Chapter 5. 두 변수의 관계

김남형 응용통계학과

가천대학교

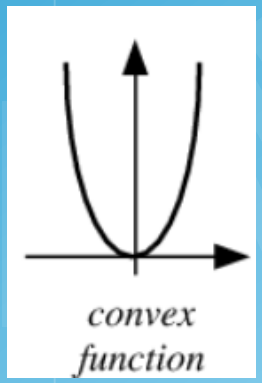
nhkim@gachon.ac.kr

- 직선화의 방법

두 변수사이에 어떤 관계가 있는가?

X(독립•설명변수), Y(종속•반응변수) X ⟶ Y

$Y = s(X) + \varepsilon$ ⇨ $s(X)$ 가 직선인 경우 $s(X) = a + bX$

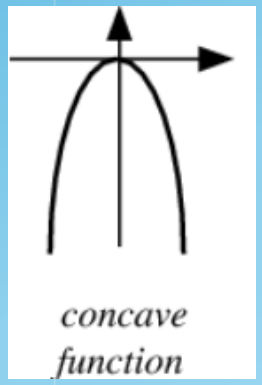


함수가 볼록 함수인 경우 : $Y \to Y^{1/2}, Y^{1/3}, \log(Y), -Y^{-1}, \cdots$

$X \to X^2, X^3, X^4, \cdots, \exp(X)$

(재표현의 사다리를 내려가는 방향으로 Y를 변환) 또는

(올라가는 방향으로 X를 변환)



함수가 오목 함수인 경우 : $Y \to Y^2, Y^3, Y^4, \cdots, \exp(Y)$

$X \to X^{1/2}, X^{1/3}, \log(X), -X^{-1}, \cdots$

(재표현의 사다리를 내려가는 방향으로 X를 변환) 또는

(올라가는 방향으로 Y를 변환)

