회귀분석론 HW1

212STG18 예지혜

2021년 3월 1일

1.

```
crime <- read.table("Crime rate.txt")
names(crime) <- c("Y","X")
head(crime)
```

	Y <int>	X <int>
1	8487	74
2	8179	82
3	8362	81
4	8220	81
5	6246	87
6	9100	66

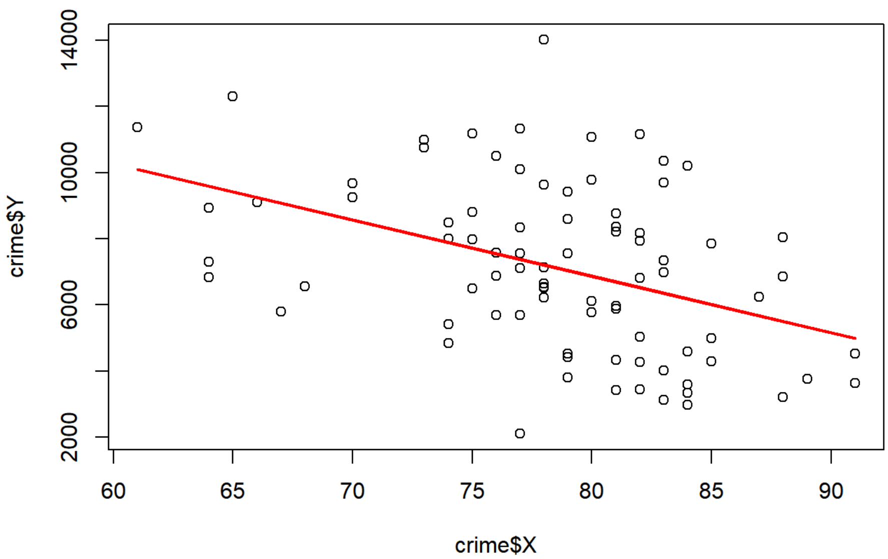
6 rows

(a)

```
lm.1 <- lm(Y~X, data = crime)
summary(lm.1)
```

```
##
## Call:
## lm(formula = Y ~ X, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## X            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```
plot(crime$X, crime$Y)
lines(crime$X, lm.1$fitted, col=2, lwd=2)
```



적합이 별로 좋지 않아 보인다.

(b)

(1) 진학률 1% 당 범죄 사건 발생이 평균적으로 10,000건 당 170.58건 감소한다.

(2)

```
lm.1$coefficients[1] + lm.1$coefficients[2]*80
```

```
## (Intercept)
##    6871.585
```

(3)

```
lm.1$residuals[10]
```

```
##       10
## 1401.566
```

(4)

```
anova(lm.1)[3]
```

	Mean Sq <dbl>
X	93462942
Residuals	5552112

2 rows

σ^2의 추정치는 5552112이다.

2.

(a)

```
summary(lm.1)
```

```
##
## Call:
## lm(formula = Y ~ X, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## X            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```
qt(0.995,df=length(crime$X)-2)
```

```
## [1] 2.637123
```

$H0 : \beta 1 = 0, H1 : \beta 1 \neq 0$

t값의 절대값이 4.103으로 기각역인 2.64보다 크므로 귀무가설을 기각한다. 따라서 선형관계가 존재한다고 할 수 있다. p-value는 9.57e-05이다.

(b)

```
c(-170.58-qt(0.995,df=length(crime$X)-2)*41.57, -170.58+qt(0.995,df=length(crime$X)-2)*41.57)
```

```
## [1] -280.20522  -60.95478
```

$\beta 1$은 99%의 경우 이 신뢰구간에 포함된다.

