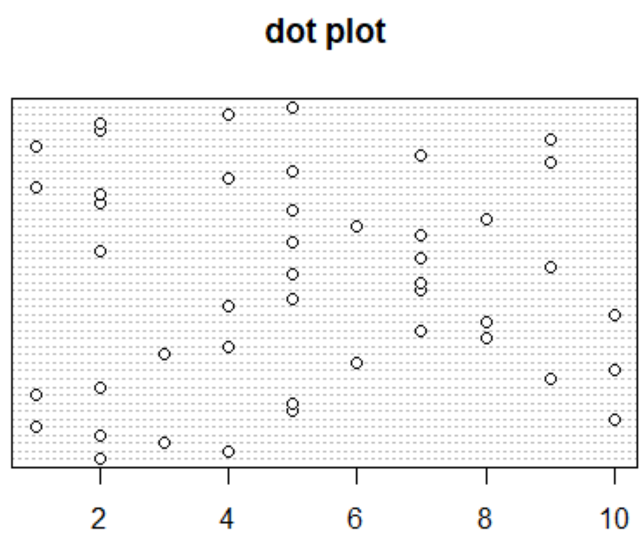
Chapter3

3.4 번

a) data 분포 파악

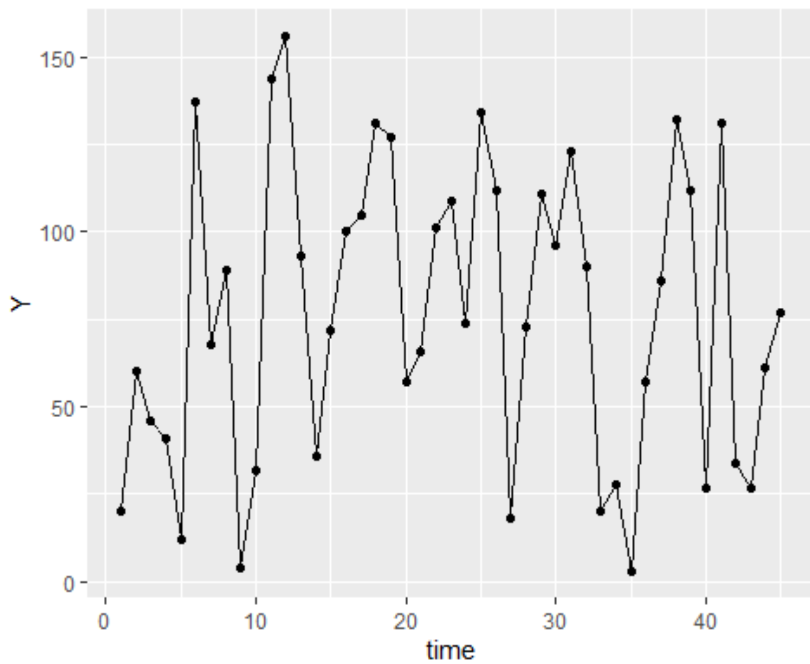
```
ta304 <- read.table("c:/Users/KimMinyoung/Documents/CH03PR04.txt")
dotchart(ta304[,2], main = "dot plot")
```



Data 가 1~10 사이에 분포해 있다. outlier 는 없어보인다.

b) time plot

```
time <- c(1:45)
library(ggplot2)
ggplot(ta304, aes(x=time, y=Y))+geom_point()+geom_line()
```



특별한 패턴이 보이지 않는다. 시간에 따른 상관관계가 없어보인다.

c) stem and leaf plot

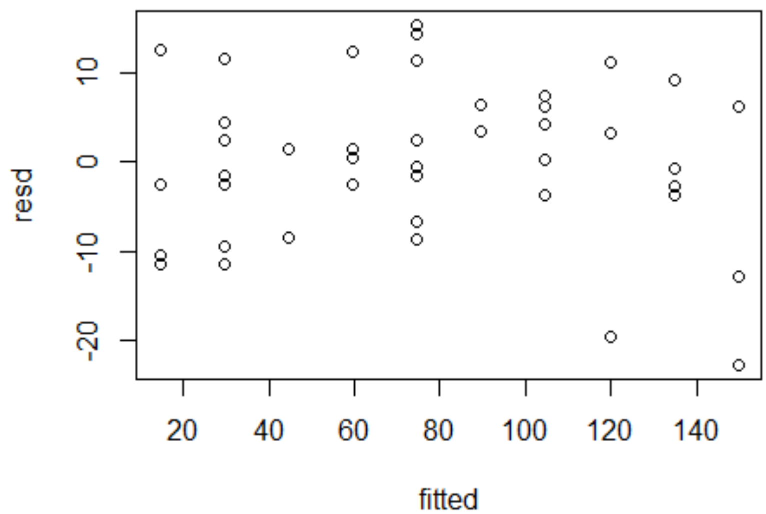
```
lm.ta304 <- lm(Y~X, data=ta304)
resd <- lm.ta304$resid
stem(resd)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -2 | 30
##   -1 |
##   -1 | 3110
##   -0 | 99997
##   -0 | 44333222111
##    0 | 001123334
##    0 | 5666779
##    1 | 112234
##    1 | 5
```

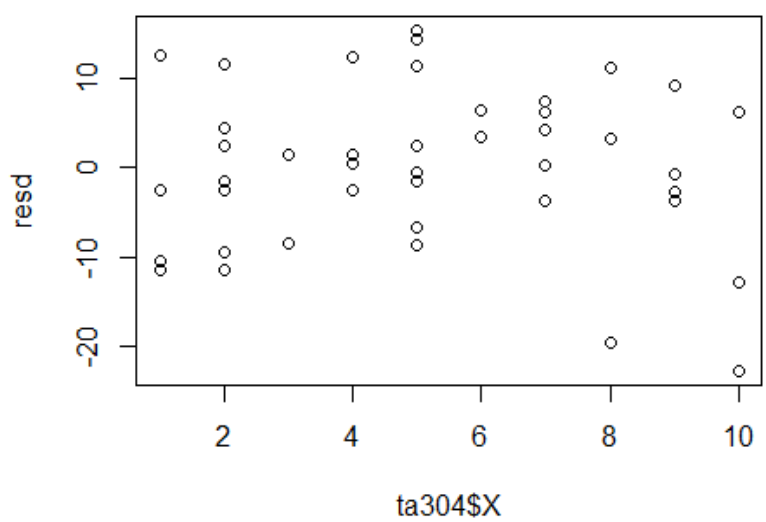
residual 의 분포가 종모양을 따른다.

