Homework 5

Luo Bingjun 2017013573 Software 71

2019-4-11

Prob-1: KNNL 3.7

```
data = read.table(file='CH01PR27.txt', header=F)
n=60
colnames(data) <- c('Y', 'X')
attach(data)
```

a.

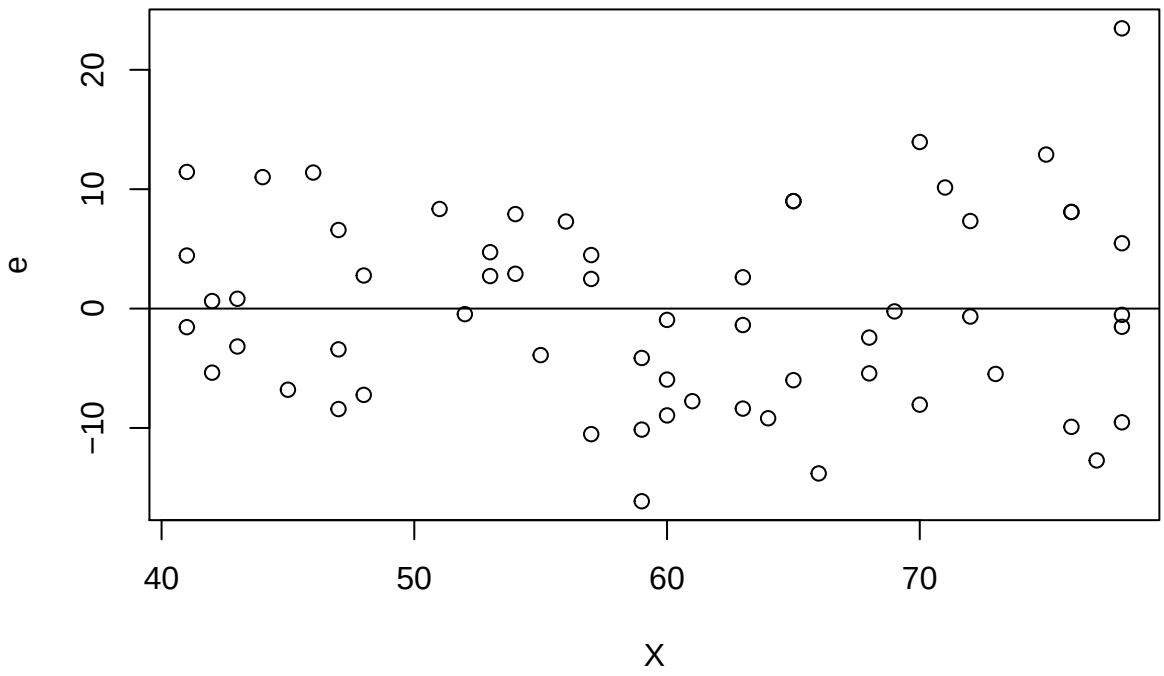
```
stem(X)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   4 | 11122334
##   4 | 5677788
##   5 | 123344
##   5 | 56777999
##   6 | 00013334
##   6 | 5556889
##   7 | 001223
##   7 | 5666788888
```

It is consistent with the random selection because the number of women in each 10-year age group is approximately the same.

