Homework 4

Luo Bingjun 2017013573

2019-04-05

Prob-1: KNNL 2.31

```
crime = read.table(file='CH01PR28.txt', header=F)
n=84
alpha=0.01
colnames(crime) <- c('Y', 'X')
attach(crime)
```

a.

```
regCrime <- lm(Y ~ X)
anova(regCrime)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.

$$H_0 : \beta_1 = 0 \leftrightarrow H_1 : \beta_1 \neq 0$$

In 2.30a,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum (X_i - X)^2)$$

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{SSE/(n-2)}{\sum (X_i - X)^2}}} \sim t_{n-2}$$

In this section,

$$F = \frac{MSR}{MSE} \sim F(1, n-2)$$

```
SST=sum((Y-mean(Y))^2)
SSE=sum((regCrime$residuals)^2)
SSR=SST-SSE
F=(SSR/1)/(SSE/(n-2))
p=qf(1-alpha,1,n-1)
c(F,p)
```

```
## [1] 16.83376  6.95044
```

Reject H_0 when $F > p = 6.95$

The P-value for the F test and the t test is not the same.

c.

The reduction of total variation is

```r
SST/(n-1)-SSE/(n-2)
```

```
## [1] 1059167
```

The reduction rate is

```r
(SST/(n-1)-SSE/(n-2))/SST/(n-1)
```

```
## [1] 2.325534e-05
```

So this is a relatively small reduction.

d.

```r
r=sqrt(SSR/SST)
r
```

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

Prob-2: KNNL 2.32

a.

Full model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Reduced model:

$$Y_i = \beta_0 + \epsilon_i$$

b.

For full model,

$$SSE(F) = SSE$$

$$df_F = n - 2$$

For reduced model,

$$SSE(R) = SST$$

$$df_R = n - 1$$

c.

$$F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} = \frac{(SSR)/1}{SSE/(n-2)} = \frac{MSR}{MSE} = F$$

Prob-3: Prove Equation 2.57 of KNNL

$$\hat{\beta_1} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \beta_1 + \frac{\sum(X_i - \bar{X})\epsilon_i}{\sum(X_i - \bar{X})^2}$$
$$Var(\hat{\beta_1}) = \frac{\sum(X_i - \bar{X})^2 Var(\epsilon_i)}{(\sum(X_i - \bar{X})^2)^2} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$
$$E(\hat{\beta_1}) = \beta_1,\ E(\hat{\beta_1}^2) = E(\hat{\beta_1}) + Var(\hat{\beta_1}) = \beta_1 + \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$
$$MSR = SSR = \sum(\hat{Y_i} - \bar{Y})^2 = \sum((\hat{\beta_0} + \hat{\beta_1}X_i) - (\hat{\beta_0} + \hat{\beta_1}\bar{X}))^2 = \hat{\beta_1}^2 \sum(X_i - \bar{X})^2$$
$$E(MSR) = E(\beta_1^2)\sum(X_i - \bar{X})^2 = \sigma^2 + \beta_1 \sum(X_i - \bar{X})^2$$

Prob-4: KNNL 2.56

```r
X<-c(1,4,10,11,14)
n=5
sigma=0.6
beta_0=5
beta_1=3
```

a.

$$E(MSR) = \sigma^2 + \beta_1 \sum(X_i - \bar{X})^2$$
$$E(MSE) = \frac{\sigma^2}{n-2}$$

```r
MSR=sigma^2+beta_1*sum((X-mean(X))^2)
MSR
```

```
## [1] 342.36
```

```r
MSE=sigma^2/(n-2)
MSE
```

```
## [1] 0.12
```

b.

It would have been worse to have made observations at X=6,7,8,9,10.

If we have made observations at X=6,7,8,9,10, $\sum(X_i - \bar{X})^2$ would have decrease. As it decrease, the variation of r^2 would be steeper as Y varied, so the result would be worse.

$$Y_i = 5 + 3X_i + \epsilon_i$$
$$\hat{Y} = \hat{\beta_0} + 8\hat{\beta_1} = \bar{Y} + \hat{\beta_1}(8 - \bar{X}) = \sum(\frac{(8 - \bar{X})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} + \frac{1}{n})Y_i$$
$$Var(\hat{Y}) = \sum(\frac{(8 - \bar{X})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} + \frac{1}{n})^2 Var(Y_i) = \sigma^2 \sum(\frac{(8 - \bar{X})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} + \frac{1}{n})^2$$

When we made observations at X=1,4,10,11,14, $Var(\hat{Y})$ is

