R Assignment 3

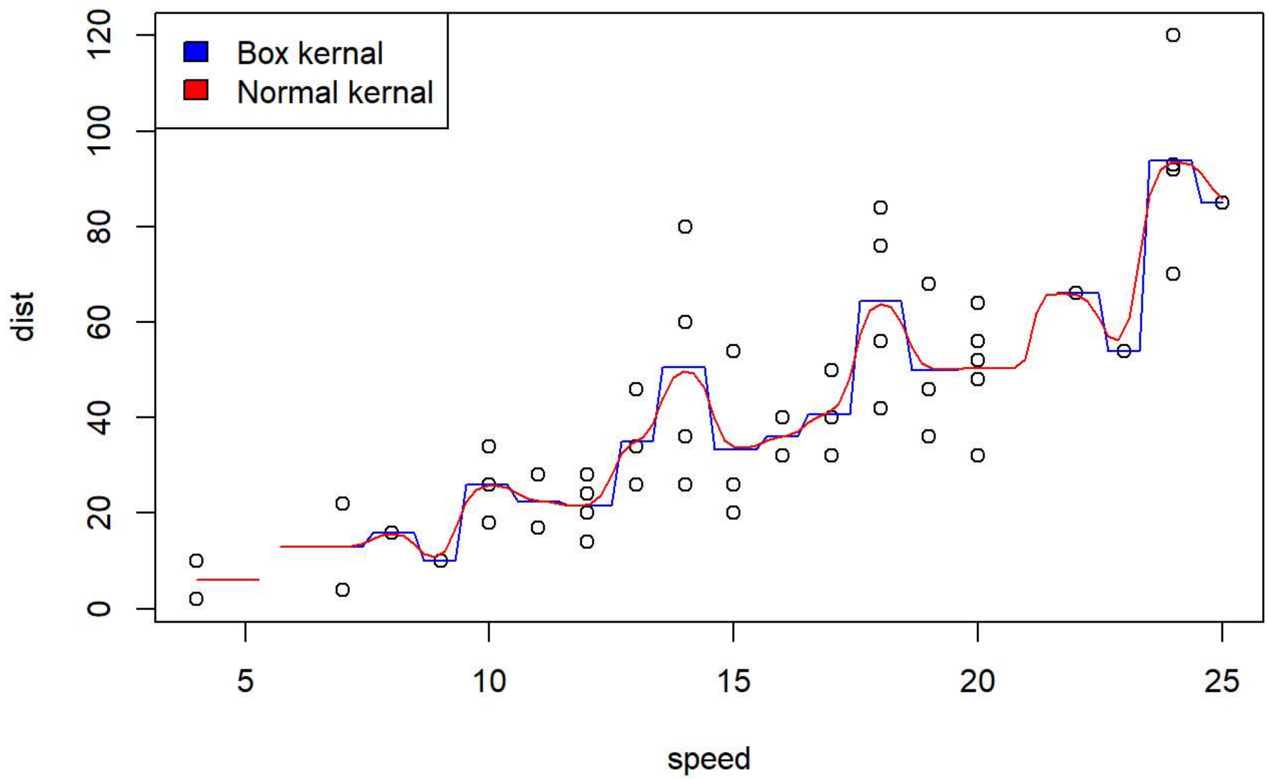
12111603 Tan Zhiheng

2023-11-19

Question 1

(a)

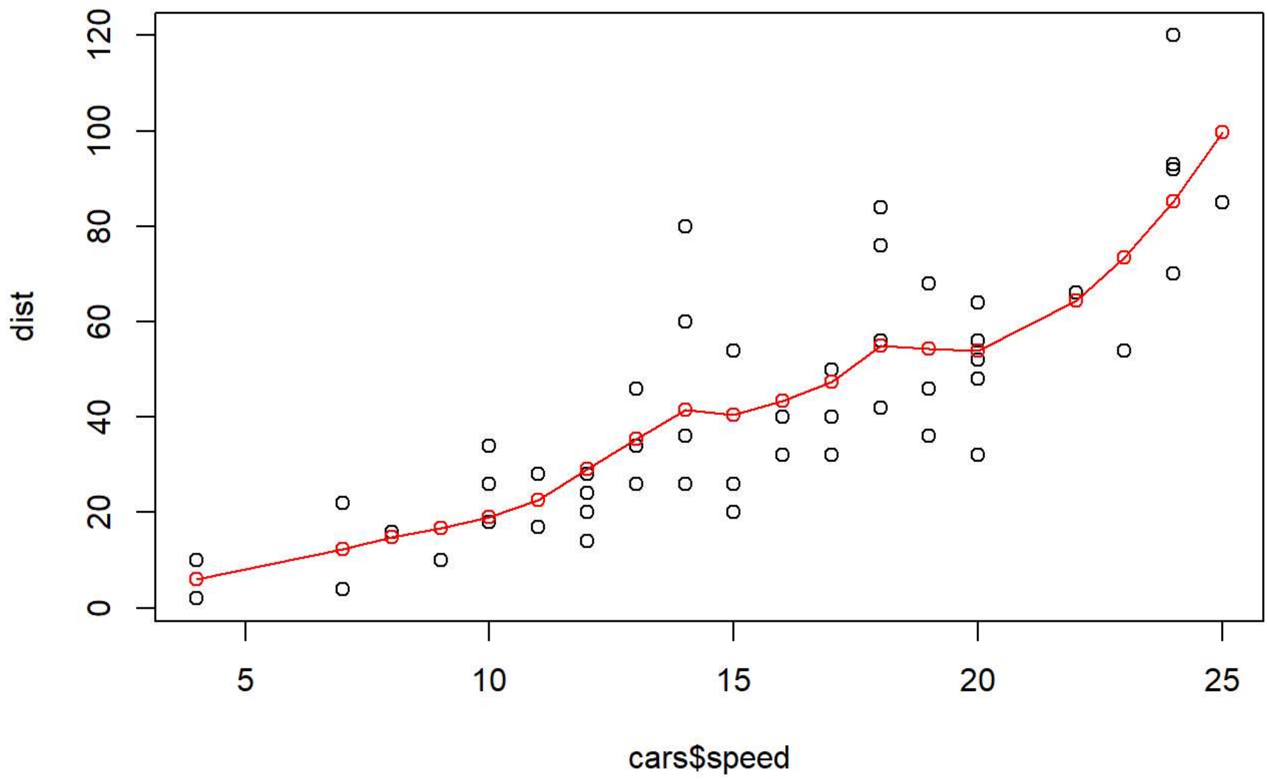
```
data("cars")
data <- na.omit(cars)
attach(data)
plot(speed, dist)
fit1 <- ksmooth(speed,dist,kernel="box",
              bandwidth=1)
fit2 <- ksmooth(speed,dist,kernel="normal",
              bandwidth=1)

lines(fit1,col="blue",lwd=1)
lines(fit2,col="red",lwd=1)
legend("topleft", legend = c("Box kernal", "Normal kernal"),
       fill = c("blue","red"))
```




```
detach(data)
```

