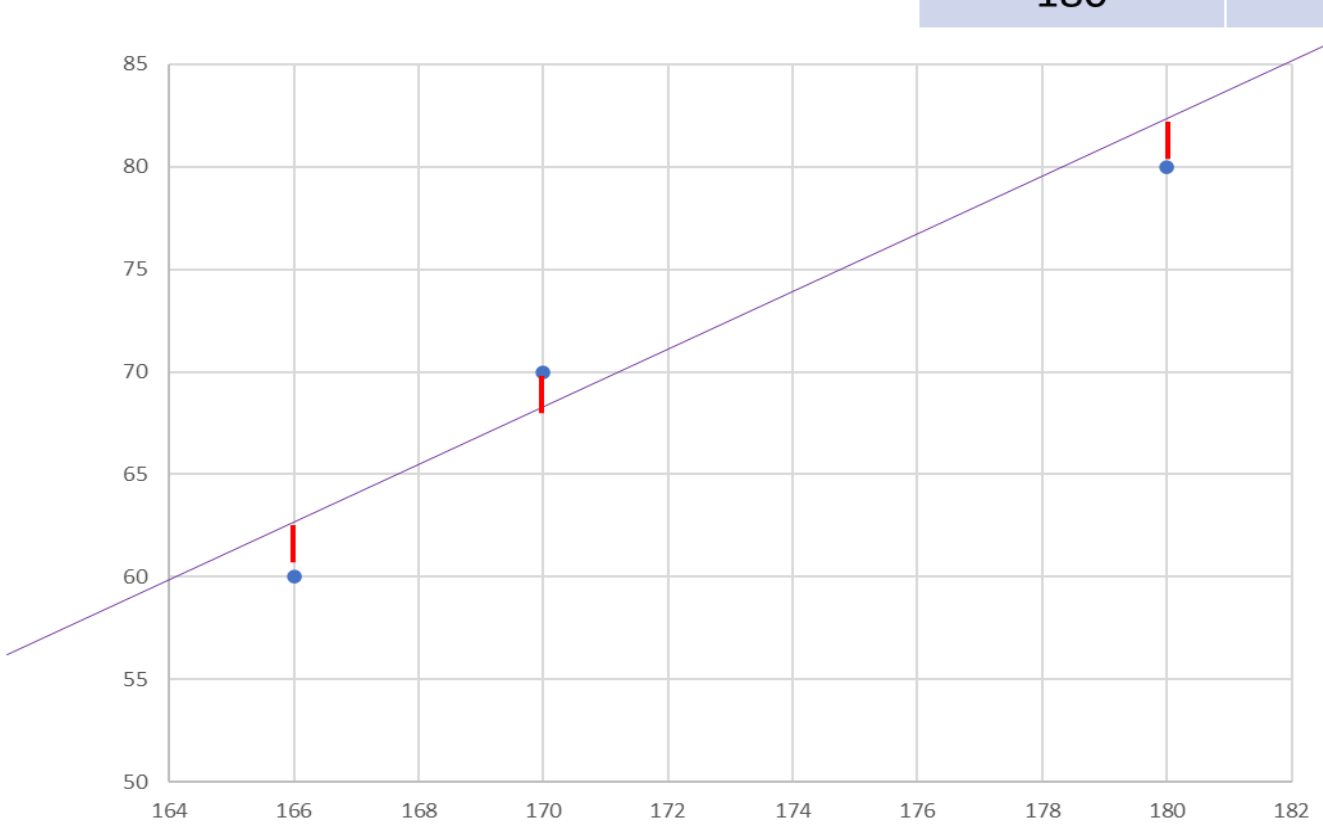


JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80



JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

$$L(w_0, w_1) = \sum_{i=1}^{3} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{3} (y_i - w_0 - w_1 x_i)^2$$

$$= \sum_{i=1}^{3} ({y_i}^2 + {w_0}^2 + {w_1}^2 {x_i}^2 - 2 y_i w_0 - 2 y_i w_1 x_i + 2 w_0 w_1 x_i)$$

$$= \sum_{i=1}^{3} {y_i}^2 + 3{w_0}^2 + {w_1}^2 \sum_{i=1}^{3} {x_i}^2 - 2 \sum_{i=1}^{3} y_i w_0$$

$$-2 \sum_{i=1}^{3} x_i y_i w_1 + 2 w_0 w_1 \sum_{i=1}^{3} x_i$$

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
173	70
180	80

$$\frac{\partial L(w_0, w_1)}{\partial w_0} = 6w_0 - 2\sum_{i=1}^{3} y_i + 2w_1 \sum_{i=1}^{3} x_i$$

$$\frac{\partial L(w_0, w_1)}{\partial w_1} = -2\sum_{i=1}^{3} x_i y_i + 2w_1 \sum_{i=1}^{3} x_i^2 + 2w_0 \sum_{i=1}^{3} x_i$$

