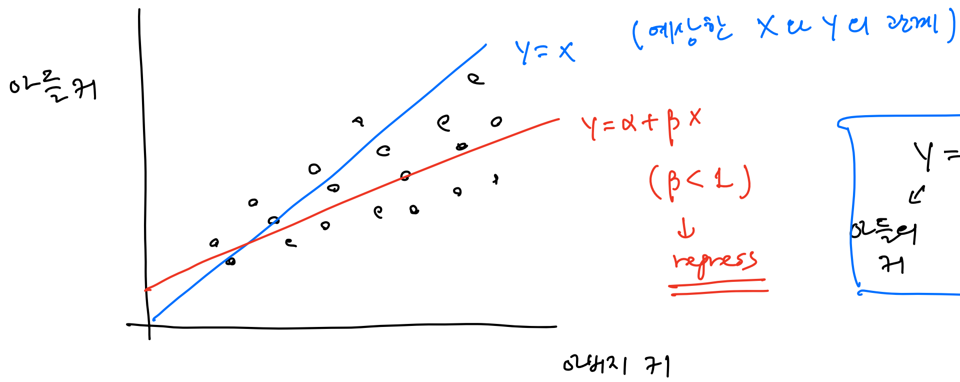
○ 회귀분석 (regression analysis) 역사

Francis Galton : (아버지 키$_i$, 아들 키$_i$) $i=1, 2, \dots, n$



$$Y = \beta_0 + 1 \cdot X$$

아들의 키 ← Y, X → 아버지 키

○ 단순 선형 회귀

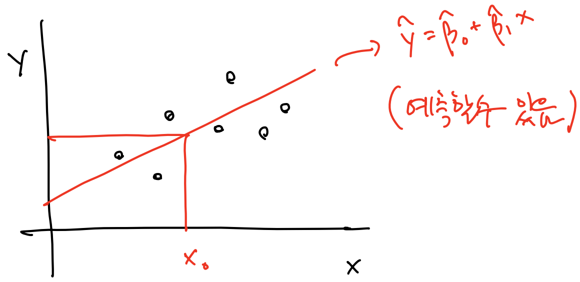
모형 : $Y = \beta_0 + \beta_1 X + \varepsilon$, $E(\varepsilon) = 0$

$Var(\varepsilon) = \sigma^2$

$\beta_0, \beta_1 \Rightarrow$ 회귀계수 (regression coefficient)

목적 ① β_0, β_1 추정 $\Rightarrow$ 예측 (prediction)

↳ $\hat{\beta}_0, \hat{\beta}_1 \Rightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



② $H_0 : \beta_1 = 0$ 검정 $\Rightarrow$ X와 Y의 관계를 규명

↳ under H_0 (H_0 is true) $\Rightarrow$ 모형 $Y = \beta_0 + \varepsilon$

$\Rightarrow$ Y는 X와 관계가 없다.

* 일반적으로 설명변수 X는 확률변수가 아니고 fixed & known

그래서 $E(Y) = E(\beta_0 + \beta_1 X + \varepsilon) = \beta_0 + \beta_1 X + E(\varepsilon)$

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

0

$= \beta_0 + \beta_1 x$

$=$ true regression line

○ Back to model

$\varepsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1, \dots, n$ $\qquad E(\varepsilon_i) = 0$ & $Var(\varepsilon_i) = \sigma^2$ for all i.

Define $\boxed{\text{fitted values}}$ and $\boxed{\text{residuals}}$ $\qquad \hat\beta_0, \hat\beta_1$

$\hat y_i = \hat\beta_0 + \hat\beta_1 x_i, \quad i=1,2,\dots,n$ (fitted values)

$e_i = y_i - \hat y_i, \quad i=1,2,\dots,n$ (residuals)

