9/16 회귀분석 실습자료.

(1) Using the four points (0,0), (1,1), (0,1), (1,2) , obtain the Simple Regression Line by using LSE (Use LSE formula)

$n=4,\ \sum_{i=1}^{n} x_i = 2,\ \sum_{i=1}^{n} y_i = 4,\ \sum_{i=1}^{n} x_i y_i = 3,\ \sum_{i=1}^{n} x_i^2 = 1$

$$\hat{y} = b_o + b_1 x,\ b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = 1$$

$b_0 = \overline{y} - b_1 \overline{x} = -0.5$

$\therefore \hat{y} = -0.5 + x$

--> 수업시간에 추정한 회귀식을 통해 좌표축에 네 개의 점을 찍고 그려보도록 하겠습니다.

(2) With the above regression line, verify that the sum of residuals is zero

$$\sum_{i=1}^{n} \hat{e_i} = \sum_{i=1}^{n} (y_i - \hat{y_i}) = (0-0.5)+(0+0.5)+(0-0.5)+(0+0.5) = 0$$

(3) Calculate the SSE for this regression line

$$\sum_{i=1}^{n} (y_i - \hat{y_i})^2 = (0+0.5)^2 + (0-0.5)^2 + (0+0.5)^2 + (0-0.5)^2 = 1$$

(4) Calculate the SSE for $\hat{y}$= 2x

$$\sum_{i=1}^{n} (y_i - \hat{y_i})^2 = (0-0)^2 + (1-2)^2 + (1-2)^2 + (0-0)^2 = 2$$

(5) Write the distribution of b1, b0, (use $\sigma^2 = MSE$ from (3))

solve) $\hat{\sigma^2} = MSE = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 = \frac{1}{4-2} \times 1 = 0.5,\ \ y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim N\left(0, \sigma^2\right)$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\rightarrow \quad b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right) = N(\beta_1, 0.5) = N(1, 0.5)$$

$b_0 = \overline{y} - b_1 \times \overline{x}$

$$\rightarrow \quad b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)\right) = N(\beta_0, 0.25) = N(0.5, 0.25)$$

p69. exercise 3.1 Production Units vs. Overhead (Hand + SAS)

Production (in 10,000) units	5	6	7	8	9	10	11
Overhead costs (in $1000)	12	11.5	14	15	15.4	15.3	17.5

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, $\epsilon_i \sim i.i.d\, N(0, \sigma^2)$

-> $n = 7, \sum_{i=1}^{n} x_i = 56, \sum_{i=1}^{n} y_i = 100.7, \sum_{i=1}^{n} x_i y_i = 831.1, \sum_{i=1}^{n} x_i^2 = 476$

--> $\hat{y} = b_o + b_1 x,\ b_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = 0.9107,$

$b_0 = \overline{y} - b_1 \overline{x} = 7.1$

$\therefore \hat{y} = 7.1 + 0.9107x$ (Least Square Estimator)

Using SAS (proc reg)

```
/** p69. exercise 3-1 Production Units vs. Overhead **/

* Input data ;

data budget ;
input production overhead @@ ;
cards ;
5 12 6 11.5 7 14 8 15 9 15.4 10 15.3 11 17.5
;
run ;

proc print data=budget ;
run ;

* Simple linear regression (proc reg) ;
```

```
proc reg data=budget ;
model overhead=production ;
plot overhead*production / conf pred ;
run ; quit ;
```

SAS 시스템 1
2015년 09월 14일 월요일 오후 07시02분39초

OBS	production	overhead
1	5	12.0
2	6	11.5
3	7	14.0
4	8	15.0
5	9	15.4
6	10	15.3
7	11	17.5

SAS 시스템 2
2015년 09월 14일 월요일 오후 07시02분39초

The REG Procedure
Model: MODEL1
Dependent Variable: overhead

Number of Observations Read	7
Number of Observations Used	7

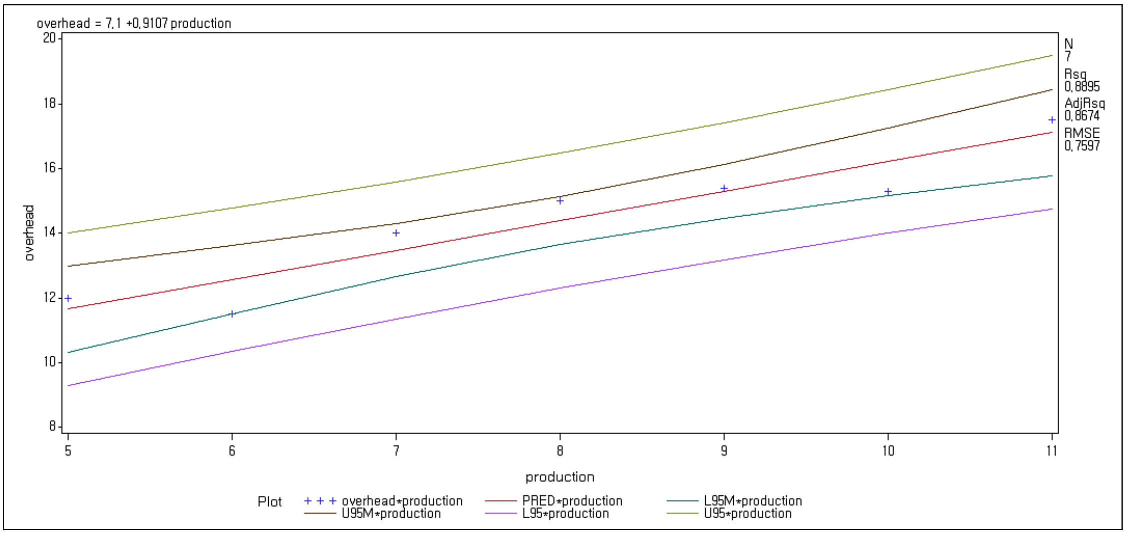
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	23.22321	23.22321	40.24	0.0014
Error	5	2.88536	0.57707		
Corrected Total	6	26.10857			

Root MSE	0.75965	R-Square	0.8895
Dependent Mean	14.38571	Adj R-Sq	0.8674
Coeff Var	5.28060		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.10000	1.18383	6.00	0.0018
production	1	0.91071	0.14356	6.34	0.0014


overhead = 7,1 +0,9107 production
N
7
Rsq
0,8895
AdjRsq
0,8674
RMSE
0,7597
overhead
production
Plot
overhead*production
PRED*production
L95M*production
U95M*production
L95*production
U95*production