(b)

```
attach(cars)
plot(cars$speed, dist)
tt1 <- loess(cars$dist ~ speed,data = cars,span=0.5, family="gaussian")
lines(tt1$x,fitted(tt1),col="red",type = "o")
```



(c)

```
speed <- unique(cars$speed)
yhat1 <- na.omit( unique(fit1$y) )
df1 <- as.data.frame(cbind(speed,yhat1))

cars1 <- merge(cars, df1, by='speed')

sum( (cars1$yhat1-cars1$dist)^2)
```

```
## [1] 6764.783
```

```
sum( (fitted(tt1)-cars$dist)^2 )
```

```
## [1] 9299.113
```

From the MSE, we can see that the Nadaraya-Watson Kernel Regression model fits better.

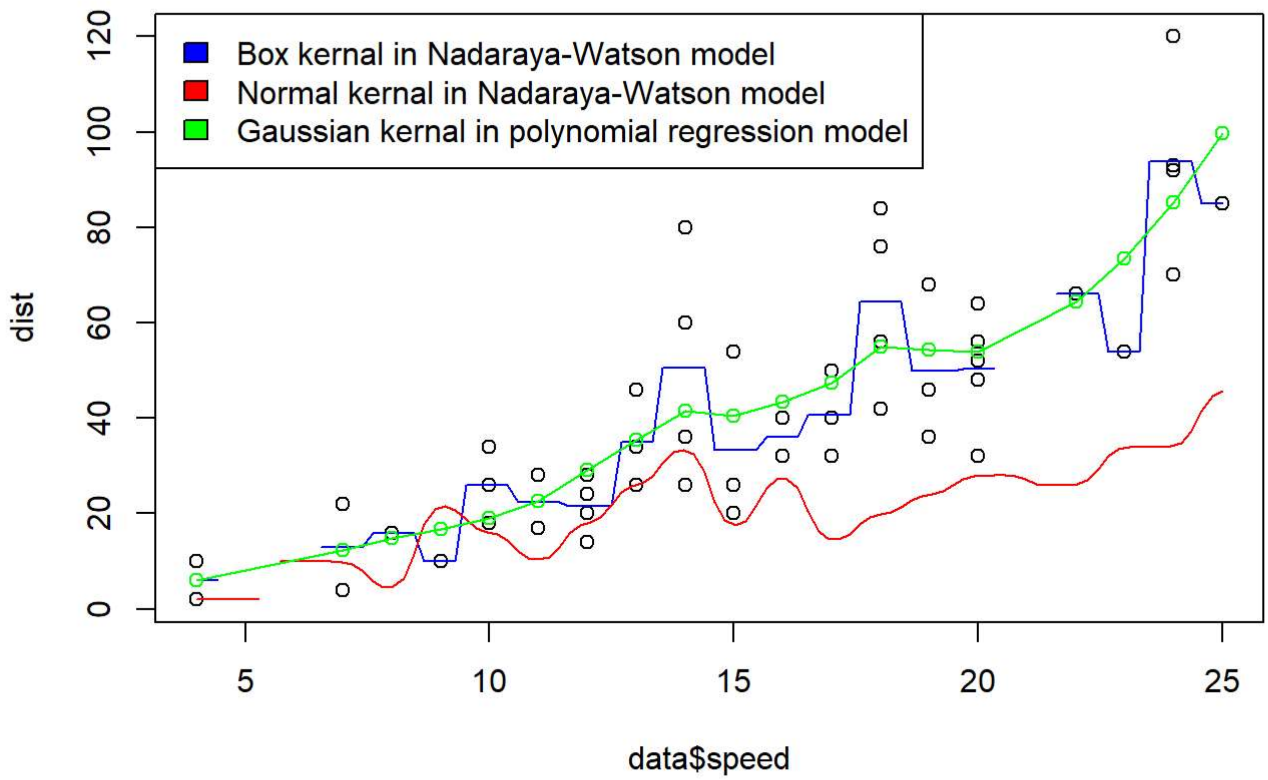
(d)

```
data("cars")
data <- na.omit(cars)
attach(data)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     speed
```

```
## The following objects are masked from cars:
##
##     dist, speed
```

```
plot(data$speed, dist)

lines(ksmooth(data$speed,dist,kernel="box",
              bandwidth=1),col="blue",lwd=1)
lines(ksmooth(speed,dist,kernel="normal",
              bandwidth=1),col="red",lwd=1)
legend("topleft", legend = c("Box kernal in Nadaraya-Watson model",
                             "Normal kernal in Nadaraya-Watson model",
                             "Gaussian kernal in polynomial regression model"),
       fill = c("blue","red", "green"))

lines(tt1$x,fitted(tt1),col="green",type = "o",lwd = 1)
```




```
detach(data)
```

