Simple Linear Regression

두 변수 사이의 관계

- 대략적 파악 : 산점도(scatter plot)
- 상관분석(correlation analysis)
  - 두 변수 사이의 상관관계 분석
  - 확률변수 $X, Y \to \rho = \mathrm{Corr}(X, Y)$ - 직선적인 관련성 파악
- 회귀분석(regression analysis)
  - 두 변수 사이의 함수관계를 분석
  - x : 독립변수 또는 설명변수, Y : 종속변수 또는 반응변수

    $$Y = f(x) + \varepsilon,\ \varepsilon : \text{오차항} \to f(x)?$$
  - 단순선형회귀분석 - 직선관계를 모형으로 분석

    $$(f(x) = a + bx)$$
  - 중회귀분석 - 두 개 이상의 설명변수 사용

    $$(f(x) = a + b_1 x_1 + \cdots + b_k x_k)$$

Example

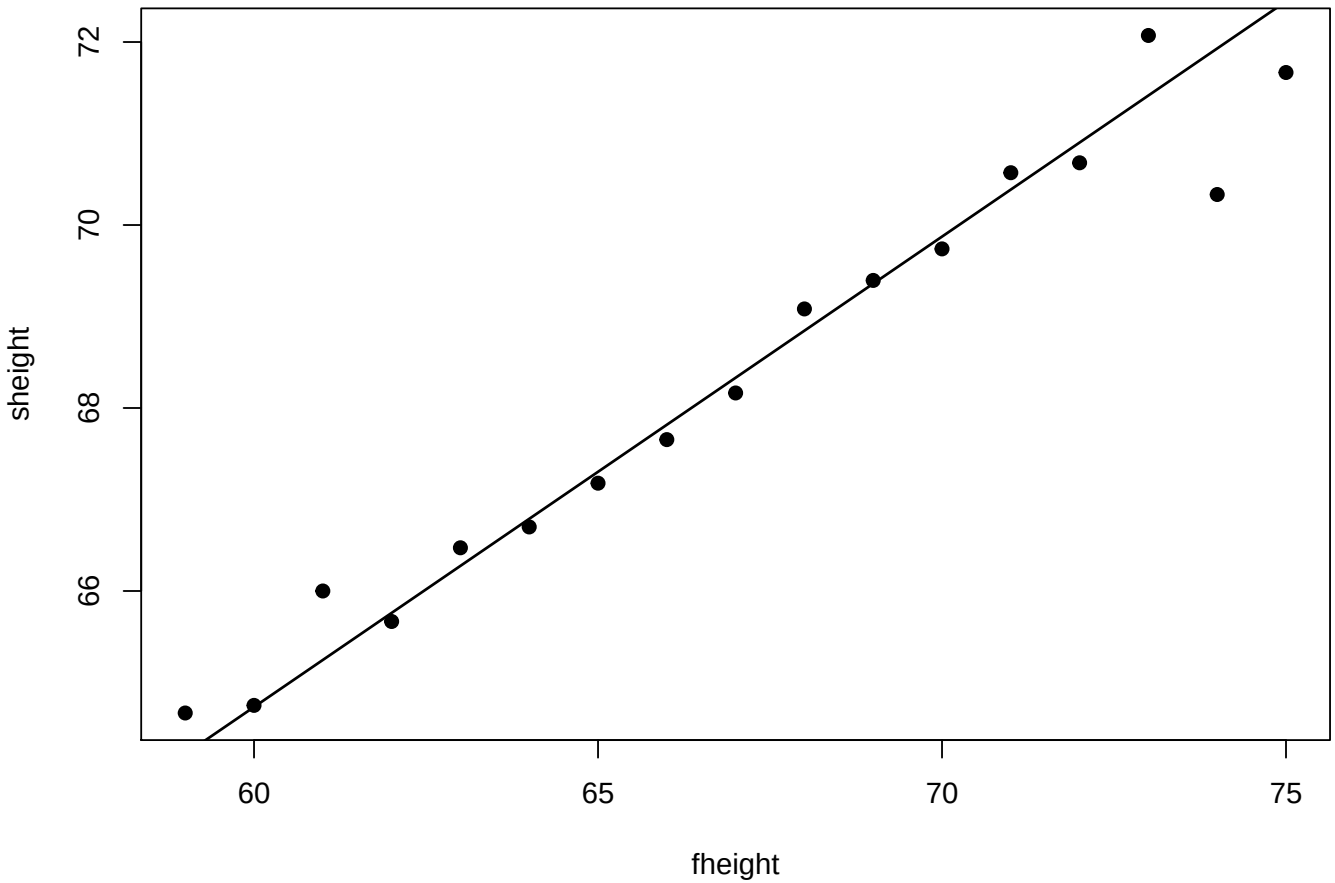


Figure: 아버지의 키(fheight)와 아들 키(sheight)간의 회귀직선

회귀분석의 기본개념

- 자료구조

  - 자료구조 : $(x_1, Y_1), ..., (x_n, Y_n)$
  - $(x_1, \ldots, x_n)$: 설명변수(explanatory variable)(또는 독립변수)

