Chapter 3

R Lab

```
boston <- Boston
head(boston)
```

```
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
  medv
1 24.0
2 21.6
3 34.7
4 33.4
5 36.2
6 28.7
```

```
names(boston)
```

```
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
 [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

Simple Linear Regression

```
summary(lm.fit <- lm(medv ~ lstat, data = boston))
```

```

Call:
lm(formula = medv ~ lstat, data = boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,	Adjusted R-squared:  0.5432 
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
coef(lm.fit)
```

```
(Intercept)       lstat 
 34.5538409  -0.9500494 
```

```r
confint(lm.fit)
```

```
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```

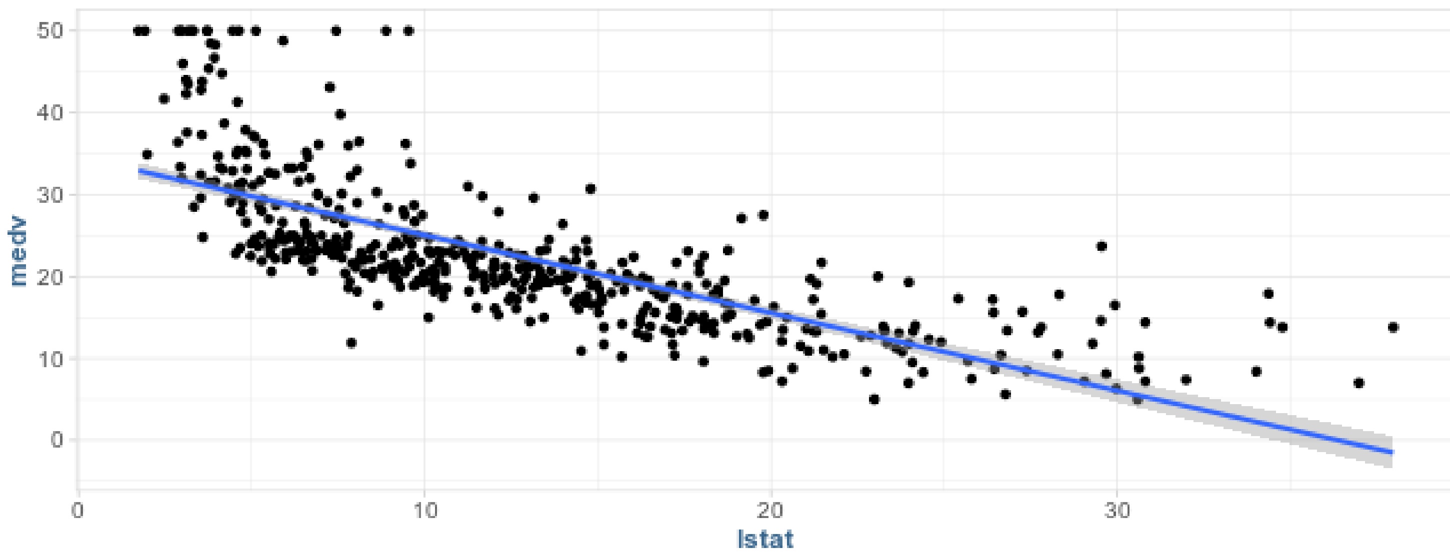
```r
predict(lm.fit, data.frame(lstat = c(5, 10, 15)),
        interval = "confidence")
```

```
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
```

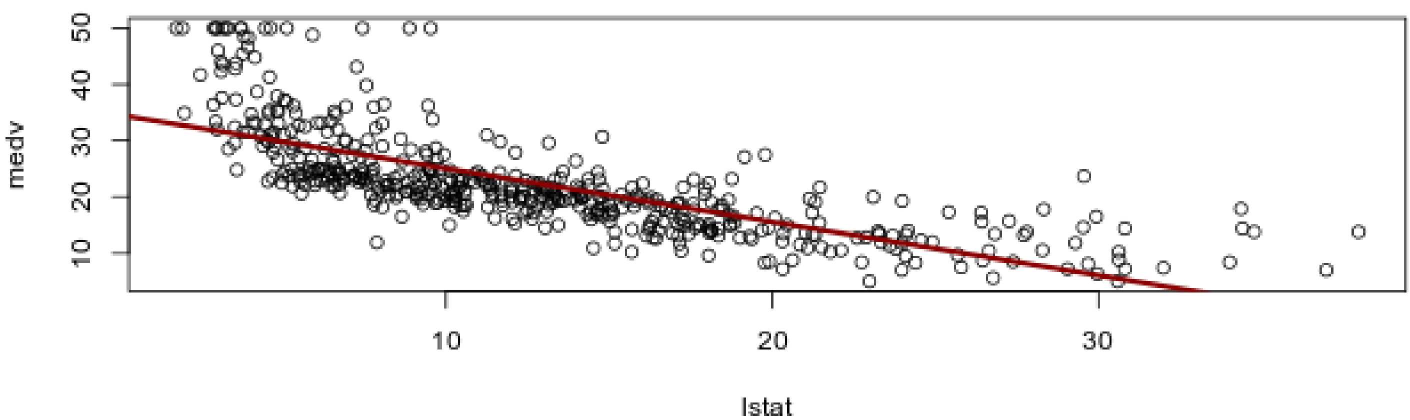
```r
predict(lm.fit, data.frame(lstat = c(5, 10, 15)),
        interval = "prediction")
```

```
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

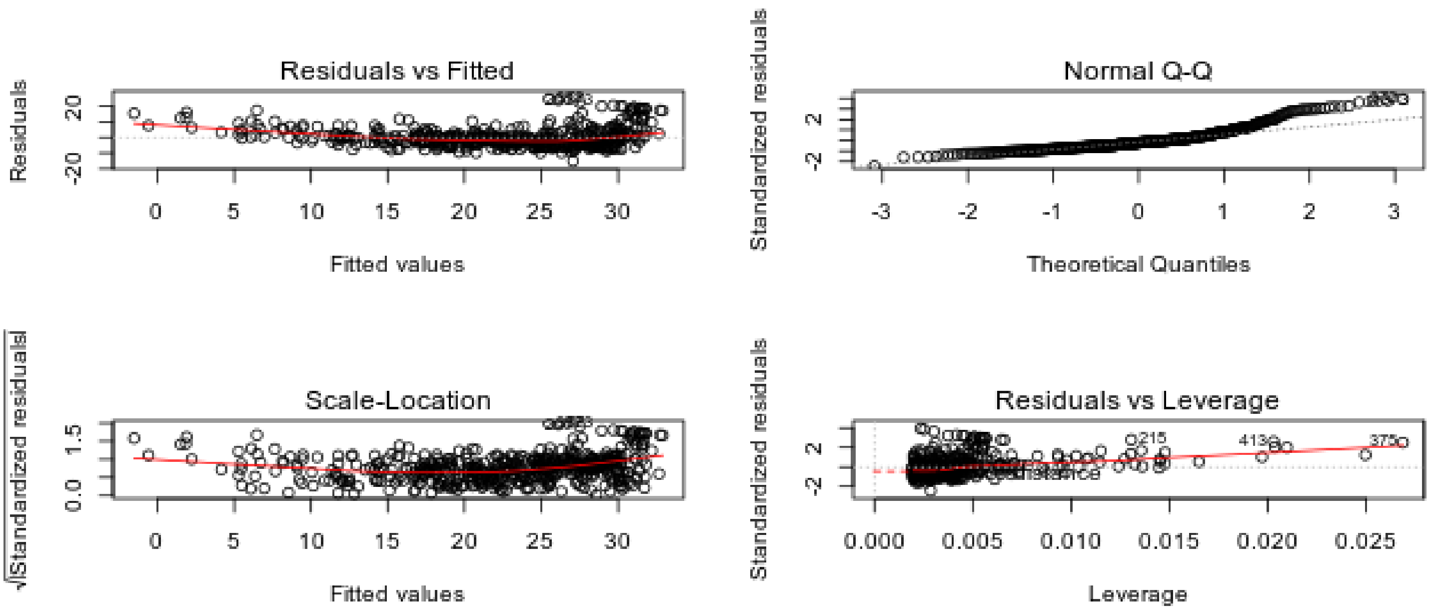
```r
ggplot(boston, aes(lstat, medv)) +
   geom_point() +
   geom_smooth(method = "lm")
```



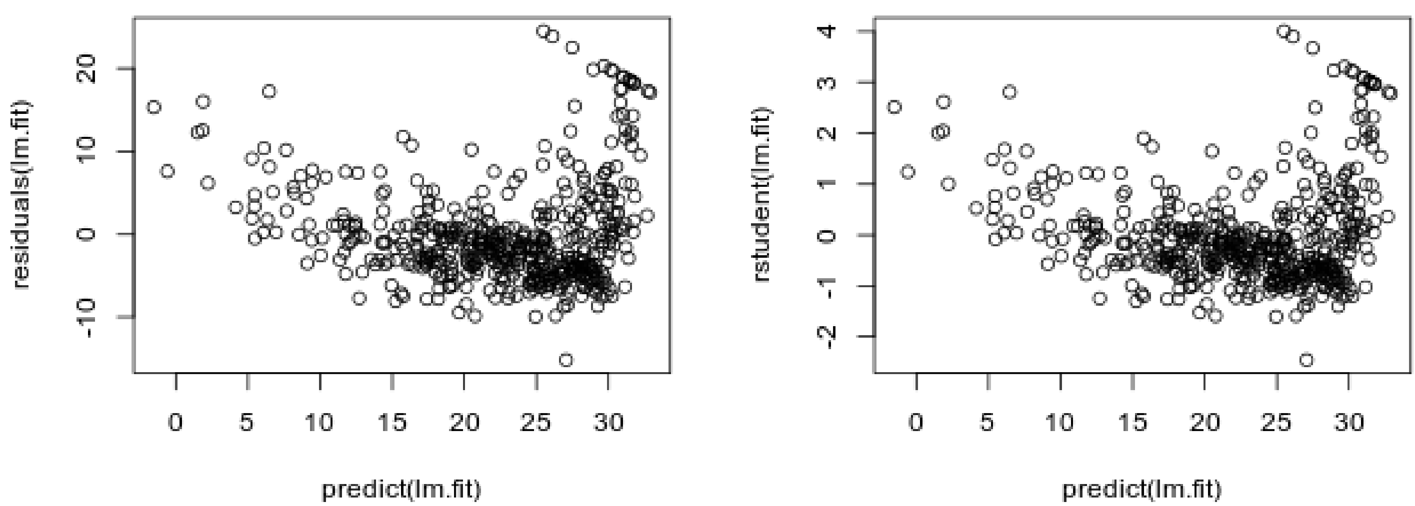

```
with(boston, {
   plot(lstat, medv)
   abline(lm.fit, col ="darkred")
   abline(lm.fit, lwd = 3, col = "darkred")
})
```



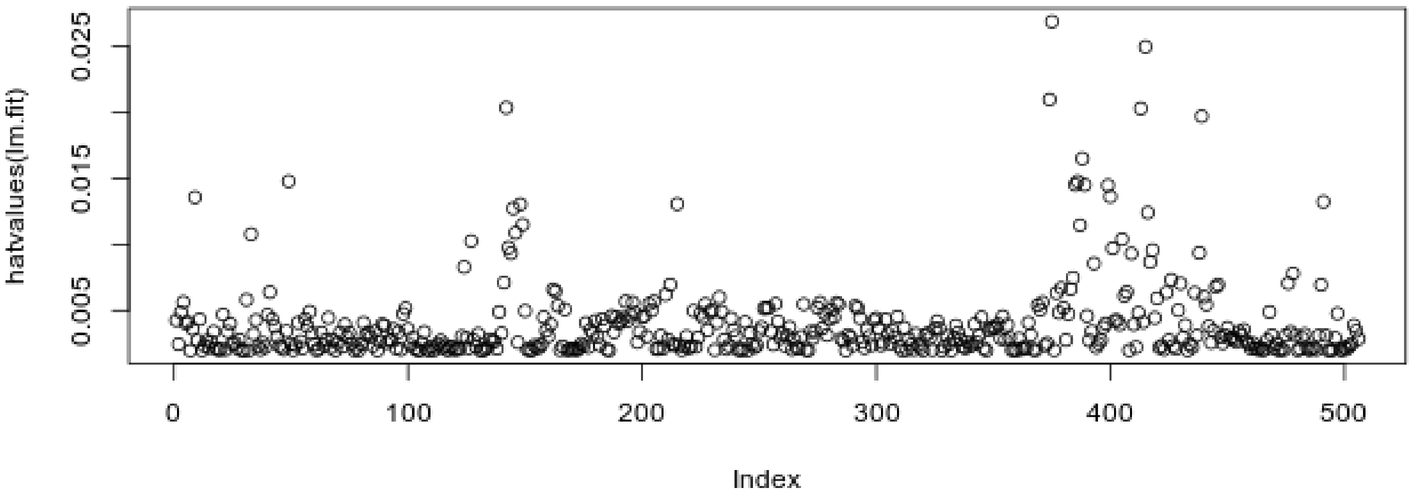

