10강

모형진단

정보통계학과 김성수교수

학습목차

1. 모형진단이란
2. 치료

1 모형진단이란

모형진단이란

✓ 모형진단

모형이나 가정에 문제점이 있는지를 알아보는 것

=> 모형진단이 필요한 이유는 우리가 사용하는 회귀모형이나 회귀모형에 대해 세운 가정이 실제 문제에서는 적절하지 못한 경우가 종종 발견되기 때문임.

=> 잔차의 형태로부터 모형과 가정의 적정성 여부를 알 수 있음.

✓ 회귀모형에 부여되는 가정

1) 오차의 등분산성 가정. 모든 X값에 대해서 Y의 분산은 같다.

2) 모형의 선형성 가정. Y와 X사이에는 선형의 관계가 있다.

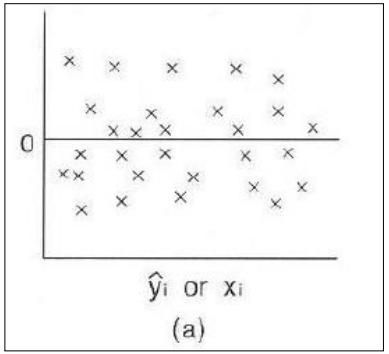
3) 오차의 정규성 가정. 모든 X값에 대해서 Y의 분포는 정규분포를 따른다.

모형진단이란

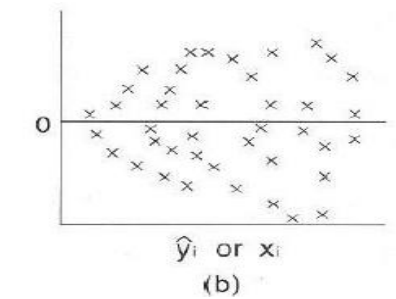
✓ **가정의 타당성을 알아볼 수 있는 가장 보편화된 방법**

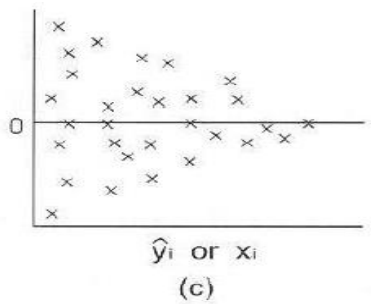
- 잔차나 스튜던트화잔차를 세로축으로 하고 $\hat{Y}$ 을 가로축으로 하는 **잔차산점도**를 그려보는 것임.
- 각각의 설명변수를 가로축으로 하는 경우에는 이러한 산점도를 **잔차-설명변수 산점도**라 부름.
- 모형에 문제가 있거나 가정이 위배되었다면 잔차는 잔차산점도에 영향을 줄 것임.

