9월 23일 실습

p74. exercise 3.3

(3)번 문제 3개의 자료를 각각 Scatter Plot 하시오.
Regression line을 각각 구하라.

첫 번째 모델.
Regression line이 양의 상관관계를 보인다.
size가 커질수록 value가 커진다.
-> $=-(50035)+(72.82038)x$ (x : SIZE, y : VALUE)

두 번째 모델.
Regression line이 양의 상관관계를 보인다.
numports가 증가할수록 cost가 증가한다.
-> $\hat{y}=(16594)+(650.16917)x$ (x : NUMPORTS, y : COST)

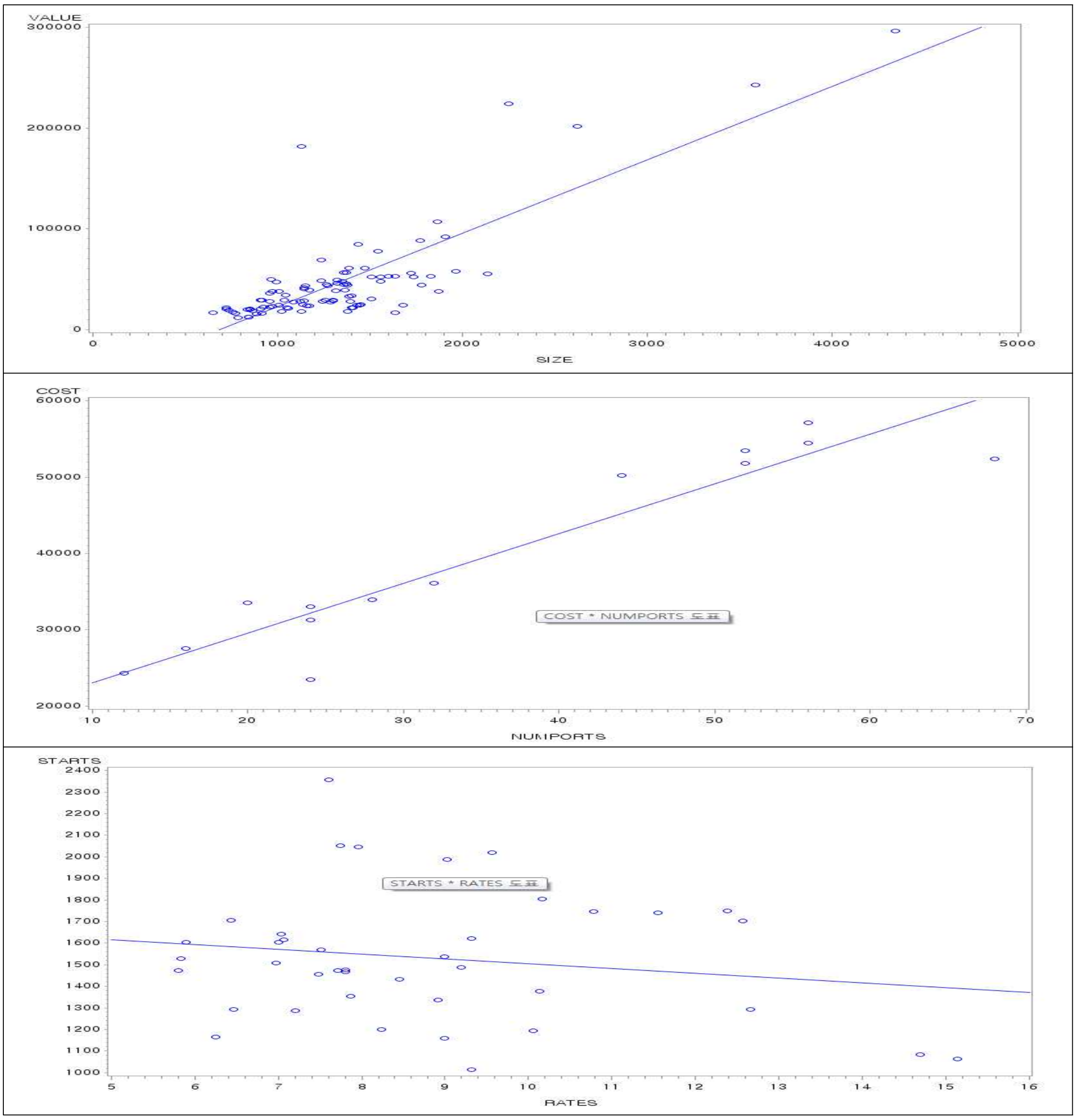
세 번째 모델.
Regression line이 음의 상관관계를 보인다.
rates가 높아질수록 starts가 낮아진다.
-> $\hat{y}=(1726.04027)-(22.23388)x$ (x : RATES, y : STARTS)