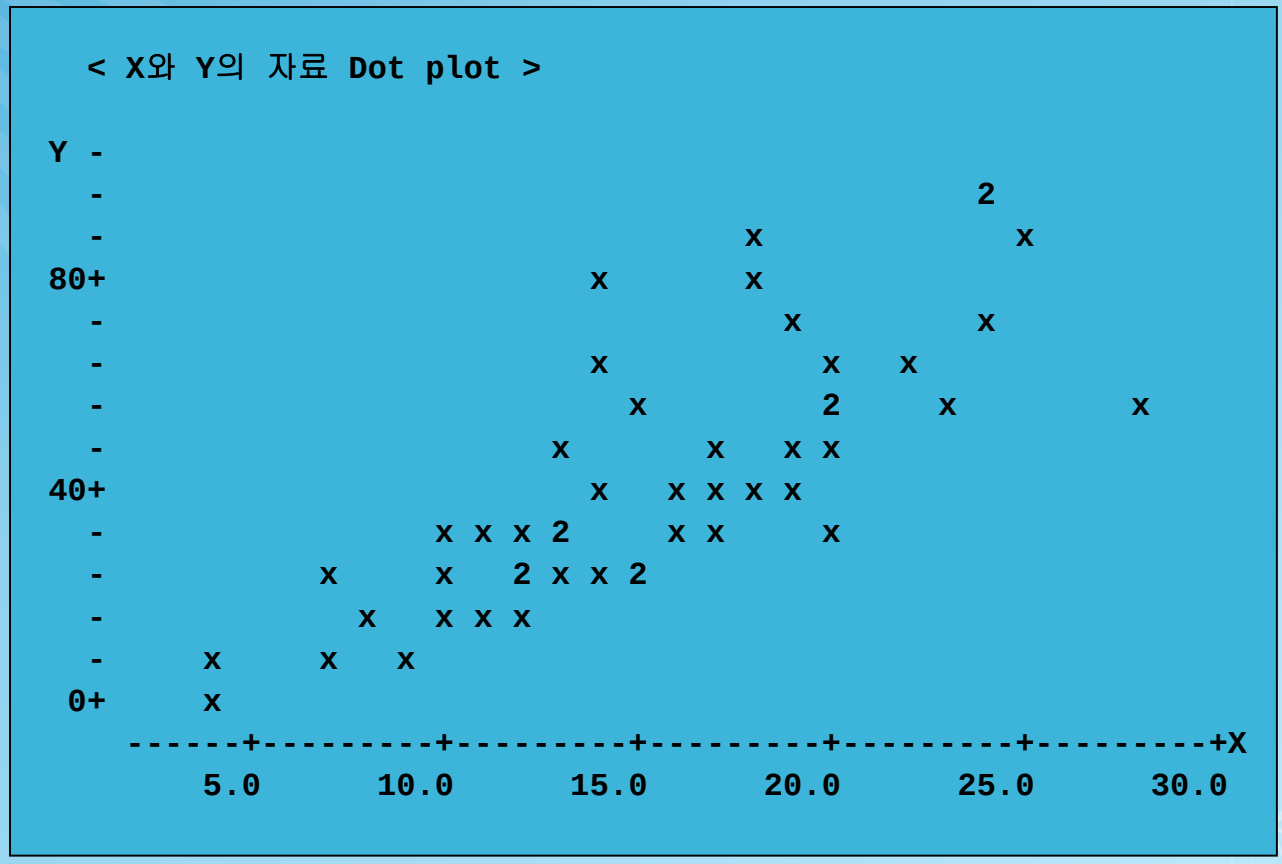
<직선화에 관한 사례>

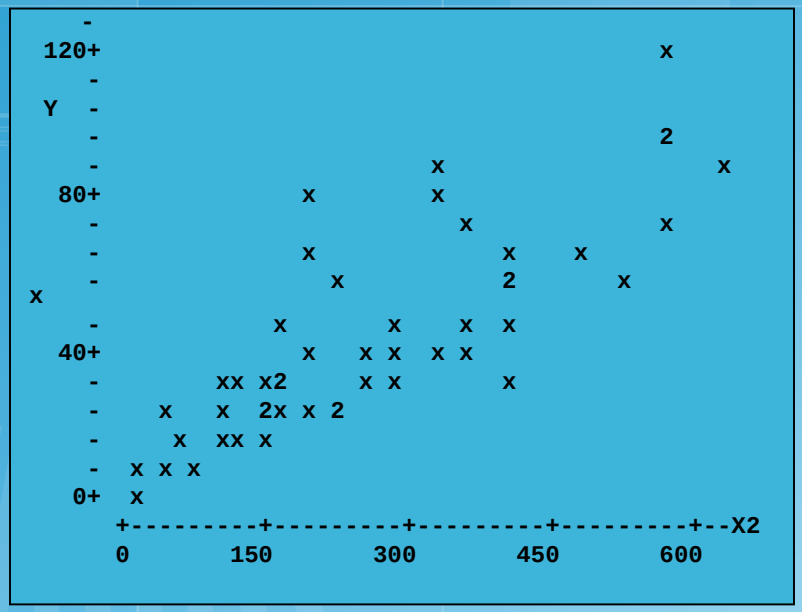
자동차 속도 X와 급제동에 실제 요구되는 거리 Y에 관한 자료

<원 자료>

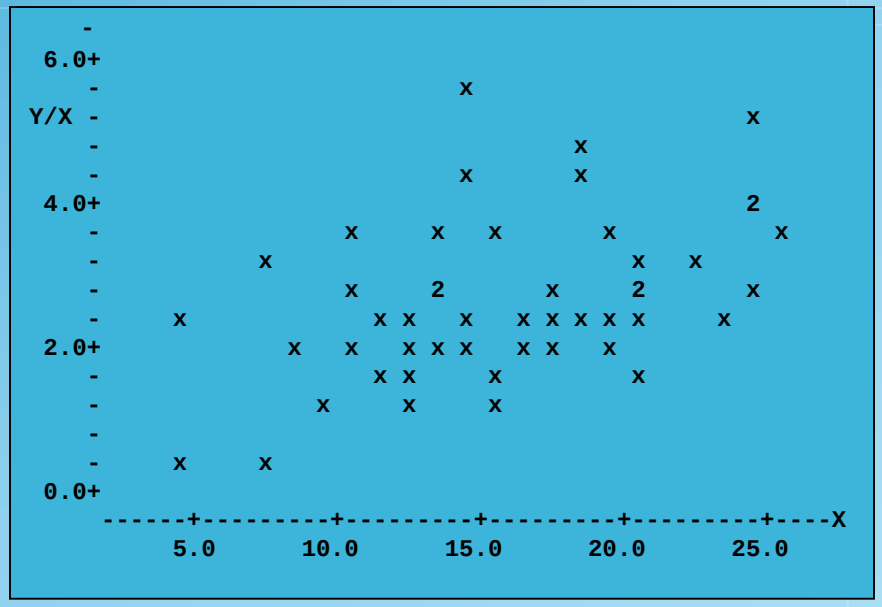
edach5-1.MTW ***

↓	C1	C2
	X	Y
1	4	2
2	7	4
3	4	10
4	7	22
5	8	16
6	9	10
7	10	18
8	11	17
9	10	26
10	11	28
11	10	34
12	12	14
13	13	26
14	14	26





$X \rightarrow X^2$



$Y \rightarrow Y / X$

원 자료의 plot이 곡선 형태를 뛰고 있다. 즉, $Y \propto X^2$ 일 가능성을 제시

$X \rightarrow X^2$ 으로 설명변수 변환 ⇒ 직선형태로 변환(산포가 커짐)

분산의 성질 $Var(aZ) = a^2 Var(Z)$ 에서

제동거리 Y 대신 속도당 제동거리 Y/X를 세로축으로 변환(산포의 균일성, 직선형태 확보)

- 세 그룹 저항성 직선의 초기 추정

  직선 경향의 두 변수간의 회귀식 ⇒ $y = a + bx$

  회귀직선의 추정방법(최소제곱법 : least square method)

  $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ 을 최소화 하는 방법

  계산이 간단하고, 정규성의 가정 아래에서 통계적 추론

  문제점) 특이점에 영향을 크게 받는다. EDA의 관점에서는 적합하지 않다.

