REPORT

수강과목	:	회귀분석(II)
담당교수	:	김충락
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2019.10.28

Regression Analysis (II)
Project 1.

Due October 28, 2019

You may use any statistical packages like R, minitab, spss, sas, etc.

1. Make your own dataset based on data in Example 6.5 (p. 251). Let $X_1 <- X_1 + \epsilon$, $X_2 <- X_2 + \epsilon$, $Y <- Y + \epsilon$, where $\epsilon \sim N(0,1)$. Do the response surface analysis with contour plot.

먼저 Example 6.5 (p. 251)의 자료를 입력해준다.

```
> x1<-c(4,20,12,12,12,12,12,6.3,6.3,17.7,17.7)
> length(x1)
[1] 11
> x2<-c(250,250,250,250,220,280,250,229,271,229,271)
> length(x2)
[1] 11
> y<-c(83.8,81.7,82.4,82.9,84.7,67.9,81.2,81.3,83.1,85.3,72.7)
> length(y)
[1] 11
```

자료를 $X_1 <- X_1 + \epsilon$, $X_2 <- X_2 + \epsilon$, $Y <- Y + \epsilon$, where $\epsilon \sim N(0,1)$의 형태로 바꿔준다.

```
> X1<-x1+rnorm(n=1,mean=0,sd=1)
> X2<-x2+rnorm(n=1,mean=0,sd=1)
> Y<-y+rnorm(n=1,mean=0,sd=1)
> data<-data.frame(X1,X2,Y)
> data
          X1       X2       Y
1   4.395786 250.3226 82.84779
2  20.395786 250.3226 80.74779
3  12.395786 250.3226 81.44779
4  12.395786 250.3226 81.94779
5  12.395786 220.3226 83.74779
6  12.395786 280.3226 66.94779
7  12.395786 250.3226 80.24779
8   6.695786 229.3226 80.34779
9   6.695786 271.3226 82.14779
10 18.095786 229.3226 84.34779
11 18.095786 271.3226 71.74779
```

R에 내장된 반응표면분석을 실시하는 패키지를 다운받는다.

```
> install.packages("rsm")
```

설명변수는 X1(공정시간)과 X2(공정온도)로 2개이다.

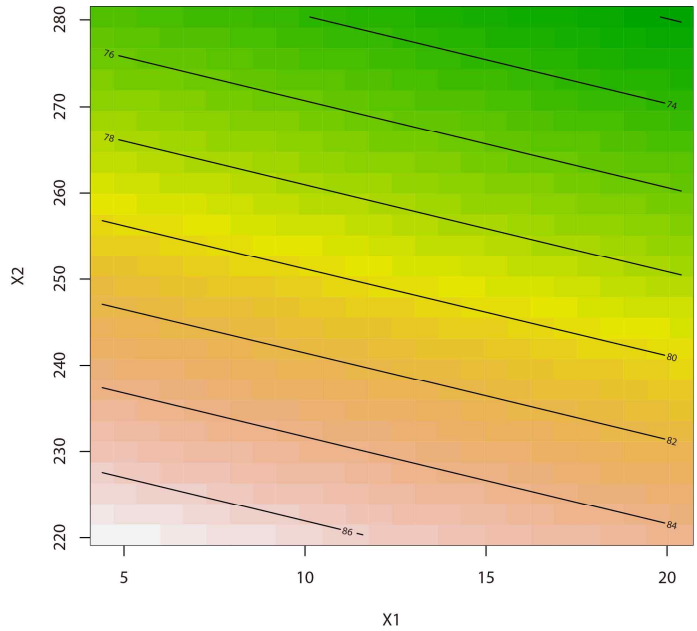
a를 first-order, b를 second-order으로 설정해서 비교한다.

```
> a<-rsm(Y~ FO(X1,X2),data=data)
> summary(a)

Call:
rsm(formula = Y ~ FO(X1, X2), data = data)

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 133.573187  17.150586  7.7883 5.294e-05 ***
X1           -0.206544   0.249835 -0.8267    0.4324
X2           -0.205051   0.067214 -3.0507    0.0158 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared:  0.5553,    Adjusted R-squared:  0.4441
F-statistic: 4.995 on 2 and 8 DF,  p-value: 0.0391

Analysis of Variance Table

Response: Y
            Df  Sum Sq Mean Sq F value  Pr(>F)
FO(X1, X2)   2 160.856  80.428  4.9951 0.03910
Residuals    8 128.810  16.101
Lack of fit  6 127.283  21.214 27.7911 0.03514
Pure error   2   1.527   0.763

Direction of steepest ascent (at radius 1):
        X1          X2
-0.7096673 -0.7045370

Corresponding increment in original units:
        X1          X2
-0.7096673 -0.7045370
```