SAS Code

```
/* p.74  3.3 세가지 자료에 대해 Regression line 각각 구하기. */

proc import out=a
datafile='C:\Users\ADMIN\Desktop\realest3.csv'
dbms=csv replace ;
run;
proc import out=b
datafile='C:\Users\ADMIN\Desktop\comnode3.csv'
dbms=csv replace ;
run;
proc import out=c
datafile='C:\Users\ADMIN\Desktop\hstarts3.csv'
dbms=csv replace ;
run;

symbol1  V=circle C=blue  I= r;

proc gplot data=a;
plot value*size ;
run;
quit;
```

```
proc gplot data=b;
plot cost*numports;
run;
quit;

proc gplot data=c;
plot starts*rates;
run;
quit;
```

SAS Output


VALUE
SIZE
COST
NUMPORTS
COST * NUMPORTS 도표
STARTS
RATES
STARTS * RATES 도표

p74. exercise 3.3 기울기에 대한 95% 신뢰구간을 SAS Output을 보고 구하라.

$$\pm t_{0.025}(n-2) \times s.e(\widehat{\beta_1})$$

첫 번째 모델.
기울기에 대한 95% confidence interval은 (62.45193,83.18883)이다.
$\rightarrow 72.82038 \pm 1.984 \times 5.22480$

두 번째 모델.
기울기에 대한 95% confidence interval은 (504.37633,795.96201)이다.
$\rightarrow 650.16917 \pm 2.179 \times 66.91389$

세 번째 모델.
기울기에 대한 95% confidence interval은 (-64.23266,19.76489)이다.
$\rightarrow -22.23388 \pm 2.024 \times 20.74634$

SAS Code

```
/* 기울기에 대한 95% 신뢰구간을 sas output 을 보고 구하라.  */

proc reg data=a ;
model value=size / clb alpha=0.05 ;
plot value*size / conf pred ;
run ; quit ;

proc reg data=b ;
model cost=numports / clb alpha=0.05 ;
plot cost*numports / conf pred ;
run ; quit ;

proc reg data=c ;
model starts=rates / clb alpha=0.05 ;
plot starts*rates / conf pred ;
run ; quit ;
```

SAS Output

SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: VALUE

Number of Observations Read	100
Number of Observations Used	100

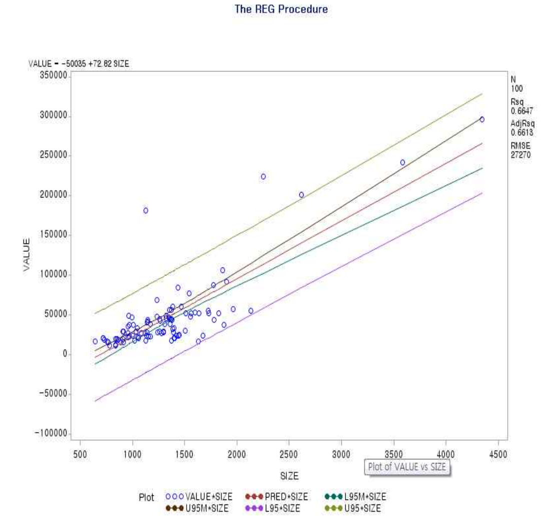
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.444589E11	1.444589E11	194.25	<.0001
Error	98	72879341312	743666748		
Corrected Total	99	2.173383E11			

Root MSE	27270	R-Square	0.6647
Dependent Mean	46184	Adj R-Sq	0.6613
Coeff Var	59.04737		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-50035	7422.67750	-6.74	<.0001	-64765	-35305
SIZE	1	72.82038	5.22480	13.94	<.0001	62.45193	83.18883



SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: COST

Number of Observations Read	14
Number of Observations Used	14

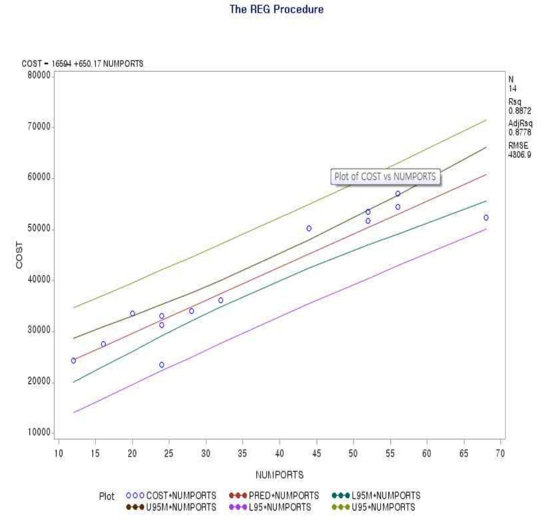
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1751268376	1751268376	94.41	<.0001
Error	12	2225[illegible]	[illegible]		
Corrected Total	13	1973862522			