```
par(mfrow = c(2,2))
with(boston, {
   plot(lm.fit)
})
```




```r
par(mfrow = c(1,2))
plot(predict(lm.fit), residuals(lm.fit))
plot(predict(lm.fit), rstudent(lm.fit))
```




```r
par(mfrow = c(1,1))
plot(hatvalues(lm.fit))
```




```r
which.max(hatvalues(lm.fit))
```

```
375
375
```

Multiple Liinear Regression

```r
summary(lm.fit <- lm(medv ~ lstat + age, data = boston))
```

```

Call:
lm(formula = medv ~ lstat + age, data = boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.981  -3.978  -1.283   1.968  23.158 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.173 on 503 degrees of freedom
```

```
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

```r
summary(lm.fit <- lm(medv ~ ., data = boston))
```

```
Call:
lm(formula = medv ~ ., data = boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```r
vif(lm.fit)
```

```
    crim       zn    indus     chas      nox       rm      age      dis
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
     rad      tax  ptratio    black    lstat
7.484496 9.008554 1.799084 1.348521 2.941491
```

```r
summary(lm.fit1 <- lm(medv ~ .-age, data = boston))
```

```
Call:
```

```
lm(formula = medv ~ . - age, data = boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.6054  -2.7313  -0.5188   1.7601  26.2243 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.436927   5.080119   7.172 2.72e-12 ***
crim         -0.108006   0.032832  -3.290 0.001075 ** 
zn            0.046334   0.013613   3.404 0.000719 ***
indus         0.020562   0.061433   0.335 0.737989    
chas          2.689026   0.859598   3.128 0.001863 ** 
nox         -17.713540   3.679308  -4.814 1.97e-06 ***
rm            3.814394   0.408480   9.338  < 2e-16 ***
dis          -1.478612   0.190611  -7.757 5.03e-14 ***
rad           0.305786   0.066089   4.627 4.75e-06 ***
tax          -0.012329   0.003755  -3.283 0.001099 ** 
ptratio      -0.952211   0.130294  -7.308 1.10e-12 ***
black         0.009321   0.002678   3.481 0.000544 ***
lstat        -0.523852   0.047625 -10.999  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.74 on 493 degrees of freedom
Multiple R-squared:  0.7406,	Adjusted R-squared:  0.7343 
F-statistic: 117.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

Interaction Terms

```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
```

Non-linear Transformations of the Predictors

```r
summary(lm.fit2 <- lm(medv ~ lstat + I(lstat^2), data = boston))
```

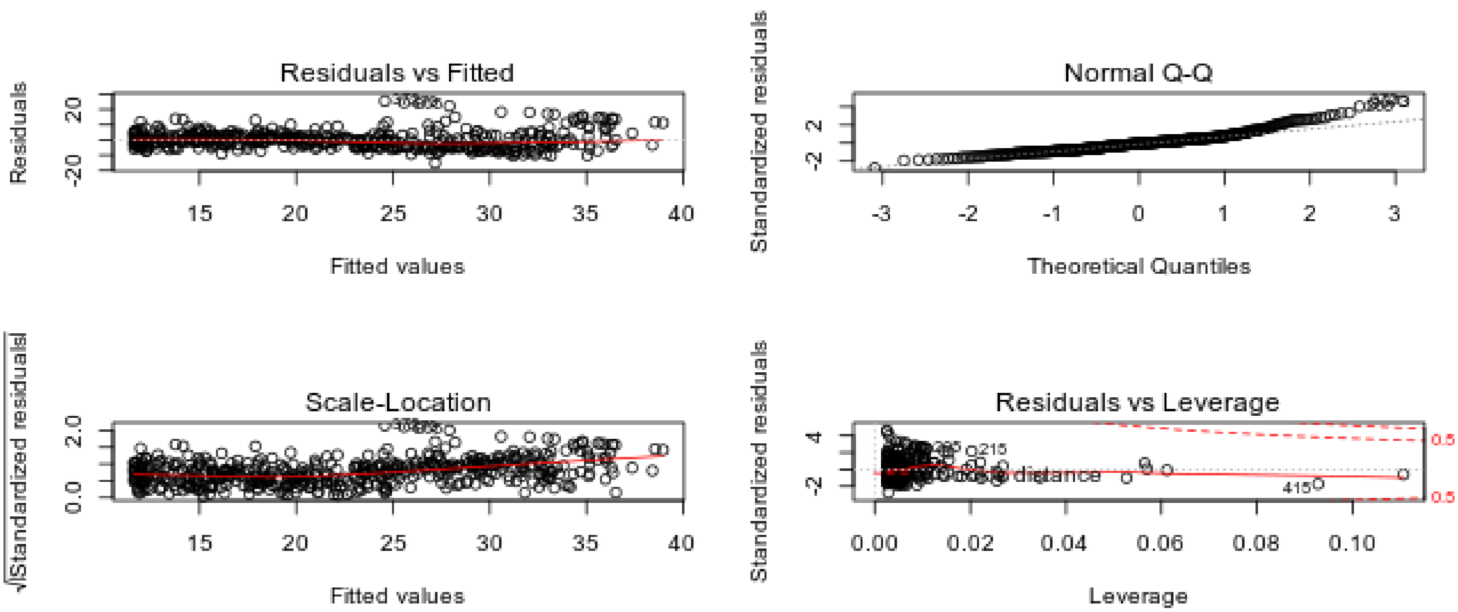
```

Call:
lm(formula = medv ~ lstat + I(lstat^2), data = boston)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2834  -3.8313  -0.5295   2.3095  25.4148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6393
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

```r
lm.fit <- lm(medv ~ lstat, data = boston)
anova(lm.fit, lm.fit2)
```

```
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
```

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

```r
par(mfrow = c(2,2))
with(boston, {
  plot(lm.fit2)
})
```




```r
summary(lm.fit5 <- lm(medv ~ poly(lstat, 5), data = boston))
```

```

Call:
lm(formula = medv ~ poly(lstat, 5), data = boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5433  -3.1039  -0.7052   2.0844  27.1153 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,    Adjusted R-squared:  0.6785
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

```r
summary(lm.fit5 <- lm(medv ~ log(rm), data = boston))
```

```
Call:
lm(formula = medv ~ log(rm), data = boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.487  -2.875  -0.104   2.837  39.816 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -76.488      5.028  -15.21   <2e-16 ***
log(rm)       54.055      2.739   19.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.915 on 504 degrees of freedom
Multiple R-squared:  0.4358,    Adjusted R-squared:  0.4347
F-statistic: 389.3 on 1 and 504 DF,  p-value: < 2.2e-16
```

Qualitative Predictors

```r
carseats <- Carseats
summary(carseats)
```

```
     Sales          CompPrice       Income        Advertising    
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000  
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000  
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000  
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635  
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000  
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000  
   Population        Price        ShelveLoc        Age          Education   
 Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0  
 1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0  
 Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0  
 Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9  
 3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0  
 Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0  
```

```
 Urban       US
 No :118   No :142
 Yes:282   Yes:258
```

```r
summary(lm.fit <- lm(Sales ~ . + Income:Advertising+Price:Age, data = carseats))
```

```

Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data = carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9208 -0.7503  0.0177  0.6754  3.3413 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6.5755654  1.0087470   6.519 2.22e-10 ***
CompPrice           0.0929371  0.0041183  22.567  < 2e-16 ***
Income              0.0108940  0.0026044   4.183 3.57e-05 ***
Advertising         0.0702462  0.0226091   3.107 0.002030 ** 
Population          0.0001592  0.0003679   0.433 0.665330    
Price              -0.1008064  0.0074399 -13.549  < 2e-16 ***
ShelveLocGood       4.8486762  0.1528378  31.724  < 2e-16 ***
ShelveLocMedium     1.9532620  0.1257682  15.531  < 2e-16 ***
Age                -0.0579466  0.0159506  -3.633 0.000318 ***
Education          -0.0208525  0.0196131  -1.063 0.288361    
UrbanYes            0.1401597  0.1124019   1.247 0.213171    
USYes              -0.1575571  0.1489234  -1.058 0.290729    
Income:Advertising  0.0007510  0.0002784   2.698 0.007290 ** 
Price:Age           0.0001068  0.0001333   0.801 0.423812    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 386 degrees of freedom
Multiple R-squared:  0.8761,    Adjusted R-squared:  0.8719 
F-statistic:   210 on 13 and 386 DF,  p-value: < 2.2e-16
```

```r
contrasts(carseats$ShelveLoc)
```

```
       Good Medium
Bad       0      0
Good      1      0
```

```
Medium    0      1
```

Conceptual

1.)

Describe the null hypothesis to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values.

a.)

$H_0 : TV = 0, H_a : TV ! = 0$

With a p-value of less than 0.0001, we reject the null hypothesis that the TV advertising budget does not effect sales.

b.)

$H_0 : radio = 0, H_a : radio! = 0$

With a p-value of less than 0.0001, we reject the null hypothesis that the radio advertising budget does not effect sales.

c.)

$H_0 : newspaper = 0, H_a : newspaper! = 0$

With a p-value of .8599, we fail to reject the null hypothesis that the newspaper advertising budget does not effect sales.

2.)

Carefully explain the differences between the KNN classifier and KNN regression methods.

KNN regression uses the same basic technique as the classifier, which is to take a specified number of neighbors (based on some distance measure, d*) and average them together to generate a value for the response. The difference here is that the regression returns a continious response variable, and the classificer results in a discrete metric that is based on the probability generated from the average of the neighbors.*

3.)

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 Female/0 Male), X_4 Interaction Between GPA and IQ, 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.)

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.

The least square line is given by

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \times IQ - 10GPA \times Gender$$

which becomes for the males

$$\hat{y} = 50 + 20GPA + 0.07IQ + 0.01GPA \times IQ,$$

and for the females

$$\hat{y} = 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ.$$

So the starting salary for males is higher than for females on average iff $50 + 20GPA \geq 85 + 10GPA$ *which is equivalent to* $GPA \geq 3.5$. *Therefore iii. is the right answer.*

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

It suffices to plug in the given values in the least square line for females given above and we obtain

$$\hat{y} = 85 + 40 + 7.7 + 4.4 = 137.1,$$

which gives us a starting salary of $137100\$$.

(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.

False. To verify if the GPA/IQ has an impact on the quality of the model we need to test the hypothesis $H_0 : \hat{\beta}_4 = 0$ *and look at the p-value associated with the* t *or the* F *statistic to draw a conclusion.*

4.)

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Without knowing more details about the training data, it is difficult to know which training RSS is lower between linear or cubic. However, as the true relationship between X *and* Y *is linear, we may expect the least squares line to be close to the true regression line, and consequently the RSS for the linear regression may be lower than for the cubic regression.*

(b) Answer (a) using test rather than training RSS.

In this case the test RSS depends upon the test data, so we have not enough information to conclude. However, we may assume that polynomial regression will have a higher test RSS as the overfit from training would have more error than the linear regression.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Polynomial regression has lower train RSS than the linear fit because of higher flexibility: no matter what the underlying true relationship is the more flexible model will closer follow points and reduce train RSS. An example of this beahvior is shown on Figure 2.9 from Chapter 2.

(d) Answer (c) using test rather than training RSS.

There is not enough information to tell which test RSS would be lower for either regression given the problem statement is defined as not knowing "how far it is from linear". If it is closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS. Or, if it is closer to cubic than linear, the cubic regression test RSS could be lower than the linear regression test RSS. It is dues to bias-variance tradeoff: it is not clear what level of flexibility will fit data better.

5.)

Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i-th fitted value takes the form $\hat{y}_i = x_i\hat{\beta}$, where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{k=1}^{n} x_k^2}.$$

Show that we can write

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j.$$

What is a_j ?

We have immediately that

$$\hat{y}_i = x_i \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} x_k^2} = \sum_{j=1}^{n} \frac{x_i x_j}{\sum_{k=1}^{n} x_k^2} y_j = \sum_{j=1}^{n} a_j y_j.$$

6.)

Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\overline{x}, \overline{y})$.

The least square line equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$, so if we substitute $\overline{x}$ for x we obtain

$$y = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}.$$

We may conclude that the least square line passes through the point $(\overline{x}, \overline{y})$.

7.)

It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\overline{x} = \overline{y} = 0$.

We have the following equalities

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_j y_j^2};$$

with $\hat{y}_i = \hat{\beta}_1 x_i$ *we may write*

$$R^2 = 1 - \frac{\sum_i (y_i - \sum_j x_j y_j / \sum_j x_j^2 x_i)^2}{\sum_j y_j^2} = \frac{\sum_j y_j^2 - (\sum_i y_i^2 - 2\sum_i y_i (\sum_j x_j y_j / \sum_j x_j^2) x_i + \sum_i (\sum_j x_j y_j / \sum_j x_j^2)^2 x_i^2)}{\sum_j y_j^2}$$

and finally

$$R^2 = \frac{2(\sum_i x_i y_i)^2 / \sum_j x_j^2 - (\sum_i x_i y_i)^2 / \sum_j x_j^2}{\sum_j y_j^2} = \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2 \sum_j y_j^2} = Cor(X, Y)^2.$$

Applied

8.)

This question involves the use of simple linear regression on the “Auto” data set.

(a) Use the lm() function to perform a simple linear regression with “mpg” as the response and “horsepower” 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 ?

