STA137 HW1

Haoming Lei Chunqiu Li, Yirui Li

R Markdown

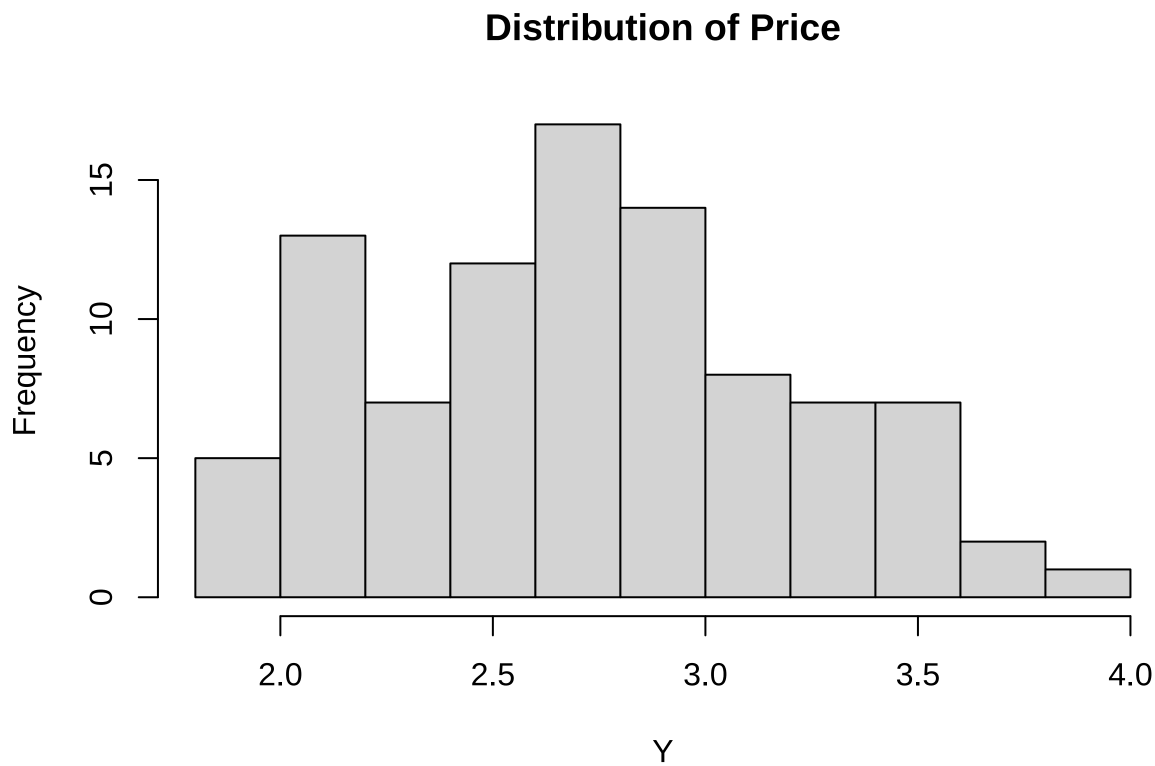
1.

(a)

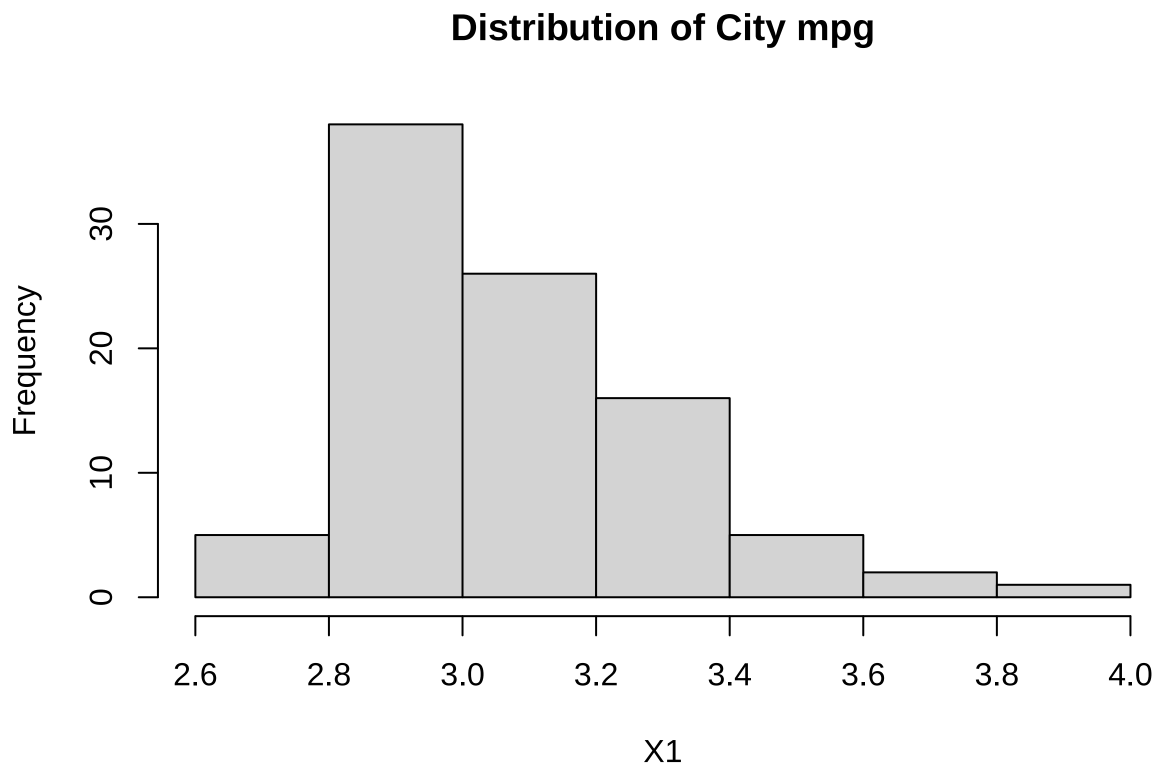
```
#Obtain a histogram for each of the variables
library("readxl")
mydata = read_excel("cars931.xlsx")
head(mydata)
```

```
## # A tibble: 6 x 8
##   Price `City mpg` `Hwy mpg` `Engine size`    HP  Tank Weight `Model/Make`
##   <dbl>      <dbl>     <dbl>         <dbl> <dbl> <dbl>  <dbl> <chr>
## 1  12.9         25        31           1.8   140  13.2   2705 Acura Integra
## 2  29.2         18        25           3.2   200  18     3560 Acura Legend
## 3  25.9         20        26           2.8   172  16.9   3375 Audi 90
## 4  30.8         19        26           2.8   172  21.1   3405 Audi 100
## 5  23.7         22        30           3.5   208  21.1   3640 BMW 535i
## 6  14.2         22        31           2.2   110  16.4   2880 Buick Century
```

