

HW3 R Markdown

Santanu Mukherjee

9/26/2021

R Markdown

Chapter 2 page 53:

Q5

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Advantages The advantages for a very flexible approach for regression or classification are obtaining a better fit for generalized non-linear models, whereby the objective is to reduce bias.

Disadvantages The disadvantages for a very flexible approach for regression or classification are that it leads to overfitting, thus requiring the introduction of greater number of parameters and hence greater amount of noise.

Preference for more flexible approach So, in prediction problems, where the main objective is to predict and the interpretation of the data (results) does not matter that much, a more flexible approach would be preferred to a less flexible approach.

Preference for less flexible approach Now, where interpretation of the data (results) is important, a less flexible approach is preferred to a more flexible approach, so that the chances of overfitting (increase of error or noise) is greatly reduced.

Q6

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Parametric VS Non-parametric statistical learning

In the parametric approach, the objective is to assume a functional form and estimate the parameters in our regression or classification study.

However, in a non-parametric approach, there is no predetermined functional form and hence no parametric form is assumed. So, non-parametric regression or classification requires larger sample sizes compared to parametric approach.

Advantages and Disadvantages

So, the advantages of a parametric approach to regression or classification are that the model is much more simplified and requires less parameters and so, not many observations are required compared to a non-parametric approach.

However, the disadvantages of a parametric approach to regression or classification are potential inaccuracies to the estimated parameters of the assumed functional form if the function assumed is wrong.

For non-parametric approach, the advantage is that no functional approach is assumed and based on data, there are times when it gives much more accurate results.

For non-parametric approach, the disadvantage is there are many parameters and there can be overfitting.

Q7

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

```
##   X1 X2 X3   Y
## 1  0  3  0  Red
## 2  2  0  0  Red
## 3  0  1  3  Red
## 4  0  1  2 Green
## 5 -1  0  1 Green
## 6  1  1  1  Red
```

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

- Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.
- What is our prediction with $K = 1$? Why?
- What is our prediction with $K = 3$? Why?
- If the Bayes decision boundary in this problem is highly non linear, then would we expect the best value for K to be large or small? Why?

Answers

- The Euclidean distance has been computed and the column Euclidean_Distance displays that.

```
##   X1 X2 X3   Y Euclidean_Distance
## 1  0  3  0  Red              3.000
## 2  2  0  0  Red              2.000
## 3  0  1  3  Red              3.162
## 4  0  1  2 Green             2.236
## 5 -1  0  1 Green             1.414
## 6  1  1  1  Red              1.732
```

- Green. Observation #5 is the single nearest neighbor for $K = 1$.
- Red. Observations #2, #5, #6 are the closest neighbors for $K = 3$. For 2 observations (# 2 and # 6) it is Red, 5 is Green. So the answer is Red.
- Small. A small K would perform better because it would be 'flexible' for a non-linear decision boundary, whereas a large K would have a smoothing effect and would try to fit a more linear boundary because it takes more points into consideration.

Chapter 3 page 120:

Q1

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.

The null hypotheses associated with table 3.4 are that advertising budgets of “TV”, “radio” or “newspaper” do not have an effect on sales. More precisely $H(1)_0 : \beta_1 = 0$, $H(2)_0 : \beta_2 = 0$ and $H(3)_0 : \beta_3 = 0$.

The corresponding p-values are highly significant for “TV” and “radio” and not significant for “newspaper”; so we reject the null hypothesis.

Q2

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

The KNN classifier is typically used to solve classification problems (those with a qualitative response) by identifying the neighborhood of x_0 and then estimating the conditional probability $P(Y = j | X = x_0)$ for class j as the fraction of points in the neighborhood whose response values equal j .

However, the KNN regression method is used to solve regression problems (those with a quantitative response) by again identifying the neighborhood of x_0 and then estimating $f(x_0)$ as the average of all the training responses in the neighborhood.

