Modern Statistics and Big Data Analysis - how an exam may look like

At the moment of posting this, not all material that this mock exam refers to is already covered!

You are not asked to submit answers to these questions, and there will be no feedback (unless I use one of these questions as an Exercise).

The exam will consist of one data analysis question ("analyse a given dataset" without precise indication what methods to use), one theoretical question, one question about interpreting computer output, and one literature question. The literature question 1 will be made available to you in good time before the exam, and has to be submitted *before* the exam starts. It will not be part of the exam on the day.

You will have access to the internet in the lab and you can use all materials. However, any use of websites or services that allow an exchange of messages with your fellow students is strictly forbidden. There is zero tolerance. If I see any of you having open such a page or app (even if you don't actually use it) the exam is failed with a mark of zero. Any actual exchange of messages is also strictly forbidden. I cannot rule out that an innocently and legitimately looking website has some hidden function that allows exchange of messages such as allowing for user comments that are immediately published; I may tolerate the use of such sites given that chances are in some cases neither you nor I can immediately check whether such a facility exists, but surely its use for exchange of messages will not be tolerated.

Obviously I cannot control whether you collaborate before the exam on question 1. However you are required to use your own words and solutions and wordings that give strong evidence that they were prepared in collaboration with fellow students will also lead to a failure of the exam with zero marks (I may deviate from this if the evidence is too weak). In other words, you can discuss the article in advance, however it is forbidden to collaborate with fellow students when it comes to the actual writing.

1. **Literature question, to do at home**
   On Moodle you can find the article "Variable Selection for Model-Based Clustering" by A. E. Raftery and N. Dean, Journal of the American Statistical Association 101, 168-178 (2006). Keep in mind that it is quite normal to not understand everything in a reseach paper. There are details in this chapter that are hard to understand based on just what you have learnt in courses, but these should not be important to understand for answering the questions at the required level. The ability to get from the paper what you need without being affected by the things you do not understand is a key competence.

   (a) Explain the approach for variable selection in model-based clustering as proposed in the article in your own words.

   (b) How is according to the authors a variable defined that is "irrelevant" for the clustering? State this in your own words; you are not asked to repeat the formalism in the paper.

   (c) On p.170, beginning of Sec. 2.4, the authors make a connection to linear regression. The formula given there follows, according to the authors, "from the standard result for conditional multivariate normal means". This standard result (which many of you may have learnt at some point) can be found on the Wikipedia page on the "Multivariate normal distribution" in the section on "Conditional distributions" (and also elsewhere; you can use another source if you want).

   Consider the linear regression model in Sec. 4.1 of the course slides on robust statistics, and explain how the Y, x, p in the regression are related to the variables in the notation of the article for the mentioned part of p.170. Also explain, using the "standard result", how the $\beta_0, \ldots, \beta_p$-parameters in the regression could be found (assuming that the theoretically true model would be known) from the within-component means $\boldsymbol{\mu}_g$ and covariance matrices $\boldsymbol{\Sigma}_g$.
   Hint: Although the $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are different for different g, it can be shown that the β-parameters will be the same regardless of which g is used; you can just use a single fixed g for this and you do not need to worry about the others. You are *not* asked to prove or argue that the regression β-parameters do not depend on g, although you may try to understand why this holds anyway, even though there are no marks assigned to this.

   (d) The paper does not mention the term cross-validation. Cross-validation is a major approach for measuring quality for regression variable selection. What do you think is the reason for not using it here? Imagine a way how splitting the dataset in two parts (say) could be useful for variable selection in clustering and explain it.

   (e) The method introduced in the paper is (with a small improvement) implemented in the R-package `clustvarsel`. Apply it to the Boston housing data set from the Virtuale page of the course (you may have to remove the "chas" variable). You can use default settings. Also apply `mclust` with default settings and without variable selection to the dataset. Compare the two clusterings and comment on whether you think the use of `clustvarsel` is advantageous here.

2. Analyse the "tombdataX00.dat" dataset.

   Background: On the Moodle page of the course, a dataset named "tombdataX00.dat" is provided. This dataset was collected by the archaeologist Flinders Petrie, and reconstructed from Petrie's handwritten notes by Alice Stevenson of UCL's Petrie museum for Egyptian archaeology. The dataset is of genuine research interest for the Petrie museum; be warned that the dataset is not a particularly nice textbook example with a single clear and easily identifiable good clustering.

   The data contains information on 77 tombs on ancient Egypt. The rows of the dataset are the tombs. The columns of the dataset refer to 145 different types of pottery artifacts. The first row of the dataset gives the codes for the artifact types (this is a letter followed by a number such as B22, C31 etc.). On the Moodle page for the course, there is also a rough "Guide to Petrie's Sequence Dating Slips" written by Alice Stevenson, which explains the artifact types and shows examples (you don't need to know this to solve the ICA; just in case you're interested; note that the first letter in the code refers to a wider class of pottery and the number to a more specific type; *I would probably not give distracting information like this in a real exam unless it's really required*).