Question 2

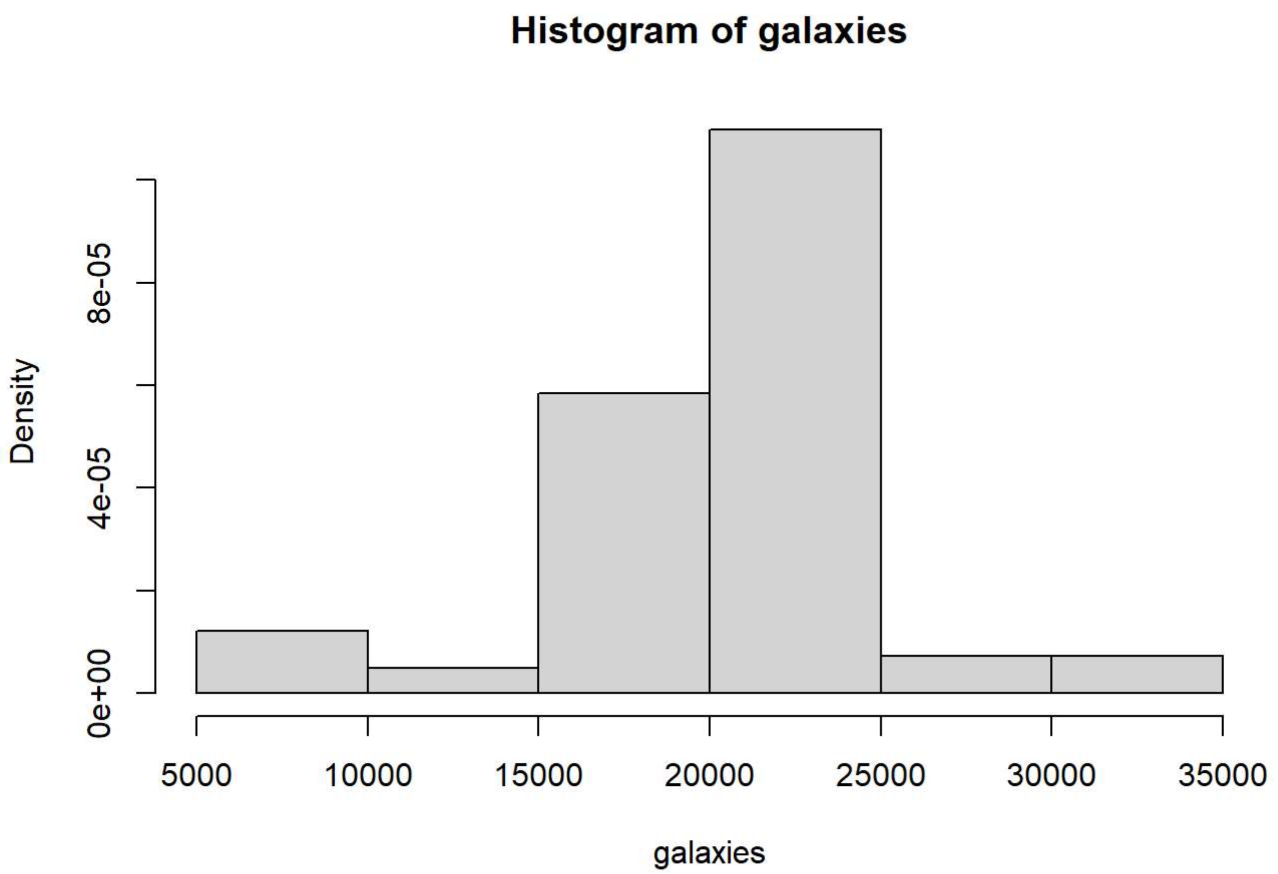
```
library(MASS)
summary(galaxies)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9172   19532   20834   20828   23133   34279
```

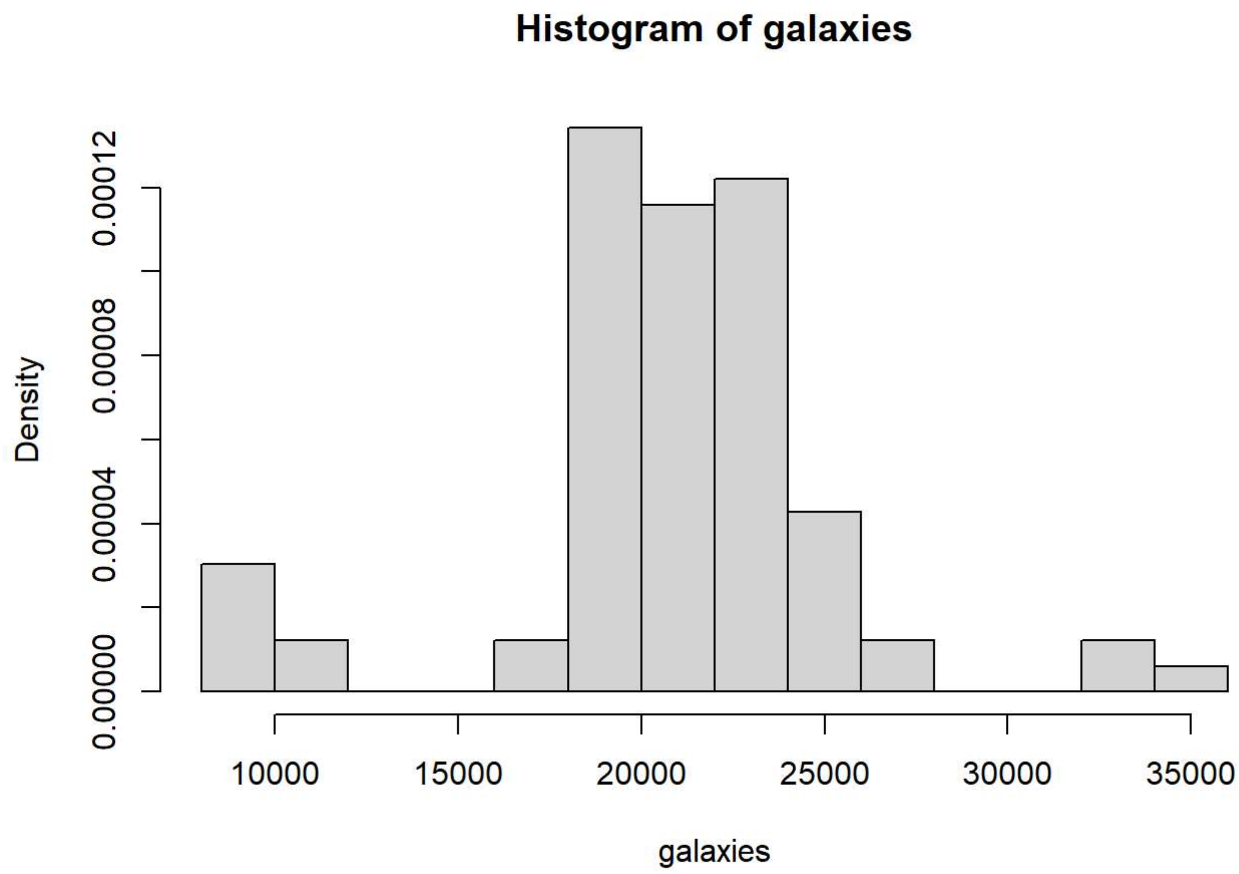
```
data(galaxies)
n <- length(galaxies)
s <- sd(galaxies)
hstars <- 3.491*s*n^{-1/3} # the best bandwidth
iqr <- IQR(galaxies)
```