(c)

```
anova(lm.1)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
X	1	93462942	93462942	16.83376	9.571396e-05
Residuals	82	455273165	5552112	NA	NA

2 rows

(d)

```
qf(0.99,1,length(crime$X)-2)
```

```
## [1] 6.95442
```

```
(-4.103)^2
```

```
## [1] 16.83461
```

$H0 : \beta 1 = 0, H1 : \beta 1 \neq 0$

F값은 16.83으로 6.95보다 크므로 귀무가설을 기각한다. 따라서 선형관계는 존재한다.

p-value는 9.571e-05로 (a) t검정과 같은 값을 가지며, t값을 제곱했을 때 F값과 동일한 값이 나오므로 같은 검정임을 알 수 있다.

(e)

```
93462942 + 455273165
```

```
## [1] 548736107
```

```
93462942 / (93462942 + 455273165)
```

```
## [1] 0.170324
```

X 변수를 통해 총변동에서 SSR (=93462942) 만큼 감소하였다. 총변동에서 X(진학률)이 차지하는 비율은 17.03% 이다. 이 회귀식이 전체의 17% 정도를 설명하므로 큰 비율은 아니라고 해석된다.

(f)

```
sqrt(summary(lm.1)$r.squared)
```

```
## [1] 0.4127033
```

3.

(a)

full model : $Yi = \beta 0 + \beta 1Xi + \varepsilon i, \varepsilon i$ are independent $N(0, \sigma 2)$

reduced model : $Yi = \beta 0 + \varepsilon i$

(b)

```
anova(lm.1)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
X	1	93462942	93462942	16.83376	9.571396e-05
Residuals	82	455273165	5552112	NA	NA

2 rows

(1) SSE(F) = SSE = 455273165

(2) SSE(R) = SST = 93462942 + 455273165 = 548736107

(3) df(F) = n-2 = 82

(4) df(R) = n-1 = 83

(5) F* = 16.834

(6)

```
qf(0.99,1,length(crime$X)-2)
```

```
## [1] 6.95442
```

(c) 기각역이 동일하며, F통계량을 제곱하면 t통계량과 같다.

4.

```
sol <- read.table("Solution concentration.txt")
names(sol) <- c("Y","X")
head(sol)
```

	Y <dbl>	X <dbl>
1	0.07	9
2	0.09	9
3	0.08	9
4	0.16	7
5	0.17	7
6	0.21	7

6 rows

(a)

```
lm.2 <- lm(Y~X, data=sol)
summary(lm.2)
```

```
##
## Call:
## lm(formula = Y ~ X, data = sol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5333 -0.4043 -0.1373  0.4157  0.8487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
## X            -0.3240     0.0433  -7.483 4.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4743 on 13 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.7971
## F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

(b)

```
new <- data.frame(mean=with(tapply(Y, factor(X),mean), data=sol))
new <- data.frame(X= as.numeric(row.names(new)), new)
full <- lm(Y~factor(X), data=sol)
smaller <- lm(Y~X, data=sol)
anova(smaller, full)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	13	2.924653	NA	NA	NA	NA
2	10	0.157400	3	2.767253	58.60342	1.194477e-06

2 rows

```
qf(0.975,3,10)
```

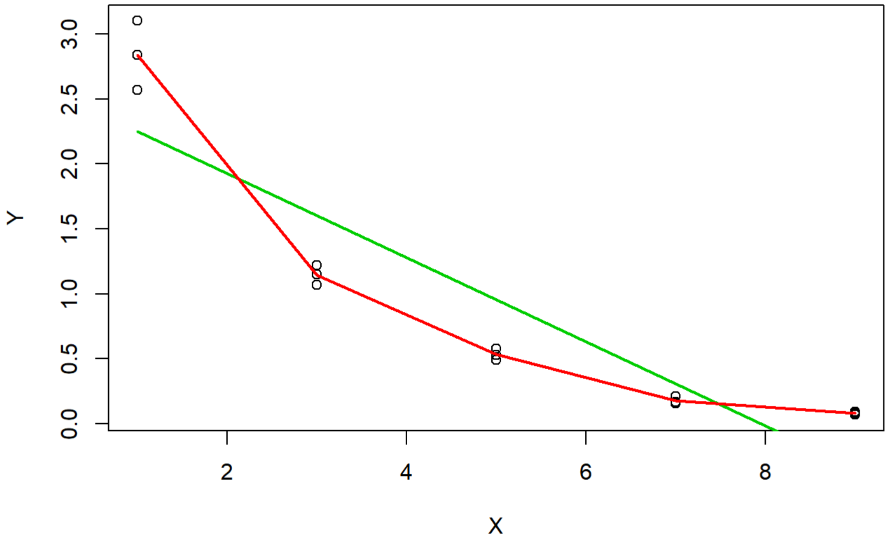
```
## [1] 4.825621
```

$H0: E(Y) = \beta 0 + \beta 1 X vs. H1: not H0$

F통계량이 58.603으로 기각역인 4.83보다 크므로 귀무가설을 기각한다. 따라서 1차선형회귀모형은 적합하지 않다.

