Tutorial 2

November 26, 2020

1 ISLR Chapter 3

1.1 Exercise 1

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

title

a) Describe the null hypothesis

In Table 3.4, the null hypothesis for "TV" is that in the presence of radio ads and newspaper ads, TV ads have no effect on sales. Similarly, the null hypothesis for "radio" is that in the presence of TV and newspaper ads, radio ads have no effect on sales.

b) Explain your conclusions based on the p-values in Table 3.4

On the one hand, the low p-values of TV and radio allow us to reject the null hypothesis for TV and radio. Hence, we believe that TV (radio) ads have an effect on sales in the presence of radio (TV) and newspaper ads.

On the other hand, the high p-value of newspaper does not allow us to reject the null-hypothesis. That is, possible effects of newspaper ads are not large enough to stand out from the estimation errors.

Here, remember! An insignificant hypothesis test-result is never informative about whether the null is true. We do not have an error-control for falsely "accepting" the null-hypothesis. We only have an error-control (by the significance level) for falsely rejecting the null-hypothesis.

1.2 Exercise 2

Carefully explain the main difference between the KNN classifier and KNN regression methods.

KNN classifier and KNN regression methods are closely related in formula. However, the final result of KNN classifier is the classification output for Y (qualitative), where as the output for a KNN regression predicts the quantitative value for f(X).

1.3 Exercise 3

Suppose we have a data set with five predictors:

$X_1 = GPA$

$X_2 = IQ$

$X_3 = Gender$ (1 for Female and 0 for Male)

$X_4 =$ Interaction between GPA and IQ

$X_5 =$ Interaction between GPA and Gender

The response variable (in thousands of dollars) is defined as:

$Y =$ starting salary after graduation

Suppose we use least squares to fit the model, and get $\hat{\beta_0} = 50$, $\hat{\beta_1} = 20$, $\hat{\beta_2} = 0.07$, $\hat{\beta_3} = 35$, $\hat{\beta_4} = 0.01$, $\hat{\beta_5} = 10$.

Which is equivalent to write:

$E[Y|X] = 50 + 20 * GPA + 0.07 * IQ + 35 * Gender + 0.01 * GPA * IQ - 10 * GPA * Gender$

3. a) Which answer is correct, and why?

```
i) For a fixed value of IQ and GPA, males earn more on average than females.
ii) For a fixed value of IQ and GPA, females earn more on average than males.
iii) For a fixed value of IQ and GPA, males earn more on average than females provided that th
iv) For a fixed value of IQ and GPA, females earn more on average than males provided that the
```

To answer this question, let us start by taking a derivative of Y wrt $Gender$ (assuming continuity):

$$\frac{\partial E[Y|X]}{\partial Gender} = 35 - 10 * GPA \Rightarrow \frac{\partial E[Y|X]}{\partial Gender} > 0 \iff GPA > 3.5 \tag{1}$$

For a fixed value of IQ and GPA, males ($Gender = 0$) earn more on average than females provided that the GPA is higher than 3.5. Thus, iii) is the correct answer.

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

[23]:
```
Y_hat <- 50 + 20*4 + 0.07*110 + 35*1 + 0.01*4*110 - 10*4
Y_hat
```

137.1

3.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. We must examine the p-value of the regression coefficient to determine if the interaction term is statistically significant or not.

1.4 Exercise 8

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

8.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.

[24]:
```
# Store data into dataframe college
Auto <- read.csv("Auto.csv", header=T, na.strings="?")

# Remove missing values from the data
Auto = na.omit(Auto)

# Perform linear regression
fit.lm <- lm(mpg ~ horsepower, data=Auto)

# Use summary function to print the results
summary(fit.lm)
```

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

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

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

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

i) Is there a relationship between the predictor and the response? Yes, there is. By looking at the estimated coefficient for horsepower (-0.157845) and its p-value ($< 2e-16$).

ii) How strong is the relationship between the predictor and the response? The relationship is strong as indicated by the very small p-value ($< 2e-16$).

iii) Is the relationship between the predictor and the response positive or negative? The relationship between mpg and horsepower is negative as indicated by the coefficient (-0.157845).