Q3

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

- a. Which answer is correct, and why?
 - i. For a fixed value of IQ and GPA, males earn more on average than females.
 - ii. For a fixed value of IQ and GPA, females earn more on average than males.
 - iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
 - iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.
- b. Predict the salary of a female with IQ of 110 and a GPA of 4.0.
- 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.

Answers

- a. The least square line is given by

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \times IQ - 10GPA \times Gender$$
 so for the *MALES*, it is

$$\hat{y} = 50 + 20GPA + 0.07IQ + 0.01GPA \times IQ,$$
 and for the *FEMALES*, it is

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

So the starting salary for *MALES* is higher than for *MALES* on average if and only if $50 + 20GPA \geq 85 + 10GPA$ which is equivalent to $GPA \geq 3.5$.

Once the GPA is high enough, males earn more on average.

Therefore *iii.* is the right answer.

$$b. Y(\text{Gender} = 1, \text{IQ} = 110, \text{GPA} = 4.0) = 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 (4 * 110) - 10 * 4 = \mathbf{137.1}$$

So, **\$137,100** is the starting salary.

- c. False. To verify if the GPA/IQ has an impact on the quality of the model we need to test the hypothesis $H_0 : \beta_4 = 0$ and look at the p-value associated with the t or the F statistic to draw a conclusion.

Q4

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.

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

- Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. 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.
- Answer (a) using test rather than training RSS.
- 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.
- Answer (c) using test rather than training RSS.

Answers

- 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.
- So, when we use test data, even in that case the **test RSS** depends upon the test data, so we do not have enough information to come to a conclusion. 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.
- Polynomial regression has lower **training RSS** than the linear fit because of higher flexibility. Now, irrespective of what the underlying true relationship is, it is obvious that the more flexible model will have closer follow points and reduce **training RSS**.
- Based on what is given, there is not enough information to tell which **test RSS** would be lower for either regression given the question where it states "*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**.
On the other hand, if it is closer to cubic than linear, the cubic regression **test RSS** could be lower than the linear regression **test RSS**. This is due to bias-variance tradeoff; it is not clear what level of flexibility will fit data better.

Q5

Consider the fitted values that result from performing linear regression without an intercept.

In this setting, the i th fitted value takes the form $\hat{y}_i = x_i \hat{\beta}$,

$$\text{where } \hat{\beta} = (\sum_{i=1}^n x_i y_i) / (\sum_{i=1}^n x_i^2).$$

Show that we can write $\hat{y}_i = \sum_{j=1}^n a_j y_j$

What is a_i ?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

Answers

So , now we have $\hat{y}_i = (x_i) * (\sum_{j=1}^n x_j y_j) / (\sum_{i=1}^n x_i^2)$,

where $a_j = (x_i) * (x_j) / (\sum_{i=1}^n x_i^2)$

Q8

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

- 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:
 - Is there a relationship between the predictor and the response?
 - How strong is the relationship between the predictor and the response?
 - Is the relationship between the predictor and the response positive or negative?
 - What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?
- Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.
- Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Answers

