8. 6. Dec. - Robust methods

Dates are fixed:

Feb. 3. Project Exercises/Assignment due.

Please start working on your assignment early. If drafts/questions are sent in before winter holiday, I will try to give my comments during the break if I find the time (no guarantee).

For those who do not prepare a written assignment: **Feb. 8. written examination**. This is just a pass/fail examination - no grades are given in this examination.

Exercise 21. *Check the software available to you and look for robust estimators of location. Trimmed mean? Winsorised mean? Median? Other?*
Apply these methods to these test cases:

$$X = (1, 2, 3, 4, 5, 6, 7, 8, 9)$$
$$X = (1, 2, 3, 4, 5, 6, 7, 8, 10)$$
$$X = (1, 2, 3, 4, 5, 6, 7, 8, 20)$$
$$X = (1, 2, 3, 4, 5, 16, 17, 18, 19)$$

Exercise 22. *(needs some software lookup) Find an implementation of Rousseeuw's maximum depth regression, and apply it to the (raw) Scottish Hill Runners' data.*
Literature
PJ Rousseeuw, M Hubert: Regression depth. Journal of the American Statistical Association 446 (1999) 388–402

Note: we are now about to leave the 20^{th} century. The next methods discussed are already from the 21^{st} century, so nearly recent.

Exercise 23. *Robust regression is more complex than robust location estimation.*
Implicitly, it needs robust estimation of scale (in replacement of residual some of squares) and there are various candidates to chose from.
The regression needs its own robustification (in replacement of Gauss-Markov, or least squares). In general, there is no closed solution, but an iterative optimisation is used. This has again several candidates to chose from.
Function `rlm()` *from* `library(MASS)` *is applied to the (raw) Scottish Hill Runners' data. No correction has been applied for Knock Hill.*
Using variants for scale estimation and for the optimisation details, we get slightly different results. Compare these results, and give your comments.
(To make layout easier, code and output are below.)

Input

```
rlm.hillshuberMADM <- rlm(time ~ dist + climb, data=hills, psi=psi.huber, scale="MAD", method="M")
summary(rlm.hillshuberMADM)
```

Output

```
Call: rlm(formula = time ~ dist + climb, data = hills, psi = psi.huber,
    scale = "MAD", method = "M")
Residuals:
      Min        1Q    Median        3Q       Max
-10.75039  -3.28395  -0.03358   3.53791  65.70100

Coefficients:
            Value   Std. Error t value
(Intercept) -9.6067  1.7545    -5.4754
dist         6.5507  0.2451    26.7237
climb        0.0083  0.0008     9.9199

Residual standard error: 5.209 on 32 degrees of freedom
```

See Figure figure 16.

