2주차 2차시 : 단순회귀분석(회귀계수의 추정)

1. 회귀계수의 추정

   (1) 보통최소자승법(OLS)

   (2) 추정된 회귀선의 특징

1.회귀계수의 추정

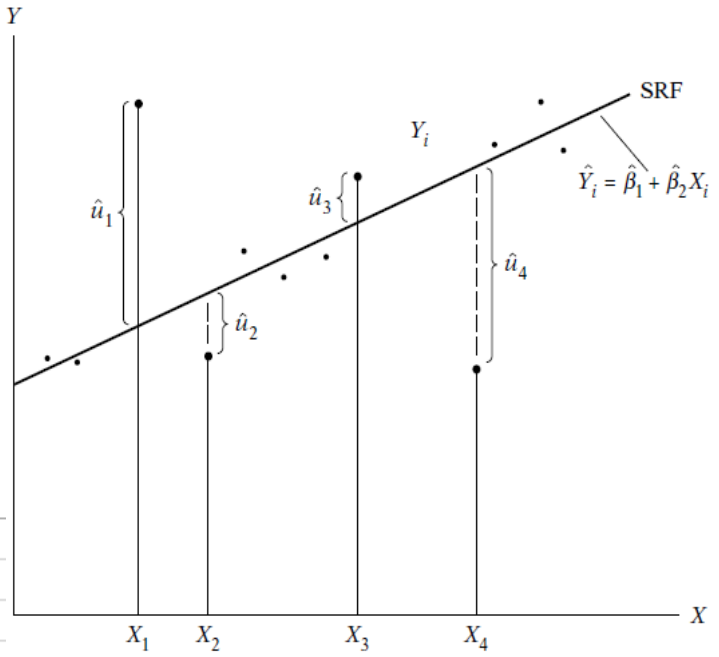
(1)보통최소자승법(Ordinary Least Squares : OLS)

-잔차(=실제치-예측치)의 합계가 최소가 되도록 하는 것이 바람직한데 잔차의 합이 0이 되는 식은 유일하지가 않으므로 잔차의 제곱의 합이 최소가 되게 하는 회귀식을 구하는 추정방법

(모형) $Y_i = \beta_0 + \beta_1 X_i + u_i$

(추정 회귀선) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

(잔차) $e_i = Y_i - \hat{Y}_i$



	A	B	C	D	E	F	G	H	I
1	Y	X		hat Y1	res Y1	sq(res Y1)	hat Y2	res Y2	sq(res Y2)
2	4	1		2.929	1.071	1.147041	4	0	0
3	5	4		7	-2	4	7	-2	4
4	7	5		8.357	-1.357	1.841449	8	-1	1
5	12	6		9.714	2.286	5.225796	9	3	9
6					0	12.21429		0	14
7									
8		Y1	Y2						
9	절편	1.572	3						
10	기울기	1.357	1						

(참고 1)합산연산자(Summation Operator)

-기호 $X_1, X_2, \ldots, X_n$이 n개의 관측치를 나타낼 때,
이들의 총합 $X_1 + X_2 + \cdots + X_n$은 총합기호 $\sum$를 사용하여 나타내면 간편함
-총합기호는 다음과 같은 특성을 가지고 있음

① $$\sum_{i=1}^{n}(X_i \pm Y_i) = \sum_{i=1}^{n} X_i \pm \sum_{i=1}^{n} Y_i$$

② $$\sum_{i=1}^{n} cX_i = c\sum_{i=1}^{n} X_i$$

③ $$\sum_{i=1}^{n} c = nc$$

④ $$\sum_{i=1}^{n}(X_i \pm c) = \sum_{i=1}^{n} X_i \pm nc$$

⑤ $$\sum_{i=1}^{n}(X_i \pm c)^2 = \sum_{i=1}^{n} X_i^2 \pm 2c\sum_{i=1}^{n} X_i + \quad nc^2$$

(보통최소자승법)

$$\underset{(\hat{\beta}_0,\hat{\beta}_1)}{\text{Min.}} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

(f.o.n.c)

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_0} = \frac{\partial \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_1} = \frac{\partial \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) = 0$$

정규방정식(normal equation)을 유도

$$\sum_{i=1}^{n} Y_i = \mathrm{n}\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

정규방정식을 연립으로 풀면,

$$\sum_{i=1}^{n} X_i Y_i = \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} X_i^2$$

$$- \quad \bar{X} \sum_{i=1}^{n} Y_i = \mathrm{n}\widehat{\beta_0}\bar{X} + \widehat{\beta_1}\bar{X} \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i - \bar{X} \sum_{i=1}^{n} Y_i = \widehat{\beta_1}\left(\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i\right)$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{X} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i} = \frac{cov(X,Y)}{var(X)}$$