(c)

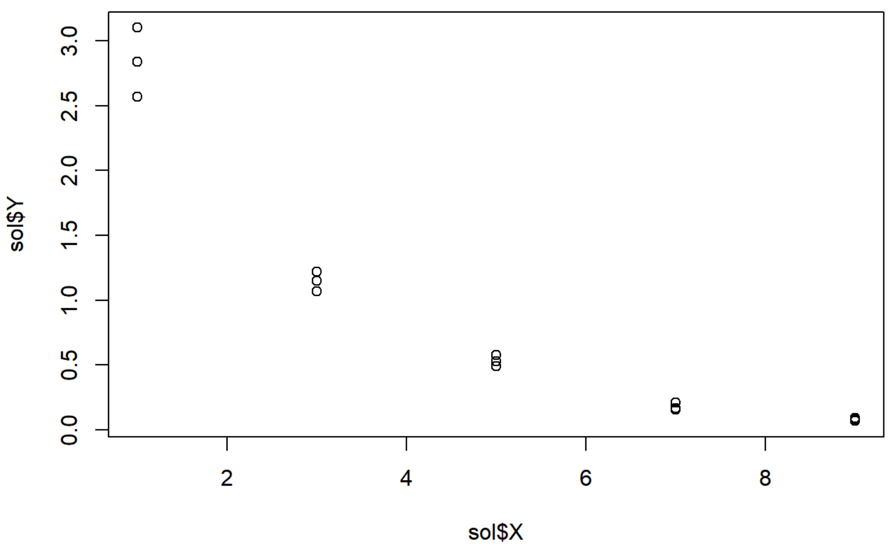
```
with(plot(X,Y), data=sol)
lines(sol$X, smaller$fitted, col=3, lwd=2)
with(lines(X,mean, col=2, lwd=2), data=new)
```



문제 (b)에서 lack of fit이 존재한다는 결과가 나왔는데, anova table을 보면, X가 무의미하다고 결론을 내리기에는 무리가 있다. 또한, 그래프를 보면 로그변환이 더 필요해보인다.

(d)

```
plot(sol$X, sol$Y)
```



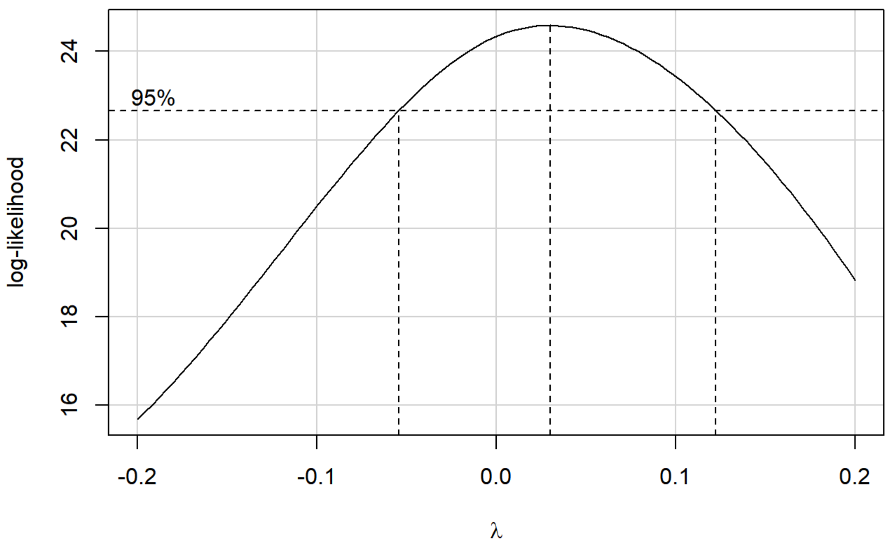
(b), 로그변환이 가장 적절해보인다.