- Performing a simple linear regression on the Auto dataset

```
##
## 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. The answer to this question can be given by testing the hypothesis $H_0: \beta_i = 0 \forall i$. The p-value is less than 0.05 and hence we can reject the null hypothesis and can conclude there is statistically significant relationship between “horsepower” and “mpg”.
- ii. The R-square is 0.6059 which means that 60.59% of the variability in “mpg” can be explained using “horsepower” and that is why this shows as a strong relationship.
- iii. As the coefficient of predictor “horsepower” is *negative*, the relationship between response and predictor is negative. The more horsepower an automobile has the linear regression indicates the less mpg fuel efficiency the automobile will have.
- iv. The below lists the predicted *mpg* associated with a “horsepower” of 98, along with the 95% confidence and prediction intervals.

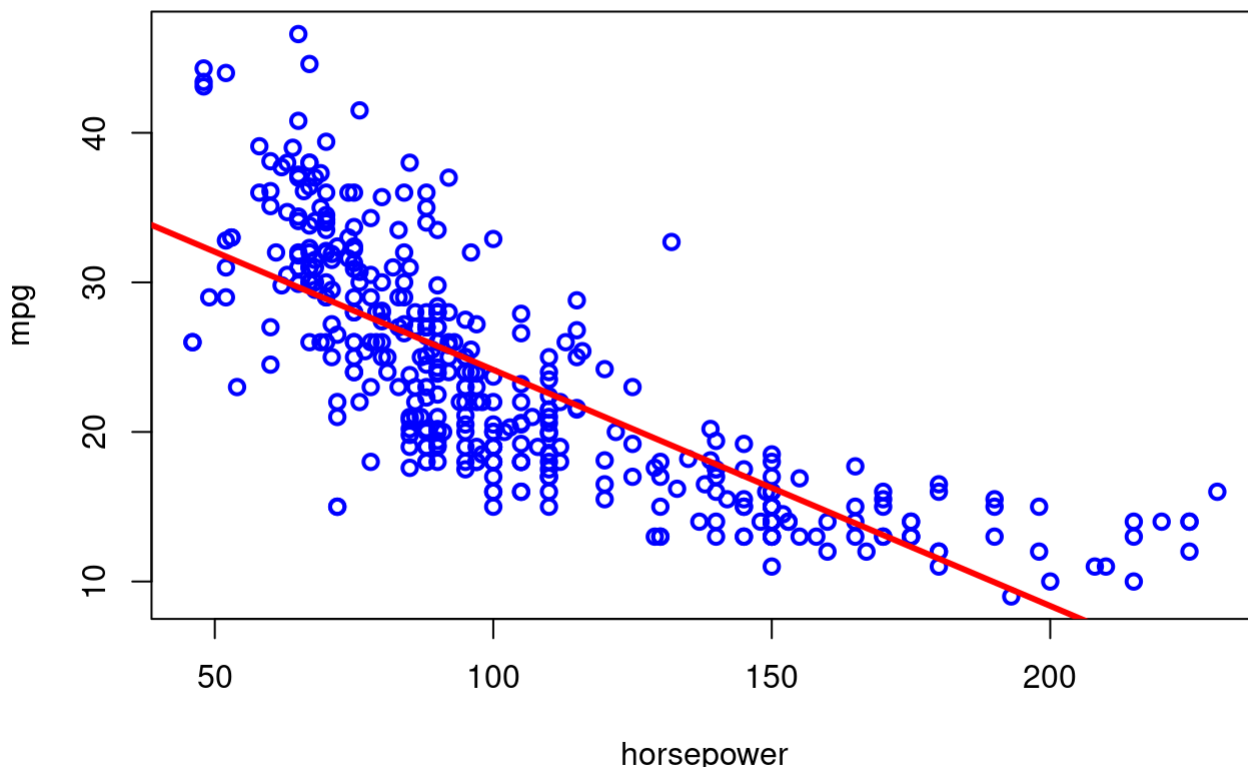
```
##          fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