```
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
```

We can answer this question by testing the hypothesis $H_0 : \beta_i = 0 \; \forall i$. The p-value corresponding to the F-statistic is 7.031989×10^{-81}, this indicates a clear evidence of a relationship between "mpg" and "horsepower".

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

To calculate the residual error relative to the response we use the mean of the response and the RSE. The mean of mpg is 23.4459184. The RSE of the lm.fit was 4.9057569 which indicates a percentage error of 20.9237141%. We may also note that as the R^2 is equal to 0.6059483, almost 60.5948258% of the variability in "mpg" can be explained using "horsepower".

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

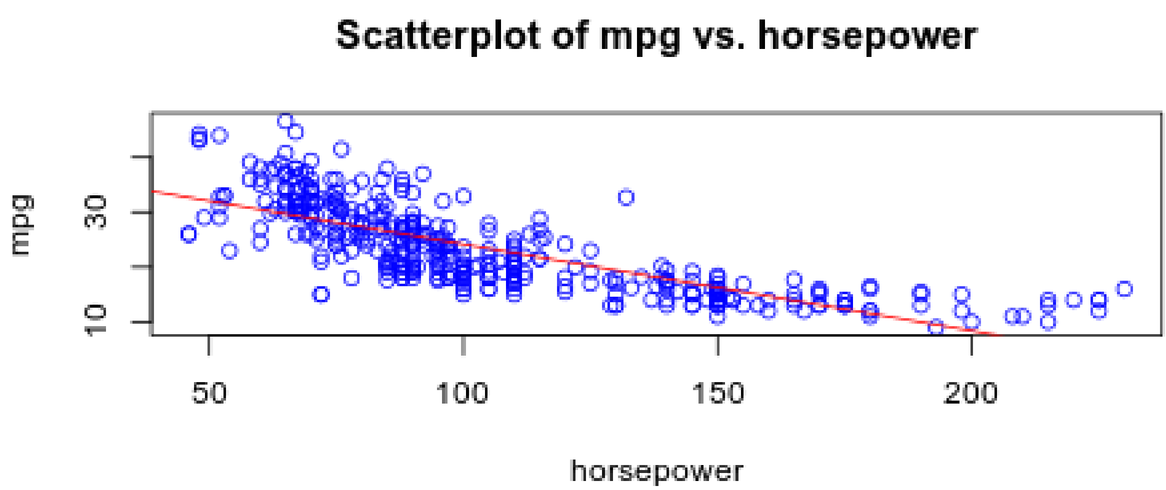
As the coeficient of "horsepower" is negative, the relationship is also negative. The more horsepower an automobile has the linear regression indicates the less mpg fuel efficiency the automobile will have.

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

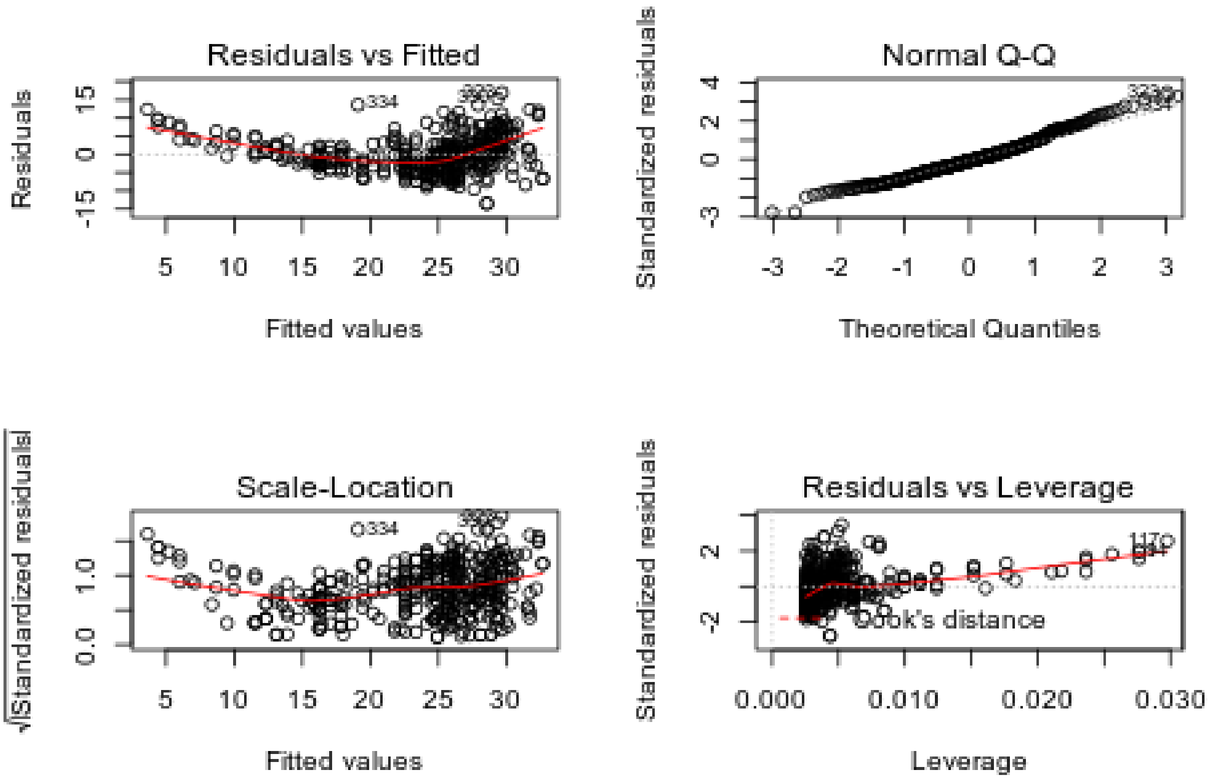
```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

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

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



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



The plot of residuals versus fitted values indicates the presence of non linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and a few high leverage points.

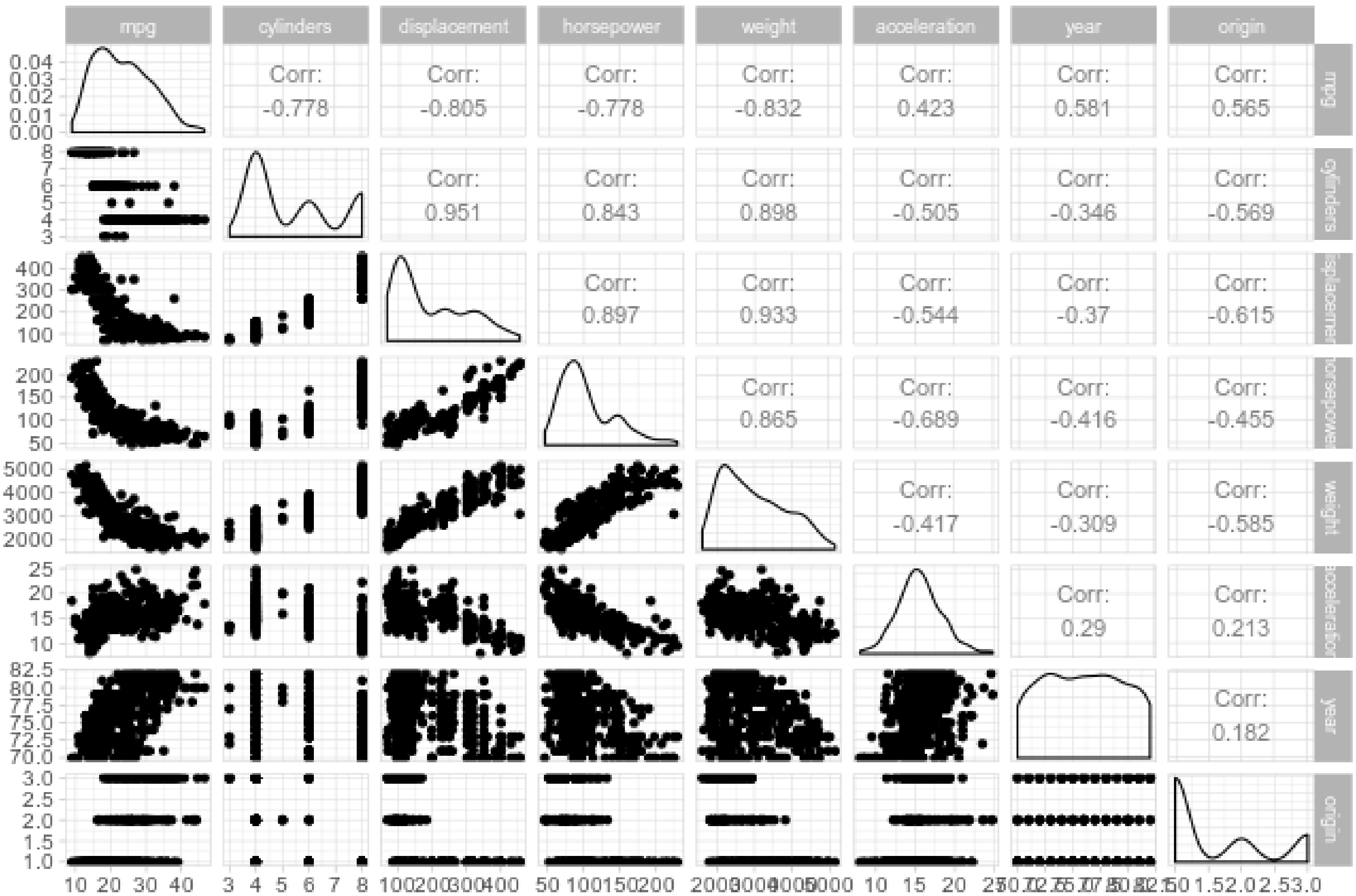
9.)

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

```
auto <- as.data.table(Auto)
```

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

```
ggpairs(auto[, -"name", with = F])
```



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

```r
cor(auto[, -"name", with = F])
```

```
                    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 :

```r
summary(fit <- lm(mpg ~ . -name, data = auto))
```

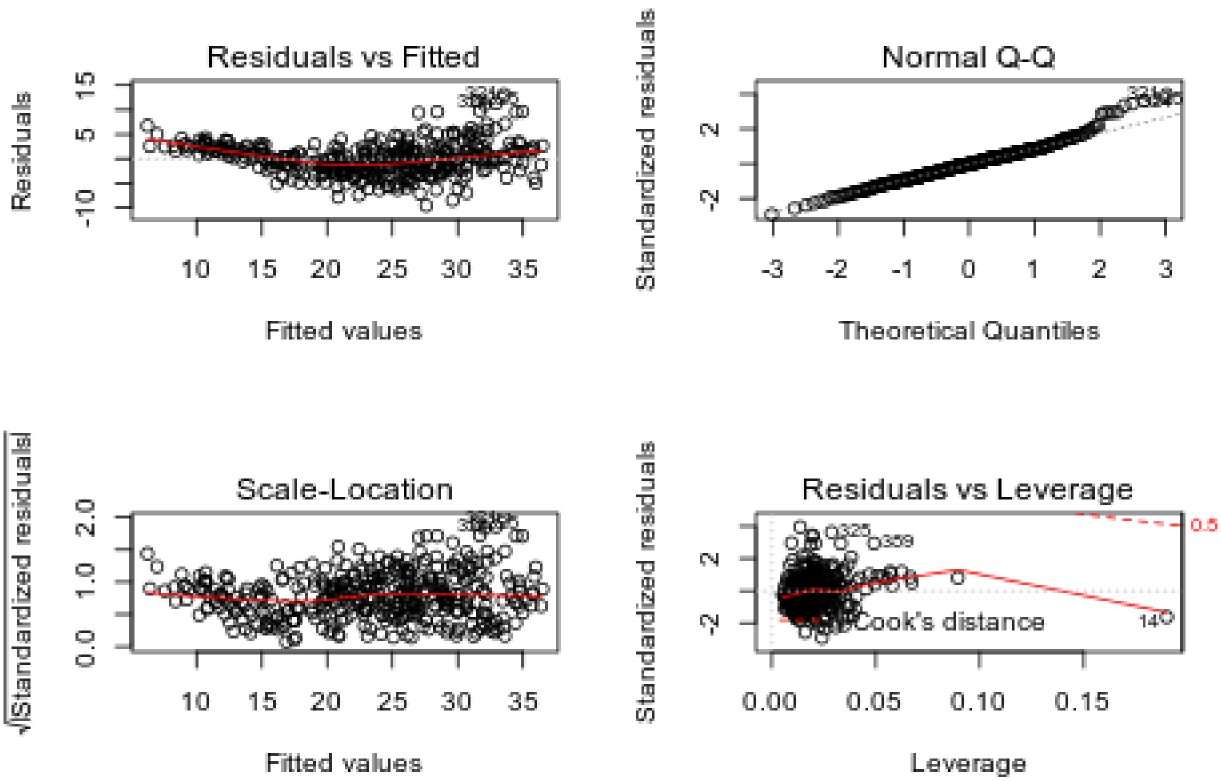
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5903 -2.1565 -0.1169  1.8690 13.0604 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

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

Displacement, weight, year and origin seem to have statistically significant impacts on the response.

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers ? Does the leverage plots identify any observations with unusually high leverages ?

