Chapter 3

R Lab

```r
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
```

```r
names(boston)
```

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

Simple Linear Regression

```r
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
```

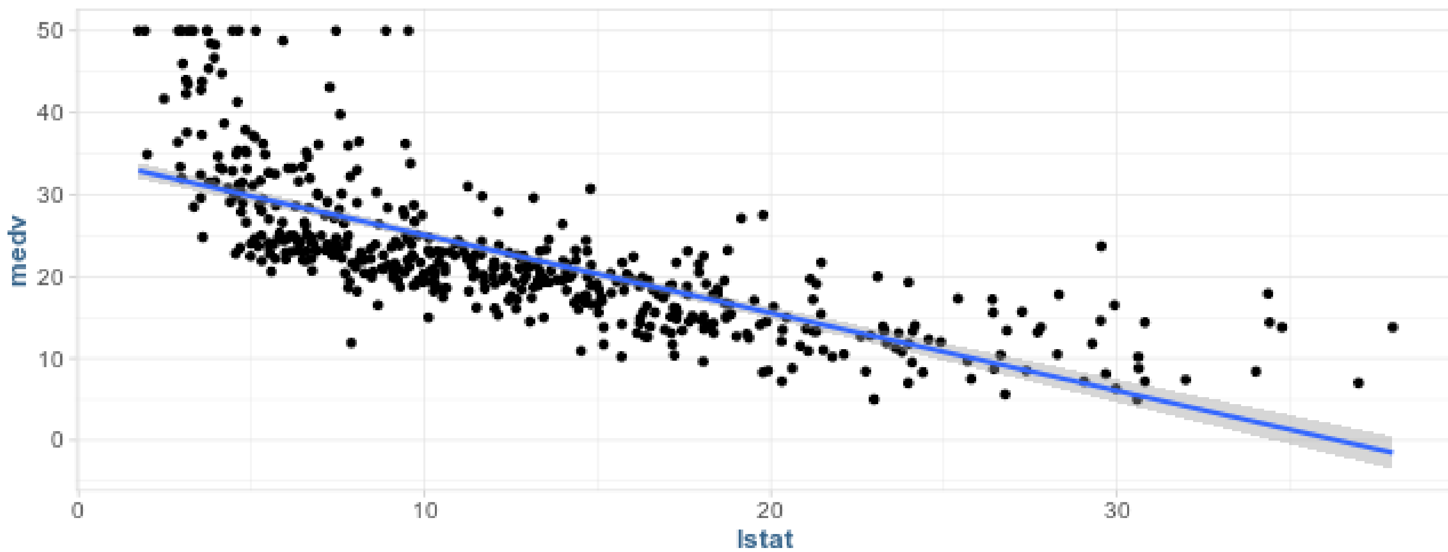
```r
predict(lm.fit, data.frame(lstat = c(5, 10, 15)),
        interval = "confidence")
```

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

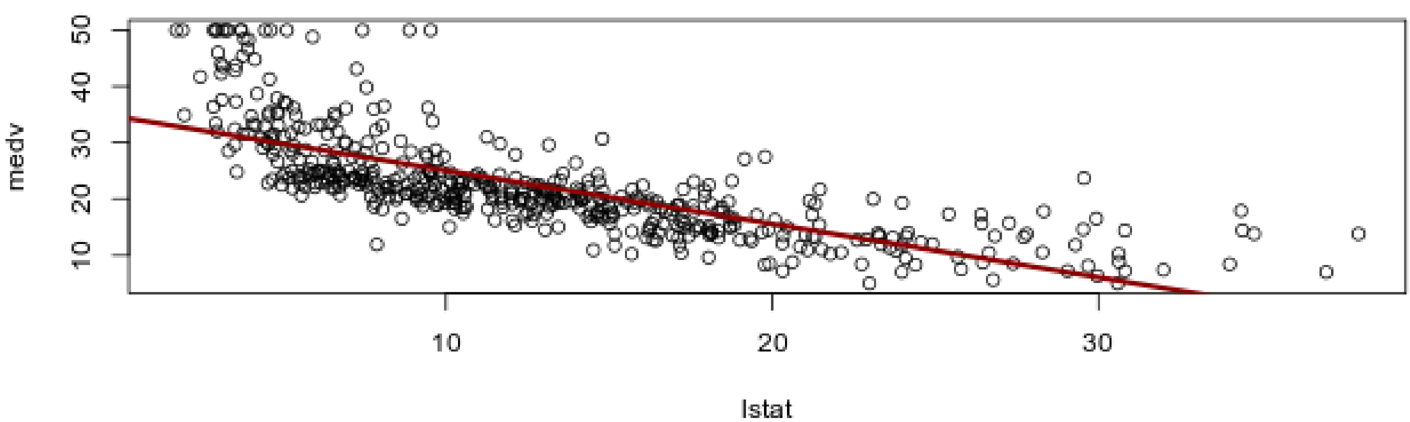
```r
predict(lm.fit, data.frame(lstat = c(5, 10, 15)),
        interval = "prediction")
```