    두 변수가 있을 때, 다른 한 변수에 영향을 주는 변수
  - $(Y_1, \ldots, Y_n)$: 반응변수(response variable)(또는 종속변수)

    두 변수가 있을 때, 다른 한 변수에 영향을 받는 변수
  - 관측값 : $(x_1, y_1), \ldots, (x_n, y_n)$

Example

상점번호	광고료	총판매액	상점번호	광고료	총판매액
1	4	9	6	12	30
2	8	20	7	6	18
3	9	22	8	10	25
4	8	15	9	6	10
5	8	17	10	9	20

Table: 표본상점의 광고료(단위:10만원)와 총판매액(단위:100만원)

기본 가정

- Linearity (선형성) : $E(Y|X=x) = \mu_{y \cdot x} = \beta_0 + \beta_1 x$
- Homoscedastic (등분산성) : $Var(Y|X=x) = \sigma^2$
- Normality (정규성) : $Y|X=x \sim N(E(Y|X=x), \sigma^2)$
- Independency (독립성) : ϵ's are mutually independent

Example

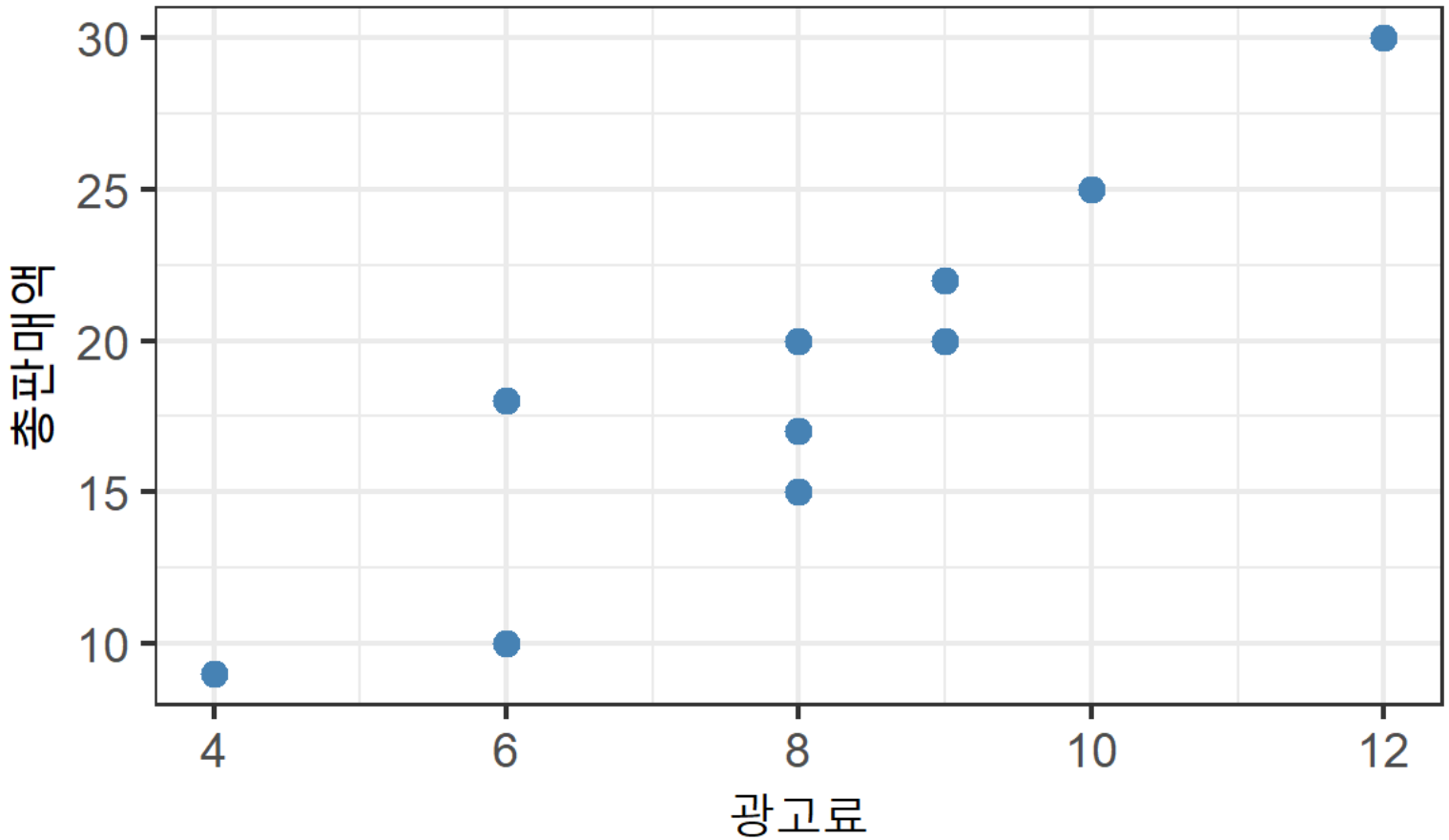


Figure: 광고료와 총판매액의 산점도

단순선형회귀 모형

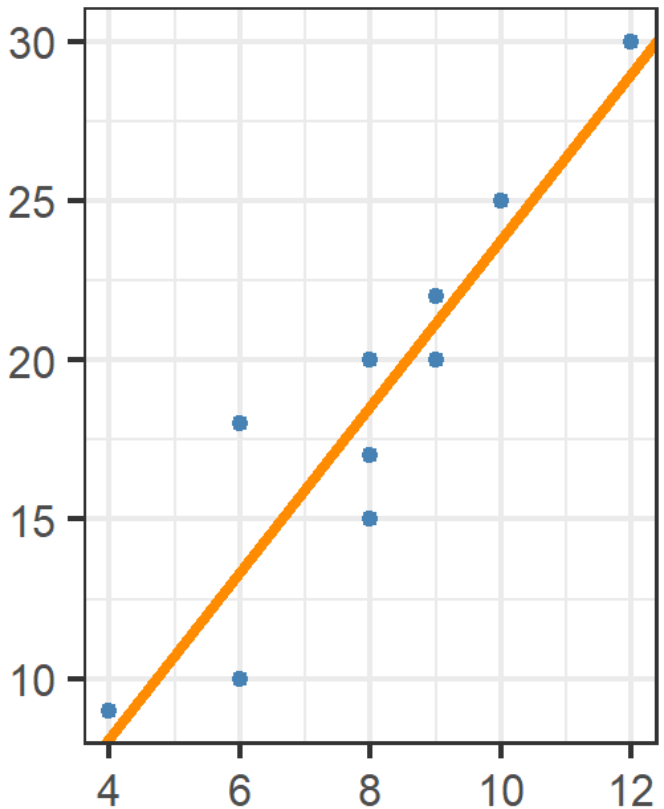
Model

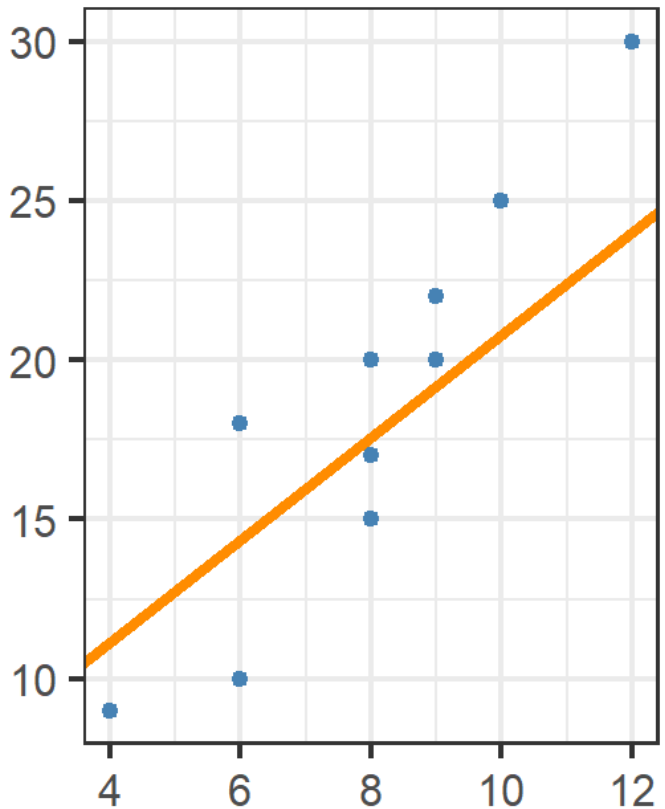
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