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

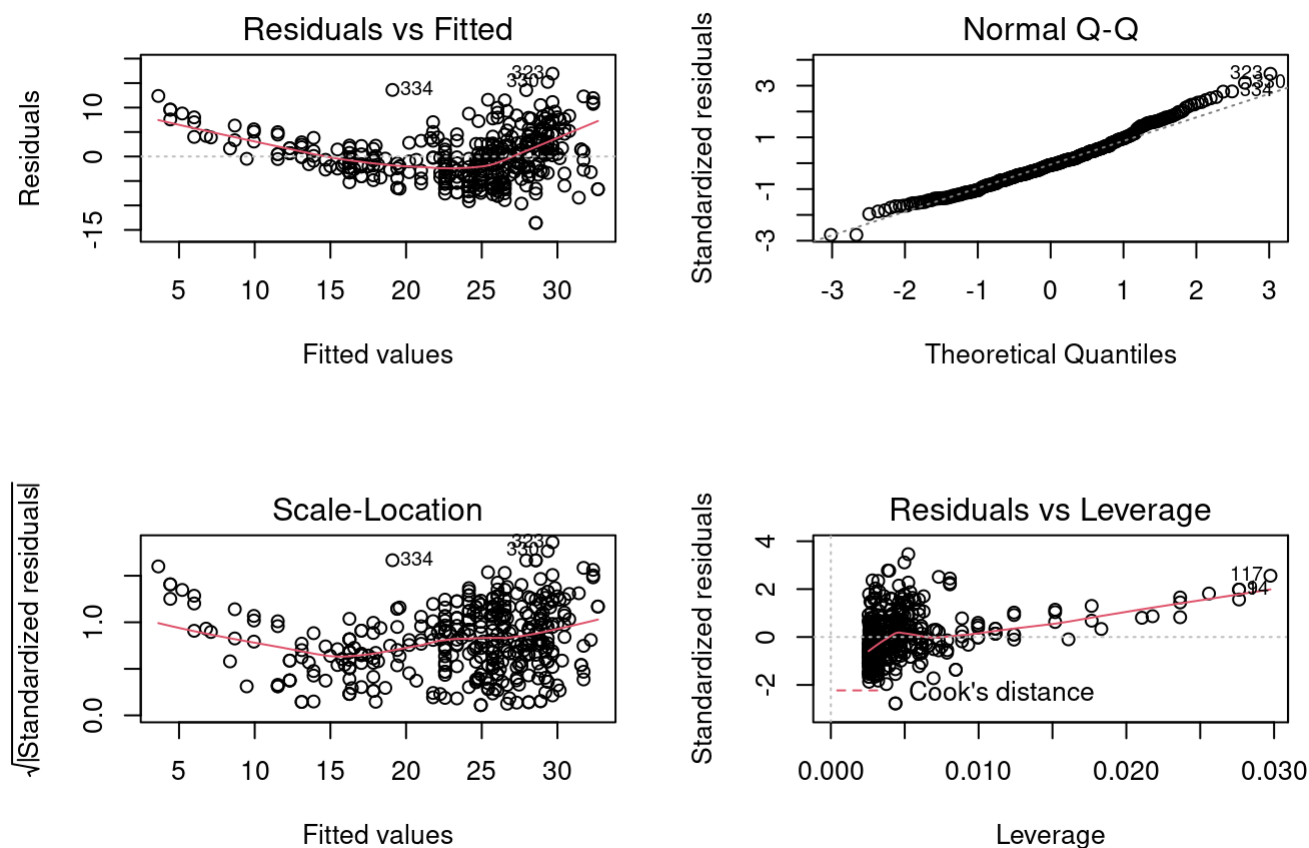
Predicted value of mpg is 24.46708. Confidence interval is (23.97308, 24.96108) and prediction interval is (14.8094, 34.12476).

- b. The below shows the graph between the response and the predictor.