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

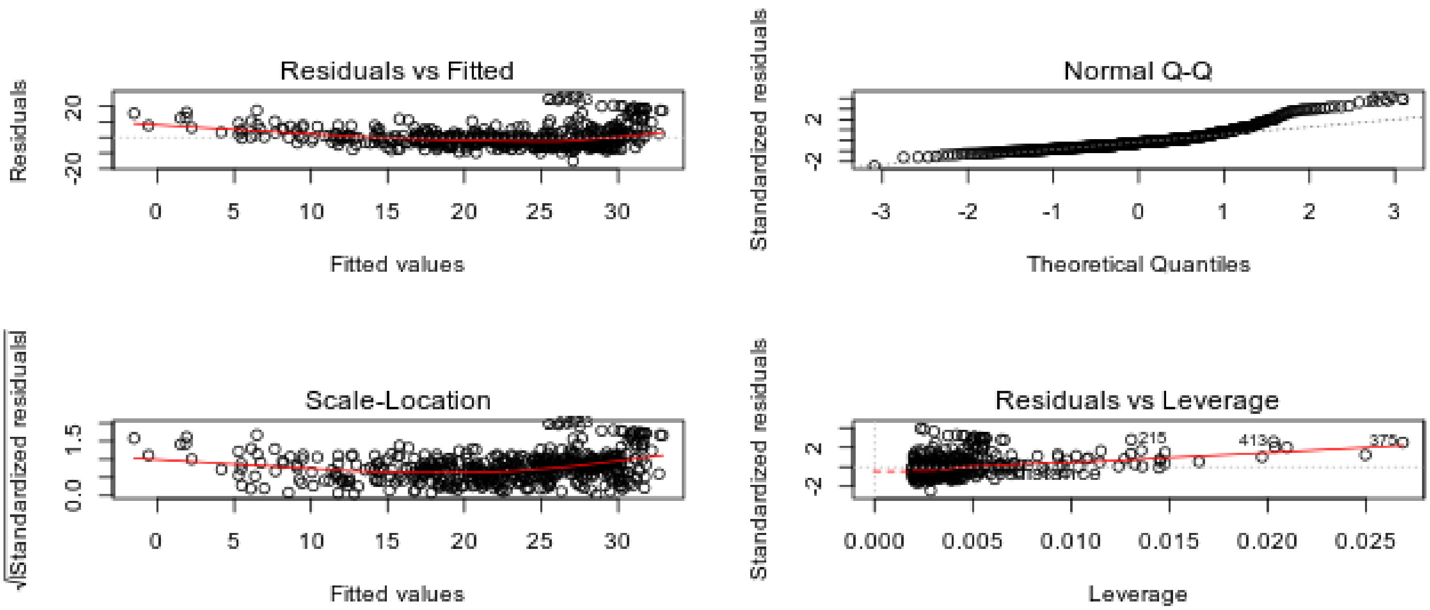
```r
ggplot(boston, aes(lstat, medv)) +
   geom_point() +
   geom_smooth(method = "lm")
```