- $(\epsilon_1, \ldots, \epsilon_n)$: 오차항(random error)

  서로 독립이면서 평균이 0, 분산이 σ^2 인 확률 변수
- 회귀계수(regression coefficient) (or 모수, parameter)
  - β_0 : 상수항 또는 절편 (constant coefficient or intercept)
  - β_1 : 기울기 (slope)
- 회귀직선, 회귀선 : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - $\hat{\beta}_0, \hat{\beta}_1, \hat{y}$: estimate of $\beta_0, \beta_1, E(Y|X = x)$

Example - 회귀직선의 비교





Least Square Estimation (LSE)

- 오차제곱합

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

- 최소제곱추정량(LSE)

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

- Least square fit : $\hat{y} \left(\equiv \widehat{E(Y|X=x)} \right) = \hat{\beta}_0 + \hat{\beta}_1 x$

Least Square Estimation (LSE)

- 정규방정식 (normal equation)

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Least Square Estimation (LSE)

- 최소제곱추정량 = 정규방정식의 해

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Least Square Estimation (LSE)

- 최소제곱추정량(LSE)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{(xy)}}{S_{(xx)}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

  - $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Example - LSE

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	4.00	9.00	-4.00	-9.60	16.00	92.16	38.40
2	8.00	20.00	0.00	1.40	0.00	1.96	0.00
3	9.00	22.00	1.00	3.40	1.00	11.56	3.40
4	8.00	15.00	0.00	-3.60	0.00	12.96	-0.00
5	8.00	17.00	0.00	-1.60	0.00	2.56	-0.00
6	12.00	30.00	4.00	11.40	16.00	129.96	45.60
7	6.00	18.00	-2.00	-0.60	4.00	0.36	1.20
8	10.00	25.00	2.00	6.40	4.00	40.96	12.80
9	6.00	10.00	-2.00	-8.60	4.00	73.96	17.20
10	9.00	20.00	1.00	1.40	1.00	1.96	1.40
sum	80	186	0	0	46	368.4	120

Example - LSE

- 최소제곱추정량 (LSE)