  해결책) 자료점에 저항성이 큰 세 그룹 저항성 직선(three-group resistant line : RLINE)

1) 세 그룹(three group)

자료 점들을 크기 순으로 나열

$x_1 \le x_2 \le \cdots \le x_n$ 을 세 그룹 – 왼쪽그룹, 가운데 그룹, 오른쪽 그룹으로 나눈다.

자료수의 할당

그룹	$n = 3k$	$n = 3k+1$	$n = 3k+2$
L	k	k	$k+1$
M	k	$k+1$	k
R	k	k	$k+1$

2) 요약점

각 그룹에서의 자료점들로 부터 x와 y의 중위수를 구한다.

예를들면, (1,8), (2,6), (3,9), (4,1), (5,10) ⟹ 요약점(3,8)

왼쪽 그룹(L 그룹), 가운데 그룹(M 그룹), 오른쪽 그룹(R 그룹)의 요약점

$$(x_L, y_L), (x_M, y_M), (x_R, y_R)$$

3) 기울기 b와 절편 a의 초기 추정

회귀직선 $y = a + bx$ 에서의 기울기 b 는 L 그룹과 R 그룹의 요약점 (x_L, y_L) 과 (x_R, y_R) 을 연결하는 직선의 기울기 즉,

$$b = \frac{y_R - y_L}{x_R - x_L}$$

세 요약점에서의 잔차 r_L, r_M, r_R 의 합을 0으로 하도록 절편 a 를 정함

$$r_L = y_L - a - bx_L, r_M = y_M - a - bx_M, r_R = y_R - a - bx_R$$

따라서 $r_L + r_M + r_R = 0$ ⟹ $a = \{(y_L - bx_L) + (y_M - bx_M) + (y_R - bx_R)\}/3$

