Summarise Course/Methods

Siger

4 janvier 2019

Si Dieu est infini, alors je suis une partie de Dieu sinon je serai sa limite...

Table des matières

1 Conceptual **1**
1.1 1) . . . 1
1.2 2. . . . 2
1.3 3. . . . 2
1.4 4. . . . 2
1.5 6. . . . 2
1.6 7. . . . 3

2 Applied **3**
2.1 8. . . . 3
2.2 9 . . . 5

1 Conceptual

1.1 1)

	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

Figure 1 – Analyse p-value

Noting $H_{0_{variable}}$ the null hypotheses to which the p-values correspond then for $variable \in \{TV, radio, newspaper\}$ H_0 is true if and only if there is no relationship between *sales* and *variable*.
Regarding $p-values$ of $intercept, TV, radio$ we observe that it is unlikely to observe such a substantial association between each of these variables and *sales* due to chance. Then we can infer that it exists a association between each of these variables and *sales*, unlike *newspaper*.

1.2 2.

K-Nearest neighbors classifier : attempts to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability.

$$\mathbb{P}_{\{X=x_0\}}(\{Y=j\}) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = j)$$

K-Nearest neighbors regression : given a value for K and a prediction point x_0 , KNN regression first identifies the K training observations that are closest to x_0 , represented by $\mathcal{N}_0$. It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x_i\in\mathcal{N}_0} y_i$$

1.3 3.

Fait sur papier

1.4 4.

For a linear regression without intercept we have the ith fitted value in the form :
$\hat{y}_i = x_i\hat{\beta}$ with $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$
Then we get $\hat{y_i} = \sum_{i=1}^{n} \dfrac{x_i}{\sum_{i=1}^{n} x_i^2} y_i = \sum_{i=1}^{n} a_i y_i$ "We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values."

1.5 6.

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ but $\beta_0 = \overline{y} - \beta_1\overline{x}$ then $\hat{y}_i = \overline{y} - \beta_1\overline{x} + \beta_1 x_i$
For $x_i = \overline{x}$ we get $\hat{y}_i = \overline{y}$, so the point $(\overline{x}, \overline{y})$ belongs to the least square line.

1.6 7.

2 Applied

2.1 8.

```
  library(MASS)
library(ISLR)
autoDF=Auto
lm.fit = lm(mpg~horsepower, data=autoDF) #simple linear regression horsepower onto mpg
summary(lm.fit) #Displaying global information

##
## Call:
## lm(formula = mpg ~ horsepower, data = autoDF)
##
## 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 Regarding the p-value, it seems there is a relationship between predictor and the response

ii This relationship is weak because $\widehat{horsepower} = -0.157845$

iii The relationship is negative because $-0.157845 < 0$

iv The predicted mpg associated with a horsepower of 98 is