   The first column of the dataset gives systematic names of the tombs (also combinations of a letter and a number such as n1828). For each of the 77 tombs, the dataset states whether a certain type of artifact is present in the tomb (1) or not (0).

   Such data are used for the purpose of dating the tombs. The idea is that tombs with similar artifacts are likely to come from the same period, so the purpose of clustering here is to find clusters of tombs that are likely to have been created in about the same period of time. Clustering could be of interest at several levels of coarseness, grouping tombs together that could be very close in time, or somewhat distant but still related.

   What is expected: Produce two distance-based clusterings of the tombs (obviously you can try out more, but you will only get credit for two). Explain why you chose the specific clustering methods and motivate all the methodological decisions you made, including your choice of a distance.

   Produce a visualisation of each clustering.

   Compare the clusterings and discuss to what extent each of them may be helpful for tomb dating so that the discussion can be understood by an archaeologist.

   Also submit the R-code that you use.

   Hint: Read the data with
   `tombdata <- read.table("tombdataX00.dat",header=TRUE)`. The tomb names are in `row.names(tombdata)`, the artifact codes are in `names(tombdata)`.

3. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n \in I\!R^p$ be distributed according to the joint density

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n) = \prod_{i=1}^{n} \varphi(\mathbf{x}_i; \mathbf{a}_{\gamma(i)}, \mathbf{\Sigma}), \tag{1}$$

$\varphi(\bullet; \mathbf{a}, \mathbf{\Sigma})$ being the pdf of the p-variate Gaussian distribution with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{\Sigma}$ (i.e., the covariance matrices for all clusters are assumed equal). $\gamma : I\!N_n \mapsto I\!N_k$ (or equivalently $\gamma = (\gamma(1), \ldots, \gamma(n)) \in I\!N_k^n$) denotes the true clusters for the n observations, and all parameters are unknown. For given $c = (c(1), \ldots, c(n)) \in I\!N_K^n$, $k \in I\!N_K$, $n_k(c) = \sum_{i=1}^{n} 1(c(i) = k)$, let

$$\mathbf{W}_k(c) = \sum_{c(i)=k} (\mathbf{x}_i - \mathbf{m}_k^{Km}(c))(\mathbf{x}_i - \mathbf{m}_k^{Km}(c))', \ \mathbf{W}(c) = \sum_{k=1}^{K} \mathbf{W}_k(c),$$

where $\mathbf{m}_k^{Km}(c) = \frac{1}{n_k(c)} \sum_{c(i)=k} \mathbf{x}_i$. (If not explicitly stated, notation is as in the course notes.)

(a) Show that

$$\hat{c} = (\hat{c}(1), \ldots, \hat{c}(n)) \text{ minimising } \det \mathbf{W}(c)$$

are ML-estimators for $\gamma(1), \ldots, \gamma(n)$ of model (1).

Hints: You may use that for given fixed γ the ML-estimators of (1) for $(\mathbf{a}_1, \ldots, \mathbf{a}_K, \mathbf{\Sigma})$ are $(\mathbf{m}_1^{Km}(\gamma), \ldots, \mathbf{m}_K^{Km}(\gamma), \mathbf{W}(\gamma)/n)$.

You may also use that for all $c, \mathbf{a}_1, \ldots, \mathbf{a}_K, \mathbf{\Sigma}$:

$$\sum_{k=1}^{K} \sum_{c(i)=k} (\mathbf{x}_i - \mathbf{a}_k)' \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{a}_k) =$$

$$\text{trace}\left(\mathbf{W}(c)\mathbf{\Sigma}^{-1}\right) + \sum_{k=1}^{K} n_k (\mathbf{m}_k^{Km}(c) - \mathbf{a}_k)' \mathbf{\Sigma}^{-1} (\mathbf{m}_k^{Km}(c) - \mathbf{a}_k).$$

(b) Along the lines of the K-Means algorithm in Section 2.2 of the course notes, propose an algorithm to find a local optimum of the likelihood of model (1), or, equivalently, of $\det \mathbf{W}(c)$. Explain why the algorithm improves the likelihood in every step (if it does) unless there is no change.

This defines a clustering method based on model (1).

4. In the Appendix for this question there is some R-code for an output from an analysis of a regression data set. The data set was published by the United Nations Children's Fund UNICEF 1997. It contains information on 121 countries in 1995/96 on the following variables:

y	`Child.Mortality`	child mortality rate (deaths before the age of 5 per 1000 births)
X1	`Literacy.Fem`	literacy among females compared to males in percent (ie, 100 means that female literacy equals male literacy)
X2	`Literacy.Ad`	literacy among adults in percent
X3	`Drinking.Water`	percentage of population with access to safe drinking water
X4	`Polio.Vacc`	percentage of one-year olds vaccinated against polio
X5	`Tetanus.Vacc.Preg`	percentage of pregnant women vaccinated against tetanus
X6	`Urban.Pop`	percentage of population living in urban areas
X7	`Foreign.Aid`	received foreign aid as percentage of GDP

The analyses carried out in the appendix aim at explaining child mortality from the variables X1-X7, including statements about which of these variables seem to be important influence factors. It was also of interest to identify countries that behave differently from the general tendency.