d)

```
resd <- lm.ta304$resid
fitted <- lm.ta304$fitted
plot(fitted, resd)
```



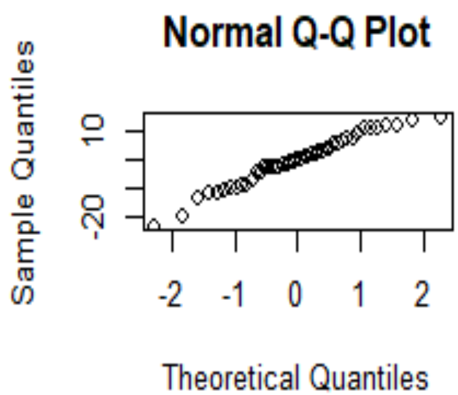

```
plot(ta304$X, resd)
```



두 plots 을 비교했을 때 X_i 나 (Y_i_hat)에 대한 residuals 의 분포가 동일함을 확인할 수 있다. 이는 regression model 2.1 의 가정인 등분산성을 만족시키고 이상치 또한 존재하지 않는다. 또한 두 plots 은 같은 정보를 제공한다.

e) Normal Probability Plot

```r
par(mfrow=c(2,2))
qqnorm(lm.ta304$resid)
```



H_0: normal

H_1: normal x

$r = \sqrt{R^2} = \sqrt{0.9575} = 0.9891$

d=0.01 일때 0.9785

r 이 0.9785 보다 크므로 귀무가설을 채택한다.

즉, Normal 가정이 합리적이라고 보인다.

<expected value>

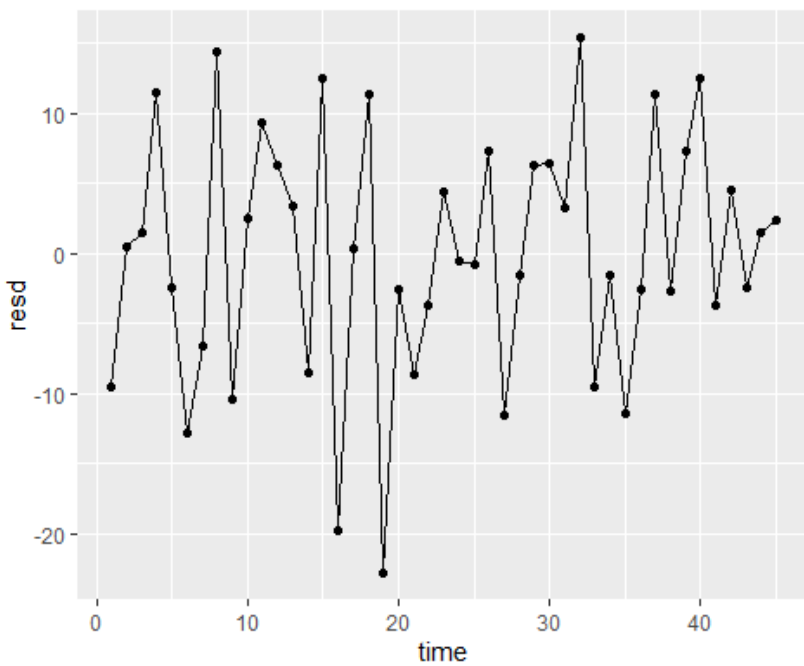
$\sqrt{MSE}\left(z\left(\frac{k-0.375}{n+0.25}\right)\right)$

```r
summary(lm.ta304)

##
## Call:
## lm(formula = Y ~ X, data = ta304)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## X            15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

f) Time Plot Of Residuals

```r
time <- c(1:45)
resd <- lm.ta304$resid
ggplot(data=ta304,aes(time,resd))+geom_point()+geom_line()
```



시간에 따른 특별한 패턴이 없어보인다.

g) $\log_e \sigma_i^2 = r_0 + r_1 X_i$

$* SSR^* = 15155$

$SSE = 3416.38$

$n = 45$

$H_0: r_1 = 0$

$H_1: r_1 \neq 0$

$$X^2_{BP} = \frac{SSR^*}{2} \Big/ \left(\frac{SSE}{n}\right)^2$$

$$= \frac{15155}{2} \Big/ \left(\frac{3416.38}{45}\right)^2 = 1.31469$$

$\chi^2(0.95, 1) = 3.84$

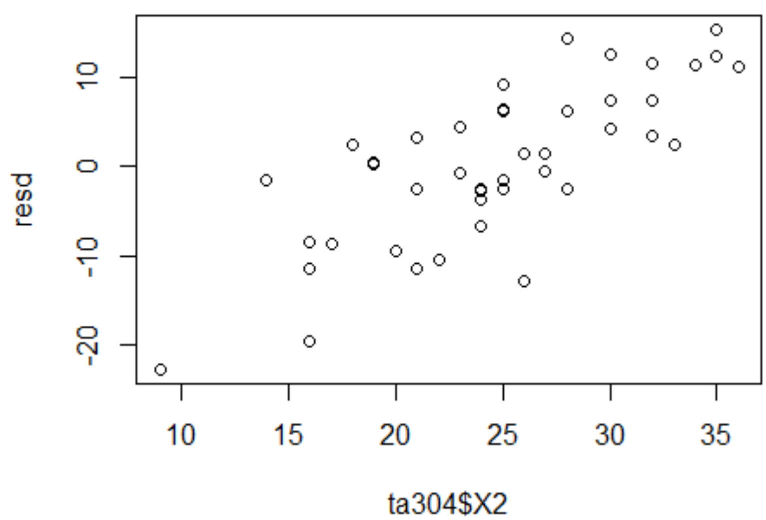
$X^2_{BP} \le 3.84$: H_0

⇒ error variance가 constant하다