적합 된 모형은 $\hat{Y}= 133.573187 - 0.206544X_1 - 0.205051X_2$이다.

이 모형은 Y의 반응 값에서 lack of fit이 0.03514으로 유의수준 0.05보다 작으므로 모형이 적합하다고 볼 수 없다.

> contour(a,~X1+X2, image = TRUE)



그래프에서 X1(공정시간)과 X2(공정온도)를 증가하면 Y(효율)는 감소한다.

> b<-rsm(Y~ SO(X1,X2),data=data)
> summary(b)

Call:
rsm(formula = Y ~ SO(X1, X2), data = data)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3.2956e+02	1.3672e+02	-2.4105	0.06083	.
X1	6.8972e+00	2.4744e+00	2.7874	0.03856	*
X2	3.1557e+00	1.0674e+00	2.9566	0.03164	*
X1:X2	-3.0075e-02	9.4237e-03	-3.1914	0.02423	*
X1^2	1.7134e-02	2.9597e-02	0.5789	0.58776	
X2^2	-5.9682e-03	2.1178e-03	-2.8181	0.03719	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared: 0.9121, Adjusted R-squared: 0.8243
F-statistic: 10.38 on 5 and 5 DF, p-value: 0.01128

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FO(X1, X2)	2	160.856	80.428	15.8022	0.006896
TWI(X1, X2)	1	51.840	51.840	10.1853	0.024226
PQ(X1, X2)	2	51.522	25.761	5.0614	0.062856
Residuals	5	25.448	5.090		
Lack of fit	3	23.922	7.974	10.4461	0.088623
Pure error	2	1.527	0.763		

Stationary point of response surface:

X1	X2
9.57779	240.24419

Eigenanalysis:

eigen() decomposition

$values

[1] 0.02454460 -0.01337914

$vectors

	[,1]	[,2]
X1	-0.8969851	-0.4420607
X2	0.4420607	-0.8969851

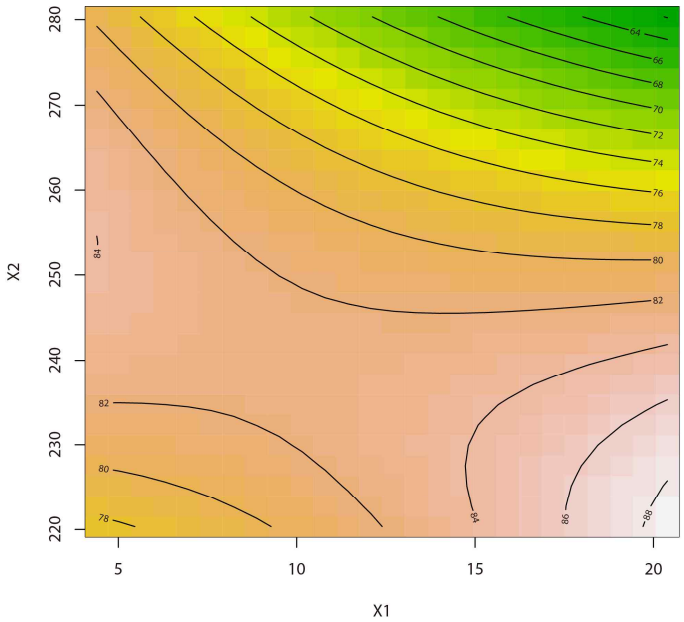
적합 된 모형은

$\hat{Y} = -329.56 + 6.8972X_1 + 3.1557X_2 + 0.017134X_1^2 - 0.0059682X_2^2 - 0.030075X_1X_2$이다.

