Inclass 5 Supplemental

Andriana Manousidaki

2023-09-23

Aim

With this assignment we will practice more on linear regression. Let us assume that the matrix X has the data values of the predictor variables, $X_1, X_2, ..., X_p$ and the matrix Y is the vector of observed data of the response variable.

We aim to find the Ordinary Least Squares Estimator, betaOLS, of the coefficients $ beta = (beta_1, beta_2,…,beta_p)^T$ of the model $Y = beta_0 + beta_1 X_1 + beta_2 X_2 + ... + beta_p X_p + e$, assuming that:

- $E(e) = 0$
- $Cov(e) = s_e^2 I_n$

When we have small sample size (n< 30) and want to do hypothesis testing on the coefficients, we will add the assumption - $ e $ follows $N(0, s_e^2 I_n)$

In-class practice

Read the data

The Boston data set, which records medv (median house value) for 506 census tracts(suburbs) in Boston.

We will seek to predict medv using 12 predictors such as rm (average number of rooms per house), age (proportion of owner-occupied units built prior to 1940), and lstat (percent of households with low socioeconomic status). For full description see here.

```
#install.packages('ISLR2')
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.1.3
```

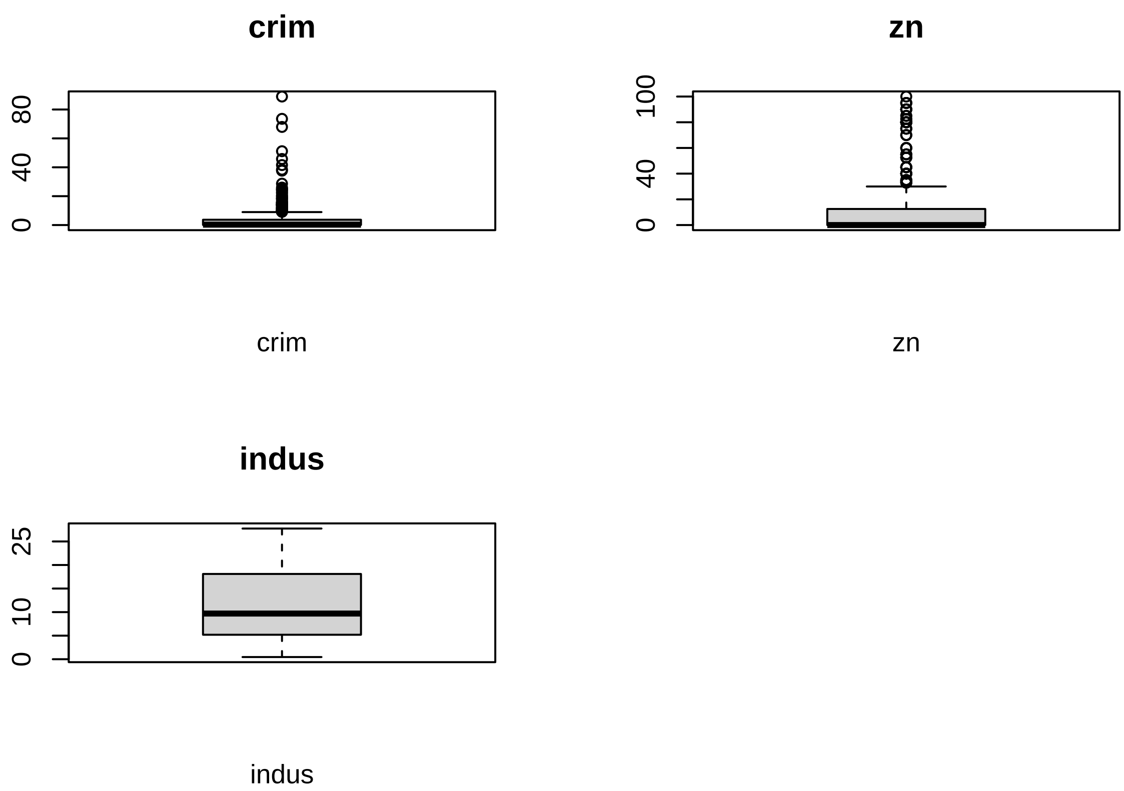
```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

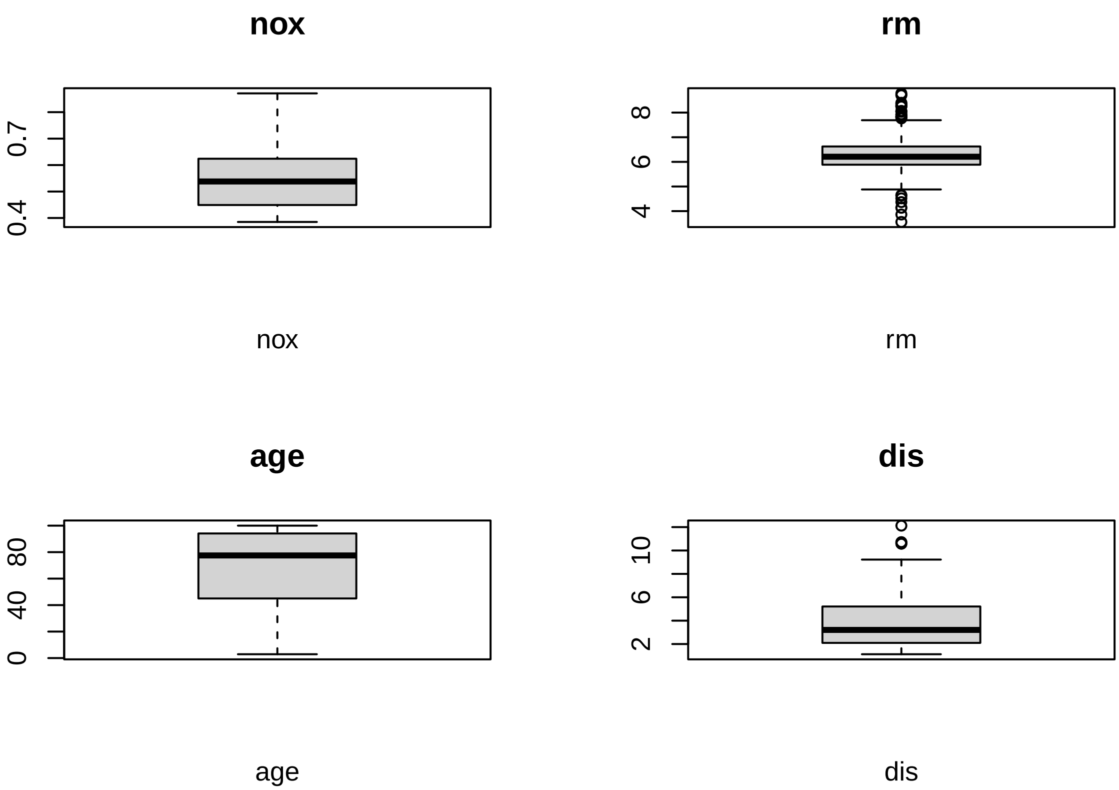
Summary of variables

Boxplots

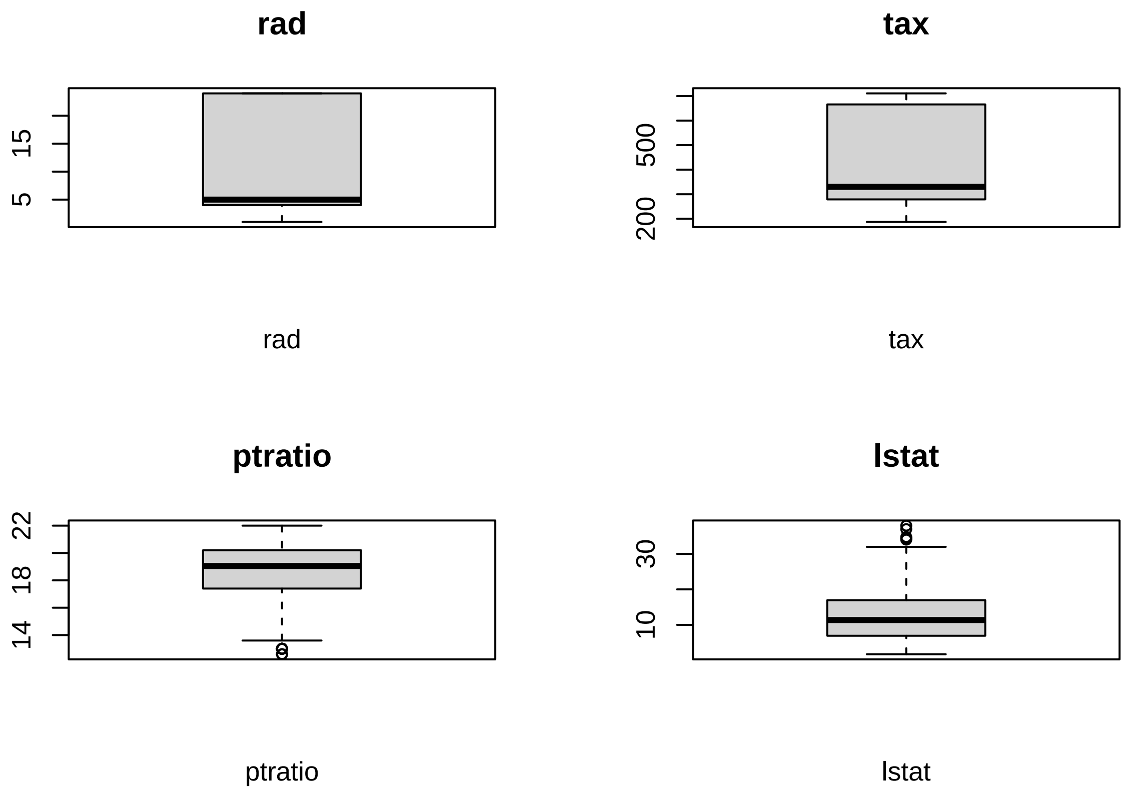
```
par(mfrow=c(2,2))
for (i in 1:3) {
boxplot(Boston[,i],main=colnames(Boston)[i],xlab=colnames(Boston)[i],data=Boston)
}