The best bandwidth is 3667.1992299, then we perform histogram smoothing.

```
nobreaks <- (max(galaxies)-min(galaxies))/hstars
hist(galaxies,breaks=round(nobreaks),probability=TRUE)
```

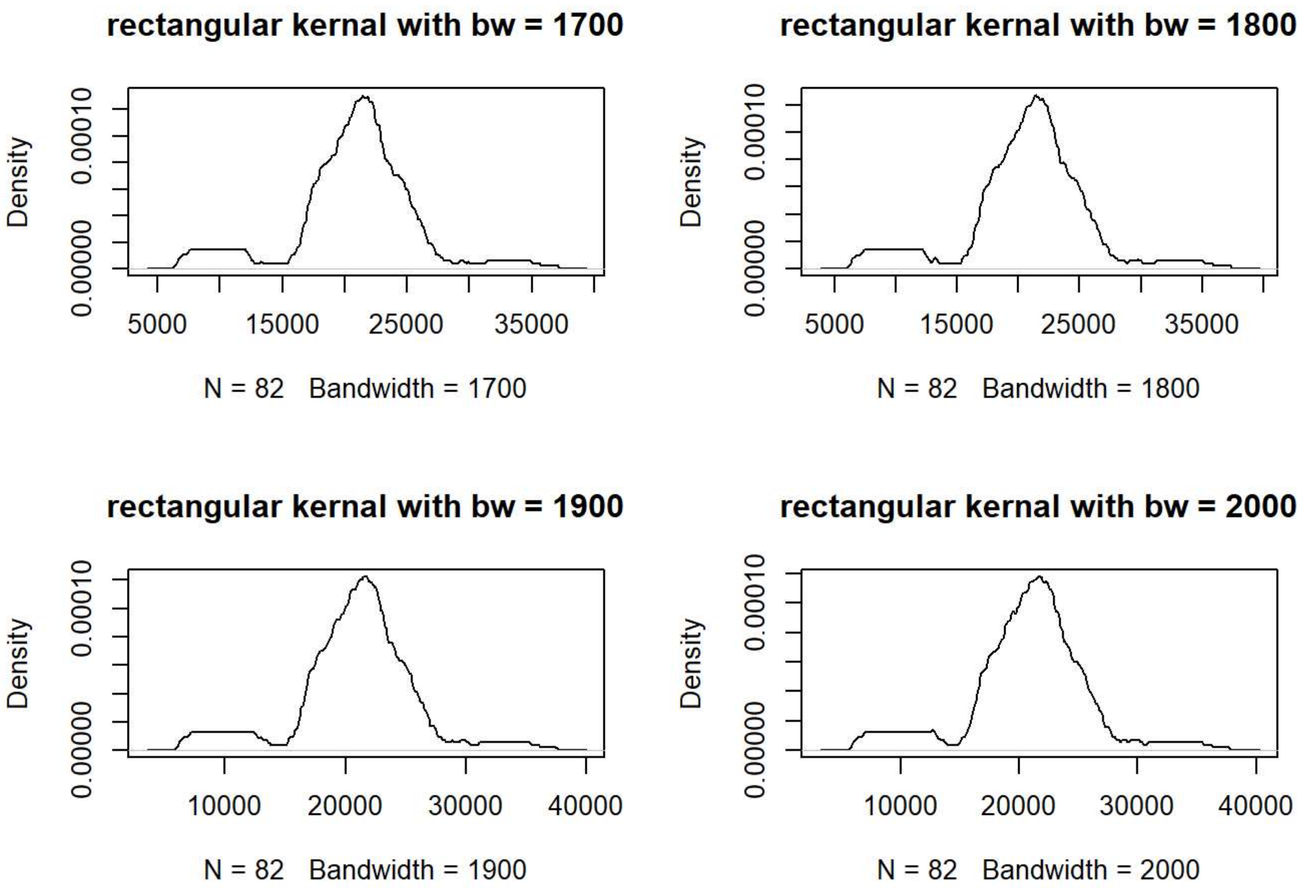



```
hstariqr <- 2.6*iqr*n^{-1/3}
nobreaks2 <- (max(galaxies)-min(galaxies))/hstariqr
hist(galaxies,breaks=round(nobreaks2),probability=TRUE)
```