이 모형은 이 모형은 Y의 반응 값에서 lack of fit이 0.088623으로 유의수준 0.05보다 높으므로 모형이 적합하다고 볼 수 있다. F-statistic의 p-value 값이 0.01128으로 유의수준 0.05내에서 의미가 있다고 볼 수 있다. Adjusted R-squared 값은 0.8243으로 82.43%의 설명력을 가진다.

Stationary point of response surface:에서 값을 확인했을 때
정상 점은 $(x_{1,}x_2)$=(9.57779, 240.24419)에서 가진다.

> contour(b,~X1+X2, image = TRUE)



Eigenanalysis에서의 values의 값이 하나는 양수, 다른 하나는 음수로 주어지므로 정상 점은 안부점이 된다.

그래프에서 보았을 때 X1(공정시간)이 증가하면 Y(효율)는 증가하고 X2(공정온도)를 증가시키면 Y(효율)는 감소한다.

2. Make your own dataset based on data in Example 6.6 (p. 254). Let $X_1 <- X_1 + \epsilon$, $Y <- Y + \epsilon$, where $\epsilon \sim N(0, 0.1^2)$. Compute the WLSE using the same method as the one in text.

먼저 Example 6.6 (p. 254)의 자료를 입력해준다.

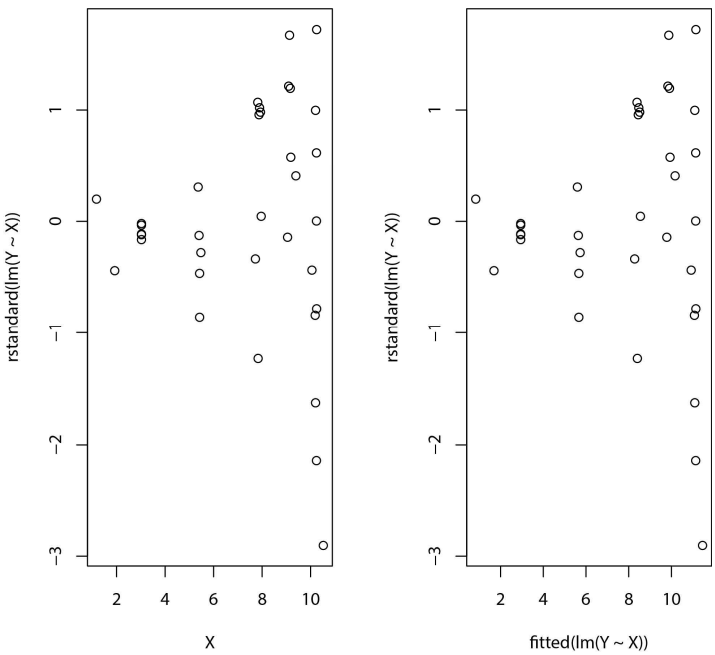
```
> x<-c(1.15,1.90,3,3,3,3,3,5.34,5.38,5.4,5.4,5.45,7.7,7.8,7.81,7.85,7.87,7.91,7.94,9.03,
+ 9.07,9.11,9.14,9.16,9.37,10.17,10.18,10.22,10.22,10.22,10.18,10.50,10.23,10.03,10.23)
> length(x)
[1] 35
> y<-c(0.99,0.98,2.6,2.67,2.66,2.78,2.8,5.92,5.35,4.33,4.89,5.21,
+ 7.68,9.81,6.52,9.71,9.82,9.81,8.5,9.47,11.45,12.14,11.5,10.65,
+ 10.64,9.78,12.39,11.03,8,11.9,8.68,7.25,13.46,10.19,9.93)
> length(y)
[1] 35
```

자료를 $X_1 <- X_1 + \epsilon$, $Y <- Y + \epsilon$, where $\epsilon \sim N(0, 0.1^2)$의 형태로 바꿔준다.

```
> X<-x+rnorm(n=1,mean=0,sd=0.1)
> Y<-y+rnorm(n=1,mean=0,sd=0.1)
```

잔차 대 예측치와 설명변수와의 산점도를 그려본다.

```
> par(mfrow=c(1,2))
> plot(x=X,y=rstandard(lm(Y~X)))
> plot(x=fitted(lm(Y~X)),y=rstandard(lm(Y~X)))
```



위의 산점도를 보았을 때 예측치와 X의 값이 커질수록 잔차의 분산이 커진다. 그러므로 등분산 가정에 위배됨을 알 수 있다. 따라서 우리는 WLSE를 적용을 시킨다.