```r
par(mfrow = c(2,2))
plot(fit)
```



The residuals appear to be approximately normally distributed, however there is a skewness in the right tail. There is one high leverage point (14).

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

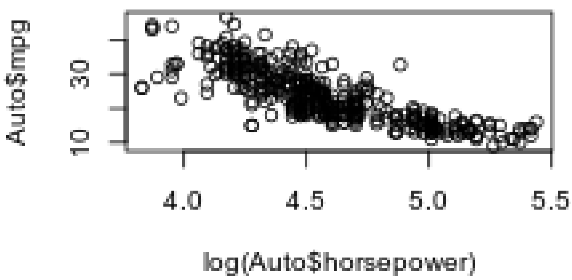
From the correlation matrix, we obtained the two highest correlated pairs and used them in picking interaction effects.

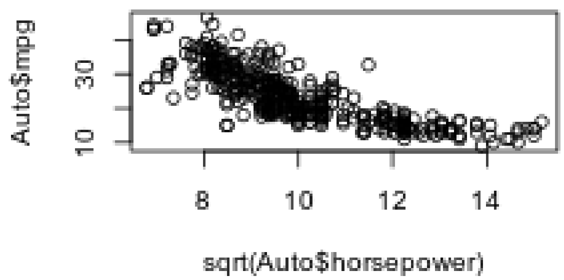
```
summary(lm.fig <- lm(mpg ~ cylinders * displacement+displacement * weight, data = auto))
```

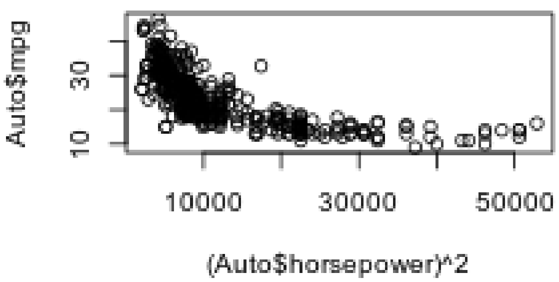
```
Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2934  -2.5184  -0.3476   1.8399  17.7723

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders               7.606e-01  7.669e-01   0.992    0.322
```

```
displacement          -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight                -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

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







We limit ourselves to examining "horsepower" as sole predictor. It seems that the log transformation gives the most linear looking plot.

10.)

This question should be answered using the "Carseats" data set.

(a) Fit a multiple regression model to predict "Sales" using "Price", "Urban" and "US

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative !

The coefficient of the "Price" variable may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.4588492 units in sales all other predictors remaining fixed. The coefficient of the "Urban" variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed. The coefficient of the "US" variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

The model may be written as

$$Sales = 13.0434689 + (-0.0544588) \times Price + (-0.0219162) \times Urban + (1.2005727) \times US + \varepsilon$$

with $Urban = 1$ if the store is in an urban location and 0 if not, and $US = 1$ if the store is in the US and 0 if not.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We can reject the null hypothesis for the "Price" and "US" variables.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
Call:
lm(formula = Sales ~ Price + US, data = Carseats)

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

```
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

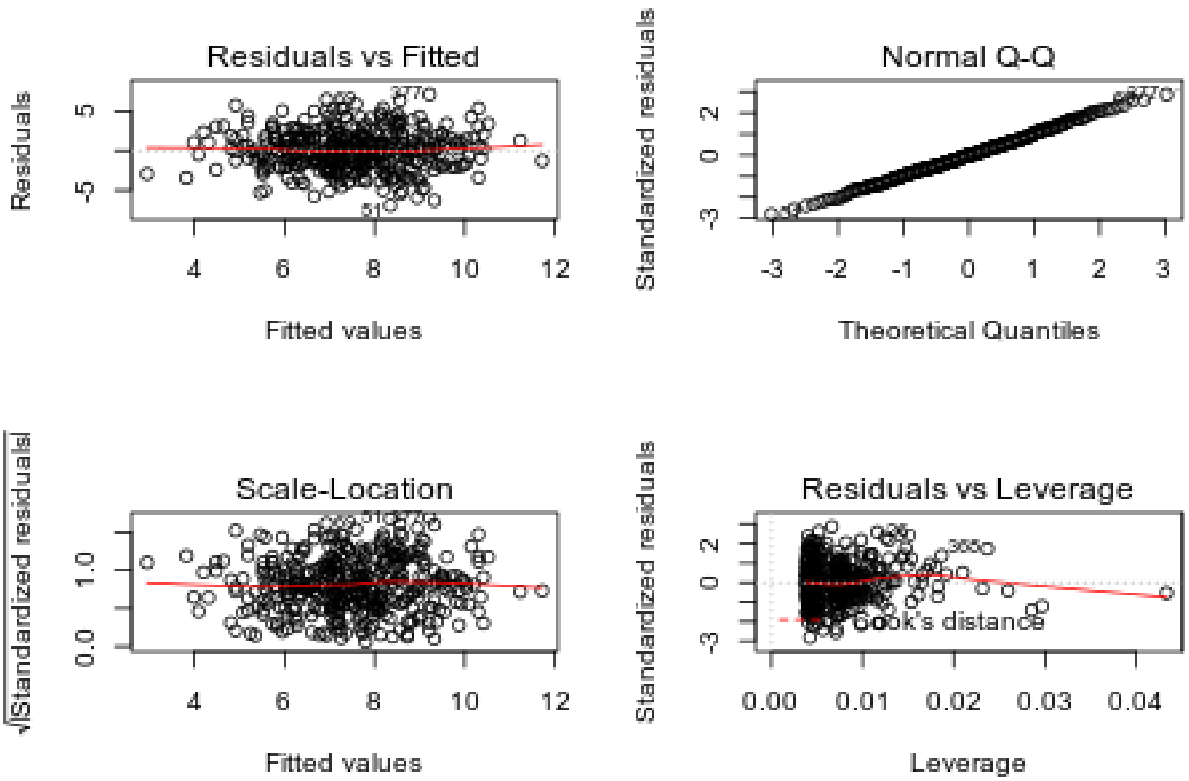
(f) How well do the models in (a) and (e) fit the data ?

The R^2 for the smaller model is marginally better than for the bigger model. Essentially about 23.9262888% of the variability is explained by the model.

(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

(h) Is there evidence of outliers or high leverage observations in the model from (e) ?



The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and some leverage points as some points exceed $(p+1)/n$ (0.01).

11.)

In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

(a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H_0. Comment on these results.

```
Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9154 -0.6472 -0.1771  0.5056  2.3109

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

```
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

According to the summary above, we have a value of 1.9938761 for $\hat{\beta}$, a value of 0.1064767 for the standard error, a value of 18.7259319 for the t-statistic and a value of $2.6421969 \times 10^{-34}$ for the p-value. The small p-value allows us to reject H_0.

(b) Now perform a simple linear regression of x onto y, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H_0. Comment on these results.

```
Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2368  0.1030  0.2858  0.8938

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

According to the summary above, we have a value of 0.3911145 for $\hat{\beta}$, a value of 0.0208863 for the standard error, a value of 18.7259319 for the t-statistic and a value of $2.6421969 \times 10^{-34}$ for the p-value. The small p-value allows us to reject H_0.

(c) What is the relationship between the results obtained in (a) and (b) ?

We obtain the same value for the t-statistic and consequently the same value for the corresponding p-value. Both results in (a) and (b) reflect the same line created in (a). In other words, $y = 2x + \varepsilon$ could also be written $x = 0.5(y - \varepsilon)$.

(d) For the regrssion of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/SE(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38), and where

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i=1}^{n} x_i^2}}.$$

Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$\frac{\sqrt{n-1}\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i y_i)}}.$$

We have

$$t = \frac{\sum_i x_i y_y / \sum_j x_j^2}{\sqrt{\sum_i (y_i - x_i \hat{\beta})^2/(n-1)\sum_j x_j^2}} = \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_i (y_i - x_i \sum_j x_j y_j / \sum_j x_j^2)^2}} = \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{(\sum_j x_j^2)(\sum_j y_j^2) - (\sum_j x_j y_j)}}.$$

Now let's verify this result numerically.

```
[1] 18.72593
```

We may see that the t above is exactly the t-statistic given in the summary of "fit6".

(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same t-statistic for the regression of x onto y.

It is easy to see that if we replace x_i by y_i in the formula for the t-statistic, the result would be the same.

(f) In R, show that when regression is performed with an intercept, the t-statistic for $H_0 : \beta_1 = 0$ is the same for the regression of y onto x as it is the regression of x onto y.

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

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8768 -0.6138 -0.1395  0.5394  2.3462 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.03769    0.09699  -0.389    0.698    
x            1.99894    0.10773  18.556   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762 
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16


Call:
lm(formula = x ~ y)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.90848 -0.28101  0.06274  0.24570  0.85736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03880    0.04266    0.91    0.365
y            0.38942    0.02099   18.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4249 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

It is again easy to see that the t-statistic for "fit7" and "fit8" are both equal to 18.5555993.

12.)

This problem involves simple linear regression without an intercept.

(a) Recall that the coefficient estimate $\hat{\beta}$ for the linear regression of Y onto X witout an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X ?

The coefficient estimate for the regression of Y onto X is

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_j x_j^2};$$

The coefficient estimate for the regression of X onto Y is

$$\hat{\beta}' = \frac{\sum_i x_i y_i}{\sum_j y_j^2}.$$

The coefficients are the same iff $\sum_j x_j^2 = \sum_j y_j^2$.

(b) Generate an example in R with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

```
[1] 338350

[1] 1353606


Call:
lm(formula = y ~ x + 0)

Residuals:
```

```
      Min        1Q    Median        3Q       Max 
-0.223590 -0.062560  0.004426  0.058507  0.230926 

Coefficients:
   Estimate Std. Error t value Pr(>|t|)    
x 2.0001514  0.0001548   12920   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09005 on 99 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 1.669e+08 on 1 and 99 DF,  p-value: < 2.2e-16


Call:
lm(formula = x ~ y + 0)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.115418 -0.029231 -0.002186  0.031322  0.111795 

Coefficients:
  Estimate Std. Error t value Pr(>|t|)    
y 5.00e-01   3.87e-05   12920   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04502 on 99 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 1.669e+08 on 1 and 99 DF,  p-value: < 2.2e-16
```

(c) Generate an example in R with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.

```
[1] 338350
```

```
[1] 338350
```

```
Call:
lm(formula = y ~ x + 0)

Residuals:
   Min     1Q Median     3Q    Max 
-49.75 -12.44  24.87  62.18  99.49 

Coefficients:
```

```
  Estimate Std. Error t value Pr(>|t|)
x   0.5075     0.0866    5.86 6.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.37 on 99 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:   0.25
F-statistic: 34.34 on 1 and 99 DF,  p-value: 6.094e-08


Call:
lm(formula = x ~ y + 0)

Residuals:
   Min     1Q Median     3Q    Max
-49.75 -12.44  24.87  62.18  99.49

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y   0.5075     0.0866    5.86 6.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.37 on 99 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:   0.25
F-statistic: 34.34 on 1 and 99 DF,  p-value: 6.094e-08
```

13.)

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure conistent results.

(a) Using the rnorm() function, create a vector, "x", containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

(b) Using the rnorm() function, create a vector, "eps", containing 100 observations drawn from a $N(0, 0.25)$ distribution.