The linear regression predicts that, on average, the more horsepower an automobile has the less mpg fuel efficiency the automobile will have.

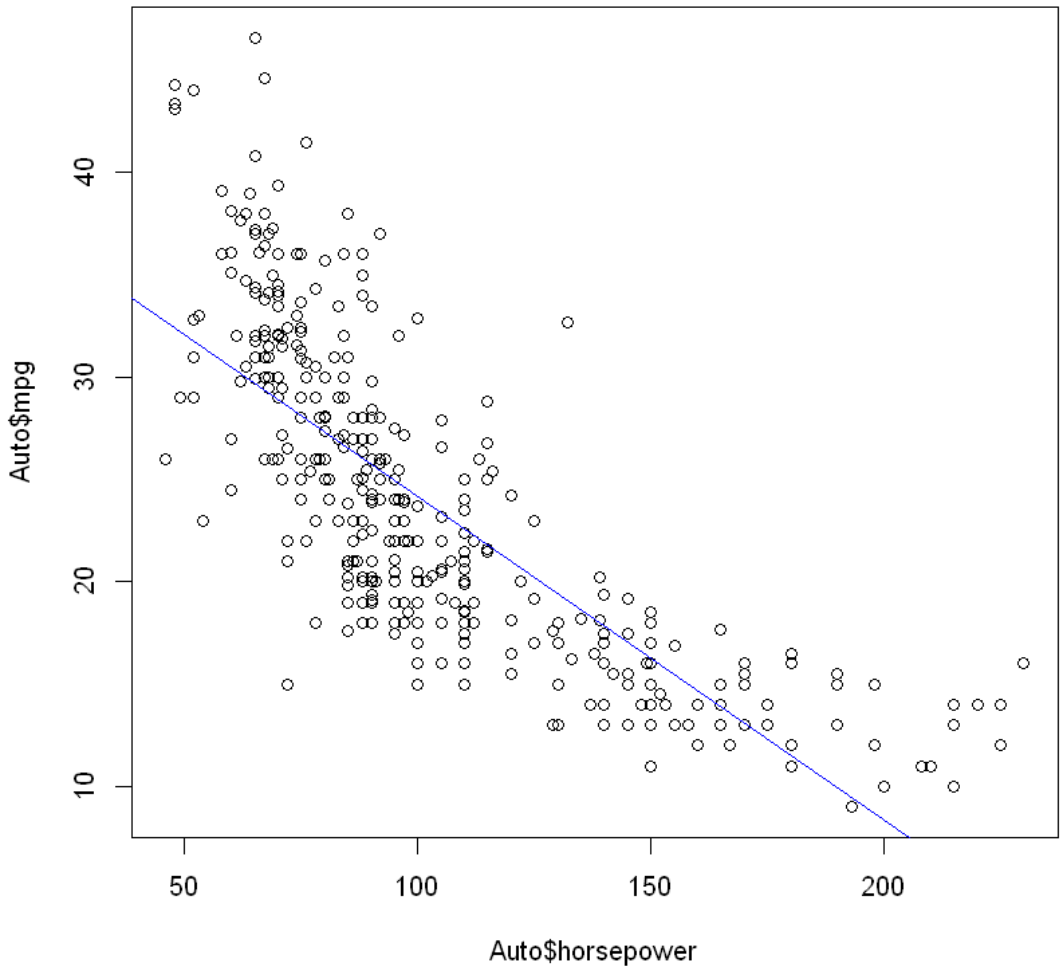
iv) What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