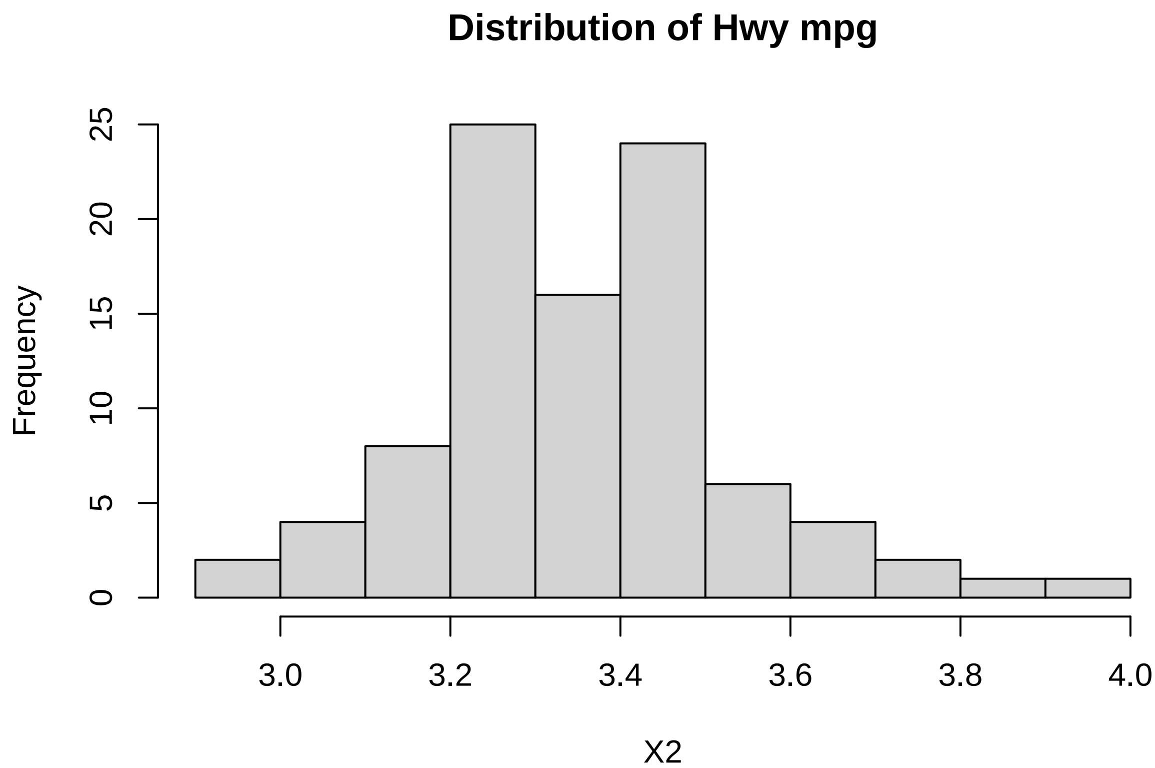
```
Y=log(mydata$Price)
X1=log(mydata$`City mpg`)
X2=log(mydata$`Hwy mpg`)
X3=log(mydata$`Engine size`)
X4=sqrt(mydata$HP)
X5=mydata$Tank
X6=mydata$Weight
hist(Y,main="Distribution of Price")#Bimodal distribution
```



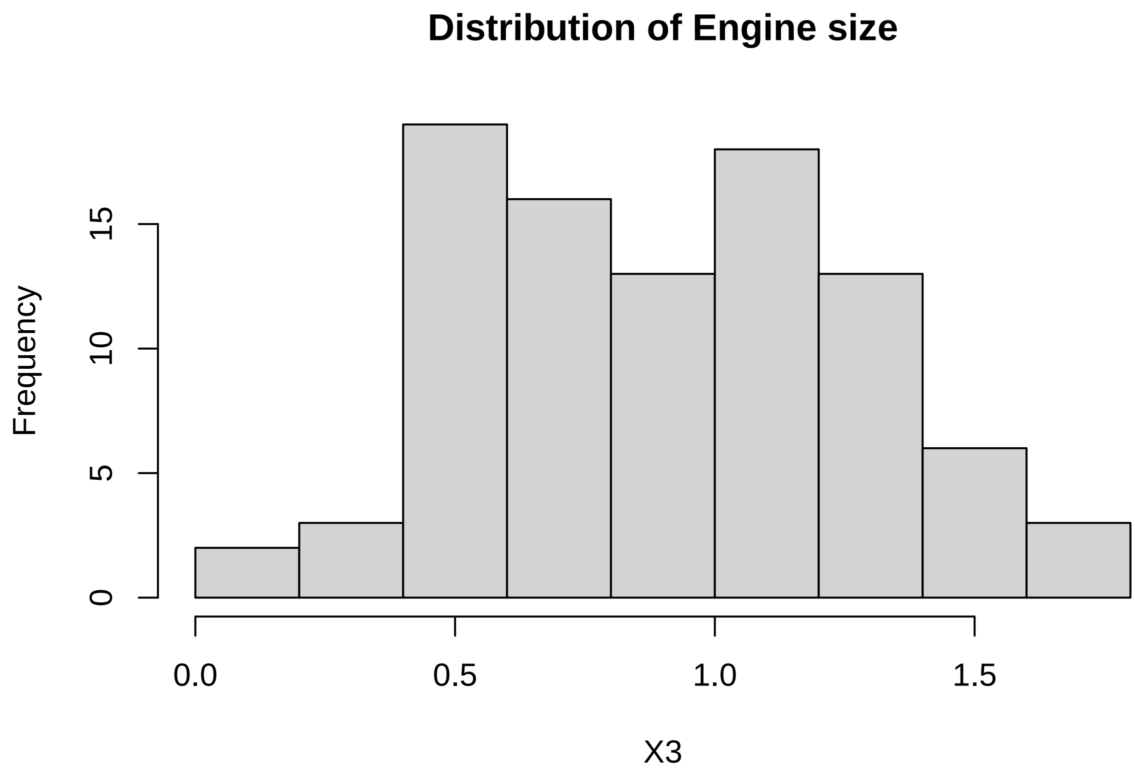

```r
hist(X1,main="Distribution of City mpg")#right-skewed distribution
```