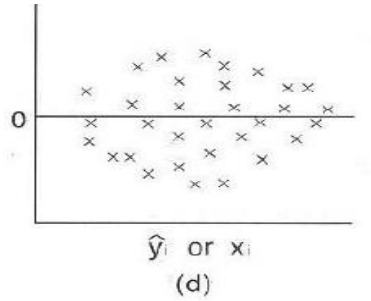
잔차산점도의 형태–1



=>회귀모형이 제대로 적합 되었을 때의 산점도

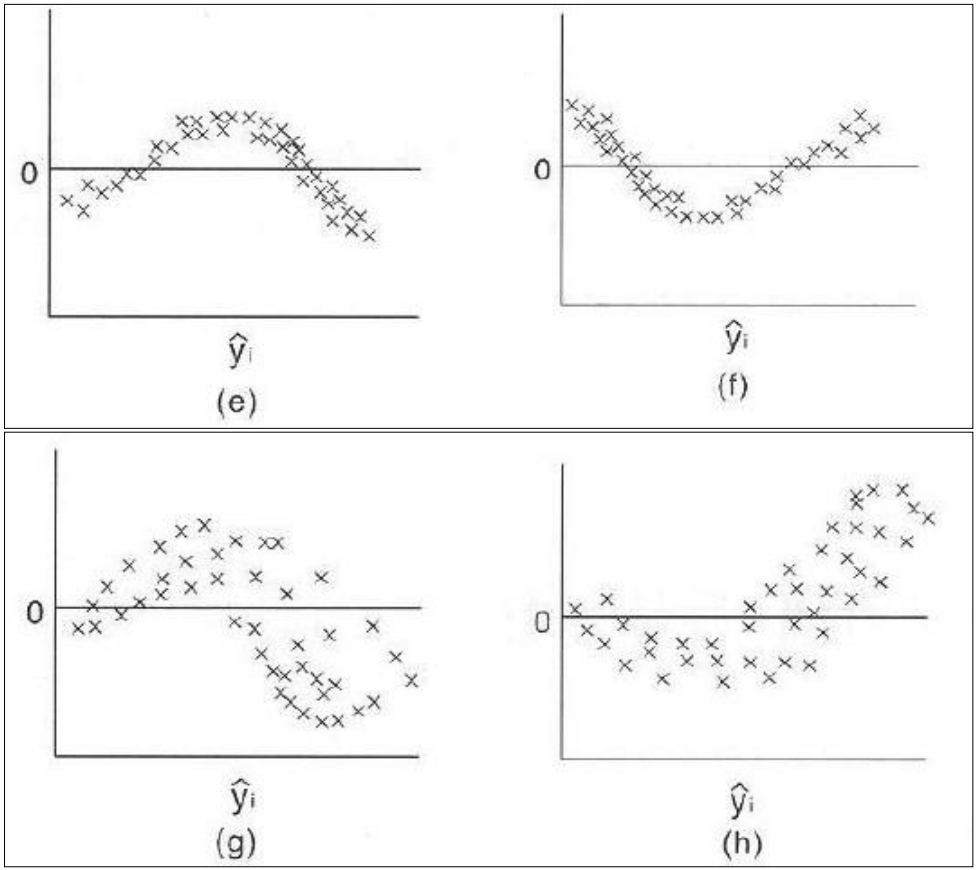






$\hat{Y}$ 이 증가함에 따라 세로축의 값인 잔차나 스튜던트화잔차의 퍼짐의 정도가
(b)에서는 증가, (c)에서는 감소, (d)에서는 증가하다가 다시 감소.
이러한 산점도는 분산이 일정하지 않은, 즉 이분산성을 나타내고 있는 것임.

잔차산점도의 형태–2



⇒ $\hat{Y}$ 이 증가함에 따라 잔차나 스튜던트화 잔차가 곡선의 형태를 보여줌.

- 회귀모형의 비선형성

⇒ 이분산성과 비선형성을 함께 나타내고 있음

오차의 등분산성 가정

모든 관측값에 대해 오차의 분산이 같다. 즉, $Var(\varepsilon_i) = \sigma^2$

- 오차의 등분산성 여부를 판단하는 방법

1) 잔차산점도를 이용. 잔차나 스튜던트잔차를 세로축으로 하고 $\hat{Y}$ 을 가로축으로 하는 산점도에서 $\hat{Y}$ 이 증가함에 따라 세로축의 값의 퍼짐의 정도가 증가 또는 감소하는 모양의 산점도는 분산이 일정치 않음을 나타냄.

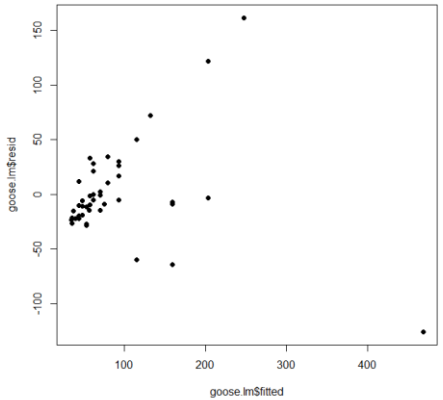
2) 스코어검정(score test) 이용 : 그림을 통한 오차의 등분산성을 찾아내는 데는 주관적이고, 경험적인 판단이 필요한 경우가 많은데 이러한 점을 보완하는 통계적인 방법으로 스코어검정을 실시. Cook과 Weisberg (1983)가 제안한 방법임. 잔차산점도로 충분히 진단이 가능한 경우에는 이러한 스코어 검정은 실시하지 않아도 무방함.

