MATH 4322 - Homework 2

Instructions

- Due September 14, 2023 at 11:59 pm
- Answer all questions fully
- Submit the answers in one file, preferably PDF, then upload in Canvas.
- These questions are from **Introduction to Statistical Learning, 2nd edition**, chapters 3 and 6.

Problem 1

The following output is based on predicting `sales` based on three media budgets, `TV`, `radio`, and `newspaper`.

```
Call:
lm(formula = sales ~ TV + radio + newspaper, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
```

```
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

a. Give the estimated model to predict sales.
   #ANS:

   sales = 2.938889 + 0.045765 * TV + 0.188530 * radio − 0.001037 * newspaper

b. Describe the null hypothesis to which the p-values given in the `Coefficients` table correspond. Explain this in terms of the `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

   #ANS:

   H0: TV is not needed in the model, if the radio and newspaper are in the model with the t = 32.809 and p-value = 0

   H0: radio is not needed in the model, if the TV and newspaper are in the model with t = 21.893 and p-value = 0

   H0: newspaper is not needed in the model, if the TV and radio are in the model with t = −0.177 and p-value = 0.86

c. Are there any variables that may not be significant in predicting `sales`?
   #ANS:

   The variable that may may not be significant in predicting sales is newspaper, since the p-value is larger than 0.05

Problem 2

Based on the previous problem,the following is the output from the full model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

```
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq   F value Pr(>F)
TV          1 3314.6  3314.6 1166.7308 <2e-16 ***
radio       1 1545.6  1545.6  544.0501 <2e-16 ***
newspaper   1    0.1     0.1    0.0312 0.8599
Residuals 196  556.8     2.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is based on the model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

```
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)
TV          1 3314.6  3314.6 1172.50 < 2.2e-16 ***
radio       1 1545.6  1545.6  546.74 < 2.2e-16 ***
Residuals 197  556.9     2.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is based on the model $\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \epsilon$

```
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq F value    Pr(>F)
TV          1 3314.6  3314.6  312.14 < 2.2e-16 ***
Residuals 198 2102.5    10.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Determine the `AIC` for all three models.

#ANS:

#Formula: AIC = 2(p+1) + n ln(SSE/n)

sales.lm:

```
#|echo = true
AIC = 2 * (4) + 200 * log(556.8/200)
AIC
```

```
[1] 212.7777
```

sales2.lm:

```
#|echo = true
AIC = 2* (3) + 200 * log(556.9/200)
AIC
```

```
[1] 210.8137
```

sales1.lm:

```
#|echo = true
AIC = 2 * (2) + 200 * log(2102.5/200)
AIC
```

```
[1] 474.513
```

b) Determine the C_p for all three models.

#ANS:

#Formula: Cp = SSEp/MSEall + 2(p+1) - n

sales.lm:

```
#|echo = true
Cp = 556.8/2.8 + 2 * (4) - 200
Cp
```

```
[1] 6.857143
```

sales2.lm:

```
#|echo = true
Cp = 556.9/2.8 + 2 * (3) - 200
Cp
```

```
[1] 4.892857
```

sales1.lm:

```
#|echo = true
Cp = 2102.5/2.8 + 2 * (2) - 200
Cp
```

```
[1] 554.8929
```

c) Determine the adjusted R^2 for all three models.

#ANS:

#Formula: Adjusted R^2 = 1 - (SSE/(n-p-1)/SST/(n-1))

sales.lm:

```
#|echo = true
AdjRsq = 1 - ((556.8/(200 - 3 - 1))/ (5417.1/199))
AdjRsq
```

```
[1] 0.8956411
```

sales2.lm:

```
#|echo = true
AdjRsq = 1 - ((556.9/(200 - 2 - 1))/ (5417.1/199))
AdjRsq
```

```
[1] 0.8961522
```

sales1.lm:

```
#|echo = true
AdjRsq = 1 - ((2102.5/(200 - 1 - 1))/ (5417.1/199))
AdjRsq
```

```
[1] 0.609917
```

d) Determine the `RSE` for all three models.
#ANS:

#Formula: RSE = sqrt(SSE/(n - p + 1))

sales.lm:

```r
#|echo = true
RSE = sqrt(556.8/(200 - (3 + 1)))
RSE
```

```
[1] 1.685472
```

sales2.lm:

```r
#|echo = true
RSE = sqrt(556.9/(200 - (2 + 1)))
RSE
```

```
[1] 1.68134
```

sales1.lm:

```r
#|echo = true
RSE = sqrt(2102.5/(200 - (1 + 1)))
RSE
```

```
[1] 3.258633
```

e) Which model best fits to predict `sales` based on these statistics?
#ANS:

The best model is sales2.lm due to its AIC, Cp, and the RSE being the lowest and its adjusted R^2 being the highest.

sales = B0 + B1 × TV + B2 × radio + e

Problem 3

Suppose we have a data set with five predictors, X_1 =GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$,$\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

#ANS:

#Model if Male: salary = 50 + 20 * GPA + 0.07 * IQ + 35 * 0 + 0.01 * GPA * IQ

#Model if Female: salary = 50 + 10 * GPA + 0.07 * IQ + 35 * 1 + 0.01 * GPA * IQ

Answer iii. is correct since a male's higher GPA would result in their salary being higher.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

#ANS:

GPA = 4.0

IQ = 110

50 + 10 * GPA + 0.07 * IQ + 35 * 1 + 0.01 * GPA * IQ

The predicted salary is $137.1 (in thousands) or $137,100

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

#ANS: This is somewhat true, but we need to clarify this by using the t-test statistic

Problem 4

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots, p$ predictors. Answer true or false to the following statements.

(a) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection.
#ANS: True

(b) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by backward stepwise selection.
#ANS: True

(c) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.
#ANS: False

(d) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

#ANS: False

(e) The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$ - variable model identified by best subset selection.

#ANS: False

Problem 5

This question involves the use of simple linear regression on the *Auto* data set. This can be found in the `ISLR2` package in `R`.

(a) Use the `lm()` function to perform a simple linear regression with *mpg* as the response and *horsepower* (*hp*) as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive or negative?

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals? Give an interpretation of these intervals.

#ANS:

```r
#|echo = true
library(ISLR2)
data("Auto")
# Fit a simple linear regression model with mpg
# as the response and horsepower (hp) as the predictor
auto.lm = lm(mpg ~ horsepower, data = Auto)
# Print the summary of the regression results
summary(auto.lm)
```

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i.

There is a relationship between mpg and horsepower. Because the p-value of $2e - 16$ is < 0.05, it suggests a statistically significant relationship.

ii.

The R2 value of 0.6059 indicates that roughly 61% of the variation in the response variable (mpg) is due to the predictor variable (horsepower).The relationship between mpg and horsepower is somewhat strong.

iii.

The relationship between mpg and horsepower is negative, horsepower's coefficient is -0.158.

iv.

```r
predict(auto.lm, newdata = data.frame(horsepower = 98),interval = "c")
```

```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```r
predict(auto.lm, newdata = data.frame(horsepower = 98),interval = "p")
```

```
       fit      lwr      upr
1 24.46708 14.8094 34.12476
```

The predicted mpg associated with a horsepower of 98 is 24.46708.