b.

```
reg <- lm(Y~X)
# e_i
e=reg$residuals
plot(X,e)
abline(0,0)
```

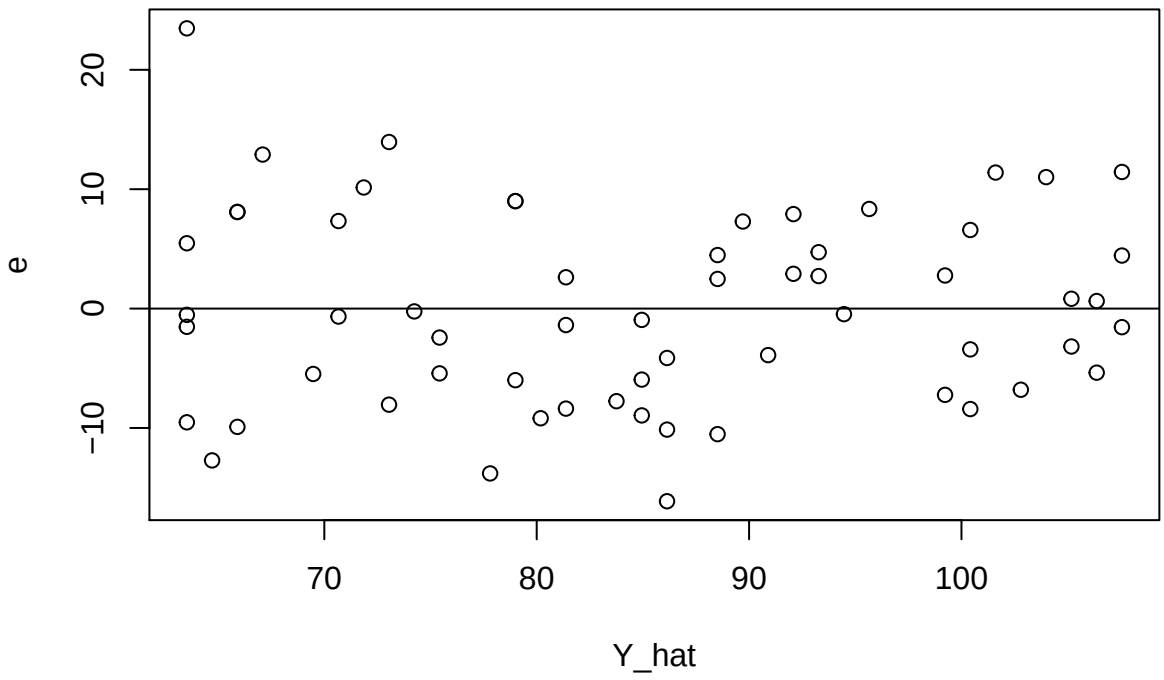


The plot shows the variance of e_i seems not to be the same, because the variance of two side is greater than center.

c.

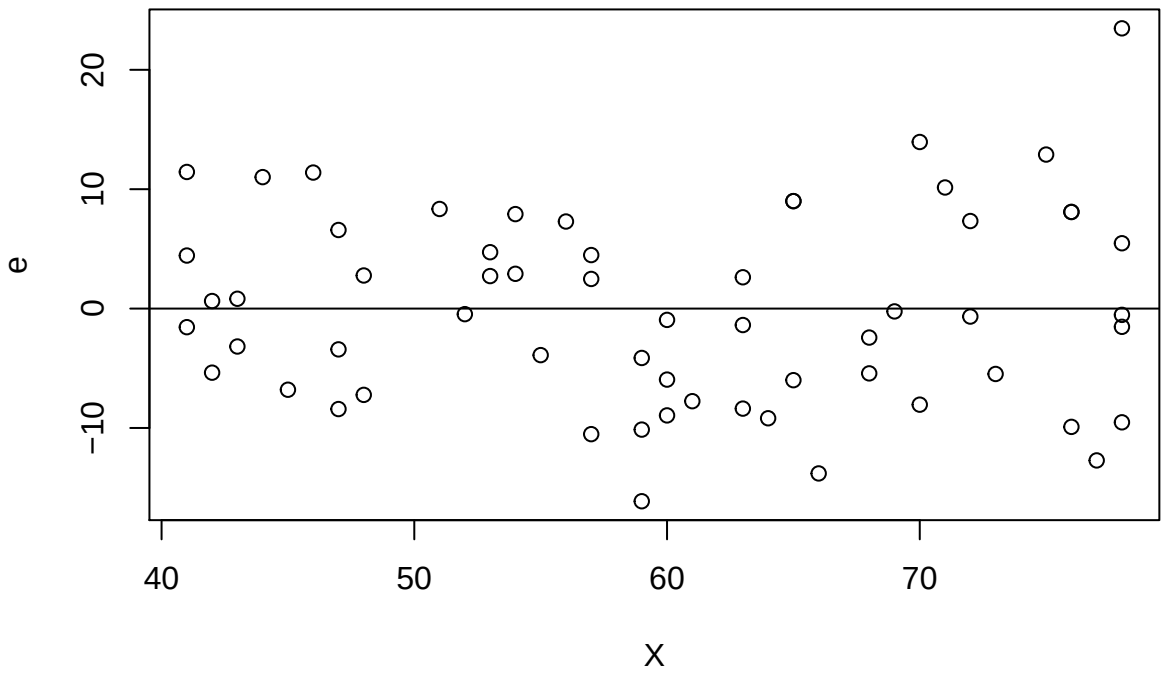
e_i against $\hat{Y}_i$:

```
Y_hat=predict(reg,data.frame(X))
plot(Y_hat,e)
abline(0,0)
```



e_i against X_i:

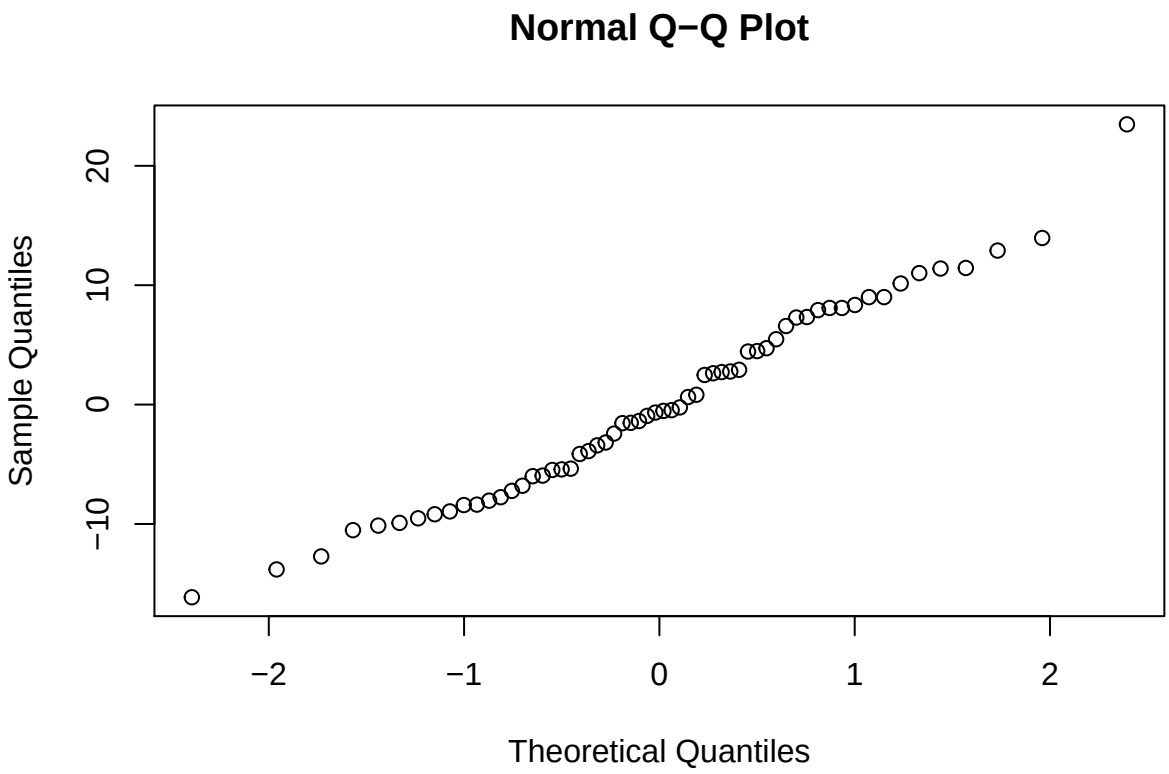
```
plot(X,e)
abline(0,0)
```



They provide the same information because $\hat{Y}_i$ has a linear relation with X_i.

d.

```
qqnorm(e)
```




```
# ordered residuals
ordered <- sort(e)
# expected values
expected <- qnorm(rank(ordered)/(length(ordered) + 1))
# coefficient of correlation
cor(ordered,expected)
```

```
## [1] 0.9897475
```

Because $0.9897475 > 0.984$, we conclude that the normal assumption is tenable.

e.

$$H_0 : \gamma_1 = 0 \leftrightarrow H_1 : \gamma_1 \neq 0$$

```
alpha=0.05
SSE=sum(e^2)
reg_star=lm(e^2~X)
SSR_star=anova(reg_star)[,2][2]
# test statistic
(SSR_star/2)/(SSE/n)^2
```

```
## [1] 47.97437
```

```
qchisq(1-alpha, 1)
```

```
## [1] 3.841459
```

```
detach(data)
```

Since $X^2_{BP} > \chi^2_{0.95;1}$, we conclude H_1, that error variance is not constant.

It is consistent with my preliminary findings.

Prob-2: KNNL 3.15

```
data = read.table(file='CH03PR15.txt', header=F)
n=15
c=5
colnames(data) <- c('Y', 'X')
attach(data)
```

a.

```
reg<-lm(Y~X)
reg
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       2.575       -0.324
```

The regression function is

$$\hat{Y} = -0.324X + 2.575$$

b.

$$H_0 : EY = \beta_0 + \beta_1 X \leftrightarrow H_1 : EY \neq \beta_0 + \beta_1 X$$

```
Reduced <- lm(Y ~ X)
Full <- lm(Y ~ 0 + as.factor(X), data)
anova(Reduced, Full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ 0 + as.factor(X)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 2.9247
## 2     10 0.1574  3    2.7673 58.603 1.194e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
alpha=0.025
qf(1-alpha,c-2,n-c)
```

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

Since $F^* > F_{0.975;3,10}$, we conclude H_1, that the regression function is not linear.

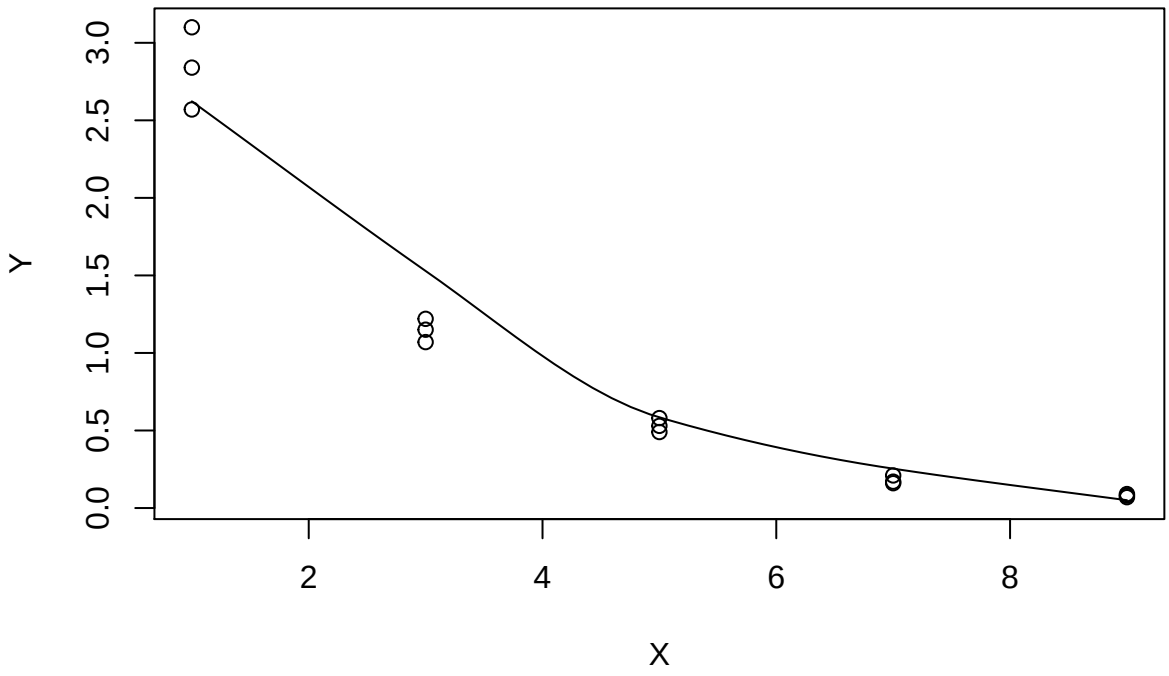
c.