R 활용 : 오차의 등분산

〈흰기러기 무리에 대한 크기 추정〉

사진 (Y)	관측자 (X)	사진 (Y)	관측자 (X)	사진 (Y)	관측자 (X)
56	50	26	30	73	50
38	25	88	75	123	75
25	30	56	35	150	150
48	35	11	9	70	50
38	25	66	55	90	60
22	20	42	30	110	75
22	12	30	25	95	150
42	34	90	40	57	40
34	20	119	75	43	25
14	10	165	100	55	100
30	25	152	150	325	200
9	10	205	120	114	60
18	15	409	250	83	40
25	20	342	500	91	35
62	40	200	200	56	20

```
> goose = read.table("c:/data/reg/goose.txt", header=T)
> head(goose,3)
  photo obsA obsB
1    56   50   40
2    38   25   30
3    25   30   40
> goose.lm = lm(photo ~ obsA, data=goose)
> plot(goose.lm$fitted, goose.lm$resid, pch=19)
```



잔차산점도는 X가 증가함에 따라 잔차의 흩어짐이 많아짐 => 이분산성이 의심됨

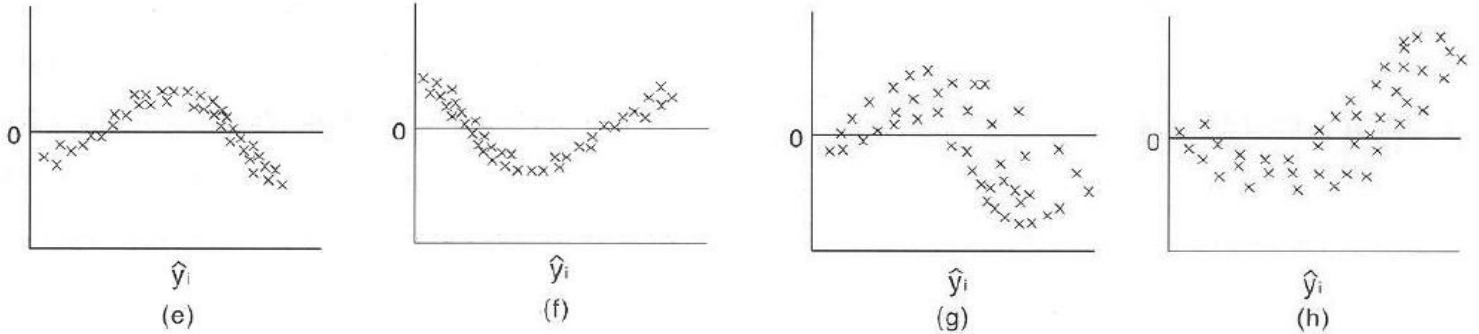
R 활용 : 오차의 등분산 – 스코어 검정

```
> library(car)
> ncvTest(goose.lm)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 81.41318    Df = 1     p = 1.831324e-19
```

스코어 검정의 χ^2=81.41 이고, 유의확률 p-값이 매우 작으므로 등분산 가정을 기각

회귀모형의 선형성

✓ 모형의 비선형성



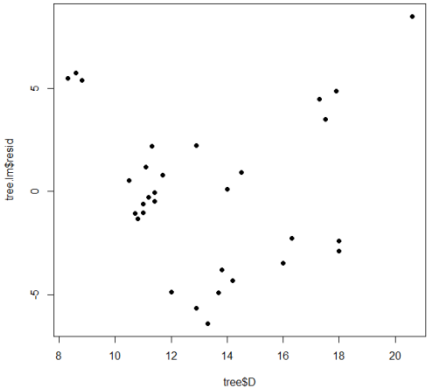
⇒ 그림 (e)-(h)에서와 같이 $\hat{Y}$ 또는 가로축의 값이 증가함에 따라 세로축의 값이 곡선의 형태를 보이는 경우 이는 반응변수와 설명변수의 관계에 비선형성이 있음을 나타냄.