Scatterplot of mpg vs. horsepower



C. Plot() function to produce diagnostics plots of the least squares regression fit.

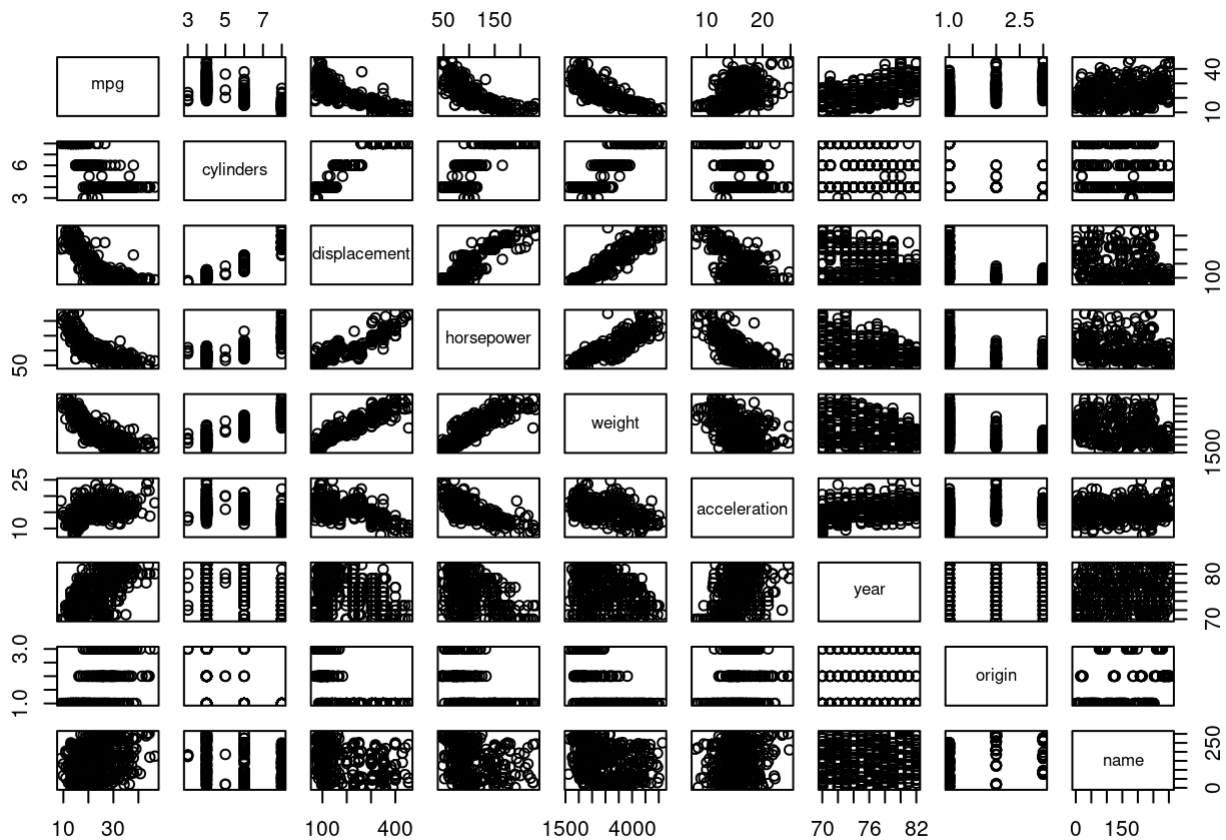


The plot of residuals versus fitted values indicates that the data is non-linear. The plot of standardized residuals versus leverage indicates the presence of a few outliers and a few high leverage points.

Q9

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

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



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

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

```
##           mpg  cylinders displacement horsepower   weight
## mpg      1.0000000 -0.7776175  -0.8051269 -0.7784268 -0.8322442
## cylinders -0.7776175  1.0000000   0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233   1.0000000  0.8972570  0.9329944
## horsepower -0.7784268  0.8429834   0.8972570  1.0000000  0.8645377
## weight     -0.8322442  0.8975273   0.9329944  0.8645377  1.0000000
## acceleration 0.4233285 -0.5046834  -0.5438005 -0.6891955 -0.4168392
## year        0.5805410 -0.3456474  -0.3698552 -0.4163615 -0.3091199
## origin      0.5652088 -0.5689316  -0.6145351 -0.4551715 -0.5850054
##
## acceleration      year      origin
## mpg              0.4233285  0.5805410  0.5652088
## cylinders         -0.5046834 -0.3456474 -0.5689316
## displacement      -0.5438005 -0.3698552 -0.6145351
## horsepower        -0.6891955 -0.4163615 -0.4551715
## weight            -0.4168392 -0.3091199 -0.5850054
## acceleration      1.0000000  0.2903161  0.2127458
## year              0.2903161  1.0000000  0.1815277
## origin            0.2127458  0.1815277  1.0000000
```