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




```
for (i in 5:8) {
boxplot(Boston[,i],main=colnames(Boston)[i],xlab=colnames(Boston)[i],data=Boston)
}
```

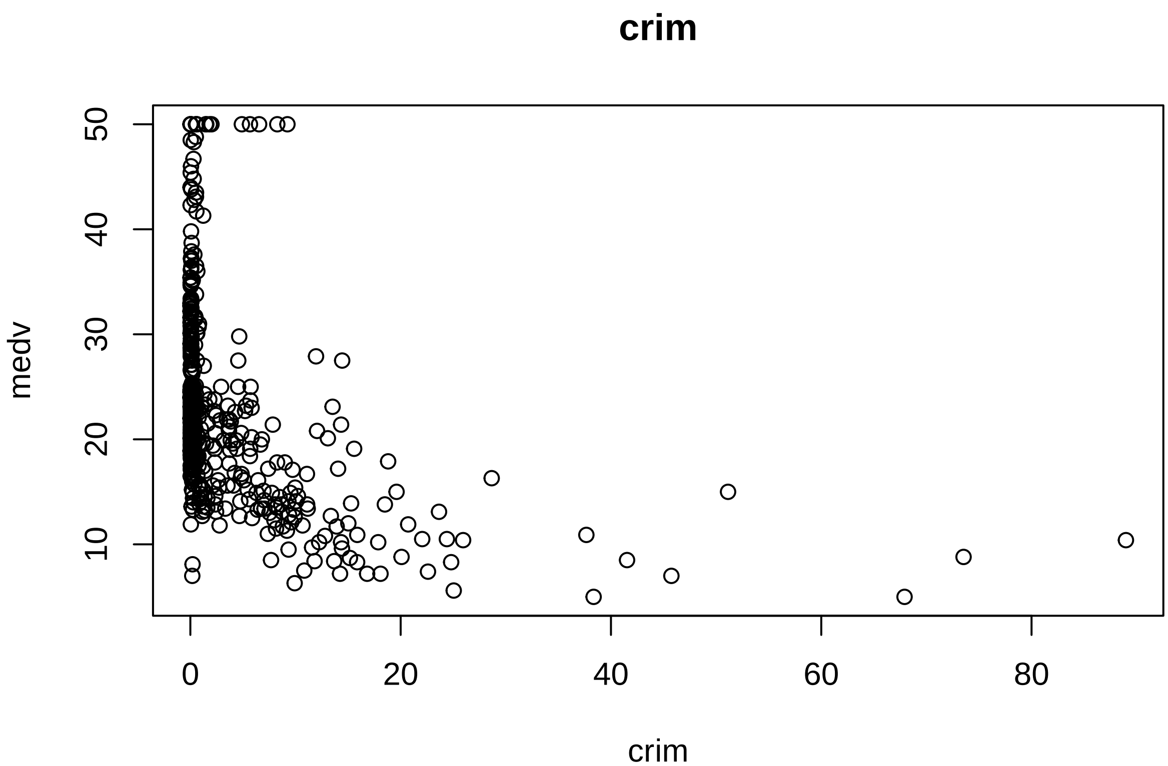


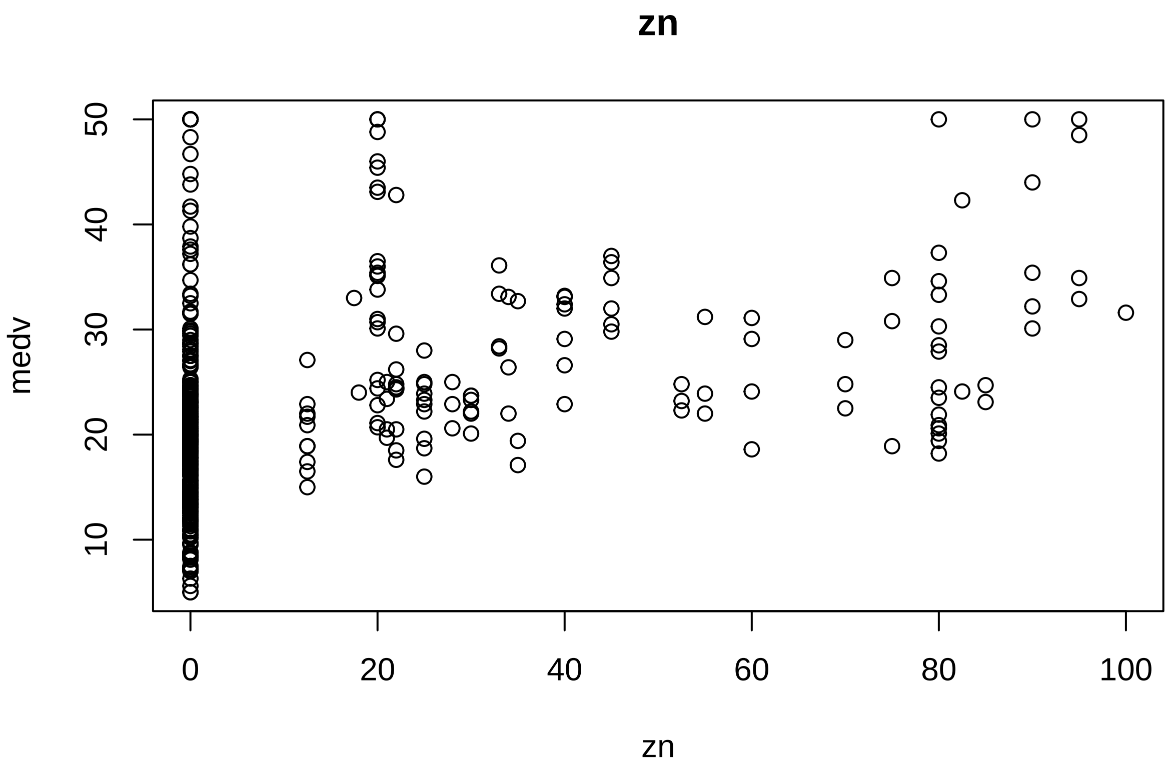

```
par(mfrow=c(2,2))
for (i in 9:12) {
boxplot(Boston[,i],main=colnames(Boston)[i],xlab=colnames(Boston)[i],data=Boston)
}
```



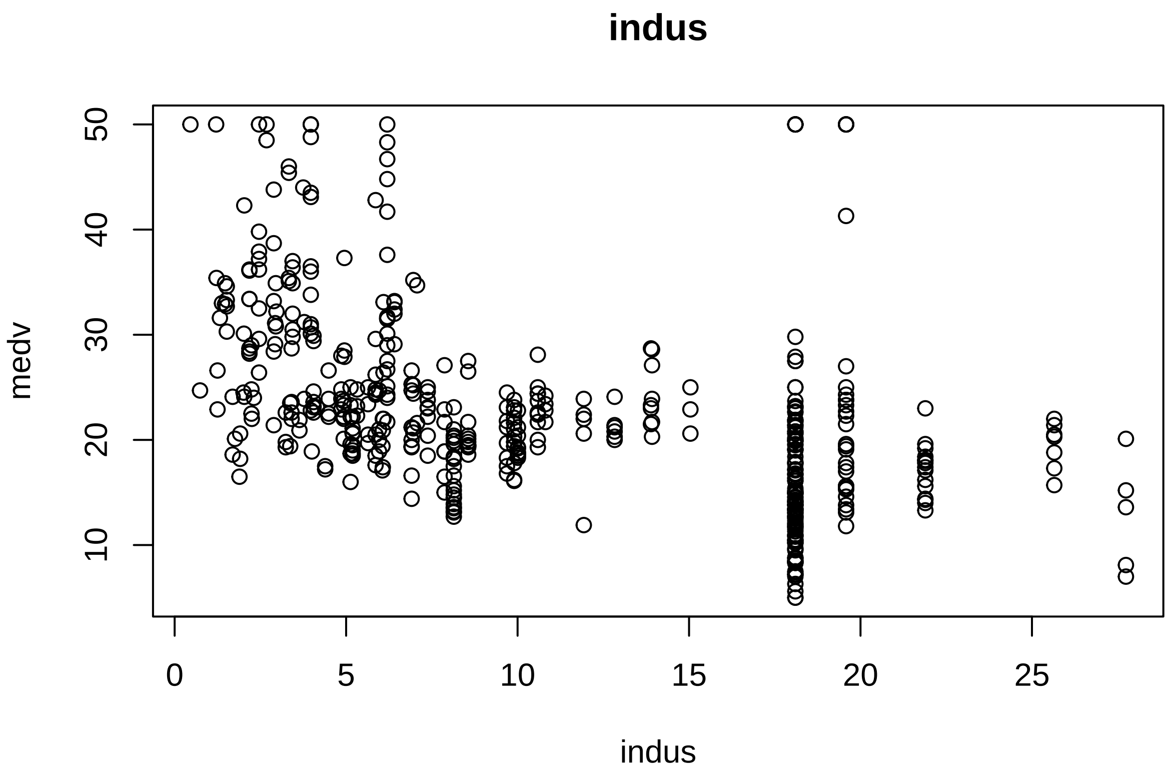
Scatterplot

```
for (i in 1:11) {
  if (i!=4) {
    plot(medv~Boston[,i],main=colnames(Boston)[i],xlab=colnames(Boston)[i],data=Boston)
  }
}
```