```
  coef(lm.fit)[1] + coef(lm.fit)[2]*98

## (Intercept)
##    24.46708
```

And confidence and prediction interval, for $horsepower = 98$ are respectively :

```
  predict(lm.fit, data.frame(horsepower=98), interval="confidence")
```

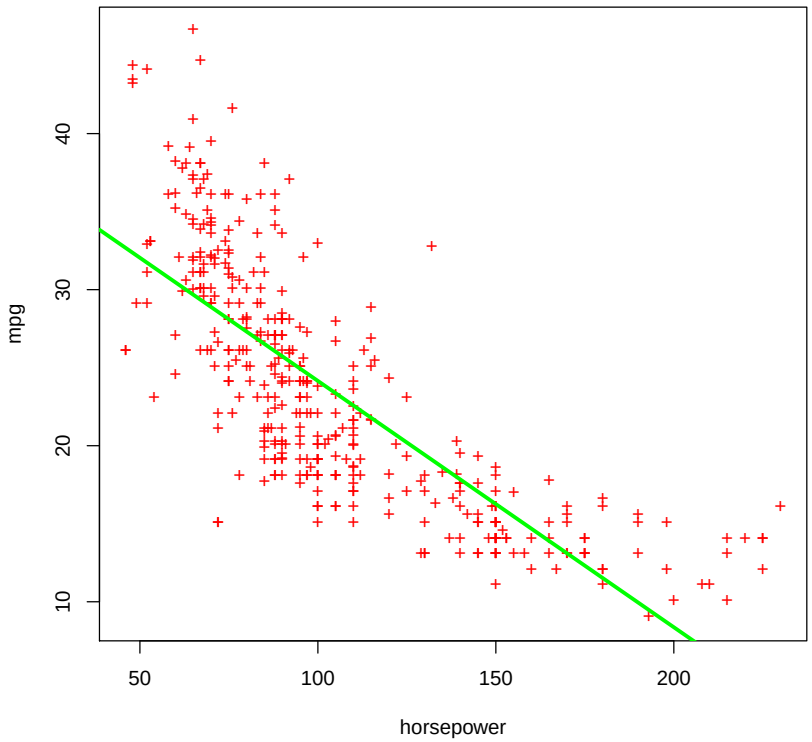
```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```
predict(lm.fit, data.frame(horsepower=98), interval="prediction")
```

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

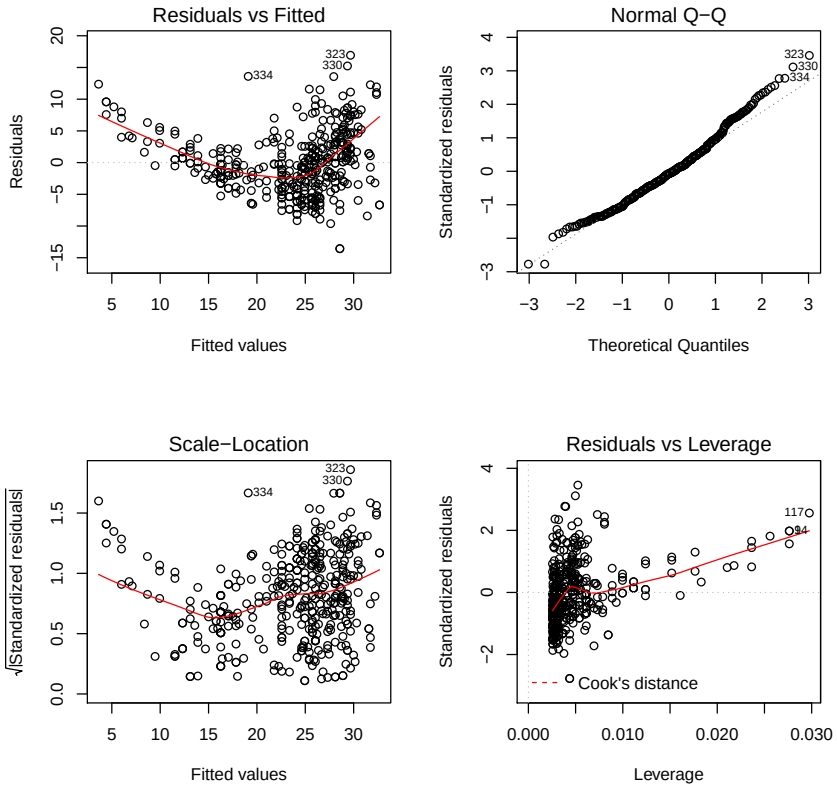
(b) Plot of the response and the predictor :

```
  attach(Auto)
plot(horsepower, mpg, col='red', pch='+')
abline(lm.fit, lwd=3, col='green')
```



(c) Plot of the response and the predictor :

```
  par(mfrow=c(2,2))
plot(lm.fit)
```

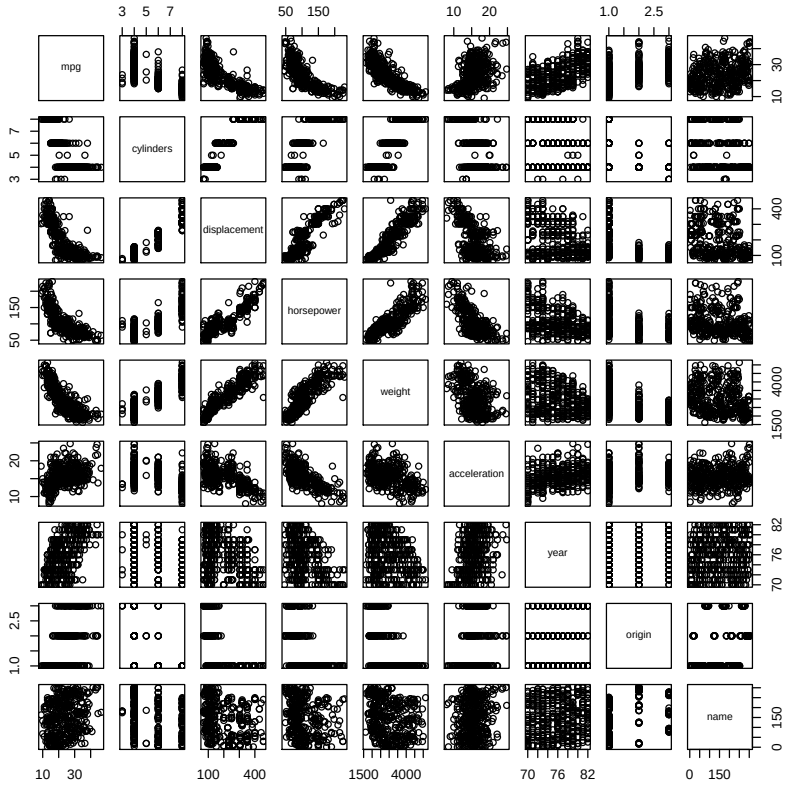


Regarding plots above, the first "Residuals vs Fitted" plots shows us that the linear model is not a very suitable model for these data.
And the "Normal Q-Q" shows us that residuals or not normally spread.
Then the "Scale-Location" shows us that residuals are not equally spread along the range of predictors, so the equal variance (homoscedasticity) is not adapted to these data.
Finally the "Residuals vs Leverage" shows us that it is not seems to exist very big outliers.

2.2 9

(a) Scatterplot matrix which including all of the variables in the data set