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

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

```
##
## 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. The answer to this question can be given by testing the hypothesis $H_0: \beta_i = 0 \forall i$. The p-value is less than 0.05 and hence we can reject the null hypothesis and can conclude there is statistically significant relationship between “mpg” and other predictors.

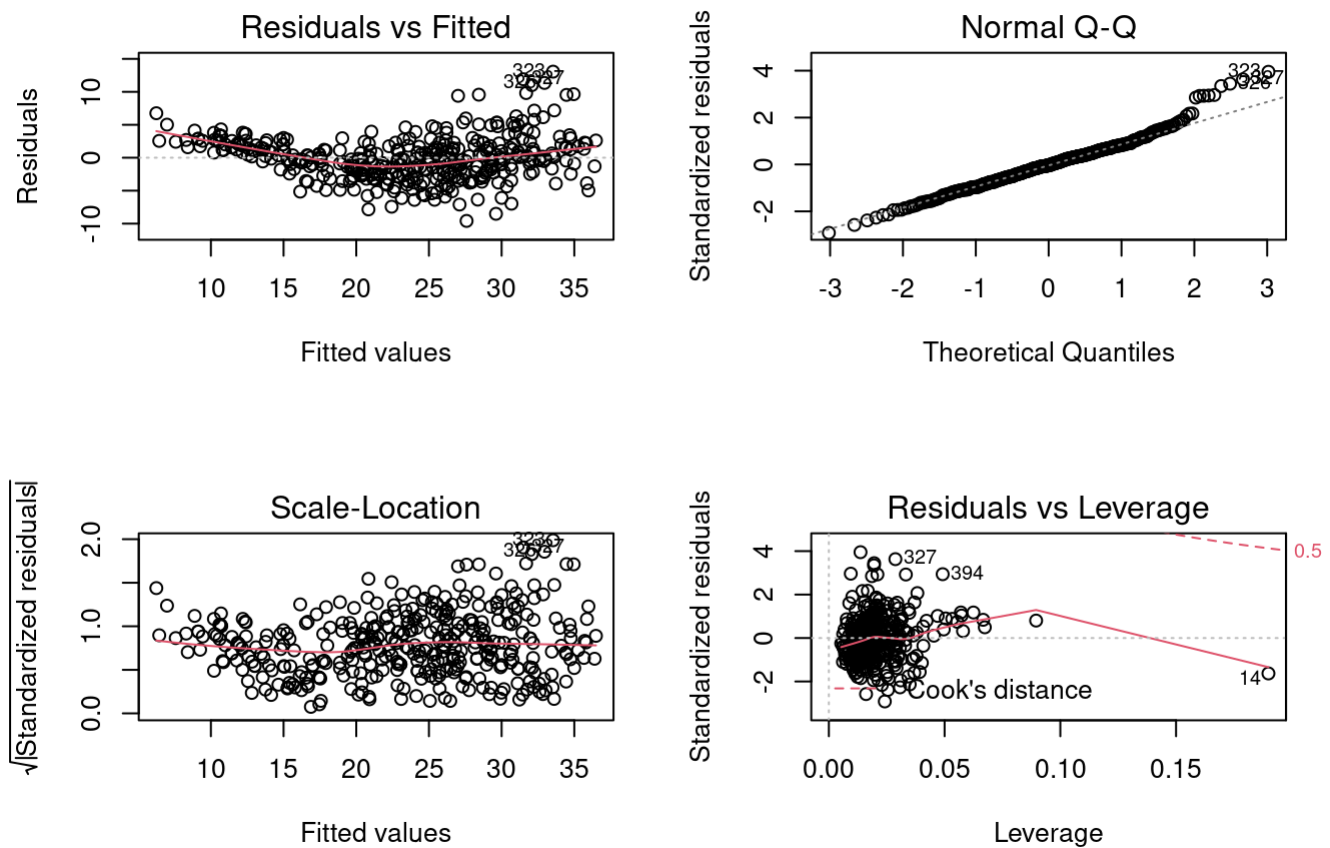
ii. Which predictors appear to have a statistically significant relationship to the response ?

The answer to this question can be given by checking the p-values associated with each predictor’s t-statistic. We may conclude that all predictors are statistically significant except “cylinders”, “horsepower” and “acceleration”.

iii. What does the coefficient for the “year” variable suggest ?

The coefficient of the “year” variable suggests that the average effect of an increase of 1 year is an increase of 0.750773 in “mpg” (all other predictors remaining constant). In other words, cars become more fuel efficient every year by almost 0.75 mpg / year.

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 ?



The plot of residuals versus fitted values indicates the presence of mild non-linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers and one high leverage point (point 14).

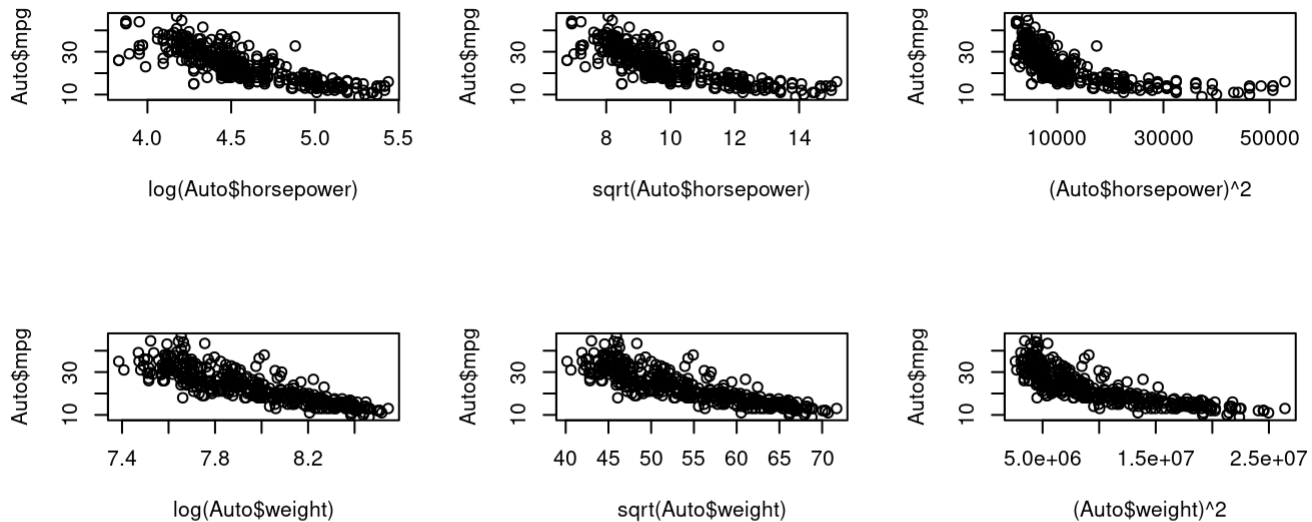
- e. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant ?

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

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto[, 1:8])
##
## 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
```

