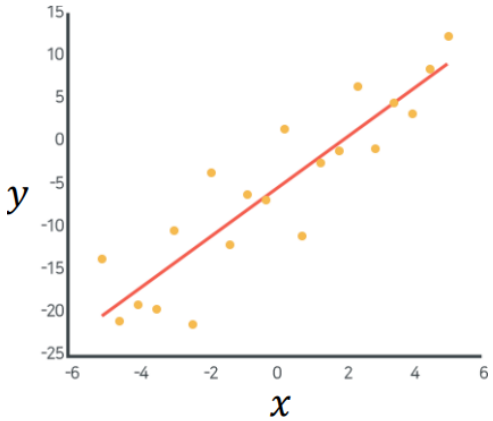
10.단순회귀분석

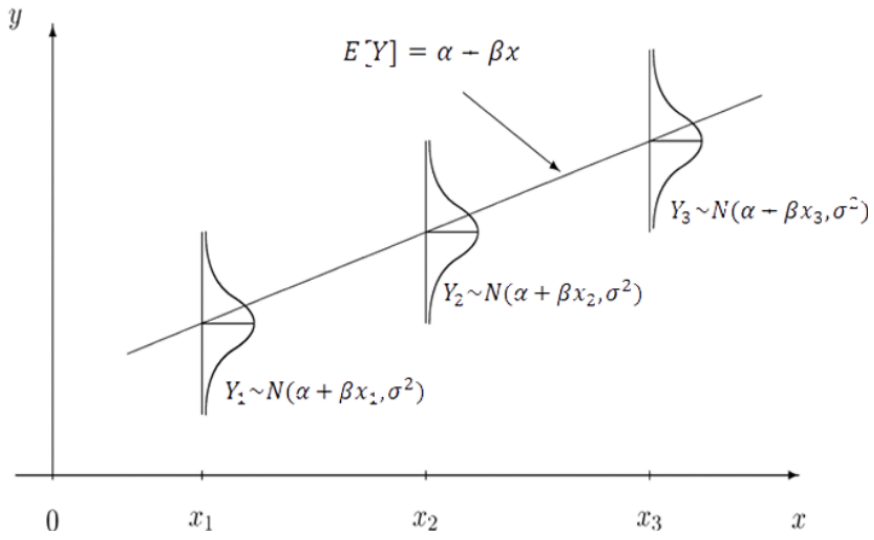
- 단순 선형 회귀(Simple Linear Regression) :
  특성변수 X 하나만 가지고, 연속형 종속변수 Y를 예측하는 모델
- 회귀분석 개요
  - 회귀분석
    - 독립변수와 종속변수 간의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 자료로부터 통계적으로 추정하는 분석방법.
    - $y=f(x)$의 함수 관계가 있을 때,
      - x를 설명변수(explanatory variable)
        또 는독립변수(independent variable)
        또는 예측변수, 특성 변수
        - 단순 회귀 : 독립변수가 1개
        - 다중 회귀 : 독립변수가 2개 이상
      - y를반응 변수(response variable)
        또는종속 변수(dependent variable)
        또는 목표변수, 타겟변수



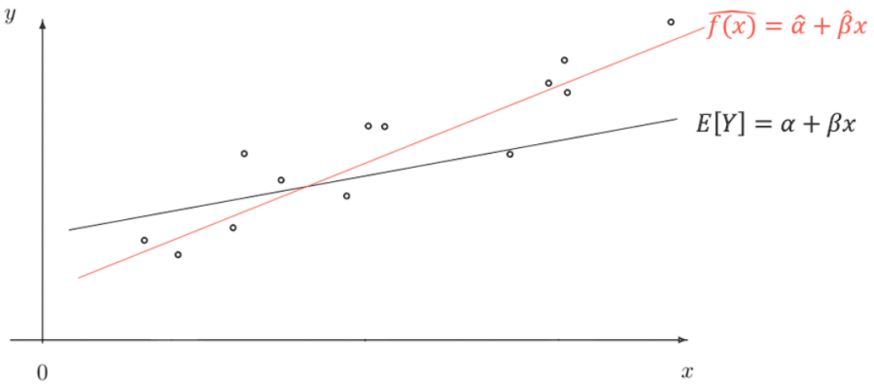
- 단순선형회귀모형
  - 모형 정의 및 가정
    - 자료(x_i, Y_i), i=1,⋯, n 에 다음의 관계식이 성립한다고 가정함

$Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, ..., n$

- 오차항인$\epsilon_1, \epsilon_2, \cdots, \epsilon_n$는 서로 독립인 확률변수로, ϵ_i $N[0, \sigma^2]$:정규, 등분산, 독립가정
- α, β는 회귀계수라 부르며 α는 절편, β는 기울기를 나타냄
- α,β, σ^2은 미지의 모수로, 상수임.
- 자료(x_i, Y_i), i=1,⋯, n 에 다음의 관계식이 성립한다고 가정함.
  Y_i $N[\alpha + \beta x_i, \sigma^2] \rightarrow E[Y_i] = \alpha + \beta x_i$

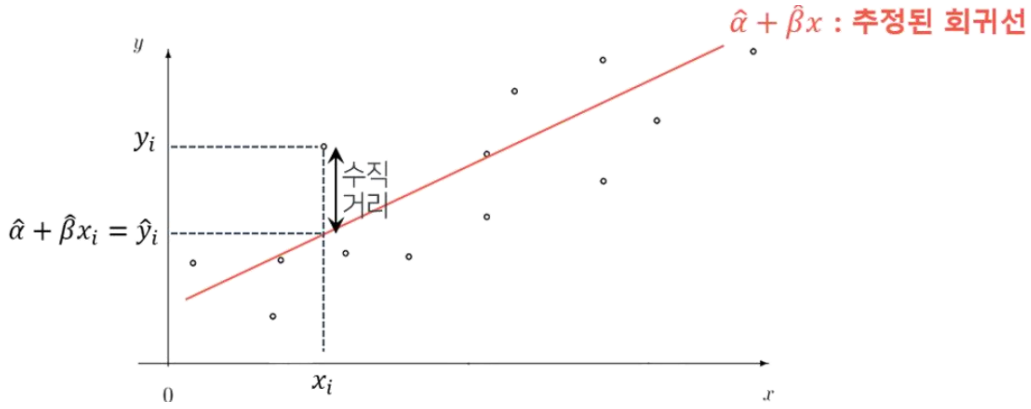


- 단순 선형회귀모형의 모수 추정
  - 모수 추정
    - 모형이 포함한 미지의 모수 α,β 를 추정하기 위하여 각 독립변수 x_i에 대응하는 종속변수 y_i로 짝지어진 n 개의 표본인 관측치 (x_i, y_i) 가 주어짐



  - 최소제곱법

- 단순회귀모형 $Y_i = \alpha + \beta x_i + \epsilon_i$에서 자료점과 회귀선 간의 수직거리 제곱합 $SS(\alpha, \beta) = \Sigma_i^n (y_i - \alpha - \beta_i)^2$이 최소가 되도록 α와 β를 추정하는 방법



  - α에 대한 최소제곱 추정량 : $\widehat{\alpha} = \widehat{y} - \widehat{\beta}\overline{x}$
  - β에 대한 최소제곱 추정량 : $\widehat{\beta} = \frac{\Sigma_i^n x_i(y_i - \overline{y})}{\Sigma_i^n x_i(x_i - \overline{x})}$
    (단, $\overline{x}$는x_i의평균, $\overline{y}$는y_i의평균)
  - y_i의 추정치 : $\widehat{y_i} = \widehat{\alpha} + \widehat{\beta}x_i, i = 1, 2, ..., n$
  - 잔차 : $e_i = y_i - \widehat{y_i} = y_i - \widehat{\alpha} - \widehat{\beta}x_i, i = 1, 2, ..., n$
    - 오차 ϵ과 대응되는 개념으로써 오차는 확인이 불가능하지만 잔차는 확인이 가능