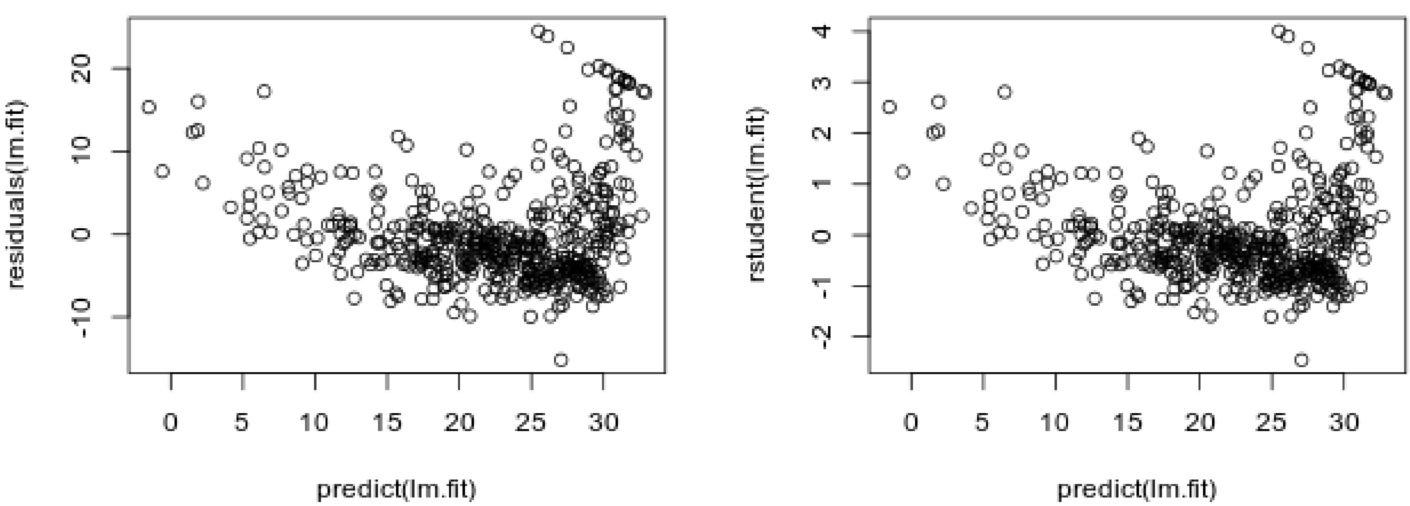

```r
with(boston, {
   plot(lstat, medv)
   abline(lm.fit, col ="darkred")
   abline(lm.fit, lwd = 3, col = "darkred")
})
```



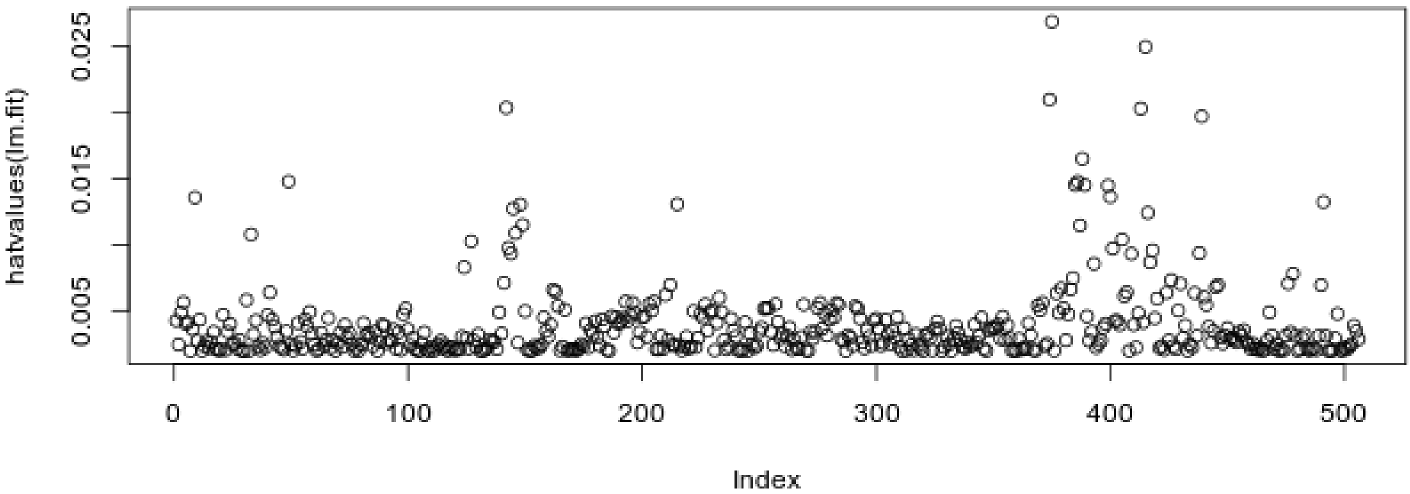

```r
par(mfrow = c(2,2))
with(boston, {
   plot(lm.fit)
})
```




```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
```