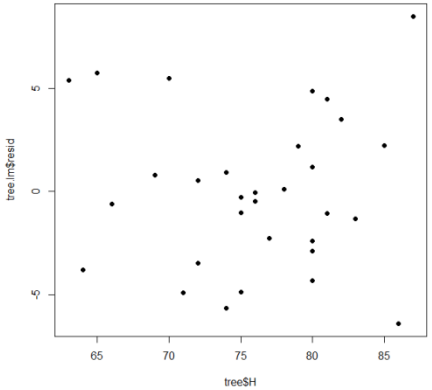
※ 비선형인 경우, 변수의 적절한 변환을 통해서 선형으로 변환

R 활용 : 선형성

〈나무자료〉

순서	지름	높이	부피	순서	지름	높이	부피	순서	지름	높이	부피
1	8.3	70	10.3	12	11.4	76	21.0	23	14.5	74	36.3
2	8.6	65	10.3	13	11.4	76	21.4	24	16.0	72	38.3
3	8.8	63	10.2	14	11.7	69	21.3	25	16.3	77	42.6
4	10.5	72	16.4	15	12.0	75	19.1	26	17.3	81	55.4
5	10.7	81	18.8	16	12.9	74	22.2	27	17.5	82	55.7
6	10.8	83	19.7	17	12.9	85	33.8	28	17.9	80	58.3
7	11.0	66	15.6	18	13.3	86	27.4	29	18.0	80	51.5
8	11.0	75	18.2	19	13.7	71	25.7	30	18.0	80	51.0
9	11.1	80	22.6	20	13.8	64	24.9	31	20.6	87	77.0
10	11.2	75	19.9	21	14.0	78	34.5				
11	11.3	79	24.2	22	14.2	80	31.7				






```
> tree = read.table("c:/data/reg/tree.txt", header=T)
> head(tree, 3)
  num    D  H    V
1   1 8.3 70 10.3
2   2 8.6 65 10.3
3   3 8.8 63 10.2
> tree.lm = lm(V ~ D+H, data=tree)
> plot(tree$D, tree.lm$resid, pch=19)
> plot(tree$H, tree.lm$resid, pch=19)
```

변수 D의 잔차산점도의 경우 2차 함수 형태의 비선형성이 나타남.

오차의 정규성

- 오차가 정규분포를 따른다는 가정.
- 회귀모형의 적합이후의 통계적 추론에 중요한 역할을 함

➢ 정규성 가정 진단 : 잔차 또는 스튜던트화 잔차의 정규성을 검토하는 정규확률그림(normal probability plot) 이나 Shapiro와 Wilk(1965)가 제안한 W통계량 이용.

➢ 정규확률그림 해석 : 만약 표본이 정규분포에서 얻은 것이라면 정규확률그림은 직선에 근접.

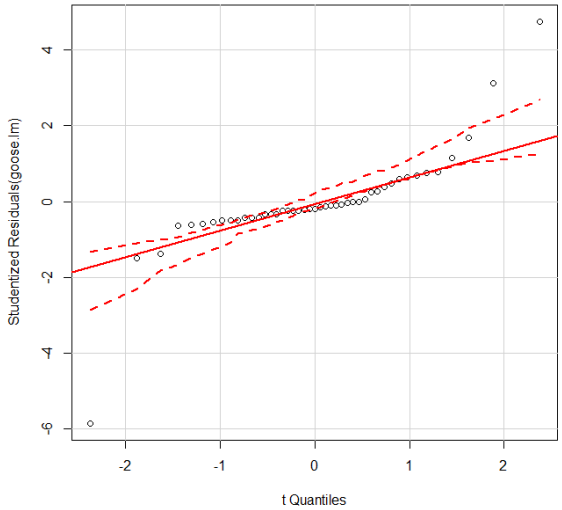
R 활용 : 오차의 정규성

```
> goose.lm = lm(photo ~ obsA, data=goose)
> qqPlot(goose.lm)
> # 정규성 검정
> library(mvnormtest)
> goose.rstudent = rstudent(goose.lm)
> shapiro.test(goose.rstudent)

        Shapiro-Wilk normality test

data:  goose.rstudent
W = 0.7192, p-value = 5.971e-08
```

W 통계량의 값은 0.7192 이고, 유의확률 p-값이 매우 작으므로 정규성 가정을 기각함.



잔차가 직선의 형태를 벗어나 곡선의 형태로 직선에서 벗어나고 있음을 보이므로 정규성 가정에 위배되는 것으로 판단.

2 치료

분산안정을 위한 변환

✓ **오차의 등분산성 가정이 위배 되는 경우**

- 최소제곱법으로 추정된 추정량은 불편추정량의 성질은 만족시키지만
  추정량의 분산이 커질 가능성이 높아짐

➢ **이분산성 치료방법**

1) 첫째 방법으로는 **가중최소제곱**을 이용. 여기서 가중값은 분산의 역에 비례하도록 제공될 수 있는데 일반적으로 가중값은 경험적으로 채택됨.

