전산통계 과제#6

컴퓨터소프트웨어 학부
2018008559
신상윤

8-1

코드

```
data exe8_1;
input x y @@;
cards;
0.150 0.154 0.090 0.082 0.110 0.078
0.100 0.085 0.090 0.072 0.120 0.097
0.900 0.079 0.090 0.080 0.100 0.088
0.140 0.144 0.095 0.090 0.060 0.053
0.080 0.078 0.040 0.050 0.080 0.072
;
run;
data exe8_1_1;
input x y @@;
cards;
0.150 0.154 0.090 0.082 0.110 0.078
0.100 0.085 0.090 0.072 0.120 0.097
0.090 0.080 0.100 0.088
0.140 0.144 0.095 0.090 0.060 0.053
0.080 0.078 0.040 0.050 0.080 0.072
;
run;
proc reg data = exe8_1;
        model y = x;
run;
proc reg data = exe8_1_1;
        model y = x / r dw;
        output out = out student=z;
        plot student.*x;
run;
proc univariate data = out;
        probplot z / normal(mu=EST sigma=EST);
run;
```

결과

주어진 데이터를 있는 그대로 회귀분석을 진행하면

y=0.0062x + 0.08587

즉, β가 0에 아주 가까운 식이 나타난다.

따라서 H0 : β = 0에서 검정해보면

p-value = 0.8706/2 = 0.4353으로

유의수준 5%에서 H0를 채택한다.

결론적으로 x가 y에 유의미한 영향을 주고 있지 않다는 뜻이다.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00002357	0.00002357	0.03	0.8706
Error	13	0.01110	0.00085406		
Corrected Total	14	0.01113			

Root MSE	0.02922	R-Square	0.0021
Dependent Mean	0.08680	Adj R-Sq	-0.0746
Coeff Var	33.66863		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.08587	0.00939	9.15	<.0001
x	1	0.00620	0.03730	0.17	0.8706

다시 data를 살펴보면, x = 0.9, y = 0.079인 data를 제외하고는 x와 y의 크기가 비슷함을 볼 수 있다. 즉, 이상치가 존재하여 회귀분석을 제대로 하지 못한 것 같다. 따라서 이를 제외하고 다시 분석한다.

추정된 회귀식은

y = 0.94077x – 0.00302이고,

p-value가 매우 작아

유의수준 5%에서 H0를 기각한다.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance

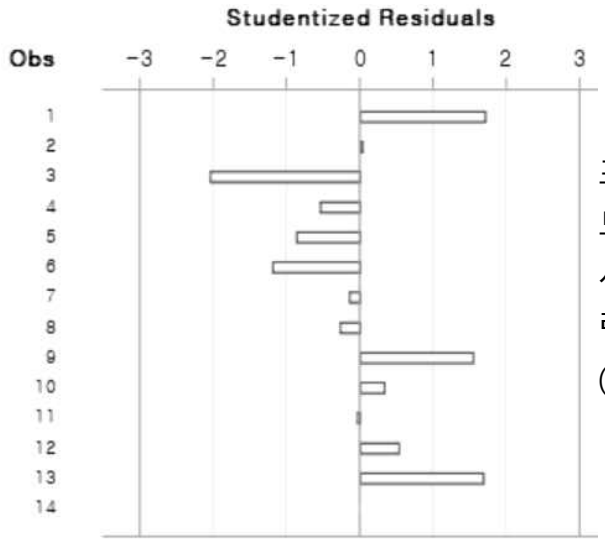
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00948	0.00948	71.83	<.0001
Error	12	0.00158	0.00013194		
Corrected Total	13	0.01106			

Root MSE	0.01149	R-Square	0.8569
Dependent Mean	0.08736	Adj R-Sq	0.8449
Coeff Var	13.14895		