(a) Which of the regression estimators do you find most trustworthy here and why (you can comment on known characteristics of the methods but you are also expected to use the data analysis for arguing your decision)?

(b) Which of the covariance matrix estimators do you find most trustworthy here and why (you can comment on known characteristics of the methods but you are also expected to use the data analysis for arguing your decision)?

(c) Which of the countries do you think are outliers in the sense that they seem to behave substantially different from the others (you can use the abbreviated country names as in the plots), based on which plots or results?

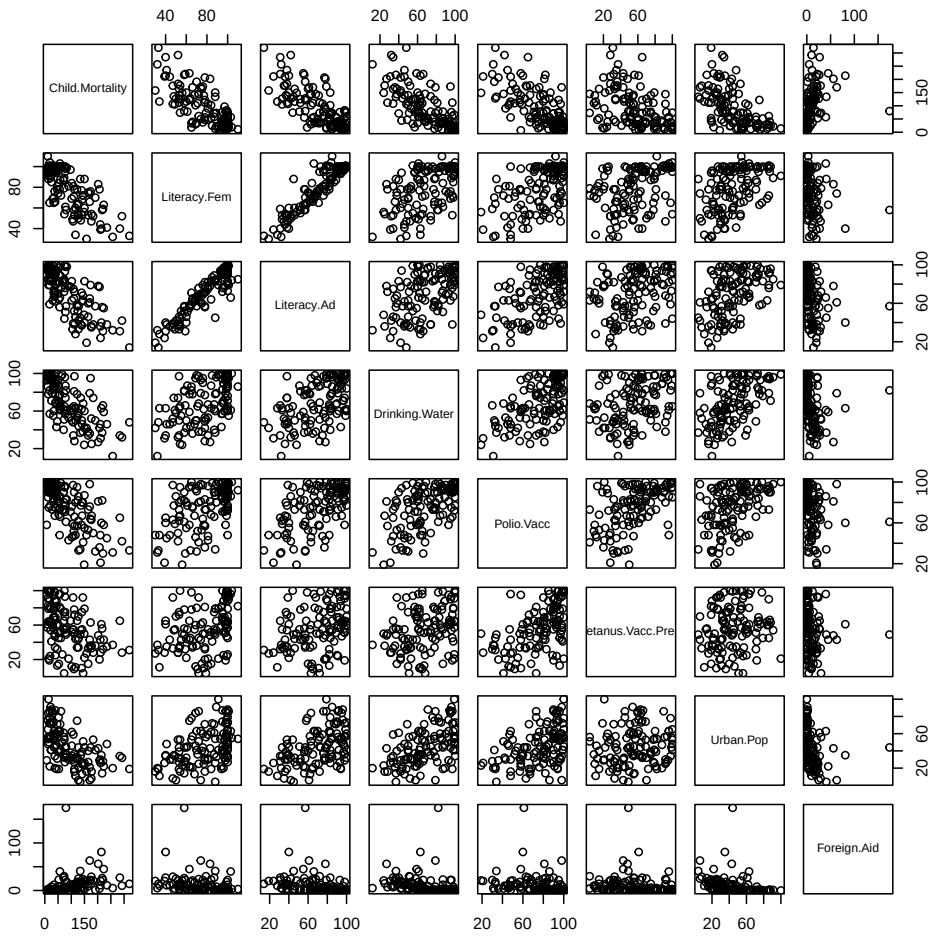
(d) For outlier identification, two different kinds of analyses were run here, namely (i) regression (least squares, MM, and S), and (ii) covariance matrix estimation based on all variables (standard, and MCD with $\alpha = 0.5$ and $\alpha = 0.75$). In what sense are outliers identified by regression different from outliers identified from covariance matrix estimation?

(e) A social scientist suggests that the observation "SaoTP" should be removed, because its level of foreign aid makes it essentially different from all other observations, and it should therefore not be used in the same analysis. What do you think of this suggestion? Which of the three regression estimators would in your opinion be most affected by such a decision?