2) 둘째 방법으로는 **등분산변환방법**(variance stabilizing trasformation) 이용. Y의 값이 양의 값이고, 분산의 크기가 $E(Y)$ 에 의존하면 거의 모든 경우에 적절한 반응변수의 변환으로 분산의 변동을 안정시키고 등분산성의 가정을 만족시켜줄 수 있음

분산안정을 위한 변환

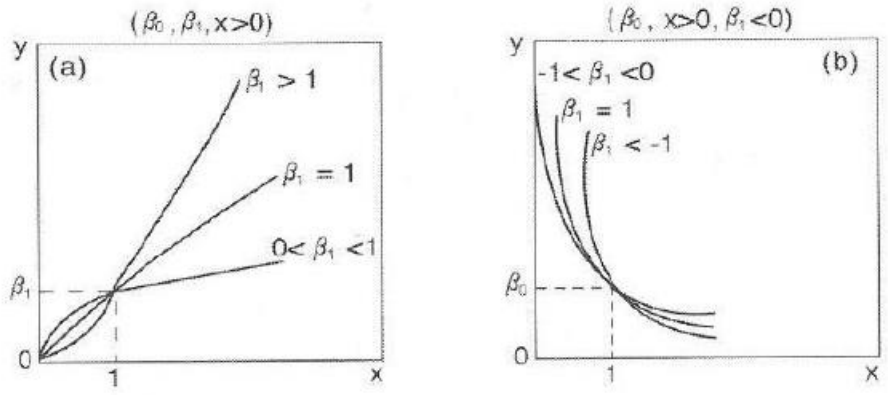
✓ **등분산변환의 예**

변환	상황	비고
$\sqrt{Y}$	$Var(\varepsilon_i) \propto E(Y_i)$	반응변수가 포아송분포를 따르는 계수(counts)의 값인 경우
$\ln Y$	$Var(\varepsilon_i) \propto [E(Y_i)]^2$	보편적으로 많이 쓰임. 특히 반응변수의 값의 범위가 넓은 경우
$1/Y$	$Var(\varepsilon_i) \propto [E(Y_i)]^4$	대부분의 반응변수의 값이 0에 가까우나 일부가 그렇지 못한 경우

선형 변환

✓ 비선형인 경우에는 적절한 변환을 통하여 선형모형으로 문제를 해결

1) ***Y***와 ***X***의 관계가 다음과 같은 멱등함수로 설정이 된 경우



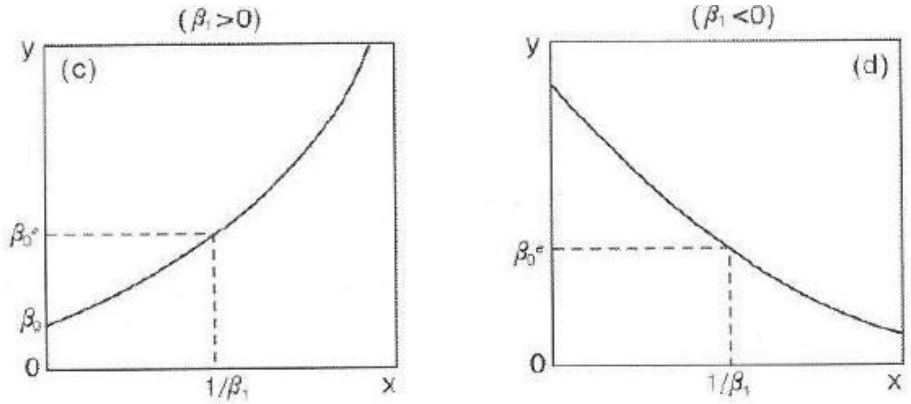
=> 양변에 로그를 취하면 다음과 같은 선형의 모형을 얻을 수 있다.

$$Y = \beta_0 X^{\beta_1} \quad \Rightarrow \quad \ln Y = \ln \beta_0 + \beta_1 \ln X$$

