R Assignment 2

12111603 Tan Zhiheng

2023-10-16

Question 1

(a)

```
## 载入需要的程辑包：MASS
```

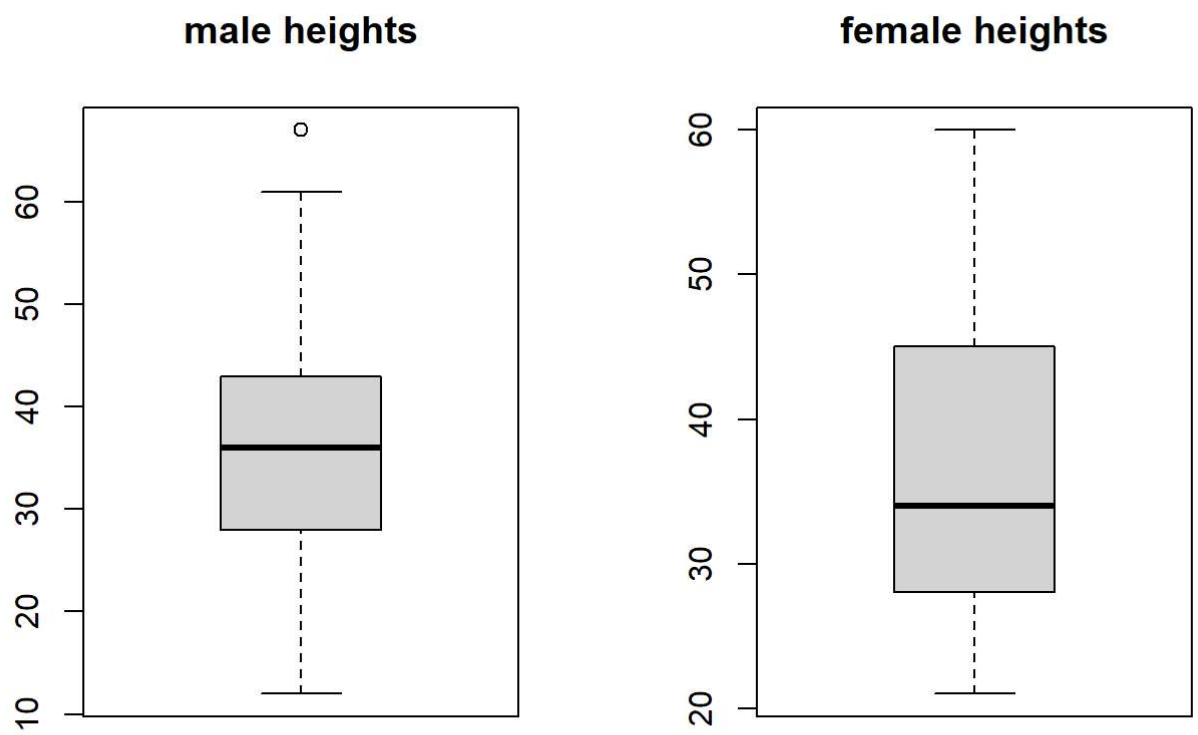
```
## 载入需要的程辑包：HistData
```

```
## 载入需要的程辑包：Hmisc
```

```
##
## 载入程辑包：'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
data("kid.weights")
male <- subset(kid.weights, gender == "M")
female <- subset(kid.weights, gender =="F")


par(mfrow = c(1,2))
boxplot(male$height,main = "male heights")
boxplot(female$height, main = "female heights")
```



With unknown variance

```
t.test(kid.weights$height, mu = 36)
```

```
##
##  One Sample t-test
##
## data:  kid.weights$height
## t = 0.77401, df = 249, p-value = 0.4397
## alternative hypothesis: true mean is not equal to 36
## 95 percent confidence interval:
##  35.19064 37.85736
## sample estimates:
## mean of x
##    36.524
```

```
wilcox.test(kid.weights$height,mu =36)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  kid.weights$height
## V = 12836, p-value = 0.9923
## alternative hypothesis: true location is not equal to 36
```

From the result of Mann Whitney test, we can conclude that $\mu = 36$.

With known variance

```
sigma <- sqrt(var(kid.weights$height))
average <- mean(kid.weights$height)
# Calculating the value of t statistic
abs(sqrt(250)*(average - 36)/sigma)
```

```
## [1] 0.7740114
```

```
# Calculating quantile
qnorm(1-0.05/2)
```

```
## [1] 1.959964
```

Since the test statistic is less than $Z_{\alpha/2}$, which is the upper-α quantile, we can conclude that $\mu = 36$ at 0.05 level of significance.

(b)

```
LR <- function(x,mu0,alpha){
  S <- sd(x); n <- length(x)
  ifelse(abs(sqrt(n)*(mean(x)-mu0)/S) > qt(1-alpha/2,df = n-1),
  "Reject Null Hypothesis","Fail to Reject Null Hypothesis")
}

LR(male$height,36,0.05)
```

```
## [1] "Fail to Reject Null Hypothesis"
```

From LR test, we can draw the conclusion that we fail to reject the null hypothesis $H_0: \mu = 36$ at 0.05 level of significance.

```
t.test(male$height,mu = 36)
```

```
##
##  One Sample t-test
##
## data:  male$height
## t = 1.0187, df = 120, p-value = 0.3104
## alternative hypothesis: true mean is not equal to 36
## 95 percent confidence interval:
##  35.01751 39.06514
## sample estimates:
## mean of x
##  37.04132
```

T test derives the same conclusion as the previous LR test: we cannot reject null hypothesis $H_0: \mu = 36$ at 0.05 level of significance.

(c)

```
wilcox.test(male$height,female$height)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male$height and female$height
## W = 8212, p-value = 0.4758
## alternative hypothesis: true location shift is not equal to 0
```

From Mann Whitney test, we can conclude that the height of the male and female have the same mean value at 0.05 level of significance.

```
t.test(height~gender, data= kid.weights, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  height by gender
## t = -0.7394, df = 248, p-value = 0.4604
## alternative hypothesis: true difference in means between group F and group M is not equal to
0
## 95 percent confidence interval:
##  -3.673137  1.668012
## sample estimates:
## mean in group F mean in group M
##        36.03876        37.04132
```

From t test, we can conclude that the height of the male and female have the same mean value at 0.05 level of significance.