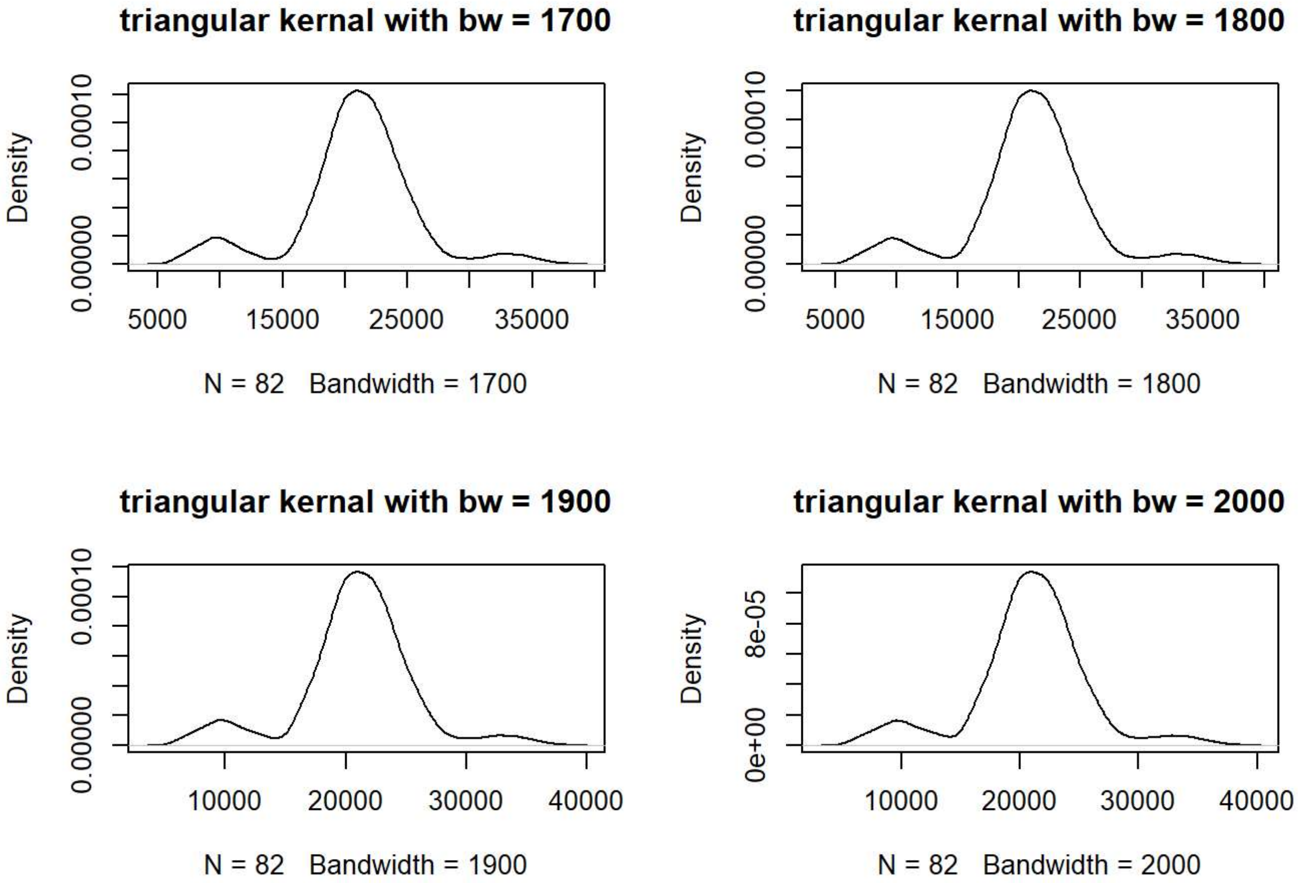


Next, we perform density function estimation with uniform, triangular, epanechnikov and gaussian kernal.

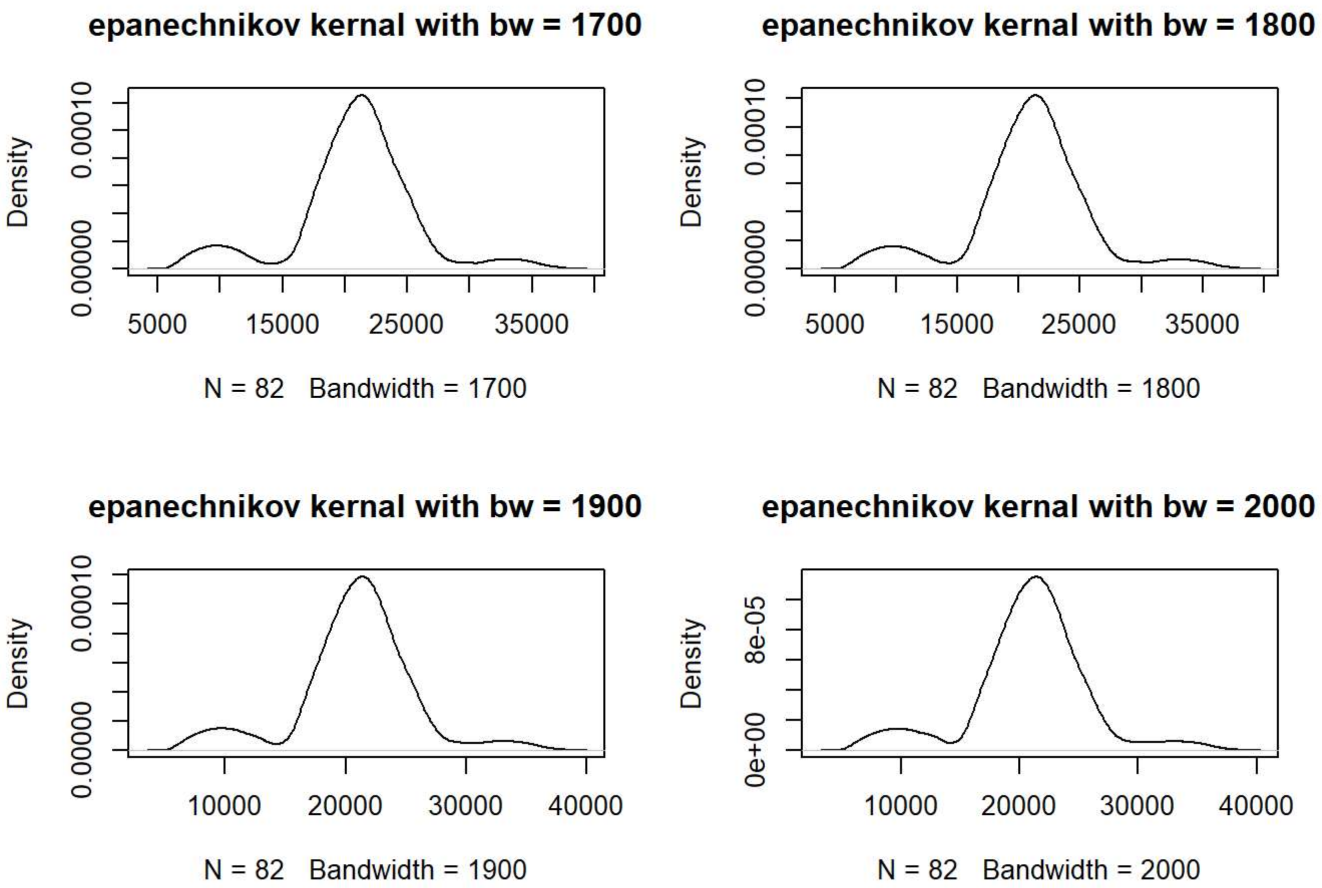
```
par(mfrow = c(2,2))
plot(density(galaxies,kernel="rectangular",bw=1700),
     main="rectangular kernal with bw = 1700")
plot(density(galaxies,kernel="rectangular",bw=1800),
     main="rectangular kernal with bw = 1800")
plot(density(galaxies,kernel="rectangular",bw=1900),
     main="rectangular kernal with bw = 1900")
plot(density(galaxies,kernel="rectangular",bw=2000),
     main="rectangular kernal with bw = 2000")
```



