6. 23. Nov. - Regression

Note: so far, five exercises have been propose as a topic for a term project. Select one of those. If you have a draft for comment, please send it tome. If find the time, I will try to give you my comments. or recommendations.

Each of the following examples shows two basic step in data analysis using linear models. Inspect the output, and report your results. If you have an idea how to use these results, for instance for next steps to take, please add them.

Exercise 17. *This is using a modified version of the Scottisch Hill Runner's data. Simple inspection. For graphical output, see figure 11.*

Output

```
     dist            climb          time
 Min.   : 2.000   Min.   : 300   Min.   : 15.95
 1st Qu.: 4.500   1st Qu.: 725   1st Qu.: 27.42
 Median : 6.000   Median :1000   Median : 36.37
 Mean   : 7.529   Mean   :1815   Mean   : 56.16
 3rd Qu.: 8.000   3rd Qu.:2200   3rd Qu.: 63.63
 Max.   :28.000   Max.   :7500   Max.   :204.62
```

Naive linear model. For graphical output, see figure 12.

Output

```
Call:
lm(formula = time ~ dist + climb, data = hills)

Residuals:
    Min      1Q  Median      3Q     Max
-16.632  -4.934   1.007   4.541  27.903

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.94198    2.58005  -5.016 1.90e-05 ***
dist          6.34556    0.36047  17.604  < 2e-16 ***
climb         0.01175    0.00123   9.555 6.83e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.8 on 32 degrees of freedom
Multiple R-squared:  0.9712,        Adjusted R-squared:  0.9694
F-statistic: 540.2 on 2 and 32 DF,  p-value: < 2.2e-16
```