수치 예 : 사립초등학교 학생 자료

그린버그의 1953년 연구에서 나온 사립 초등학교 학생의 나이와 키에 관한 연구

나이의 변수 : x (측정단위 : 달수), 키의 변수 : y (측정단위 : cm) n=18명

	학생번호	나이(x)	키(y)	잔차 r (r=y−83.11−0.4933x)
L	1	109	137.6	0.72
	2	113	147.8	8.95
	3	115	136.8	−3.04
	4	116	140.7	0.37
	5	119	132.9	−9.11
	6	120	145.4	3.09
M	7	121	135.0	−7.80
	8	124	133.0	−11.28
	9	126	148.5	3.23
	10	129	148.3	1.55
	11	130	147.5	0.26
	12	133	148.8	0.08
R	13	134	133.2	−16.01
	14	135	148.7	−1.01
	15	137	152.0	1.31
	16	139	150.6	−1.08
	17	141	165.3	12.63
	18	142	149.9	−3.26

세 그룹 중 M 그룹의 중위수를 구해보자

x의 중위수는 (126+129)/2=127.5, y의 중위수는 (147.5+148.3)/2=147.9

세 그룹의 요약점

$$(x_L, y_L) = (115.5, 139.15), (x_M, y_M) = (127.5, 147.9), (x_R, y_R) = (138.0, 150.25)$$

기울기 b와 절편 a의 초기 추정치는

$$b = \frac{150.25 - 139.15}{138.0 - 115.5} = 0.4933$$

$$a = [(139.15 - 115.5b) + (147.90 - 127.5b) + (150.25 - 138b)]/3 = 83.11$$

결국, 세 그룹 저항성 직선 RLINE의 초기 추정식은

$$\hat{y} = 83.11 + 0.4933x$$

잔차 r은

$$r = y - 83.11 - 0.4933x$$

잔차

```
-1. | 6
 1* | 1
-0. | 97
-0* | 3113
+0* | 00331001
 0. | 8
+1* | 2
 1. |
```

- 미니탭에서의 저항성 직선 RLINE

  좀 더 적합도가 나은 직선을 구해보자.

  a 와 b 의 초기 추정치를 a_0 와 b_0 라고 하자.

  그 때의 잔차 $r_i = y_i - a_0 - b_0 x_i$ 를 구하게 되는데 이 때 자료 $(x_1, r_1), \cdots, (x_n, r_n)$ 에 세 그룹 저항성 직선을 같은 방법으로 적합

$$r = \alpha_0 + \beta_0 x \quad \Rightarrow \quad y - a_0 - b_0 x = \alpha_0 + \beta_0 x$$
$$\Rightarrow y = a_0 + b_0 x + \alpha_0 + \beta_0 x = (a_0 + \alpha_0) + (b_0 + \beta_0) x$$