```
library(car)
ncvTest(lm.ta304)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.31468, Df = 1, p = 0.25155
```

h) X2-잔차 Graph

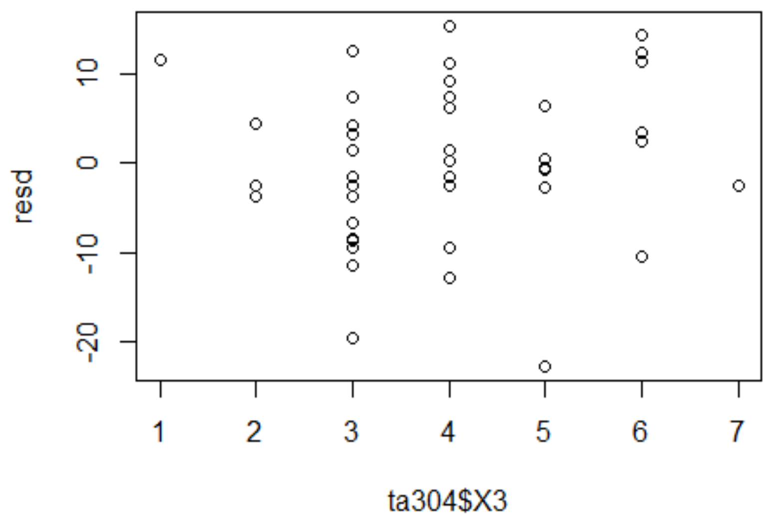
```
resd <- lm.ta304$resid
plot(ta304$X2, resd)
```



양의 상관관계 : model 이 X2 를 포함함으로써 improved

X3-잔차 Graph

```
plot(ta304$X3, resd)
```

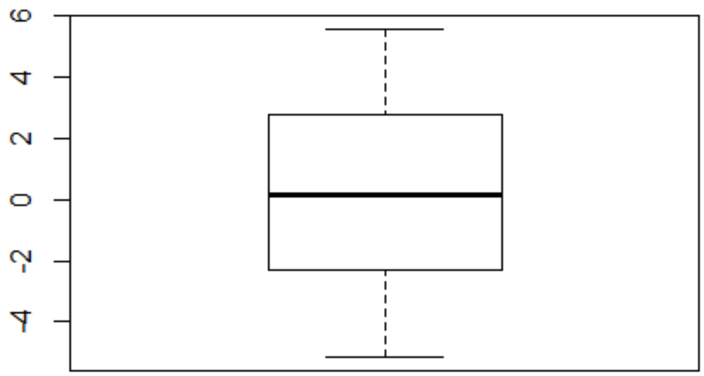


특별한 패턴이 없어보인다. : not bring any imporvedment

3.6 번

a) residual box plot

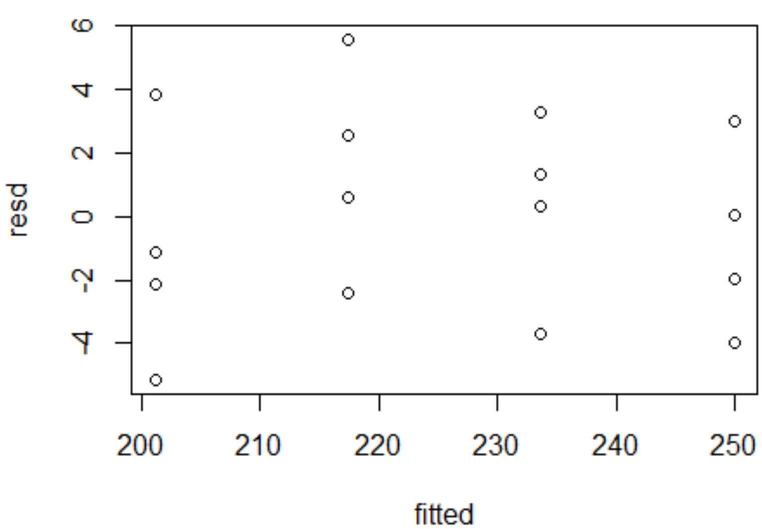
```
ta306 <- read.table("c:/Users/KimMinyoung/Documents/CH01PR22.txt")
names(ta306) <- c("Y", "X")
lm.ta306 <- lm(Y~X, data=ta306)
resd <- lm.ta306$resid
resd

##      1      2      3      4      5      6      7      8      9     10
## -2.150  3.850 -5.150 -1.150  0.575  2.575 -2.425  5.575  3.300  0.300
##     11     12     13     14     15     16
##  1.300 -3.700  0.025 -1.975  3.025 -3.975

boxplot(resd)
```



boxplot 을 통해 어디로 치우치지 않음을 확인할 수 있다. 어느정도 normality 하다고 볼 수 있다.

b) residual

```
fitted <- lm.ta306$fitted
plot(fitted, resd)
```

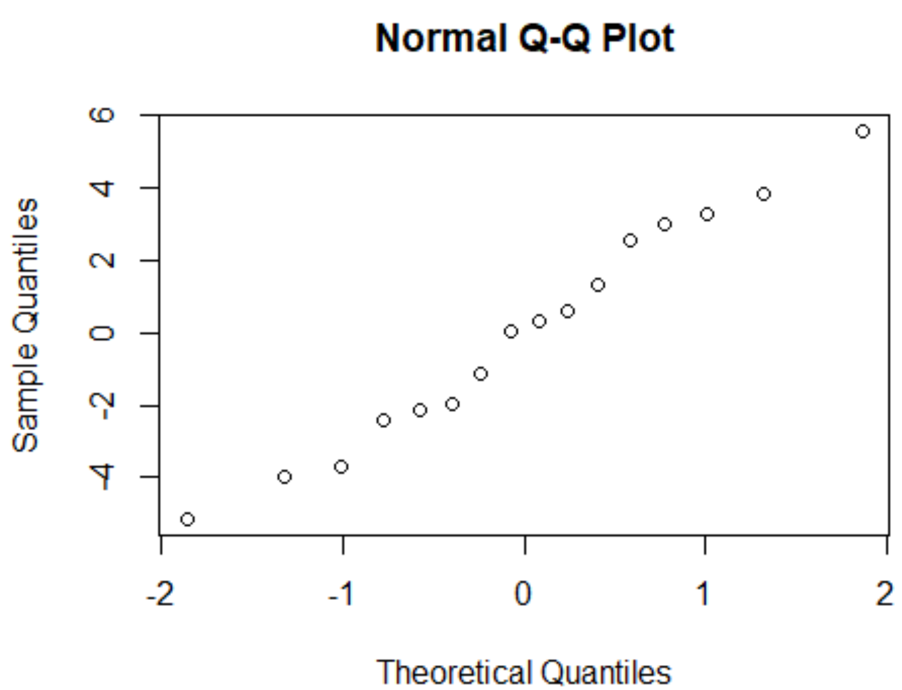


normal 가정이 적당해보인다.

c)

normal Probability plot of Residuals