(e)

```
library(car)
```

```
## Loading required package: carData
```

```
box_sol<-boxCox(lm.2,lambda = c(-0.2,-0.1,0,0.1,0.2))
```




```
library(ALSM)
```

```
## Warning: package 'ALSM' was built under R version 3.6.3
```

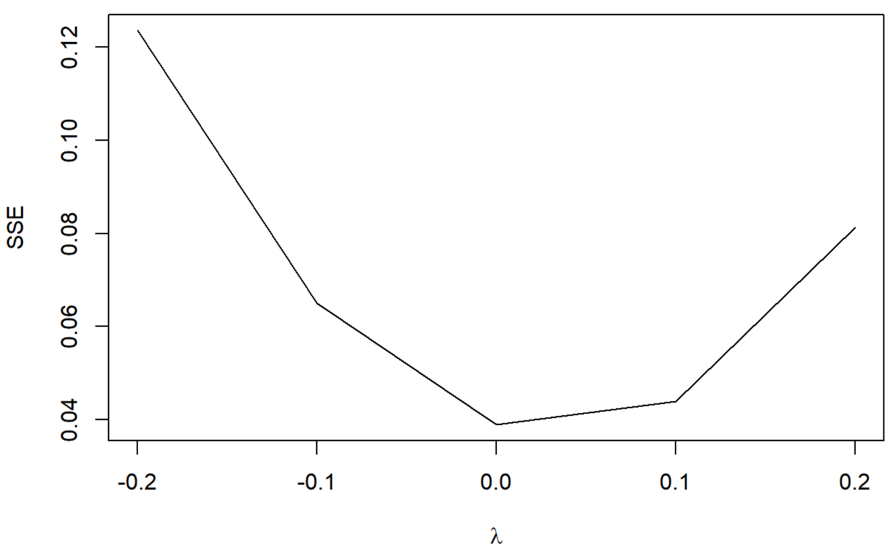
```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

```
## Loading required package: SuppDists
```

```
## Warning: package 'SuppDists' was built under R version 3.6.3
```

```
boxcox.sse(sol$X, sol$Y,l = c(-0.2,-0.1,0,0.1,0.2))
```



	lambda <dbl>	SSE <dbl>
1	-0.2	0.12353047

	lambda <dbl>	SSE <dbl>
2	-0.1	0.06505067
5	0.0	0.03897303
3	0.1	0.04396062
4	0.2	0.08131793

5 rows

람다가 0일때, 즉 로그 변환이 가장 적합해보인다.

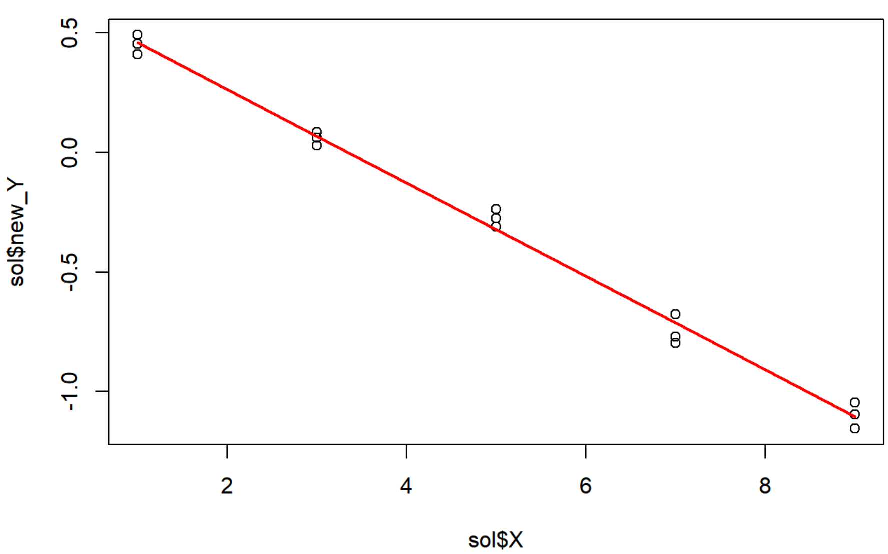
(f)

```
sol$new_Y <- log(sol$Y, base = 10)
new.lm.2 <- lm(new_Y~X, data = sol)
summary(new.lm.2)
```