Appendix for Question 4

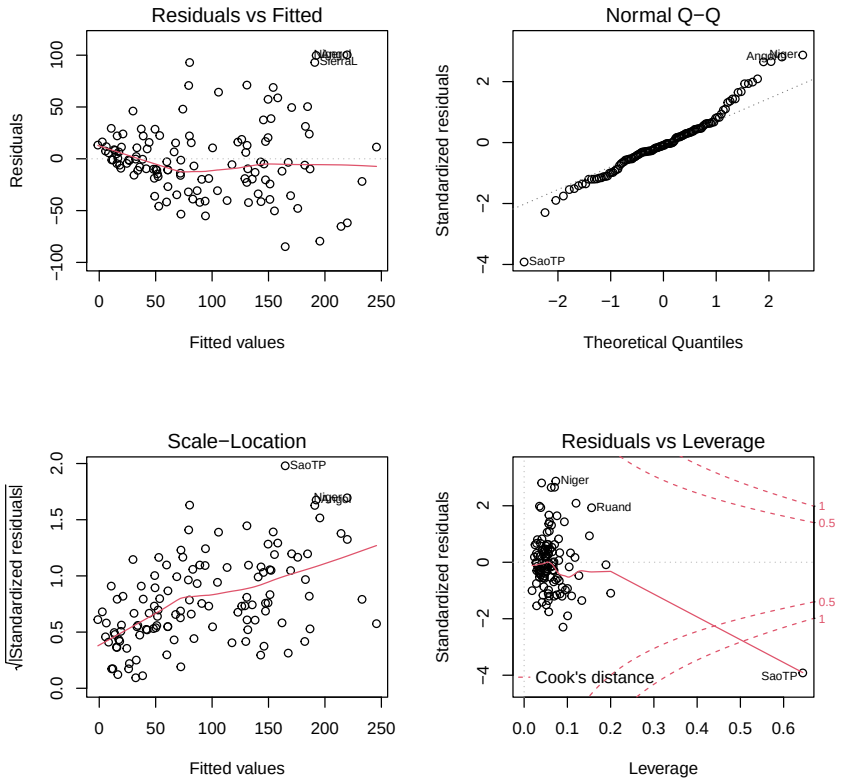
Overview of the dataset

```
library(robustbase)
unicef <- read.table("unicef97.dat",header=TRUE)
pairs(unicef,pch=rownames(unicef))
> str(unicef)
'data.frame': 121 obs. of  8 variables:
 $ Child.Mortality  : int  257 78 39 292 173 25 177 22 112 12 ...
 $ Literacy.Fem     : int  32 61 66 52 76 100 54 89 53 99 ...
 $ Literacy.Ad      : int  32 51 62 42 79 96 36 85 38 97 ...
 $ Drinking.Water   : int  12 87 78 32 95 71 25 96 97 100 ...
 $ Polio.Vacc       : int  31 77 75 42 64 90 67 98 66 85 ...
 $ Tetanus.Vacc.Preg: int  37 55 34 28 63 63 36 54 72 100 ...
 $ Urban.Pop        : int  20 45 57 32 44 88 16 91 19 48 ...
 $ Foreign.Aid      : int  5 4 1 10 22 0 15 1 4 0 ...
```



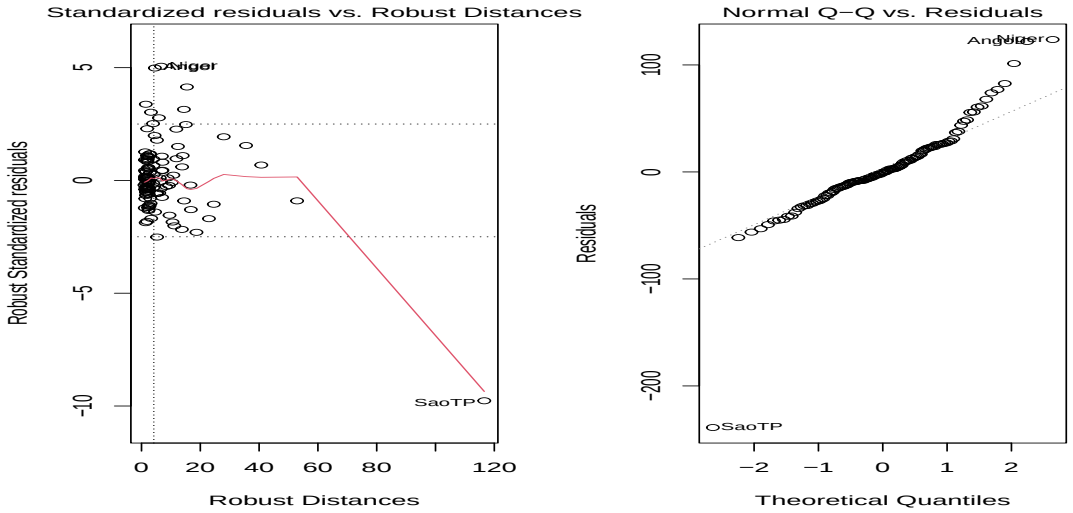
Least Squares regression

```
uniceflm <- lm(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+
               Polio.Vacc+Tetanus.Vacc.Preg+Urban.Pop+Foreign.Aid, data=unicef)
summary(uniceflm)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       333.4750    16.7638  19.893  < 2e-16 ***
Literacy.Fem       -1.1577     0.4432  -2.612  0.01021 *
Literacy.Ad        -0.2405     0.4167  -0.577  0.56497
Drinking.Water     -0.8695     0.2004  -4.339 3.13e-05 ***
Polio.Vacc         -0.7159     0.2362  -3.031  0.00302 **
Tetanus.Vacc.Preg  -0.0985     0.1593  -0.618  0.53750
Urban.Pop          -0.4112     0.1952  -2.107  0.03736 *
Foreign.Aid         0.2878     0.1759   1.636  0.10459
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.27 on 113 degrees of freedom
Multiple R-Squared: 0.7587,     Adjusted R-squared: 0.7437
F-statistic: 50.75 on 7 and 113 DF,  p-value: < 2.2e-16

# Diagnostic plots
plot(uniceflm,ask=FALSE)
```