```
par(mforw=c(2,2))

## Warning in par(mforw = c(2, 2)): "mforw"는 그래픽 매개변수가 아닙니다

qqnorm(lm.ta306$resid)
```



coefficient correalation

```
resid<-lm.ta306$resid
newresid = sort(resid)
k=1:16
z= qnorm((k-0.375)/(16+0.25))
mse = sum(resid^2)/(16-2)
expected = z*sqrt(mse)
sxy = sum((expected - mean(expected))*(newresid-mean(newresid)))
syy = sum((newresid- mean(newresid))^2)
sxx = sum((expected - mean(expected))^2)
```

```r
r=sxy/sqrt(sxx*syy)
r

## [1] 0.9916733
```

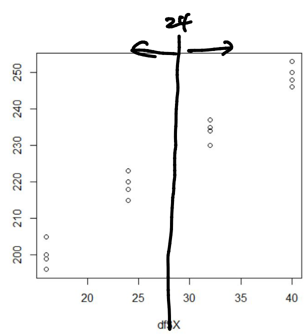
H_0: normal
H_1: not normal
$r = 0.9916$
$r \geq 0.941$ 이므로, H_0 채택
즉, normality 하다고 할 수 있다.

d)

	expected		Actual
25%	$t(0.25, 14) = -0.692$	$t(0.25, 14) \times \sqrt{MSE} = -2.24$	4번째: -2.625
50%	$t(0.5, 14) = 0$	$t(0.5, 14) \times \sqrt{MSE} = 0$	8번째: 0.025
75%	$t(0.75, 14) = 0.692$	$t(0.75, 14) \times \sqrt{MSE} = 2.24$	12번째: 2.575

$\Rightarrow$ consistent

e) Brown-Forsythe Test



$n = n_1 + n_2$

$$\begin{cases} X \leq 24: & n_1 = 8 \quad \bar{d}_1 = \frac{\sum d_{i1}}{n_1} = 2.93125 \doteq 2.931 \\ X > 24: & n_2 = 8 \quad \bar{d}_2 = \frac{\sum d_{i2}}{n_2} = 2.19375 \doteq 2.194 \end{cases}$$

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n-2}$$

$$= \frac{(2.931 - 2.194)}{1.724\sqrt{\frac{1}{8} + \frac{1}{8}}} \qquad = (1.724)^2$$

$$= 0.86$$

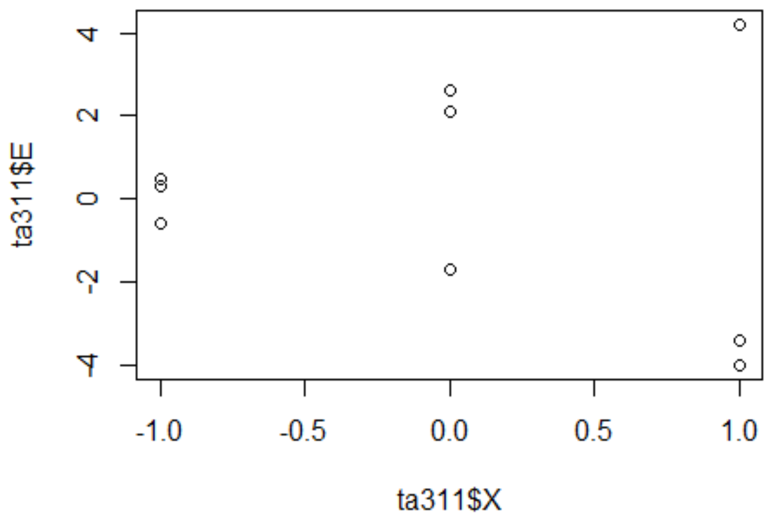
$t(0.975, 14) = 2.145$
If $t^* \leq 2.145$: constant
If $t^* > 2.145$: constant X

```r
df <- read.table("c:/Users/KimMinyoung/Documents/CH01PR22.txt")
names(df)<-c("Y","X")
ord <- order(df$X)
df<-df[ord,]
attach(df)

lm.df <-lm(Y~X, data= df)
resid <- lm.df$resid
abs.r0 <-abs(resid[df$X<=24]-median(resid[df$X<=24]))
abs.r1 <-abs(resid[df$X>24]-median(resid[df$X>24]))
abs.r <-c(abs.r0, abs.r1)
t.test(abs.r0,abs.r1,var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  abs.r0 and abs.r1
## t = 0.85579, df = 14, p-value = 0.4065
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.110838  2.585838
## sample estimates:
## mean of x mean of y
##   2.93125   2.19375
```

3.11 번

a)