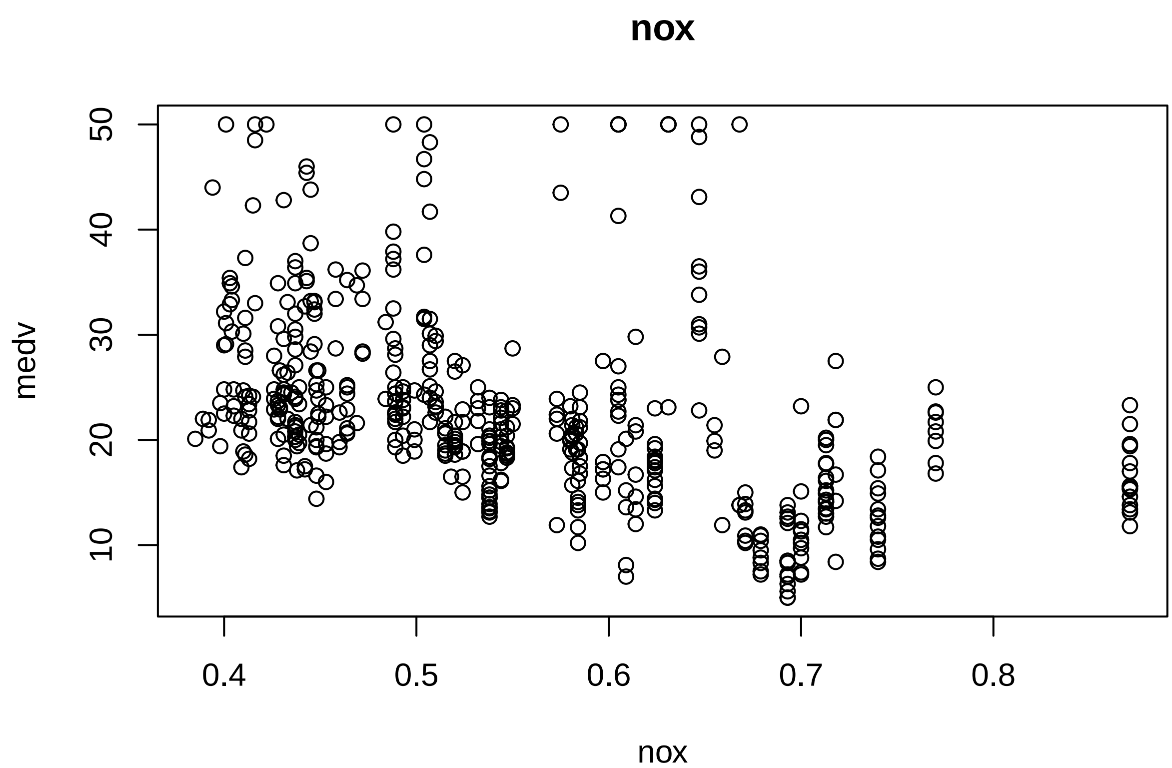

crim
medv
50
40
30
20
10
0
20
40
60
80
crim

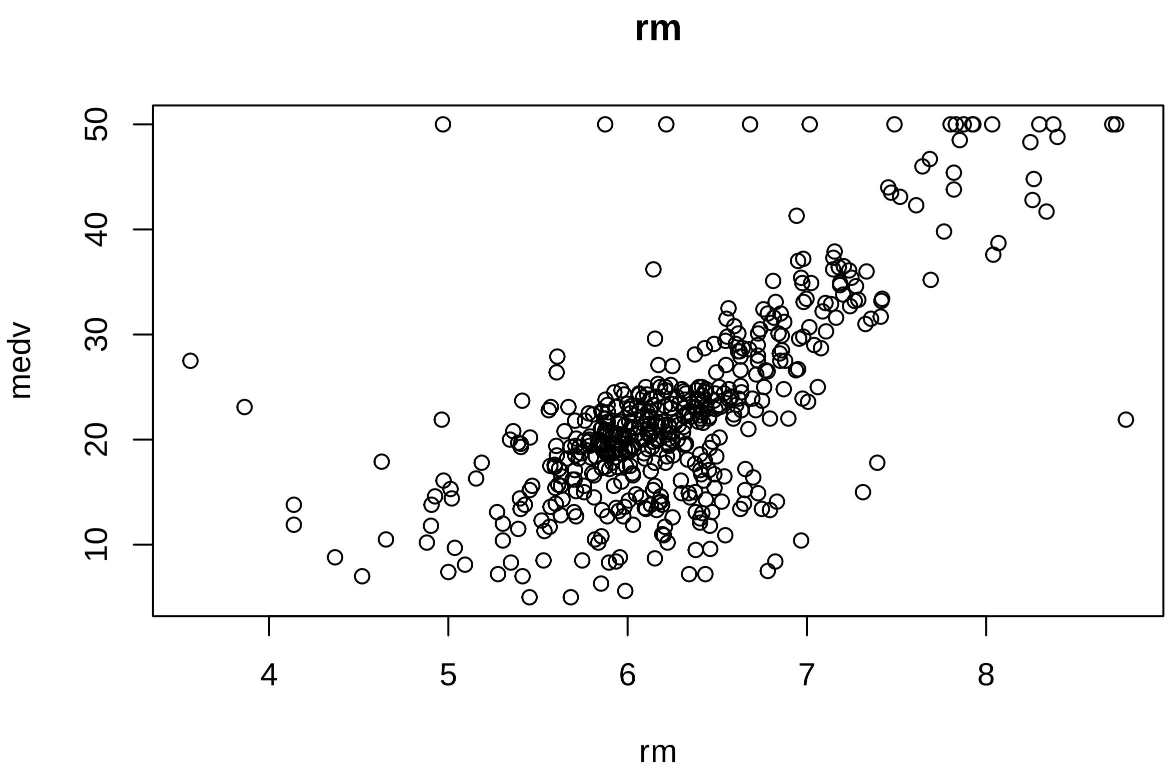


zn
50
40
30
20
10
medv
0
20
40
60
80
100
zn

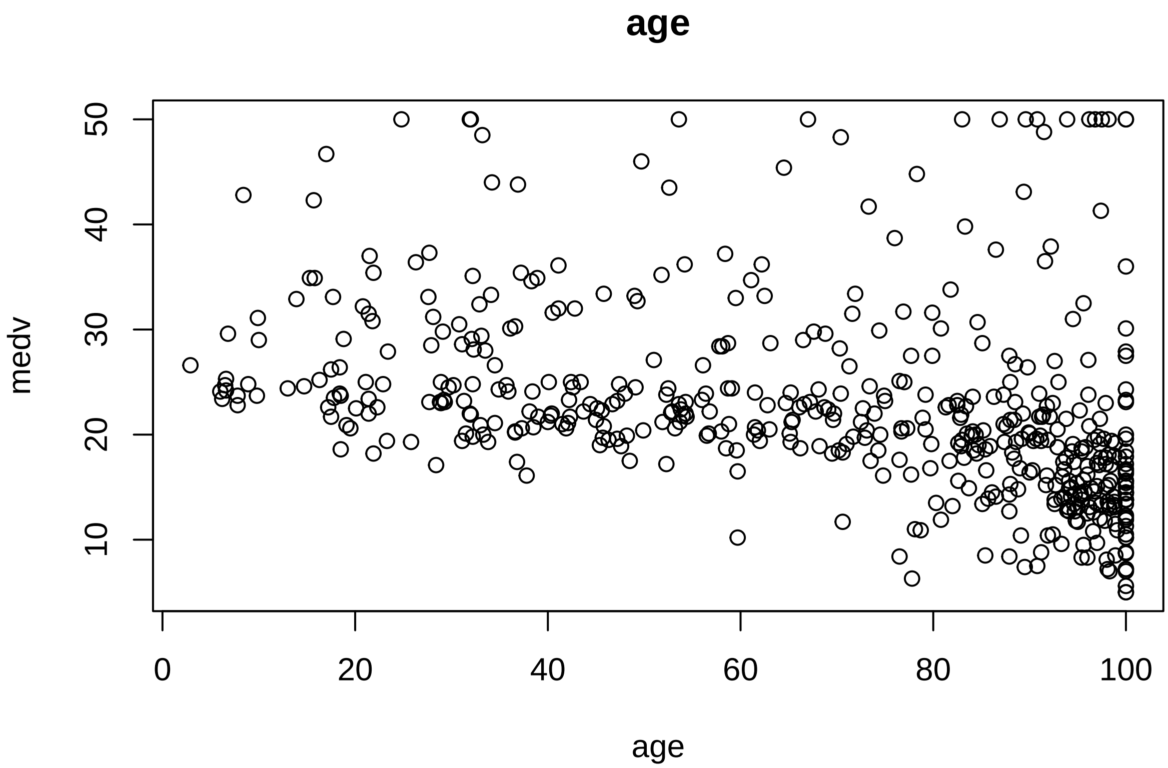


indus
medv
50
40
30
20
10
0
5
10
15
20
25
indus

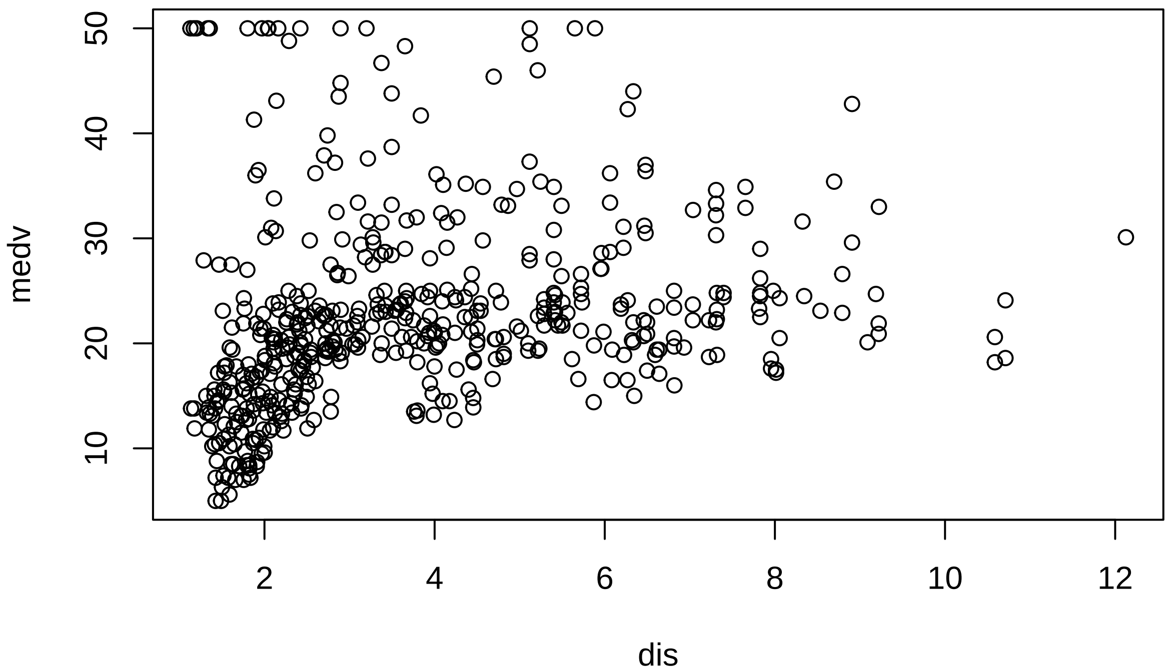

nox
50
40
30
20
10
medv
0.4
0.5
0.6
0.7
0.8
nox

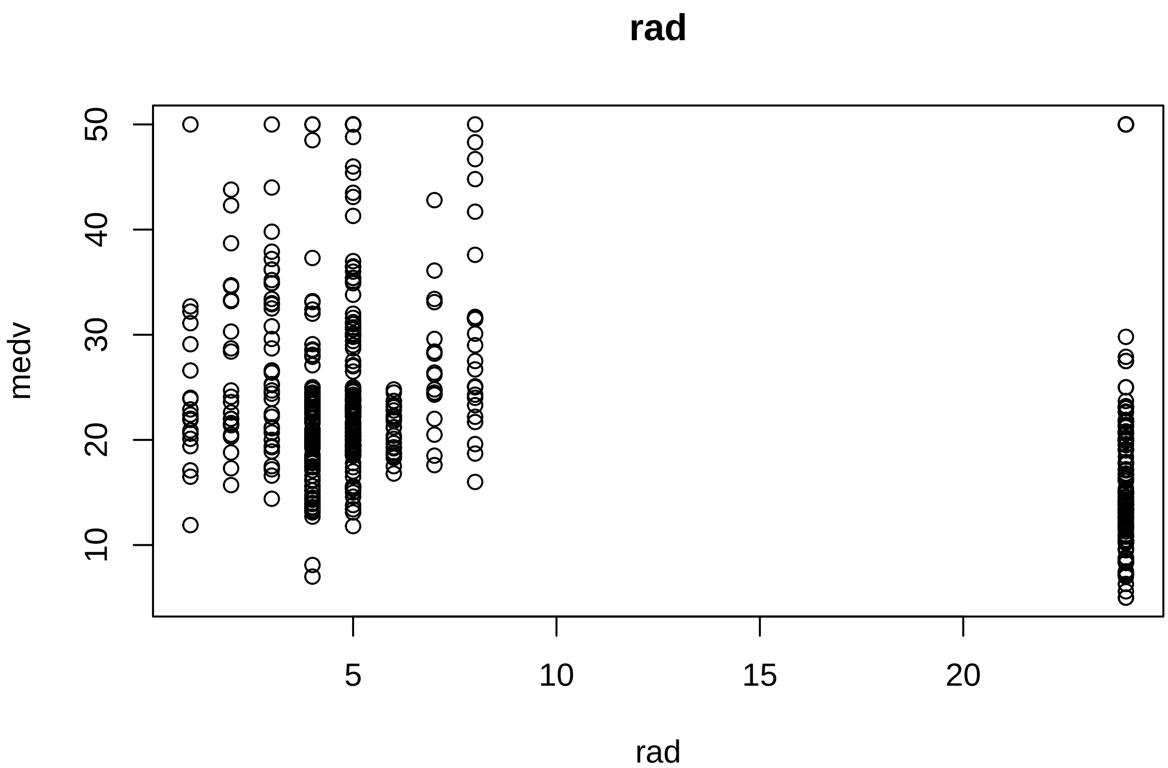