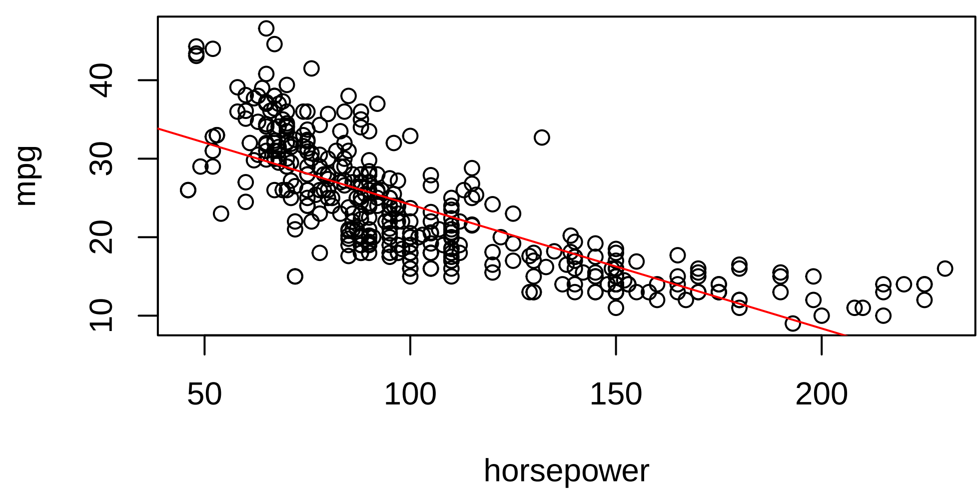
The associated 95% confidence interval is [23.97308, 24.96108], means that we are 95% confident that all automobiles with the horsepower of 98 have the mean mpg between 23.97308 and 24.96108.

The associated 95% prediction interval is [14.8094, 34.12476], meaning we are 95% confident that for one automobile with the horsepower of 98, have the mpg falls between 14.8094 and 34.12476.

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

#ANS

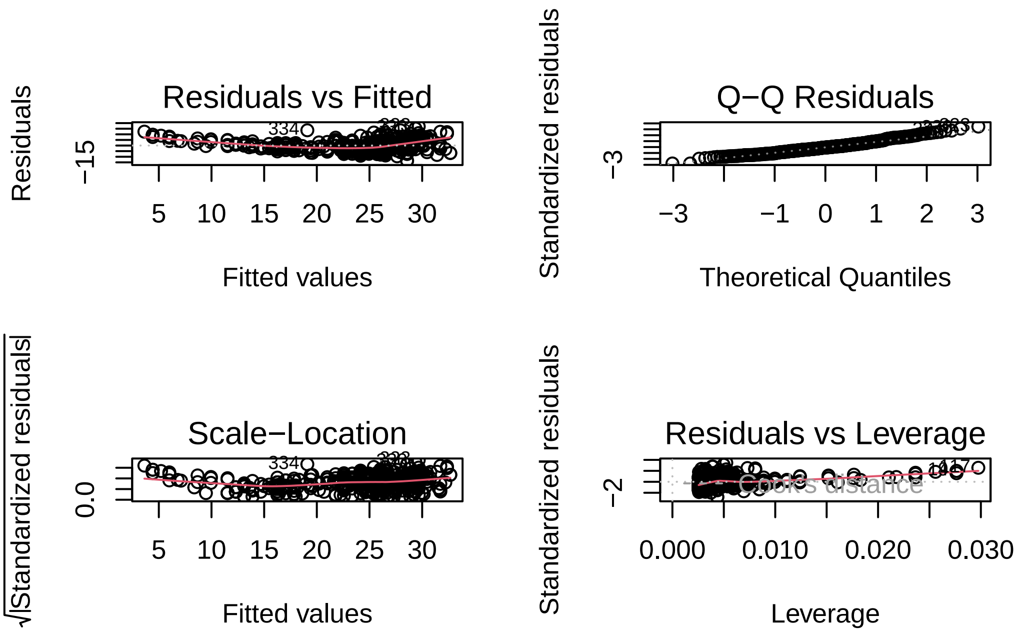
```
#|echo = true
attach(Auto)
plot(horsepower,mpg)
abline(auto.lm, col = "red")
```



(c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

#ANS

```
#|echo = true
par(mfrow = c(2,2))
plot(auto.lm)
```



There may not be a linear relationship between mpg and horsepower. There are some possible outliers in the Residuals vs. Leverage plot.

Problem 6

This question involves the use of multiple linear regression on the *Auto* data set.

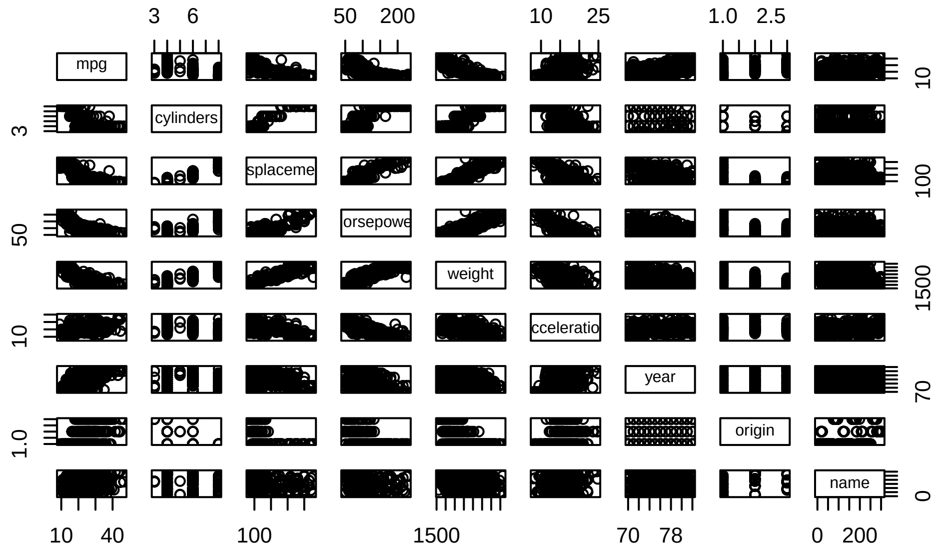
(a) Produce a scatterplot matrix which includes all of the variables in the data set.

#ANS:

```
#|echo = true
colnames(Auto)
```

```
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
[6] "acceleration" "year"         "origin"       "name"
```

```r
pairs(mpg~., data = Auto)
```



(b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the name variable, `cor()` which is qualitative.

#ANS

```r
#|echo = true
cor(Auto[, names(Auto) !="name"])
```

```
                    mpg  cylinders displacement horsepower     weight
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
```

```
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
year            0.2903161  1.0000000  0.1815277
origin          0.2127458  0.1815277  1.0000000
```

(c) Use the `lm()` function to perform a multiple linear regression with *mpg* as the response and all other variables except name as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

#ANS

```r
#|echo = true
auto = Auto[,1:8]
auto$origin = as.factor(auto$origin)
auto$cylinders = as.factor(auto$cylinders)
auto.lm = lm(mpg~.,data = auto)
summary(auto.lm)
```

```

Call:
lm(formula = mpg ~ ., data = auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6797 -1.9373 -0.0678  1.6711 12.7756 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.208e+01  4.541e+00  -4.862 1.70e-06 ***
cylinders4    6.722e+00  1.654e+00   4.064 5.85e-05 ***
cylinders5    7.078e+00  2.516e+00   2.813  0.00516 ** 
cylinders6    3.351e+00  1.824e+00   1.837  0.06701 .  
cylinders8    5.099e+00  2.109e+00   2.418  0.01607 *  
displacement  1.870e-02  7.222e-03   2.590  0.00997 ** 
horsepower   -3.490e-02  1.323e-02  -2.639  0.00866 ** 
weight       -5.780e-03  6.315e-04  -9.154  < 2e-16 ***
acceleration  2.598e-02  9.304e-02   0.279  0.78021    
year          7.370e-01  4.892e-02  15.064  < 2e-16 ***
origin2       1.764e+00  5.513e-01   3.200  0.00149 ** 
origin3       2.617e+00  5.272e-01   4.964 1.04e-06 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.098 on 380 degrees of freedom
Multiple R-squared:  0.8469,    Adjusted R-squared:  0.8425
F-statistic: 191.1 on 11 and 380 DF,  p-value: < 2.2e-16
```

```
i. Is there a relationship between the predictors and the response?
```

By the F-statistic at least one of the Bj is not zero due to p-value = 0. Therefore there is at least one predictor significant in relation to mpg.

```
ii. Which predictors appear to have a statistically significant relationship to the response?
```

By the T-test majority of the predictors have a statistically significant relationship to mpg, except cylinders6 and acceleration.