[25]:
```
# Horsepower of 98
new <- data.frame(horsepower = 98)

# confidence interval
print(predict(fit.lm, new, interval = "confidence"))

#prediction interval
print(predict(fit.lm, new, interval = "prediction"))
```

```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

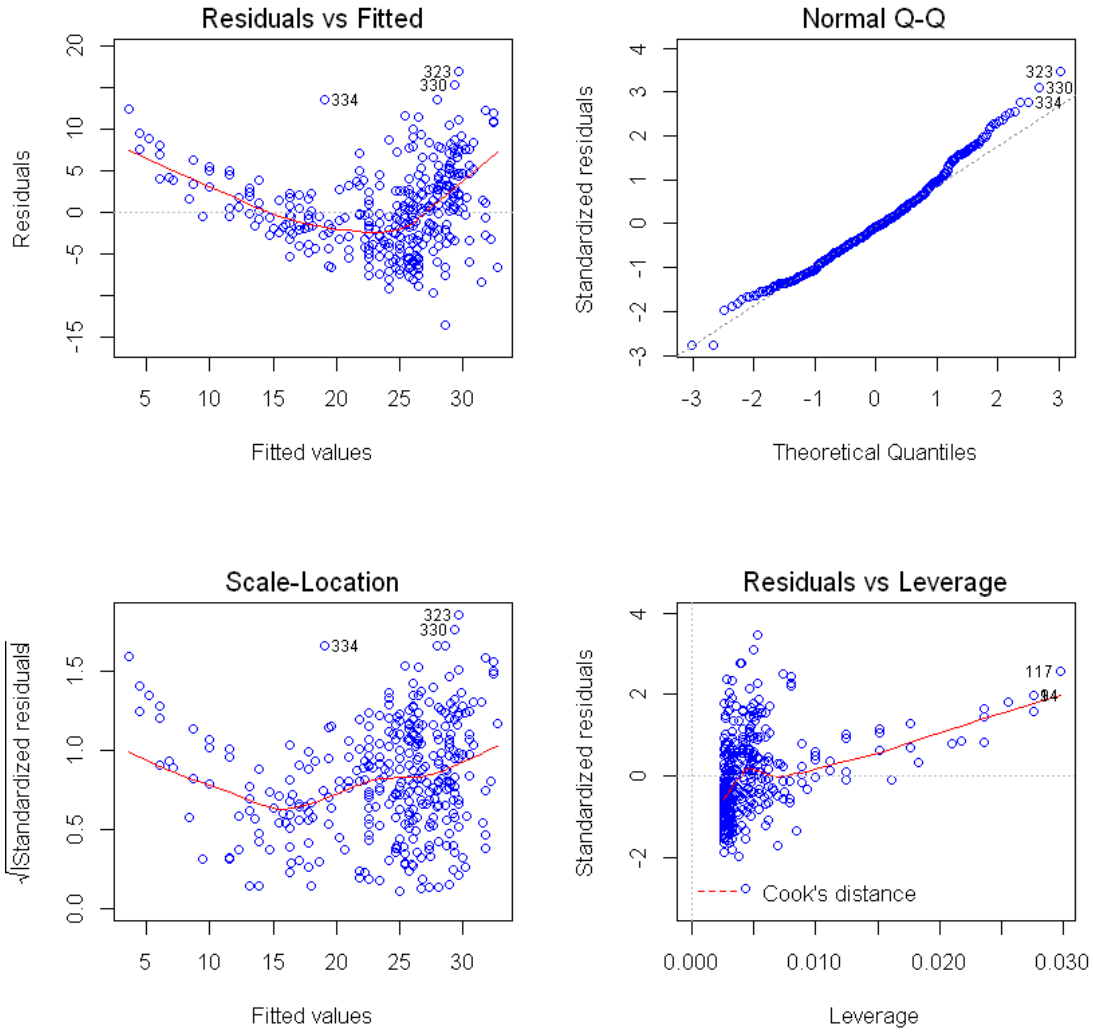
8. b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

[26]:
```
plot(Auto$horsepower, Auto$mpg)
abline(fit.lm, col="blue")
```