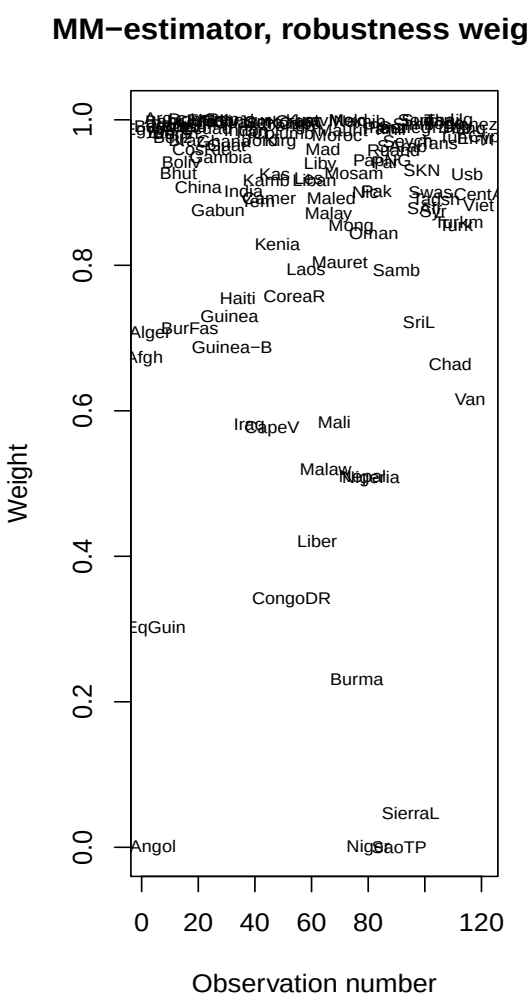


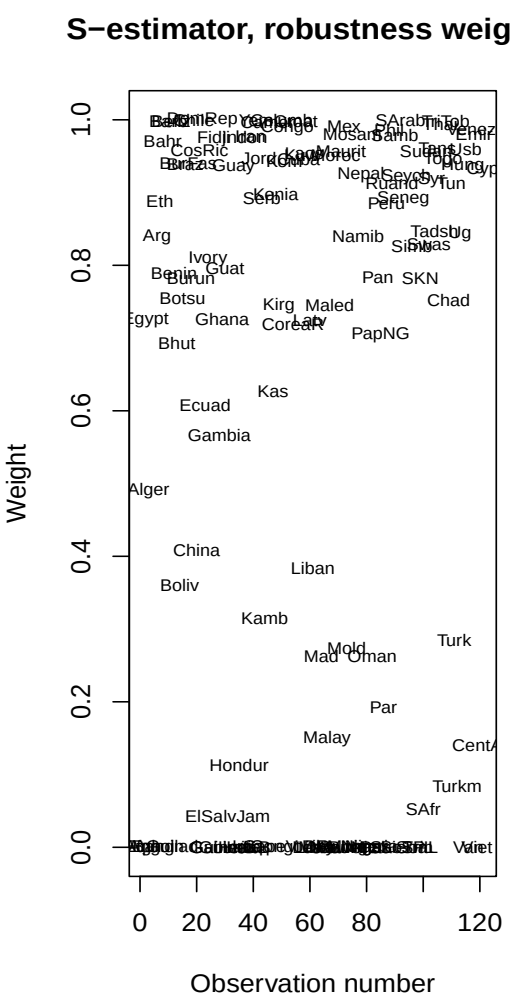
Robust regression

```
unicefmm <- lmrob(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+
                Polio.Vacc+Tetanus.Vacc.Preg+Urban.Pop+Foreign.Aid,
                data=unicef)
summary(unicefmm)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       277.88469   34.15661   8.136 5.91e-13 ***
Literacy.Fem       -1.14738    0.55415  -2.071 0.040683 *
Literacy.Ad         0.01122    0.43620   0.026 0.979529
Drinking.Water     -0.61264    0.19972  -3.067 0.002702 **
Polio.Vacc         -0.63284    0.36036  -1.756 0.081775 .
Tetanus.Vacc.Preg  -0.15987    0.13705  -1.166 0.245872
Urban.Pop          -0.32653    0.16752  -1.949 0.053752 .
Foreign.Aid         1.25256    0.31866   3.931 0.000146 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 24.46
Multiple R-squared:  0.8142,Adjusted R-squared:  0.8027
Convergence in 24 IRWLS iterations

Robustness weights:
 3 observations c(4,80,91) are outliers with |weight| = 0 ( < 0.00083);
 9 weights are ~= 1. The remaining 109 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04766 0.85490 0.94130 0.87120 0.98690 0.99900

# Diagnostic plots
plot(unicefmm,which=c(1,2),ask=FALSE)
```