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

$$\frac{\partial L(w_0, w_1)}{\partial w_0} = 6w_0 - 2\sum_{i=1}^{3} y_i + 2w_1 \sum_{i=1}^{3} x_i = 0$$

$$6w_0 = 2\sum_{i=1}^{3} y_i - 2w_1 \sum_{i=1}^{3} x_i$$

$$w_0 = 1/3 \sum_{i=1}^{3} y_i - 1/3 w_1 \sum_{i=1}^{3} x_i$$

$$w_0 = \bar{Y} - w_1 \bar{X}$$

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

$$\frac{\partial L(w_0, w_1)}{\partial w_1} = -2\sum_{i=1}^{3} x_i y_i + 2w_1 \sum_{i=1}^{3} x_i^2 + 2w_0 \sum_{i=1}^{3} x_i = 0$$

$$-2\sum_{i=1}^{3} x_i y_i + 2w_1 \sum_{i=1}^{3} x_i^2 + 2\left(\frac{\sum_{i=1}^{3} y_i}{3} - \frac{\sum_{i=1}^{3} x_i}{3} w_1\right)\sum_{i=1}^{3} x_i = 0$$

$$-2\sum_{i=1}^{3} x_i y_i + 2w_1 \sum_{i=1}^{3} x_i^2 + 2\left(\frac{\sum_{i=1}^{3} y_i \sum_{i=1}^{3} x_i}{3} - \frac{(\sum_{i=1}^{3} x_i)^2}{3} w_1\right) = 0$$

$$-\sum_{i=1}^{3} x_i y_i + w_1 \sum_{i=1}^{3} x_i^2 + \left(\frac{\sum_{i=1}^{3} y_i \sum_{i=1}^{3} x_i}{3} - \frac{(\sum_{i=1}^{3} x_i)^2}{3} w_1\right) = 0$$

JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $y_i = w_0 + w_1 x_i$

키	몸무게
166	60
170	70
180	80

$$\frac{\partial L(w_0, w_1)}{\partial w_1} = -2\sum_{i=1}^{3} x_i y_i + 2w_1 \sum_{i=1}^{3} x_i^2 + 2w_0 \sum_{i=1}^{3} x_i = 0$$

$$-\sum_{i=1}^{3} x_i y_i + w_1 \sum_{i=1}^{3} x_i^2 + \left(\frac{\sum_{i=1}^{3} y_i \sum_{i=1}^{3} x_i}{3} - \frac{(\sum_{i=1}^{3} x_i)^2}{3} w_1 \right) = 0$$

$$-\sum_{i=1}^{3} x_i y_i + w_1 [\sum_{i=1}^{3} x_i^2 - \frac{(\sum_{i=0}^{3} x_i)^2}{3}] + \frac{\sum_{i=1}^{3} y_i \sum_{i=0}^{3} x_i}{3} = 0$$

$$w_1 = (\sum_{i=1}^{3} x_i y_i - \frac{\sum_{i=1}^{3} y_i \sum_{i=0}^{3} x_i}{3}) / [\sum_{i=1}^{3} x_i^2 - \frac{(\sum_{i=0}^{3} x_i)^2}{3}]$$

2. 회귀분석

단순회귀

$$y_i = w_0 + w_1 x_i$$

키	몸무게
166	60
170	70
180	80

$$w_1 = \left(\sum_{i=1}^{3} x_i y_i - \frac{\sum_{i=1}^{3} y_i \sum_{i=0}^{3} x_i}{3}\right) / \left[\sum_{i=1}^{3} x_i^2 - \frac{\left(\sum_{i=0}^{3} x_i\right)^2}{3}\right]$$