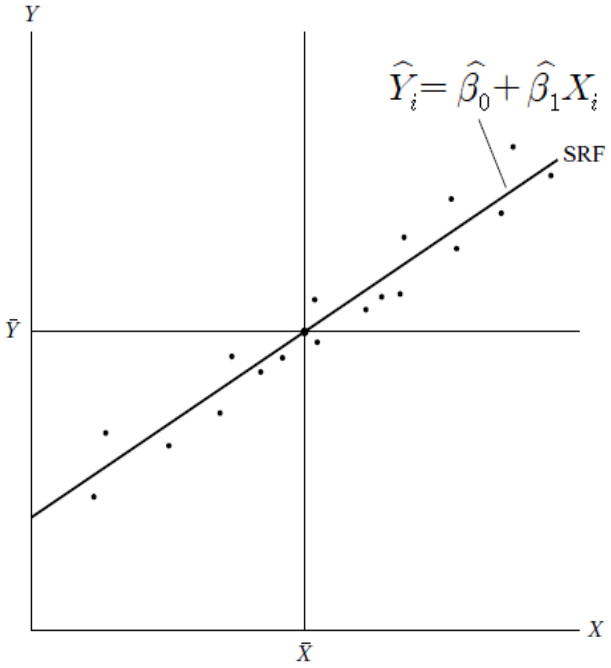
또는 $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

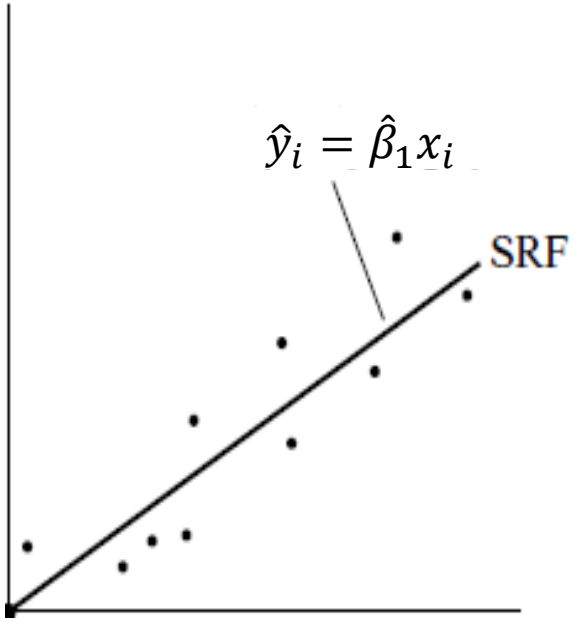
$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

$$\therefore \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

(2)추정된 회귀선의 특징

① 최소자승법으로 구한 회귀선은 항상 X, Y의 표본평균점을 지난다





② Y의 추정치 $\hat{Y}_i$의 평균값은 실제 Y의 평균값과 같다. 즉 $\bar{\hat{Y}}=\bar{Y}$이다

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ &= (\bar{Y} + \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 X_i \\ &= \bar{Y} + \hat{\beta}_1 (X_i - \bar{X})\end{aligned}$$

양변에 총합연산을 하면 다음과 같게 된다

$$\sum_{i=1}^{n} \hat{Y}_i = \mathrm{n}\bar{Y} + \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})$$

$\sum_{i=1}^{n}(X_i - \bar{X})$=0 이므로 양변을 n으로 나누어 주면 $\bar{\hat{Y}}=\bar{Y}$이다.

③ 잔차의 합 $\sum_{i=1}^{n} e_i$은 0이다. 즉, 잔차의 평균 $\bar{e}$은 0이다.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