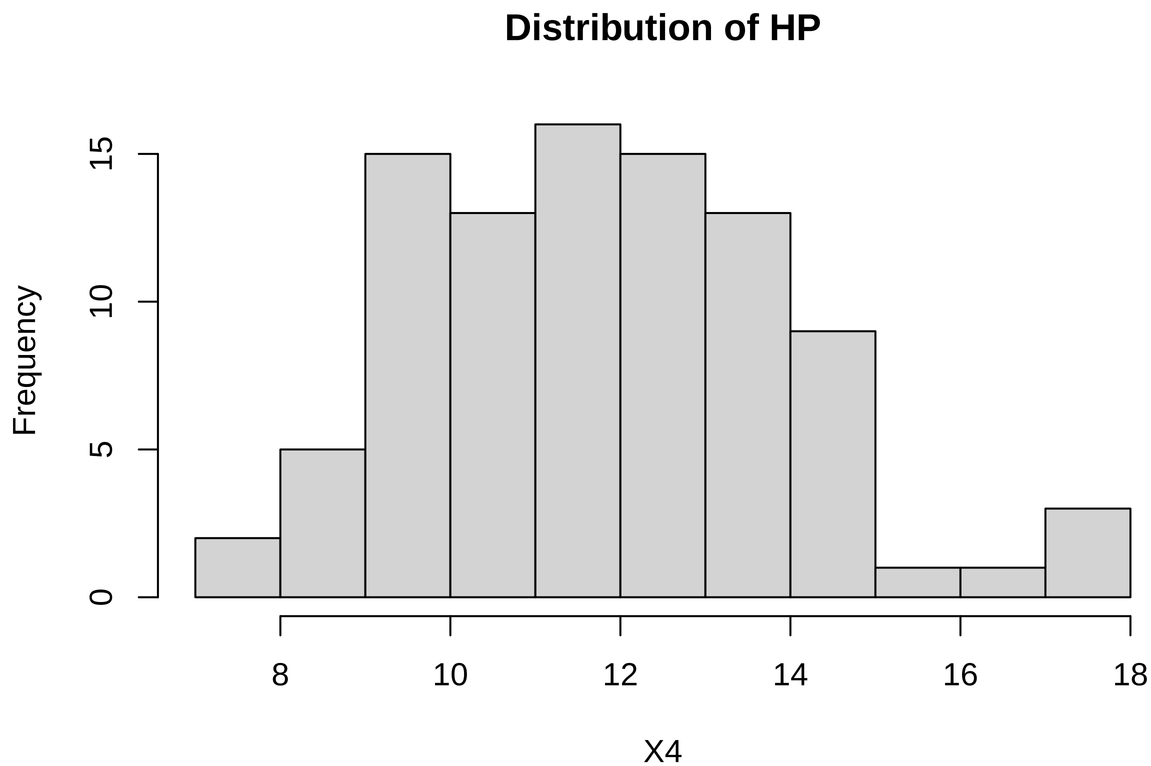

```r
hist(X2,main="Distribution of Hwy mpg")#Bimodal distribution
```



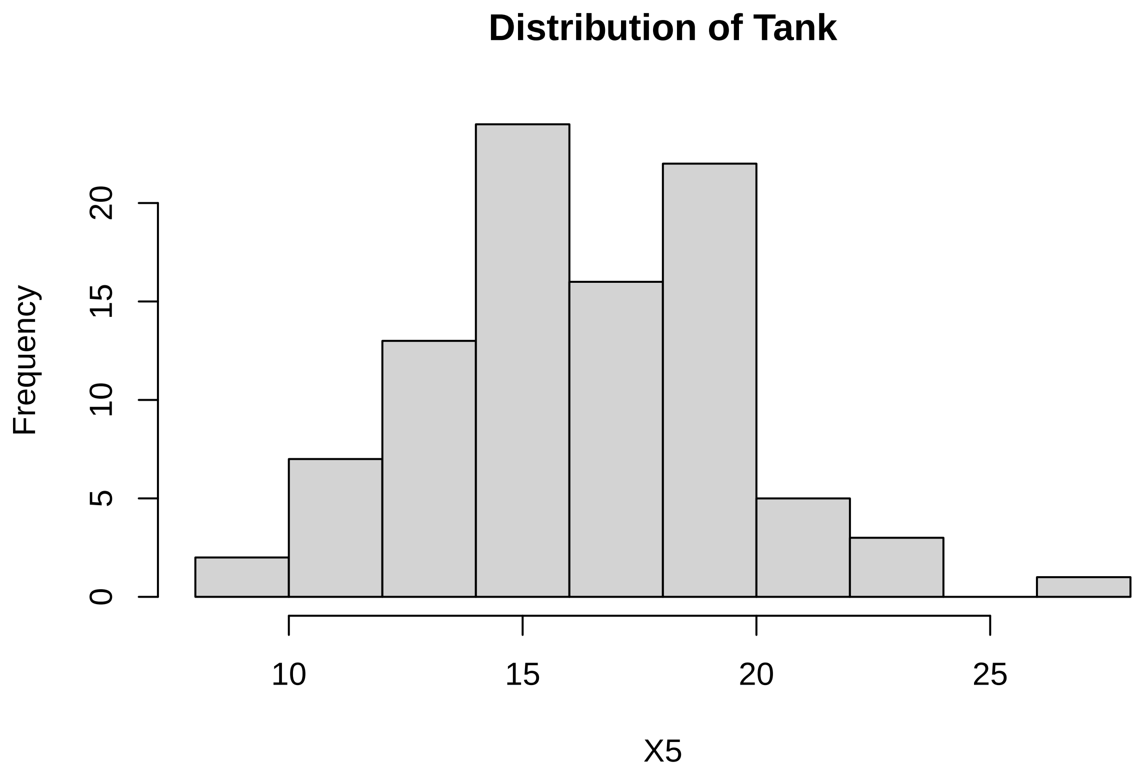

```r
hist(X3,main="Distribution of Engine size")#Plateau distribution
```