```r
ta311 <- read.table("c:/Users/KimMinyoung/Documents/CH03PR11.txt")
names(ta311) <- c("X", "E")
library(ggplot2)
plot(ta311$X, ta311$E)
```



대략적으로 보기에는 X 가 커질수록 e 가 조금 퍼져 not constant 할 것 같다. 하지만 data 의 수가 적으므로 어떻게 결론을 내릴지는 유의수준을 어떻게 설정하냐에 따라 갈릴 것 같다. 아래와 같이 alpha=0.05 로 해서 Test 하면 constant 하다고 할 수 있다.

b) (3.10) : $\log_e \delta_i^2 = r_0 + r_1 X_i$ 가 적용가능하다고 했을때, Bruech-Pagan Test

$\Rightarrow SSR^* = 330.042$

$SSE = 59.960$

$X^2_{BP} = \frac{SSR^*}{2} \Big/ \left(\frac{SSE}{n}\right)^2 = \frac{330.042}{2} \Big/ \left(\frac{59.960}{9}\right)^2 = 3.72$

$\chi^2(0.95, 1) = 3.84$

$H_0: r_1 = 0$ (Constant)

$H_1: r_1 \neq 0$ (constant x)

If $X^2_{BP} \leq \chi^2(0.95, 1)$ 귀무가설 기각 x

$X^2_{BP} > \chi^2(0.95, 1)$ 귀무가설 기각

$\underset{3.72}{X^2_{BP}} \leq \underset{3.84}{\chi^2(0.95, 1)}$ 이므로 귀무가설은 기각하지않는다.

따라서 유의수준 $\alpha=0.05$에서는 constant 하다고 할 수 있다.

```r
df <- ta311
df$ei <- (df$E^2)
lmssr<-lm(df$ei~df$X)
summary(lmssr)
```

```
##
## Call:
## lm(formula = df$ei ~ df$X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7722 -2.2522  0.8444  1.1144  3.5611
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6622     0.8451   7.883 0.000100 ***
## df$X          7.4167     1.0350   7.166 0.000183 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 7 degrees of freedom
## Multiple R-squared:   0.88,  Adjusted R-squared:  0.8629
## F-statistic: 51.35 on 1 and 7 DF,  p-value: 0.0001828
```

```r
anova(lmssr)
```

```
## Analysis of Variance Table
##
## Response: df$ei
##           Df Sum Sq Mean Sq F value    Pr(>F)
## df$X       1 330.04  330.04  51.348 0.0001828 ***
## Residuals  7  44.99    6.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.14 번

a) Lack Of Fit Test

H_0 : $E(Y_{ij}|X_{ij}) = B_0 + B_1 X_{ij}$ (reduced model)
H_1 : $E(Y_{ij}|X_{ij}) = \mu_i$ (Full model)

H_0 : $E(Y) = B_0 + B_1 X$
H_1 : $E(Y) \neq B_0 + B_1 X$

$SSE(F) = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 = SSPE$

$= 128.750$

$df_F = n - c = 16 - 4 = 12$

$SSE(R) = \sum\sum (Y_{ij} - (b_0 + b_1 X_j))^2 = \sum\sum (Y_{ij} - \hat{Y}_{ij})^2 = SSE$

$= 146.427$

$df_R = n - 2 = 16 - 2 = 14$

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \Big/ \frac{SSE(F)}{df_F} = \frac{146.43 - 128.75}{14 - 12} \Big/ \frac{128.75}{12}$$

$$= 0.8237$$

$F(0.99, 2, 12) = 6.93$

If $F^* \leq F(1-\alpha, c-2, n-c)$: H_0 채택
$F^* > F(1-\alpha, c-2, n-c)$: H_1 채택

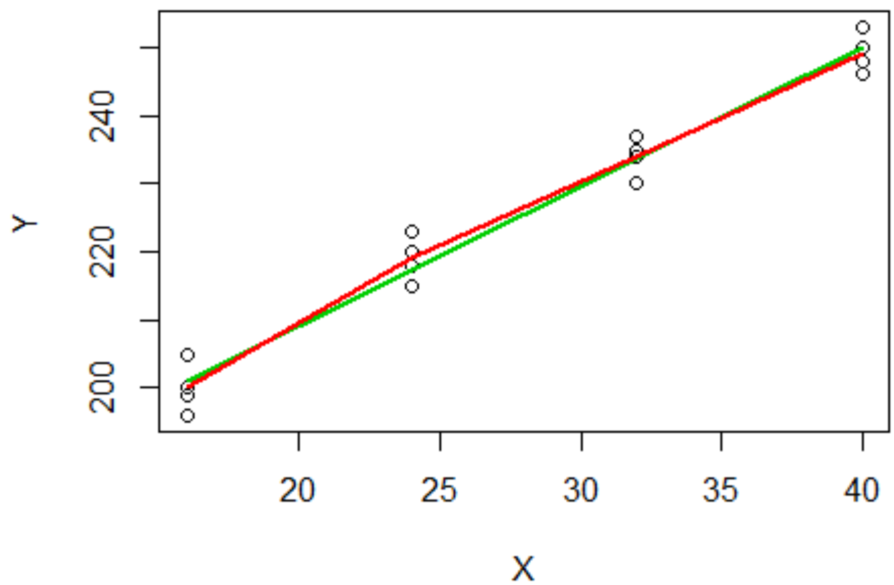
If $F^* \leq F(0.99, 2, 12)$: H_0 채택
$F^* > F(0.99, 2, 12)$: H_1 채택

$\Rightarrow$ $\underset{0.824}{F^*} \leq \underset{6.93}{F(0.99, 2, 12)}$ 이므로 H_0을 채택한다.

따라서 reduced model을 채택한다.