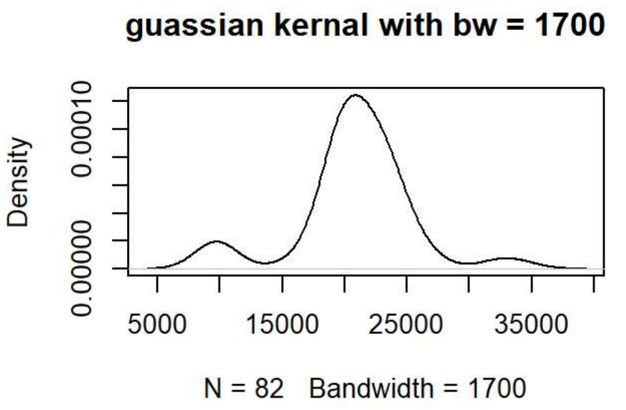

```
par(mfrow = c(2,2))
plot(density(galaxies,kernel="triangular",bw=1700),
     main="triangular kernal with bw = 1700")
plot(density(galaxies,kernel="triangular",bw=1800),
     main="triangular kernal with bw = 1800")
plot(density(galaxies,kernel="triangular",bw=1900),
     main="triangular kernal with bw = 1900")
plot(density(galaxies,kernel="triangular",bw=2000),
     main="triangular kernal with bw = 2000")
```

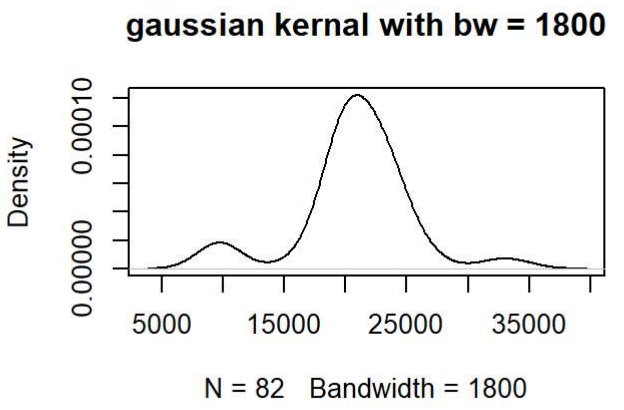


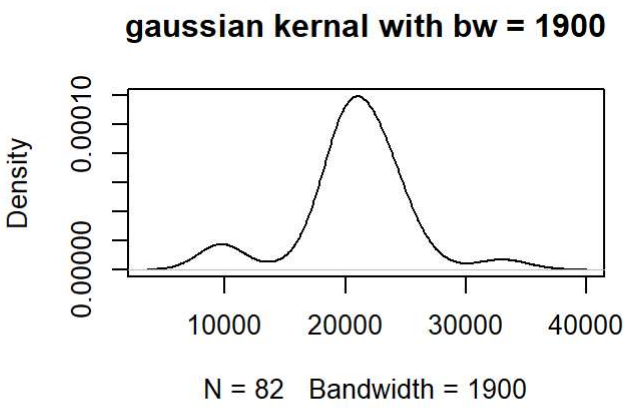

```
par(mfrow = c(2,2))
plot(density(galaxies,kernel="epanechnikov",bw=1700),
     main="epanechnikov kernal with bw = 1700")
plot(density(galaxies,kernel="epanechnikov",bw=1800),
     main="epanechnikov kernal with bw = 1800")
plot(density(galaxies,kernel="epanechnikov",bw=1900),
     main="epanechnikov kernal with bw = 1900")
plot(density(galaxies,kernel="epanechnikov",bw=2000),
     main="epanechnikov kernal with bw = 2000")
```

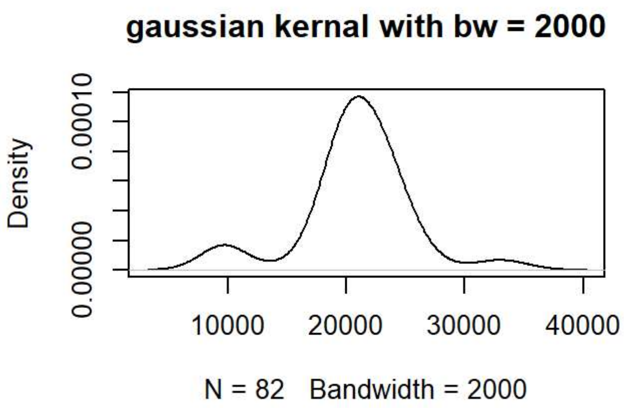



```
par(mfrow = c(2,2))
plot(density(galaxies,kernel="gaussian",bw=1700),
     main="guassian kernal with bw = 1700")
plot(density(galaxies,kernel="gaussian",bw=1800),
     main="gaussian kernal with bw = 1800")
plot(density(galaxies,kernel="gaussian",bw=1900),
     main="gaussian kernal with bw = 1900")
plot(density(galaxies,kernel="gaussian",bw=2000),
     main="gaussian kernal with bw = 2000")
```