$$\hat{\beta}_1 = \frac{120}{46} = 2.6087,$$

$$\hat{\beta}_0 = 18.6 - 2.6087 \times 8 = -2.2696$$

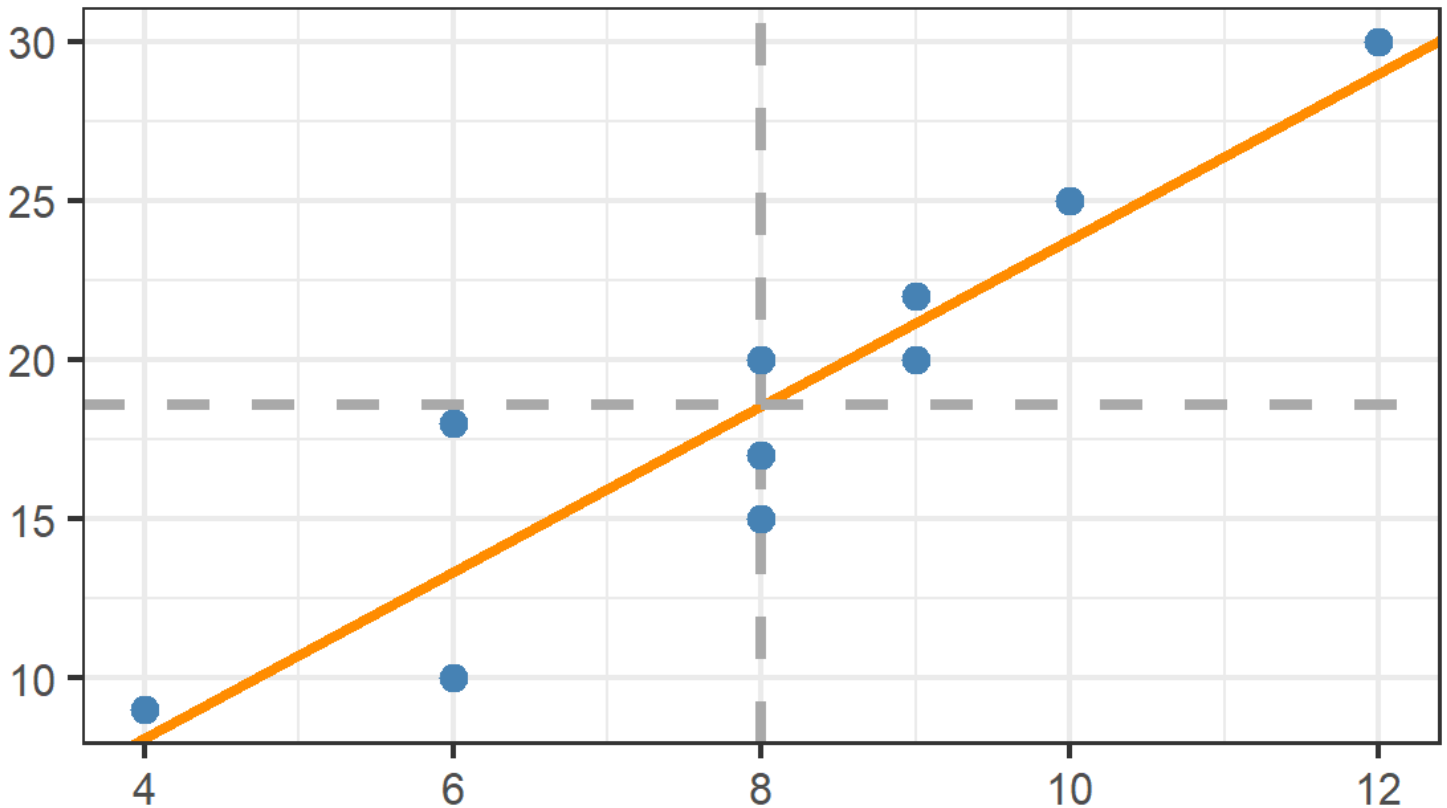
- 추정된 회귀직선: Least square fit

$$\hat{y} = -2.2696 + 2.6087 \cdot x$$

Properties of fitted regression line

- 잔차 (residual) : $e_i = y_i - \hat{y}_i, \;\; i = 1, 2, \ldots, n$
  - (1) 잔차의 합은 0이다. $(\sum_{i=1}^{n} e_i = 0)$
  - (2) $\Sigma_{i=1}^{n} e_i^2$ 은 최소값을 갖는다.
  - (3) 잔차의 x_i에 의한 가중합은 0이다. $(\sum_{i=1}^{n} x_i e_i = 0)$
  - (4) 잔차의 $\hat{y}_i$에 의한 가중합은 0이다. $(\sum_{i=1}^{n} \hat{y}_i e_i = 0)$
  - (5) $(\bar{x}, \bar{y})$는 적합된 회귀직선 위에 있다.

Example



Estimation of error variance

- 오차분산 (σ^2)의 추정:
  - 잔차(오차) 제곱합 (residual (or error) sum of squares) :

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

  - 평균제곱오차 (mean squared error) : $MSE = \dfrac{SSE}{n-2}$
  - 오차분산의 추정값 : $\hat{\sigma}^2 = MSE$