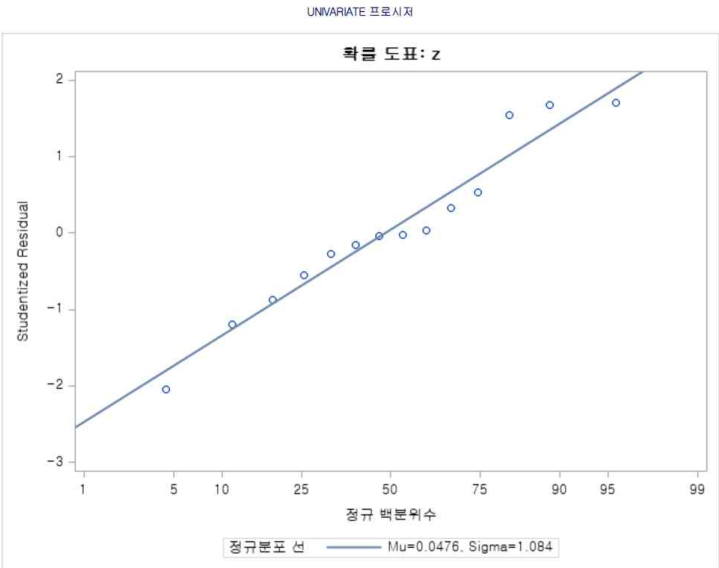
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00302	0.01110	-0.27	0.7899
x	1	0.94077	0.11100	8.48	<.0001

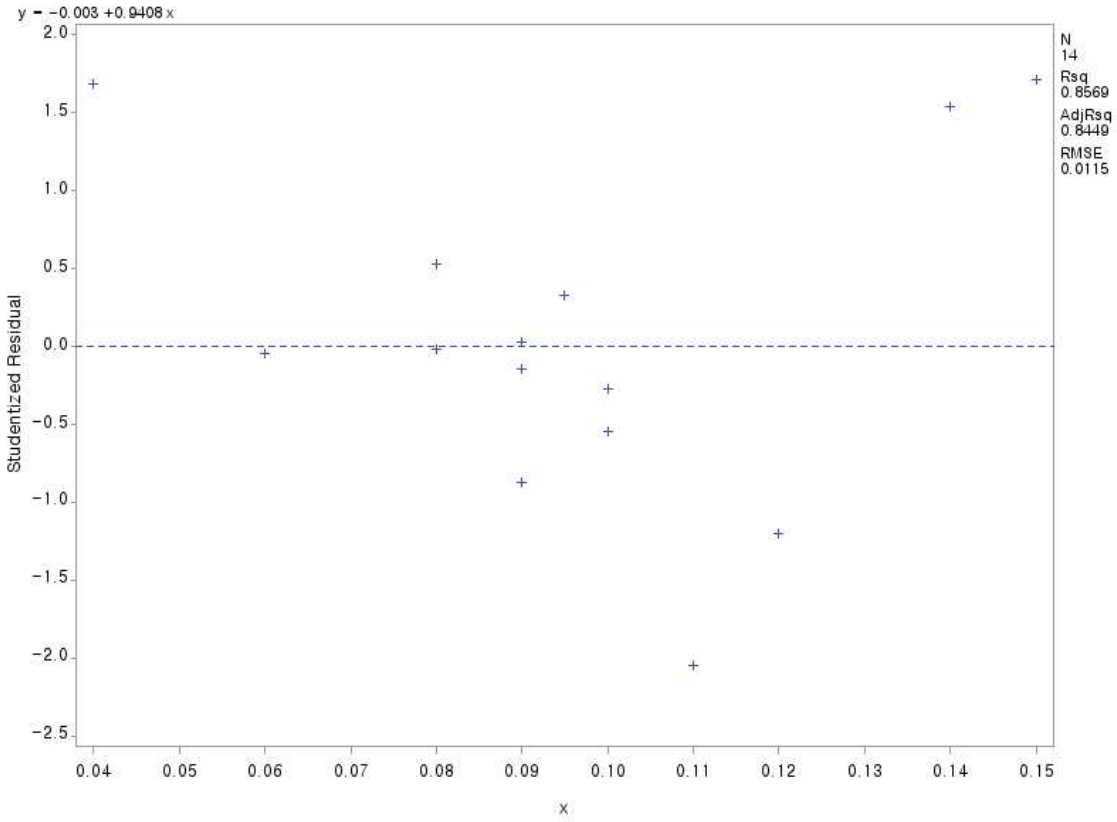
<잔차분석>



표준화 잔차 도표를 보면
모든 잔차가 0을 중심으로 –2 ~ 2
사이에서 특별한 규칙 없이
랜덤하게 잘 분포하고 있다.
(특이값 없음)



표준화 잔차의
정규확률도표를 보면
대체적으로 일직선을
이루고있어 정규성 가정에
큰 문제가 없다.