Question 3

(a)

```
library(HSAUR3)
```

```
## Warning: 程辑包'HSAUR3'是用R版本4.3.2 来建造的
```

```
## 载入需要的程辑包：tools
```

```
data(foster)

attach(foster)
aggregate(weight, by = list(motgen,litgen), FUN = mean)
```

Group.1 <fct>	Group.2 <fct>	x <dbl>
A	A	63.68000
B	A	52.40000
I	A	54.12500

Group.1 <fct>	Group.2 <fct>	x <dbl>
J	A	48.96000
A	B	52.32500
B	B	60.64000
I	B	53.92500
J	B	45.90000
A	I	47.10000
B	I	64.36667

1-10 of 16 rows Previous **1** 2 Next

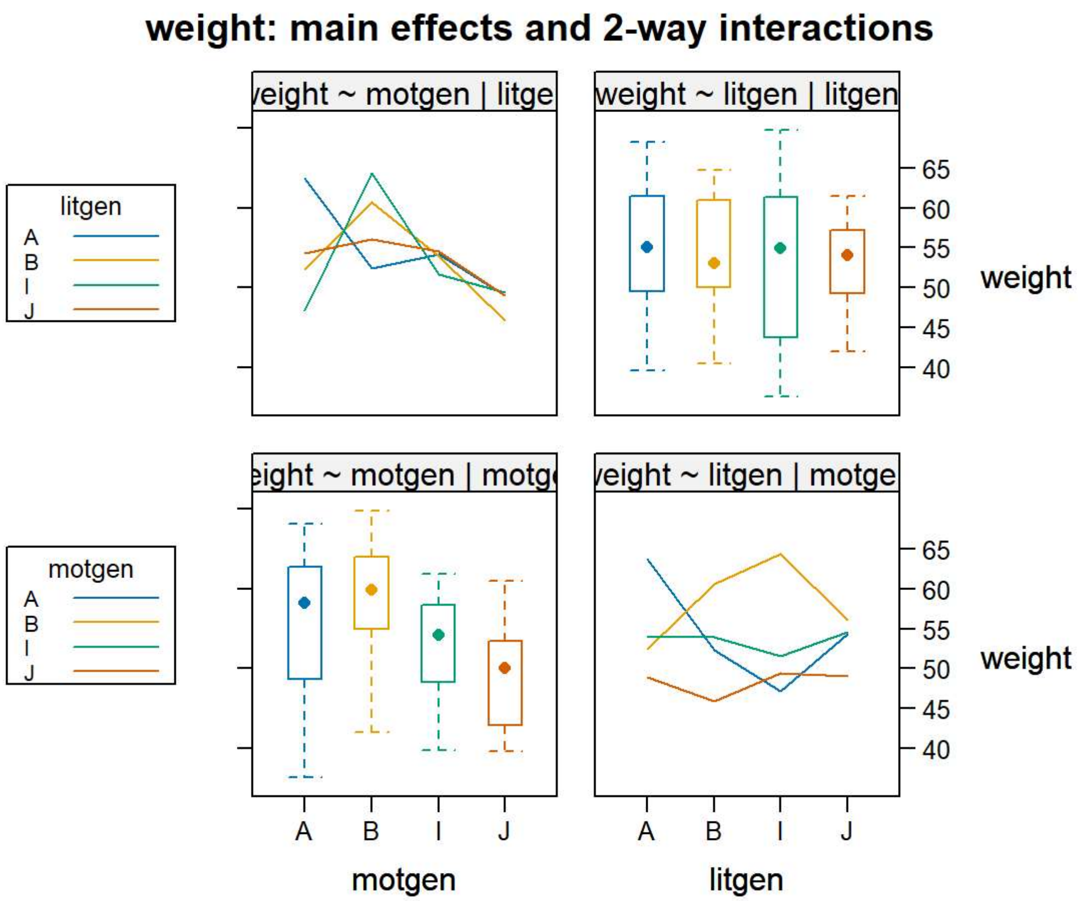
```
aggregate(weight, by = list(motgen,litgen), FUN = sd)
```

Group.1 <fct>	Group.2 <fct>	x <dbl>
A	A	3.273683
B	A	9.374433
I	A	5.321889
J	A	8.760594
A	B	5.533158
B	B	5.647389
I	B	5.114277
J	B	7.636753
A	I	18.103315
B	I	7.124839

1-10 of 16 rows Previous **1** 2 Next

(b)

```
interaction2wt(weight ~ motgen*litgen)
```



See the figures on the diagonal line, lines where have different trends, which means there exists interaction between motgen and litgen.

(c)

```
fit1 <- aov(weight ~ motgen + litgen)
summary(fit1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## motgen       3    772  257.20   4.254 0.00905 **
## litgen       3     64   21.21   0.351 0.78870
## Residuals   54   3265   60.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mean weight differs significantly for different motgen categories, but does not differ significantly for different litgen categories.

```
fit2 <- aov(weight ~ motgen * litgen)
summary(fit2)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)   
## motgen         3  771.6  257.20   4.742 0.00587 **
## litgen         3   63.6   21.21   0.391 0.76000   
## motgen:litgen  9  824.1   91.56   1.688 0.12005   
## Residuals     45 2440.8   54.24                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mean weight differs significantly for different motgen categories, but does not differ significantly for different litgen categories. Moreover, the interaction is not significant at 0.1 significance level.

(d)

In a one-way ANOVA, the dependent variable is assumed to be normally distributed, and have equal variance in each group.