가중치는 관측치를 대략 5개의 그룹으로 묶고 관측 치에 대한 표본평균(설명변수)과 표본분산(반응변수) 사이의 회귀모형을 찾는다.

```
> x1<-X[3:7]
> x2<-X[8:12]
> x3<-X[13:19]
> x4<-X[20:25]
> x5<-X[26:35]
> y1<-Y[3:7]
> y2<-Y[8:12]
> y3<-Y[13:19]
> y4<-Y[20:25]
> y5<-Y[26:35]
> xbar<-c(mean(x1),mean(x2),mean(x3),mean(x4),mean(x5))
> s<-c(var(y1),var(y2),var(y3),var(y4),var(y5))
> xsq<-xbar^2
> summary(lm(s~xbar+xsq))

Call:
lm(formula = s ~ xbar + xsq)

Residuals:
      1       2       3       4       5
-0.1206  0.2014  0.5487 -1.3122  0.6827

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.53948    3.85759   0.399    0.728
xbar        -0.73102    1.30484  -0.560    0.632
xsq          0.08734    0.09766   0.894    0.466

Residual standard error: 1.128 on 2 degrees of freedom
Multiple R-squared:  0.7372,    Adjusted R-squared:  0.4744
F-statistic: 2.805 on 2 and 2 DF,  p-value: 0.2628
```

다음과 같은 이차모형을 얻었다. $\widehat{s^2} = 1.53948 - 0.73102\overline{x} + 0.08734\overline{x^2}$ 이 회귀식에서 $\overline{x}$대신 x값을 대입하여 나온 표본분산의 역수를 가중치로 만든다.

```
> w<-1/(lm(s~xbar+xsq)$coefficients%*%t(cbind(1,x,x^2)))
```

만든 가중치와 데이터들을 데이터프레임으로 정렬시켰다.

```
> data2<-data.frame(X,Y,t(w))
> data2
```

	X	Y	t.w.
1	1.17297	1.102078	1.2280157
2	1.92297	1.092078	2.1465604
3	3.02297	2.712078	7.5452624
4	3.02297	2.782078	7.5452624
5	3.02297	2.772078	7.5452624
6	3.02297	2.892078	7.5452624
7	3.02297	2.912078	7.5452624
8	5.36297	6.032078	7.9022868
9	5.40297	5.462078	7.4206806
10	5.42297	4.442078	7.1958088
11	5.42297	5.002078	7.1958088
12	5.47297	5.322078	6.6754057
13	7.72297	7.792078	0.9179898
14	7.82297	9.922078	0.8683421
15	7.83297	6.632078	0.8635995
16	7.87297	9.822078	0.8450111
17	7.89297	9.932078	0.8359405
18	7.93297	9.922078	0.8182309
19	7.96297	8.612078	0.8053145
20	9.05297	9.582078	0.4852980
21	9.09297	11.562078	0.4774212
22	9.13297	12.252078	0.4697343
23	9.16297	11.612078	0.4640901
24	9.18297	10.762078	0.4603835
25	9.39297	10.752078	0.4240030
26	10.19297	9.892078	0.3185690
27	10.20297	12.502078	0.3175105
28	10.24297	11.142078	0.3133289
29	10.24297	8.112078	0.3133289
30	10.24297	12.012078	0.3133289
31	10.20297	8.792078	0.3175105
32	10.52297	7.362078	0.2862384
33	10.25297	13.572078	0.3122963
34	10.05297	10.302078	0.3339603
35	10.25297	10.042078	0.3122963

```
> WLSE<-lm(Y~X, data=data2, weights = t(w))
> summary(WLSE)

Call:
lm(formula = Y ~ X, data = data2, weights = t(w))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.8592 -0.5526  0.1686  0.9953  1.6700

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.81266    0.30922  -2.628   0.0129 *
X            1.16552    0.06027  19.337   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.156 on 33 degrees of freedom
Multiple R-squared:  0.9189,    Adjusted R-squared:  0.9164
F-statistic: 373.9 on 1 and 33 DF,  p-value: < 2.2e-16
```

