

〈 회귀 분석 〉

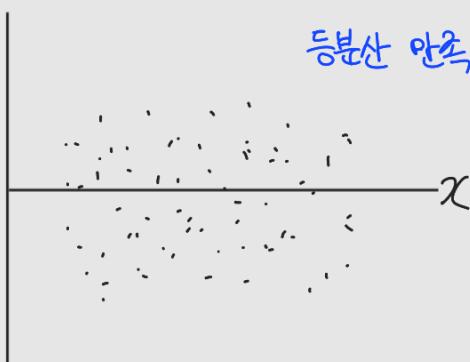
회귀분석 : 주어진 자료를 통해 변수 간의 함수관계를 밝히고, 이 함수관계를 이용하여 독립변수 값에 대응되는 종속변수의 값을 예측 또는 설명하는 분석 방법

회귀분석의 기본 가정 검토

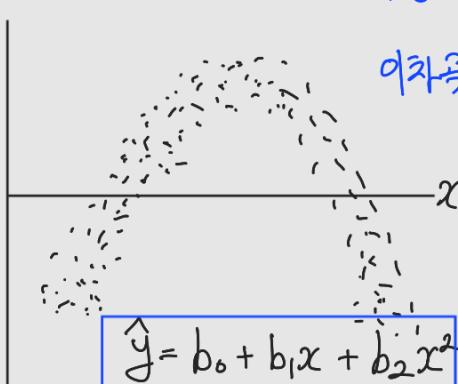
① 오차항의 정규성 검토 : 정규확률도표 (P-P plot)

② 등분산성 검정 : 잔차들의 산점도

⑦



㉡



$$\hat{y} = b_0 + b_1 x + b_2 x^2 \quad \text{가 적합}$$

㉢



㉣



③ 독립성 검토 (독립변수가 시계열자료일 경우) : 더빈-왓슨 통계량을 이용하여 자기상관성을 검토한다

더빈-왓슨 통계량이 2에 가까우면 독립성을 만족한다

$H_0 : \rho = 0$ 0에 가까우면 오차항 간에 양의 상관관계가 존재한다

$H_1 : \rho > 0$ 4에 가까우면 오차항 간에 음의 상관관계가 존재

최소제곱법 (Method of Least Squares)

잔차들의 제곱합 $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$ 을 S라고 하면

S를 최소로 하는 a와 b는 다음과 같이 구한다.

- S를 a와 b로 각각 편미분하여 0으로 놓으면 $\frac{\partial S}{\partial a} = -2 \sum (y_i - a - bx_i) = 0$
- 두식을 풀면 다음과 같은 방정식을 얻을 수 있다. $\frac{\partial S}{\partial b} = -2 \sum x_i (y_i - a - bx_i) = 0$

$$\textcircled{1} \quad \underline{an + b\sum x_i = \sum y_i}, \quad \textcircled{2} \quad \underline{a\sum x_i + b\sum x_i^2 = \sum x_i y_i}$$

이 두 방정식을 정규방정식이라 하며, 정규방정식으로부터 구한 a와 b는 다음과 같다.

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}, \quad a = \bar{y} - b \bar{x}$$

• 이 추정된 a와 b를 회귀추정량이라 한다. 즉, 추정된 회귀선은 $\hat{y}_i = a + bx_i$ 이다.

• 표본자료로부터 $y_i = \alpha + \beta x_i + \epsilon_i$ 을 추정하여 얻은 직선 $\hat{y}_i = a + bx_i$ 을 회귀선이라 한다.

• a, b, \hat{y}_i 은 α, β, y_i 의 추정값이며 a를 절편, b를 기울기라 한다.

• b는 x_i 가 한 단위 증가할 때에 \hat{y}_i 의 증가량을 나타낸다.

• 총변동 = 오차변동 + 회귀변동 $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

$$SST = SSE + SSR \quad \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$(\because \sum e_i = 0, \sum e_i \hat{y}_i = 0)$$

• 자료에서 회귀식을 추정하는 방법

① \bar{x}, \bar{y} 를 구한다

② $S_{xx} = \sum (x_i - \bar{x})^2, S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y}$ 를 구한다

③ $b = \frac{S_{xy}}{S_{xx}}$ 를 계산

④ $a = \bar{y} - b \bar{x}$ 를 계산

$$\textcircled{5} \quad \hat{y} = a + bx$$

- 적합성 : 자료의 크기에 따라 표준오차는 변화 하므로 좋은 적합성의 특도는 아니다.

결정계수 (Coefficient of Determination ; R^2) : 추정된 회귀선이 관측값들을 얼마나 잘 설명하고 있는가를 나타내는 척도로서 총변동 중에서 회귀선에 의해 설명되는 비율이다.