즉, 우리의 회귀함수는 선형관계에 있다고 할 수 있다.

```r
ta01 <- read.table("c:/Users/KimMinyoung/Documents/CH01PR22.txt")
names(ta01) <- c("Y", "X")
new <- data.frame(mean=with(tapply(Y, factor(X), mean), data=ta01))
new <- data.frame(X=as.numeric(row.names(new)), new)
full <- lm(Y~factor(X), data=ta01)
smaller <- lm(Y~X, data=ta01)
with(plot(X, Y), data=ta01)
lines(ta01$X, smaller$fitted, col=3, lwd=2)
with(lines(X, mean, col=2, lwd=2), data=new)
```




```r
anova(smaller, full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ factor(X)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 146.43
## 2     12 128.75  2    17.675 0.8237 0.4622
```

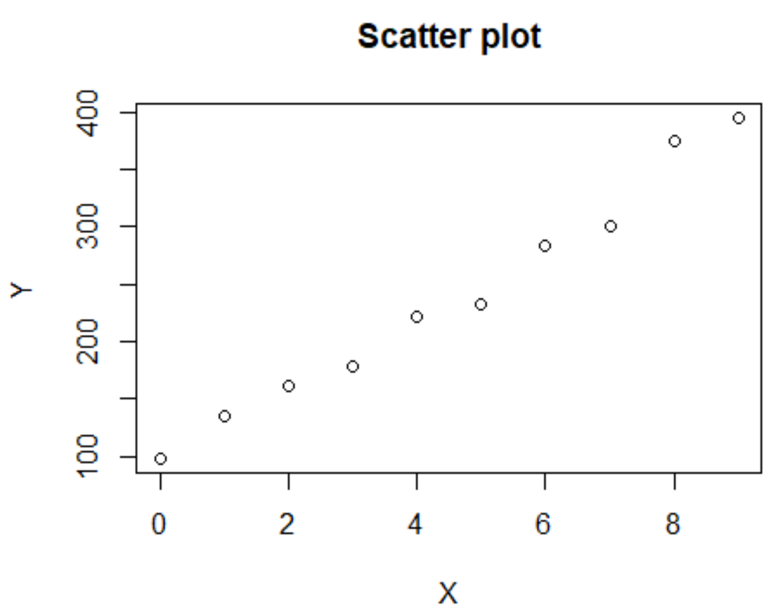
b) 장점이 있다. B_0 B_1 두개만 추정하면 되고 단순해서 좋다.
나중에 여러가지 계산하기에도 편하다..
큰 단점은 딱히 없는것 같다.

c) lack of fit test에서는 선형인지 비선형인지를 알려준다.
lack of fit test에서 not linear 하다는 결론을 도출할때,
어떤 회귀함수가 적절한지에 대한 정보는 나타내지는 않는다.
그래서 뒤에 단원에서 배운 transform을 해서 회귀함수를 정해야 할것 같다

3.17 번

a) scatter plot of data

```r
ta317<-read.table("c:/Users/KimMinyoung/Documents/CH03PR17.txt")
names(ta317) <- c("Y", "X")
plot(ta317$X, ta317$Y, main="Scatter plot", xlab="X", ylab="Y")
```



선형관계에 있는 것 처럼 보인다.

b) Box Cox

$$W_i = k_1(Y_i^{\lambda} - 1), \text{ if } \lambda \neq 0 \qquad k_1 = 1/\lambda k_2^{\lambda-1}$$
$$= k_2(\log_e Y_i), \text{ if } \lambda = 0 \qquad k_2 = \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}$$

⇒ 이공식 그대로 적용해서 함수를 구현해 풀어보자

```r
df<-ta317
transform_sse<-function(lambda){ # transform 후 SSE 계산
  n<-10
  k2<-(prod(df$Y))^(1/n)
  k1<-1/(lambda*k2^(lambda-1))
  wi<-k1*(df$Y^(lambda)-1)
  lm<-lm(wi~df$X)
  ei<-lm$resid
  SSE<-sum(ei^2)
  paste("lambda:",lambda,"SSE:",SSE)
}

transform_sse(0.3)

## [1] "lambda: 0.3 SSE: 1099.70927132159"

transform_sse(0.4)

## [1] "lambda: 0.4 SSE: 967.908780410858"

transform_sse(0.5)

## [1] "lambda: 0.5 SSE: 916.404798150164"

transform_sse(0.6)

## [1] "lambda: 0.6 SSE: 942.449764952538"

transform_sse(0.7)

## [1] "lambda: 0.7 SSE: 1044.2384001476"
```

→ $\lambda = 0.5$일때 적당해보인다.

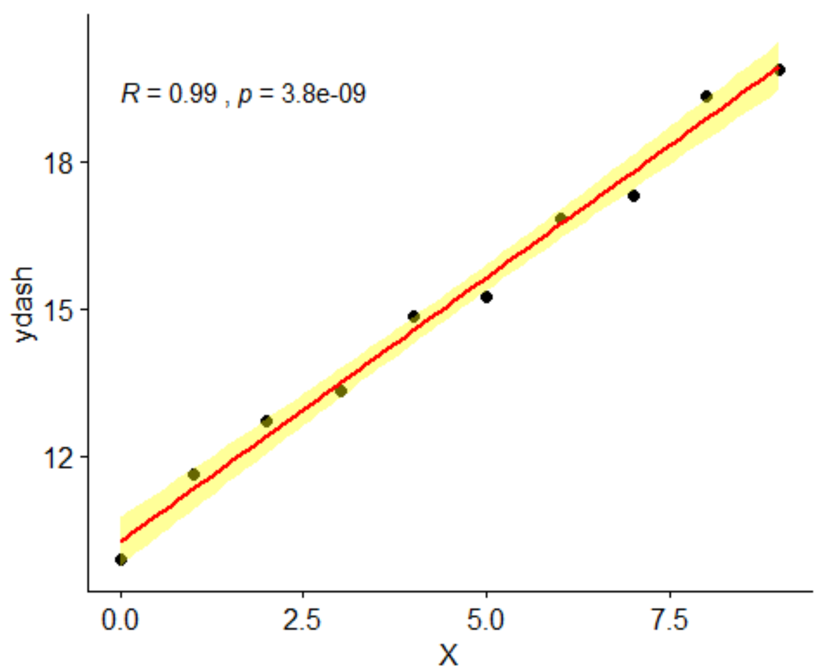
c) $Y' = \sqrt{Y}$

$$\hat{Y}' = 10.261 + 1.076X$$

```
ta317$ydash <- sqrt(ta317$Y)
fit <- lm(ydash~X, data=ta317)
summary(fit)
## Call:
## lm(formula = ydash ~ X, data = ta317)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47447 -0.30811  0.01549  0.29541  0.46781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.26093    0.21290   48.20 3.80e-11 ***
## X            1.07629    0.03988   26.99 3.83e-09 ***
## ---
## Residual standard error: 0.3622 on 8 degrees of freedom
## Multiple R-squared:  0.9891, Adjusted R-squared:  0.9878
## F-statistic: 728.4 on 1 and 8 DF,  p-value: 3.826e-09
```