The test in (b) doesn't indicate what regression function is approriate because it is all under H_0, that the regression function is linear.

Prob-3: KNNL 3.16

a.

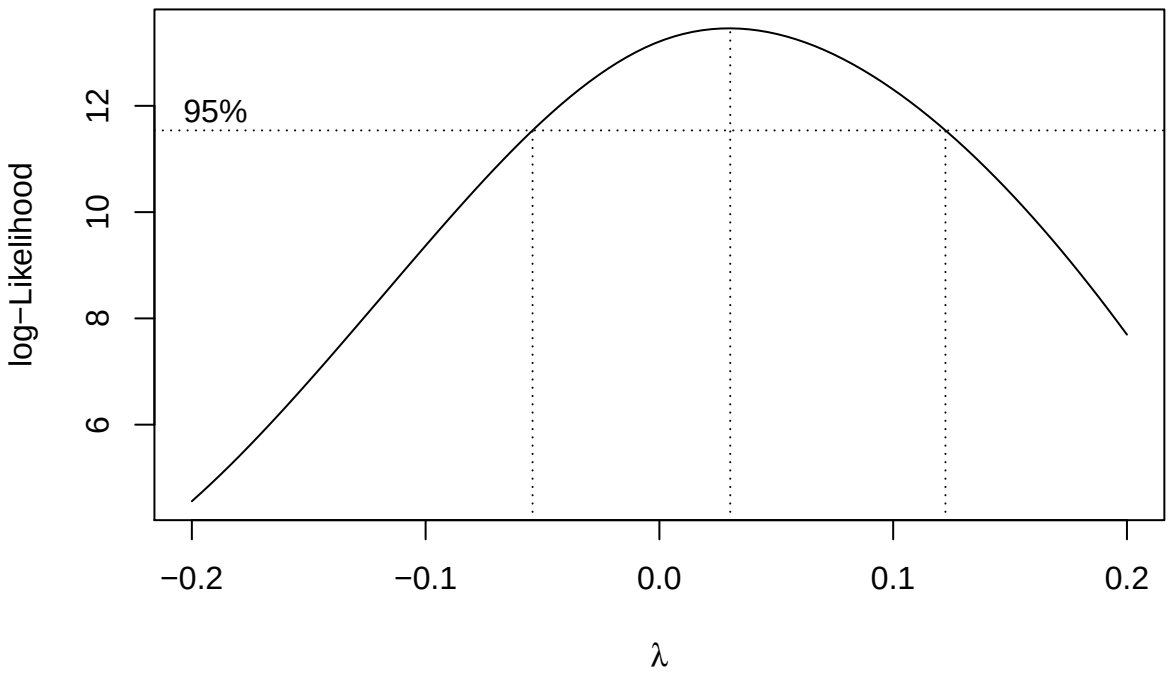
```
scatter.smooth(Y~X)
```



I might try $Y' = \log_{10} Y$.

b.

```
library(MASS)
x= boxcox(Y ~ X, lambda = c(-.2, -.1, 0, .1, .2))
```



It suggests that $Y' = Y^{0.03}$.