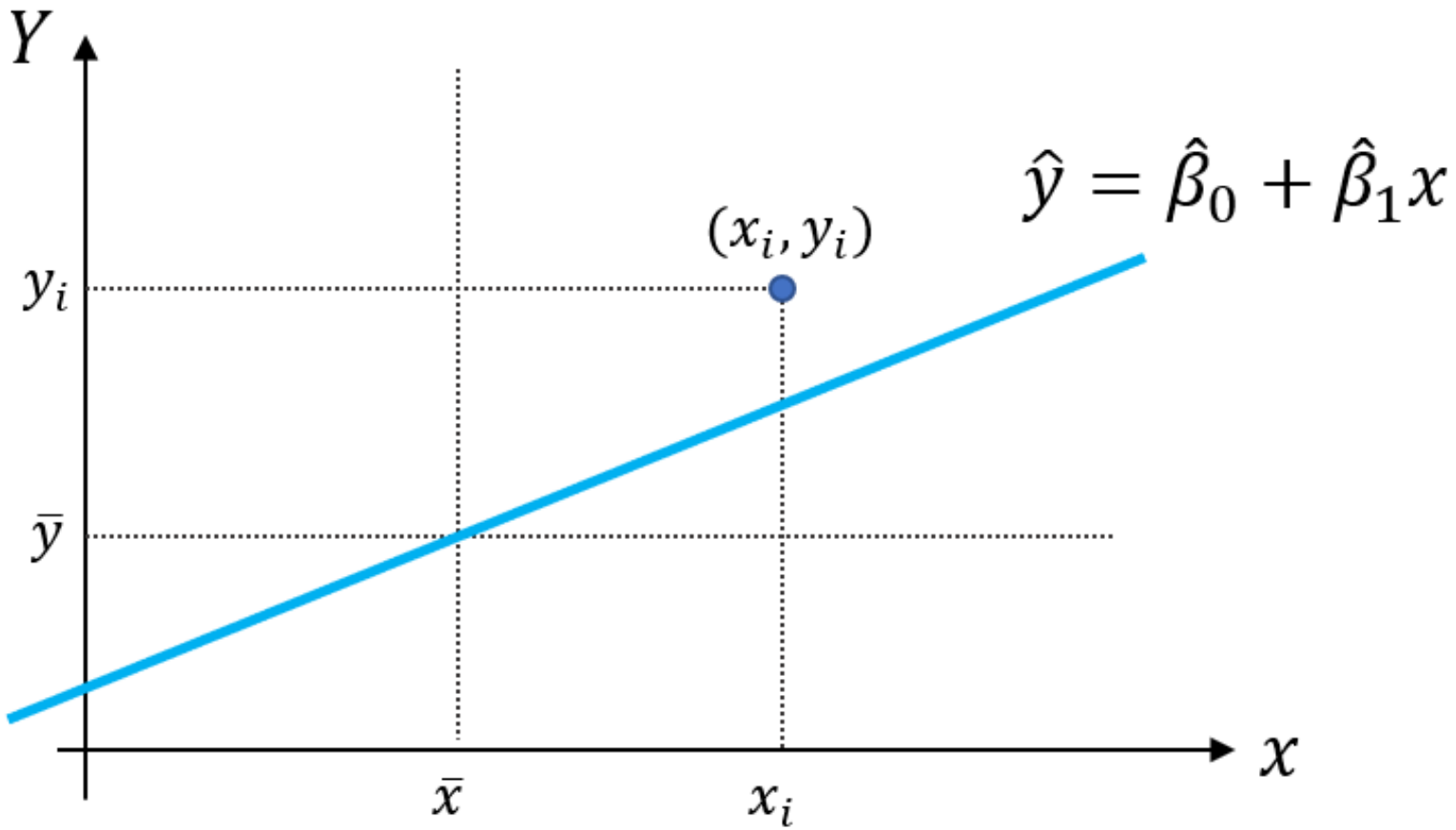
Decomposition of deviations

- 총편차의 분해
  - $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}), \ \ \forall i$
  - 총편차(total deviation) $= y_i - \bar{y}$
  - 추측값의 편차 $= (\hat{y}_i - \bar{\hat{y}}) = (\hat{y}_i - \bar{y})$