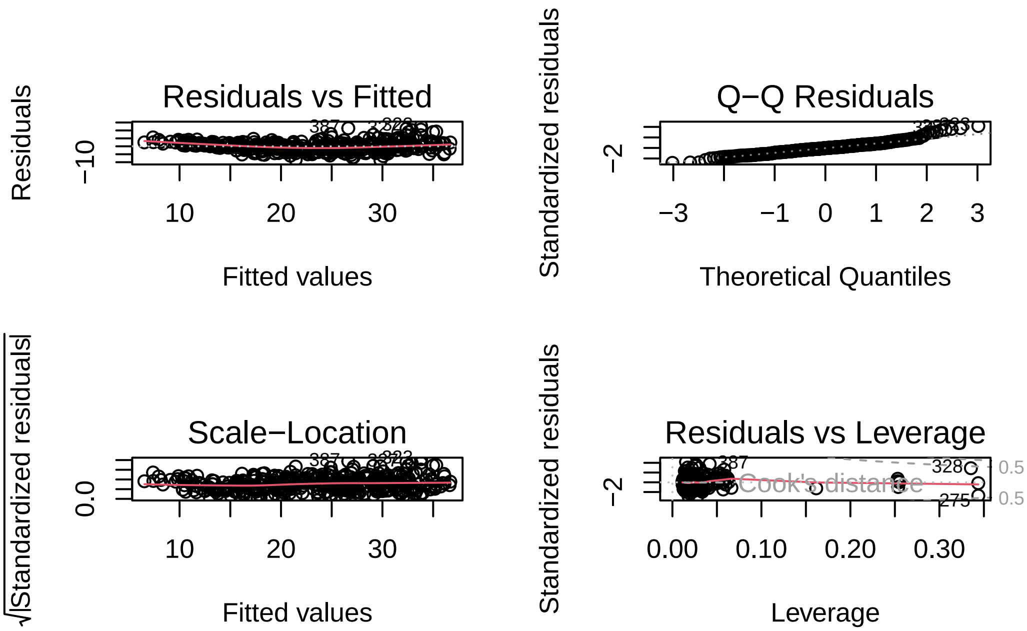
```
iii. What does the coefficient for the year variable suggest?
```

The coefficient for the year 7.370e-01 suggests for every additional year, mpg is estimated on average to increase by 7.370e-01 assuming all of the other predictors are fixed.

(d) Use the `plot()` function to produce diagnostic plots of the linear regression fit based on the predictors that appear to have a statistically significant relationship to the response. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

#ANS

```r
#|echo = true
auto2 = auto[,-6]
auto.lm2 = lm(mpg~., data = auto2)
par(mfrow=c(2,2))
plot(auto.lm2)
```



#The residual plots display possible outliers, particularly involving observations numbered 387, 327, and 323.

#The leverage plot identifies observations numbered 387, 328, and 275 as possible high leverage.

#The model exhibits a strong linear fit.

#Observation 387 is possibly a outlier and has high leverage.

e. Use the * and/or : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

ANS#

```
#|echo = true
auto3.lm = lm(mpg~cylinders * horsepower+ displacement* horsepower + weight* horsepower +
summary(auto3.lm)
```

```
Call:
lm(formula = mpg ~ cylinders * horsepower + displacement * horsepower +
    weight * horsepower + horsepower + year + origin, data = auto2)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.7296 -1.5516 -0.1096  1.3943 12.3188 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -3.204e+01  1.973e+01  -1.624  0.10519    
cylinders4               3.209e+01  1.953e+01   1.643  0.10115    
cylinders5               6.095e+01  2.158e+01   2.824  0.00500 ** 
cylinders6               2.924e+01  1.978e+01   1.478  0.14025    
cylinders8               3.404e+01  2.026e+01   1.680  0.09386 .  
horsepower               7.401e-02  1.971e-01   0.375  0.70754    
displacement            -2.023e-02  1.975e-02  -1.024  0.30649    
weight                  -8.250e-03  1.571e-03  -5.252 2.53e-07 ***
year                     7.418e-01  4.575e-02  16.215  < 2e-16 ***
origin2                  1.023e+00  5.256e-01   1.946  0.05240 .  
origin3                  1.639e+00  4.991e-01   3.285  0.00112 ** 
cylinders4:horsepower   -2.656e-01  1.965e-01  -1.352  0.17724    
cylinders5:horsepower   -5.945e-01  2.241e-01  -2.652  0.00834 ** 
cylinders6:horsepower   -2.452e-01  1.979e-01  -1.239  0.21627    
cylinders8:horsepower   -2.785e-01  2.007e-01  -1.388  0.16605    
horsepower:displacement  2.000e-04  1.289e-04   1.551  0.12180    
horsepower:weight        3.092e-05  1.115e-05   2.774  0.00581 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.81 on 375 degrees of freedom
Multiple R-squared:  0.8757,    Adjusted R-squared:  0.8704 
F-statistic: 165.1 on 16 and 375 DF,  p-value: < 2.2e-16
```

Yes, the interaction between horsepower and weight, cylinders5 and horsepower seems significant.

(f) Try a few different transformations of the variables, such as *log(X)*, $\sqrt{X}$, X^2. Comment on your findings.

#ANS

```
#|echo = true
    auto4.lm = lm(mpg~cylinders * log(horsepower)+ displacement* horsepower^2 + weight* sq
    summary(auto4.lm)
```

```
Call:
lm(formula = mpg ~ cylinders * log(horsepower) + displacement *
    horsepower^2 + weight * sqrt(horsepower) + horsepower + year +
    origin, data = auto2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7188 -1.5062 -0.0928  1.3959 12.4963

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -9.819e+01  1.212e+02  -0.810  0.41822
cylinders4                  1.264e+02  9.091e+01   1.391  0.16514
cylinders5                  2.547e+02  9.932e+01   2.564  0.01074 *
cylinders6                  1.187e+02  9.132e+01   1.300  0.19443
cylinders8                  1.318e+02  9.371e+01   1.406  0.16052
log(horsepower)             1.194e+01  5.156e+01   0.232  0.81705
displacement               -2.686e-02  2.230e-02  -1.204  0.22923
horsepower                 -2.932e-01  4.764e-01  -0.615  0.53861
weight                     -1.161e-02  2.935e-03  -3.956 9.12e-05 ***
sqrt(horsepower)            4.763e+00  1.872e+01   0.254  0.79932
year                        7.451e-01  4.618e-02  16.135  < 2e-16 ***
origin2                     9.632e-01  5.468e-01   1.761  0.07899 .
origin3                     1.581e+00  5.073e-01   3.118  0.00197 **
cylinders4:log(horsepower) -2.625e+01  1.979e+01  -1.326  0.18562
cylinders5:log(horsepower) -5.499e+01  2.177e+01  -2.526  0.01194 *
cylinders6:log(horsepower) -2.474e+01  1.985e+01  -1.246  0.21362
cylinders8:log(horsepower) -2.735e+01  2.032e+01  -1.345  0.17929
displacement:horsepower     2.489e-04  1.501e-04   1.658  0.09813 .
weight:sqrt(horsepower)     6.572e-04  2.550e-04   2.578  0.01033 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.817 on 373 degrees of freedom
Multiple R-squared:  0.8757,    Adjusted R-squared:  0.8697
F-statistic:   146 on 18 and 373 DF,  p-value: < 2.2e-16
```

Using different transformations of the variables, the R^2 is actually the same while the Adjusted R^2 increased a little bit. The interaction of horsepower and weight, cylinders5 and horsepower is still statistically significant.