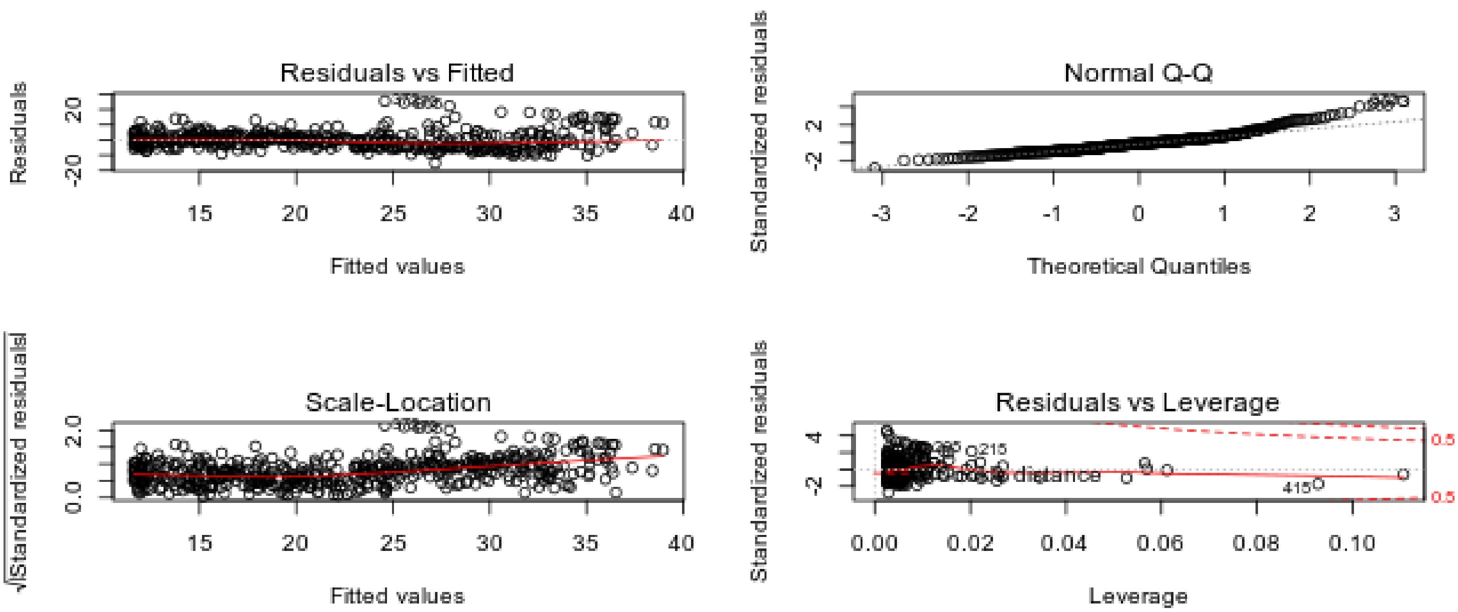
```r
lm.fit <- lm(medv ~ lstat, data = boston)
anova(lm.fit, lm.fit2)
```

```
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

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 ?

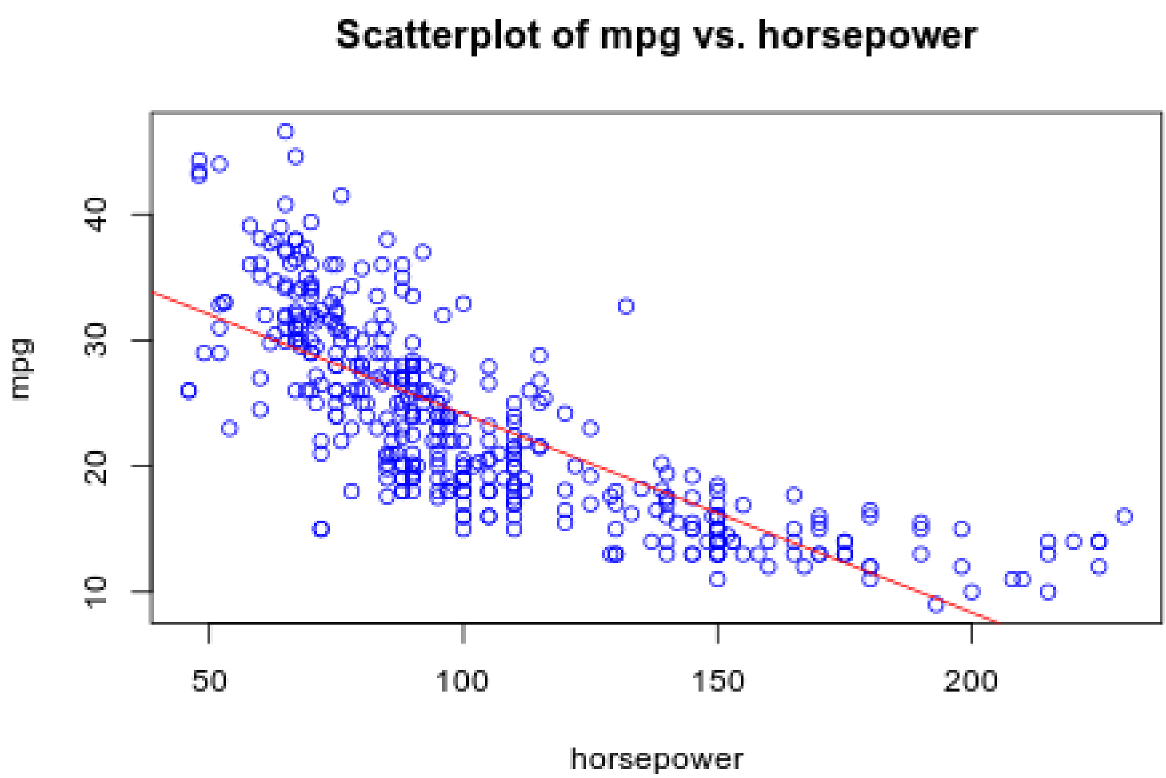
As the coeficient of "horsepower" is negative, the relationship is also