표준화 잔차 도표를 보면 오른쪽 위 모서리 부분에도 데이터가 분포되어 있어 잔차가 잘 흩어져있다고 할 수 있다. 즉, 등분산성을 가정할 수 있다.

Durbin–Watson D	1.294
Number of Observations	14
1st Order Autocorrelation	0.273

마지막으로 DW값이 1.294로 0보다 2에 가까우므로 독립성 가정에 문제가 없다.

8-3

코드

```
data  exe8_3;
input  x1  x2  x3  x4  y;
cards;
21     1    71.0    12.7    170
22     6    56.5     8.0    120
```

```
24     5   56.0    4.3   125
24     1   61.0    4.3   148
25     1   65.0   20.7   140
27    19   62.0    5.7   106
28     5   53.0    8.0   120
28    25   53.0    0.0   108
31     6   65.0   10.0   124
32    13   57.0    6.0   134
33    13   66.5    8.3   116
33    10   59.1   10.3   114
34    15   64.0    7.0   130
35    18   69.5    7.0   118
35     2   64.0    6.7   138
36    12   56.5   11.7   134
36    15   57.0    6.0   120
37    16   55.0    7.0   120
37    17   57.0   11.7   114
38    10   58.0   13.0   124
38    18   59.5    7.7   114
(중략)

;
run;
proc corr data = exe8_3;
        var y x1 x2 x3 x4;
run;
proc reg data = exe8_3;
        model y = x1 x2 x3 x4 / stb;
run;
proc reg data = exe8_3;
        model y = x1 x2 x3 x4 / selection=stepwise slentry=0.15
slstay=0.15;
run;
```

결과

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	39
Number of Observations Used	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2791.32176	697.83044	6.34	0.0006
Error	34	3740.11414	110.00336		
Corrected Total	38	6531.43590			

Root MSE	10.48825	R-Square	0.4274
Dependent Mean	127.41026	Adj R-Sq	0.3600
Coeff Var	8.23187		

anova table에서 p-value는 0.0006이므로 유의수준 5%에서 귀무가설을 기각한다. (H0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$) 따라서 4개의 독립변수 중 적어도 하나 이상의 변수는 종석변수를 설명하는데에 유의하게 기여할 것이다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	51.51366	17.29240	2.98	0.0053	0
x1	1	-0.15304	0.28178	-0.54	0.5906	-0.08974
x2	1	-0.53129	0.22197	-2.39	0.0224	-0.40904
x3	1	1.43708	0.31458	4.57	<.0001	0.77814
x4	1	-0.17484	0.47127	-0.37	0.7129	-0.05417

각 회귀계수에 대한 p-value를 보면 x1,x4는 유의수준 5%에서 귀무가설(H0 : β=0)을 기각할 수 없고, x2, x3는 기각한다.

표준화 회귀계수를 살펴보아도 x1과 x4의 계수는 매우 작다. (영향이 없다) 즉, x1,x4는 유의하게 기여하지 않고, x2,x3는 유의하게 기여한다고 할 수 있다.

피어슨 상관 계수, N = 39
H0: Rho=0 가정하에서 Prob > |r|

	y	x1	x2	x3	x4
y	1.00000	0.00584 0.9718	-0.08748 0.5964	0.52136 0.0007	0.25079 0.1236
x1	0.00584 0.9718	1.00000	0.58821 <.0001	0.43166 0.0061	-0.00537 0.9741
x2	-0.08748 0.5964	0.58821 <.0001	1.00000	0.48115 0.0019	0.00110 0.9947
x3	0.52136 0.0007	0.43166 0.0061	0.48115 0.0019	1.00000	0.39187 0.0136
x4	0.25079 0.1236	-0.00537 0.9741	0.00110 0.9947	0.39187 0.0136	1.00000

피어슨 상관계수를 보아도 x1,x4는 y와 상관이 없어보인다. 대신 x1은 x2와 관련이 있어 보이는데, 이는 이주후 경과기간이 길수록 당연히 나이가 많을 확률이 높은 것이고, x3와 x4의 상관성은 복부 피부 두께가 두꺼울수록 몸무게가 많이 나갈 확률이 높을 것이기 때문이다.

따라서 x2와 x3를 이용하여 y를 나타낼 것이다.