선형 변환

✓ 비선형인 경우에는 적절한 변환을 통하여 선형모형으로 문제를 해결

2) 지수함수를 취하는 경우



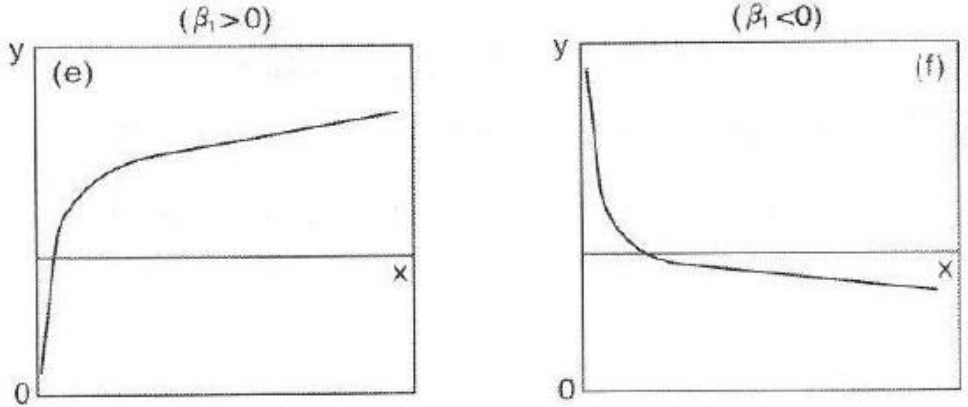
=> 양변에 로그

$$Y = \beta_0 e^{\beta_1 X} \quad \Rightarrow \quad \ln Y = \ln \beta_0 + \beta_1 X$$

선형 변환

✓ 비선형인 경우에는 적절한 변환을 통하여 선형모형으로 문제를 해결

3) 로그함수 형태인 경우

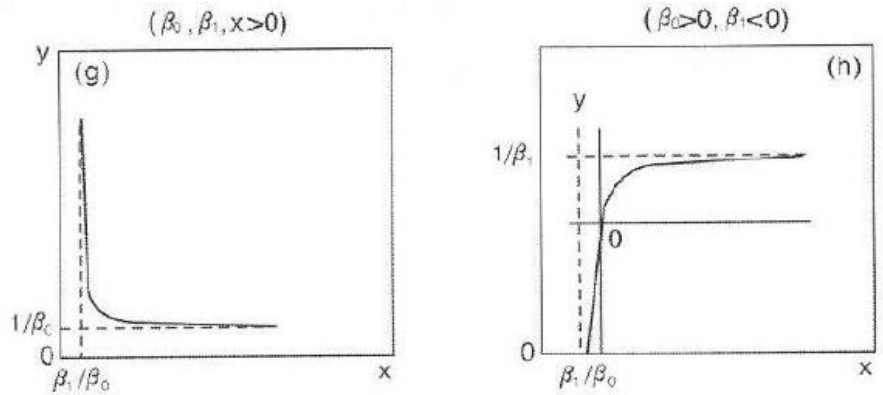


$$Y = \beta_0 + \beta_1 \ln X = \beta_0 + \beta_1 X^*$$

선형 변환

✓ 비선형인 경우에는 적절한 변환을 통하여 선형모형으로 문제를 해결

4) 쌍곡선 형태인 경우



$$Y=\frac{X}{\beta_0 X-\beta_1} \quad \Rightarrow \quad Y^*=\frac{1}{Y}=\frac{\beta_0 X-\beta_1}{X}=\beta_0-\beta_1\left(\frac{1}{X}\right)=\beta_0-\beta_1 X^*$$

반응변수의 변환

✓ Box와 Cox 멱등변환

Box와 Cox(1964)는 변환의 선택 문제를 해결할 수 있는 절차적인 방법을 제시
: 이들은 주어진 모형을 확대하여 변환방법을 찾는 문제를 모수의 추정의 문제로 접근함.

➢ **Box와 Cox 멱등변환**

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda \, [\, GM(\boldsymbol{Y})\,]^{\lambda - 1}} \;, & \lambda \neq 0 \\ GM(\boldsymbol{Y}) \, \ln(y_i) \;, & \lambda = 0 \end{cases}$$

, 여기서 $GM(Y) = (Y_1 Y_2 \cdots Y_n)^{1/n}$ **는 기하평균.**

반응변수의 변환 예

✓ λ 에 따른 SSE_λ 의 변환

λ	SSE_λ
-2.0	34101.038
-1.0	986.042
-0.5	291.583
0.0	134.094
0.5	96.949
1.0	126.866
2.0	1275.556

=> 최소의 λ 의 값을 가져다 주는 SSE_λ 의 값은 약 0.5정도임을 알 수 있음.