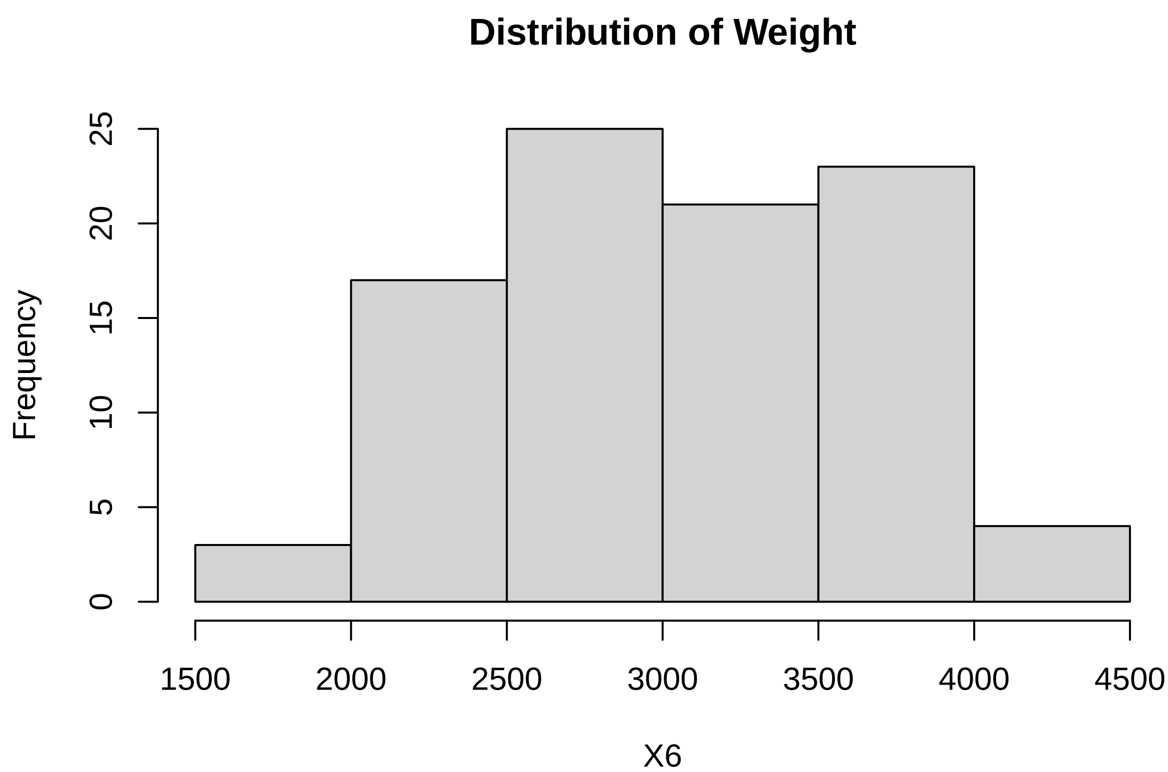

```
hist(X4,main="Distribution of HP")#Edge peak distribution
```




```r
hist(X5,main="Distribution of Tank")#Bimodal distribution
```

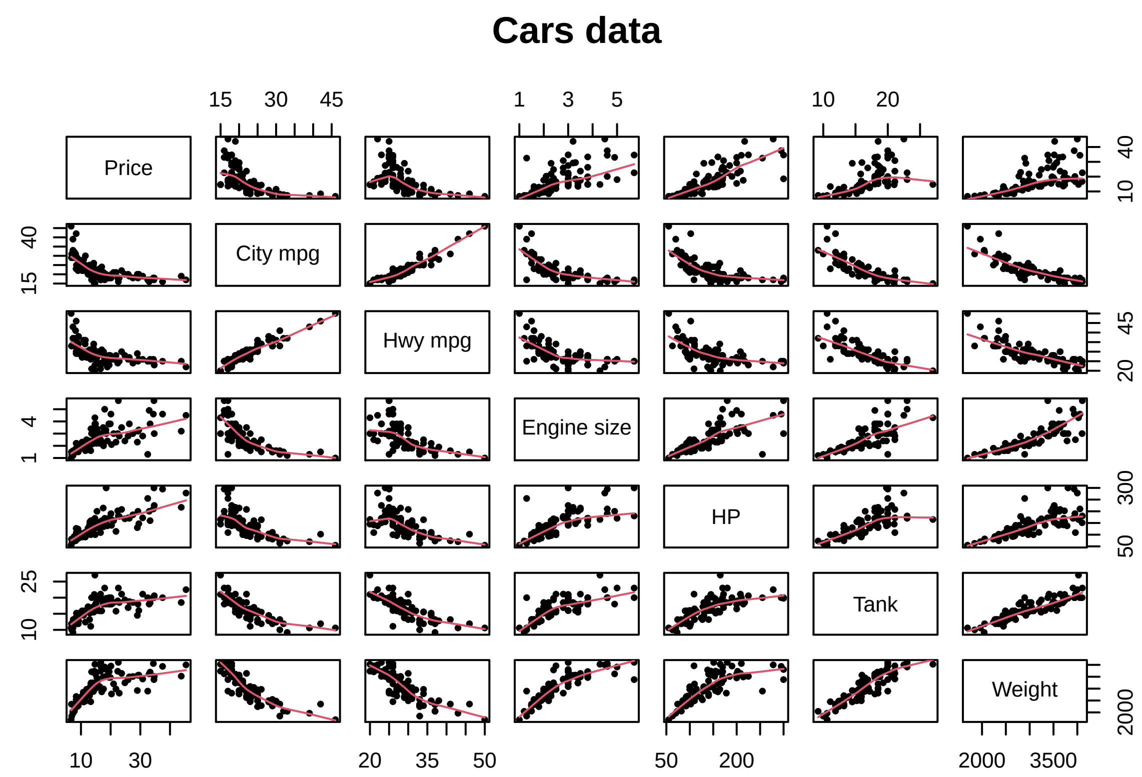



```
hist(X6,main="Distribution of Weight")#Plateau distribution
```



(b)

```r
#Obtain an matrix of the data
pairs(mydata[,c(1,2,3,4,5,6,7)], panel = panel.smooth, main = "Cars data", pch = 19, cex=0.5)
```




```r
#Compute correlation matrix
cor(X1,Y)
```

```
## [1] -0.7653709
```

```r
cor(X2,Y)
```

```
## [1] -0.6771546
```

```r
cor(X3,Y)
```

```
## [1] 0.7334954
```

```r
cor(X4,Y)
```

```
## [1] 0.8421865
```

```r
cor(X5,Y)
```

```
## [1] 0.7309787
```

```r
cor(X6,Y)
```

```
## [1] 0.7710072
```

#What do the plots suggest about the nature of relationship between Y and each of the predictor variables? City mpg and price seem to be negative correlated. Hgy mpg and price seem to be negative correlated. Eng size and price seem to be positive correlated. HP and price seem to be positive correlated. Tank and price seem to be positive correlated. Weight and price seem to be positive correlated.

#Does it seem that there is a problem of multicollinearity? Yes, there is a problem of multicollinearity. For the predictor of the engine size, HP, Tank, Weight, they are closed to 1. Thus indicated a further investigation.

(c)

```r
#compute parameter estimates
model1=lm(Y~X1+X2+X3+X4+X5+X6)
model1
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4           X5
##   1.8118325   -0.7831210    0.4153257   -0.0938671    0.1303312   -0.0005864
##          X6
##   0.0001587
```

```r
#standard errors,R2,Adj_R2
summary(model1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69711 -0.15954 -0.00492  0.12869  0.64015
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.8118325  1.2249902   1.479   0.1428
## X1          -0.7831210  0.4408392  -1.776   0.0792 .
## X2           0.4153257  0.4674927   0.888   0.3768
## X3          -0.0938671  0.1732330  -0.542   0.5893
## X4           0.1303312  0.0207344   6.286  1.3e-08 ***
## X5          -0.0005864  0.0187063  -0.031   0.9751
## X6           0.0001587  0.0001456   1.090   0.2787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2463 on 86 degrees of freedom
## Multiple R-squared:  0.7545, Adjusted R-squared:  0.7374
## F-statistic: 44.06 on 6 and 86 DF,  p-value: < 2.2e-16
```

```
#analysis of variance table
aov1<-anova(model1)
aov1
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## X1         1 12.4472 12.4472 205.2310 < 2.2e-16 ***
## X2         1  0.2483  0.2483   4.0941  0.046139 *
## X3         1  0.5321  0.5321   8.7730  0.003952 **
## X4         1  2.7165  2.7165  44.7895 2.128e-09 ***
## X5         1  0.0165  0.0165   0.2718  0.603484
## X6         1  0.0721  0.0721   1.1882  0.278742
## Residuals 86  5.2159  0.0606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d) We show the summary table of the model below:

```
sum1=summary(model1)$coefficients
sum1
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  1.8118324888 1.2249901620  1.4790588 1.427783e-01
## X1          -0.7831210130 0.4408391636 -1.7764325 7.919814e-02
## X2           0.4153257160 0.4674927031  0.8884111 3.767978e-01
## X3          -0.0938670970 0.1732330119 -0.5418546 5.893199e-01
## X4           0.1303312322 0.0207344476  6.2857345 1.301515e-08
## X5          -0.0005863680 0.0187063109 -0.0313460 9.750662e-01
## X6           0.0001587298 0.0001456187  1.0900373 2.787415e-01
```

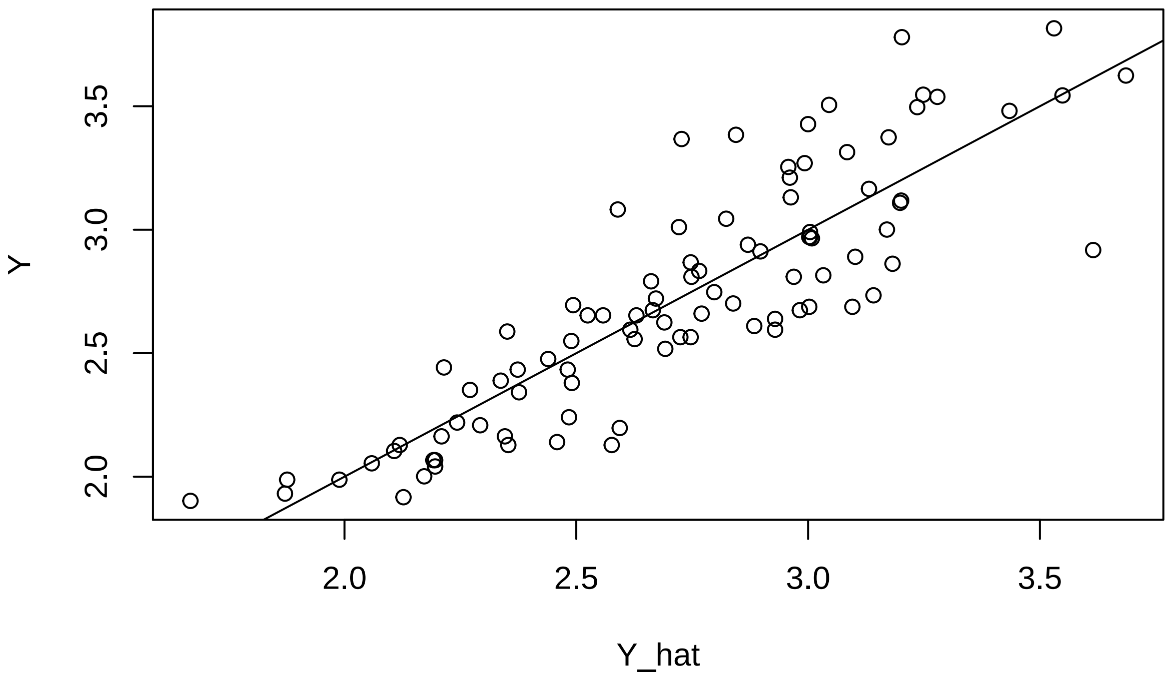
According to the t-statistics and their corresponding p-value table, we consider deleting the predictor of City mpg, which is X1, because the t-value of x1 is -1.7764325, which is the smallest, therefore, its less related to Y.

2.

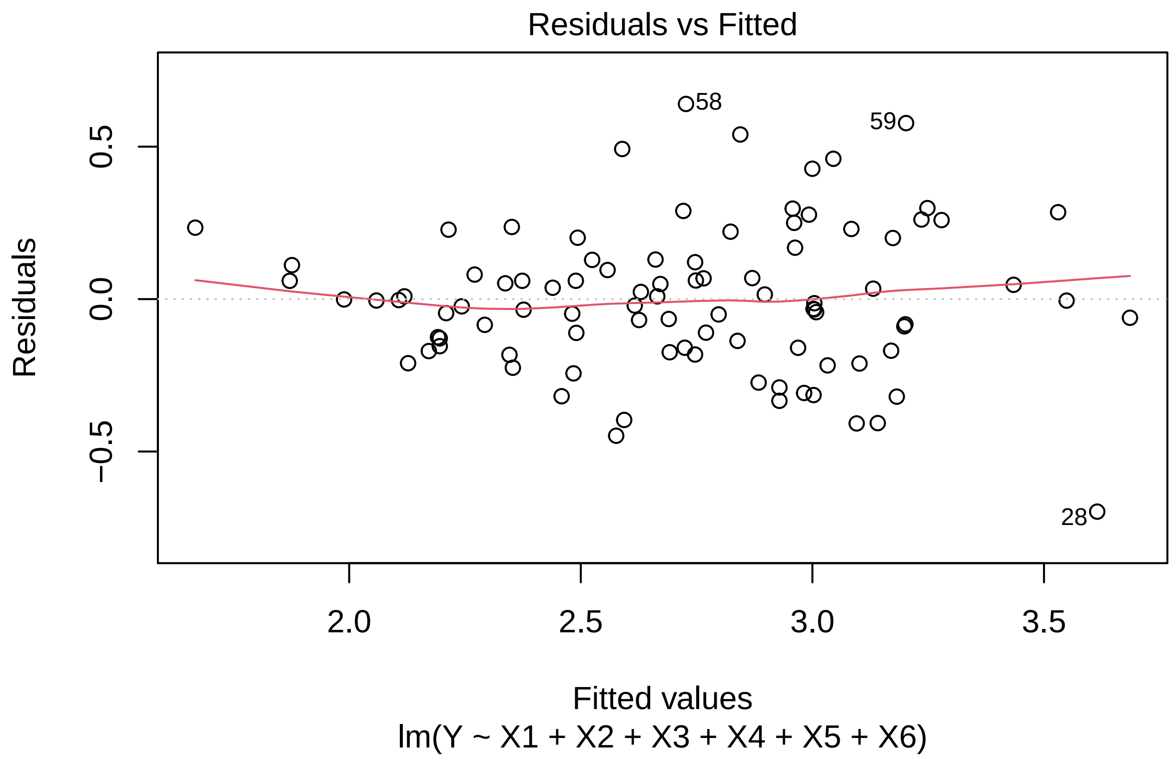
(a)

##Does it seem that the fitted model is reasonable? Do you suspect any nonlinearity? Is the assumption of equal variance of the errors (ie, " i 's) reasonable here? Explain your answers