Problem 7

This problem involves the `Boston` data set, from the `ISLR2` package. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

#ANS

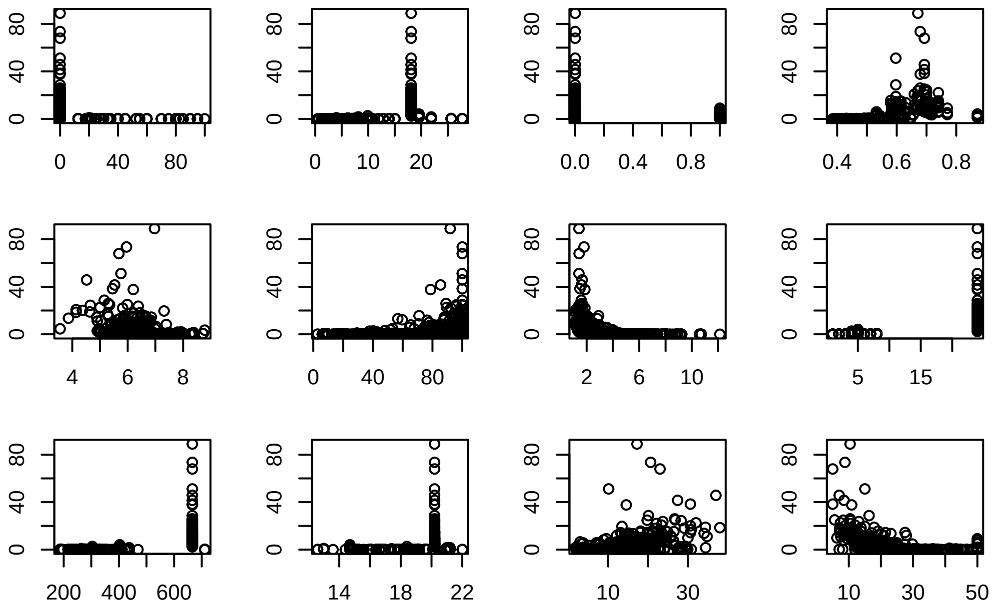
```r
#|echo = true
    # Create a data frame to store results
    results_df <- data.frame(Predictor = character(0), Coefficient = numeric(0), P_Value =

    # Iterate through predictor variables
    for (predictor in colnames(Boston)[-1]) {
      # Fit the linear regression model
      lm.fit <- lm(crim ~ ., data = Boston[, c(predictor, "crim")])

      # Extract the coefficient for the predictor and its p-value
      coef_predictor <- coef(lm.fit)[predictor]
      p_value <- summary(lm.fit)$coefficients[predictor, 4]

      # Add the results to the data frame
      results_df <- rbind(results_df, data.frame(Predictor = predictor, Coefficient = coef
    }

    # Print the results
    print(results_df)
```

```
        Predictor Coefficient      P_Value
zn             zn -0.07393498 5.506472e-06
indus       indus  0.50977633 1.450349e-21
chas         chas -1.89277655 2.094345e-01
nox           nox 31.24853120 3.751739e-23
rm             rm -2.68405122 6.346703e-07
age           age  0.10778623 2.854869e-16
dis           dis -1.55090168 8.519949e-19
rad           rad  0.61791093 2.693844e-56
tax           tax  0.02974225 2.357127e-47
ptratio   ptratio  1.15198279 2.942922e-11
```

```
lstat        lstat  0.54880478 2.654277e-27
medv          medv -0.36315992 1.173987e-19
```

The predictor variable model that appears to lack significance in explaining the per capita crime rate is whether the suburb borders the Charles River. In contrast, all other predictor variables appear to have a significant impact on the per capita crime rate.

```r
#|echo = true
# Set the layout for multiple plots
par(mfrow = c(3, 4))
# Adjust the outer margins
par(oma = c(0.5, 0.5, 0.5, 0.5))  # Adjust outer margins as needed

# Adjust the inner margins
par(mar = c(2, 2, 2, 2))  # Adjust inner margins as needed

for (i in 1:12) {
  plot(Boston[, i + 1], Boston$crim, xlab = "", ylab = "Crim")
}
```



```r
# Reset par settings to the default values after creating the plots
# par(mfrow = c(1, 1))
# par(oma = c(0, 0, 0, 0))  # Reset outer margins to default
# par(mar = c(5.1, 4.1, 4.1, 2.1))  # Reset inner margins to default
```

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

#ANS

```r
#|echo = true
fit.lm = lm(crim ~ ., data = Boston)
summary(fit.lm)
```

```

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-8.534 -2.248 -0.348  1.087 73.923 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.7783938  7.0818258   1.946 0.052271 .  
zn           0.0457100  0.0187903   2.433 0.015344 *  
indus       -0.0583501  0.0836351  -0.698 0.485709    
chas        -0.8253776  1.1833963  -0.697 0.485841    
nox         -9.9575865  5.2898242  -1.882 0.060370 .  
rm           0.6289107  0.6070924   1.036 0.300738    
age         -0.0008483  0.0179482  -0.047 0.962323    
dis         -1.0122467  0.2824676  -3.584 0.000373 ***
rad          0.6124653  0.0875358   6.997 8.59e-12 ***
tax         -0.0037756  0.0051723  -0.730 0.465757    
ptratio     -0.3040728  0.1863598  -1.632 0.103393    
lstat        0.1388006  0.0757213   1.833 0.067398 .  
medv        -0.2200564  0.0598240  -3.678 0.000261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.46 on 493 degrees of freedom
Multiple R-squared:  0.4493,    Adjusted R-squared:  0.4359
```

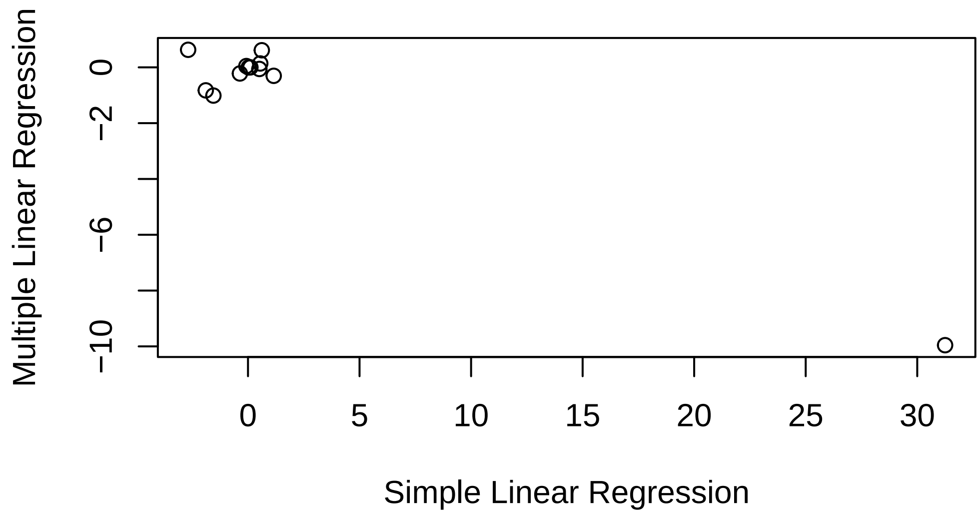
```
F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

It appears that the predictors zn, dis, rad, and medv are significant in predicting crim.

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

#ANS