c.

```
Y1=log10(Y)
reg<-lm(Y1~X)
reg
```

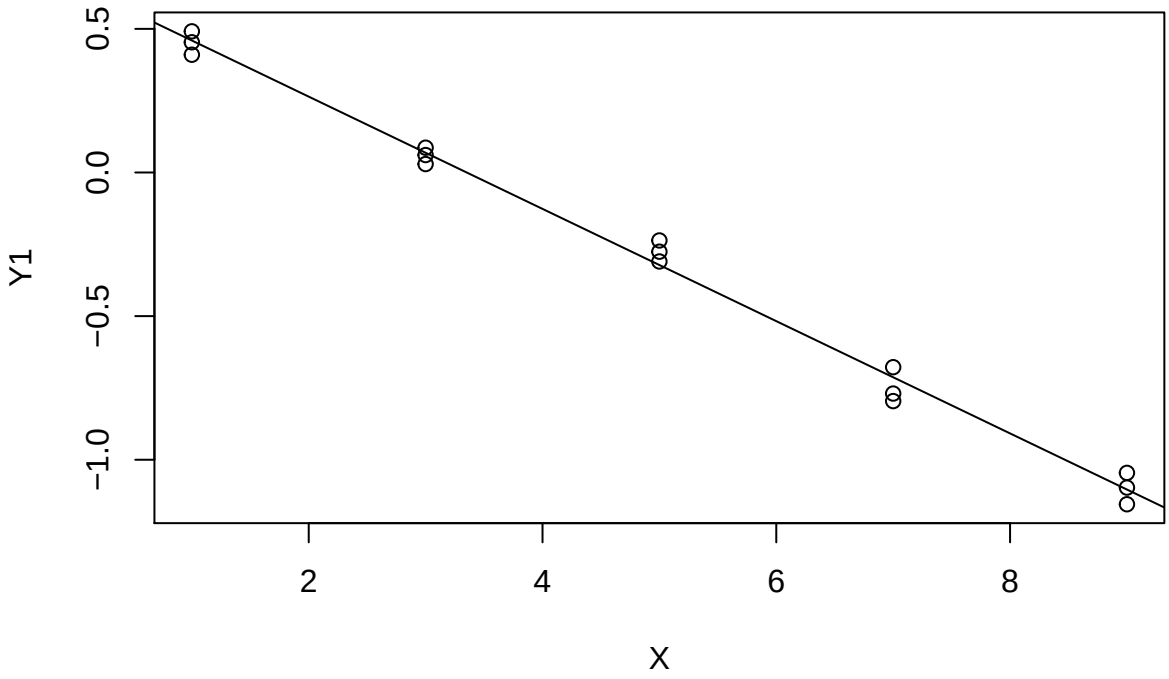
```
##
## Call:
## lm(formula = Y1 ~ X)
##
## Coefficients:
## (Intercept)            X
##      0.6549      -0.1954
```

The regression function is

$$\hat{Y'} = -0.1954X + 0.6549$$

d.