(c) Using "x" and "eps", generate a vector "y" according to the model

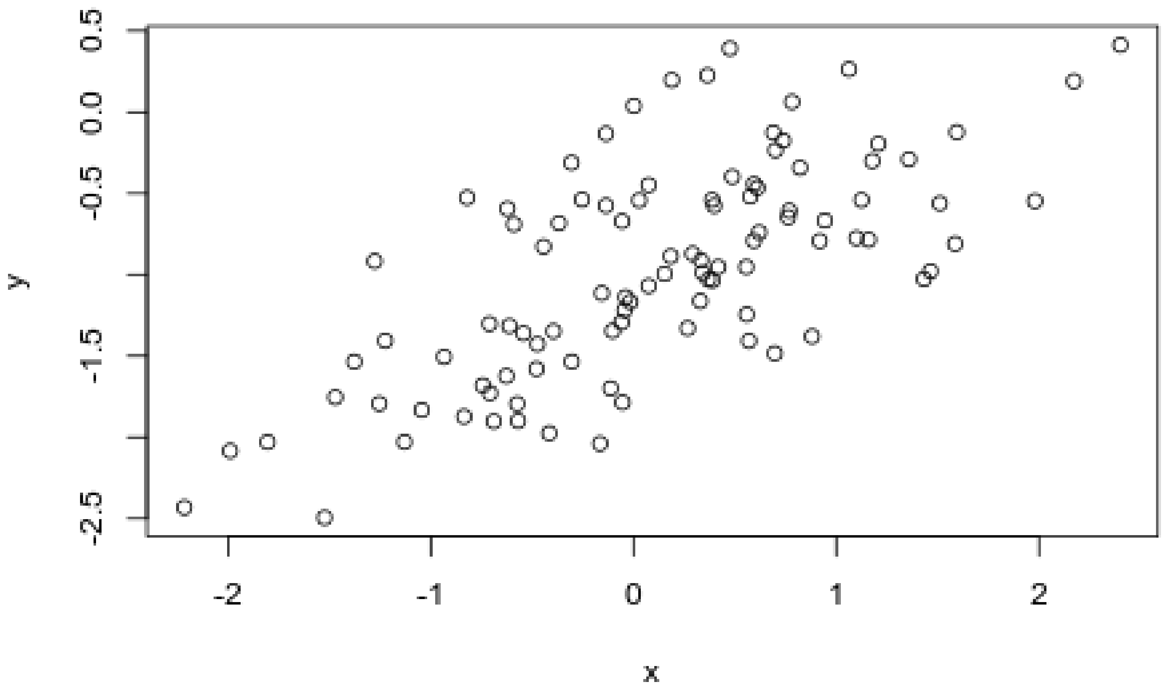
$$Y = -1 + 0.5X + \varepsilon.$$

What is the length of the vector "y" ? What are the values of β_0 and β_1 in this linear model ?

```
[1] 100
```

The values of β_0 and β_1 are -1 and 0.5 respectively.

(d) Create a scatterplot displaying the relationship between "x" and "y". Comment on what you observe.



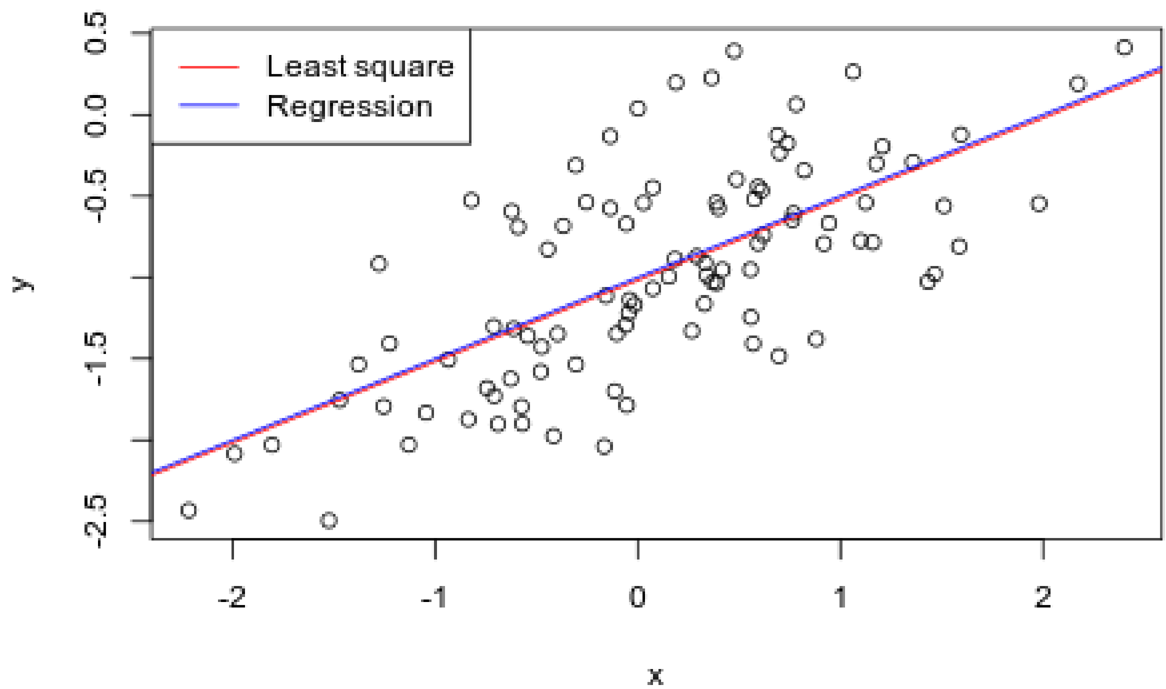
The relationship between "x" and "y" looks linear with some noise introduced by the "eps" variable.

(e) Fit a least squares linear model to predict "y" using "x". Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1 ?

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.93842 -0.30688 -0.06975  0.26970  1.17309

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
x            0.49947    0.05386   9.273 4.58e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared:  0.4674,    Adjusted R-squared:  0.4619
F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are pretty close to β_0 and β_1. The model has a large F-statistic with a near-zero p-value so the null hypothesis can be rejected.

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() function to create an appropriate legend.



(g) Now fit a polynomial regression model that predicts "y" using "x" and "x^2". Is there evidence that the quadratic term improves the model fit ? Explain your answer.

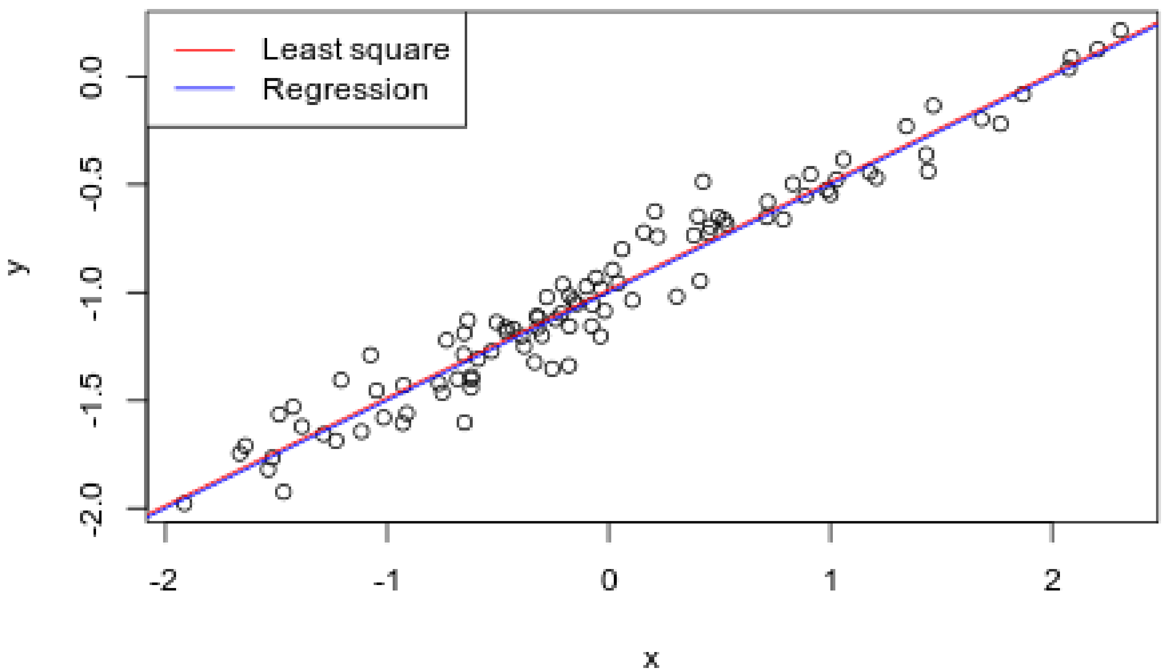
```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.98252 -0.31270 -0.06441  0.29014  1.13500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
x            0.50858    0.05399   9.420  2.4e-15 ***
I(x^2)      -0.05946    0.04238  -1.403    0.164
---
```

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

Residual standard error: 0.479 on 97 degrees of freedom
Multiple R-squared:  0.4779,	Adjusted R-squared:  0.4672 
F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

The coefficient for "x^2" is not significant as its p-value is higher than 0.05. So there is not sufficient evidence that the quadratic term improves the model fit even though the R^2 is slightly higher and RSE slightly lower than the linear model.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The initial model should remain the same. Describe your results.

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29052 -0.07545  0.00067  0.07288  0.28664 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.98639    0.01129  -87.34   <2e-16 ***
x            0.49988    0.01184   42.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1128 on 98 degrees of freedom
Multiple R-squared:  0.9479,	Adjusted R-squared:  0.9474 
F-statistic:  1782 on 1 and 98 DF,  p-value: < 2.2e-16
```

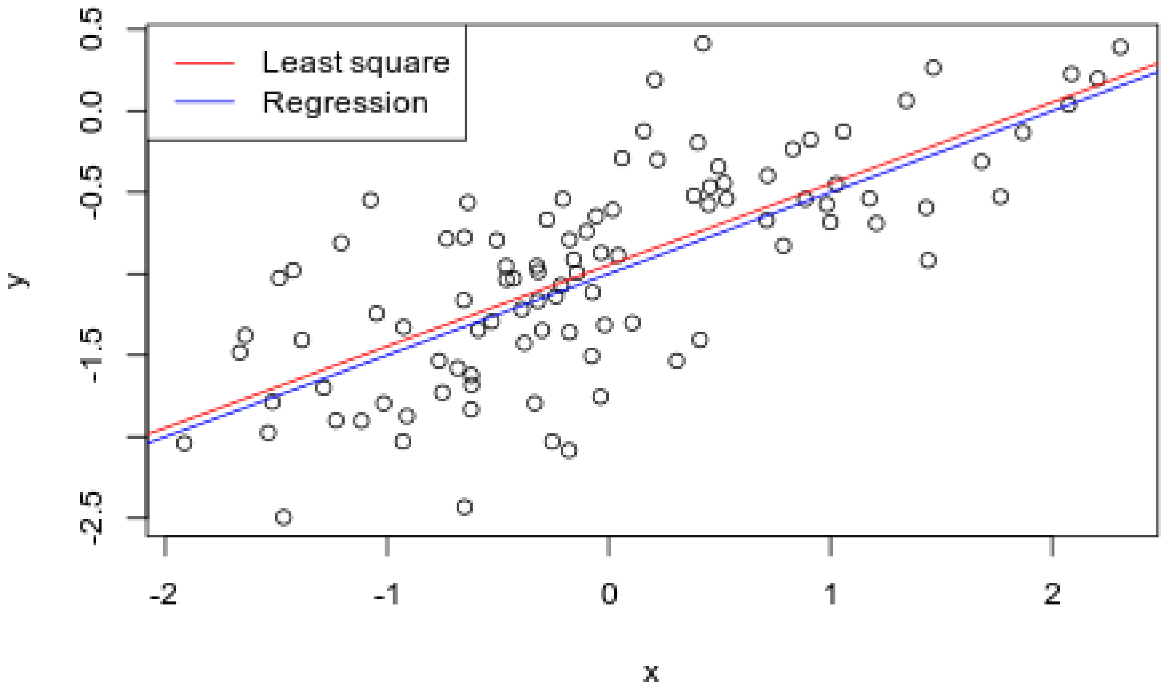


We reduced the noise by decreasing the variance of the normal distribution used to generate the error term ε. We may see that the coefficients are very close to the previous ones, but now, as the relationship is nearly linear, we have a much higher R^2 and much lower RSE. Moreover, the two lines overlap each other as we have very little noise.

(i) Repeat (a)-(f) after modifying the data generation process in such a way that there is more noise in the data. The initial model should remain the same. Describe your results.

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.16208 -0.30181  0.00268  0.29152  1.14658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.94557    0.04517  -20.93   <2e-16 ***
x            0.49953    0.04736   10.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4514 on 98 degrees of freedom
```

```
Multiple R-squared:  0.5317,    Adjusted R-squared:  0.5269
F-statistic: 111.2 on 1 and 98 DF,  p-value: < 2.2e-16
```



We increased the noise by increasing the variance of the normal distribution used to generate the error term ε. We may see that the coefficients are again very close to the previous ones, but now, as the relationship is not quite linear, we have a much lower R^2 and much higher RSE. Moreover, the two lines are wider apart but are still really close to each other as we have a fairly large data set.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set ? Comment on your results.

```
                 2.5 %     97.5 %
(Intercept) -1.1150804 -0.9226122
x            0.3925794  0.6063602

                2.5 %     97.5 %
(Intercept) -1.008805 -0.9639819
x            0.476387  0.5233799

                 2.5 %     97.5 %
(Intercept) -1.0352203 -0.8559276
x            0.4055479  0.5935197
```

All intervals seem to be centered on approximately 0.5. As the noise increases, the confidence intervals widen. With less noise, *there is more predictability in the data set.*

14.)