```
Y_hat<-model1$fitted.values
plot(Y_hat,Y)
abline(lm(Y~Y_hat))
```



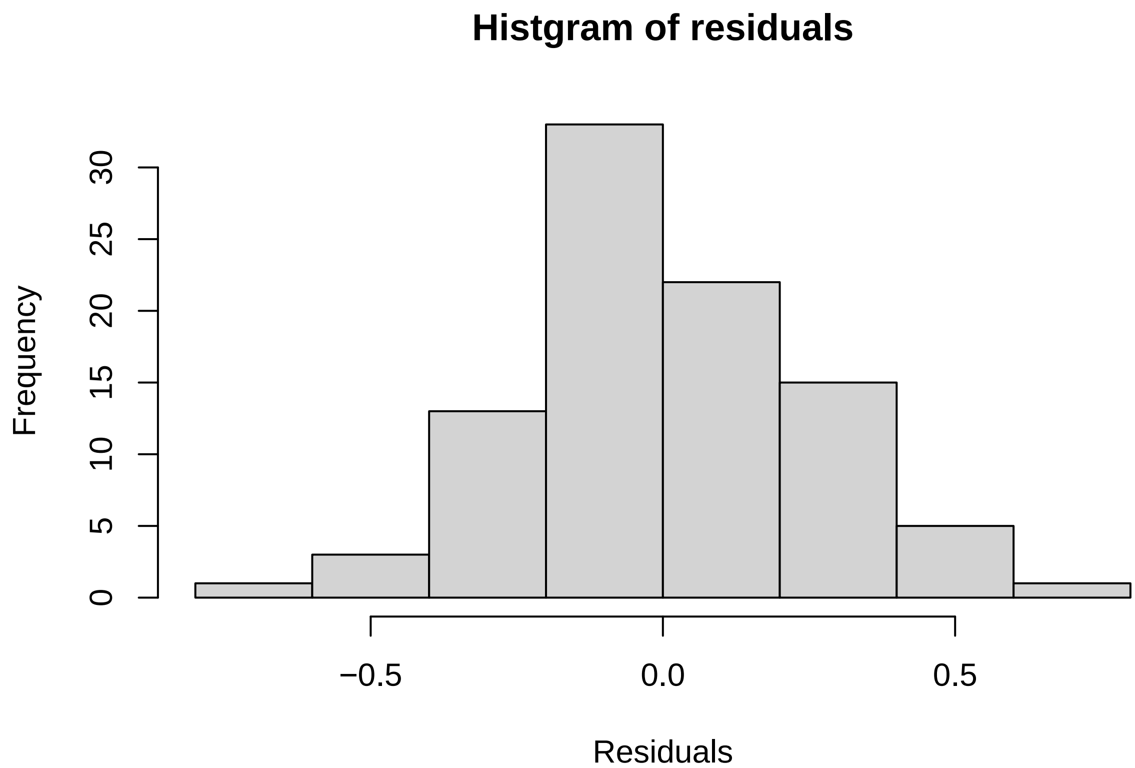

```
plot(model1,which=1)
```



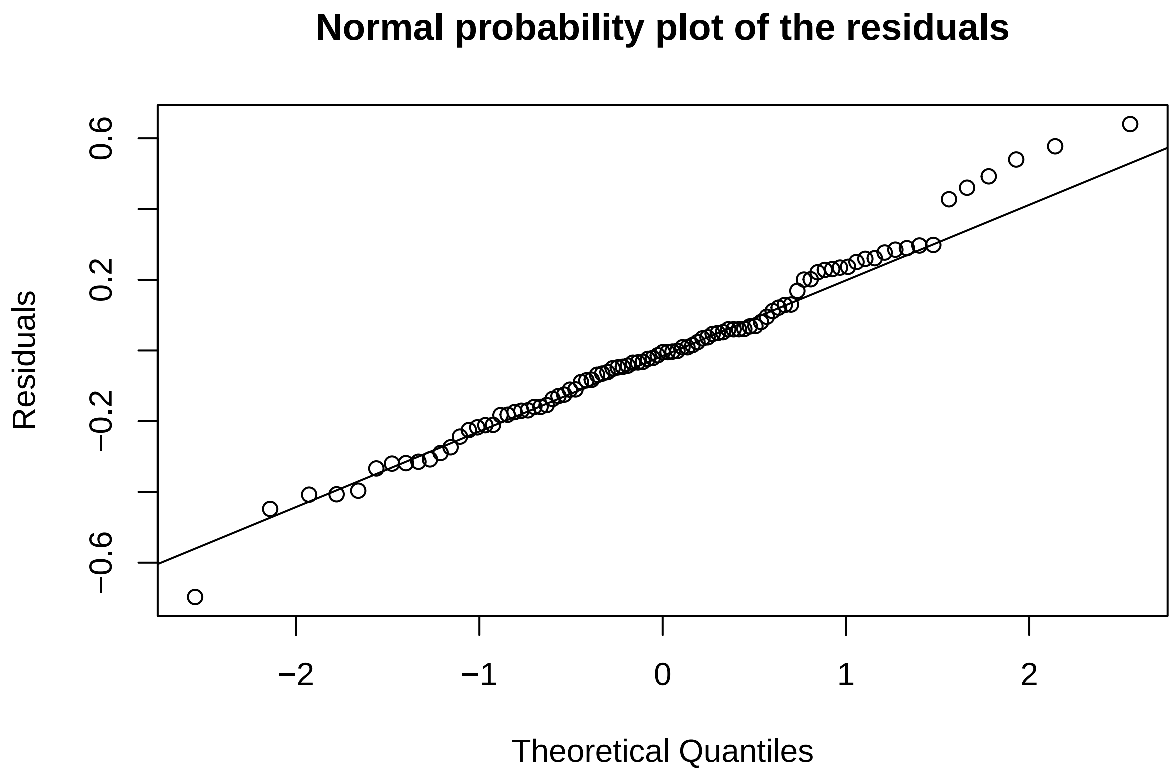
It seems that the fitted model is reasonable, and we do not worrry about nonlinearity. The assumption of equal variance of the errors is reasonable here because even some points are away from the line, the trend shows that points continuously approach the line, most of the points fall approximately along the regression line.

(c) ##Is the assumption of normality of the errors reasonable? Explain. The assumption of normality of the errors is reasonable because in the second graph, we can see that points located closer to the straight line.