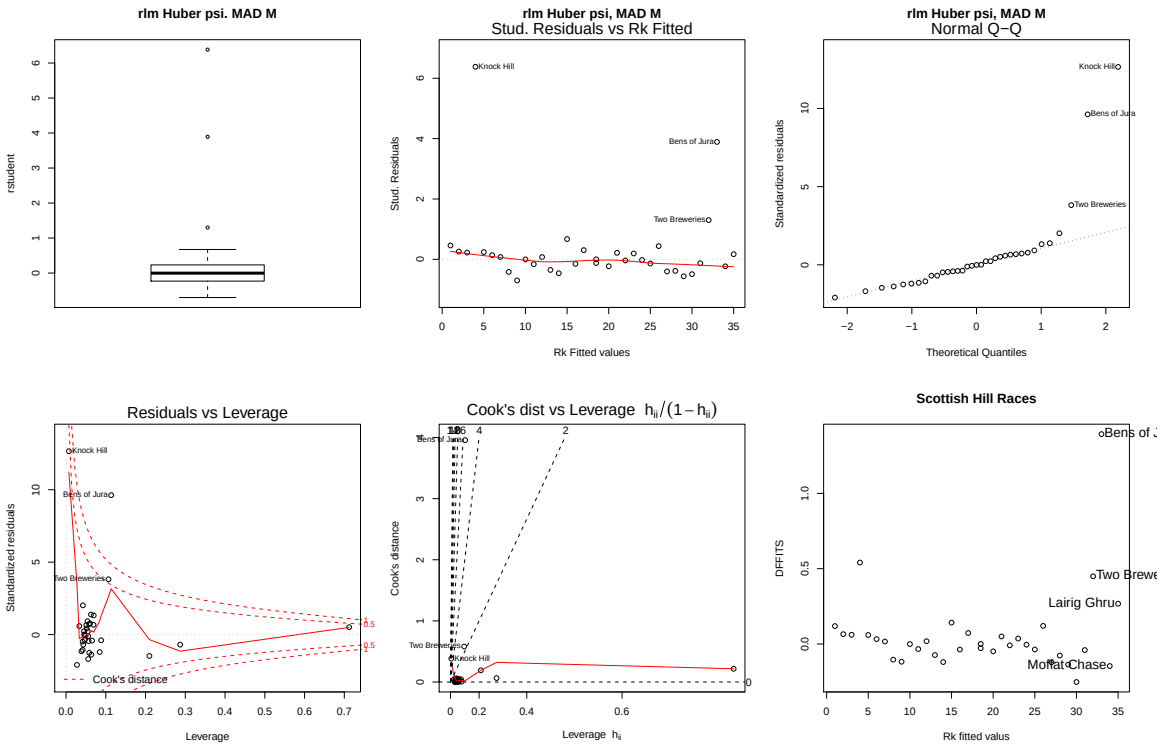

FIGURE 16. Diagnostics from rlm(). MAD, Huber's psi. Hills data.

Input

```
rlm.hillshampelMADM <- rlm(time ~ dist + climb, data=hills,
        psi=psi.hampel, scale.est="MAD", method="M")
summary(rlm.hillshampelMADM)
```

Output

```
Call: rlm(formula = time ~ dist + climb, data = hills, psi = psi.hampel,
    scale.est = "MAD", method = "M")
Residuals:
     Min       1Q   Median       3Q      Max
-10.5493  -3.0876  -0.1103   3.8871  65.5395

Coefficients:
            Value   Std. Error t value
(Intercept) -9.5099  1.6259    -5.8489
dist         6.6693  0.2272    29.3592
climb        0.0075  0.0008     9.6317

Residual standard error: 5.496 on 32 degrees of freedom
```

Input

```
rlm.hillsbiMADM <- rlm(time ~ dist + climb, data=hills,
        psi=psi.bisquare, scale.estim="MAD", method="M")
summary(rlm.hillsbiMADM)
```

Output

```
Call: rlm(formula = time ~ dist + climb, data = hills, psi = psi.bisquare,
    scale.estim = "MAD", method = "M")
Residuals:
     Min       1Q   Median       3Q      Max
-10.8333  -2.6738   0.3937   3.6004  64.5814

Coefficients:
            Value   Std. Error t value
(Intercept) -8.1207  1.4883    -5.4563
dist         6.6382  0.2079    31.9241
climb        0.0065  0.0007     9.1614

Residual standard error: 4.781 on 32 degrees of freedom
```

Input

```
rlm.hillsbiHubM <- rlm(time ~ dist + climb, data=hills,
        psi=psi.bisquare, scale.est="Huber")
summary(rlm.hillsbiHubM)
```

Output

```
Call: rlm(formula = time ~ dist + climb, data = hills, psi = psi.bisquare,
    scale.est = "Huber")
Residuals:
    Min      1Q  Median      3Q     Max
-10.825  -2.690   0.364   3.618  64.606

Coefficients:
            Value   Std. Error t value
(Intercept) -8.1516  1.4845    -5.4910
dist         6.6374  0.2074    32.0013
climb        0.0065  0.0007     9.2218
```

```
Residual standard error: 5.453 on 32 degrees of freedom
```

Input

```
rlm.hillsMM <- rlm(time ~ dist + climb, data=hills,
        method="MM")
summary(rlm.hillsMM)
```

Output

```
Call: rlm(formula = time ~ dist + climb, data = hills, method = "MM")
Residuals:
     Min       1Q   Median       3Q      Max
-10.8328  -2.6754   0.3909   3.6018  64.5833

Coefficients:
            Value   Std. Error t value
(Intercept) -8.1231  1.4864    -5.4649
dist         6.6381  0.2077    31.9643
climb        0.0065  0.0007     9.1764

Residual standard error: 4.844 on 32 degrees of freedom
```

This is just using default settings. Included only for information – ignore for now.