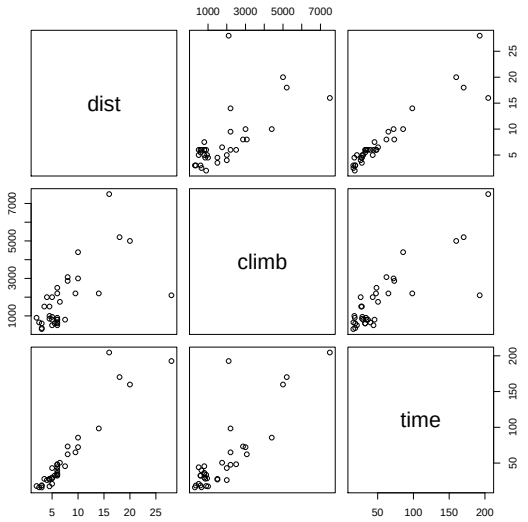

FIGURE 11. Raw data inspection

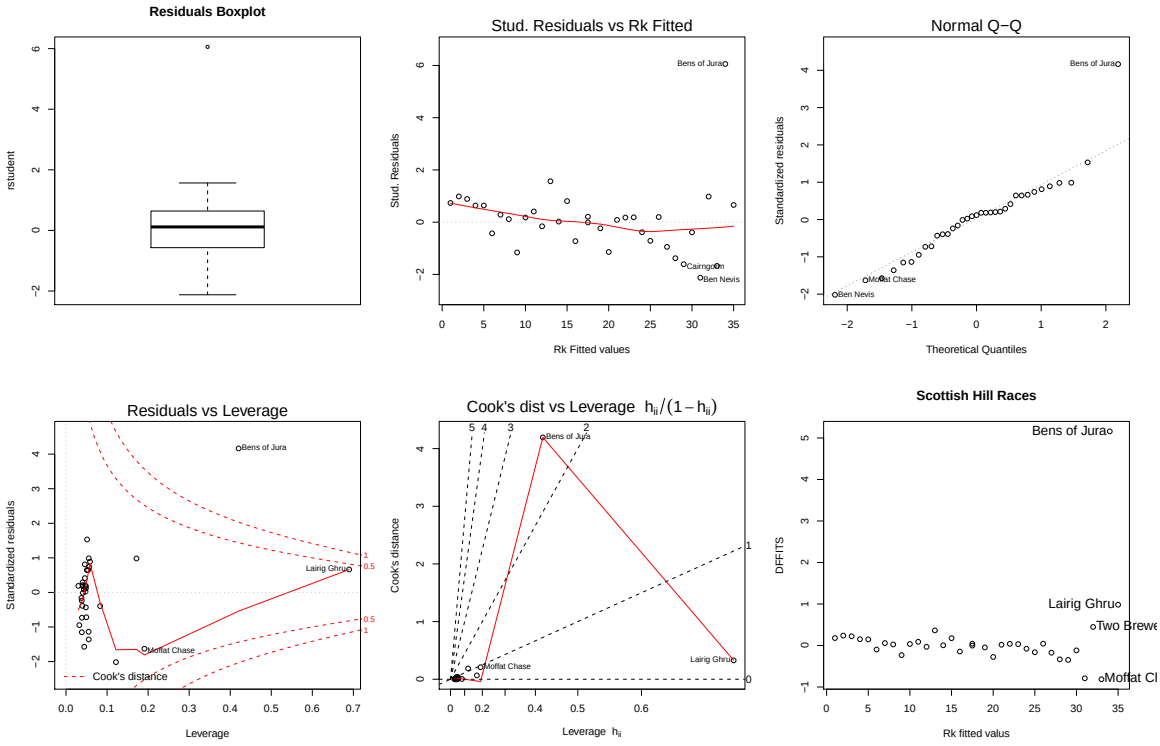

FIGURE 12. Diagnostics from a naive lm()

Exercise 18. *This is using a modified version of the Scottisch Hill Runner's data, and used a different approach to modelling. Fit try to fit the speed, and allow the climb to enter as a quadratic term.*

Input

```
hills$speed <- hills$time/hills$dist
hills$time <- NULL
hills$climb2 <- hills$climb^2
```

Simple inspection. For graphical output, see figure 13.

Output

```
      dist            climb          speed            climb2
 Min.   : 2.000   Min.   : 300   Min.   : 3.870   Min.   :   90000
 1st Qu.: 4.500   1st Qu.: 725   1st Qu.: 5.968   1st Qu.:  531250
 Median : 6.000   Median :1000   Median : 6.625   Median : 1000000
 Mean   : 7.529   Mean   :1815   Mean   : 6.986   Mean   : 5842110
 3rd Qu.: 8.000   3rd Qu.:2200   3rd Qu.: 7.955   3rd Qu.: 4840000
 Max.   :28.000   Max.   :7500   Max.   :12.789   Max.   :56250000
```

Naive linear model. For graphical output, see figure 14.

Output

```
Call:
lm(formula = speed ~ dist + climb + climb2, data = hills)

Residuals:
     Min       1Q   Median       3Q      Max
-2.52653 -0.58174 -0.05889  0.58727  2.50198

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.803e+00  3.925e-01  14.786 1.37e-15 ***
dist        -6.263e-02  4.151e-02  -1.509   0.1415
climb        8.597e-04  3.554e-04   2.419   0.0216 *
climb2       1.608e-08  4.855e-08   0.331   0.7427
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 31 degrees of freedom
Multiple R-squared:  0.6702,        Adjusted R-squared:  0.6382
F-statistic:    21 on 3 and 31 DF,  p-value: 1.291e-07
```

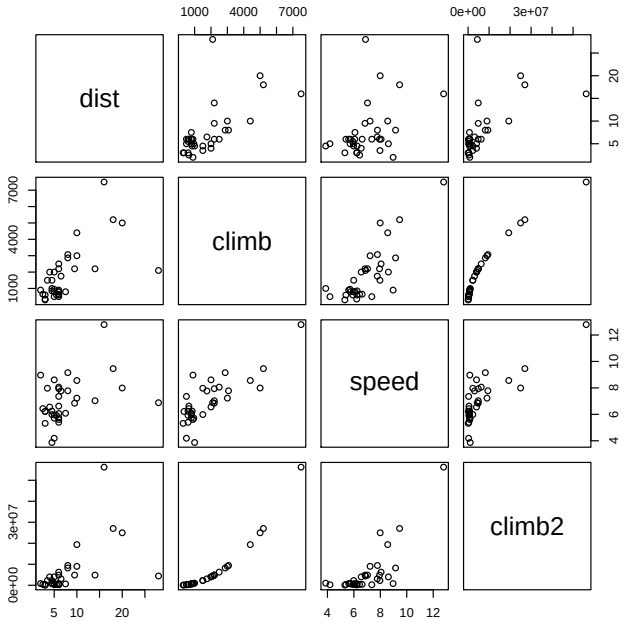

FIGURE 13. Raw data inspection. Scottish Hills data VS2.