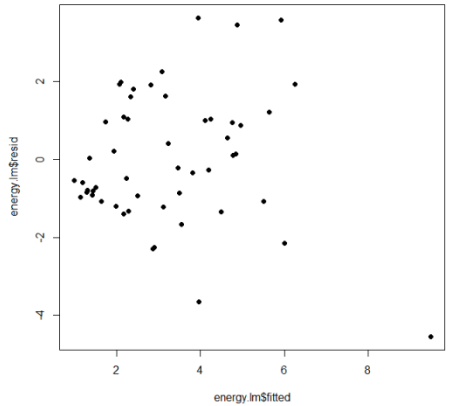
이는 $\sqrt{}$ 변수변환에 해당

R 활용 : Box–Cox 변환

〈53명의 주거지역 고객에 대한 수요(Y)와 에너지사용량(X)〉

고객	X(KWH)	Y(KW)	고객	X(KWH)	Y(KW)	고객	X(KWH)	Y(KW)
1	679	0.79	19	745	0.77	37	770	1.74
2	292	0.44	20	435	1.39	38	724	4.10
3	1012	0.56	21	540	0.56	39	808	3.94
4	493	0.79	22	874	1.56	40	790	0.96
5	582	2.70	23	1543	5.28	41	783	3.29
6	1156	3.64	24	1029	0.64	42	406	0.44
7	997	4.73	25	710	4.00	43	1242	3.24
8	2189	9.50	26	1434	0.31	44	658	2.14
9	1097	5.34	27	837	4.20	45	1746	5.71
10	2078	6.85	28	1748	4.88	46	468	0.64
11	1818	5.84	29	1381	3.48	47	1114	1.90
12	1700	5.21	30	1428	7.58	48	413	0.51
13	747	3.25	31	1255	2.63	49	1787	8.33
14	2030	4.43	32	1777	4.99	50	3560	4.94
15	1643	3.16	33	370	0.59	51	1495	5.11
16	414	0.50	34	2316	8.19	52	2221	3.85
17	354	0.17	35	1130	4.79	53	1526	3.93
18	1276	1.88	36	463	0.51			

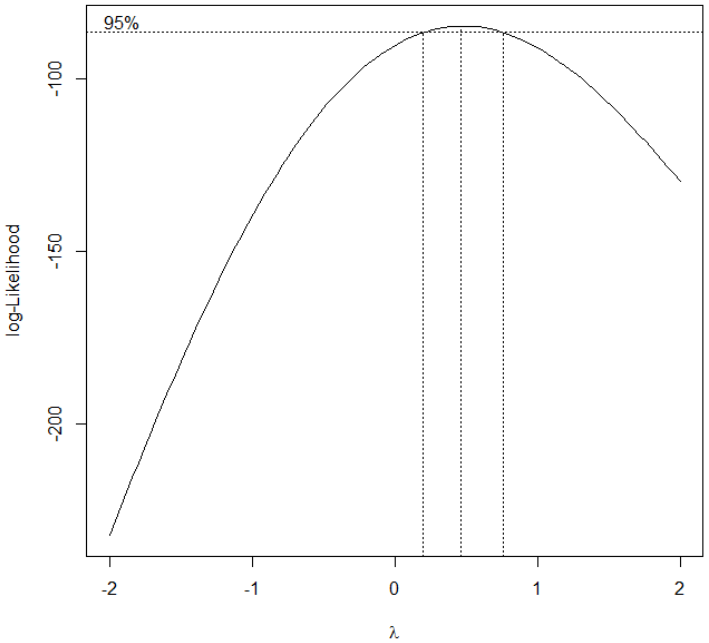
```
> energy = read.table("c:/data/reg/energy.txt", header=T)
> head(energy,3)
  customer    X    Y
1        1  679 0.79
2        2  292 0.44
3        3 1012 0.56
> energy.lm = lm(Y ~ X, data=energy)
> plot(energy.lm$fitted, energy.lm$resid, pch=19)
```



잔차산점도는 X가 증가함에 따라 잔차의 흩어짐이 많아짐 => 이분산성이 의심됨

R 활용 : Box-Cox 변환

```
> library(MASS)
> boxcox(Y~X, data=energy, lambda=seq(-2,2, 1/2), plotit=TRUE)
```



Box-Cox 변환그림에서는 log-likelihood 값이 최대가 되는 λ 값을 찾으면 됨. 그림에서 λ 는 0.5 가 됨.
이는 $\sqrt{}$ 변환에 해당.

다음시간 안내

11강. SAS와 SPSS 활용 (7~10강)