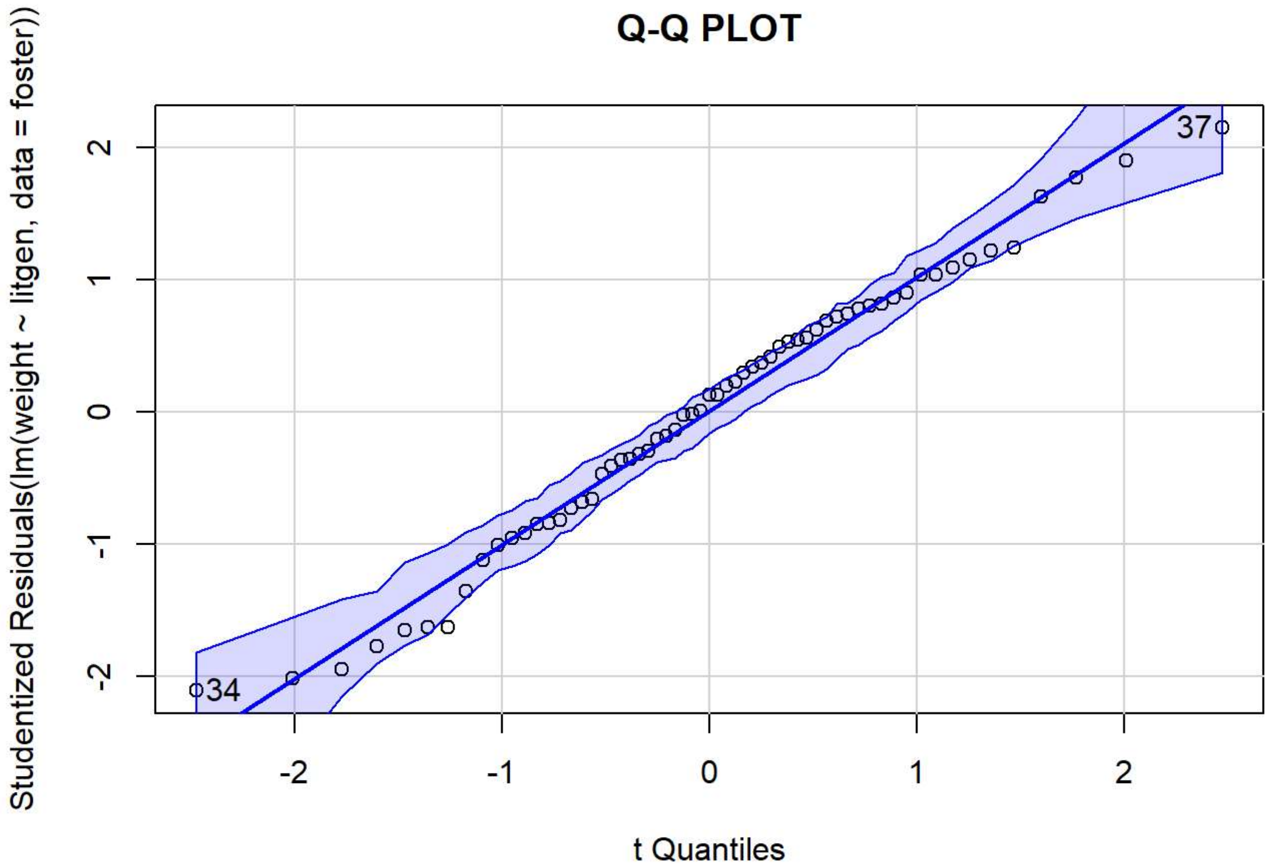
```
library(car)
```

```
## 载入需要的程辑包：carData
```

```
## 
## 载入程辑包：'car'
```

```
## The following objects are masked from 'package:HH':
## 
##     logit, vif
```

```
fit <- aov(weight ~ litgen)
qqPlot(lm(weight ~ litgen, data=foster)
       ,simulate=TRUE,main="Q-Q PLOT",labels=FALSE)
```




```
## [1] 34 37
```

```
bartlett.test(weight ~ litgen, data=foster)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight by litgen
## Bartlett's K-squared = 6.1503, df = 3, p-value = 0.1045
```

```
outlierTest(fit)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 37 2.147241           0.036118           NA
```

From qq plot we can discover that dependent variable obeys approximately Gausssian distribution. Then from the bartlett test, p-value = 0.1045 > 0.05, which means variance in each group do not differ significantly. Thus, the assumptions are satisfied.

(e)

```
library(lmPerm)
```

```
## Warning: 程辑包'lmPerm'是用R版本4.3.2 来建造的
```

```
set.seed(1234)
model <- aovp(weight ~ motgen * litgen, data = foster, perm = "prob")
```

```
## [1] "Settings:  unique SS "
```

```
summary(model)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## motgen         3  671.7  223.91   4.128 0.0114 *
## litgen         3   27.7    9.22   0.170 0.9161
## motgen:litgen  9  824.1   91.56   1.688 0.1201
## Residuals     45 2440.8   54.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
detach(foster)
```

The results derived by permutation test are almost consistent with those from two-way ANOVA. The permutation test tells us motgen is significant at 0.05 significance level while litgen and interaction does not.

Question 4

(a)

```
library(ISLR)
data(Default)
attach(Default)
summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

```
model <- glm(default ~ student + balance + income, family = binomial())
summary(model)
```

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = binomial())
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

- The standard error of coefficients associated with student-Yes is 2.363e-01;
- The standard error of coefficients associated with balance is 2.319e-04;
- The standard error of coefficients associated with income is 8.203e-06.

(b)

```
boot.fn=function(formula,data,indices){
  d=data[indices,]
  fit=glm(formula,data=d, family = binomial())
  return(coef(fit))
}
```

(c)

```
library(boot)
```

```
##
## 载入程辑包：'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:HH':
##
##     logit
```

```
## The following object is masked from 'package:survival':
##
##     aml
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
set.seed(1234)
results=boot(data=Default,statistic=boot.fn, R=500, formula = default ~ student
             + balance + income )
print(results)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Default, statistic = boot.fn, R = 500, formula = default ~
##     student + balance + income)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1* -1.086905e+01 -1.165716e-02 5.127709e-01
## t2* -6.467758e-01 -1.240014e-02 2.441341e-01
## t3*  5.736505e-03  1.221147e-05 2.403813e-04
## t4*  3.033450e-06 -2.598701e-07 8.640181e-06
```

```
detach(Default)
```

The standard errors derived by glm() and our bootstrap function are closed enough to each other. Though our bootstrap function merely replicates 500 times, admittedly it leads to some randomness to some extent, however, its results are numerically consistent with those derived by glm().