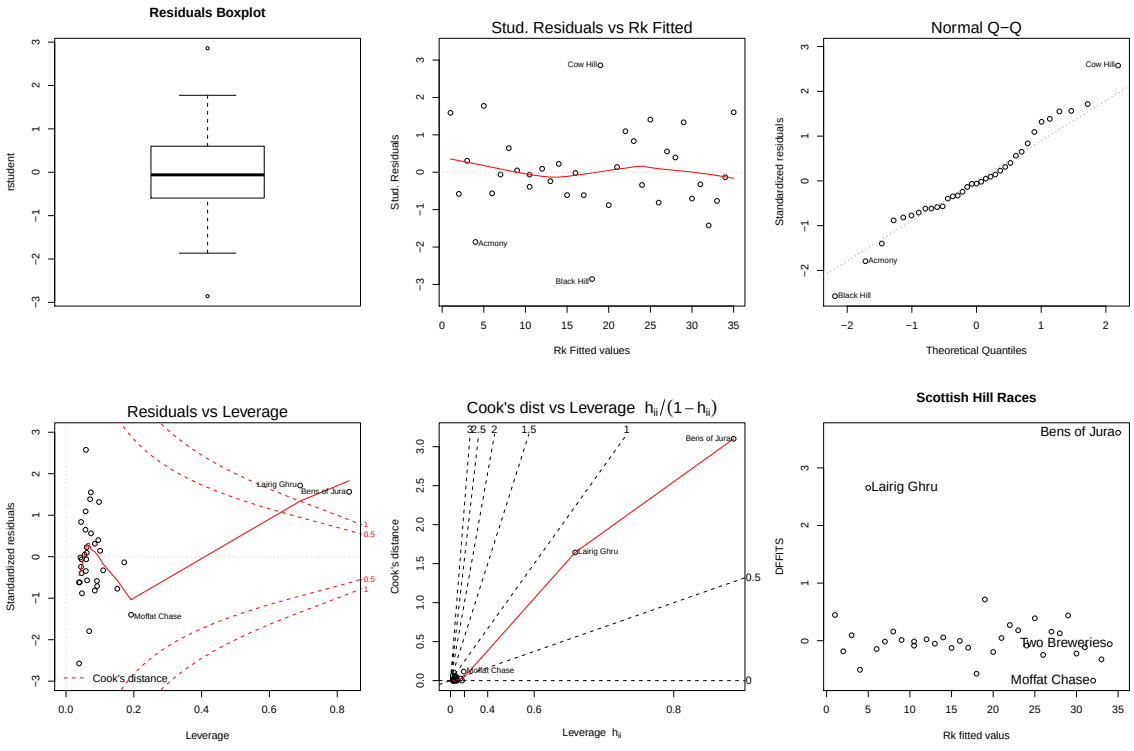

FIGURE 14. Diagnostics from a naive lm(). Scottish Hills data VS2.

Exercise 19. *This is using the Cars93 data from MASS: data from 93 Cars on Sale in the USA in 1993. The total data set has has 93 rows and 27 columns. We select the milage in city as a response for modelling.*
Simple inspection. For graphical output, see figure 15.

Output

```
   MPG.city       EngineSize     Cylinders    Horsepower
 Min.   :15.00   Min.   :1.000   3     : 3   Min.   : 55.0
 1st Qu.:18.00   1st Qu.:1.800   4     :49   1st Qu.:103.0
 Median :21.00   Median :2.400   5     : 2   Median :140.0
 Mean   :22.37   Mean   :2.668   6     :31   Mean   :143.8
 3rd Qu.:25.00   3rd Qu.:3.300   8     : 7   3rd Qu.:170.0
 Max.   :46.00   Max.   :5.700   rotary: 1   Max.   :300.0
      RPM           Weight
 Min.   :3800   Min.   :1695
 1st Qu.:4800   1st Qu.:2620
 Median :5200   Median :3040
 Mean   :5281   Mean   :3073
 3rd Qu.:5750   3rd Qu.:3525
 Max.   :6500   Max.   :4105
```

Naive linear model. For graphical output, see figure 16.

Output

```
Call:
lm(formula = MPG.city ~ EngineSize + Cylinders + Horsepower +
    RPM + Weight, data = Cars93)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3110 -1.2118  0.0307  1.1456 14.4749

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.573e+01  6.428e+00   7.114 3.66e-10 ***
EngineSize      -9.829e-02  9.844e-01  -0.100 0.920706
Cylinders4      -8.795e+00  1.732e+00  -5.079 2.29e-06 ***
Cylinders5      -9.647e+00  2.813e+00  -3.429 0.000945 ***
Cylinders6      -8.926e+00  2.273e+00  -3.926 0.000177 ***
Cylinders8      -7.508e+00  2.986e+00  -2.514 0.013856 *
Cylindersrotary -1.423e+01  3.832e+00  -3.715 0.000368 ***
Horsepower      -1.731e-02  1.648e-02  -1.050 0.296701
RPM              9.343e-04  9.837e-04   0.950 0.345028
Weight          -5.536e-03  1.127e-03  -4.914 4.42e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.628 on 83 degrees of freedom
Multiple R-squared:  0.8026,        Adjusted R-squared:  0.7812
F-statistic: 37.51 on 9 and 83 DF,  p-value: < 2.2e-16
```