```
##
## Call:
## lm(formula = new_Y ~ X, data = sol)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082958 -0.044421  0.006813  0.033512  0.085550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.654880   0.026181   25.01 2.22e-12 ***
## X           -0.195400   0.004557  -42.88 2.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04992 on 13 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9924
## F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15
```

(g)

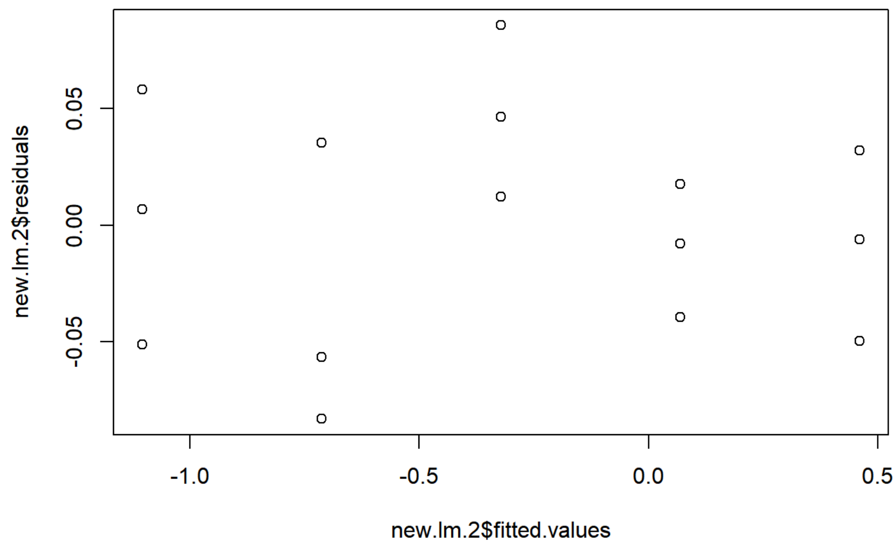
```
plot(sol$X, sol$new_Y)
lines(sol$X, new.lm.2$fitted, col=2, lwd=2)
```



적합이 아주 잘 되어 보인다.

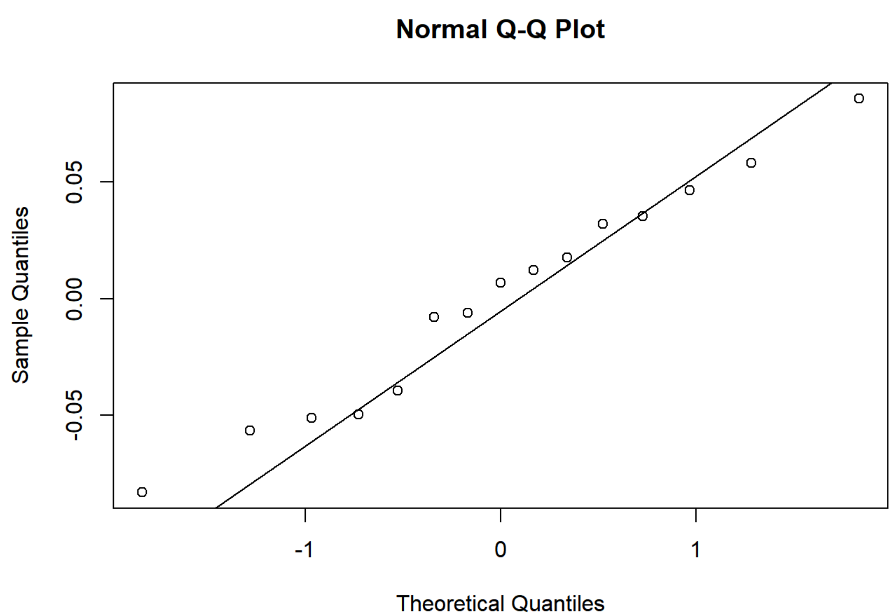
(h)

```
plot(new.lm.2$fitted.values, new.lm.2$residuals)
```



패턴 없이 고르게 잘 분포되어 있으므로 잘 적합되었다고 볼 수 있다.

