评分： _____

SHANGHAI UNIVERSITY

COURSE PAPER

Test 2

Information	student1	student2
Academy	Management	Management
Specialty	Management Science	Information Science
Student ID	18121216	19121255
Name	Lulu Shi	Jingwen Zhang
Course	Machine Learning in Business	
Teacher	Si Zhang | Liyang Xiao	

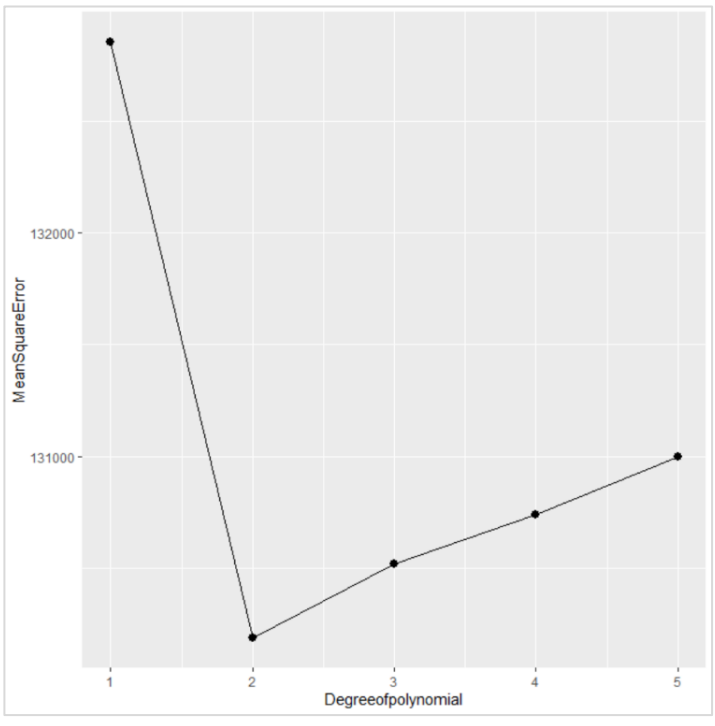
1) Sample 1,000 observations from the original data set without replacement and create a dataset *dat* to record them. In the following steps, we use dat to conduct our analysis.

```
library(AER)
library(boot)
library(ggplot2)
dev.new()
##Q1
data1988=data("CPS1988")
summary(data1988)
set.seed(1)
train=sample(28155,1000)
dat=CPS1988[train,]
```

2) We are interested in using the variable education to predict wage. Perform the leave-one outcross-validation on the linear regression models with degree of polynomial ranging from 1 to5 and plot the result.

```
##Q2
cv.err=rep(0:5)
for(i in 1:5){
  glm.fit=glm(wage~poly(education,i),data=dat)
  cv.err[i]=cv.glm(dat,glm.fit)$delta[1]
}
cv.err
## [1] 132854.7 130187.5 130517.2 130738.4 130998.9      5.0
Degreeofpolynomial=seq(1:5)
MeanSquareError=cv.err[1:5]
photo=data.frame(Degreeofpolynomial,MeanSquareError)
ggplot(photo, aes(x=Degreeofpolynomial, y=MeanSquareError),xlab="Degree of polynomial",ylab="Mean Square Error") + geom_line() + geom_point(size=4, shape=20)
```

plot:



3) Use the bootstrap method to draw a histogram for the correlation between education and wage. (Tips: In R, if we have a vector X and a vector Y, we can use cor(X,Y) to calculate the correlation between X and Y.)