```r
#|echo = true
# Load the Boston dataset if it's not already loaded
if (!exists("Boston", envir = .GlobalEnv)) {
  library(MASS)
  data(Boston)
}

# Create an empty vector to store the coefficients
b.boston <- numeric(length = ncol(Boston) - 1)

# Iterate through predictor variables
for (i in 1:(ncol(Boston) - 1)) {
  lm.fit <- lm(crim ~ Boston[, i + 1], data = Boston)
  b.boston[i] <- lm.fit$coef[2]
}

# Fit a multiple linear regression model
lm.fit.multi <- lm(crim ~ ., data = Boston)

# Plot the coefficients
plot(b.boston, coef(lm.fit.multi)[-1], xlab = "Simple Linear Regression", ylab = "Mul
```



(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

#ANS

```
lm.zn = lm(crim~poly(zn,3),data=Boston)
summary(lm.zn)
```

```
Call:
lm(formula = crim ~ poly(zn, 3), data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-4.821 -4.614 -1.294  0.473 84.130 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    3.6135     0.3722   9.709  < 2e-16 ***
poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
```

```
poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 **
poly(zn, 3)3 -10.0719     8.3722  -1.203  0.22954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.372 on 502 degrees of freedom
Multiple R-squared:  0.05824,   Adjusted R-squared:  0.05261
F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
```

```r
lm.nox = lm(crim~poly(nox,3),data=Boston)
summary(lm.nox)
```

```
Call:
lm(formula = crim ~ poly(nox, 3), data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.110 -2.068 -0.255  0.739 78.302

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.6135     0.3216  11.237  < 2e-16 ***
poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.234 on 502 degrees of freedom
Multiple R-squared:  0.297, Adjusted R-squared:  0.2928
F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
lm.dis = lm(crim~poly(dis,3),data=Boston)
summary(lm.dis)
```

```
Call:
lm(formula = crim ~ poly(dis, 3), data = Boston)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-10.757  -2.588   0.031   1.267  76.378 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.6135     0.3259  11.087  < 2e-16 ***
poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.331 on 502 degrees of freedom
Multiple R-squared:  0.2778,	Adjusted R-squared:  0.2735 
F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
lm.rad = lm(crim~poly(rad,3),data=Boston)
summary(lm.rad)
```

```
Call:
lm(formula = crim ~ poly(rad, 3), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.381  -0.412  -0.269   0.179  76.217 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.6135     0.2971  12.164  < 2e-16 ***
poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 ** 
poly(rad, 3)3   4.6985     6.6824   0.703  0.48231    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.682 on 502 degrees of freedom
Multiple R-squared:    0.4,	Adjusted R-squared:  0.3965 
F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
lm.ptratio = lm(crim~poly(ptratio,3), data=Boston)
summary(lm.ptratio)
```

```
Call:
lm(formula = crim ~ poly(ptratio, 3), data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-6.833 -4.146 -1.655  1.408 82.697 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.614      0.361  10.008  < 2e-16 ***
poly(ptratio, 3)1   56.045      8.122   6.901 1.57e-11 ***
poly(ptratio, 3)2   24.775      8.122   3.050  0.00241 ** 
poly(ptratio, 3)3  -22.280      8.122  -2.743  0.00630 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.122 on 502 degrees of freedom
Multiple R-squared:  0.1138,	Adjusted R-squared:  0.1085 
F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13
```

```r
lm.lstat = lm(crim~poly(lstat,3),data=Boston)
summary(lm.lstat)
```

```
Call:
lm(formula = crim ~ poly(lstat, 3), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.234  -2.151  -0.486   0.066  83.353 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.6135     0.3392  10.654   <2e-16 ***
poly(lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
poly(lstat, 3)2  15.8882     7.6294   2.082   0.0378 *  
```

```
poly(lstat, 3)3 -11.5740     7.6294  -1.517   0.1299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.629 on 502 degrees of freedom
Multiple R-squared:  0.2179,    Adjusted R-squared:  0.2133
F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
lm.medv = lm(crim~poly(medv,3), data=Boston)
summary(lm.medv)
```

```
Call:
lm(formula = crim ~ poly(medv, 3), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-24.427  -1.976  -0.437   0.439  73.655

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.614      0.292  12.374  < 2e-16 ***
poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
poly(medv, 3)3  -48.033      6.569  -7.312 1.05e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.569 on 502 degrees of freedom
Multiple R-squared:  0.4202,    Adjusted R-squared:  0.4167
F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
lm.indus = lm(crim ~ poly(indus,3), data = Boston)
summary(lm.indus)
```

```
Call:
lm(formula = crim ~ poly(indus, 3), data = Boston)

Residuals:
```

```
   Min     1Q Median     3Q    Max 
-8.278 -2.514  0.054  0.764 79.713 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3.614      0.330  10.950  < 2e-16 ***
poly(indus, 3)1   78.591      7.423  10.587  < 2e-16 ***
poly(indus, 3)2  -24.395      7.423  -3.286  0.00109 ** 
poly(indus, 3)3  -54.130      7.423  -7.292  1.2e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.423 on 502 degrees of freedom
Multiple R-squared:  0.2597,	Adjusted R-squared:  0.2552 
F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
lm.rm = lm(crim ~ poly(rm,3), data = Boston)
summary(lm.rm)
```

```
Call:
lm(formula = crim ~ poly(rm, 3), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.485  -3.468  -2.221  -0.015  87.219 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    3.6135     0.3703   9.758  < 2e-16 ***
poly(rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
poly(rm, 3)2  26.5768     8.3297   3.191  0.00151 ** 
poly(rm, 3)3  -5.5103     8.3297  -0.662  0.50858    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.33 on 502 degrees of freedom
Multiple R-squared:  0.06779,	Adjusted R-squared:  0.06222 
F-statistic: 12.17 on 3 and 502 DF,  p-value: 1.067e-07
```

```r
lm.age = lm(crim ~ poly(age,3), data = Boston)
summary(lm.age)
```

```

Call:
lm(formula = crim ~ poly(age, 3), data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-9.762 -2.673 -0.516  0.019 82.842 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.6135     0.3485  10.368  < 2e-16 ***
poly(age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
poly(age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
poly(age, 3)3  21.3532     7.8397   2.724  0.00668 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.84 on 502 degrees of freedom
Multiple R-squared:  0.1742,	Adjusted R-squared:  0.1693 
F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
lm.tax = lm(crim ~ poly(tax,3), data = Boston)
summary(lm.tax)
```

```

Call:
lm(formula = crim ~ poly(tax, 3), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.273  -1.389   0.046   0.536  76.950 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.6135     0.3047  11.860  < 2e-16 ***
poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
```

```
poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.854 on 502 degrees of freedom
Multiple R-squared:  0.3689,	Adjusted R-squared:  0.3651
F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

It appears that the medv predictor may exhibit a more pronounced non-linear relationship, whereas the others exhibit considerably smaller R-squared values.