Backward Elimination: Step 2

Variable x1 Removed: R-Square = 0.4208 and C(p) = 1.3913

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2748.27852	1374.13926	13.08	<.0001
Error	36	3783.15738	105.08770		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	50.31913	15.81839	1063.39240	10.12	0.0030
x2	-0.57184	0.18794	972.89889	9.26	0.0044
x3	1.35408	0.26722	2698.29454	25.68	<.0001

Forward Selection: Step 2

Variable x2 Entered: R-Square = 0.4208 and C(p) = 1.3913

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2748.27852	1374.13926	13.08	<.0001
Error	36	3783.15738	105.08770		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	50.31913	15.81839	1063.39240	10.12	0.0030
x2	-0.57184	0.18794	972.89889	9.26	0.0044
x3	1.35408	0.26722	2698.29454	25.68	<.0001

Stepwise Selection: Step 2

Variable x2 Entered: R-Square = 0.4208 and C(p) = 1.3913

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2748.27852	1374.13926	13.08	<.0001
Error	36	3783.15738	105.08770		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	50.31913	15.81839	1063.39240	10.12	0.0030
x2	-0.57184	0.18794	972.89889	9.26	0.0044
x3	1.35408	0.26722	2698.29454	25.68	<.0001

각 변수선택 방법에서도 x2, x3를 선택함을 볼 수 있다.

따라서 다음과 같은 추정식을 쓸 수 있다.

y = -0.57184x_2 + 1.35408x_3 + 50.31913

결론적으로 최고혈압은 나이와 복부 피부두께와는 관계가 없었고, 이주후 경과기간이 길수록 최고혈압이 작아지는 경향을 보이고, 몸무게가 많이 나갈수록 최고혈압도 커지는 경향을 보였다.

8-4

결과

(가) $\hat{\beta}$ = 68.3/70.6 = 0.967, $\hat{\alpha}$ = 122.7 – 0.967*10.8 = 112.252

y = 0.967x + 112.252

(나)

1.

$$\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_iy_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}$$
$$= \sum x_iy_i - n\bar{x}\bar{y} = 68.3$$

$\sum x_i y_i = 68.3 + 15 \times 10.8 \times 122.7 = 19945.7$
$\sum x_i^2 = 70.6 + 15 \times 10.8 \times 10.8 = 1820.2$
$\sum y_i^2 = 98.5 + 15 \times 122.7 \times 122.7 = 225927.85$
$\sum (y_i - (ax_i + b))^2$ (추정된 식이 $y = ax + b$일 때)
$= \sum y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i$
$= 32.42537$
오차제곱합 = 32.43
분산 : $\hat{\sigma}^2 = 32.43/13$ = 2.495

2.

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{68.3}{70.6} = 0.967$$
$$SSE = S_{yy} - \widehat{\beta_1} S_{xy} = 98.5 - 0.967 \times 68.3 = 32.4$$
$$\widehat{\sigma^2} = \frac{SSE}{n-2} = \frac{32.4}{13} = 2.49$$

(다) x가 한 단위 증가할 때 E(Y)의 증가분은 기울기와 같으므로 기울기 검정을 실행한다.

H0 : $\beta_1 \geq 1.5$ H1 : $\beta_1 < 1.5$

검정통계량 $T = \frac{0.967 - 1.5}{\sqrt{\frac{2.495}{S_{xx}}}}$ = -2.83527, 기각역 $T \leq$ -1.771이므로

유의수준 5%에서 H0가 기각된다. 따라서 기울기가 1.5 미만이라 할 수 있다.

(라)

0.967 * 12 + 112.252 = 123.856

$t_{0.025}(13) = 2.16037$

$\sqrt{2.495(\frac{1}{15} + \frac{1.44}{70.6})} = 0.466$

신뢰구간 : (122.8491, 124.8629)

8-5

코드

```
data exe8_5;
input iq gpa;
cards;
100      3.0
120      3.8
110      3.1
105      2.9
 85      2.6
 95      2.9
130      3.6
100      2.8
105      3.1
 90      2.4
;
run;

proc corr data = exe8_5;
        var iq gpa;
run;
proc reg data = exe8_5 alpha=0.05;
        model gpa = iq / clb cli;
run;
```

결과

(가)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.05854	0.47199	0.12	0.9044
iq	1	0.02848	0.00450	6.32	0.0002

p-value가 0.0001이므로 유의수준 5%에서 H0를 기각한다. 즉, iq는 성적에 유의미한 영향을 준다.