This problem focuses on the collinearity problem.

(a) Perform the following commands in R.

The last line corresponds to creating a linear model in which “y” is a function of “x1” and “x2”. Write out the form of the linear model. What are the regression coefficients ?

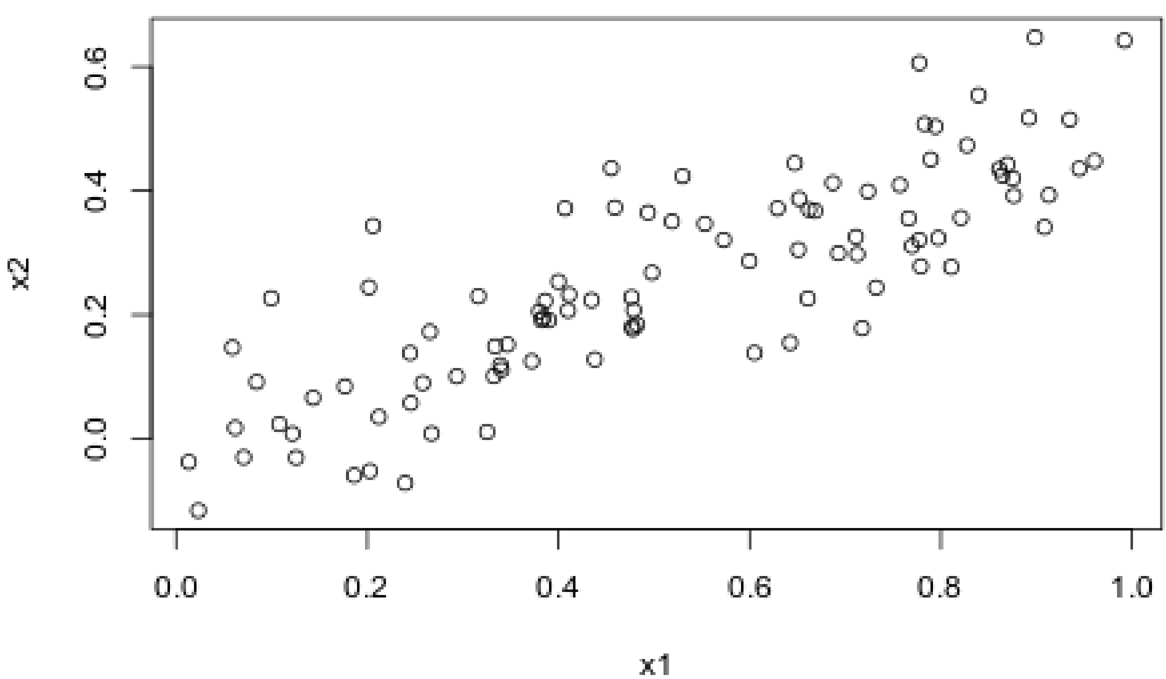
The form of the linear model is

$$Y = 2 + 2X_1 + 0.3X_2 + \varepsilon$$

with ε *a* $N(0, 1)$ *random variable. The regression coefficients are respectively 2, 2 and 0.3.*

(b) What is the correlation between “x1” and “x2” ? Create a scatterplot displaying the relationship between the variables.

```
[1] 0.8351212
```



The variables seem highly correlated.

(c) Using this data, fit a least squares regression to predict “y” using “x1” and “x2”. 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 $H_0 : \beta_2 = 0$?

```
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
```

The coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are respectively 2.1304996, 1.4395554 and 1.0096742. Only $\hat{\beta}_0$ is close to β_0. As the p-value is less than 0.05 we may reject H_0 for β_1, however we may not reject H_0 for β_2 as the p-value is higher than 0.05.

(d) Now fit a least squares regression to predict "y" using only "x1". Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
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 coefficient for "x1" in this last model is very different from the one with "x1" and "x2" as predictors. In this case "x1" is highly significant as its p-value is very low, so we may reject H_0.

(e) Now fit a least squares regression to predict “y” using only “x2”. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
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 coefficient for “x2” in this last model is very different from the one with “x1” and “x2” as predictors. In this case “x2” is highly significant as its p-value is very low, so we may again reject H_0.

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

No, the results do not contradict each other. As the predictors “x1” and “x2” are highly correlated we are in the presence of collinearity, in this case it can be difficult to determine how each predictor separately is associated with the response. Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_1$ *to grow (we have a standard error of 0.7211795 and 1.1337225 for “x1” and “x2” respectively in the model with two predictors and only of 0.3962774 and 0.6330467 for “x1” and “x2” respectively in the models with only one predictor). Consequently, we may fail to reject* H_0 *in the presence of collinearity. The importance of the “x2” variable has been masked due to the presence of collinearity.*

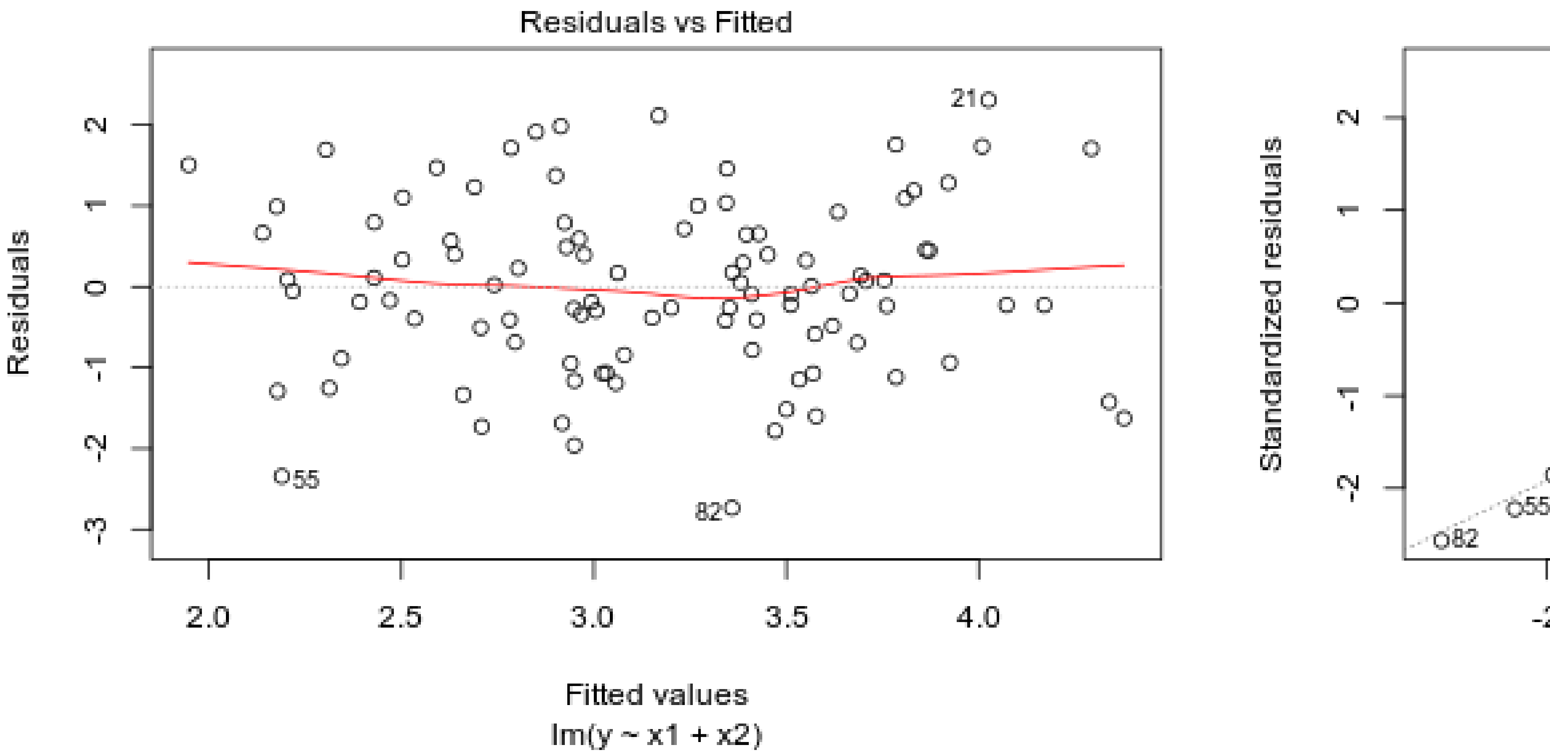
(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

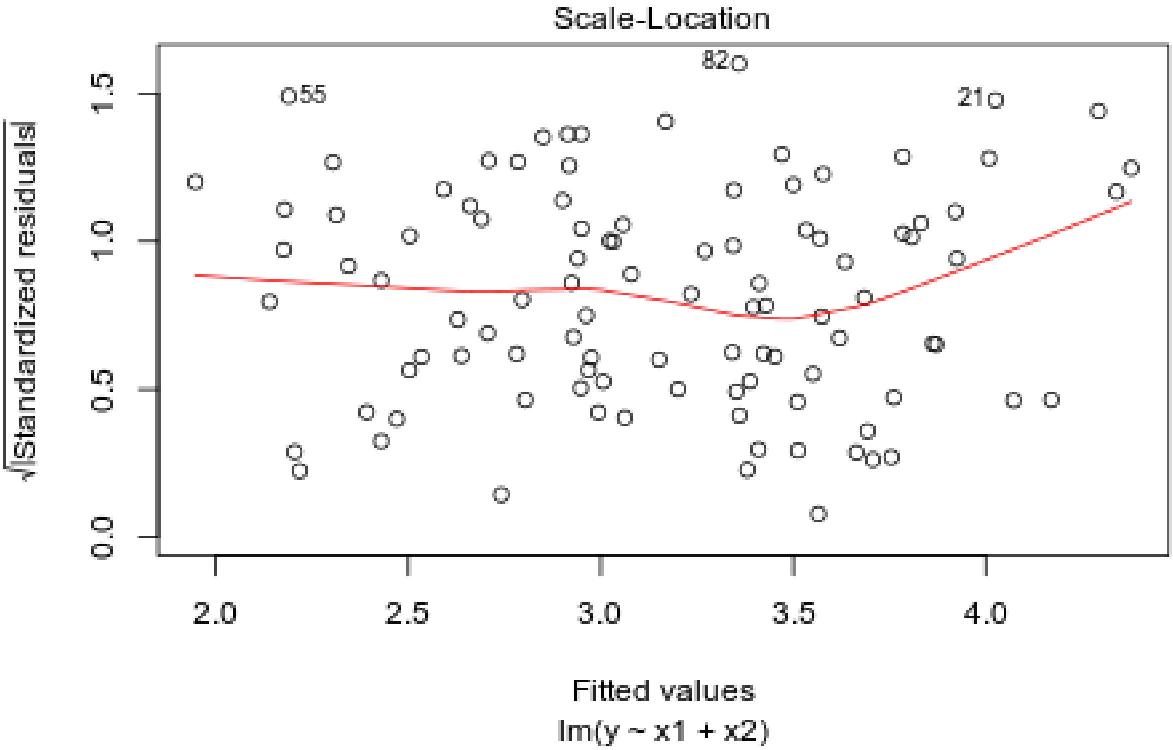
Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on each of the models ? In each model, is this observation an outlier ? A high-leverage point ? Explain your answers.