Note

$y_i = \boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i$

$y_i = \boxed{\hat\beta_0 + \hat\beta_1 x_i} + e_i$

즉 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ 라는 조건을

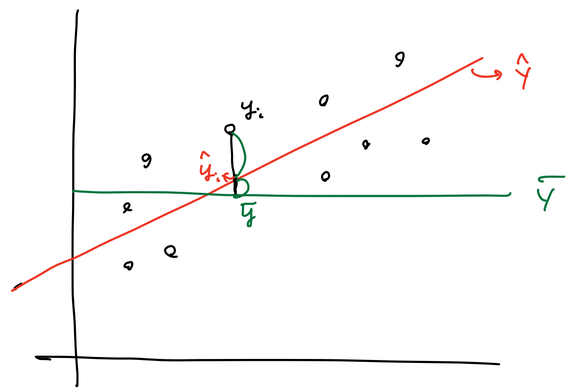
e_i 가 만족하는지 check 해야 한다

// residual analysis (잔차분석)

○ 제곱합 분해, R^2 (결정 계수)

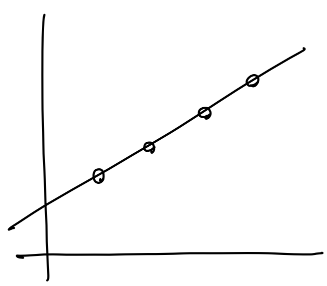
$$\sum_{i=1}^{n} (y_i - \bar y)^2 = \sum_{i=1}^{n} (y_i - \hat y_i)^2 + \sum_{i=1}^{n} (\hat y_i - \bar y)^2$$

$$SS_T = SS_E + SS_R$$



$$R^2 = \frac{SS_R}{SS_T}$$

* $R^2 = 1$ 인 경우 $\qquad SS_E = 0$



* R^2의 문제점

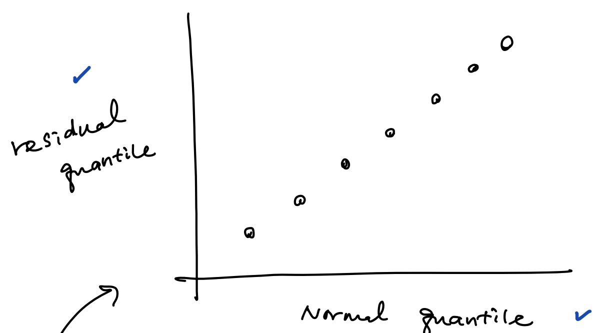
F-stat. $\quad F_0 = \dfrac{MS_R}{MS_E} = \dfrac{SS_R / df_R}{SS_E / df_E} \underset{H_0}{\sim} F_{df_R, df_E}$

$F_0 \uparrow \Rightarrow$ reject $H_0 : \beta_1 = 0 \iff$ t-test

$\Rightarrow$ 그래서 Multiple regression model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, $H_0: \beta_1 = \beta_2 = 0$ use F-test.

- Normal probability plot



$\varepsilon_i \sim Normal$

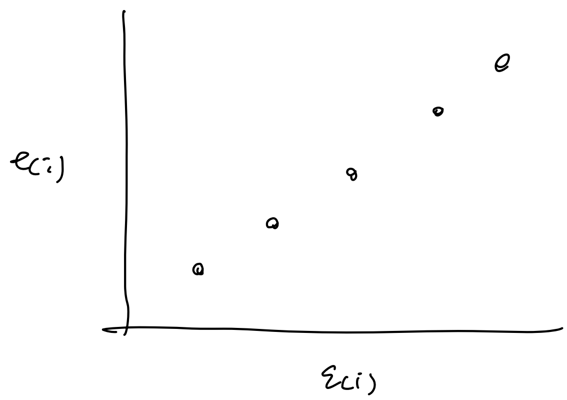
⇒ 즉 Normal 분포를 따르는 데이터 $\varepsilon_1, \dots, \varepsilon_n$ 추출한 후

$\varepsilon_{(1)} < \varepsilon_{(2)} < \dots < \varepsilon_{(n)}$ 순서

⇒ 얻은 잔차 $e_1, \dots, e_n$에 대해

$e_{(1)} < e_{(2)} < \dots < e_{(n)}$ 순서

만약 $e_1, \dots, e_n$이 Normal 분포를 따른다면



- 범주형 설명변수

만약 X가 범주형 변수 (여자, 남자) 일때, 수치화가 필요하면 이를 위해

$$X = \begin{cases} 1 & \text{if 여자} \\ 0 & \text{if 남자} \end{cases}$$

Back to iris data.