추정한 회귀식은 평균평점 = 0.02848 * IQ + 0.05854이다.

(나)

$$t_{0.025}(8)\frac{0.1824}{\sqrt{109800-10\times 104\times 104}}=0.010387$$

신뢰구간 : (0.01809, 0.03886)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.05854	0.47199	0.12	0.9044	-1.02987	1.14694
iq	1	0.02848	0.00450	6.32	0.0002	0.01809	0.03886

(다)

$$\text{추정값}=0.02848\times 125+0.05854=3.61854$$

$$2.306\times 0.1824\times\sqrt{1+\frac{1}{10}+\frac{(125-104)^2}{1640}}=0.49212$$

예측구간 : (3.12642, 4.1106)

(라)

피어슨 상관 계수, N = 10 H0: Rho=0 가정하에서 Prob > |r|		
	iq	gpa
iq	1.00000	0.91281 0.0002
gpa	0.91281 0.0002	1.00000

피어슨 상관계수는 0.91281로 매우 높다.

p-value도 0.0001로 유의수준 5%에서

H0 : ρ=0를 기각한다.

즉, iq와 gpa는 선형 관계가 있다고 볼 수 있다.

(마)

Root MSE	0.18241	R-Square	0.8332
Dependent Mean	3.02000	Adj R-Sq	0.8124
Coeff Var	6.04009		

결정계수는 0.8332이다.
1에 가까우므로 데이터를 잘 설명하고 있다고 할 수 있다.
상관계수의 제곱값이다.

8-7

코드

```
data exe8_7;
input ver math gpa;
cards;
623     509     2.6
593     611     2.8
584     738     3.0
669     701     2.9
578     635     2.9
520     583     2.8
578     614     3.0
695     634     3.3
613     693     2.3
726     800     3.9
(중략)
;
run;

proc reg data = exe8_7;
        model gpa = ver;
        model gpa = math;
run;
```

결과

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.11077	0.55538	1.75	0.2043
Error	17	5.40673	0.31804		
Corrected Total	19	6.51750			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.09493	1.52257	0.06	0.9510
ver	1	0.00357	0.00216	1.65	0.1166
math	1	0.00068221	0.00195	0.35	0.7309

gpa = ver math로 회귀분석했을 때 model의 p-value는 0.2로 유의수준 5%에서 H0를 채택한다. 이는 gpa에 ver, math가 유의미한 영향을 주지 않는다는 것을 의미한다.

따라서 각각 회귀분석을 실시한다.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.07189	1.07189	3.54	0.0761
Error	18	5.44561	0.30253		
Corrected Total	19	6.51750			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.40144	1.21411	0.33	0.7447
ver	1	0.00379	0.00202	1.88	0.0761

먼저 gpa = ver에 대하여 회귀분석을 하면
기울기 검정에 대한 p-value = 0.03805로 유의수준 5% 하에서 H0를 기각한다. 따라서 ver는 gpa에 유의미한 영향을 준다고 할 수 있다.
추정된 식은 gpa = 0.00379 * ver + 0.40144이다.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.62946	1.26382	1.29	0.2136
math	1	0.00163	0.00195	0.83	0.4164

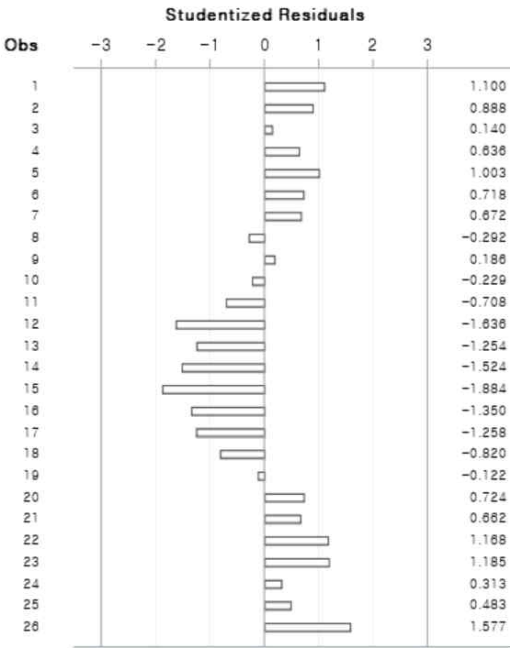
다음은 gpa = math에 대하여 회귀분석을 하면
기울기 검정에 대한 p-value = 0.2082로 유의수준 5% 하에서 H0를 채택한다. 따라서 math는 gpa에 유의미한 영향을 준다고 할 수 없다.
실제로 data를 보면 어학능력점수는 클수록 gpa가 큰 경향이 있지만, 수리능력점수는 gpa가 1.2인 학생이 701점을 받은 경우도 있었고, gpa가 3.0인 학생이 529점을 받은 경우도 있었다.