  가 된다. 이 관계로 부터 절편 a 와 기울기 b 의 새로운 추정치

$$a_1 = a_0 + \alpha_0, b_1 = b_0 + \beta_0$$

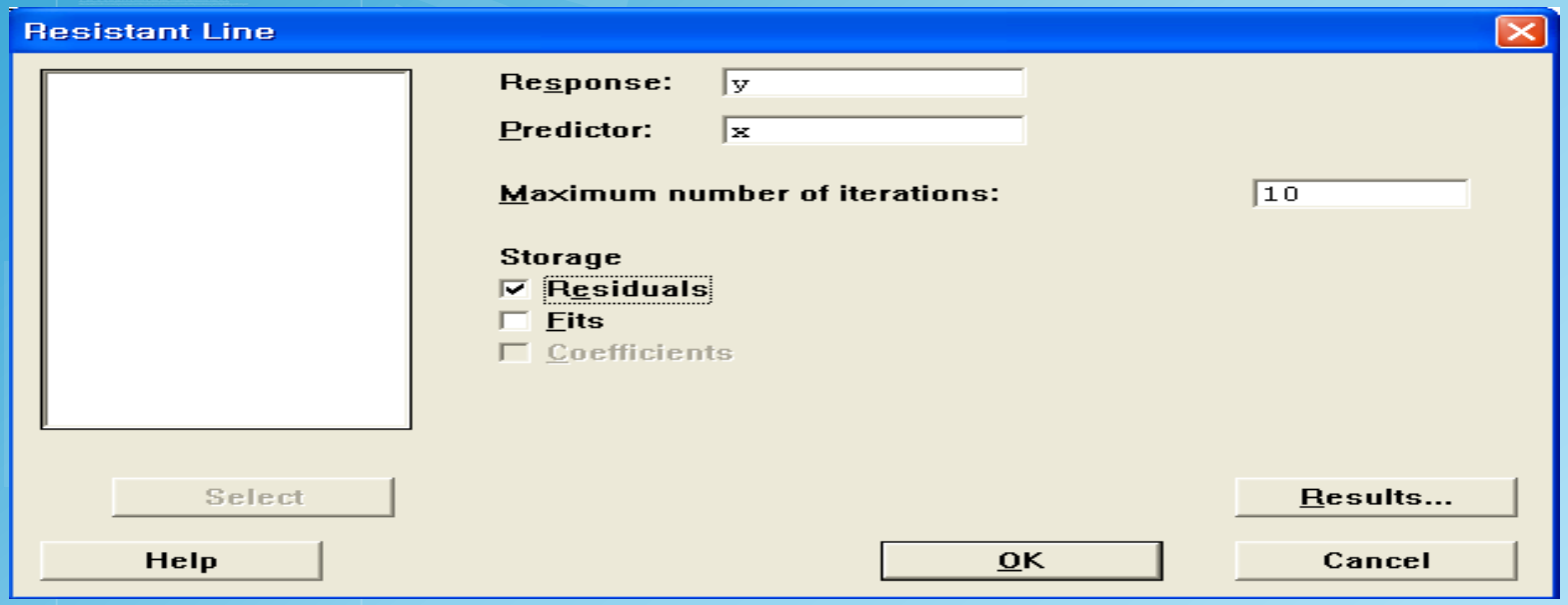
  를 얻게 된다.

  새로운 잔차를 구하고, 이들 잔차에 새 저항성 직선을 적합시켜 a 와 b 의 추정치 수정 반복 계산을 계속하되 추정치의 변화량이 아주 작으면 더 이상의 작업을 정지

[미니탭 활용]

저항성 직선 RLINE명령 사용

Stat > EDA > Resistant Line....



Resistant Line Fit: y versus x(OUTPUT)

```
Slope    = 0.4286        Level = 91.0071        Half-slope ratio = 0.307
```

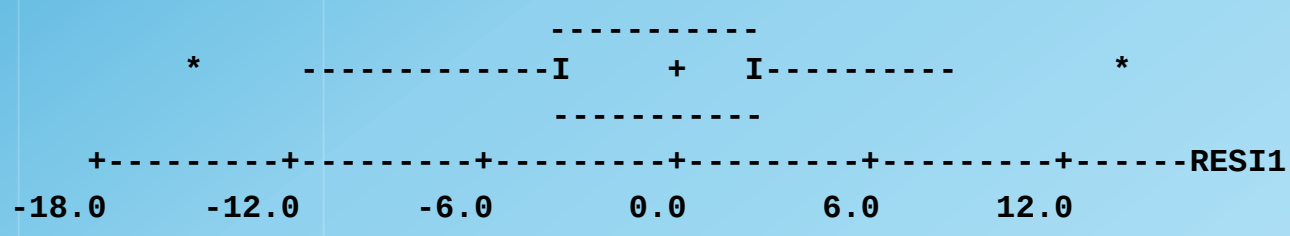
$$b_L = (y_M - y_L)/(x_M - x_L), b_R = (y_R - y_M)/(x_R - x_M)$$