$$w_0 = \bar{Y} - w_1 \bar{X}$$

$$w_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

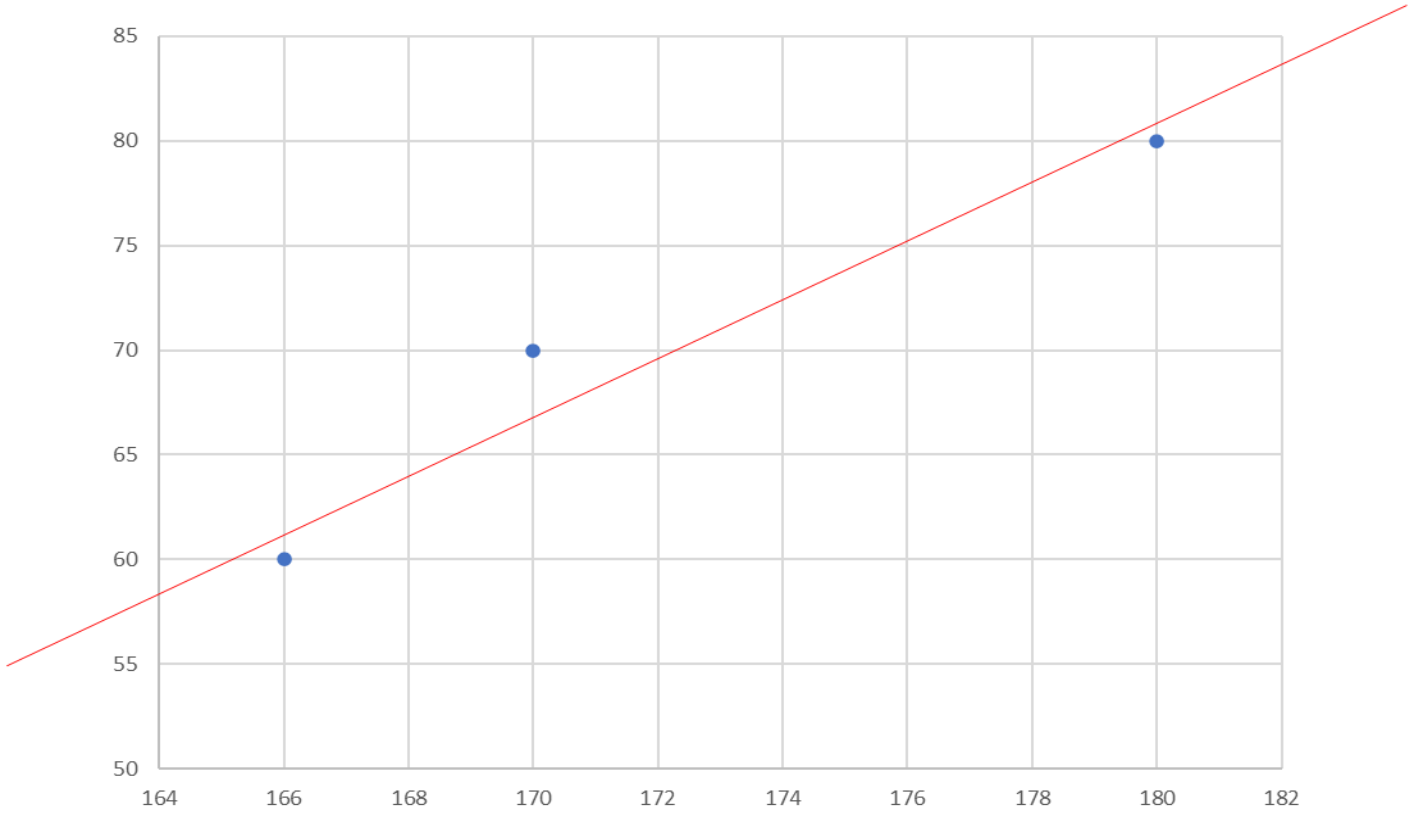
$$y_i = -161.5 + 1.35 x_i$$

2. 회귀분석

단순회귀

$$y_i = -161.5 + 1.35x_i$$

키	몸무게
166	60
170	70
180	80



JEONJU UNIVERSITY

2. 회귀분석

단순회귀 $$y_i = w_0 + w_1 x_i$$

$$w_1 = \left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=0}^{n} x_i}{n}\right) / \left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=0}^{n} x_i)^2}{n}\right]$$

$$w_0 = \bar{Y} - w_1 \bar{X}$$

$$w_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$w_1 = \frac{Cov(x, y)}{Var(x)}$$

2. 회귀분석

단순회귀

$$y_i = w_0 + w_1 x_i$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$w_1 = \frac{S_{xy}}{S_{xx}}$$