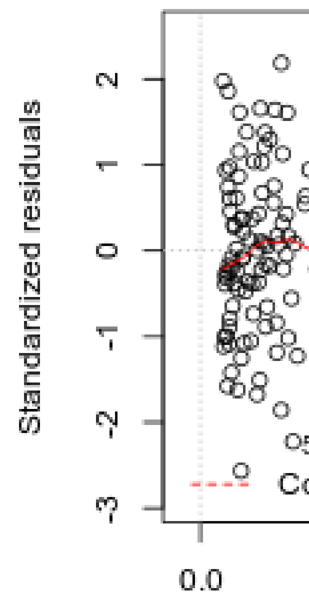
```
Call:
lm(formula = y ~ x1 + x2)

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

Coefficients:
```

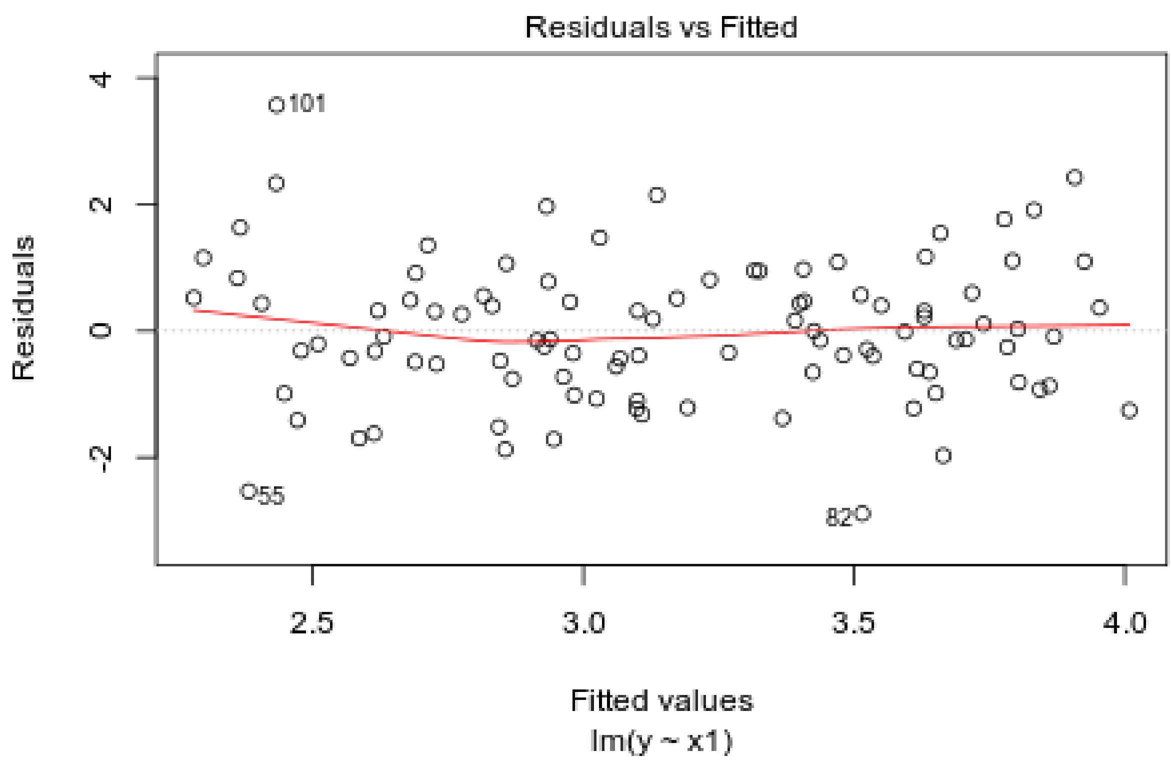
```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1            0.5394     0.5922   0.911  0.36458    
x2            2.5146     0.8977   2.801  0.00614 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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


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


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
```

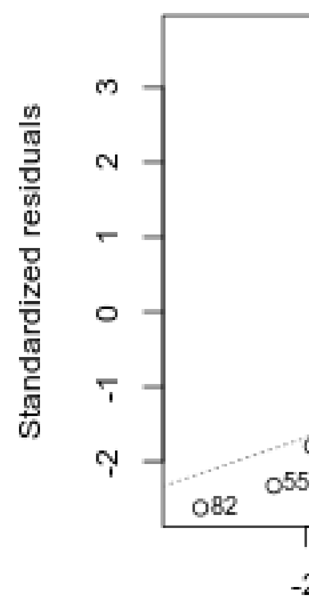
```
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2042
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

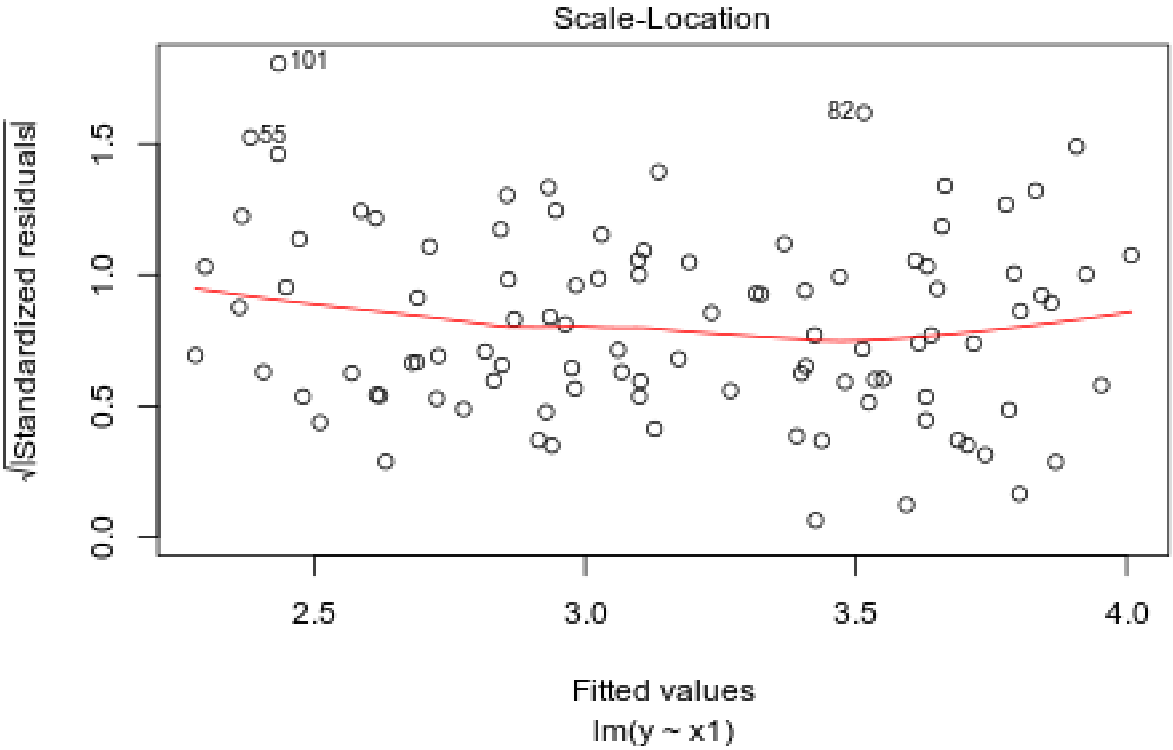


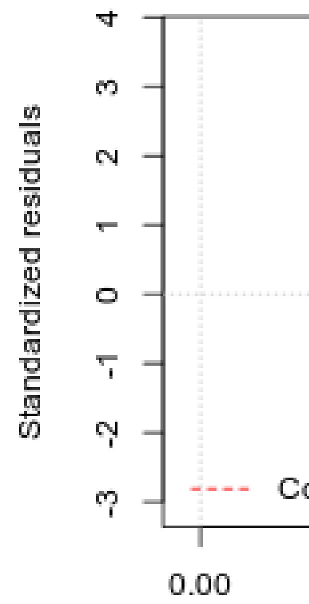

Scale-Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + x2)

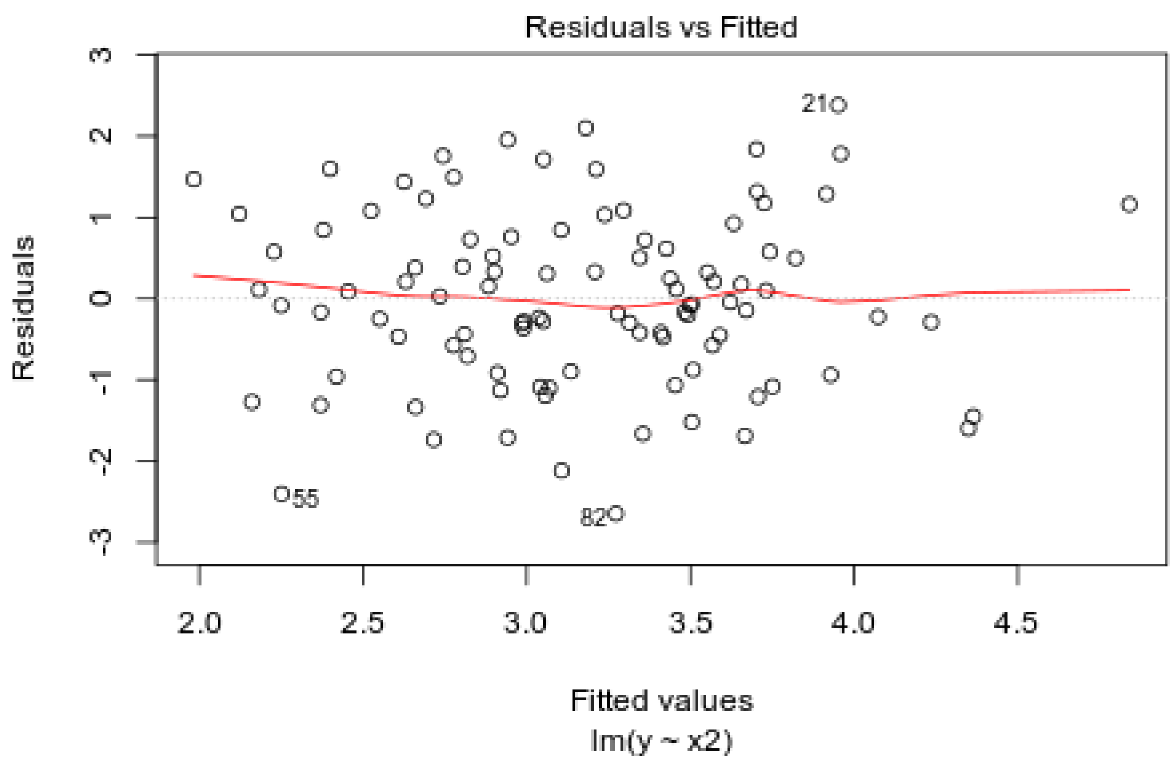

Standardized residuals

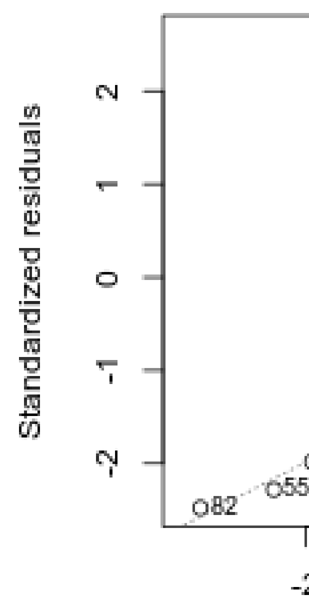

Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1)

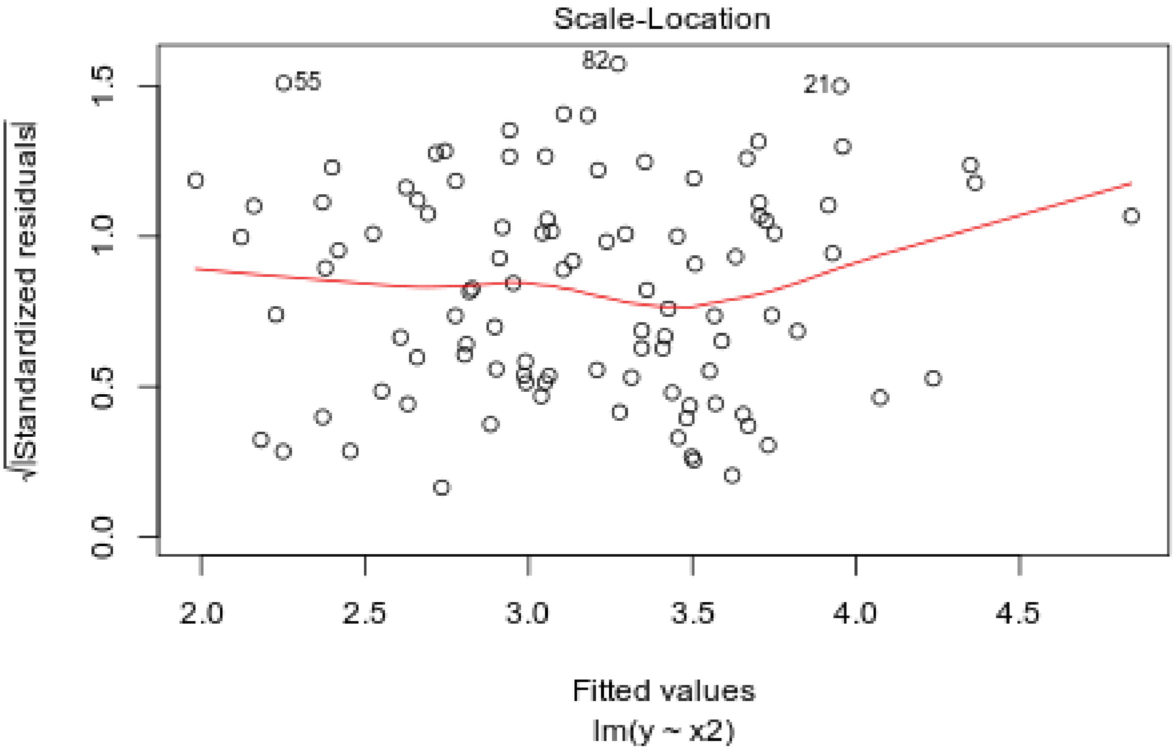

Standardized residuals

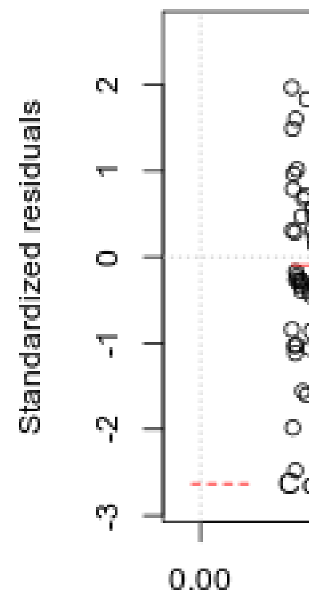

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.5
3.0
3.5
4.0
101
55
82
Fitted values
lm(y ~ x1)


Standardized residuals
4
3
2
1
0
-1
-2
-3
0.00


Residuals vs Fitted
Residuals
3
2
1
0
-1
-2
-3
2.0
2.5
3.0
3.5
4.0
4.5
21
55
82
Fitted values
lm(y ~ x2)


Standardized residuals
2
1
0
-1
-2
82
55





In the model with two predictors, the last point is a high-leverage point. In the model with "x1" as sole predictor, the last point is an outlier. In the model with "x2" as sole predictor, the last point is a high leverage point.

15.)

This problem involves the "Boston" data set, which we saw in the lab for this chapter. 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.