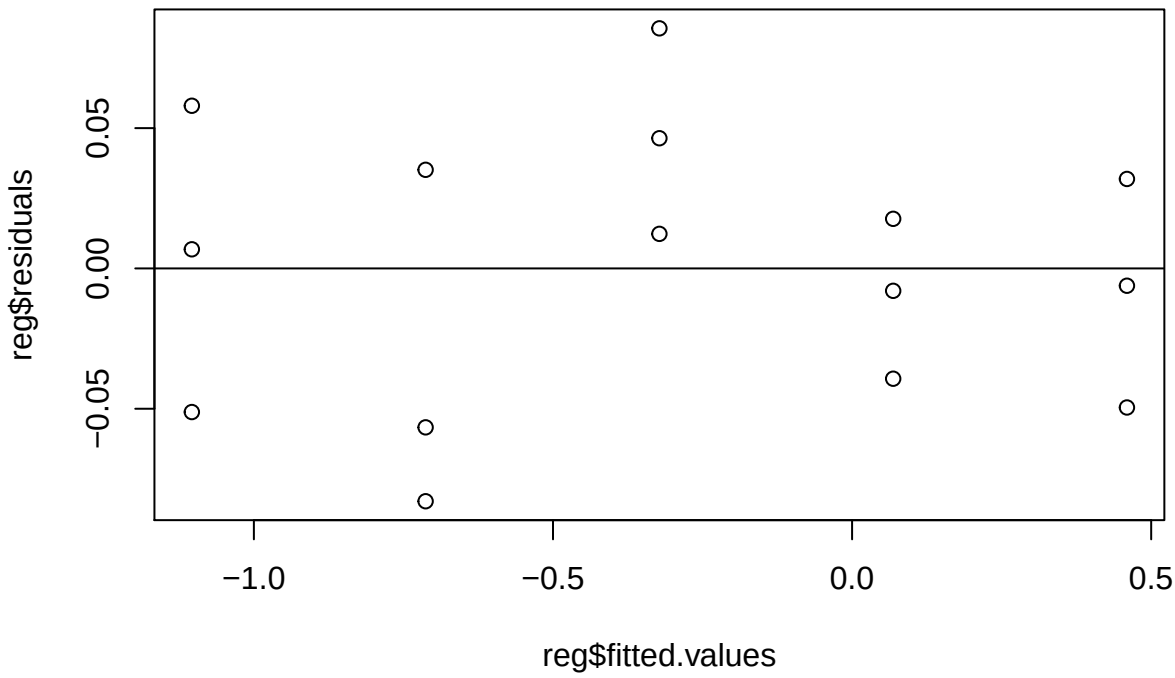
```
plot(X,Y1)
abline(reg)
```



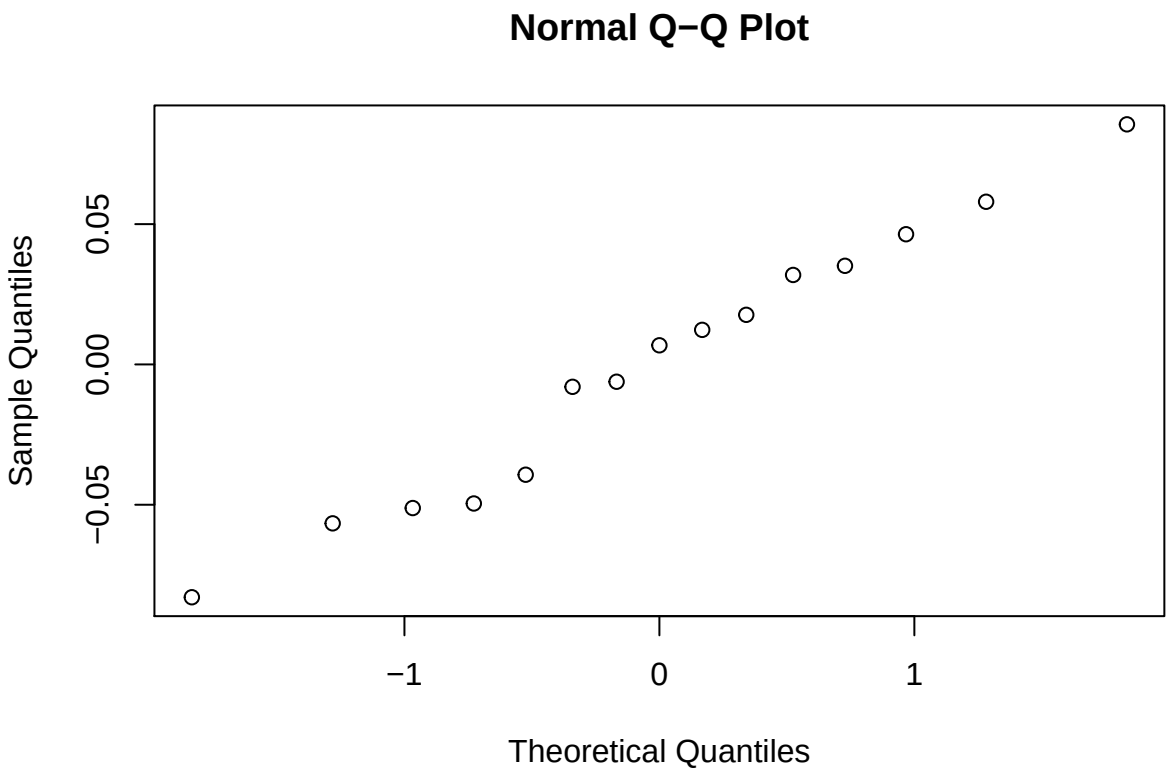
It appears to be a good fit.

e.

```
plot(reg$fitted.values,reg$residuals)
abline(0,0)
```




```
qqnorm(reg$residuals)
```




```
detach(data)
```

The plots show that residuals appear to be linear and normally distributed.

f.

The original regression function is

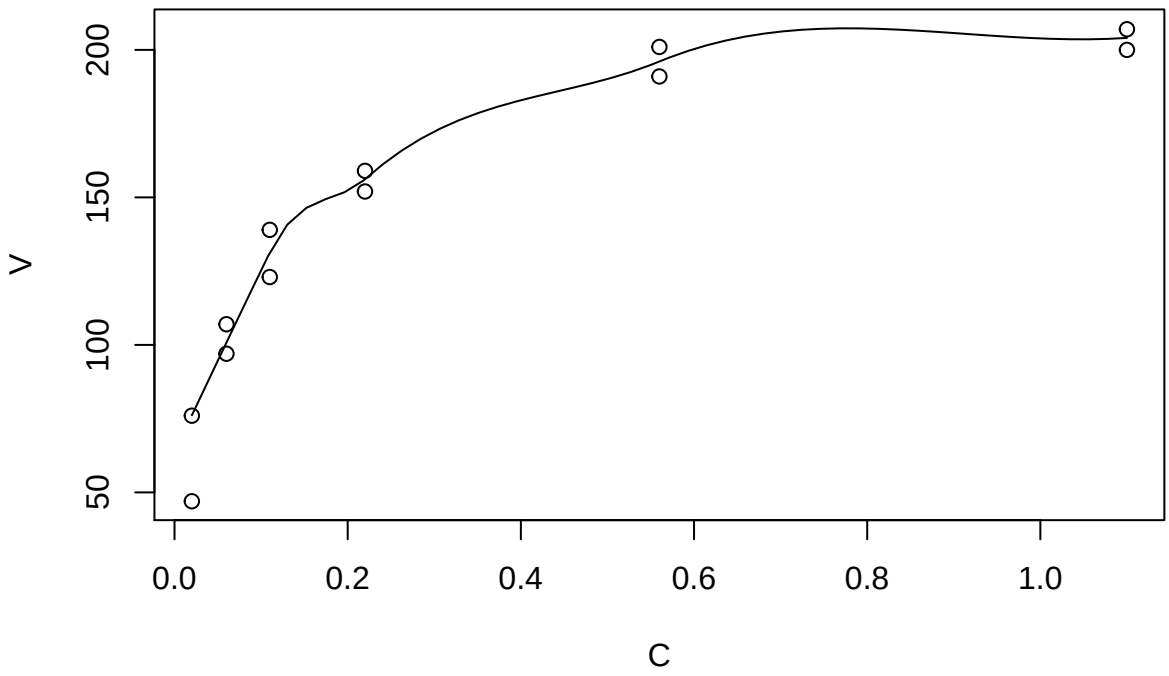
$$\hat{Y} = 10^{-0.1954X+0.6549}$$

Prob-4:

```
data = read.table(file='PROB4.txt', header=T)
colnames(data) <- c('C', 'V')
n=12
attach(data)
```

a.