d) plot

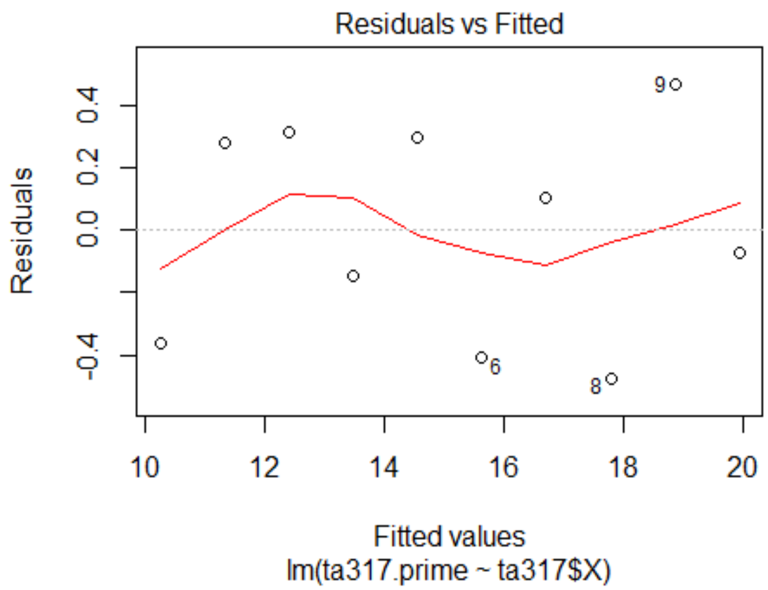
```
ta317$ydash <- sqrt(ta317$Y)
library(ggpubr)

ggscatter(ta317, x="X", y="ydash", add="reg.line", conf.int=TRUE, add.params=
list(color="red", fill="yellow"))+stat_cor(method="pearson")
```

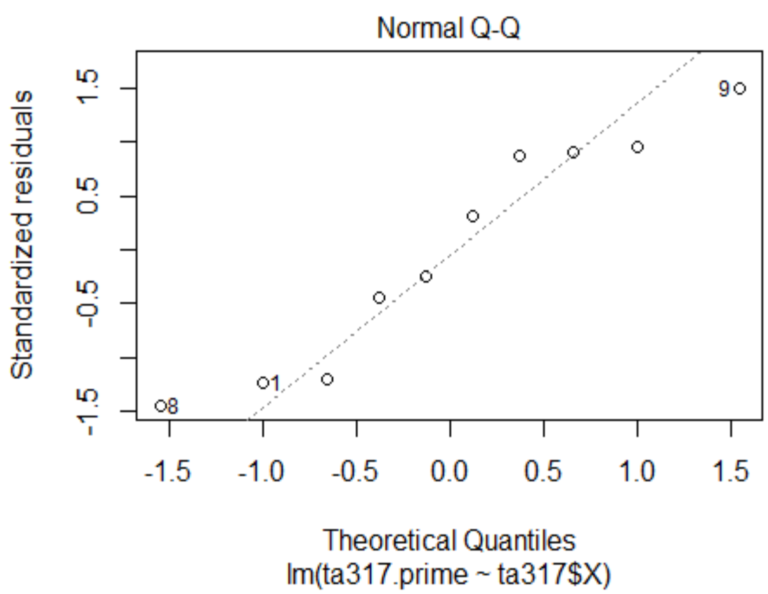


적절하게 fit 된 것 같다.

e)

```
library(MASS)
ta317.prime <- sqrt(ta317$Y)
ta317.prime.lm <- lm(ta317.prime~ta317$X)
ta317.res <- ta317.prime.lm$residuals
ta317.fitted <- ta317.prime.lm$fitted.values
plot(ta317.prime.lm, which=c(1,2))
```





이를 통해 어느정도 normal 한 것을 확인할 수 있다.

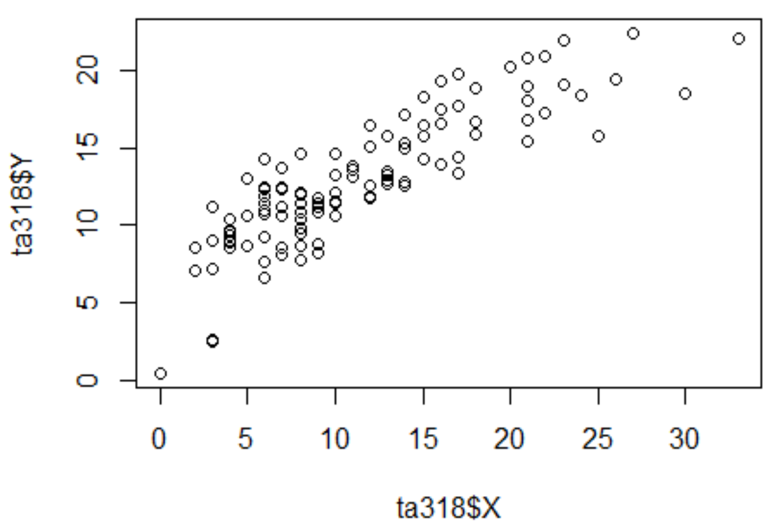
f) $\hat{Y}=(10.261+1.076X)^2$

```
ta317.prime.lm

##
## Call:
## lm(formula = ta317.prime ~ ta317$X)
##
## Coefficients:
## (Intercept)      ta317$X
##      10.261        1.076
```

3.18 번

a) Scatter Plot

```
ta318 <- read.table("c:/Users/KimMinyoung/Documents/CH03PR18.txt")
names(ta318) <- c("Y", "X")
plot(ta318$X, ta318$Y)
```