    $\Rightarrow$ 총편차 $=$ 잔차 $+$ 추측값의 편차

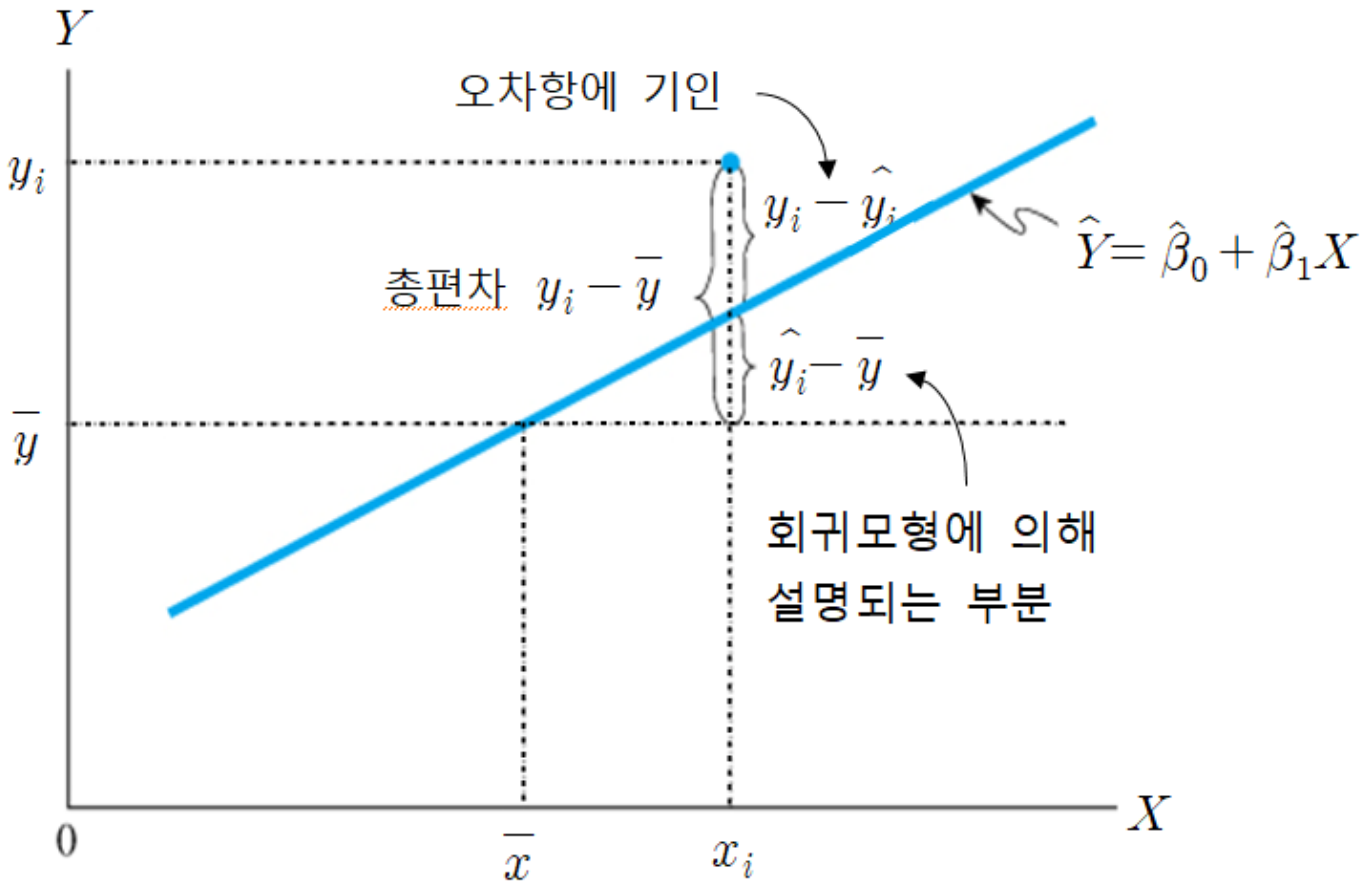
Decomposition of deviations

Figure: 편차의 분해



Decomposition of deviations

Figure: 편차의 분해



Decomposition of sum of squares

- 제곱합의 분해 : SST $=$ SSE $+$ SSR

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

제곱합의 종류	정의 및 기호
총제곱합 (total sum of squares)	$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$
잔차제곱합 (residual sum of squares)	$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
회귀제곱합 (regression sum of squares)	$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

Coefficient of determination

- 결정계수 (Coefficient of determination)
  - 정의 : $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$
  - 의미 : 회귀직선의 기여율

    (총변동 가운데 회귀직선으로 설명되는 변동의 비율)
  - 성질
    - $0 \leq R^2 \leq 1$
    - R^2 값이 1에 가까울수록 회귀에 의한 설명이 잘 됨을 뜻함
    - $R^2 = r^2$ (r : sample correlation)

      (단순선형회귀모형에서만 성립)

상관분석

- X, Y : random variables
- 모상관계수 (population coefficient of correlation)