From the p-values, we can see that the interaction between displacement and weight is statistically significant, while the interaction between cylinders and displacement is not.

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



Examined both “horsepower” and “cylinders” as predictors. It seems that the log transformation gives the most linear looking plot.

Q11

Q11. 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.

- 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
```

Based on the summary above, here we have a value of 1.9939 for $\hat{\beta}$, a value of 0.1065 for the standard error, a value of 18.73 for the t-statistic and a very small p-value. As the p-value is very small, we 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
```

Based on the summary above, here we have a value of 0.39111 for $\hat{\beta}$, a value of 0.02089 for the standard error, a value of 18.73 for the t-statistic and a very small p-value. As the p-value is very small, we reject H_0 .

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

Based on the linear regression results, We see that there is 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, $y = 2x + \varepsilon$ can also be written as $x = 0.5(y - \varepsilon)$.

- d. Question

- (d) For the regression 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

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

Question 11 D - Question

(d) We have

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

Question 11 D - Answer

Now let us verify this result numerically.

[1] 18.72593

We see that the t above is exactly the t -statistic given in the summary of “lmQ11b”.

- 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 .

Answer : Based on the data, it can be said 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
```

Answer : Based on the data, we see that t -statistic for “lmQ11f1” and “lmQ11f2” are both equal to 18.56.