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

[27]:
```
par(mfrow=c(2,2))
plot(fit.lm, col='blue')
```



Based on the residuals plots, there is some evidence of non-linearity and existance of outliers.

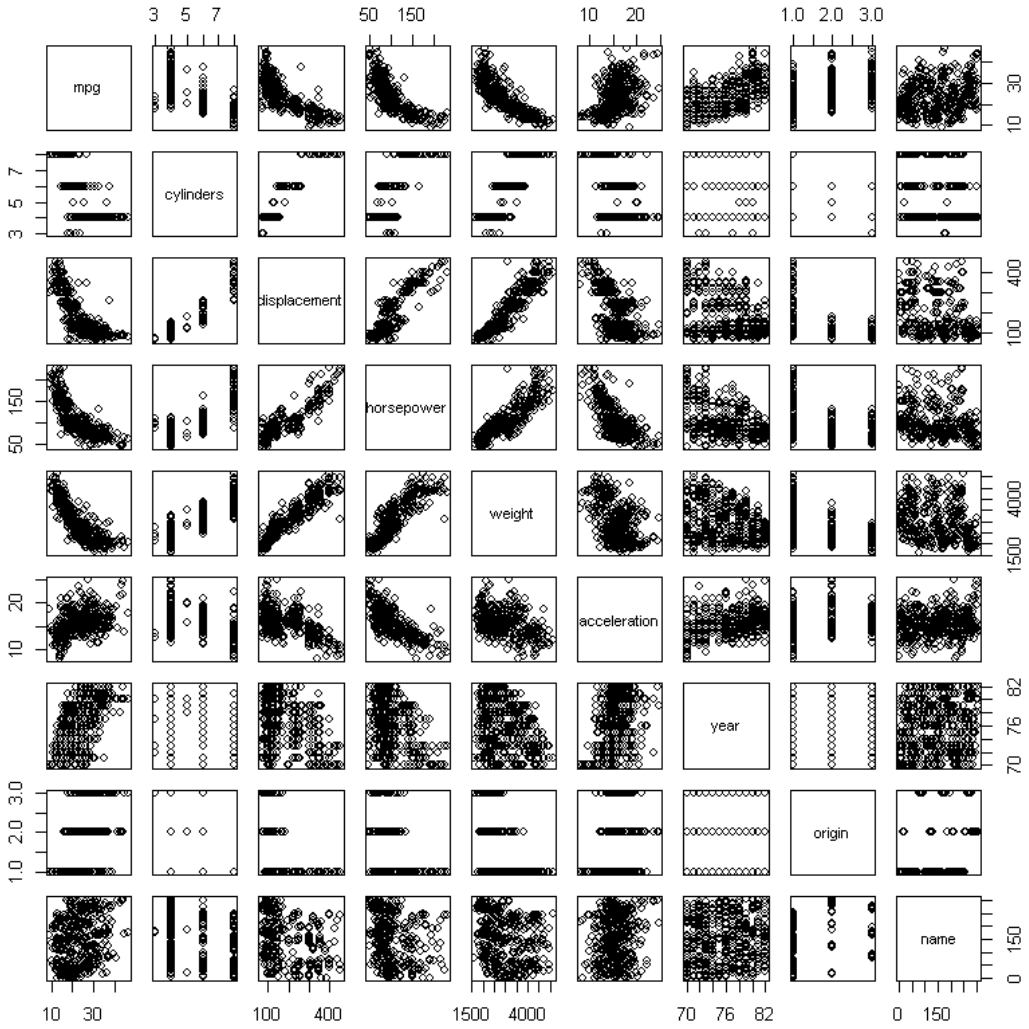
1.5 Exercise 9

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

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

[28]:
```
# Store data into dataframe college
Auto <- read.csv("Auto.csv", header=T, na.strings="?")