(d)

```
ks.test(male$height,female$height,exact = FALSE, correct = FALSE)
```

```
## Warning in ks.test.default(male$height, female$height, exact = FALSE, correct =
## FALSE): Parameter(s) correct ignored
```

```
## Warning in ks.test.default(male$height, female$height, exact = FALSE, correct =
## FALSE): 并列的时候P-值将近似
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  male$height and female$height
## D = 0.091678, p-value = 0.6703
## alternative hypothesis: two-sided
```

p-value is large, so we cannot reject the null hypothesis and consequently we can conclude that the spread of height for the male and female are the same.

Question 2

(a)

```
Carseats <- read.csv("C:/Users/Lenovo/Desktop/R/Assignment 2/Carseats.csv")
ks.test(Carseats$Sales,"pnorm", mean = mean(Carseats$Sales),sd = sd(Carseats$Sales))
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Carseats$Sales
## D = 0.032533, p-value = 0.791
## alternative hypothesis: two-sided
```

P-value is large, which means we cannot reject null hypothesis, i.e., we can conclude that Sales follows a normal distribution.

(b)

```
fit <- lm(Sales ~ Price + Advertising + Age + Urban, data = Carseats)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age + Urban, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.630 -1.534  0.019  1.516  6.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.992823   0.731610  21.860  < 2e-16 ***
## Price       -0.058047   0.004839 -11.997  < 2e-16 ***
## Advertising  0.123051   0.017130   7.183 3.41e-12 ***
## Age         -0.048865   0.007060  -6.921 1.82e-11 ***
## UrbanYes     0.020186   0.249659   0.081    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.271 on 395 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3531
## F-statistic: 55.44 on 4 and 395 DF,  p-value: < 2.2e-16
```

(c)

- 15.99 is the estimated intercept, denoted as β_0, which means for non-urban area, sales mean is 15.99.
- -0.058 is the estimated coefficient before $Price$, denoted as β_1, which means one unit increase in Price will cause 0.058 decrease in Sales.
- 0.123 is the estimated coefficient before $Advertising$, denoted as β_2, which means one unit increase in Advertising will cause 0.123 increase in Sales.
- -0.049 is the estimated coefficient before Age, denoted as β_3, which means one unit increase in Age will cause 0.049 decrease in Sales.

- When $Urban$ is YES, then the estimated coefficient is 0.020, denoted as β_{41}; Otherwise is 0, denoted as β_{42}, which means the sales mean in urban area is 0.020 larger than that in non-urban area.

(Here we let $x_4 = 1$ when $Urban$ is YES and let $x_4 = 0$ when $Urban$ is NO.)

(d)

For simplicity, denote $Price$ as x_1, $Advertising$ as x_2 and Age as x_3.

$$\hat{y} = \begin{cases} (15.99 + 0.020) - 0.058x_1 + 0.123x_2 - 0.049x_3, & when\ 'Urban' is\ YES; \\ 15.99 - 0.058x_1 + 0.123x_2 - 0.049x_3, & when\ 'Urban' is\ NO. \end{cases}$$

(e)

```
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age + Urban, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.630 -1.534  0.019  1.516  6.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.992823   0.731610  21.860  < 2e-16 ***
## Price       -0.058047   0.004839 -11.997  < 2e-16 ***
## Advertising  0.123051   0.017130   7.183 3.41e-12 ***
## Age         -0.048865   0.007060  -6.921 1.82e-11 ***
## UrbanYes     0.020186   0.249659   0.081    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.271 on 395 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3531
## F-statistic: 55.44 on 4 and 395 DF,  p-value: < 2.2e-16
```

According to the p-value in the table above, we can maintain at $\alpha = 0.001$ significance level that we are able to reject the null hypothesis $H_0: \beta_j = 0$ for $j = 0, 1, 2, 3$, since the p-values of them are less than $\alpha = 0.001$. In other words, for predictor variables Intercept, Price, Advertising and Age we can reject its cofficient $\beta_j = 0$.

(f)

```
fit2 <- lm(Sales ~ Price + Advertising + Age, data = Carseats)
summary(fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6247 -1.5288  0.0148  1.5220  6.2925
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.003472   0.718754  22.266  < 2e-16 ***
## Price       -0.058028   0.004827 -12.022  < 2e-16 ***
## Advertising  0.123106   0.017095   7.201 3.02e-12 ***
## Age         -0.048846   0.007047  -6.931 1.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 396 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3547
## F-statistic:  74.1 on 3 and 396 DF,  p-value: < 2.2e-16
```

Thus the model becomes

$$\hat{y} = 16.003 - 0.058x_1 + 0.123x_2 - 0.049x_3$$

where x_1 denotes $Price$, x_2 denotes $Advertising$ and x_3 denotes Age.

(g)

```
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age + Urban, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.630 -1.534  0.019  1.516  6.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.992823   0.731610  21.860  < 2e-16 ***
## Price       -0.058047   0.004839 -11.997  < 2e-16 ***
## Advertising  0.123051   0.017130   7.183 3.41e-12 ***
## Age         -0.048865   0.007060  -6.921 1.82e-11 ***
## UrbanYes     0.020186   0.249659   0.081    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.271 on 395 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3531
## F-statistic: 55.44 on 4 and 395 DF,  p-value: < 2.2e-16
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6247 -1.5288  0.0148  1.5220  6.2925
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.003472   0.718754  22.266  < 2e-16 ***
## Price       -0.058028   0.004827 -12.022  < 2e-16 ***
## Advertising  0.123106   0.017095   7.201 3.02e-12 ***
## Age         -0.048846   0.007047  -6.931 1.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 396 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3547
## F-statistic:  74.1 on 3 and 396 DF,  p-value: < 2.2e-16
```

```
AIC(fit)
```

```
## [1] 1798.466
```

```
AIC(fit2)
```

```
## [1] 1796.472
```

The adjusted R^2 of model in (b) is about 0.3531, which means only 35.31% variance of data can be explained by the model.