- 단순 선형회귀모형의 유의성 검정
  - 모형의 유의성 t 검정
    - 독립변수 x가 종속변수 Y를 설명하기에 유용한 변수인가에 대한 통계적 추론은 회귀계수 β에 대한 검정을 통해 파악할 수 있음.
    - 가설

      H_0 : $\beta = 0$
      H_1 : $\beta \neq 0$
    - 검정통계량과 표본분포
      - 귀무가설 H_0 이 사실일 때,
        $T = \frac{\widehat{\beta}}{S.E.[\widehat{\beta}]} \sim t[n-2]$
        $|T| = |\frac{\widehat{\beta}}{S.E.[\widehat{\beta}]}| > t_{\alpha/2, n-2}$ 또는 p-value $(= P[T > |t_0|] \times 2) < \alpha$ 면

귀무가설을 기각
→ 독립변수x가종속변수Y를설명하기에유용한변수라고해석할수있음.

> $t\alpha$ 2,$n-2$ 또는 p-value (=P[T>|t0|] × 2) < α면
> → 독립변수x가종속변수Y를설명하기에유용한변수라고해석할수있음.

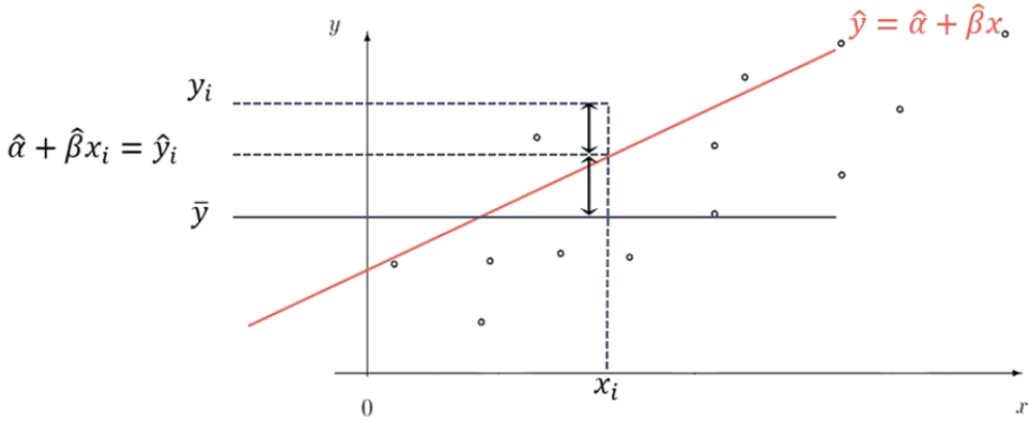
- 단순선형회귀모형 사례
  - 예제
    - 베타에 관한 유의성을 유의수준 5%로 검정할 것.

	추정치	표준오차	T 통계량	p-value
절편	-17.5791	6.7584	-2.601	0.0123
주행속도	3.9324	0.4155	9.464	1.49E-12

    - 가설 H_0 : $\beta = 0$
      H_1 : $\beta \neq 0$
    - 귀무가설(H_0)이사실일때,
    - 검정통계량의관찰값(x_0)는9.464로t[48]의분포에서 유의확률은1.49E-12
    - p-value(=1.49E-12) ≤ α(=0.05) 이므로, H_0를기각
- 단순선형회귀모형의 적합도
  - Y의 변동성 분해
    - 제곱합:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SSE}}$$

(y_i의 변동) (모형으로 설명되는 변동) (모형으로 설명되지 않는 변동)



- 모형의 적합성
  - 결정계수 R^2
    $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
  - SST=SSR+SSE이므로 항상 0과1 사이의 값을 가짐(0≤R^2≤1).
  - y_i의 변동 가운데 추정된 회귀모형으로 통해 설명되는 변동의 비중을 의미함.
  - 0에 가까울수록 추정된 모형의 설명력이 떨어지는 것으로,
    1에 가까울수록 추정된 모형이 y_i의 변동을 완벽하게 설명하는 것으로 해석할 수 있음.
  - R^2는 두 변수 간의 상관계수 r의 제곱과 같음.