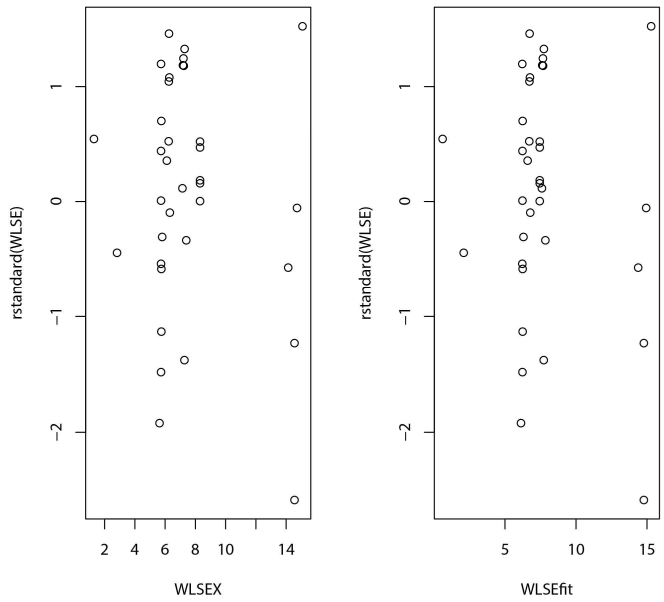
가중치를 적용 시킨 결과 모형은 $\hat{y}=-0.81266-1.16552x$이 되고 p-value는 거의 0에 가깝다. 따라서 모형은 적절하다고 볼 수 있다.

이에 따른 ANOVA TABLE은 다음과 같다.

```
> anova(WLSE)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 499.63  499.63  373.92 < 2.2e-16 ***
Residuals 33  44.09    1.34
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

등분산성을 만족하는지 확인하기 위하여 새로운 잔차 $\hat{z_i}$ 대 예측치 $\widehat{w^{1/2}}\hat{y_i}$ 와 설명변수 $\widehat{w^{1/2}}x_i$ 와의 산점도를 그려본다.

```
> WLSEfit<-diag(t(sqrt(w))%*%fitted(WLSE))
> WLSEX<-diag(t(sqrt(w))%*%X)
> par(mfrow=c(1,2))
> plot(x=WLSEX,y=rstandard(WLSE))
> plot(x=WLSEfit,y=rstandard(WLSE))
```



위의 산점도를 봤을 때 등분산성을 만족한다.

3. Make your own dataset based on data in Example 6.7 (p. 259). Let $Y <- Y + \epsilon$, where $\epsilon \sim N(0,1)$.

(1) Fit the data to the multiple linear regression model.

먼저 Example 6.7 (p. 259)의 자료를 입력해준다.

```
> y<-c(26,38,50,76,108,157,
+        17,26,37,53,83,124,
+        13,20,27,37,57,87,
+        NA,15,22,27,41,63)
> x1<-c(rep(0,6),rep(10,6),rep(20,6),rep(30,6))
> x2<-c(rep(seq(0,60,12),4))
```

자료를 $Y <- Y + \epsilon$, where $\epsilon \sim N(0,1)$의 형태로 바꿔준다.