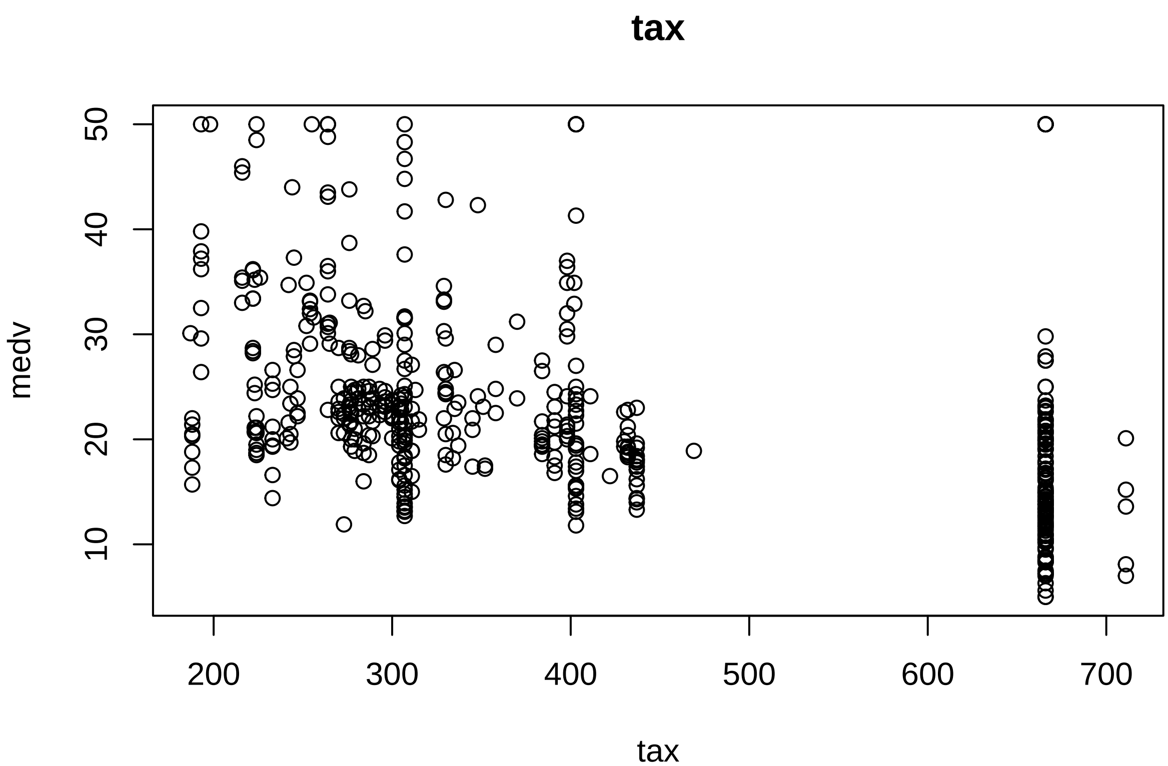
rm
medv
rm
50
40
30
20
10
4
5
6
7
8

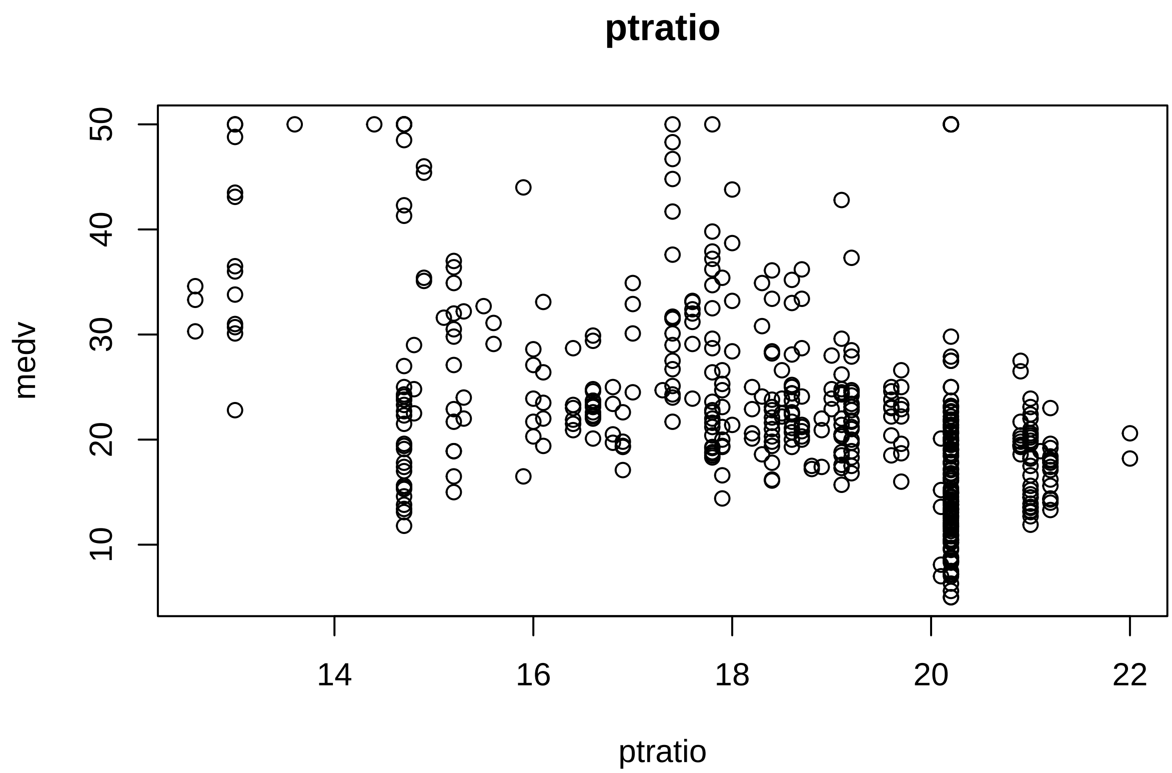

age
medv
50
40
30
20
10
0
20
40
60
80
100
age

dis




rad
medv
50
40
30
20
10
5
10
15
20
rad


tax
medv
50
40
30
20
10
200
300
400
500
600
700
tax



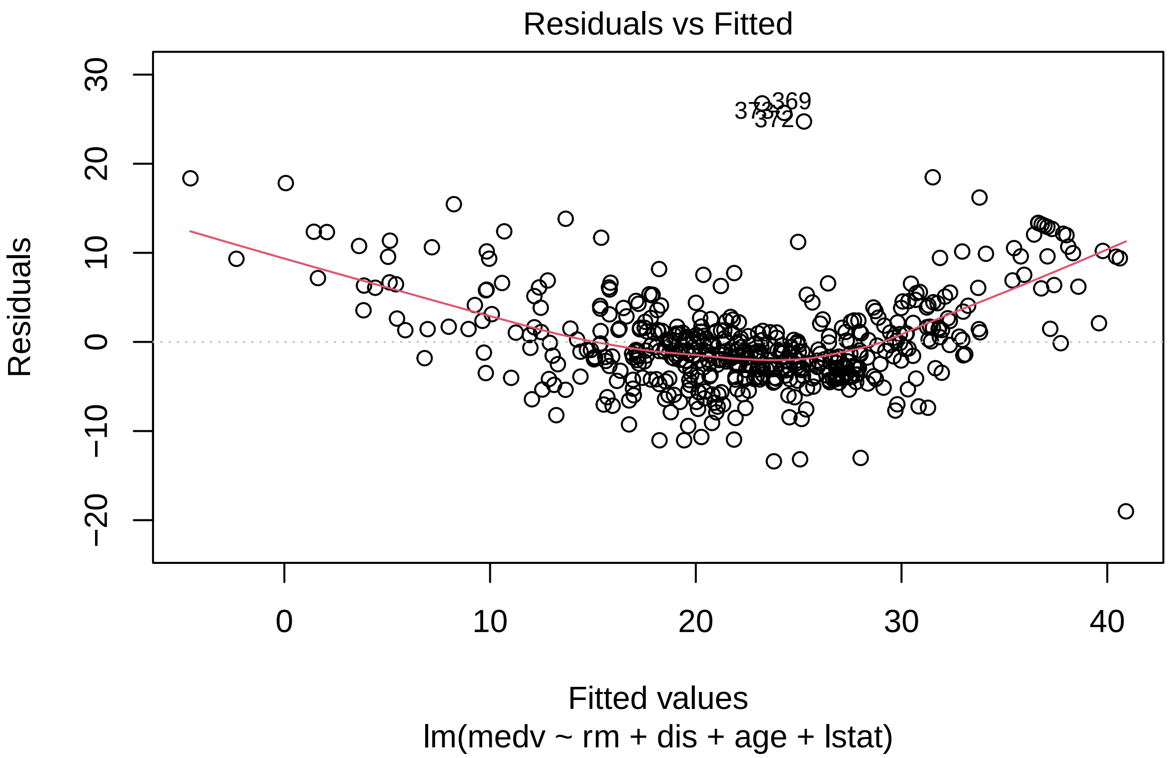
Fit the model

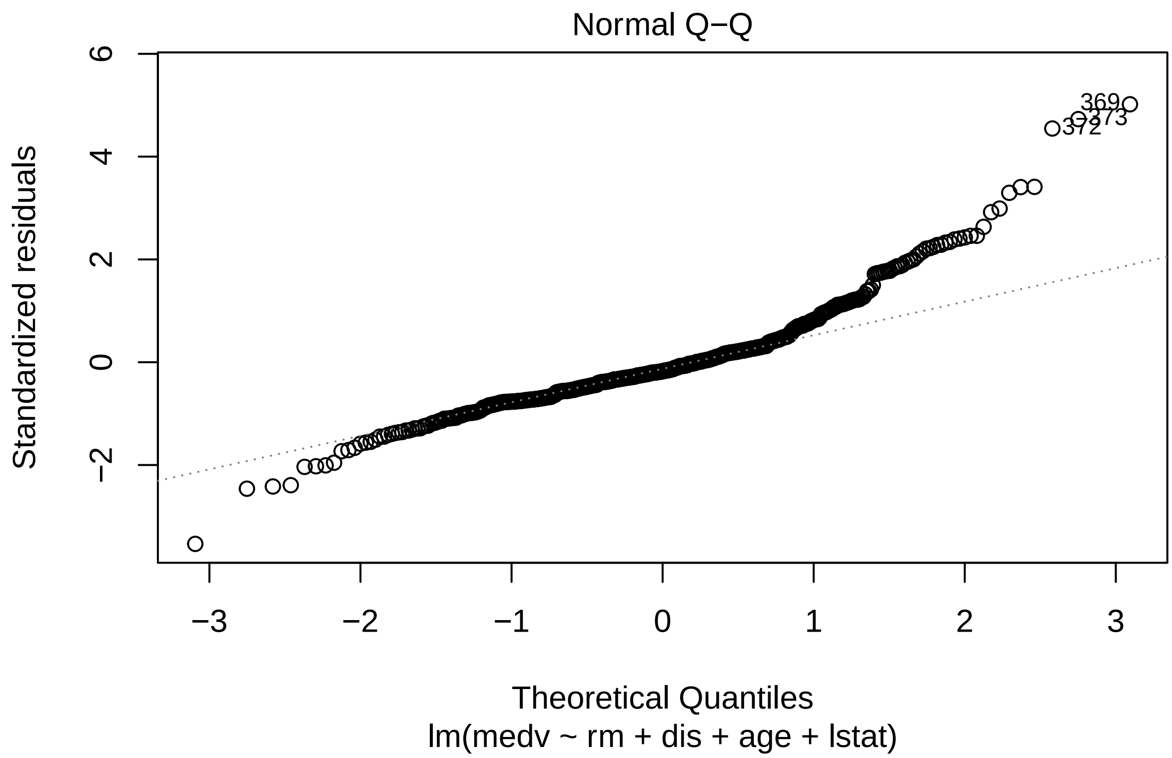
```