The adjusted R^2 of model in (f) is about 0.3547, which means only 35.47% variance of data can be explained by the model.

(h)

```
confint(fit2,level = 0.95)
```

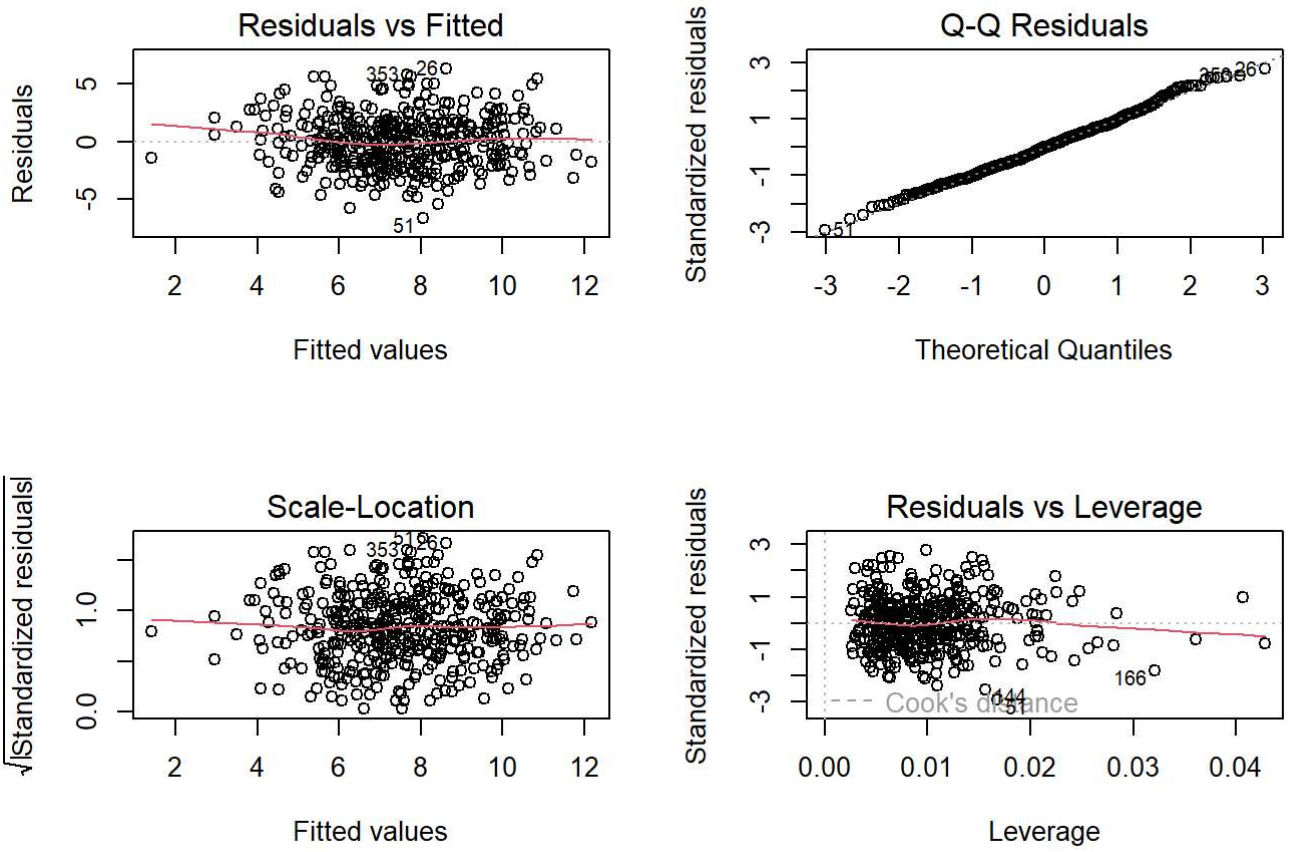
```
##                   2.5 %      97.5 %
## (Intercept) 14.59042068 17.41652325
## Price       -0.06751743 -0.04853857
## Advertising  0.08949838  0.15671410
## Age         -0.06270141 -0.03499112
```

- The 0.95 confidence interval of β_0 is $[14.590, 17.417]$.

- The 0.95 confidence interval of β_1 is $[-0.068, -0.049]$.
- The 0.95 confidence interval of β_2 is $[0.089, 0.157]$.
- The 0.95 confidence interval of β_3 is $[-0.063, -0.035]$.