```
> Y<-y+rnorm(n=1,mean=0,sd=1)
> data3<-data.frame(x1,x2,Y)
> data3
   x1 x2         Y
1   0  0  25.30623
2   0 12  37.30623
3   0 24  49.30623
4   0 36  75.30623
5   0 48 107.30623
6   0 60 156.30623
7  10  0  16.30623
8  10 12  25.30623
9  10 24  36.30623
10 10 36  52.30623
11 10 48  82.30623
12 10 60 123.30623
13 20  0  12.30623
14 20 12  19.30623
15 20 24  26.30623
16 20 36  36.30623
17 20 48  56.30623
18 20 60  86.30623
19 30  0        NA
20 30 12  14.30623
21 30 24  21.30623
22 30 36  26.30623
23 30 48  40.30623
24 30 60  62.30623
```

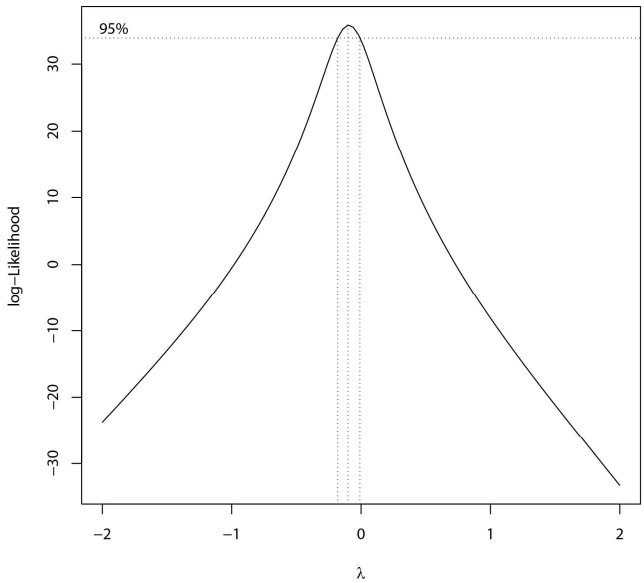
```
a<-lm(Y~x1+x2,data=data3)
> summary(a)

Call:
lm(formula = Y ~ x1 + x2, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-15.592  -9.695  -3.722   6.713  35.296

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.4899     6.3322   4.341 0.000317 ***
x1           -1.7166     0.2640  -6.502 2.44e-06 ***
x2            1.5587     0.1452  10.735 9.48e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.82 on 20 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8793,    Adjusted R-squared:  0.8673
F-statistic: 72.87 on 2 and 20 DF,  p-value: 6.543e-10
```

모형의 적합치는 $\hat{Y}=27.4899-1.7166X_1+1.5587X_2$이다. 그리고 이 모형은 p-value가 6.543e-10으로 0.05의 유의수준에서 유의하다고 할 수 있다. 그리고 Adjusted R-squared은 0.8673으로 86.73%의 설명력을 가진다.

(2) Fit the data to the Box-Cox transformation model.

λ를 구하는 boxcox함수를 사용하기 위해 MASS 라이브러리를 다운받는다.