Root MSE	4306.91446	R-Square	0.8872
Dependent Mean	40186	Adj R-Sq	0.8778
Coeff Var	10.71758		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	16594	2687.05000	6.18	<.0001	10739	22448
NUMPORTS	1	650.16917	66.91389	9.72	<.0001	504.37633	795.96201



SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: STARTS

Number of Observations Read	40
Number of Observations Used	40

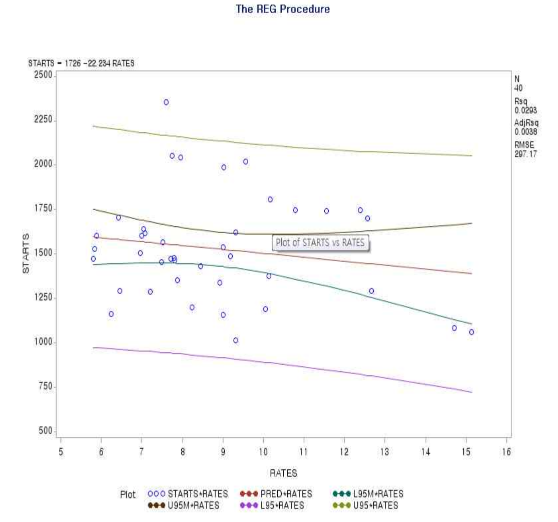
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	101430	101430	1.15	0.2906
Error	38	3355858	88312		
Corrected Total	39	3457289			

Root MSE	297.17346	R-Square	0.0293
Dependent Mean	1530.61000	Adj R-Sq	0.0038
Coeff Var	19.41536		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1726.04027	188.31147	9.17	<.0001	1344.82363	2107.25691
RATES	1	-22.23388	20.74634	-1.07	0.2906	-64.23266	19.76489



4) p. 90 (4)번 문제를 풀어라.

p69. exercise 3.1 Production Units vs. Overhead

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

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

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

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

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

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

$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = 28$

$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 = 1474.75 - \frac{100.7^2}{7} = 26.10857$

$MSE = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - y_i)^2$

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

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

$= \frac{1}{n-2} \left[\sum_{i=1}^{n} (y_i - \overline{y})^2 + b_1^2 S_{xx} - 2b_1 S_{xy} \right]$

$= \frac{1}{n-2} \left[\sum_{i=1}^{n} (y_i - \overline{y})^2 + b_1^2 S_{xx} - 2b_1 \frac{S_{xy}}{S_{xx}} S_{xx} \right]$

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

$= \frac{1}{5} (26.10857 - 0.9107^2 \times 28) = 0.5772$

Question a and b

Hypotheses -> $\beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$, significance level (α) : 0.05

Decision rule : $T = \frac{\beta_1 - \beta_1^*}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$ under H_0 -> $t_0 = \frac{0.9107 - 0}{\sqrt{\frac{0.5772}{28}}} = 6.343$

-> Reject H_0 if $t_0 > t_{0.025}(5)$ or if $t_0 < -t_{0.025}(5)$

-> $t_0 > t_{0.025}(5) = 2.571$

Result : Reject H_0! The slope of regression line is significant.

Interpretation : We can say production and overhead costs are linearly related.

Question c and d

Hypotheses : $H_0 : \beta_1 = 1 \;\; vs \;\; H_1 : \beta_1 \neq 1$, significance level (α) : 0.05

Decision rule : $T = \frac{\widehat{\beta_1} - \beta_1^*}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$ under H_0 -> $t_0 = \frac{0.9107 - 1}{\sqrt{\frac{0.5772}{28}}} = -0.622$

-> Reject H_0 if $t_0 > t_{0.025}(5)$ or if $t_0 < -t_{0.025}(5)$

-> $-t_{0.025}(5) = -2.571 < t_0 < t_{0.025}(5) = 2.571$

Result : Not reject H_0!

Interpretation : We can say slope of regression line is equal to 1.
That means if production is increase 10,000 units, overhead cost is increase $1000.