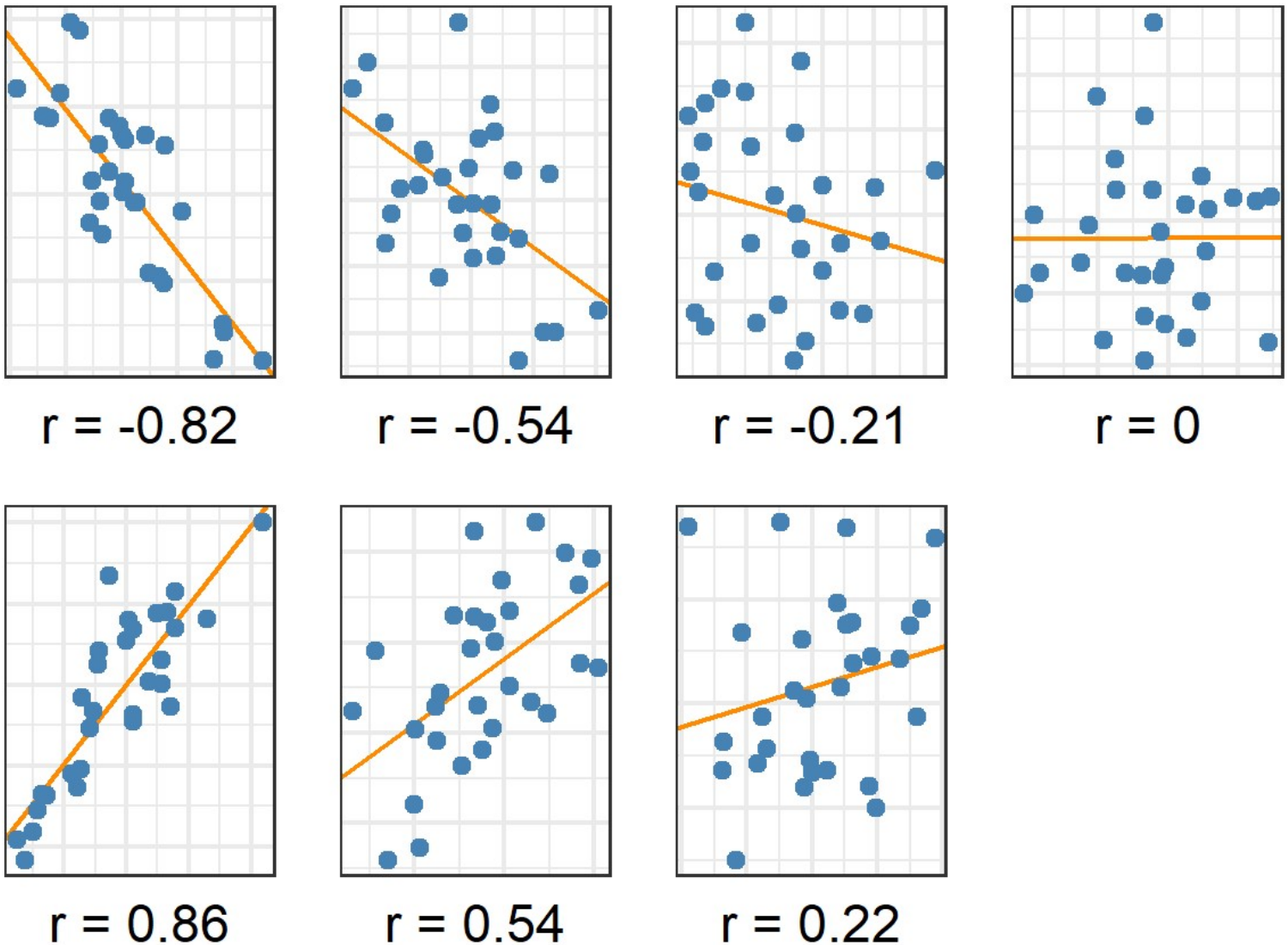
$$\rho_{XY} = \frac{\mathrm{Cov(X, Y)}}{\sqrt{\mathrm{Var(X)Var(Y)}}} := \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- $(x_1, y_1), \ldots, (x_n.y_n)$: sample
- 표본상관계수 (sample coefficient of correlation)

$$r_{xy} = \frac{S_{(xy)}}{\sqrt{S_{(xx)} S_{(yy)}}}$$

- $-1 \leq \rho \leq 1, \ -1 \leq r_{xy} \leq 1$

표본상관계수와 산점도



Example

- 광고료와 총판매액
  - 표본상관계수

$$r_{xy} = \frac{S_{(xy)}}{\sqrt{S_{(xx)}S_{(yy)}}} = \frac{120}{\sqrt{46 \times 368.4}} = 0.92$$

표본상관계수와 단순선형회귀모형

- 단순선형회귀모형에서는
  - 표본상관계수와 결정계수

$$R^2 = r_{xy}^2$$

  - 표본상관계수와 회귀직선의 기울기

$$r_{xy} = \hat{\beta}_1 \frac{s_x}{s_y}$$

s_x, s_y : 표본표준편차 (sample standard deviation)

$$s_x = \sqrt{\frac{S_{(xx)}}{n-1}}, \quad s_y = \sqrt{\frac{S_{(yy)}}{n-1}}$$

분산분석

- 단순회귀직선의 유의성 검정을 위한 분산분석표

요인	제곱합(SS)	자유도(df)	평균제곱(MS)	F_0	유의확률
회귀	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$	$P(F \geq F_0)$
잔차	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
계	SST	$n-1$			

- $F \sim F(1, n-2)$
- $F_0 > F(1, n-2; \alpha) \Rightarrow$ 유의수준 α 하에서, 회귀선이 유의함.