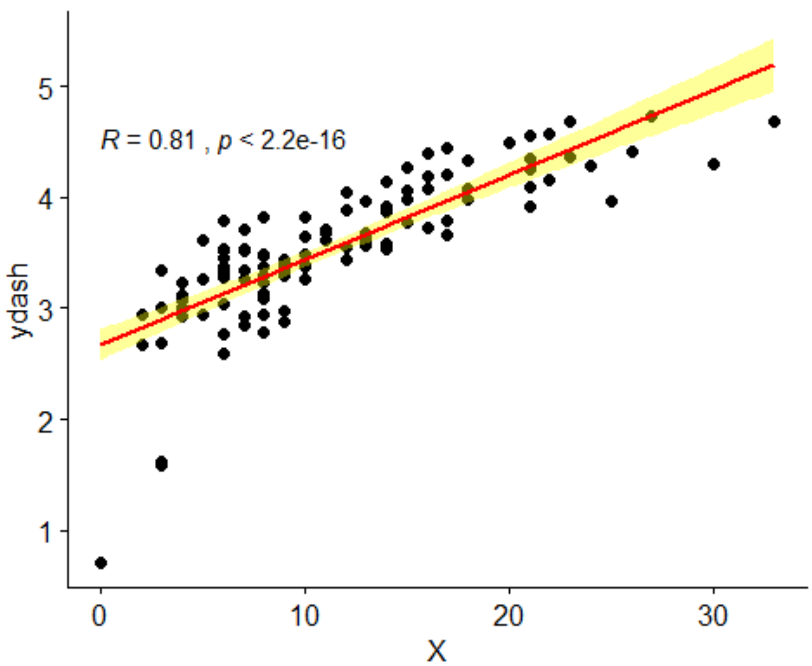
약간 곡선형태를 띄는 것 같으므로, transform 하면 더 좋을 것 같다.

b)

```
ta318 <- read.table("c:/Users/KimMinyoung/Documents/CH03PR18.txt")
names(ta318) <- c("Y", "X")
library(MASS)
ta318.prime <- sqrt(ta318$Y)
ta318.prime.lm <- lm(ta318.prime~ta318$X)
ta318.prime.lm

##
## Call:
## lm(formula = ta318.prime ~ ta318$X)
## Coefficients:
## (Intercept)      ta318$X
##     2.67386      0.07621
```

c)

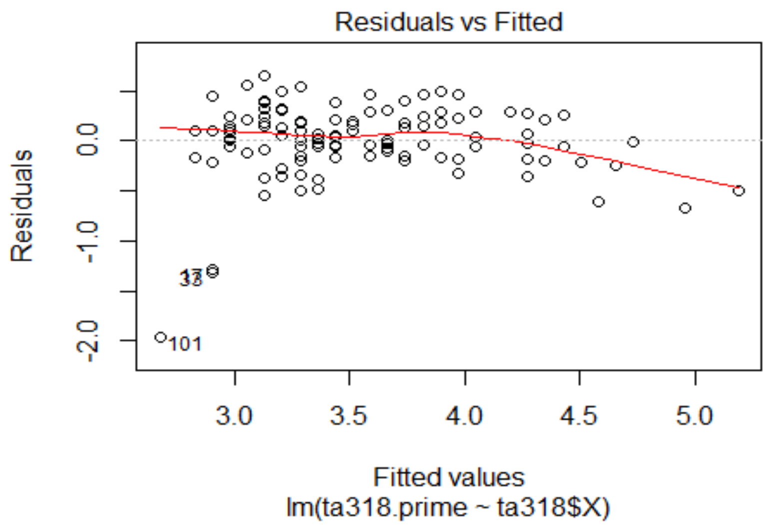
```
library(MASS)
ta318$ydash <- sqrt(ta318$Y)
fit <- lm(ydash~X, data=ta318)
ta318$ydash <- sqrt(ta318$Y)
library(ggpubr)
ggscatter(ta318, x="X", y="ydash", add="reg.line", conf.int=TRUE, add.params=
list(color="red", fill="yellow"))+stat_cor(method="pearson")
```

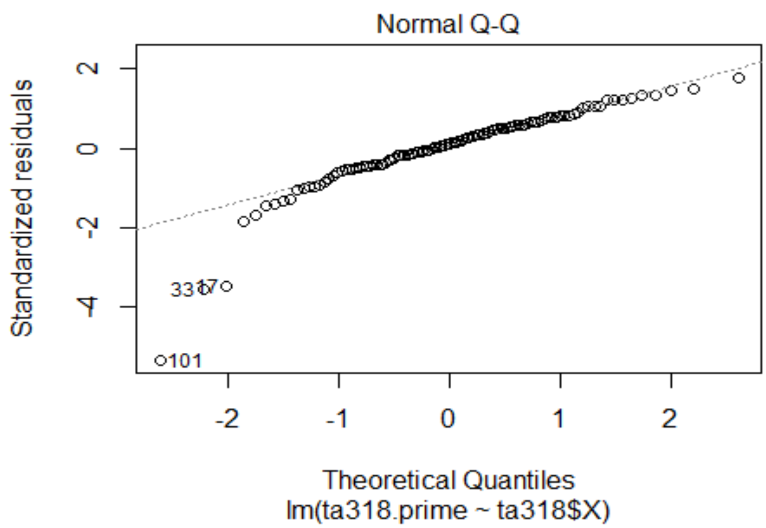


잘 fit 된 것 같다.

d)

```
ta318 <- read.table("c:/Users/KimMinyoung/Documents/CH03PR18.txt")
names(ta318) <- c("Y", "X")
library(MASS)
ta318.prime <- sqrt(ta318$Y)
ta318.prime.lm <- lm(ta318.prime~ta318$X)
ta318.res <- ta318.prime.lm$residuals
ta318.fitted <- ta318.prime.lm$fitted.values
plot(ta318.prime.lm, which=c(1,2))
```





위를 통해 normal 한 것을 확인할 수 있다.

e) $\hat{Y} = 1.254170 + 3.623352\sqrt{X}$

```
ta318.prime.lm

##
## Call:
## lm(formula = ta318.prime ~ ta318$X)
##
## Coefficients:
## (Intercept)      ta318$X
##     2.67386      0.07621
```