3주차 2차시 : 단순회귀분석(분산추정 및 결정계수)

1. 분산 추정

2. 결정계수

1.분산 추정

(1)교란항의 분산

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}$$

단, $\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

(2)회귀계수의 분산

$$\begin{aligned} \mathrm{var}(\hat{\beta}_1) &= \mathrm{E}[\left(\hat{\beta}_1 - E(\hat{\beta}_1)\right)^2] \\ &= \mathrm{E}[\left(\hat{\beta}_1 - \beta_1\right)^2] \\ &= \hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

$$\therefore \hat{\beta}_1 \sim \left(\beta_1,\ \hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^{n} x_i^2}\right)$$

(예제-계속) 다음의 홍보비 지출액 X(단위:천만 원)와 연간 매출액 Y(단위:억 원)에 관한 자료를 이용하여 회귀계수와 추정 회귀식을 구하고 해석하라.

X	2	3	4	5	6
Y	4	4	6	6	10

(기초 계산)

$\sum X_i = 20,\ \sum Y_i = 30,\ \sum X_i Y_i = 134,$
$\bar{X}=4,\ \bar{Y}=6,\ \sum X_i^2 = 90,\ \sum Y_i^2 = 204$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i = 90-(4)(20)=10$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \bar{Y}\sum_{i=1}^{n} Y_i = 204-(6)(30)=24$$

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \bar{X}\sum_{i=1}^{n} Y_i = 134-(4)(30)=14$$

$$\sum_{i=1}^{n}\hat{e}_i^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})(Y_i-\bar{Y}) = 24 - (1.4)(14) = 4.4$$

$$\therefore\ \hat{\sigma}_u^2 = \frac{4.4}{3} = 1.4667$$

$$\mathrm{var}(\hat{\beta}_1) = \hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = (1.4667)\left(\frac{1}{10}\right)=0.14667$$

```
> (dx<-x-mx)
[1] -2 -1  0  1  2
> (dy<-y-my)
[1] -2 -2  0  0  4
> (sumdxsq<-sum(dx^2))
[1] 10
> (sumdysq<-sum(dy^2))
[1] 24
> (sumdxdy<-sum(dx*dy))
[1] 14
> (ssr<-sumdysq-beta1*sumdxdy)
[1] 4.4
> (sigusq<-ssr/3)
[1] 1.466667
>
> (vbeta1<-sigusq/sumdxsq)
[1] 0.1466667
```

2.결정계수(coefficient of determination)

-표본으로부터 추정한 회귀선이 변수의 표본관측에 얼마나 적합한 지를 측정하는 계수로 설명력(explanatory power)이라고도 함

-결정계수는 표본관측이 추정된 회귀식에 가까운 정도를 계수로 나타낸 것으로 종속변수의 전변동과 회귀변동의 비율로 측정

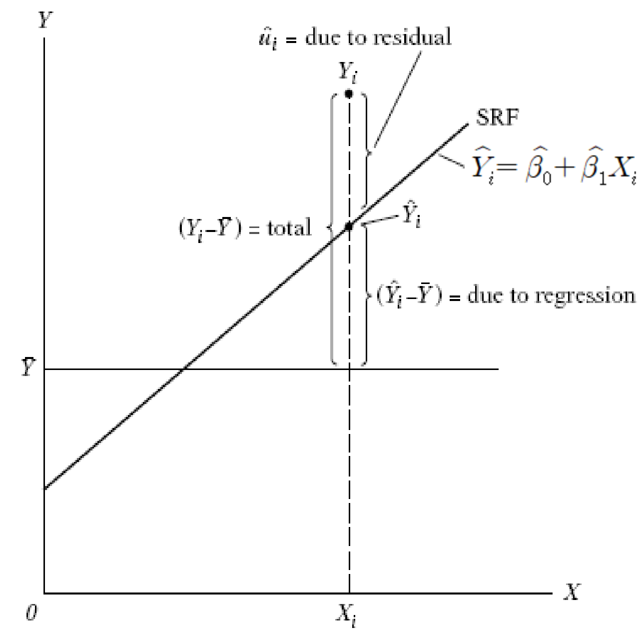
-결정계수의 값은 0과 1 사이로, 큰 값일수록 적합도 또는 설명력이 높다는 것을 의미하고 작은 값일수록 적합도 또는 설명력이 낮다는 것을 의미함

$$Y_i = \hat{Y}_i + e_i$$

또는 $Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + e_i$

$$\underset{(\text{총변동})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \underset{(\text{회귀변동})}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} + \underset{(\text{잔차변동})}{\sum_{i=1}^{n} e_i^2}$$

$$R^2 = \frac{\text{회귀변동}}{\text{총변동}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$



(예제-계속)

$$R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{19.6}{24} = 0.817$$

```
> (rsq<-(beta1^2*sumdxsq)/sumdysq)
[1] 0.8166667
```

- **실증분석 시 유의사항**

- R-square에 대한 해석(크기 및 크기 비교)