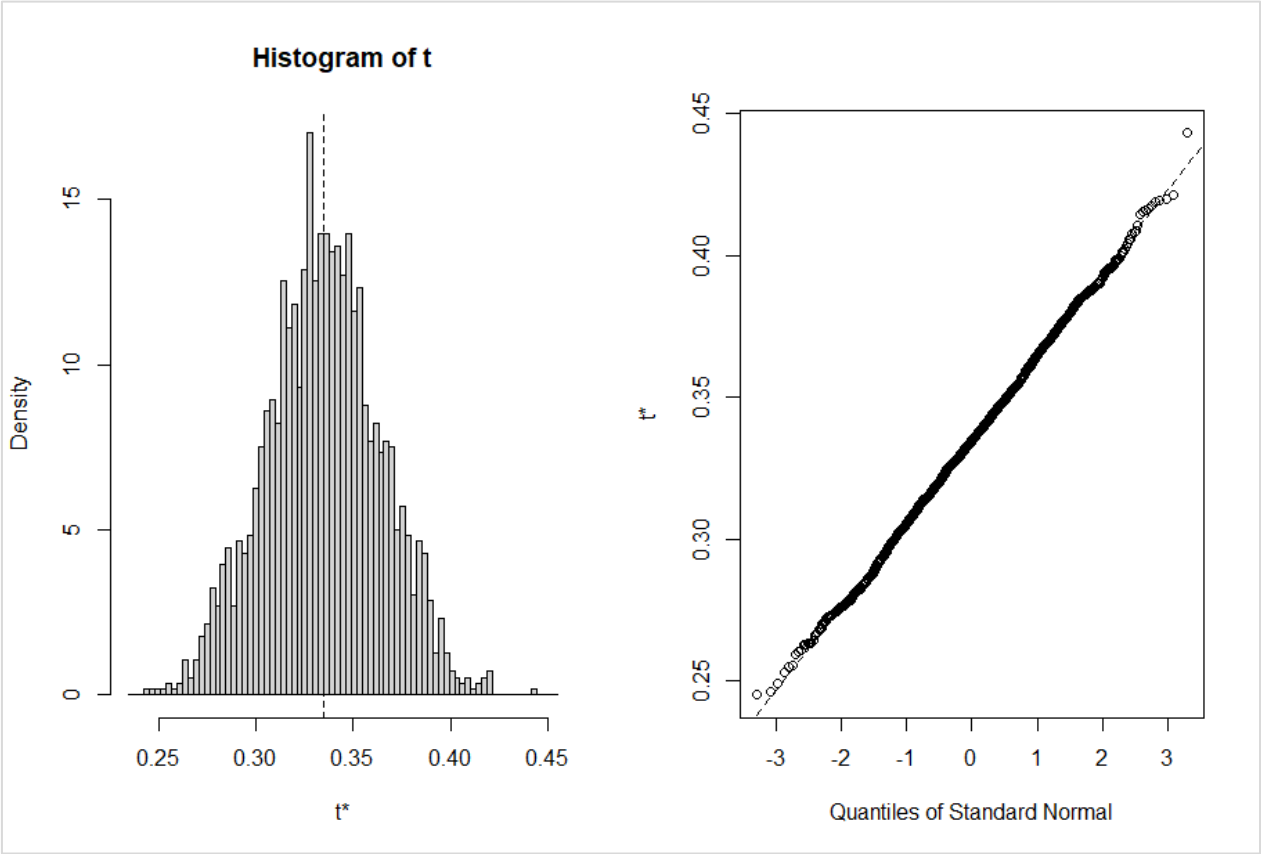
```
##Q3
##Method1:
bootC = function(data,i) {
  cor(dat$education[i],dat$wage[i],
      use = "complete.obs", method = "pearson")
}
set.seed(2)
bootResult=boot(dat,bootC,2000)
plot(bootResult)
bootResult

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
## Call:
## boot(data = dat, statistic = bootC, R = 2000)
##
## Bootstrap Statistics :
##     original       bias    std. error
## t1* 0.3349048 -0.0001264087  0.02937318

## Method2:
cor.fn=function (data ,index){
  education=data$education[index]
  wage=data$wage[index]
  return(cor(education,wage))
}
b=boot(dat,cor.fn,R=2000)
dev.new()
plot(b,freq = F)
```

From the results, we can see that **the bootstrap method** has an estimated value of **0.3349048** for **Pearson's** correlation coefficient, with a standard error of **0.02937318**.

For the generated bootstrap object, we use **plot()** to view the sampling distribution obtained by bootstrap.

The abscissa represents ***the distribution*** of all **correlation coefficients**, the ordinate represents ***the density*** of correlation coefficients, the maximum value is approximately equal to **0.33.**

plot:


Histogram of t
Density
0
5
10
15
0.25
0.30
0.35
0.40
0.45
t*
t*
0.25
0.30
0.35
0.40
0.45
-3
-2
-1
0
1
2
3
Quantiles of Standard Normal