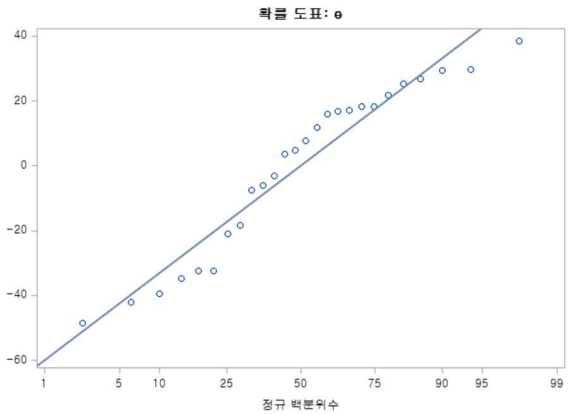
8-8

코드

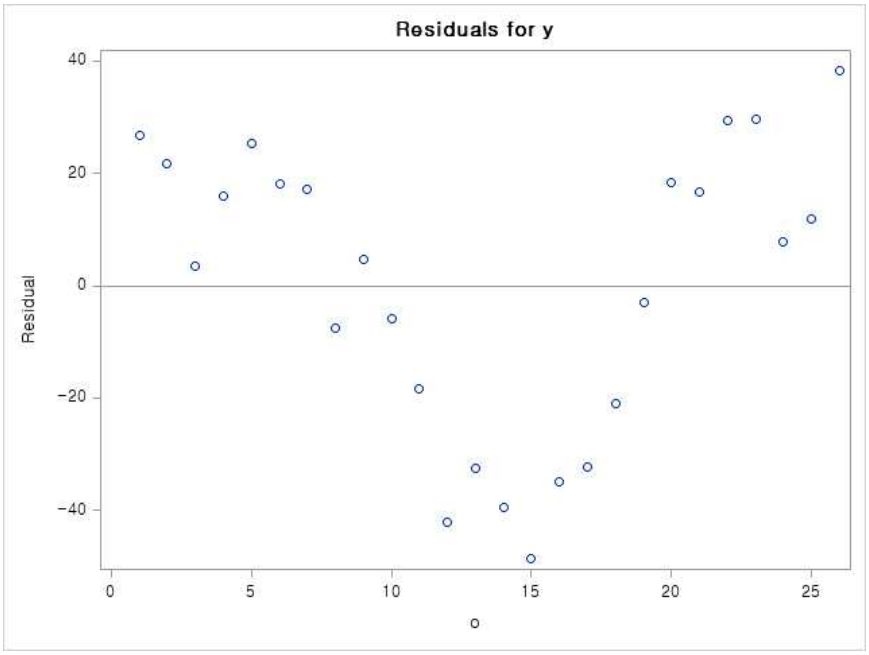
```
data exe8_8;
input y yhat;
cards;
(data)
;
run;
data exe8_8_1;
set exe8_8;
e = y - yhat;
o = _N_;
run;
proc reg data = exe8_8_1;
        model y = o / R DW;
        plot student.*o;
run;

proc univariate data = exe8_8_1 noprint;
        probplot e / normal(mu=est sigma=est);
run;
```

결과





Durbin-Watson D	0.297
Number of Observations	26
1st Order Autocorrelation	0.786



잔차에서 어떠한 경향성이 확인되었다. 표준화잔차 도표를 보면 이차함수의 경향이 보인다. 더빈왓슨 값도 0에 가깝다.
이는 년도가 어떠한 모집단에서 가져온 값이 아닌, 1씩 증가하는 값이기 때문으로 예상된다.