```
qqnorm(new.lm.2$residuals)
qqline(new.lm.2$residuals)
```



qqplot또한 직선의 형태에 가까우므로 정규성 가정을 해치지 않는다고 볼 수 있다.

(i) $Y = 10^{(} - 0.1954X + 0.65488)$

5.

```
prod = read.table("Production time.txt")
names(prod) <- c("Y","X")
head(prod)
```

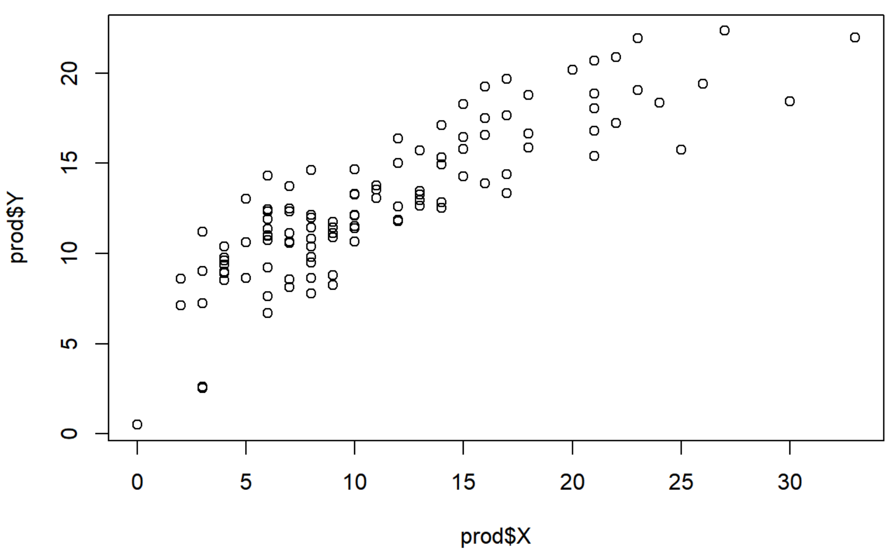
	Y <dbl>	X <int>
1	14.28	15
2	8.80	9

	Y <dbl>	X <int>
3	12.49	7
4	9.38	4
5	10.89	9
6	15.39	21

6 rows

(a)

```
plot(prod$X, prod$Y)
```



살짝 위로 볼록한 형태를 띠므로 X에 변환이 필요해보인다.

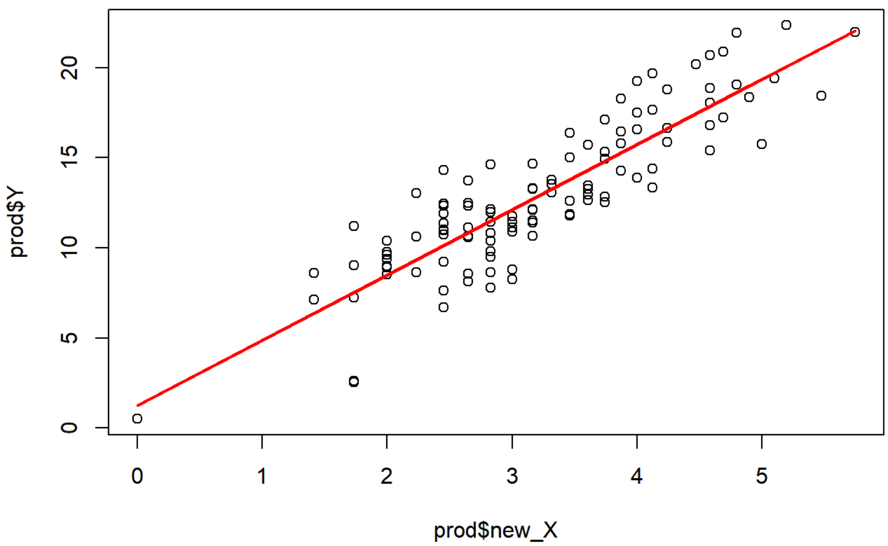
(b)

```
prod$new_X <- sqrt(prod$X)
lm.3 <- lm(Y~new_X, data = prod)
summary(lm.3)
```

