2장 | 회귀 분석의 요약

SAS를 이용한 실험 계획과 분산 분석 (자유아카데미)

RECAP

두 확률변수의 관계를 설명하는데 있어서 상관계수만으로는 부족하여 두 변수의 관계를 수학적 관계식(혹은 모형식)으로 표현하고 싶은 경우가 있다.

어떤 확률변수 Y 를 변수 X 의 함수식으로 아래와 같이 표현한다고 가정하자.

$$y = a + b\,x + \epsilon$$

이를 '단순회귀모형(simple regression model)'[2]이라고 하며, ϵ은 오차를 의미한다.

만일 n 개의 $(x_i,\, y_i)$ 자료값에 적용하면, 이는

$$y_i = a + b\,x_i + \epsilon_i, \qquad i = 1,\, 2,\, \cdots,\, n \tag{2.7}$$

이 되며 a는 회귀선의 절편(intercept)이고 b는 기울기(slope)를 나타낸다.

단순 회귀 모형의 가정

1. X 와 Y 간에 $y_i = a + b\,x_i + \epsilon_i$ 관계를 가정하고$(i = 1, 2, \cdots, n)$,

2. 오차 ϵ_i 는 서로 독립이고 분산이 σ^2 인 정규분포를 따른다.

단순회귀 모형의 최소 제곱 추정법

오차제곱합을 수식으로 표현하면 $(x_i, y_i)_{i=1}^{n}$ 에 대해

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - b\,x_i)^2 \tag{2.8}$$

이라고 쓸 수 있다. 이 Q를 최소화시키는 a, b 얻으려면 이를 a, b에 대해 각각 편미분한 값이 0이 되는 아래와 같은 연립방정식의 해를 구하면 된다.

$$\begin{aligned} \frac{\partial Q}{\partial a} &= \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - b\,x_i)^2 = 0 \\ \frac{\partial Q}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - b\,x_i)^2 = 0 \end{aligned} \tag{2.9}$$

단순 회귀모형의 최소 제곱 추정량

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

결정 계수와 분산 분석

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

$$\text{SST} = \text{SSR} + \text{SSE}$$

결정 계수와 분산 분석

Source	d.f.	S.S.	M.S.	F_0
Regression	1	SSR	MSR	$\frac{\text{MSR}}{MSE}$
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		

표 2.3: 단순회귀분석의 분산분석표(ANOVA table)

단순회귀선의 유의성검정

만일 $F_0 = \frac{\text{MSR}}{\text{MSE}} > F_{0.05,\,1,n-2}$ 이면, 회귀선은 유의하다.

결정계수(Coefficient of Determination), $(0 \le R^2 \le 1)$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{2.14}$$

회귀 계수의 유의성 검정

기울기에 대한 t 검정

$H_0 : b = 0 \quad \text{vs.} \quad H_1 : b \neq 0$ 에 대해, 만일

$$t_0 = \frac{|\hat{b}|}{\text{S.E.}(\hat{b})} = \frac{|\hat{b}|}{\sqrt{\text{MSE}/\sum_{i=1}^{n}(x_i - \bar{x})^2}} > t_{\frac{\alpha}{2}, n-2} \tag{2.15}$$

이면, H_0 를 기각한다.

절편에 대한 t 검정

$H_0 : a = 0 \quad \text{vs.} \quad H_1 : a \neq 0$ 에 대해, 만일

$$t_0 = \frac{|\hat{a}|}{\text{S.E.}(\hat{a})} = \frac{|\hat{a}|}{\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}} > t_{\frac{\alpha}{2}, n-2} \tag{2.16}$$

이면, H_0 를 기각한다.

행렬을 이용한 단순 회귀 분석

단순회귀모형인

$$y_i = a + b\,x_i + \epsilon_i, \qquad i = 1, 2, \cdots, n \tag{2.17}$$

을 벡터로 표현하면,

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\boldsymbol{\epsilon}} \tag{2.18}$$

이 되어, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ 이라고 표현된다.

행렬을 이용한 단순 회귀 분석의 추정

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$$
$$= (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$$
$$= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T X^T \boldsymbol{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}$$

$$X^T X \boldsymbol{\beta} = X^T \boldsymbol{y}$$

SST의 분할

$$\underset{\text{SST}}{\left(\boldsymbol{y}^T \boldsymbol{y} - n(\bar{y})^2\right)} = \underset{\text{SSE}}{\left(\boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T X^T \boldsymbol{y}\right)} + \underset{\text{SSR}}{\left(\hat{\boldsymbol{\beta}}^T X^T \boldsymbol{y} - n(\bar{y})^2\right)} \qquad (2.22)$$

회귀분석 SAS 코드

```
data a ; input x y ; cards;
1    1
1.5   3
2    4.5
2.5   6
3 5   5
;
proc reg data=a
   model y=x;
run;
```