(i)

```
par(mfrow = c(2,2))
plot(fit2)
```




```
library(car)
```

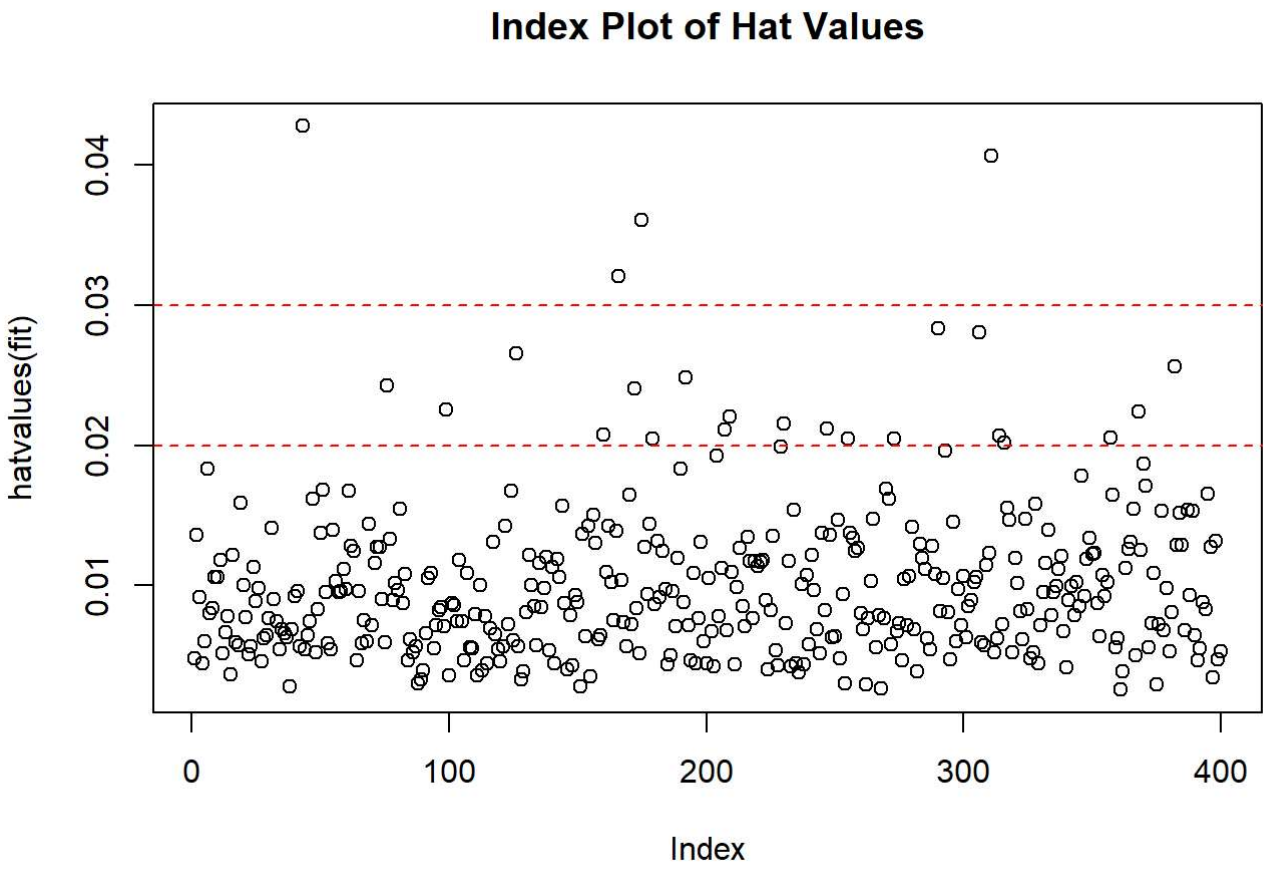
```
## 载入需要的程辑包：carData
```

```
outlierTest(fit2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 51 -2.974015          0.0031196           NA
```

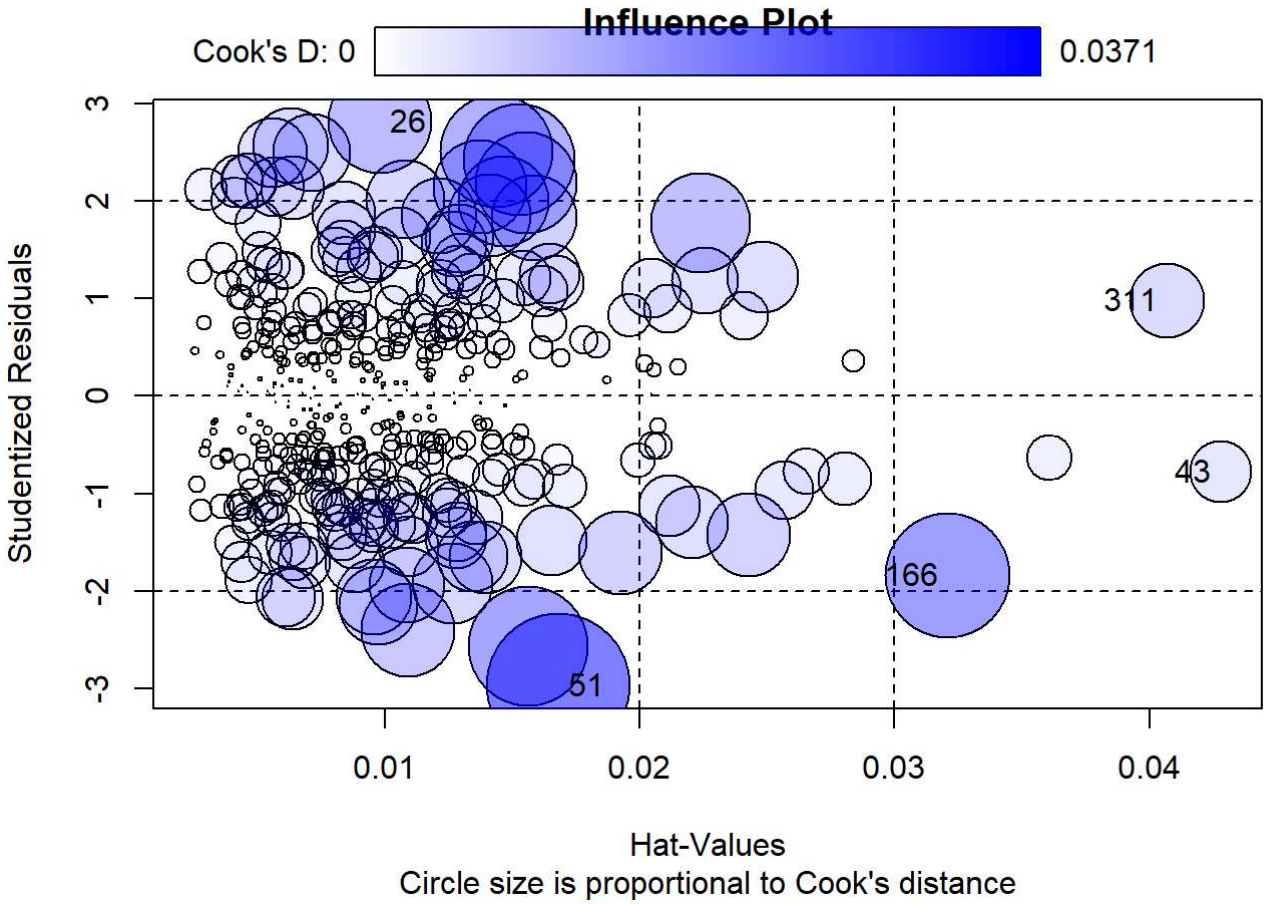
Thus, the 51^{th} carseat is an outlier.

```
library(MASS)
library(stats)
hat.plot <- function(fit) {
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main="Index Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red", lty=2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(fit2)
```




```
## integer(0)
```

```
library(car)
influencePlot(fit2, id.method="identify", main="Influence Plot",
              sub="Circle size is proportional to Cook's distance")
```