$$w_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

$$r = S_{xy} / (\sqrt{S_{xx}}\sqrt{S_{yy}})$$

2. 회귀분석

❖ 단순 회귀분석

$$e_i = y_i - \hat{y}_i$$

➢ 잔차(residual) 의 성질

✓ 잔차의 합은 0

$$\sum_{i=0}^{n} e_i = 0$$

✓ 잔차의 제곱합은 항상 최소

✓ 관찰값과 추정값의 합은 같음

$$\sum_{i=0}^{n} y_i = \sum_{i=0}^{n} \hat{y}_i$$

✓ 잔차와 추정값의 가중합은 항상 0

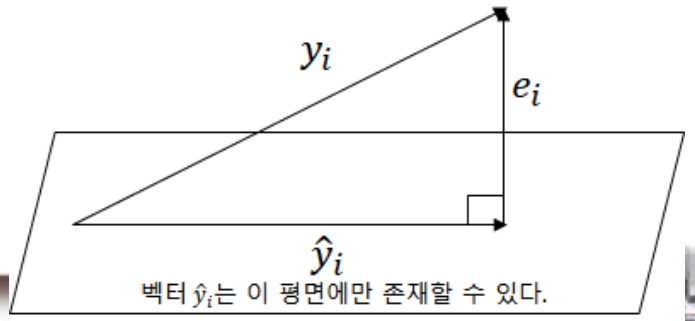
$$\sum_{i=0}^{n} e_i \hat{y}_i = 0$$

Orthogonal principle

✓ 잔차와 x 의 가중합도 항상 0

$$\sum_{i=0}^{n} x_i e_i = 0$$

✓ 점 $(\bar{x}, \bar{y})$는 항상 적합된 회귀직선상에 존재



JEONJU UNIVERSITY

2. 회귀분석

❖ 다중회귀모델

- 설명변수(독립변수)가 2 개 이상인 회귀모형
- 기본가정: 각설명변수는 종속변수와 선형관계에 있음
- 다중분석의 의의: 분석내용 향상
  - 추가적인 독립변수를 도입함으로써 오차항의 값을 줄일 수 있음
  - 단순회귀분석의 단점 극복

❖ 다중회귀모델

- 다중회귀분석 단계
  - ✓ 1단계. 이론의 가정: Multicollinearity 가정의 추가
  - ✓ 2단계. 회귀직선 도출: 최소제곱법
  - ✓ 3단계. 모형의 통계적 유의성: F통계량
  - ✓ 4단계. 회귀계수의 유의성: 개별 회귀계수의 t통계량
  - ✓ 5단계. 모델의 정확도 평가: 수정된 R^2
  - ✓ 6단계. 모델의 적합성 점검
  - ✓ 7단계. 예측

❖ 다중회귀모델

- 1단계. 이론의 가정: Multicollinearity 가정의 추가
  - ✓ 다중공선성이란, 다중선형회귀에서 설명변수(X) 사이에 강한 상관관계가 성립하는 문제
  - ✓ 다중공선선의 문제점
    - 예측값의 분산이 커짐(회귀모형의 적합성이 떨어짐)
    - 다른 중요한 독립변수가 모형에서 제거될 가능성이 높음
    - 결정계수의 값이 과대하게 나타날 수 있음
    - 설명력은 좋은데 예측력이 떨어질 수 있음

2. 회귀분석

❖ 다중회귀모델

➢ 다중 공선성

- ✓ 독립변수 간의 상관관계가 존재하는 것을 의미
- ✓ 다중 공선성을 알아보기 위한 가장 간단한 지표
  - 독립변수들간의 상관관계 조사
  - 특히, 상관계수가 0.9 이상
- ✓ 공차한계(tolerance)와 분산팽창요인(VIF)
  - 변수 i 의 공차한계(1-R_i^2): 한 독립변수가 다른 독립변수들에 의해서 설명되지 않는 부분을 의미(0.1이하)
  - VIF(variance inflation factor, 분산확대인자): 공차한계의 역수(10 이상일 경우)
- ✓ 공진성 진단의 상태지수: 15이상 (독립변수 개수 검토)

2. 회귀분석

❖ 다중회귀모델

- ➢ 다중 공선성
  - ✓ 독립변수간에 상관관계가 높아서 회귀계수의 추정이 불안정하게 이루어지는 현상
  - ✓ 제거방법
    - 덜 중요한 변수를 제거
    - 독립변수들의 결합(요인분석, 단순 평균화)
    - 표본의 수를 많이 뽑음
  - ✓ 최소표본수
    - 일반적으로 변수 수의 10배 정도가 적절
    - 표본수에 비해 과도한 독립변수가 있는 경우 Lasso, Ridge regression 사용

❖ 다중회귀모델

➢ 모형

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

➢ 관심

- ✓ 회귀계수의 추정
- ✓ 유의성 검정
  - 어떤 독립변수가 종속변수를 설명하는가
- ✓ 변수선택(모형의 선택)

❖ 다중회귀모델

➢ 모형

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & & x_{2k} \\ & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

❖ 다중회귀모델

➢ 회귀선의 추정

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2$$

$$\hat{\beta} = \arg\min \ (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta})$$

$$S = (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta})$$

$$= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta$$

$$= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

❖ 다중회귀모델

➢ 회귀선의 추정

$$\frac{\partial S}{\partial \beta} = -2X^T y + 2X^T X\beta$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = \mathrm{Hy}$$

❖ 다중회귀모델

- 수정결정계수
  - ✓ 독립변수를 추가하면 결정계수 R제곱은 증가함(SSE 감소)
  - ✓ 결정계수가 증가했다고 좋은 모형으로 판단하기 어려움(유의성)
  - ✓ 오차제곱합(SSE)을 줄이는 변수가 아닌 평균오차제곱합(MSE)을 줄이는 변수를 추가
  - ✓ 수정결정계수가 최대일때 최적의 모형이라고 판단함

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$adj.R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST} = 1 - (n-1)\frac{MSE}{SST}$$

Thank You !

JEONJU UNIVERSITY
The Place for Superstars