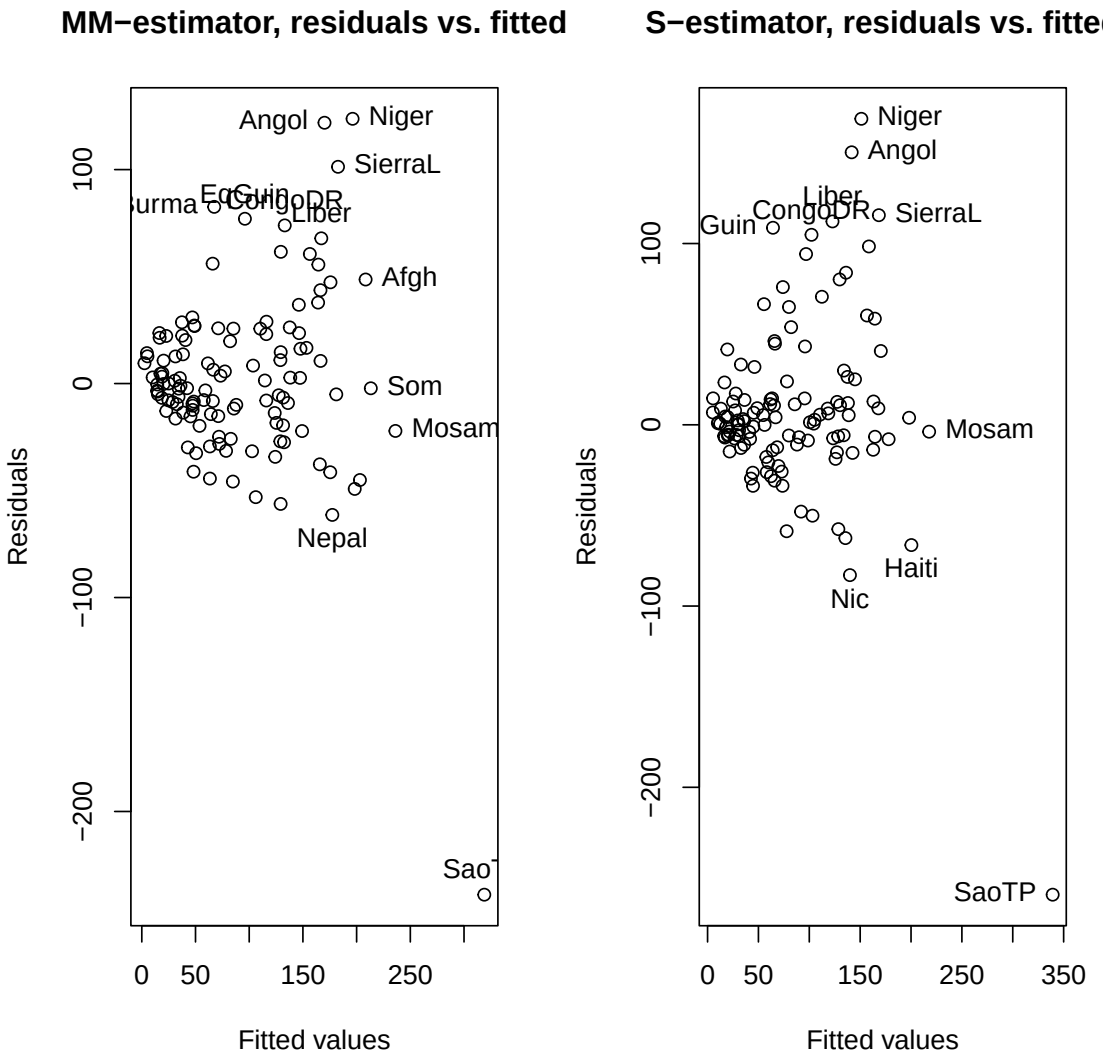


```
# S-estimator
unicefmm$init.S

Coefficients:
     (Intercept)       Literacy.Fem        Literacy.Ad     Drinking.Water
       188.65529            0.30652           -0.81904           -1.06404
      Polio.Vacc  Tetanus.Vacc.Preg          Urban.Pop        Foreign.Aid
         0.05852           -0.13186           -0.29665            1.62463
scale =  24.46 ;  converged  in  77  refinement steps

# Robustness weights
plot(1:121,unicefmm$rweights,xlab="Observation number",ylab="Weight",
  main="MM-estimator, robustness weights",type="n")
text(1:121,unicefmm$rweights,rownames(unicef),cex=0.7)
plot(1:121,unicefmm$init.S$rweights,xlab="Observation number",ylab="Weight",
  main="S-estimator, robustness weights",type="n")
text(1:121,unicefmm$init.S$rweights,rownames(unicef),cex=0.7)
```