negative. The more horsepower an automobile has the linear regression indicates the less mpg fuel efficiency the automobile will have.

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

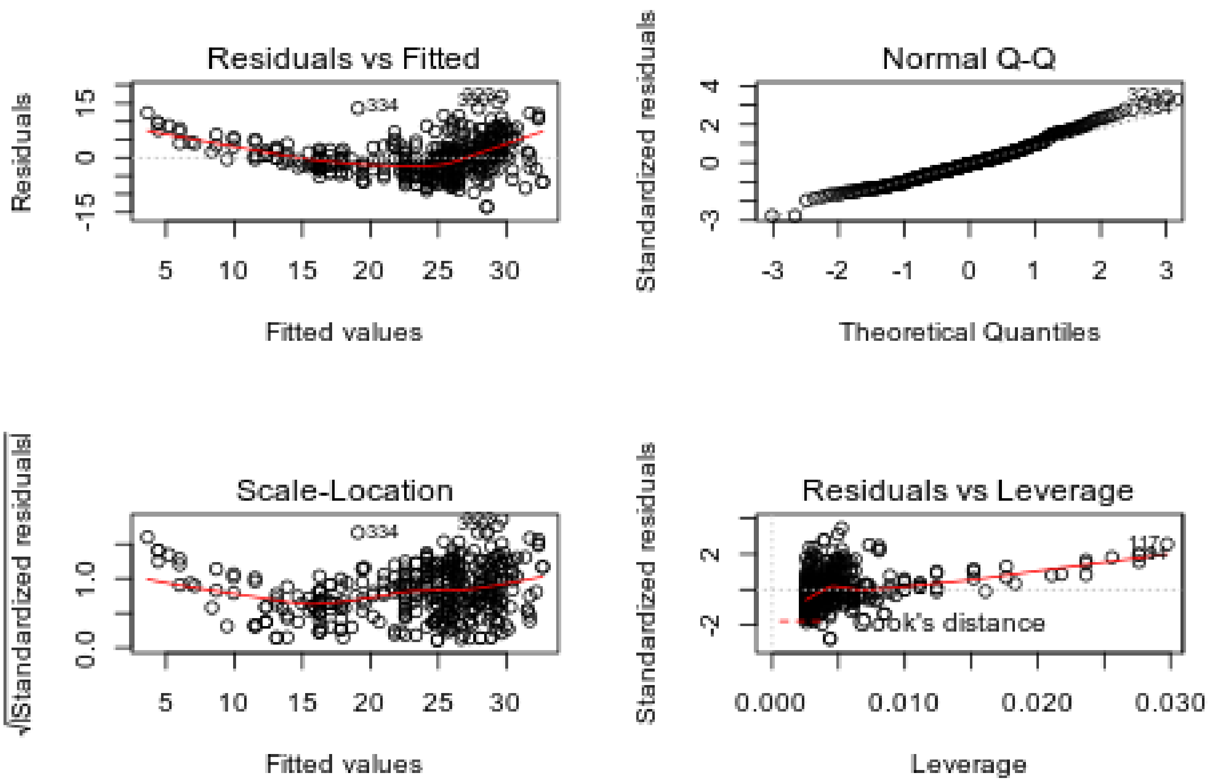
```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```
       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.