Half-slope ratio= b_R / b_L

@ 잔차의 줄기 그림(Graph > Character Graphs > Stem and Leaf....)

```
Stem-and-leaf of RESI1     N  = 18
Leaf Unit = 1.0
     1    -1 5
     2    -1 1
     4    -0 97
     9    -0 31000
     9     0 0002223
     2     0 8
     1     1 3
```

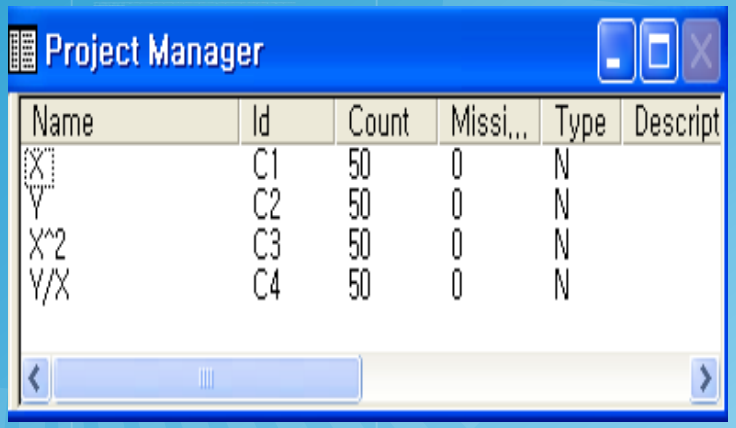
@ 상자 그림 (Graph > Character Graphs > Boxplot....)

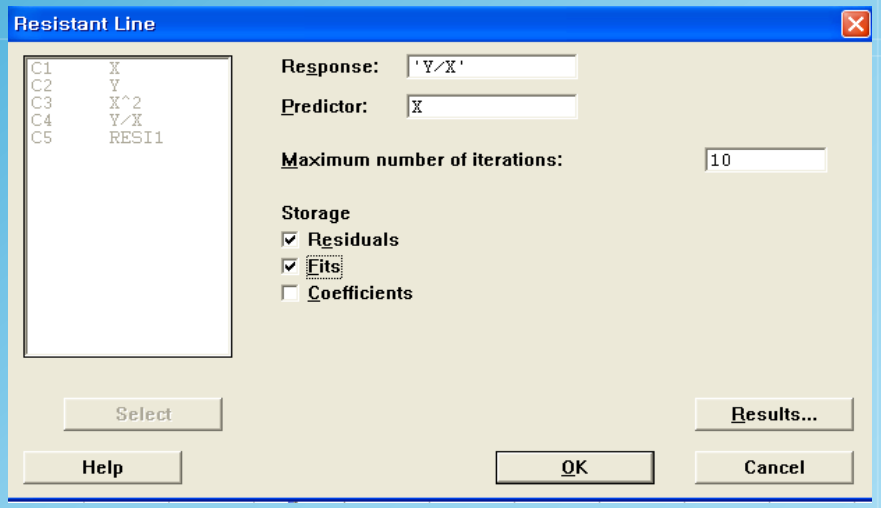


최종적인 저항성 직선의 적합식은

$$\hat{y} = 91.0071 + 0.4286x$$

<자동차의 속도와 제동거리의 자료>





위 자료를 이용한 저항성 직선 (Stat > EDA > Resistant Line)

<출력>

Resistant Line Fit: Y/X versus X

```
Slope    =   0.0667     Level =   1.4500      Half-slope ratio =    0.920
```