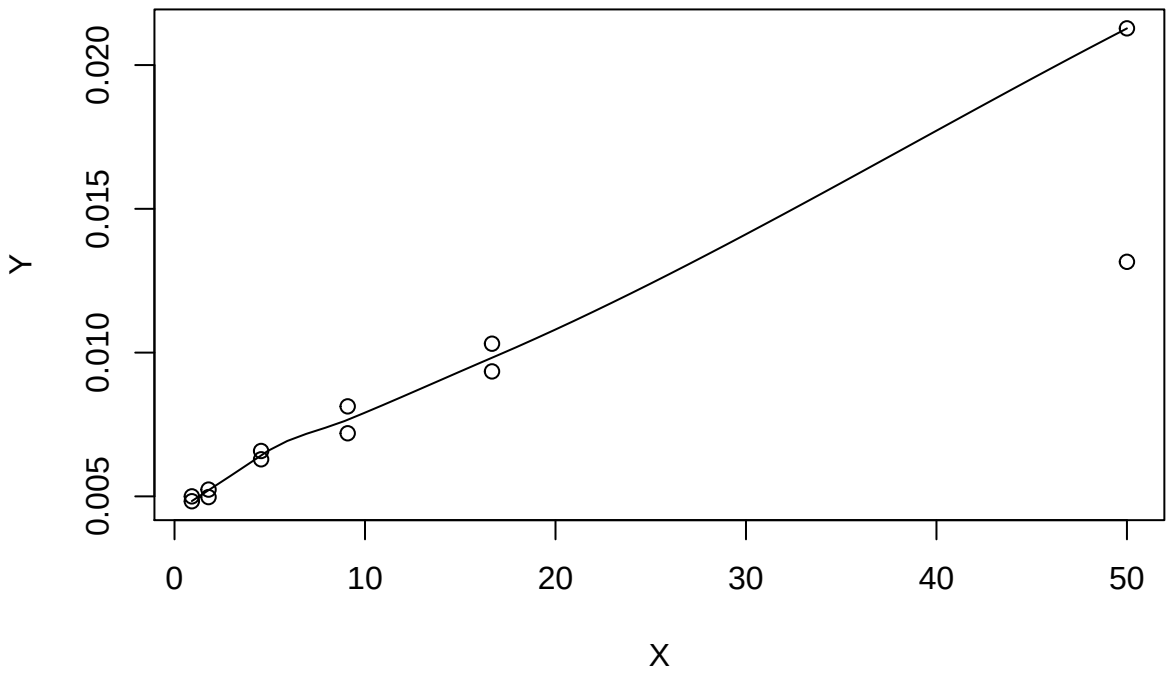
```
scatter.smooth(V~C)
```



It isn't linear at all, but fits one of the prototype in Figure 3.15.

b.

```
# 1/V
Y=1/V
# 1/C
X=1/C
scatter.smooth(Y~X)
```



The fit appears linear, but the independent variance assumption appears unconstant.

c.

They are different from each other. The distribution of 1/V may be more influential in determining the fit.

d.

```
reg<-lm(Y~X)
reg
```

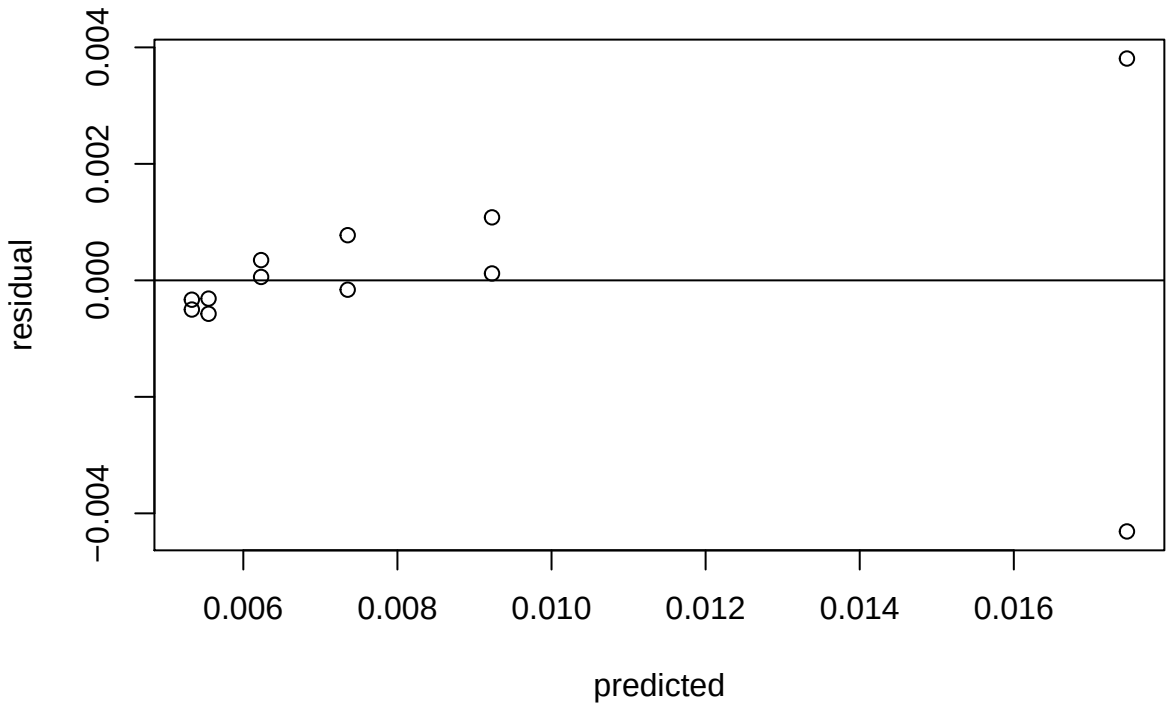
```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##   0.0051072    0.0002472
```