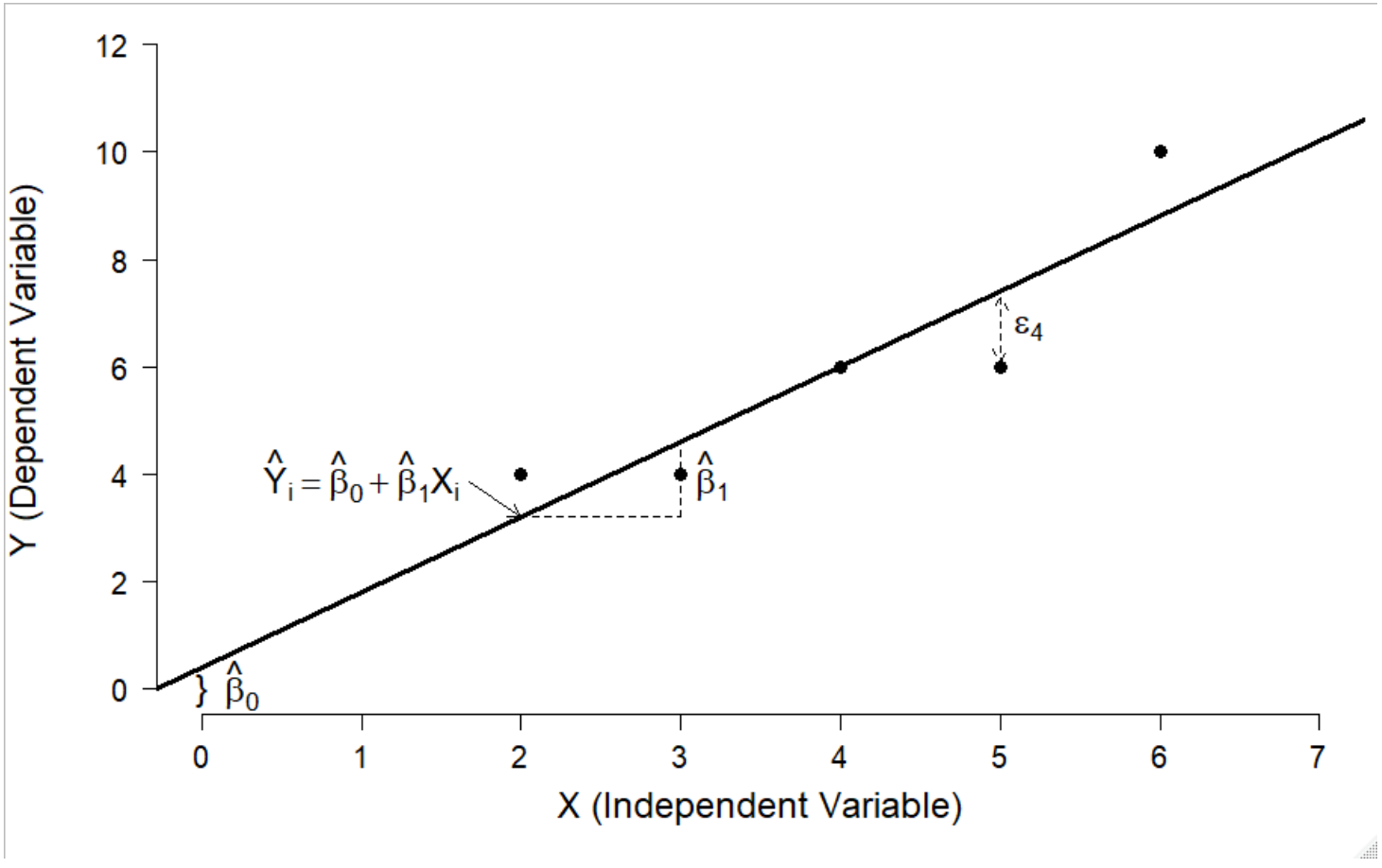
양변에 총합연산을 하면 다음과 같게 된다

$$\sum_{i=1}^{n} Y_i = \mathrm{n}\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} e_i$$

위 식이 정규방정식과 동일하기 위해서는 $\sum_{i=1}^{n} e_i$ 이 성립해야 한다.

(예제) 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)

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

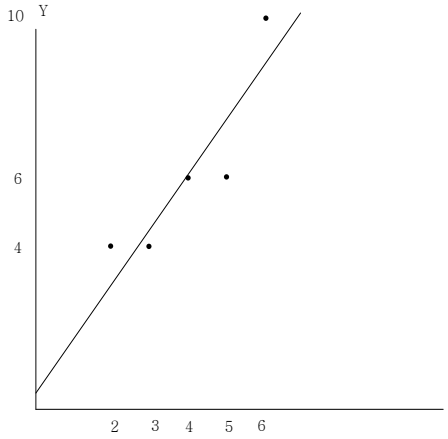
(회귀계수 계산)

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i} = \frac{134-4\times30}{90-4\times20} = \frac{14}{10} = 1.4$$

$$\therefore \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 6 - (1.4)(4) = 0.4$$

(추정 회귀식)

$$\hat{Y}_i = 0.4 + 1.4X_i$$




```
> x<-c(2,3,4,5,6)
> y<-c(4,4,6,6,10)
> x;y
[1] 2 3 4 5 6
[1]  4  4  6  6 10
>
> (n<-length(x))
[1] 5
>
> (sumx<-sum(x))
[1] 20
> (sumy<-sum(y))
[1] 30
> (mx=mean(x))
[1] 4
> (my=mean(y))
[1] 6
> (xy<-x*y)
[1]  8 12 24 30 60
> (sumxy<-sum(xy))
[1] 134
> (sumxsq<-sum(x^2))
[1] 90
> (sumysq<-sum(y^2))
[1] 204
>
```

```
> beta1<-(sumxy-mx*sumy)/(sumxsq-mx*sumx)
> beta0<-my-beta1*mx
> beta0;beta1
[1] 0.4
[1] 1.4
```

(해석)

-홍보비 지출액이 천 만원 증가하면 연간 매출액은 평균 1.4억 원 증가한다