	StudRes	Hat	CookD
	<dbl>	<dbl>	<dbl>
26	2.8116323	0.009824792	0.01927350
43	-0.7828760	0.042797306	0.00685746
51	-2.9740147	0.016822337	0.03709892
166	-1.8393501	0.032086833	0.02787106
311	0.9801292	0.040695399	0.01018916

5 rows

Thus, if hat value is greater than $\frac{2k}{n} = 0.02$, then we maintain it is a high leverage point. From the Influential Plot above, we discover that #166, #311 and #43 are high leverage

(j)

```
urban <- subset(Carseats, Urban == "Yes")
nonurban <- subset(Carseats, Urban == "No")
mean(urban$Sales)
```

```
## [1] 7.468191
```

```
mean(nonurban$Sales)
```

```
## [1] 7.563559
```

Likelihood ratio test

```
alpha = 0.05
n1 <- length(urban$Sales)
n2 <- length(nonurban$Sales)

xbar1 <- mean(urban$Sales)
xbar2 <- mean(nonurban$Sales)

S1 <- var(urban$Sales)
S2 <- var(nonurban$Sales)
Sp <- sqrt( ((n1-1)*S1 + (n2-1)*S2) /(n1+n2-2) )

ifelse( abs(xbar1 - xbar2) / (sqrt(1/n1 + 1/n2)*Sp) > qt(df=n1+n2-2, 1-alpha/2) ,
       'Reject Hypothesis H_0', 'Fail to Reject Hypothesis H_0')
```

```
## [1] "Fail to Reject Hypothesis H_0"
```

According to LRT, we should reject null hypothesis at 0.05 significance level.

T.test

```
t.test(Sales~Urban,data = Carseats,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Sales by Urban
## t = 0.30765, df = 398, p-value = 0.7585
## alternative hypothesis: true difference in means between group No and group Yes is not equal
to 0
## 95 percent confidence interval:
##  -0.5140440  0.7047797
## sample estimates:
##  mean in group No mean in group Yes
##          7.563559          7.468191
```

According to t-test, we should not reject null hypothesis at 0.05 significance level.

Mann Whitney test

```
wilcox.test(Sales~Urban, data = Carseats)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sales by Urban
## W = 17225, p-value = 0.5784
## alternative hypothesis: true location shift is not equal to 0
```

According to Mann Whitney test, we should not reject null hypothesis at 0.05 significance level.

We cannot use wilcoxon's Signed-Rank test because the data are not paired.

(k)

We first perform stepwise methods.

```
library(MASS)
fit3 <- lm(Sales ~ .-X, data = Carseats)
summary(fit3)
```

```
##
## Call:
## lm(formula = Sales ~ . - X, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285
## UrbanYes         0.1228864  0.1129761   1.088    0.277
## USYes           -0.1840928  0.1498423  -1.229    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