Problem 8

This problem focuses on the **collinearity** problem.

(a) Perform the following commands in R:

```r
set.seed (1)
x1=runif (100)
x2 =0.5* x1+rnorm (100) /10
y=2+2* x1 +0.3* x2+rnorm (100)
```

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

#ANS

Linear Model: y = B0 + B1x1 + B2x2 + e

Regression coefficients: B0 = 2, B1 = 2, B2 = 0.3

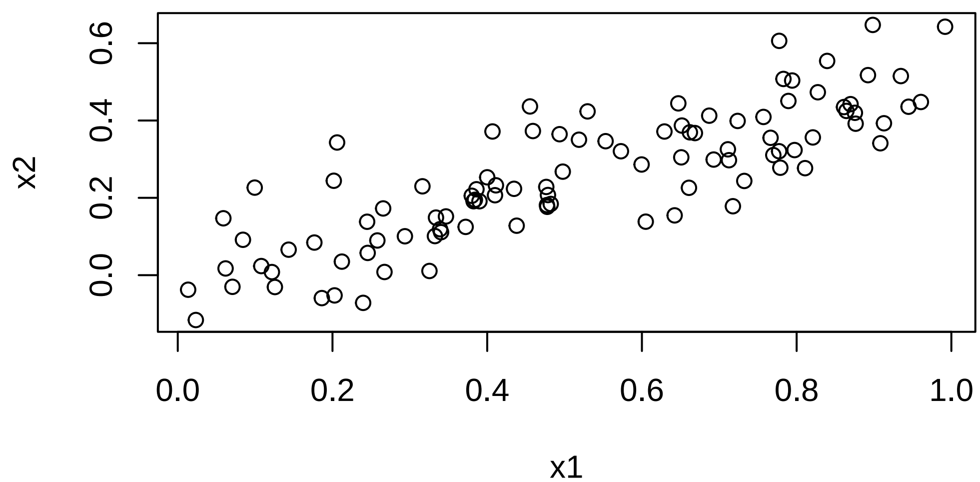
(b) What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

#ANS

```r
#|echo = true
cor(x1,x2)
```

```
[1] 0.8351212
```

```r
#|echo = true
plot(x1,x2)
```



(c) Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

#ANS

```r
#|echo = true
summary(lm(y ~ x1 + x2))
```

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8311 -0.7273 -0.0537  0.6338  2.3359 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *
x2            1.0097     1.1337   0.891   0.3754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared:  0.2088,    Adjusted R-squared:  0.1925
F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

#Ho: B1= B2 = 0

#Ha: At least one is Bj != 0 (At least one predictor is significant)

By the F-statistic, the p-value is 1.164e-05 which is close to 0. So at least one of the predictors x1,x2 is significant to y.

B0 = 2.1305

B1 = 1.4396

B2 = 1.0097

Derived from the true values of B0, B1, and B2, this estimation closely approximates B0 and moderately approximates B1, but it does not closely approximate B2.

By testing H0 : B1 = 0 we reject the null hypothesis with a p-value = 0.0487 < 0.05.

By testing H0 : B2 = 0 we fail to reject the null hypothesis with a p-value = 0.3754 > 0.05.

(d) Now fit a least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

#ANS

```r
#|echo = true
summary(lm(y~x1))
```

```
Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-2.89495 -0.66874 -0.07785  0.59221  2.45560

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,	Adjusted R-squared:  0.1942
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

The estimated coefficients B0 and B1 closely approximate the true coefficients. By examining the T-statistic and the associated p-value for predictor x1, we find that the p-value is very close to 0. This provides strong evidence to reject the null hypothesis that B1 = 0.

(e) Now fit a least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

#ANS

```
#|echo = true
summary(lm(y~x2))
```

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.62687 -0.75156 -0.03598  0.72383  2.44890 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,	Adjusted R-squared:  0.1679
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

The estimate coefficients of B2 is greater the the actual B2. By examining the T-statistic and the associated p-value for predictor x2, we find that the p-value is very close to 0. This provides strong evidence to reject the null hypothesis that B2 = 0.

f. Do the results obtained in (c)–(e) contradict each other? Explain your answer.

   #ANS

   I wouldn't say it really contradicts each other. What (c) implies is that if x1 is included in the model as a predictor for y, there is no need for x2. This holds true because x2 was derived or calculated based on the values of x1.

g. Now suppose we obtain one additional observation, which was unfortunately miss-measured.

```
x1=c(x1 , 0.1)
x2=c(x2 , 0.8)
y=c(y,6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

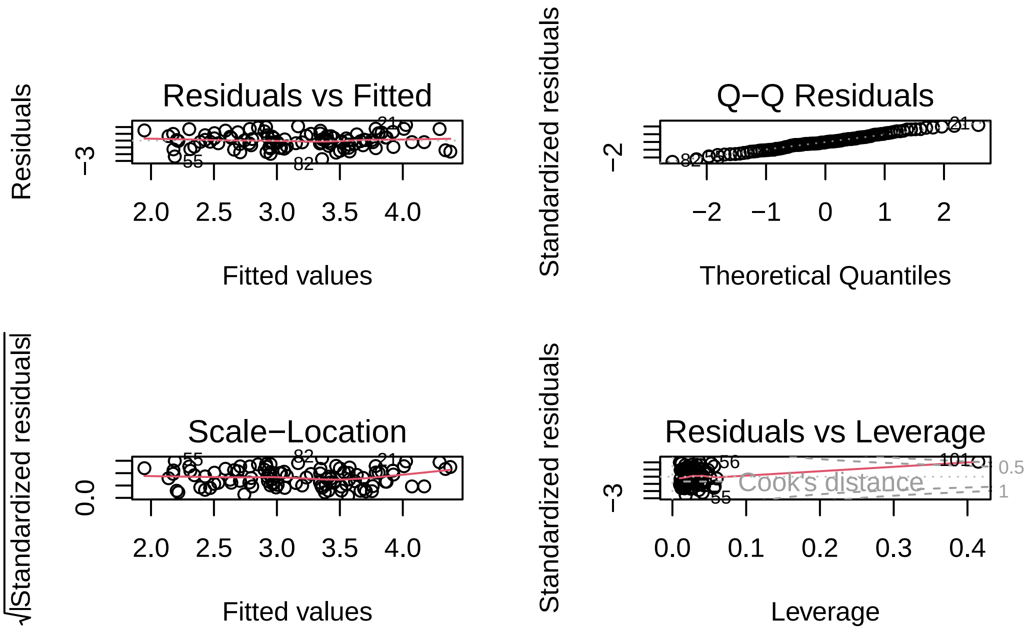
#ANS

```
#|echo = true
summary(lm(y~x1+x2))
```