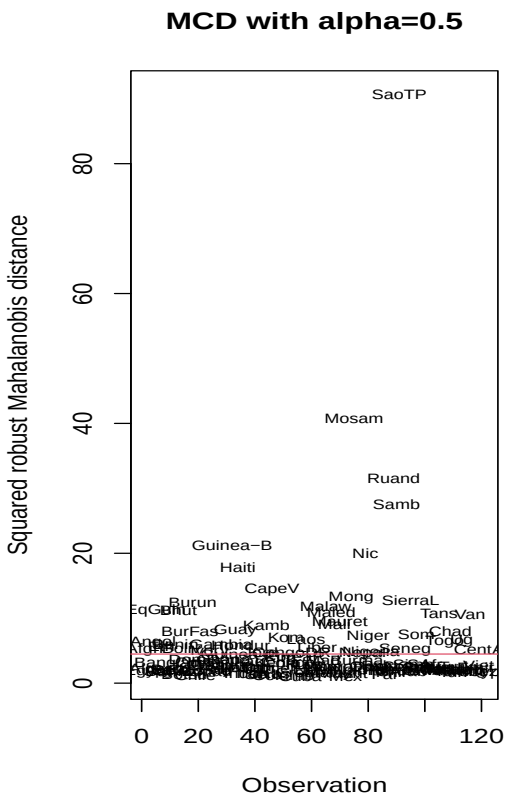


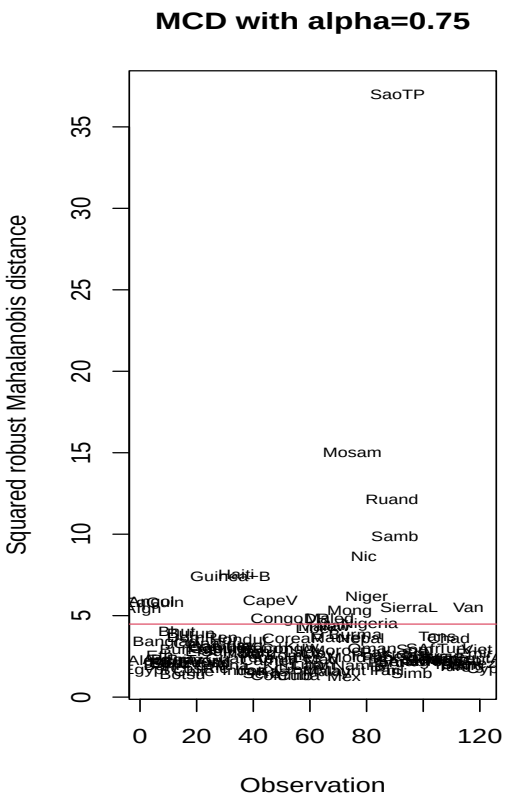
```
# Residuals vs. fitted
plot(unicefmm$fitted,unicefmm$residuals,xlab="Fitted values",ylab="Residuals",
  main="MM-estimator, residuals vs. fitted")
plot(unicefmm$init.S$fitted,unicefmm$init.S$residuals,xlab="Fitted values",
  ylab="Residuals",main="S-estimator, residuals vs. fitted")
```