# Remove missing values from the data
Auto = na.omit(Auto)
```

```
# Produce scatterplot matrix
pairs(Auto)
```



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

[29]:
```
print(cor(subset(Auto, select=-name)))
```

```
                    mpg  cylinders displacement horsepower     weight
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
```

```
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
year            0.2903161  1.0000000  0.1815277
origin          0.2127458  0.1815277  1.0000000
```

9. 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.

[30]:
```
# Perform multiplie linear regression
fit.lm <- lm(mpg ~ .-name, data=Auto)

# Print results
summary(fit.lm)
```

```
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
```

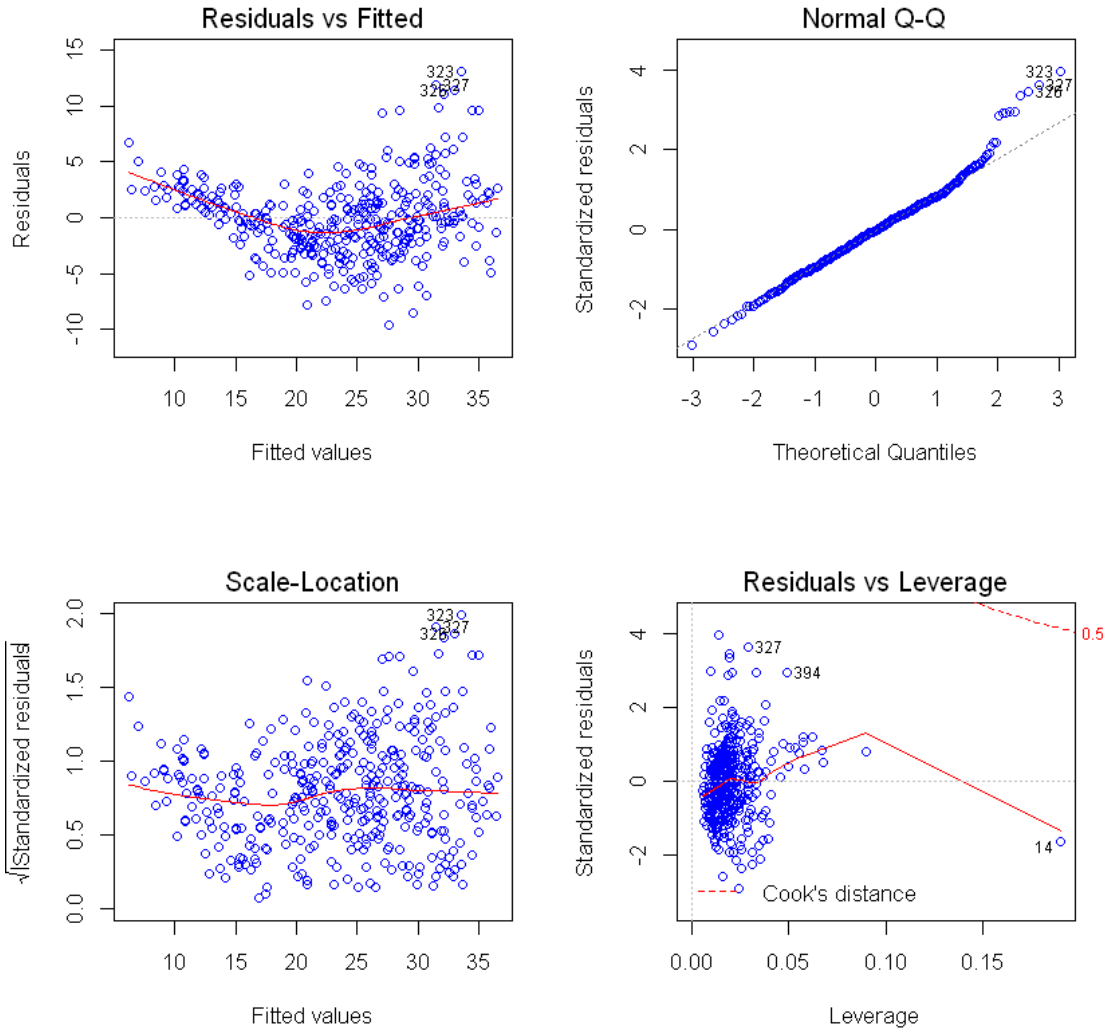
9. c. i) Is there a relationship between the predictors and the response? Yes, there is a relationship between the predictors and the response. By testing the null hypothesis of whether all the regression coefficients are zero, we can see that the F-statistic is big and its p-value is close to zero, indicating evidence against the null hypothesis.

9. c. ii) Which predictors appear to have a statistically significant relationship to the response? Looking at the p-values associated with each predictor's t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.

9. c. iii) What does the coefficient for the year variable suggest? The regression coefficient for year suggests that one more year increases mpg, on average, 0.75. In other words: on average, cars become more fuel efficient every year by 0.75 mpg / year.

9. 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 plot identify any observations with unusually high leverage?

[31]:
```
par(mfrow=c(2,2))
plot(fit.lm, col='blue')
```



We can learn the following four main messages from the diagnostic plots:

First: The residual plot (Residuals vs Fitted, top left) shows some systematic deviations of the residuals from 0. The reason is that we are imposing a straight line fit for a conditional mean function $E[Y|X] = f(X)$ which is non-linear. This results in a systematic underestimation of the true conditional mean function for large and small values of 'horsepower'.

Second: The "Residuals vs Leverage" plot (bottom right) shows that there are some potential outliers that we can see when: standardized residuals are below -2 or above +2.

Third: Furthermore, the "Residuals vs Leverage" plot shows also potentially problematic "high-leverage" points with leverage values heavily exceeding the rule-of-thumb threshold (p+ 1)/n=8/392=0.02.

Hence, all points with simultaneously high-leverages and large absolute standardized residuals should be handled with care since these may distort the estimation.

Fourth: The Normal-qq plot (top right) suggests non-normally distributed residuals - particularly the upper tail deviates from that of a normal distribution.

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

[32]:
```
fit.lm0 <- lm(mpg ~ horsepower+cylinders+year+weight:displacement, data=Auto)
summary(fit.lm0)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-9.1046 -2.8861 -0.2415  2.3967 15.3221