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.73348 -0.69318 -0.05263  0.66385  2.30619 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1            0.5394     0.5922   0.911  0.36458    
x2            2.5146     0.8977   2.801  0.00614 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared:  0.2188,    Adjusted R-squared:  0.2029
F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```r
#|echo = true
par(mfrow = c(2,2))
plot(lm(y~x1+x2))
```



The effect of the new observation shows:

 # Changed the estimations of B0, B1, and B2, resulting in a increase of both the multiple and adjusted R^2 values.

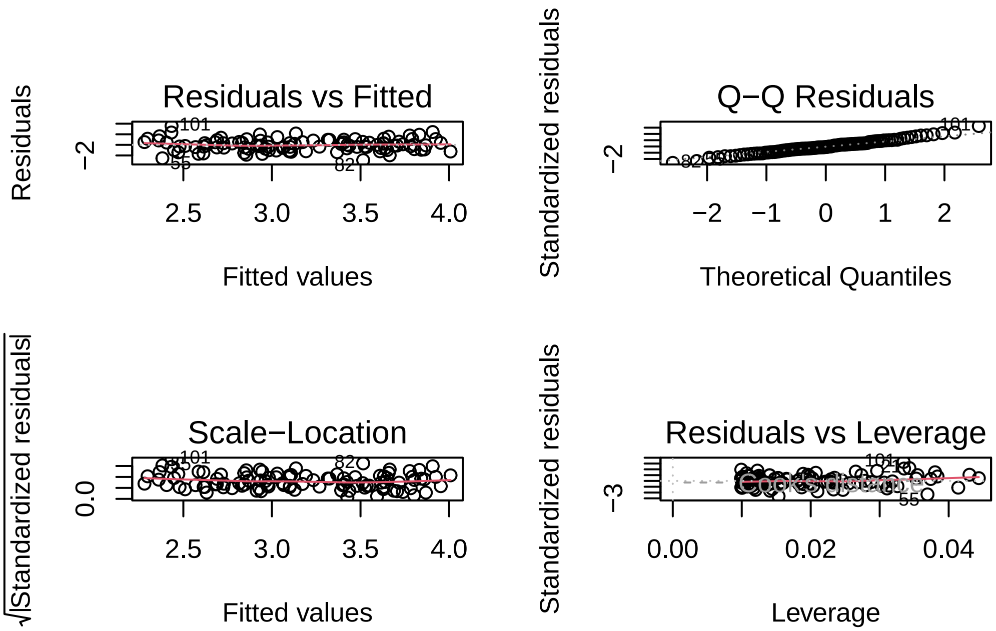
 # Notably, when testing the null hypothesis H0: B1 = 0, we now fail to reject it, as the p-value is 0.36458, exceeding the significance level of 0.05. Conversely, the null hypothesis H0: B2 = 0 is rejected with a p-value of 0.00614, which is less than 0.05.

 # The plots highlights that observations 101 is not a outlier possess notably high leverage.

```r
#|echo = true
summary(lm(y~x1))
```

```
Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8897 -0.6556 -0.0909  0.5682  3.5665 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
x1            1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,	Adjusted R-squared:  0.1477 
F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```r
#|echo = true
par(mfrow = c(2,2))
plot(lm(y~x1))
```



The effect of the new observation shows:

#Changed the estimates of B0 and B1, resulting in a decrease of both the multiple and adjusted R^2 values.

#It has a higher max value: 3.5665

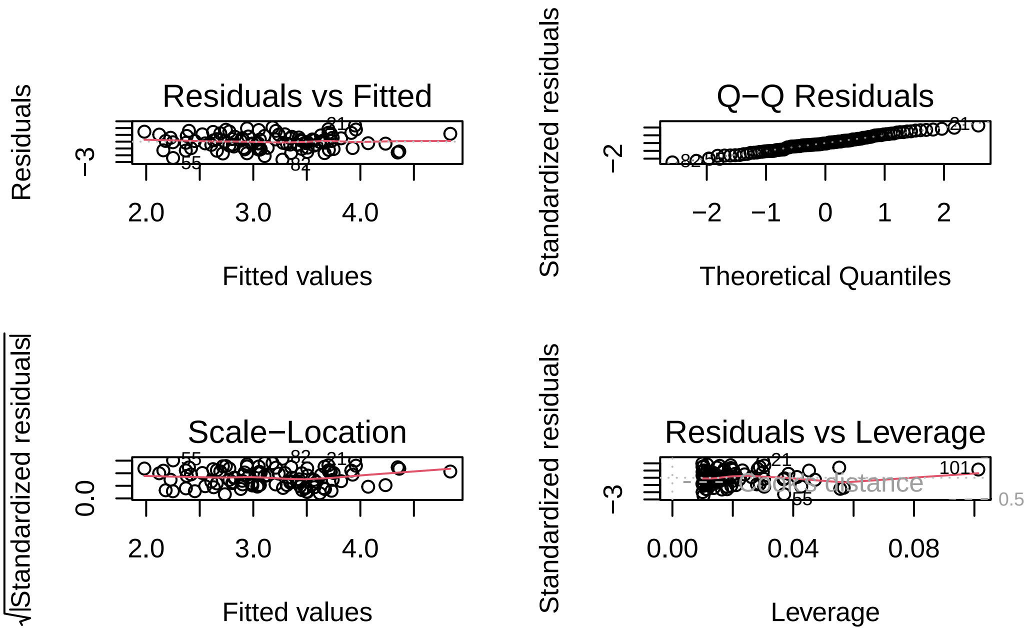
 # The plots highlights that observation 101 is both possibly outlier and a high-leverage point..

```
#|echo = true
summary(lm(y~x2))
```

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.64729 -0.71021 -0.06899  0.72699  2.38074 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
x2            3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2042
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

```r
#|echo = true
par(mfrow = c(2,2))
plot(lm(y~x2))
```



The effect of the new observation shows:

#Changed the estimates of B0 and B1 resulting in a increase of both the multiple and adjusted R^2 values.

#The median is lower with -0.06899

 # The plots highlights that observation 101 is possibly a high-leverage point.

Problem 9

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

```
#|echo = true
set.seed(1)
X = rnorm(100)
e = rnorm(100)
```

(b) Generate a response vector Y of length n = 100 according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

where β_0, β_1, β_2, and β_3 are constants of your choice.

```
#|echo = true
B0 = 2
B1 = 3
B2 = -4
B3 = 5
Y = B0 + B1 * X + B2 * X^2 + B3 * X^3 + e
```

(c) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both X and Y.

```
#|echo = true
library(leaps)
new.data = data.frame(cbind(Y,X))
fit.y = regsubsets(Y ~ poly(X,10),data = new.data)
(fit.res = summary(fit.y))
```

```
Subset selection object
Call: regsubsets.formula(Y ~ poly(X, 10), data = new.data)
10 Variables  (and intercept)
              Forced in Forced out
poly(X, 10)1      FALSE      FALSE
```

```
poly(X, 10)2      FALSE      FALSE
poly(X, 10)3      FALSE      FALSE
poly(X, 10)4      FALSE      FALSE
poly(X, 10)5      FALSE      FALSE
poly(X, 10)6      FALSE      FALSE
poly(X, 10)7      FALSE      FALSE
poly(X, 10)8      FALSE      FALSE
poly(X, 10)9      FALSE      FALSE
poly(X, 10)10     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)4 poly(X, 10)5
1  ( 1 ) "*"          " "          " "          " "          " "
2  ( 1 ) "*"          " "          "*"          " "          " "
3  ( 1 ) "*"          "*"          "*"          " "          " "
4  ( 1 ) "*"          "*"          "*"          " "          "*"
5  ( 1 ) "*"          "*"          "*"          "*"          "*"
6  ( 1 ) "*"          "*"          "*"          "*"          "*"
7  ( 1 ) "*"          "*"          "*"          "*"          "*"
8  ( 1 ) "*"          "*"          "*"          "*"          "*"
         poly(X, 10)6 poly(X, 10)7 poly(X, 10)8 poly(X, 10)9 poly(X, 10)10
1  ( 1 ) " "          " "          " "          " "          " "
2  ( 1 ) " "          " "          " "          " "          " "
3  ( 1 ) " "          " "          " "          " "          " "
4  ( 1 ) " "          " "          " "          " "          " "
5  ( 1 ) " "          " "          " "          " "          " "
6  ( 1 ) " "          " "          " "          " "          "*"
7  ( 1 ) " "          "*"          " "          " "          "*"
8  ( 1 ) " "          "*"          " "          "*"          "*"
```