```r
pairs(Auto)
```



(b) The matrix of correlations between the variables, and we have excluded then "name" variable :

```r
  Table = data.frame(Auto)
cor(Table[, -9])
```

```
##                     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) We perform a multiple linear regression with mpg as the response and all other variables except name as the predictors.

```
  lm.fit = lm(mpg~.-name, data=Auto)
summary(lm.fit)

##
## 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
```

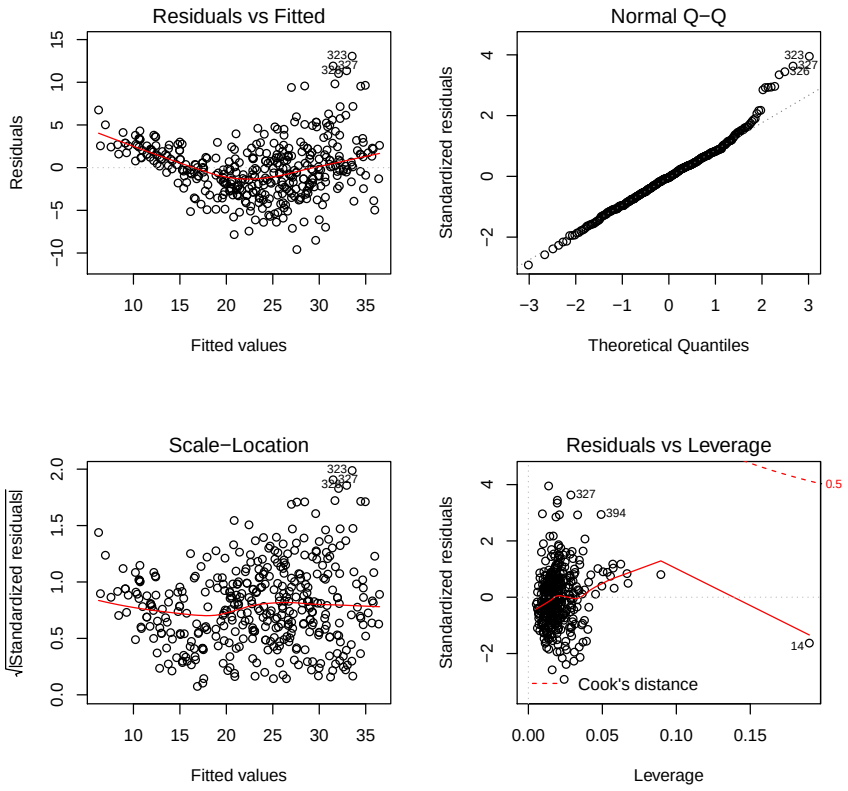
i. The predictors displacement, horsepower, and weight do not seem to have a relationship with the response.

ii. It seems that there is a relationship is statically significant between the response and the following predictors : origin, year, weight and in lesser measure displacement.

iii. The coefficient for the year variable suggests that an increasing of 10 years, bring on average 75 miles per gallon to a car.

(d) We produce diagnostic plots of the linear regression fit :

```
  lm.fit = lm(mpg~.-name, data=Auto)
par(mfrow=c(2,2))
plot(lm.fit)
```



The “Residuals vs Fitted” plot shows us that relationship is not really linear.
And the “Normal Q-Q” plot shows that the very most part of residuals are normally spread.
Then the “Scale-Location” shows us that the homoscedasticity assumption is enough respected.
Finally the “Residuals vs Leverage” shows us that there is not significant of outliers.