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.343e+01  5.043e+00  -2.663  0.00807 **
horsepower          -3.914e-02  1.278e-02  -3.063  0.00234 **
cylinders           -1.358e+00  3.233e-01  -4.201 3.31e-05 ***
year                 6.661e-01  6.019e-02  11.067  < 2e-16 ***
weight:displacement -3.354e-06  1.352e-06  -2.480  0.01355 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.985 on 387 degrees of freedom
Multiple R-squared:  0.7419,Adjusted R-squared:  0.7393
F-statistic: 278.2 on 4 and 387 DF,  p-value: < 2.2e-16
```

[33]:
```
fit.lm1 <- lm(mpg~horsepower+cylinders+year+weight*displacement, data=Auto)
summary(fit.lm1)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-9.7530 -1.8228 -0.0602  1.5780 12.6133

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.210e+00  3.819e+00  -0.579  0.56316
horsepower          -3.396e-02  9.560e-03  -3.552  0.00043 ***
cylinders            2.072e-01  2.914e-01   0.711  0.47756
year                 7.858e-01  4.555e-02  17.250  < 2e-16 ***
weight              -1.084e-02  6.346e-04 -17.076  < 2e-16 ***
```

```
displacement        -7.947e-02  9.905e-03  -8.023 1.26e-14 ***
weight:displacement  2.431e-05  2.141e-06  11.355  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.976 on 385 degrees of freedom
Multiple R-squared:  0.8568,Adjusted R-squared:  0.8546
F-statistic: 384.1 on 6 and 385 DF,  p-value: < 2.2e-16
```

The two regressions we ran teach us several important lessons: **First**, there is a difference between using the symbol $A : B$ and the symbol $A * B$ when runnig a regression. While the first includes only the interaction term between the variable A and B, the second one also includes the stand-alone variables A and B.

Second, if one wants to evaluate the "true" effect of an interaction term, one should always include the stand-alone terms. We can see exactly that by comparing the two regression outputs for *weight* and *displacement*. While the first shows a significant negative coefficient for their interaction, the second shows a significant positive coefficient.

Why is that the case? We know from exercise 9.b) that both variables are negatively correlated with mpg but positively correlated between themselves, thus, if we only include the interaction term in the regression this will capture the negative relation. However, we will only obtain the "true" interaction term coefficient by including the stand-alone terms, in the second output we can see that contrarily to the stand-alone coefficients which are significantly negative, the interaction one is significantly positive.

Third, we have that *displacement* "measures overall volume in the engine as a factor of cylinder circumfrance, depth and total number of *cylinders*". Hence, capturing the cylinders effect via the displacement variable should be expected. We see this first by observing that cylinders is highly correlated with displacement. And second, by noting that when including both variables (second regression) we render the coefficient of cylinders insignificant.