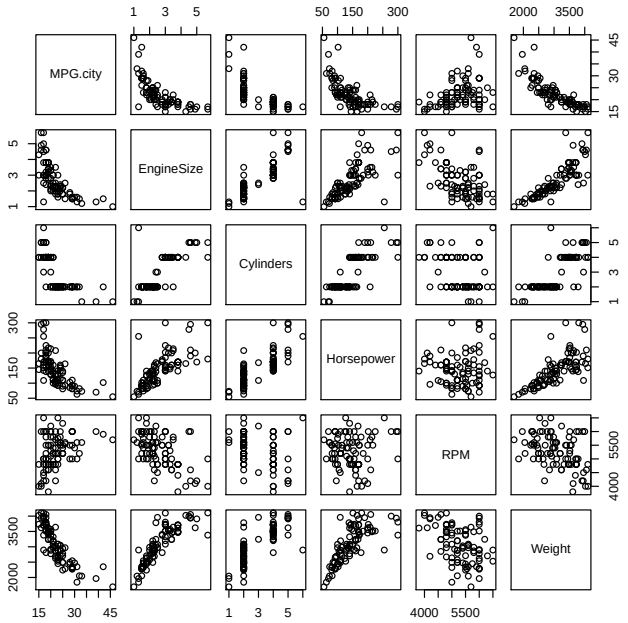

FIGURE 15. Raw data inspection. Cars93 data.

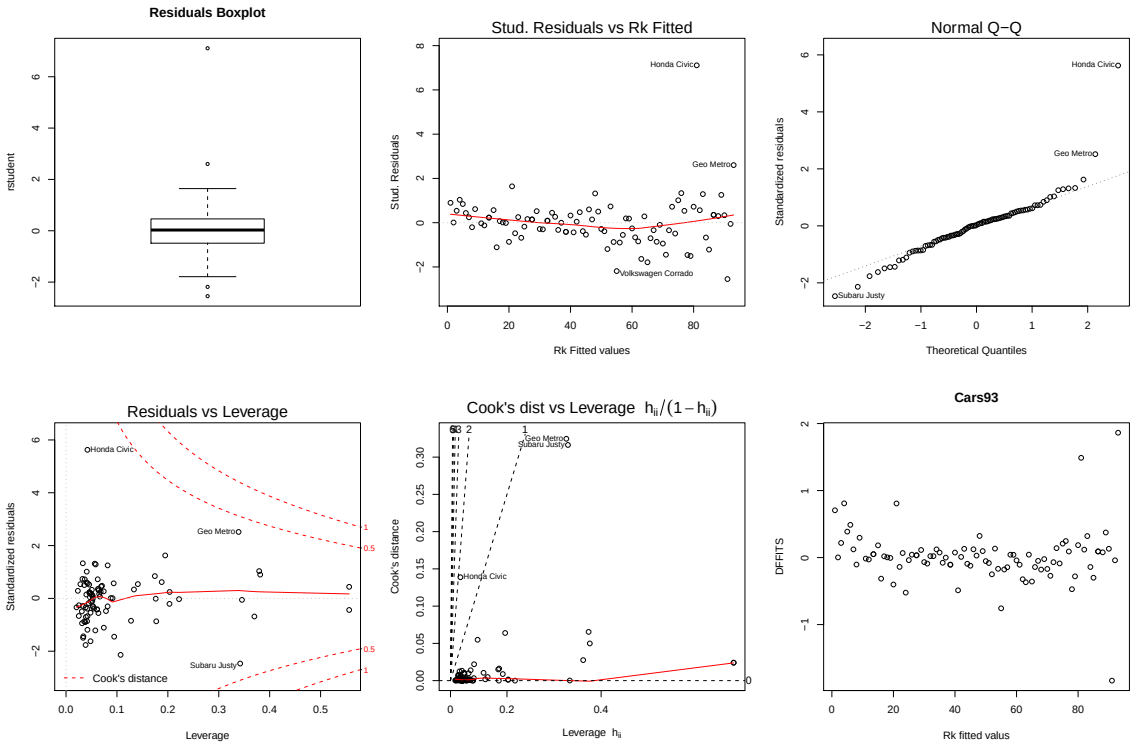

FIGURE 16. Diagnostics from a naive lm(). Cars93 data.

Exercise 20. **** R programming ****
Note: these plots use variants of the residuals. Is it feasible to unify the use of residuals, i. e. use one preferred variant? Why?
Try to give R code to implement this.