```
res<-model1$residuals
hist(res,main="Histgram of residuals",xlab="Residuals")
```




```r
qqnorm(res,ylab="Residuals",main="Normal probability plot of the residuals")
qqline(res)
```



3 (a)

```
null=lm(Y~1)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.3
```

```
#base on AIC
stepAIC(model1, scope=list(upper=model1,lower=~1),diretion="backward",k=2,trace=FALSE)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4)
##
## Coefficients:
## (Intercept)           X1           X4
##      3.0821      -0.6501       0.1393
```

```
#base on BIC
n=93
stepAIC(model1, scope=list(upper=model1,lower=~1),diretion="backward",k=log(n),trace=FALSE)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4)
```

```
##
## Coefficients:
## (Intercept)           X1           X4
##      3.0821      -0.6501       0.1393
```

Then we decide our final model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_4 X_{i4} + \epsilon_i$$

```r
#obtain the parameter estimates
final= lm(Y~X1+X4)
final$coefficients
```

```
## (Intercept)          X1          X4
##   3.0821247  -0.6501018   0.1393028
```

```r
#standard errors, R^2 and adj_R^2
summary(final)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69811 -0.15300 -0.01137  0.14849  0.64441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.08212    0.70764   4.356 3.51e-05 ***
## X1          -0.65010    0.17233  -3.772 0.000289 ***
## X4           0.13930    0.01821   7.649 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2434 on 90 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7434
## F-statistic: 134.3 on 2 and 90 DF,  p-value: < 2.2e-16
```

```r
"standard erro is 0.2434"
```

```
## [1] "standard erro is 0.2434"
```

```r
"R^2 is 0.749, and Adjusted R-squared is 0.7434  "
```

```
## [1] "R^2 is 0.749, and Adjusted R-squared is 0.7434  "
```

(b) ##Compare your result with the model obtained in part (a).

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.3
```

```r
out<-regsubsets(Y~X1+X2+X3+X4+X5+X6,data=mydata,nbest=10)#obtain all the models
sout<-summary(out)
sout
```

```
## Subset selection object
## Call: regsubsets.formula(Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = mydata,
##     nbest = 10)
## 6 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X4     FALSE      FALSE
## X5     FALSE      FALSE
## X6     FALSE      FALSE
## 10 subsets of each size up to 6
## Selection Algorithm: exhaustive
##           X1  X2  X3  X4  X5  X6
## 1  ( 1 )  " " " " " " "*" " " " "
## 1  ( 2 )  " " " " " " " " " " "*"
## 1  ( 3 )  "*" " " " " " " " " " "
## 1  ( 4 )  " " " " "*" " " " " " "
## 1  ( 5 )  " " " " " " " " "*" " "
## 1  ( 6 )  " " "*" " " " " " " " "
## 2  ( 1 )  "*" " " " " "*" " " " "
## 2  ( 2 )  " " " " " " "*" " " "*"
## 2  ( 3 )  " " "*" " " "*" " " " "
## 2  ( 4 )  " " " " " " "*" "*" " "
## 2  ( 5 )  " " " " "*" "*" " " " "
## 2  ( 6 )  "*" " " " " " " " " "*"
## 2  ( 7 )  "*" " " "*" " " " " " "
## 2  ( 8 )  "*" " " " " " " "*" " "
## 2  ( 9 )  " " " " "*" " " " " "*"
## 2  ( 10 ) " " " " " " " " "*" "*"
## 3  ( 1 )  "*" " " " " "*" " " "*"
## 3  ( 2 )  "*" "*" " " "*" " " " "
## 3  ( 3 )  "*" " " " " "*" "*" " "
## 3  ( 4 )  "*" " " "*" "*" " " " "
## 3  ( 5 )  " " "*" " " "*" " " "*"
## 3  ( 6 )  " " " " " " "*" "*" "*"
## 3  ( 7 )  " " " " "*" "*" " " "*"
## 3  ( 8 )  " " "*" "*" "*" " " " "
## 3  ( 9 )  " " "*" " " "*" "*" " "
## 3  ( 10 ) " " " " "*" "*" "*" " "
## 4  ( 1 )  "*" "*" " " "*" " " "*"
## 4  ( 2 )  "*" " " "*" "*" " " "*"
## 4  ( 3 )  "*" " " " " "*" "*" "*"
## 4  ( 4 )  "*" "*" " " "*" "*" " "
## 4  ( 5 )  "*" "*" "*" "*" " " " "
## 4  ( 6 )  "*" " " "*" "*" "*" " "
## 4  ( 7 )  " " "*" " " "*" "*" "*"
## 4  ( 8 )  " " "*" "*" "*" " " "*"
## 4  ( 9 )  " " " " "*" "*" "*" "*"
```

```
## 4  ( 10 ) " " "*" "*" "*" "*" " "
## 5  ( 1 )  "*" "*" "*" "*" " " "*"
## 5  ( 2 )  "*" "*" " " "*" "*" "*"
## 5  ( 3 )  "*" " " "*" "*" "*" "*"
## 5  ( 4 )  "*" "*" "*" "*" "*" " "
## 5  ( 5 )  " " "*" "*" "*" "*" "*"
## 5  ( 6 )  "*" "*" "*" " " "*" "*"
## 6  ( 1 )  "*" "*" "*" "*" "*" "*"
```

```
p<-apply(sout$which,1,sum)
n<-length(Y)
bic=sout$bic
cp=sout$cp

test<-sout$which
for (i in 1:43){
  sh<-test[i,]
  show<-names(sh)[sh][-1]
  show<-paste(show,collapse="+")
  show<-paste("Y=",show," ","Cp :",round(cp[i],digits=5),"BIC :",round(bic[i],digits=5))
  print(show)
}
```

```
## [1] "Y= X4   Cp : 12.85372 BIC : -105.82587"
## [1] "Y= X6   Cp : 53.08272 BIC : -74.8688"
## [1] "Y= X1   Cp : 56.11651 BIC : -72.90394"
## [1] "Y= X3   Cp : 72.85511 BIC : -62.75163"
## [1] "Y= X5   Cp : 74.14639 BIC : -62.01262"
## [1] "Y= X2   Cp : 100.69973 BIC : -47.98873"
## [1] "Y= X1+X4   Cp : 0.94731 BIC : -114.94565"
## [1] "Y= X4+X6   Cp : 3.19726 BIC : -112.59637"
## [1] "Y= X2+X4   Cp : 5.81707 BIC : -109.93364"
## [1] "Y= X4+X5   Cp : 6.50717 BIC : -109.24473"
## [1] "Y= X3+X4   Cp : 8.55894 BIC : -107.22615"
## [1] "Y= X1+X6   Cp : 43.749 BIC : -78.06728"
## [1] "Y= X1+X3   Cp : 46.17912 BIC : -76.35464"
## [1] "Y= X1+X5   Cp : 50.75951 BIC : -73.20989"
## [1] "Y= X3+X6   Cp : 51.94676 BIC : -72.41183"
## [1] "Y= X5+X6   Cp : 52.05002 BIC : -72.34274"
## [1] "Y= X1+X4+X6   Cp : 1.83343 BIC : -111.59845"
## [1] "Y= X1+X2+X4   Cp : 2.54291 BIC : -110.84167"
## [1] "Y= X1+X4+X5   Cp : 2.69521 BIC : -110.68002"
## [1] "Y= X1+X3+X4   Cp : 2.74346 BIC : -110.62887"
## [1] "Y= X2+X4+X6   Cp : 4.23149 BIC : -109.06491"
## [1] "Y= X4+X5+X6   Cp : 4.94686 BIC : -108.3223"
## [1] "Y= X3+X4+X6   Cp : 5.19547 BIC : -108.06561"
## [1] "Y= X2+X3+X4   Cp : 5.80985 BIC : -107.43428"
## [1] "Y= X2+X4+X5   Cp : 6.2244 BIC : -107.01069"
## [1] "Y= X3+X4+X5   Cp : 7.13805 BIC : -106.0839"
## [1] "Y= X1+X2+X4+X6   Cp : 3.29375 BIC : -107.64566"
## [1] "Y= X1+X3+X4+X6   Cp : 3.79399 BIC : -107.1081"
## [1] "Y= X1+X4+X5+X6   Cp : 3.83119 BIC : -107.06825"
## [1] "Y= X1+X2+X4+X5   Cp : 4.20704 BIC : -106.66657"
## [1] "Y= X1+X2+X3+X4   Cp : 4.45996 BIC : -106.39724"
```

```
## [1] "Y= X1+X3+X4+X5   Cp : 4.57447 BIC : -106.27555"
## [1] "Y= X2+X4+X5+X6   Cp : 6.17698 BIC : -104.58914"
## [1] "Y= X2+X3+X4+X6   Cp : 6.21193 BIC : -104.55271"
## [1] "Y= X3+X4+X5+X6   Cp : 6.94684 BIC : -103.78972"
## [1] "Y= X2+X3+X4+X5   Cp : 7.1932 BIC : -103.53536"
## [1] "Y= X1+X2+X3+X4+X6   Cp : 5.00098 BIC : -103.42912"
## [1] "Y= X1+X2+X4+X5+X6   Cp : 5.29361 BIC : -103.11322"
## [1] "Y= X1+X3+X4+X5+X6   Cp : 5.78927 BIC : -102.58055"
## [1] "Y= X1+X2+X3+X4+X5   Cp : 6.18818 BIC : -102.15408"
## [1] "Y= X2+X3+X4+X5+X6   Cp : 8.15571 BIC : -100.07873"
## [1] "Y= X1+X2+X3+X5+X6   Cp : 44.51046 BIC : -68.27229"
## [1] "Y= X1+X2+X3+X4+X5+X6   Cp : 7 BIC : -98.89758"
```

```
summary(show)
```

```
##    Length     Class      Mode
##         1 character character
```

```
newmodel=lm(Y~X1+X4)
summary(newmodel)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69811 -0.15300 -0.01137  0.14849  0.64441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.08212    0.70764   4.356 3.51e-05 ***
## X1          -0.65010    0.17233  -3.772 0.000289 ***
## X4           0.13930    0.01821   7.649 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2434 on 90 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7434
## F-statistic: 134.3 on 2 and 90 DF,  p-value: < 2.2e-16
```

```
"standard erro is 0.2434"
```

```
## [1] "standard erro is 0.2434"
```

```
"R^2 is 0.749, and Adjusted R-squared is 0.7434  "
```

```
## [1] "R^2 is 0.749, and Adjusted R-squared is 0.7434  "
```