```
stepAIC(fit3, direction="both")
```

```
## Start:  AIC=26.82
## Sales ~ (X + CompPrice + Income + Advertising + Population +
##     Price + ShelveLoc + Age + Education + Urban + US) - X
##
##               Df Sum of Sq     RSS    AIC
## - Population   1      0.33  403.16  25.15
## - Education    1      1.19  404.02  26.00
## - Urban        1      1.23  404.06  26.04
## - US           1      1.57  404.40  26.38
## <none>                      402.83  26.82
## - Income       1     76.16  478.99  94.09
## - Advertising  1    127.14  529.97 134.54
## - Age          1    217.44  620.27 197.48
## - CompPrice    1    519.91  922.74 356.35
## - ShelveLoc    2   1053.20 1456.03 536.80
## - Price        1   1323.23 1726.06 606.85
##
## Step:  AIC=25.15
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Urban        1      1.15  404.31  24.29
## - Education    1      1.36  404.52  24.49
## - US           1      1.89  405.05  25.02
## <none>                      403.16  25.15
## + Population   1      0.33  402.83  26.82
## - Income       1     75.94  479.10  92.18
## - Advertising  1    145.38  548.54 146.32
## - Age          1    218.52  621.68 196.38
## - CompPrice    1    521.69  924.85 355.27
## - ShelveLoc    2   1053.18 1456.34 534.89
## - Price        1   1323.51 1726.67 605.00
##
## Step:  AIC=24.29
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + US
##
##               Df Sum of Sq     RSS    AIC
## - Education    1      1.44  405.76  23.72
## - US           1      1.85  406.16  24.12
## <none>                      404.31  24.29
## + Urban        1      1.15  403.16  25.15
## + Population   1      0.25  404.06  26.04
## - Income       1     76.64  480.96  91.73
## - Advertising  1    146.03  550.34 145.63
## - Age          1    217.59  621.91 194.53
## - CompPrice    1    526.17  930.48 355.69
## - ShelveLoc    2   1053.93 1458.25 533.41
## - Price        1   1322.80 1727.11 603.10
##
## Step:  AIC=23.72
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + US
##
```

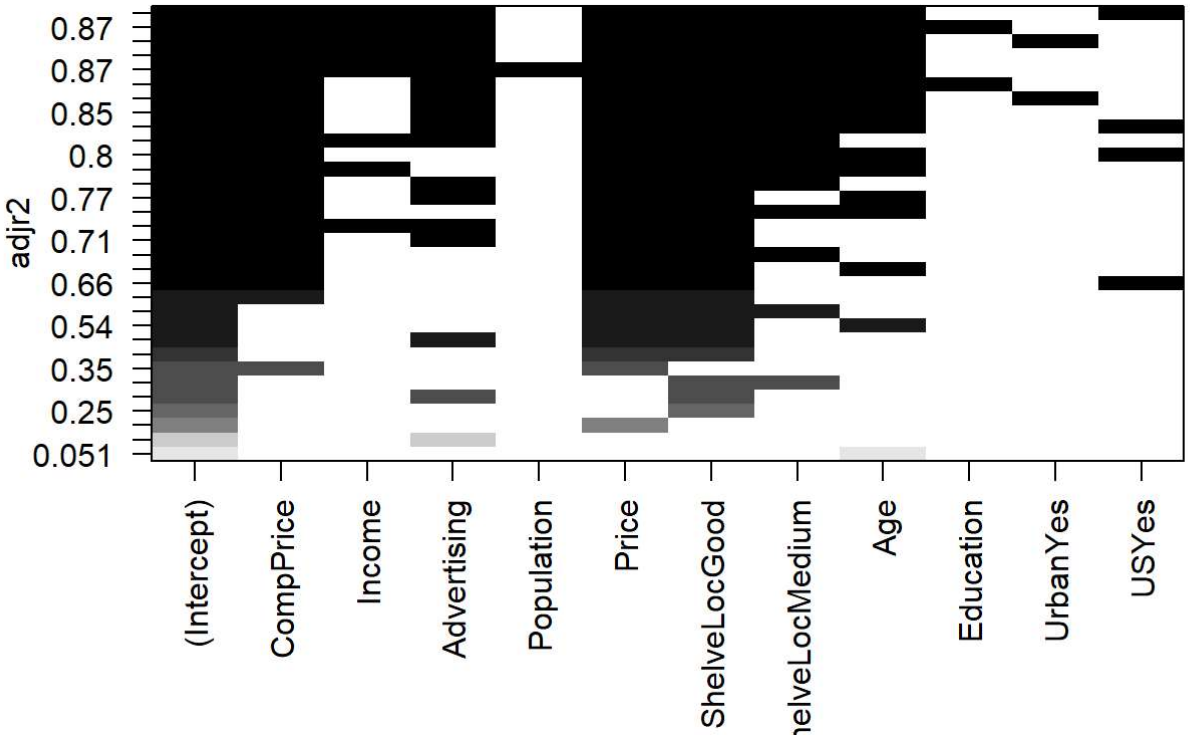
```
##               Df Sum of Sq     RSS    AIC
## - US           1      1.63  407.39  23.32
## <none>                      405.76  23.72
## + Education    1      1.44  404.31  24.29
## + Urban        1      1.24  404.52  24.49
## + Population   1      0.41  405.35  25.32
## - Income       1     77.87  483.62  91.94
## - Advertising  1    145.30  551.06 144.15
## - Age          1    217.97  623.73 193.70
## - CompPrice    1    525.25  931.00 353.92
## - ShelveLoc    2   1056.88 1462.64 532.61
## - Price        1   1322.83 1728.58 601.44
##
## Step:  AIC=23.32
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##
##               Df Sum of Sq     RSS    AIC
## <none>                      407.39  23.32
## + US           1      1.63  405.76  23.72
## + Education    1      1.22  406.16  24.12
## + Urban        1      1.19  406.20  24.15
## + Population   1      0.72  406.67  24.62
## - Income       1     76.68  484.07  90.30
## - Age          1    219.12  626.51 193.48
## - Advertising  1    234.03  641.42 202.89
## - CompPrice    1    523.83  931.22 352.01
## - ShelveLoc    2   1055.51 1462.90 530.68
## - Price        1   1324.42 1731.81 600.18
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = Carseats)
##
## Coefficients:
##     (Intercept)        CompPrice           Income      Advertising
##         5.47523          0.09257          0.01578          0.11590
##           Price    ShelveLocGood  ShelveLocMedium              Age
##        -0.09532          4.83567          1.95199         -0.04613
```

By using the both direction method, we can discover that the best model contains the variables ComPrice, Income, Advertising, Price, ShelveLocGood, ShelveLocMedian and Age.

Then we perform all-subsets methods.

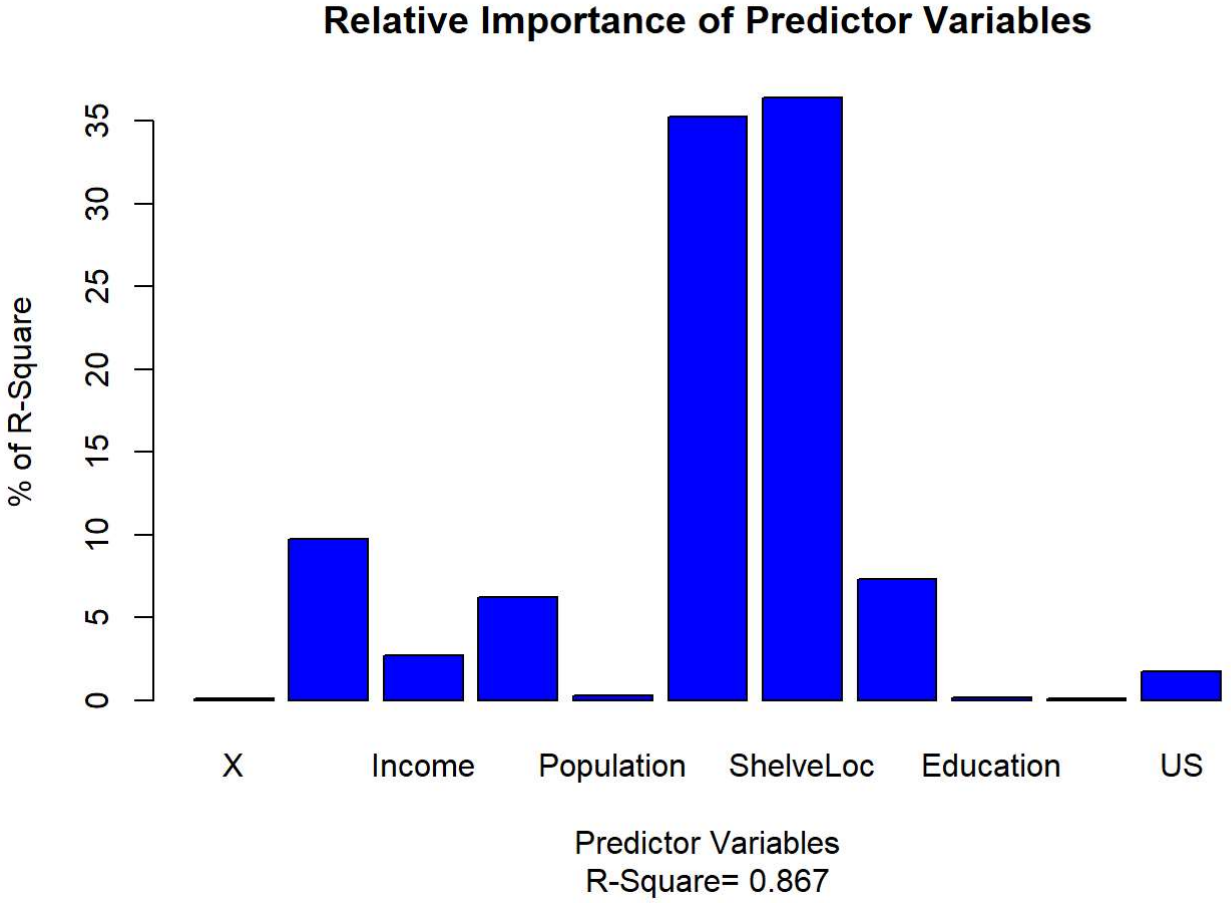
```
library(leaps)
leaps <-regsubsets(Sales ~ .-X, data = Carseats, nbest=4)
plot(leaps, scale="adjr2")
```



As we can discover from the figure above, the model with independent variables comPrice, Income, Advertising, Price,ShelveLocGood, ShelveLocMedian, Age and USYes is the best model.

(l)