Input

```
rlm.hillsDflt <- rlm(time ~ dist + climb,data=hills)
summary(rlm.hillsDflt)
```

Output

```
Call: rlm(formula = time ~ dist + climb, data = hills)
Residuals:
      Min        1Q    Median        3Q       Max
-10.75039  -3.28395  -0.03358   3.53791  65.70100

Coefficients:
            Value   Std. Error t value
(Intercept) -9.6067  1.7545    -5.4754
dist         6.5507  0.2451    26.7237
climb        0.0083  0.0008     9.9199

Residual standard error: 5.209 on 32 degrees of freedom
```

See Figure figure 17.

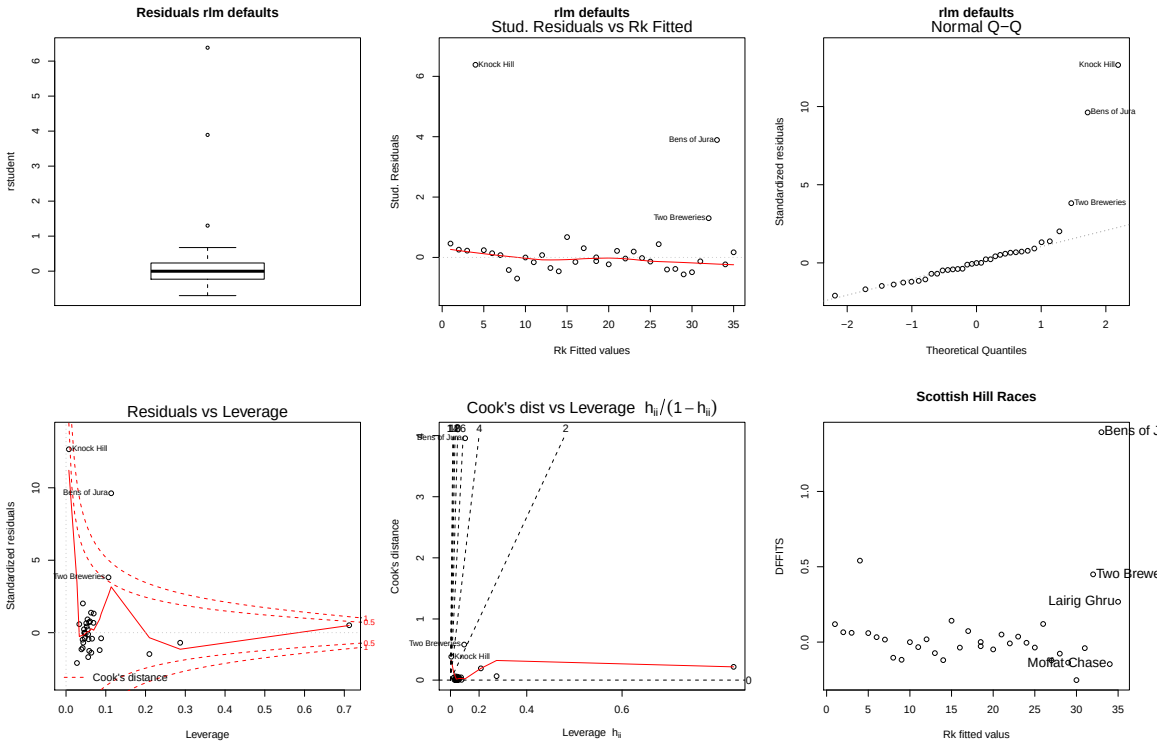

FIGURE 17. Diagnostics from rlm(). Hills data. This is just using default settings. Included only for information – ignore for now.

Exercise 24. *The diagnostics developed for least squares estimation in a Gaussian linear model need to be modified if we use robust methods.*
Which diagnostics are still meaningful? Which can be adjusted?
Can you write a list of the preferred diagnostics which you would like to see?

Exercise 25. *Compare the robust regression from 23 with the least squares regression 17.*