#After using the different selection method,the result of part (b) is same as part (a), the best fit model is Y~X1+X4

```r
knitr::opts_chunk$set(echo = TRUE)
#Obtain a histogram for each of the variables
library("readxl")
mydata = read_excel("cars931.xlsx")
head(mydata)
Y=log(mydata$Price)
X1=log(mydata$`City mpg`)
X2=log(mydata$`Hwy mpg`)
X3=log(mydata$`Engine size`)
X4=sqrt(mydata$HP)
X5=mydata$Tank
X6=mydata$Weight
hist(Y,main="Distribution of Price")#Bimodal distribution
hist(X1,main="Distribution of City mpg")#right-skewed distribution
hist(X2,main="Distribution of Hwy mpg")#Bimodal distribution
hist(X3,main="Distribution of Engine size")#Plateau distribution
hist(X4,main="Distribution of HP")#Edge peak distribution
hist(X5,main="Distribution of Tank")#Bimodal distribution
hist(X6,main="Distribution of Weight")#Plateau distribution
#Obtain an matrix of the data
pairs(mydata[,c(1,2,3,4,5,6,7)], panel = panel.smooth, main = "Cars data", pch = 19, cex=0.5)
#Compute correlation matrix
cor(X1,Y)
cor(X2,Y)
cor(X3,Y)
cor(X4,Y)
cor(X5,Y)
cor(X6,Y)
#compute parameter estimates
model1=lm(Y~X1+X2+X3+X4+X5+X6)
model1
#standard errors,R2,Adj_R2
summary(model1)
#analysis of variance table
aov1<-anova(model1)
aov1
sum1=summary(model1)$coefficients
sum1
Y_hat<-model1$fitted.values
plot(Y_hat,Y)
abline(lm(Y~Y_hat))
plot(model1,which=1)
res<-model1$residuals
hist(res,main="Histgram of residuals",xlab="Residuals")
qqnorm(res,ylab="Residuals",main="Normal probability plot of the residuals")
qqline(res)
null=lm(Y~1)
library(MASS)
#base on AIC
stepAIC(model1, scope=list(upper=model1,lower=~1),direstion="backward",k=2,trace=FALSE)
#base on BIC
n=93
stepAIC(model1, scope=list(upper=model1,lower=~1),direstion="backward",k=log(n),trace=FALSE)
```

```r
#obtain the parameter estimates
final= lm(Y~X1+X4)
final$coefficients
#standard errors, R^2 and adj_R^2
summary(final)
"standard erro is 0.2434"
"R^2 is 0.749, and Adjusted R-squared is 0.7434  "
library(leaps)
out<-regsubsets(Y~X1+X2+X3+X4+X5+X6,data=mydata,nbest=10)#obtain all the models
sout<-summary(out)
sout
p<-apply(sout$which,1,sum)
n<-length(Y)
bic=sout$bic
cp=sout$cp

test<-sout$which
for (i in 1:43){
  sh<-test[i,]
  show<-names(sh)[sh][-1]
  show<-paste(show,collapse="+")
  show<-paste("Y=",show," ","Cp :",round(cp[i],digits=5),"BIC :",round(bic[i],digits=5))
  print(show)
}

summary(show)
newmodel=lm(Y~X1+X4)
summary(newmodel)
"standard erro is 0.2434"
"R^2 is 0.749, and Adjusted R-squared is 0.7434  "
```