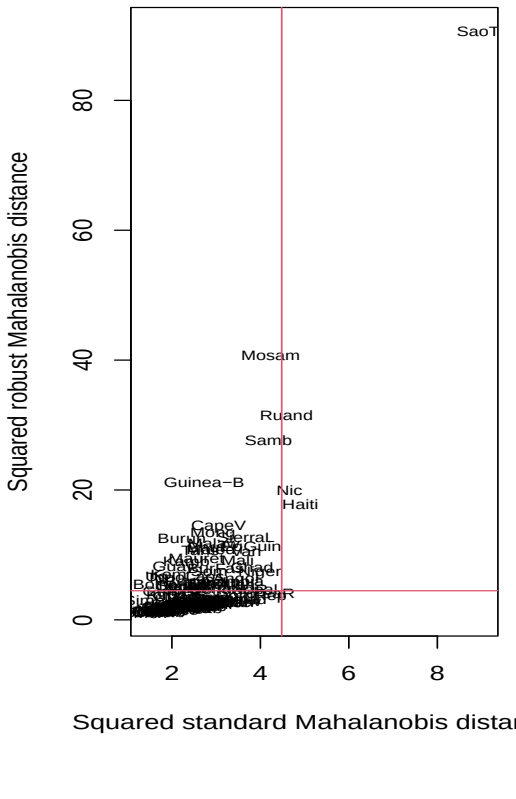


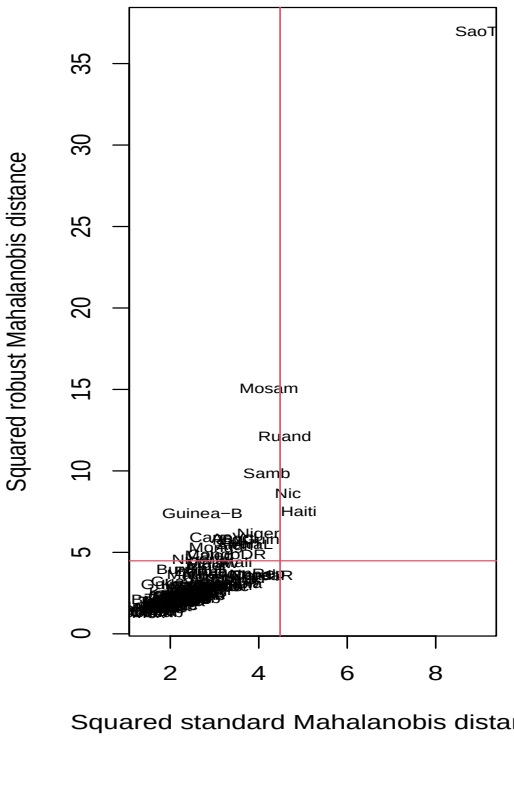
Finding outliers using covariance matrix of all variables

```
cunicef <- cov(unicef) # Covariance matrix
mcdunicef <- covMcd(unicef) # MCD with alpha=0.5
mcdu75 <- covMcd(unicef,alpha=0.75) # MCD with alpha=0.75

# Squared robust Mahalanobis distances
plot(1:121,sqrt(mcdunicef$mah),type="n",xlab="Observation",
     ylab="Squared robust Mahalanobis distance",main="MCD with alpha=0.5")
text(1:121,sqrt(mcdunicef$mah),rownames(unicef),cex=0.7)
abline(sqrt(qchisq(0.99,8)),0,col=2)
plot(1:121,sqrt(mcdu75$mah),type="n",xlab="Observation",
     ylab="Squared robust Mahalanobis distance",main="MCD with alpha=0.75")
text(1:121,sqrt(mcdu75$mah),rownames(unicef),cex=0.7)
abline(sqrt(qchisq(0.99,8)),0,col=2)

# Together with standard Mahalanobis distances
plot(sqrt(mahalanobis(unicef,colMeans(unicef),cunicef)),sqrt(mcdunicef$mah),
  type="n",xlab="Squared standard Mahalanobis distance",
ylab="Squared robust Mahalanobis distance",main="MCD with alpha=0.5")
text(sqrt(mahalanobis(unicef,colMeans(unicef),cunicef)),sqrt(mcdunicef$mah),
rownames(unicef),cex=0.7)
abline(sqrt(qchisq(0.99,8)),0,col=2)
abline(v=sqrt(qchisq(0.99,8)),col=2)
plot(sqrt(mahalanobis(unicef,colMeans(unicef),cunicef)),sqrt(mcdu75$mah),
  type="n",xlab="Squared standard Mahalanobis distance",
ylab="Squared robust Mahalanobis distance",main="MCD with alpha=0.75")
text(sqrt(mahalanobis(unicef,colMeans(unicef),cunicef)),sqrt(mcdu75$mah),
rownames(unicef),cex=0.7)
abline(sqrt(qchisq(0.99,8)),0,col=2)
abline(v=sqrt(qchisq(0.99,8)),col=2)
```


MCD with alpha=0.5
Squared robust Mahalanobis distance
Observation
SaoTP
Mosam
Ruand
Samb
Guinea−B
Nic
Haiti
CapeV
Mong
SierraL
Tans
Van
Niger
Chad
0
20
40
60
80
120


MCD with alpha=0.75
Squared robust Mahalanobis distance
Observation
SaoTP
Mosam
Ruand
Samb
Nic
CapeV
Niger
Mong
SierraL
Van
0
5
10
15
20
25
30
35
0
20
40
60
80
120


MCD with alpha=0.5
Squared robust Mahalanobis distance
Squared standard Mahalanobis distance
SaoT
Mosam
Ruand
Samb
Guinea−B
Nic
Haiti
CapeV
0
20
40
60
80
2
4
6
8


MCD with alpha=0.75
Squared robust Mahalanobis distance
Squared standard Mahalanobis distance
SaoT
Mosam
Ruand
Samb
Nic
Haiti
Guinea−B
0
5
10
15
20
25
30
35
2
4
6
8

```
# Data with outliers according to robust Mahalanobis distances (alpha=0.75)
# red, filled circles are also outliers w.r.t. standard Mahalanobis distance
```