```
v=sigma^2*sum(((8-mean(X))*(X-mean(X))/sum((X-mean(X))^2)+1/n)^2)
v
```

```
## [1] 0.072
```

When we made observations at X=6,7,8,9,10, $Var(\hat{Y})$ is

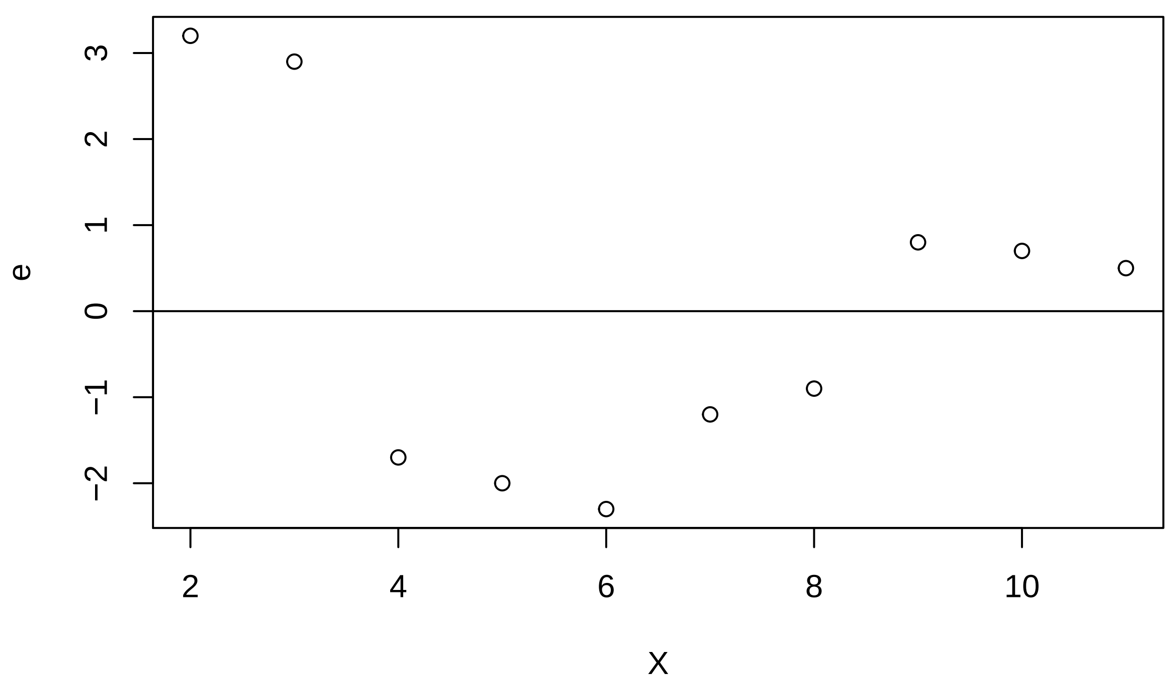
```
X<-c(6,7,8,9,10)
v=sigma^2*sum(((8-mean(X))*(X-mean(X))/sum((X-mean(X))^2)+1/n)^2)
v
```

```
## [1] 0.072
```

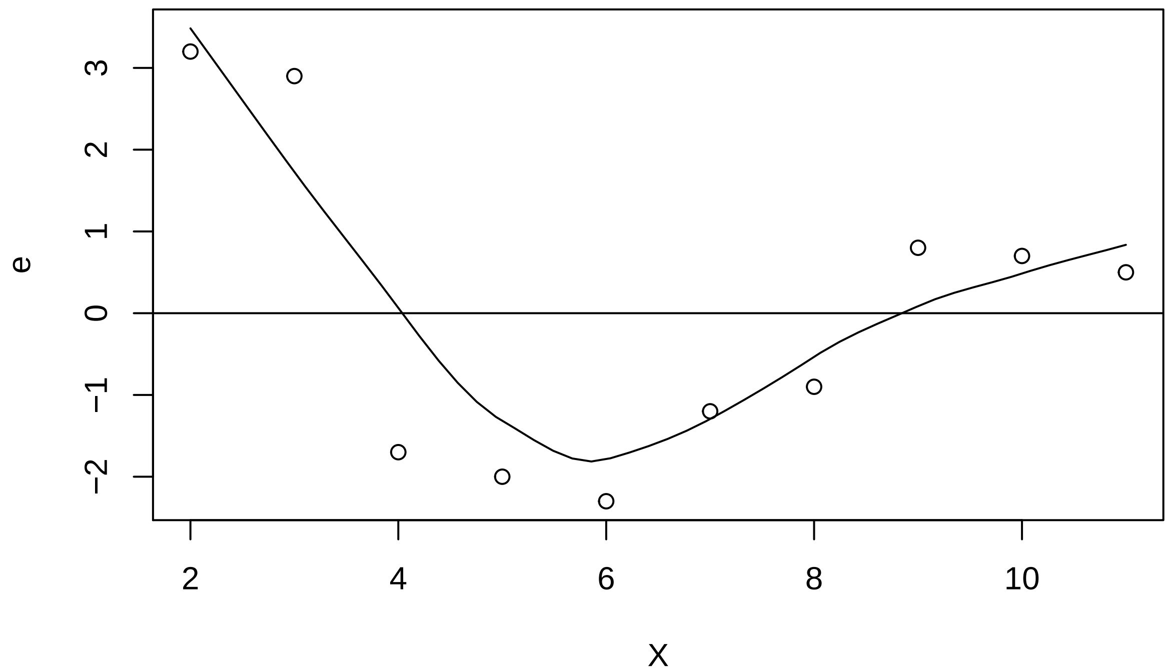
Both observations made the same variance of $\hat{Y}$, so it would have been neither better nor worse.

Prob-5: KNNL 3.9

```
X<-c(2,3,4,5,6,7,8,9,10,11)
e<-c(3.2,2.9,-1.7,-2.0,-2.3,-1.2,-0.9,0.8,0.7,0.5)
plot(X,e)
abline(a=0,b=0)
```




```
scatter.smooth(X,e)
abline(a=0,b=0)
```



The variance of Y_i seems not to be the same, but dependent on X_i, so the variance assumption might be wrong.

We can make some transformations to the model like adding $\beta_2 X_i^2$ term.