```
Call:
lm(formula = crim ~ zn)

Residuals:
   Min     1Q Median     3Q    Max
-4.429 -4.222 -2.620  1.250 84.523

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,	Adjusted R-squared:  0.03828 
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06


Call:
lm(formula = crim ~ indus)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.972  -2.698  -0.736   0.712  81.813 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.06374    0.66723  -3.093  0.00209 ** 
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653,	Adjusted R-squared:  0.1637 
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ chas)

Residuals:
   Min     1Q Median     3Q    Max 
-3.738 -3.661 -3.435  0.018 85.232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.7444     0.3961   9.453   <2e-16 ***
chas1        -1.8928     1.5061  -1.257    0.209    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124,	Adjusted R-squared:  0.001146 
```

```
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094


Call:
lm(formula = crim ~ nox)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.371  -2.738  -0.974   0.559  81.728 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
nox           31.249      2.999  10.419  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared:  0.1772,	Adjusted R-squared:  0.1756 
F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ rm)

Residuals:
   Min     1Q Median     3Q    Max 
-6.604 -3.952 -2.654  0.989 87.197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   20.482      3.365   6.088 2.27e-09 ***
rm            -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807,	Adjusted R-squared:  0.04618 
F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07


Call:
lm(formula = crim ~ age)

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

```
-6.789 -4.257 -1.230  1.527 82.849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
age          0.10779    0.01274   8.463 2.85e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.057 on 504 degrees of freedom
Multiple R-squared:  0.1244,    Adjusted R-squared:  0.1227
F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16


Call:
lm(formula = crim ~ dis)

Residuals:
   Min     1Q Median     3Q    Max
-6.708 -4.134 -1.527  1.516 81.674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4993     0.7304  13.006   <2e-16 ***
dis          -1.5509     0.1683  -9.213   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,    Adjusted R-squared:  0.1425
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ rad)

Residuals:
    Min      1Q  Median      3Q     Max
-10.164  -1.381  -0.141   0.660  76.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
rad          0.61791    0.03433  17.998  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared:  0.3913,    Adjusted R-squared:   0.39
F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ tax)

Residuals:
    Min      1Q  Median      3Q     Max
-12.513  -2.738  -0.194   1.065  77.696

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
tax          0.029742   0.001847   16.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.3383
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ ptratio)

Residuals:
   Min     1Q Median     3Q    Max
-7.654 -3.985 -1.912  1.825 83.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,   Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11


Call:
lm(formula = crim ~ black)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-13.756  -2.299  -2.095  -1.296  86.822 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.553529   1.425903  11.609   <2e-16 ***
black       -0.036280   0.003873  -9.367   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483,	Adjusted R-squared:  0.1466 
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.925  -2.822  -0.664   1.079  82.862 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
lstat        0.54880    0.04776  11.491  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared:  0.2076,	Adjusted R-squared:  0.206 
F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16


Call:
lm(formula = crim ~ medv)

Residuals:
   Min     1Q Median     3Q    Max 
-9.071 -4.022 -2.343  1.298 80.957 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.79654    0.93419   12.63   <2e-16 ***
```

```
medv        -0.36316    0.03839   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508,    Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

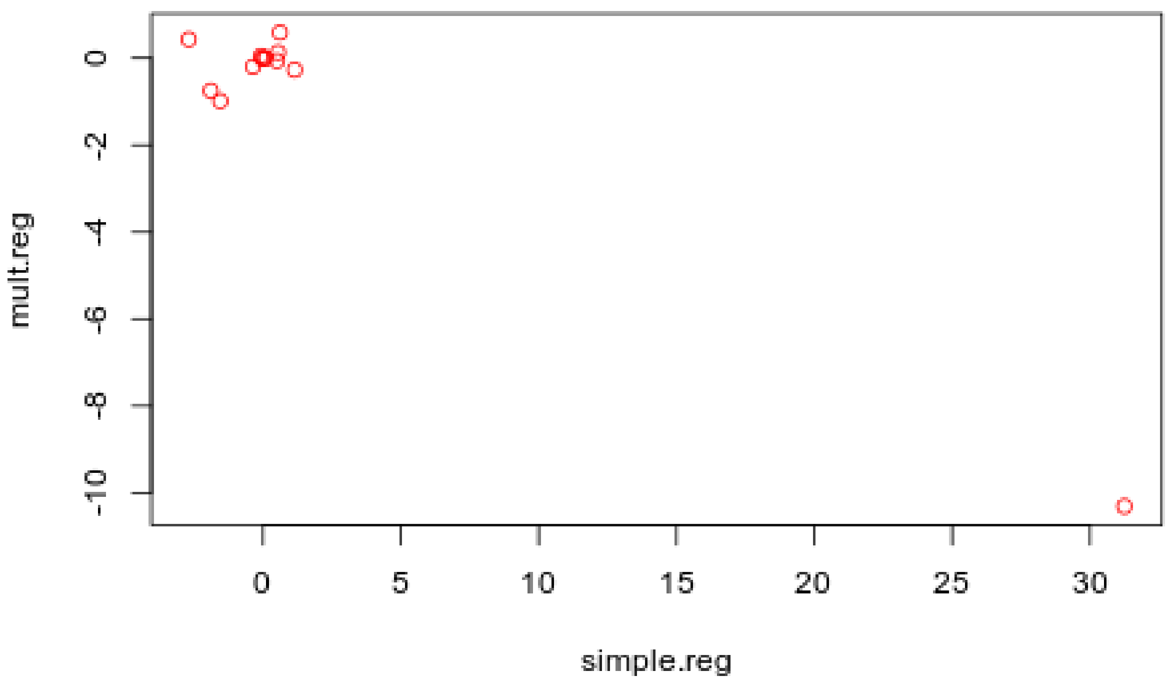
To find which predictors are significant, we have to test $H_0 : \beta_1 = 0$. All predictors have a p-value less than 0.05 except "chas", so we may conclude that there is a statistically significant association between each predictor and the response except for the "chas" predictor.

(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 $H_0 : \beta_j = 0$?

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

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454, Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

We may reject the null hypothesis for "zn", "dis", "rad", "black" and "medv".

(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 on 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.



There is a difference between the simple and multiple regression coefficients. This difference is due to the fact that in the simple regression case, the slope term represents the average effect of an increase in the predictor, ignoring other predictors. In contrast, in the multiple regression case, the slope term represents the average effect of an increase in the predictor, while holding other predictors fixed. It does make sense for the multiple regression to suggest no relationship between the response and some of the predictors while the simple linear regression implies the opposite because the correlation between the predictors show some strong relationships between some of the predictors.

```
                zn      indus        nox         rm        age        dis
zn       1.0000000 -0.5338282 -0.5166037  0.3119906 -0.5695373  0.6644082
indus   -0.5338282  1.0000000  0.7636514 -0.3916759  0.6447785 -0.7080270
nox     -0.5166037  0.7636514  1.0000000 -0.3021882  0.7314701 -0.7692301
rm       0.3119906 -0.3916759 -0.3021882  1.0000000 -0.2402649  0.2052462
age     -0.5695373  0.6447785  0.7314701 -0.2402649  1.0000000 -0.7478805
dis      0.6644082 -0.7080270 -0.7692301  0.2052462 -0.7478805  1.0000000
rad     -0.3119478  0.5951293  0.6114406 -0.2098467  0.4560225 -0.4945879
tax     -0.3145633  0.7207602  0.6680232 -0.2920478  0.5064556 -0.5344316
ptratio -0.3916785  0.3832476  0.1889327 -0.3555015  0.2615150 -0.2324705
black    0.1755203 -0.3569765 -0.3800506  0.1280686 -0.2735340  0.2915117
```

```
lstat   -0.4129946  0.6037997  0.5908789 -0.6138083  0.6023385 -0.4969958
medv     0.3604453 -0.4837252 -0.4273208  0.6953599 -0.3769546  0.2499287
               rad        tax    ptratio      black      lstat       medv
zn      -0.3119478 -0.3145633 -0.3916785  0.1755203 -0.4129946  0.3604453
indus    0.5951293  0.7207602  0.3832476 -0.3569765  0.6037997 -0.4837252
nox      0.6114406  0.6680232  0.1889327 -0.3800506  0.5908789 -0.4273208
rm      -0.2098467 -0.2920478 -0.3555015  0.1280686 -0.6138083  0.6953599
age      0.4560225  0.5064556  0.2615150 -0.2735340  0.6023385 -0.3769546
dis     -0.4945879 -0.5344316 -0.2324705  0.2915117 -0.4969958  0.2499287
rad      1.0000000  0.9102282  0.4647412 -0.4444128  0.4886763 -0.3816262
tax      0.9102282  1.0000000  0.4608530 -0.4418080  0.5439934 -0.4685359
ptratio  0.4647412  0.4608530  1.0000000 -0.1773833  0.3740443 -0.5077867
black   -0.4444128 -0.4418080 -0.1773833  1.0000000 -0.3660869  0.3334608
lstat    0.4886763  0.5439934  0.3740443 -0.3660869  1.0000000 -0.7376627
medv    -0.3816262 -0.4685359 -0.5077867  0.3334608 -0.7376627  1.0000000
```

So for example, when "age" is high there is a tendency in "dis" to be low, hence in simple linear regression which only examines "crim" versus "age", we observe that higher values of "age" are associated with higher values of "crim", even though "age" does not actually affect "crim". So "age" is a surrogate for "dis"; "age" gets credit for the effect of "dis" on "crim".

(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

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

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

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
```

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

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


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

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


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

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


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

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


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

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


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

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


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

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


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

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


Call:
lm(formula = crim ~ poly(black, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-13.096  -2.343  -2.128  -1.439  86.790

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.6135     0.3536  10.218   <2e-16 ***
poly(black, 3)1 -74.4312     7.9546  -9.357   <2e-16 ***
poly(black, 3)2   5.9264     7.9546   0.745    0.457
poly(black, 3)3  -4.8346     7.9546  -0.608    0.544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.955 on 502 degrees of freedom
Multiple R-squared:  0.1498,    Adjusted R-squared:  0.1448
F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16


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

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


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

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
```

For “zn”, “rm”, “rad”, “tax” and “lstat” as predictor, the p-values suggest that the cubic coefficient is not statistically significant; for “indus”, “nox”, “age”, “dis”, “ptratio” and “medv” as predictor, the p-values suggest the adequacy of the cubic fit; for “black” as predictor, the p-values suggest that the quandratic and cubic coefficients are not statistically significant, so in this latter case no non-linear effect is visible.