lm.fit<- lm(medv ~ rm + dis + age + lstat, data = Boston)
summary(lm.fit)
```

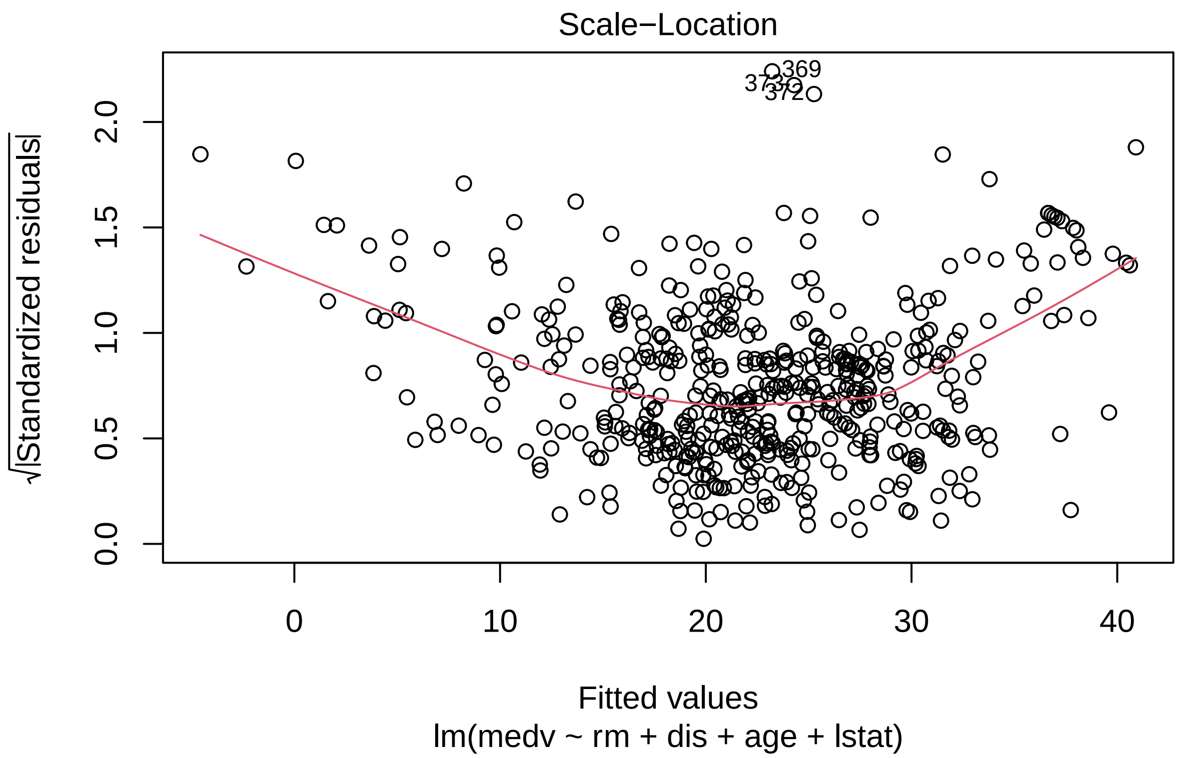
```
##
## Call:
## lm(formula = medv ~ rm + dis + age + lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0084  -3.0948  -0.9635   1.7074  26.7732
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.07985    3.43673   1.187 0.235738
## rm           4.99039    0.44852  11.126  < 2e-16 ***
## dis         -0.65876    0.17486  -3.767 0.000185 ***
## age         -0.02545    0.01437  -1.771 0.077236 .
## lstat       -0.68478    0.05383 -12.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.471 on 501 degrees of freedom
## Multiple R-squared:  0.649,  Adjusted R-squared:  0.6462
## F-statistic: 231.6 on 4 and 501 DF,  p-value: < 2.2e-16
```