```r
#|echo = true
fit.stat = cbind(fit.res$adjr2,fit.res$cp,fit.res$bic)
colnames(fit.stat) = c("Adjr2","Cp","BIC")
print(fit.stat)
```

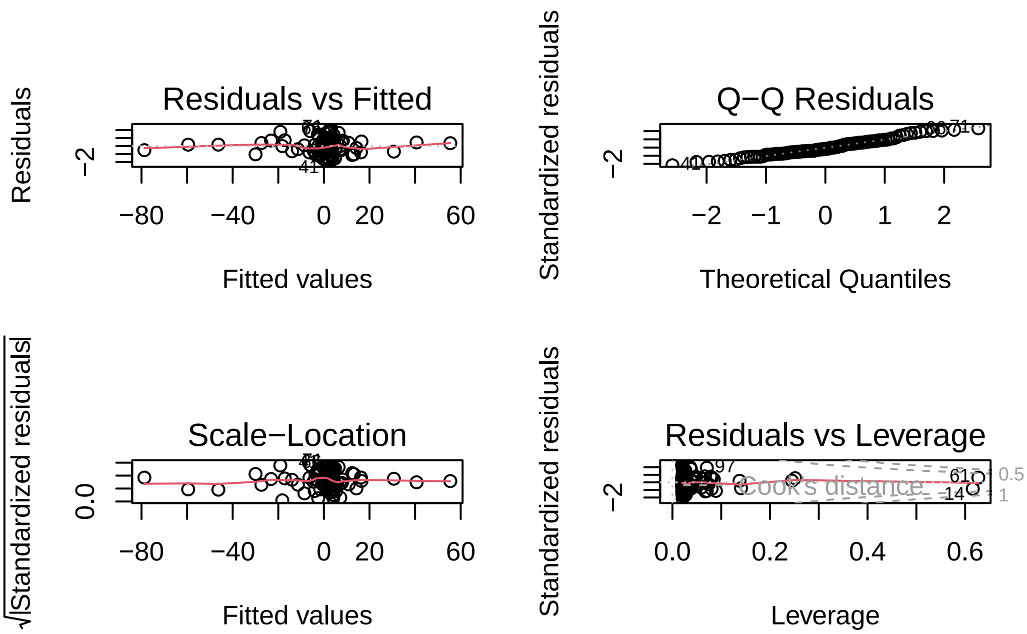
```
         Adjr2          Cp       BIC
[1,] 0.7112875 6988.565358 -116.0373
[2,] 0.9543640 1014.410886 -296.9312
[3,] 0.9960817    2.185943 -538.8672
[4,] 0.9961380    1.866261 -536.7558
[5,] 0.9961680    2.193128 -533.9887
[6,] 0.9961679    3.235128 -530.4513
```

```
[7,] 0.9961313    5.119994 -525.9753
[8,] 0.9960928    7.027330 -521.4741
```

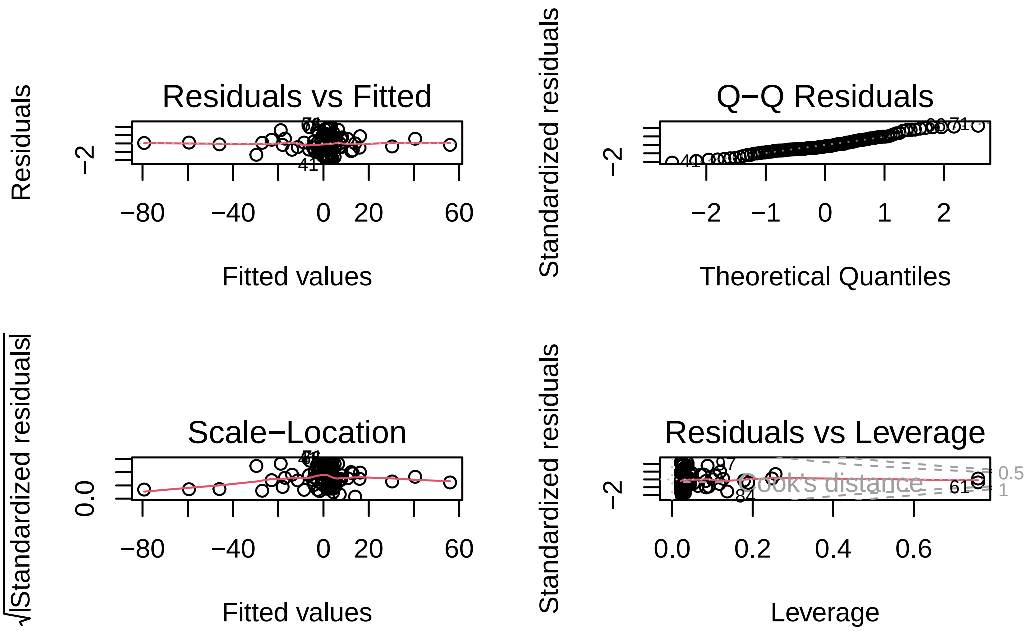
```r
summary(lm(Y ~ poly(X,4), data = new.data ))$coef
```

```
                 Estimate Std. Error    t value      Pr(>|t|)
(Intercept)    0.09982912 0.09590514   1.040915  3.005568e-01
poly(X, 4)1 129.31301080 0.95905143 134.834281 2.608371e-110
poly(X, 4)2 -30.95065970 0.95905143 -32.272158  5.406891e-53
poly(X, 4)3  75.13004937 0.95905143  78.337873  4.100697e-88
poly(X, 4)4   1.25709501 0.95905143   1.310769  1.930956e-01
```

```r
#|echo = true
par(mfrow = c(2,2))
plot(lm(Y ~ poly(X,4)))
```




```r
#|echo = true
par(mfrow = c(2,2))
plot(lm(Y ~ poly(X,5)))
```



The model with the fourth-degree polynomial appears to be the optimal choice among the subsets.

(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

```r
#|echo = true
step(lm(Y ~ poly(X,10)), direction = "backward")
```

```
Start:  AIC=4.64
Y ~ poly(X, 10)

              Df Sum of Sq     RSS    AIC
<none>                        84.1   4.64
- poly(X, 10) 10     23329 23413.3 547.59


Call:
lm(formula = Y ~ poly(X, 10))

Coefficients:
  (Intercept)   poly(X, 10)1   poly(X, 10)2   poly(X, 10)3   poly(X, 10)4
```

```
     0.09983      129.31301      -30.95066       75.13005        1.25710
 poly(X, 10)5   poly(X, 10)6   poly(X, 10)7   poly(X, 10)8   poly(X, 10)9
     1.48019        0.11900       -0.32977       -0.10795       -0.29584
poly(X, 10)10
    -0.95123
```

```
#|echo = true
step(lm(Y ~ poly(X,10)), direction = "forward")
```

```
Start:  AIC=4.64
Y ~ poly(X, 10)



Call:
lm(formula = Y ~ poly(X, 10))

Coefficients:
  (Intercept)   poly(X, 10)1   poly(X, 10)2   poly(X, 10)3   poly(X, 10)4
     0.09983      129.31301      -30.95066       75.13005        1.25710
 poly(X, 10)5   poly(X, 10)6   poly(X, 10)7   poly(X, 10)8   poly(X, 10)9
     1.48019        0.11900       -0.32977       -0.10795       -0.29584
poly(X, 10)10
    -0.95123
```

Compared to the results in (C), both forward and backward show that all of the terms is used in the regression.