```
> library(MASS)#boxcox
> par(mfrow=c(1,1))
> box_cox<-boxcox(a)
```




```
> lambda<-box_cox$x
> likeli_value<-box_cox$y
> order_table<-cbind(lambda,likeli_value)
> sorted<-order_table[order(-likeli_value),]
> sorted[1,]
      lambda  likeli_value
-0.02020202   34.33260366
```

람다가 최댓값을 가지는 점은 -0.02020202이다.

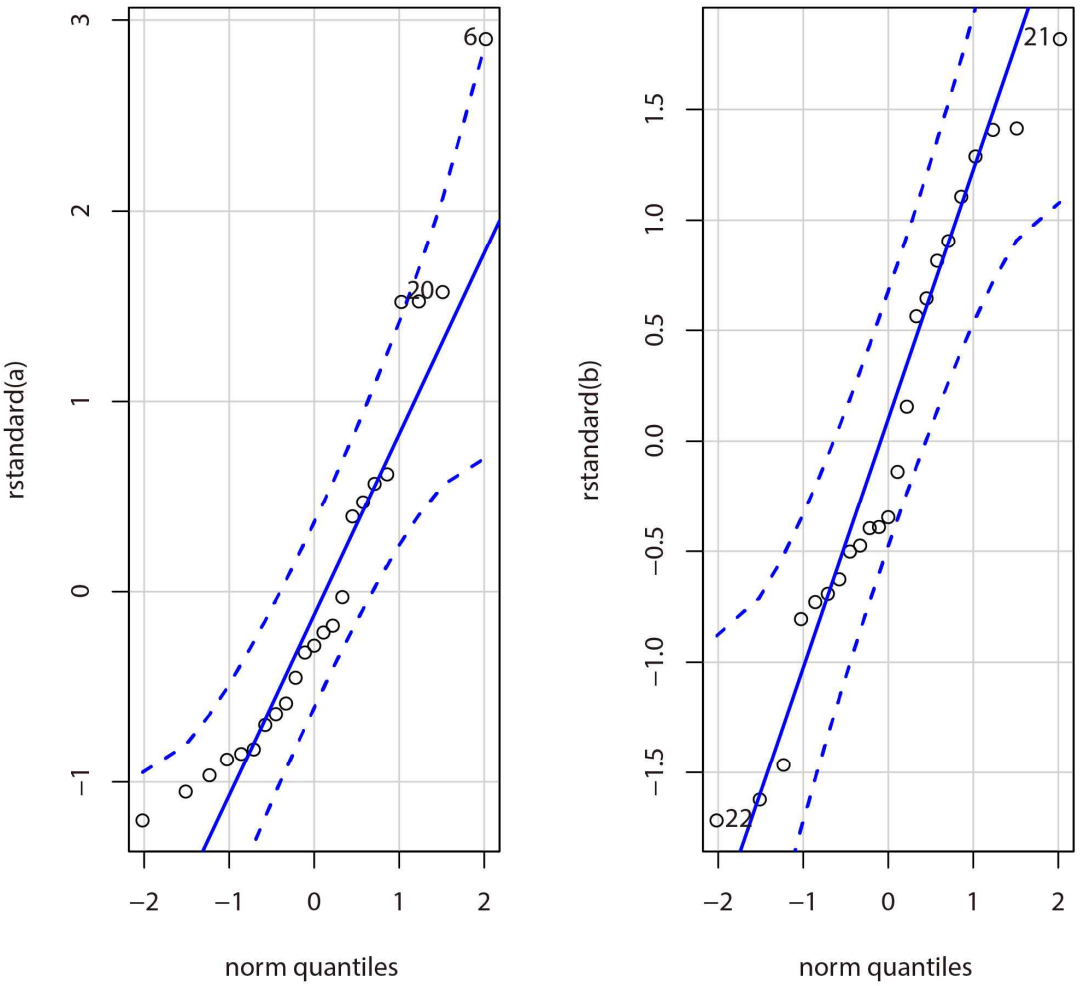
데이터를 log우도함수에 적용을 하면

```
> b<-lm(log(Y)~x1+x2,data=data3)
> summary(b)

Call:
lm(formula = log(Y) ~ x1 + x2, data = data3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.08136 -0.03067 -0.01613  0.04163  0.08552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.1810706  0.0232919  136.57   <2e-16 ***
x1          -0.0320987  0.0009711  -33.06   <2e-16 ***
x2           0.0314739  0.0005341   58.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05084 on 20 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9953,    Adjusted R-squared:  0.9948
F-statistic:  2119 on 2 and 20 DF,  p-value: < 2.2e-16
```

모형의 적합치는 $\hat{Y}= 3.1810706 - 0.0320987X_1 + 0.0314739X_2$이다. 그리고 이 모형은 p-value가 거의 0에 가까워 0.05의 유의수준에서 유의하다고 할 수 있다. 그리고 Adjusted R-squared은 0.8673으로 99.48%의 설명력을 가지므로 (1)의 모형보다 더 적합하다.

(3) Compare two models in (1) and (2) by using the $Q-Q$ plot of residuals in each model.

```
> par(mfrow=c(1,2))
> qqPlot(rstandard(a))
 6 20
 6 19
> qqPlot(rstandard(b))
21 22
20 21
```



(1)의 Q-Q plot에서는 신뢰구간 밖으로 나가는 점들이 있고, 직선위에 잘 모여 있지 않다. 하지만 (2)의 Q-Q plot은 점들이 직선위에 잘 모여 있다. 그러므로 Box-Cox변환모형을 로그변환 시켰을 때의 잔차의 정규 확률도가 더 정규분포의 형태에 근사한다.