Plots to evaluate residual assumptions

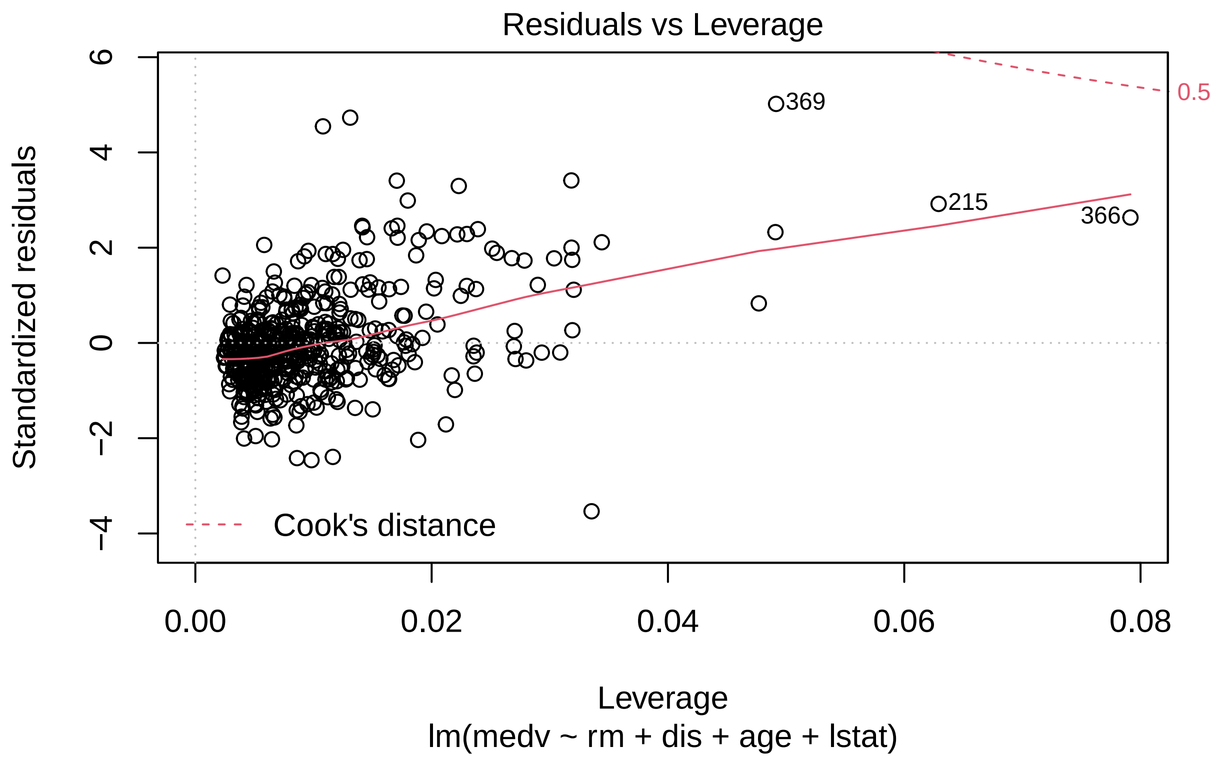
```r
plot(lm.fit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(medv ~ rm + dis + age + lstat)
369
373
372
6
4
2
0
−2
−3
−2
−1
0
1
2
3



Scale−Location
369
373
372
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Fitted values
lm(medv ~ rm + dis + age + lstat)



F-test 1: Assuming that the assumptions were true, is at least one of the predictors X1, X2,. . . ,Xp useful in predicting the response? or are all coefficients are zero and there is only the intercept in the model?

```
n=dim(Boston)[1]
H0 = lm(medv ~ 1, data = Boston)
Ha = lm(medv ~ rm + dis + age + lstat, data = Boston)

RSS0 = sum(residuals(H0)^2)
RSSA = sum(residuals(Ha)^2)

MSS = RSS0 - RSSA

df1 = length(coef(Ha)) - length(coef(H0))
df1
```

```
## [1] 4
```

```
df2 = n-length(coef(Ha))
df2
```

```
## [1] 501
```

```r
Fstat = (MSS/df1)/(RSSA/df2)
Fstat
```

```
## [1] 231.5659
```

```r
pval= pf(Fstat, df1 =df1, df2=df2, lower.tail = F)

print(c('Ftest' = Fstat, 'df1'= df1, 'df2'=df2, 'pval' = pval ))
```

```
##         Ftest           df1           df2          pval
##  2.315659e+02  4.000000e+00  5.010000e+02 2.088092e-112
```

```r
anova(H0,Ha)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ 1
## Model 2: medv ~ rm + dis + age + lstat
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    505 42716
## 2    501 14994  4     27722 231.57 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correlation coefficient

$R^2 = 1 - SSE/SST$

```r
R2= 1 - RSSA/RSS0

R2
```

```
## [1] 0.6489787
```

F-test 2: Test if age and dis are not useful in predicting medv.

```r
n=dim(Boston)[1]
H0 = lm(medv ~ rm + lstat, data = Boston)
Ha = lm(medv ~ rm + dis + age + lstat, data = Boston)

RSS0 = sum(residuals(H0)^2)
RSSA = sum(residuals(Ha)^2)

MSS = RSS0 - RSSA

df1 = length(coef(Ha)) - length(coef(H0))
df1
```

```
## [1] 2
```

```r
df2 = n-length(coef(Ha))
df2
```

```
## [1] 501
```

```r
Fstat = (MSS/df1)/(RSSA/df2)
Fstat
```

```
## [1] 7.433938
```

```r
pval= pf(Fstat, df1 =df1, df2=df2, lower.tail = F)

print(c('Ftest' = Fstat, 'df1'= df1, 'df2'=df2, 'pval' = pval ))
```

```
##        Ftest          df1          df2         pval
## 7.433938e+00 2.000000e+00 5.010000e+02 6.583551e-04
```

```r
anova(H0, Ha)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ rm + lstat
## Model 2: medv ~ rm + dis + age + lstat
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    503 15439
## 2    501 14994  2    444.98 7.4339 0.0006584 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding interactions in a model

It is easy to include interaction terms in a linear model using the lm() function. The syntax lstat:age tells R to include an interaction term between lstat and age. The syntax lstat * age simultaneously includes lstat, age, and the interaction term lstat×age as predictors; it is a shorthand for lstat + age + lstat:age.

```r
summary(lm(medv ~ lstat * age, data = Boston))
```

```
##
## Call:
## lm(formula = medv ~ lstat * age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.806  -4.045  -1.333   2.085  27.552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
## lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
## age         -0.0007209  0.0198792  -0.036   0.9711
```

```
## lstat:age    0.0041560  0.0018518   2.244   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 502 degrees of freedom
## Multiple R-squared:  0.5557, Adjusted R-squared:  0.5531
## F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Submit by Tuesday midnight

We will attempt to predict Sales (child car seat sales) in 400 locations based on a number of predictors.The Carseats data includes qualitative predictors such as Shelveloc, an indicator of the quality of the shelving location—that is, the space within a store in which the car seat is displayed—at each location. The predictor Shelveloc takes on three possible values: Bad, Medium, and Good.

More on Carseats dataset here

Question 1: Explore the data set Carseats, create boxplots for each variable and a scatterplot of sales vs each of the other variable. Which of the variables do you expect to have a linear relationship with sales?

```r
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

Question 2: Predict Carseat sales using all other variables

Question 3: Explore evaluation plots for the above model. What do you observe, are the assumption regarding the model errors satisfied?

Question 4: Explore evaluation plots for the above model. What do you observe, are the assumption regarding the model errors satisfied?

Question 5: Assuming that assumptions about the error term hold. Use an Ftest to answer the following question.Is at least one of the predictors X1, X2,. . . ,Xp useful in predicting the response? or are all coefficients are zero and there is only the intercept in the model?

Question 6: Use the Ftest to test if the variables Population, Education, Urban, US, Age are useful in predicting Sales.

Question 7: Fit the model Sales vs the remaining useful variables and adding some interatcions in the model. Write the regression equation with the estimated coefficients and interpret the coefficients of the intereaction terms.