```
Carseats <- read.csv("C:/Users/Lenovo/Desktop/R/Assignment 2/Carseats.csv")

Carseats$ShelveLoc[Carseats$ShelveLoc == "Good"] <- "2"
Carseats$ShelveLoc[Carseats$ShelveLoc == "Medium"] <- "1"
Carseats$ShelveLoc[Carseats$ShelveLoc == "Bad"] <- "0"
Carseats$ShelveLoc <- as.numeric(Carseats$ShelveLoc)

Carseats$Urban[Carseats$Urban == "Yes"] <- 1
Carseats$Urban[Carseats$Urban == "No"] <- 0
Carseats$Urban <- as.numeric(Carseats$Urban)

Carseats$US[Carseats$US == "Yes"] <- 1
Carseats$US[Carseats$US == "No"] <- 0
Carseats$US <- as.numeric(Carseats$US)



relweights0 <- function(fit,...){
  R <- cor(fit$model)
  nvar <- ncol(R)
  rxx <- R[2:nvar, 2:nvar]
  rxy <- R[2:nvar, 1]
  svd <- eigen(rxx)
  evec <- svd$vectors
  ev <- svd$values
  delta <- diag(sqrt(ev))
  lambda <- evec %*% delta %*% t(evec)
  lambdasq <- lambda ^ 2
  beta <- solve(lambda) %*% rxy
  rsquare <- colSums(beta ^ 2)
  rawwgt <- lambdasq %*% beta ^ 2
  import <- (rawwgt / rsquare) * 100
  lbls <- names(fit$model[2:nvar])
  rownames(import) <- lbls
  colnames(import) <- "Weights"
  barplot(t(import),names.arg=lbls,
          ylab="% of R-Square",
          xlab="Predictor Variables",
          main="Relative Importance of Predictor Variables",
          sub=paste("R-Square=", round(rsquare, digits=3)),
          ...)
  return(import)
}


fit4 <- lm(Sales ~. ,data = Carseats)
relweights0(fit4, col="blue")
```




```
##                 Weights
## X            0.10165500
## CompPrice    9.72817947
## Income       2.69600442
## Advertising  6.23233471
## Population   0.25257160
## Price       35.27829096
## ShelveLoc   36.40792800
## Age          7.32298477
## Education    0.13579129
## Urban        0.08534937
## US           1.75891041
```

```
scaled = as.data.frame(scale(Carseats))
zfit <- lm(Sales ~ ., data=scaled)
coef(zfit)
```