붓꽃 종류 (species) → (setosa, versicolor, virginica) 범주형 자료

[수치화 하는 방법]

$$X_1 = \begin{cases} 1 & \text{if versicolor} \\ 0 & \text{o.w} \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if virginica} \\ 0 & \text{o.w.} \end{cases}$$

	X_1	X_2
Setosa	0	0

Versicolor	1	0
Virginica	0	1

꽃받침 너비

모형 : $y(\text{Sepal.width}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$$E(Y \mid \text{Setosa}) = \beta_0$$
$$E(Y \mid \text{versicolor}) = \beta_0 + \beta_1$$
$$E(Y \mid \text{virginica}) = \beta_0 + \beta_2$$

$H_0: \beta_1 = \beta_2 = 0 \Rightarrow$ under H_0, 3종류 붓꽃의 꽃받침 너비는 차이가 없다.

만약에 $X_1 = \begin{cases} 1 & \text{if Setosa} \\ 2 & \text{if Versicolor} \\ 3 & \text{if virginica} \end{cases}$ $y = \beta_0 + \beta_1 X_1 + \varepsilon$

$$E(Y \mid \text{Setosa}) = \beta_0 + \beta_1$$
$$E(Y \mid \text{Versicolor}) = \beta_0 + 2\beta_1$$
$$E(Y \mid \text{virginica}) = \beta_0 + 3\beta_1$$

β_1 차이

β_1 차이

?

o 분산분석 (Analysis of variance, ANOVA)

금속가공품 인장강도 (반응변수, 특성값) : y

⇒ 여러개의 공법 1, 2, . . . , K

$$\begin{array}{ccccc} & 1 & 2 & \cdots & K \\ \text{반복} & y_{11} & y_{21} & & y_{K1} \\ & \vdots & \vdots & & \vdots \\ & y_{1n} & y_{2n} & & y_{Kn} \\ \hline & \bar{y}_{1\cdot} & \bar{y}_{2\cdot} & & \bar{y}_{K\cdot} \end{array}$$

K×n개의 y값

$\bar{y}_{\cdot\cdot} = \frac{1}{nK} \sum_i \sum_j y_{ij}$ 전체평균

= grand average

$$\sum_i\sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i\sum_j (y_{ij} - \bar{y}_{i.})^2 + n\sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2$$ 분해

$$SS_T = SS_W \text{ (공법내)} + SS_B \text{ (공법간)}$$

반응변수의 변동량
(고정된 값이다.)

변동량 SS_E i번째 공법 내애서의 변동량

변동량 SS_{trt}

공법에 의해 설명되는 변동량
i번째 공법으로 인한..
공법간 차이가 크다면 값이 커지고,
공법간 차이가 없다면 값이 작아질 것이다.

$\Rightarrow$ $SS_B \uparrow$ 공법간 차이가 크다

$SS_B = 0$ 〃 없다

○ 일원배치법 모형

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \iff y_{ij} = \mu_i + \varepsilon_{ij}$$

μ : 전체 평균 $\qquad \mu_i \triangleq \mu + \tau_i$

τ_i : i번째 처리 효과 $\qquad \hat{\mu} + \hat{\tau}_i$

$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ $\longrightarrow$ 잔차분석이 필요함 $e_{ij} = y_{ij} - \hat{y}_{ij}$

목적 : $H_0 : \tau_i = 0$ for all i $\qquad$ $H_0 : \mu_1 = \cdots = \mu_k = \mu$

Under $H_0 \Rightarrow$

$$y_{ij} = \mu + \varepsilon_{ij}$$

$$E(Y_{ij}) = \mu \quad (\because E(\varepsilon_{ij}) = 0)$$

= 모든 처리 수준에서 특성값이 같다

= 공법간 차이가 없다.

$\Rightarrow$ F-test

ANOVA table

SV	SS	DF	MS	F_0		
Trt	SS_{trt}	$k-1$	$MS_{trt} = \frac{SS_{trt}}{k-1}$	$F_0 = \frac{MS_{trt}}{MSE}$	p-value 이용	H_0
Err	SS_E	$N-k$	$MS_E = \frac{SS_E}{N-k}$			
Total	SS_T	$N-1$				