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

- 결정계수의 특성

단순선형회귀에서는 상관계수 r 의 제곱이 R^2 가 된다.

$$r = b \frac{S_x}{S_y} = b \sqrt{\frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}}, \text{ 여기서, } b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- $SSR = b^2 S_{xx}$, 여기서, $S_{xx} = \sum (x_i - \bar{x})^2$
- $SSE = SST(1 - r^2)$, 즉, $\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 (1 - r^2)$

- 단순회귀모형의 가설 설정 :
 - 귀무가설 (H_0) : 회귀모형은 유의하지 않다 ($\beta = 0$)
 - 대립가설 (H_1) : 회귀모형은 유의하다 ($\beta \neq 0$)

단순회귀모형의 검정통계량 결정 : $F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F_{(1, n-2)}$

단순회귀모형의 분산분석표

요인	제곱합	자유도	평균제곱	검정통계량 F	F_α
회귀	SSR	1	$\frac{1}{SSR} = MSR$	$F = \frac{MSR}{MSE}$	$F_{(\alpha, 1, n-2)}$
잔차	SSE	$n-2$	$\frac{SSE}{n-2} = MSE$		
합계	SST	$n-1$			

단순회귀모형의 결과 해석 : $F(1, n-2) \geq F(\alpha, 1, n-2) \Rightarrow$ 귀무가설 기각

$$\text{단순회귀계수의 검정통계량 결정} : t = \frac{b - \beta}{\sqrt{\text{Var}(b)}} = \frac{b - \beta}{\sqrt{MSE/S_{xx}}} \sim t(n-2),$$

$$(t \text{ 검정통계량})^2 = F$$

○ 잔차의 성질 : 잔차는 오차의 추정값이다. 잔차 $e_i = y_i - \hat{y}_i = \text{관측값} - \text{예측값}$

① $\sum e_i = 0$ 잔차들의 합은 0이다.

② $\sum y_i = \sum \hat{y}_i$ 관측값 y_i 의 합과 예측값 \hat{y}_i 의 합은 같다.

③ $\sum x_i e_i = 0$ 잔차들의 x_i 에 대한 가중합은 0이다.

④ $\sum \hat{y}_i e_i = 0$ 잔차들의 \hat{y}_i 에 대한 가중합은 0이다.

• 회귀계수 α 와 b 는 다음의 분포를 따른다.

① σ^2 을 알 경우, $b \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$, $\alpha \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

② σ^2 을 모를 경우, $b \sim t_{n-2}\left(\beta, \frac{MSE}{S_{xx}}\right)$, $\alpha \sim t_{n-2}\left(\alpha, MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

• 원점을 지나는 단순회귀선의 성질

① 절편이 있는 원점을 지나지 않는 회귀선의 경우 잔차들의 합은 0이 되지만 원점을 통과하는 회귀선에 대해서는 잔차들의 합이 반드시 0인 것은 아니다.

② 잔차제곱합인 $\sum_{i=1}^n e_i^2 = \sum (Y_i - \bar{Y}_i)^2$ 의 자유도는 $(n-1)$ 이다. $\therefore \hat{Y} = bX_i$

③ 원점을 통과하는 회귀선의 유의성 검정은 검정통계량 $F = \frac{SSR/1}{SSE/(n-1)} = \frac{MSR}{MSE}$ 과 기각치 $F(\alpha, 1, n-1)$ 을 비교 검정한다.

④ 추정된 회귀직선이 항상 (\bar{X}, \bar{Y}) 을 지나는 것은 아니다.

⑤ 절편이 있는 회귀모형에서의 결정계수는 항상 0 이상이지만, 원점을 지나는 회귀모형에서의 결정계수는 음의 값을 가질 수 있다.

다중회귀분석

$$i=1, 2, \dots, n$$

$$\text{다중회귀모형} : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i, \quad \epsilon_i \sim i.i.d N(0, \sigma^2)$$

다중회귀모형을 행렬로 표현하면 $Y = X\beta + \epsilon$ 이고 다음과 같이 표현할 수 있다.

$$I = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$n \times 1$ $n \times (k+1)$ $n \times 1$ $(k+1) \times 1$ $n \times 1$

이라 하면,

$$\pi_X = X(X'X)^{-1}X', \quad \pi_I = I(I'I)^{-1}I' \text{ 로 정의할 때, } \bar{Y} = (I'I)^{-1}I'Y \text{ 이므로,}$$

각축의 제곱합을 다음과 같이 행렬로 표현할 수 있다.

$$SST = \sum (y_i - \bar{y})^2 = Y'Y - n(\bar{y})^2 = Y'(I - \pi_I)Y$$

$$SSE = \sum (y_i - \hat{y}_i)^2 = (Y - \hat{Y})'(Y - \hat{Y}) = (Y - Xb)'(Y - Xb) = Y'(I - \pi_X)Y$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n(\bar{y})^2 = \hat{Y}'\hat{Y} - n(\bar{y})^2 = Y'(\pi_X - \pi_I)Y$$

최소제곱법은 잔차제곱합 $e'e$ 을 최소로 하는 추정량 b 를 찾는 것이다.

$$SSE = e'e = (Y - Xb)'(Y - Xb) = Y'Y - 2b'X'Y + b'X'Xb$$

$$\therefore b'X'Y = Y'Xb : \text{Scalar}$$

$$\frac{\partial SSE}{\partial b'} = -2X'Y + 2X'Xb = 0$$

$$\therefore b = (X'X)^{-1}X'Y$$

- 다중회귀 계수의 추정량 b 로부터 추정된 다중회귀식은 $\hat{Y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$ 이고, 행렬로 표현하면 $\hat{Y} = Xb$ 이다.