```
##   (Intercept)             X     CompPrice        Income   Advertising
## -5.248489e-16 -1.374617e-02  5.035880e-01  1.585225e-01  2.852866e-01
##    Population         Price     ShelveLoc           Age     Education
##  1.557519e-02 -7.985200e-01  5.748821e-01 -2.696199e-01 -2.071999e-02
##         Urban            US
##  2.364970e-02 -2.217110e-02
```

We can figure out from the result from calculating relative weights and standardized regression coefficients that important variables are Price and ShelveLoc. The result is consistent with the result in (k) by performing stepwise methods and all-subsets method.

Question 3

(a)

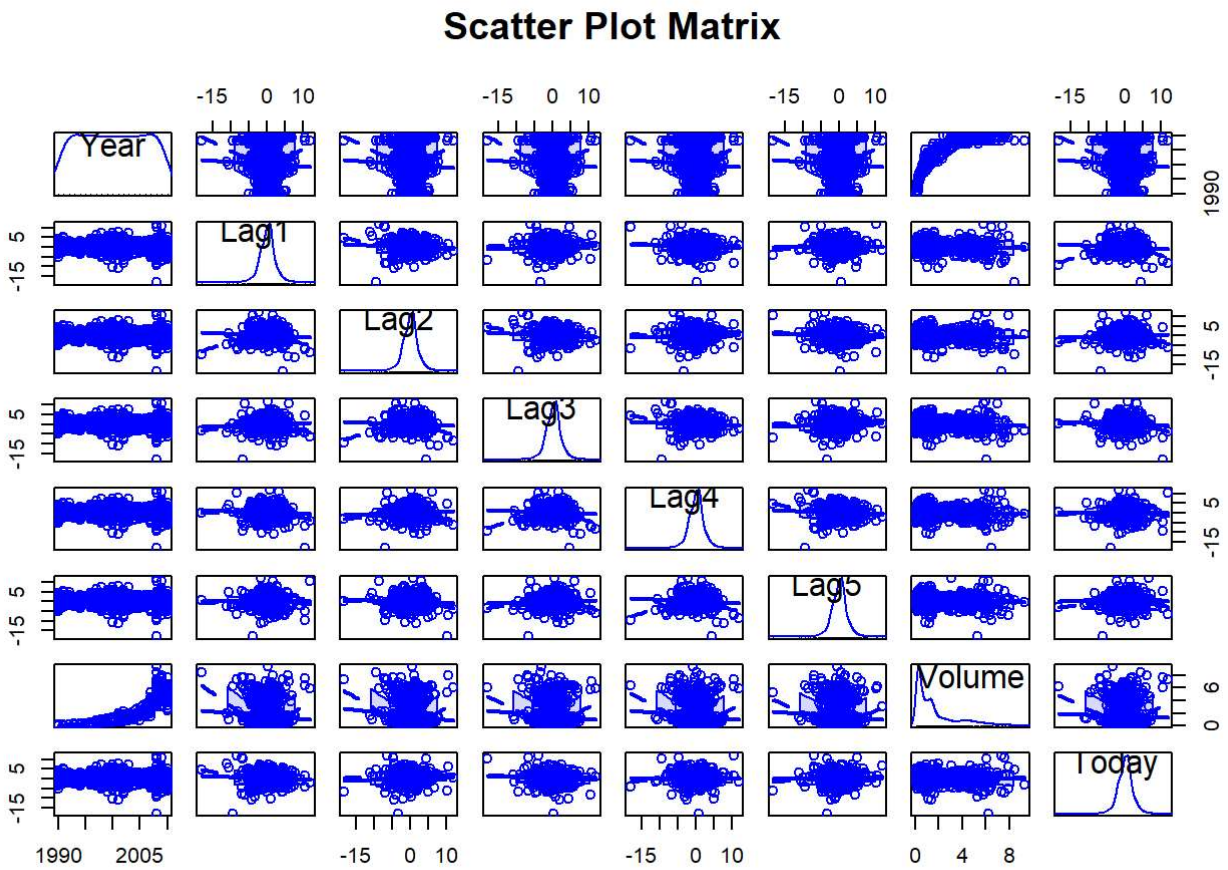
```
data <- read.csv("C:/Users/Lenovo/Desktop/R/Assignment 2/weekly.csv")

summary(data)
```

```
##        X              Year           Lag1               Lag2         
##  Min.   :   1   Min.   :1990   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.: 273   1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540  
##  Median : 545   Median :2000   Median :  0.2410   Median :  0.2410  
##  Mean   : 545   Mean   :2000   Mean   :  0.1506   Mean   :  0.1511  
##  3rd Qu.: 817   3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090  
##  Max.   :1089   Max.   :2010   Max.   : 12.0260   Max.   : 12.0260  
##       Lag3               Lag4               Lag5              Volume       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2410   Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1472   Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today           Direction        
##  Min.   :-18.1950   Length:1089        
##  1st Qu.: -1.1540   Class :character   
##  Median :  0.2410   Mode  :character   
##  Mean   :  0.1499                      
##  3rd Qu.:  1.4050                      
##  Max.   : 12.0260
```

```
cor(data[,-c(1,10)])
```

```
##                Year         Lag1        Lag2        Lag3         Lag4
## Year     1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1    -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2    -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3    -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4    -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5    -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume   0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today   -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                 Lag5       Volume         Today
## Year    -0.030519101  0.84194162 -0.032459894
## Lag1    -0.008183096 -0.06495131 -0.075031842
## Lag2    -0.072499482 -0.08551314  0.059166717
## Lag3     0.060657175 -0.06928771 -0.071243639
## Lag4    -0.075675027 -0.06107462 -0.007825873
## Lag5     1.000000000 -0.05851741  0.011012698
## Volume  -0.058517414  1.00000000 -0.033077783
## Today    0.011012698 -0.03307778  1.000000000
```

```
library(car)
scatterplotMatrix(data[,-c(1,10)],spread=FALSE, lty.smooth=2,main="Scatter Plot Matrix")
```



(b)

```
data$Direction[data$Direction == "Up"] <- 1
data$Direction[data$Direction == "Down"] <- 0
data$Direction <- as.numeric(data$Direction)
log.fit <- glm(Direction~Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = data, family = binom
ial())
summary(log.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial(), data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

The predictor Lag2 is statistically significant.

(c)

```
glm.probs <- predict(log.fit,type = "response")
glm.pred <- rep(0, 1089)
glm.pred[glm.probs>0.5] <- 1
table(data$Direction, glm.pred)
```

```
##    glm.pred
##       0   1
##   0  54 430
##   1  48 557
```

```
mean(glm.pred == data$Direction)
```

```
## [1] 0.5610652
```

The confusion matrix is shown above and the overall fraction of correct predictions is approximately 0.56.

The number 430 on the topright corner indicates there are 430 samples that are actually 0 but wrongly predicted to be 1. Additionally, the number 48 on the bottomleft corner indicates there are 48 samples that are actually 1 but wrongly predicted to be 0.

(d)

```
train <- subset(data, Year %in% 1990:2009)
test <- subset(data, Year == 2010)


log.fit2 <- glm(Direction ~ Lag2, data = train, family = binomial())
glm.probs2 <- predict(log.fit2, test, type = "response")
glm.pred2 <- rep(0, nrow(test))
glm.pred2[glm.probs2 > 0.5] <- 1


table(test$Direction, glm.pred2)
```

```
##     glm.pred2
##       0  1
##   0   3 17
##   1   0 32
```

```
mean(glm.pred2 == test$Direction)
```

```
## [1] 0.6730769
```

The confusion matrix is shown above and the overall fraction of correct predictions is approximately 0.67.