The regression line is

$$\frac{1}{V} = 0.002472\frac{1}{C} + 0.0051072$$

```
residual=reg$residuals
predicted=reg$fitted.values
```

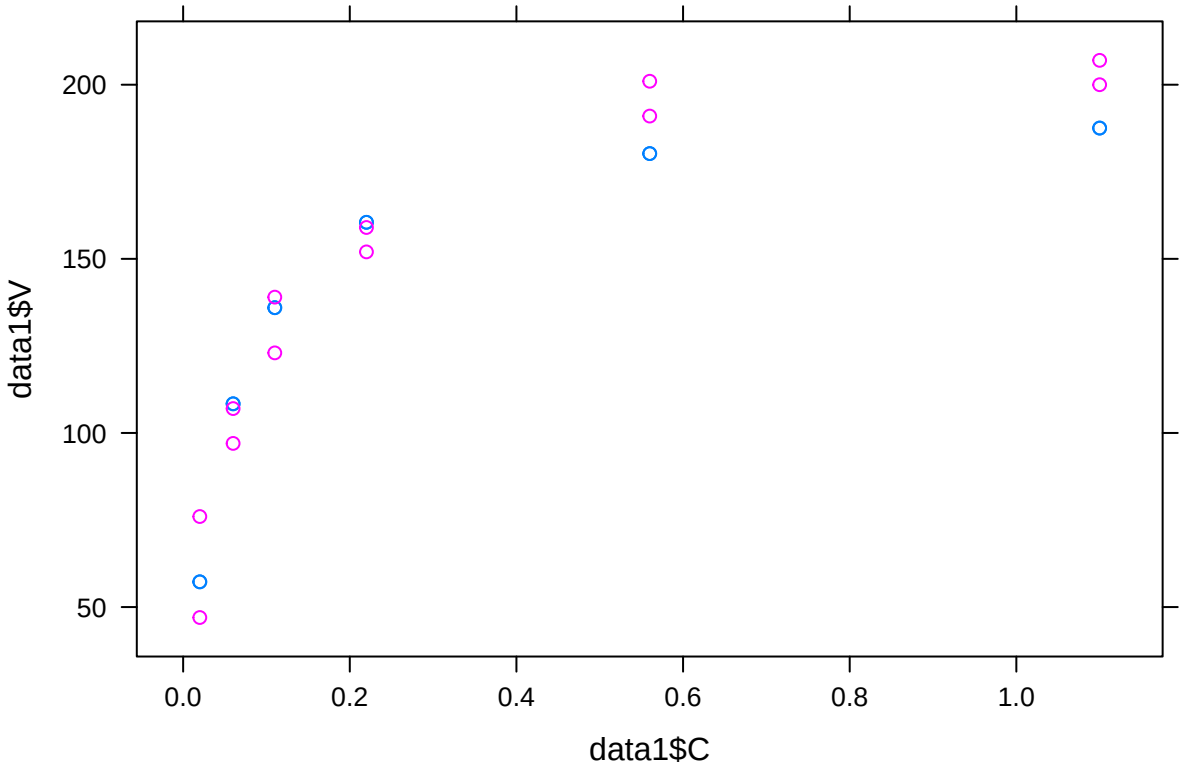
```
plot(predicted,residual)
abline(0,0)
```



The plot suggets that the variance of residuals may be not constant.

e.

```
V_hat=1/predicted
data1=rbind(data.frame(V=V_hat, group='V_hat',C=C), data.frame(V=V, group='V', C=C))
library(lattice)
xyplot(data1$V~data1$C, groups=data1$group)
```



The pink points in the plot are V, while the blue ones are the predicted values of V.

The fit appears well when $C < 0.2$, but as C increases, it becomes worse and worse, because the residual variance varies with C.