⇒ 적합 된 회귀식 : Y/X = 1.4500 + 0.0667X

Y = 1.4500X + 0.0667X^2

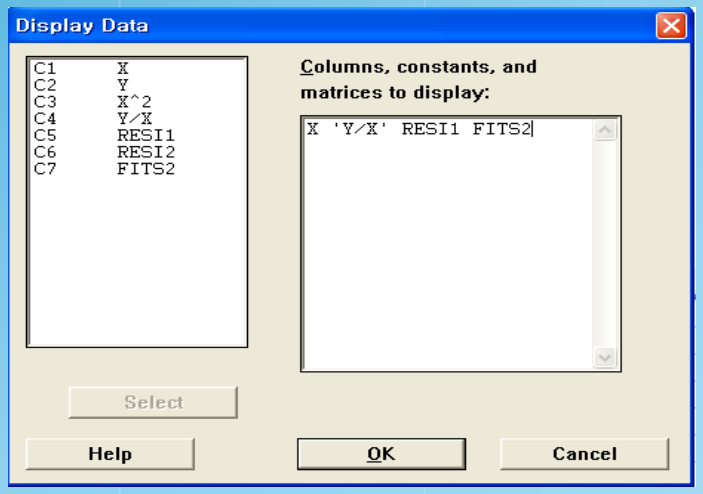
<잔차 줄기 그림과 상자 그림>

```
Stem-and-leaf of RESI1     N  = 50
Leaf Unit = 0.10
    6    -1 332110
   15    -0 987766555
   25    -0 3333222110
   25     0 000012223344
   13     0 7788
    9     1 1222
    5     1 599
    2     2 0
    1     2
    1     3 3
```

```
          --------------
   -------I      +       I-----------              *
          --------------
-------+---------+---------+---------+---------+--------RESI1
     -1.0       0.0       1.0       2.0       3.0
```

<자료의 표현>

Manip > Display Data....



Data Display

Row	X	Y/X	RESI1	FITS2
1	4	0.50000	-1.21667	1.71667
2	7	0.57143	-1.34524	1.91667
3	4	2.50000	0.78333	1.71667
4	7	3.14286	1.22619	1.91667
5	8	2.00000	0.01667	1.98333
25	13	3.53846	1.22179	2.31667
26	14	5.71429	3.33095	2.38333
27	16	2.00000	-0.51667	2.51667
28	17	1.88235	-0.70098	2.58333
47	25	3.40000	0.28333	3.11667
48	24	3.83333	0.78333	3.05000
49	24	3.87500	0.82500	3.05000
50	24	5.00000	1.95000	3.05000

❑ 요약

○변환(재표현) 볼록 함수인 경우 : $Y \to Y^{1/2}, Y^{1/3}, \log(Y), -Y^{-1}, \cdots$

$X \to X^2, X^3, X^4, \cdots, \exp(X)$

(재표현의 사다리를 내려가는 방향으로 Y를 변환) 또는 (올라가는 방향으로 X를 변환)

○변환(재표현) 오목 함수인 경우 : $Y \to Y^2, Y^3, Y^4, \cdots, \exp(Y)$

$X \to X^{1/2}, X^{1/3}, \log(X), -X^{-1}, \cdots$

(재표현의 사다리를 내려가는 방향으로 X를 변환) 또는 (올라가는 방향으로 Y를 변환)

○세 그룹 저항성 직선 RLINE를 하는 EDA의 절차

1) 자료를 크기가 가급적 같은 세 그룹으로 나누고

2) 각 그룹의 요약 점을 정한다

3) 세 요약 점 중 양끝 점을 이용하여 기울기를 구하고,
세 점의 잔차의 합이 0이되도록하여 절편을 구한다.

4) 잔차를 구하고, 같은 방법으로 잔차를 회귀시켜 얻은 결과를 이용하여 보다 나은 추정치를 구한다.

○잔차를 분석하여 회귀분석에서의 특이점을 식별해 낸다.(EDA 기법)