```
##
## Call:
## lm(formula = Y ~ new_X, data = prod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0008 -1.2161  0.0383  1.3367  4.1795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2547     0.6389   1.964   0.0521 .
## new_X         3.6235     0.1895  19.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.99 on 109 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7683
## F-statistic: 365.7 on 1 and 109 DF,  p-value: < 2.2e-16
```

(c)

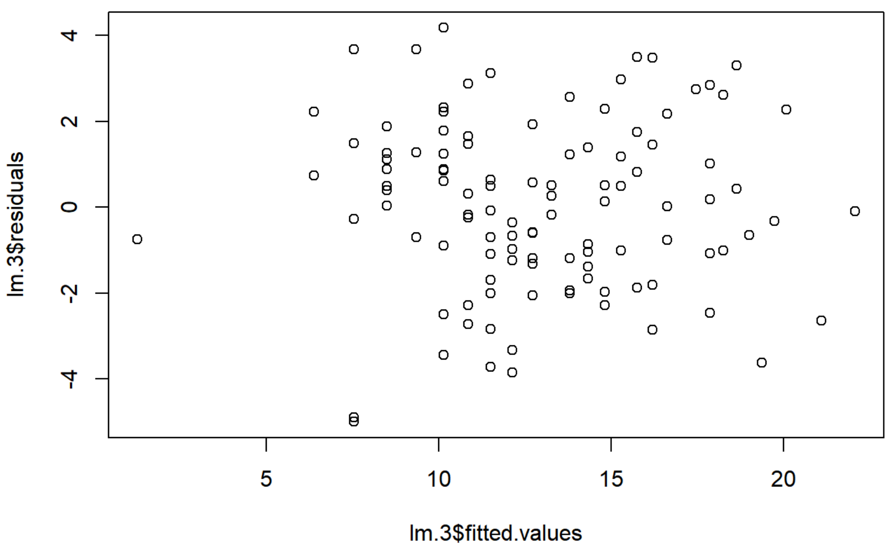
```
plot(prod$new_X, prod$Y)
lines(prod$new_X, lm.3$fitted, col=2, lwd=2)
```



적합이 잘 되어 보인다.

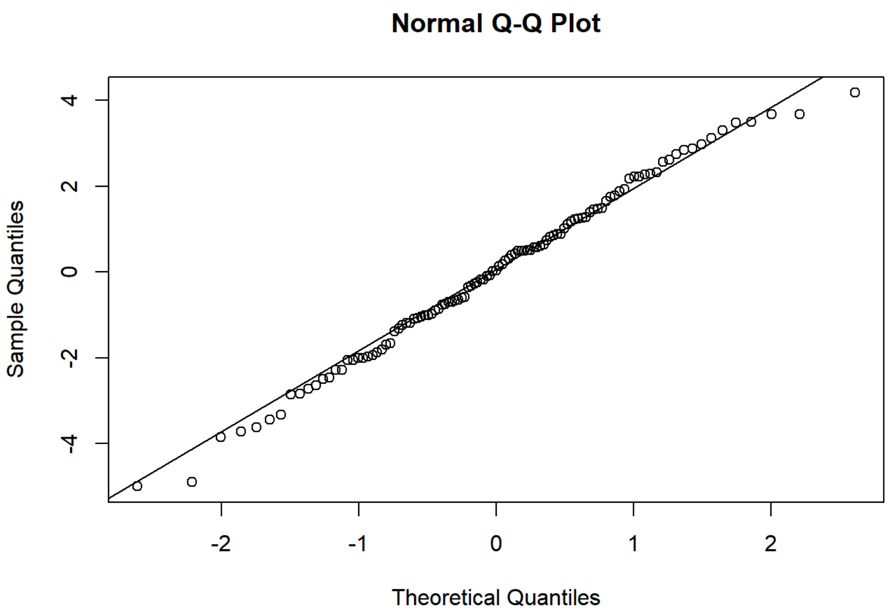
(d)

```
plot(lm.3$fitted.values, lm.3$residuals)
```



뚜렷한 패턴은 보이지 않는다.

```
qqnorm(lm.3$residuals)
qqline(lm.3$residuals)
```



qqplot은 직선의 형태에 가까우므로 정규성 가정을 해치지 않는다고 볼 수 있다.

(e) $Y = 3.6235\sqrt{X} + 1.2547$