Robust regression in R

Eva Cantoni

Research Center for Statistics and Geneva School of Economics and Management,
University of Geneva, Switzerland

April 4th, 2017

Against what is robust statistics robust?

Robust Statistics aims at producing consistent and possibly efficient estimators and test statistics with stable level when the model is *slightly* misspecified.

Model misspecification encompasses a relatively large set of possibilities, and robust statistics cannot deal with all types of model misspecifications.

By "slight model misspecification", we suppose that the data generating process lies in a *neighborhood* of the true (postulated) model, the one that is considered as "useful" for the problem under investigation.

Against what is robust statistics robust?

This neighborhood is formalized as

$$F_\varepsilon = (1 - \varepsilon)F_\theta + \varepsilon G, \tag{1}$$

- F_θ is the postulated model,
- θ is a set of parameters of interest,
- G is an arbitrary distribution and
- $0 \leq \varepsilon \leq 1$ captures "the amount of model misspecification"

Against what is robust statistics robust?

Inference	Classical		
G	$0 << \varepsilon < 1$	$0 < \varepsilon << 1$	$\varepsilon = 0$
arbitrary	?	?	$F_{\boldsymbol{\theta}}$
$G = \Delta_z$	F_ε	F_ε	$F_{\boldsymbol{\theta}}$
$G = F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	F_ε	F_ε	$F_{\boldsymbol{\theta}}$
G such that $F_\varepsilon = F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	$F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	$F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	$F_{\boldsymbol{\theta}}$
	Robust		
arbitrary	?	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$
$G = \Delta_z$	F_ε	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$
$G = F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	F_ε	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$
G such that $F_\varepsilon = F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	$F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$

Table: Models at which inference can be at best done.

Robustness measures

Robust estimators protect against:

- bias under contamination
- breakdown point

They imply a trade-off between efficiency and robustness!

Linear model and classical estimation

For $i = 1, \ldots, n$ consider

$$y_i = x_i^T \beta + \epsilon_i,$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

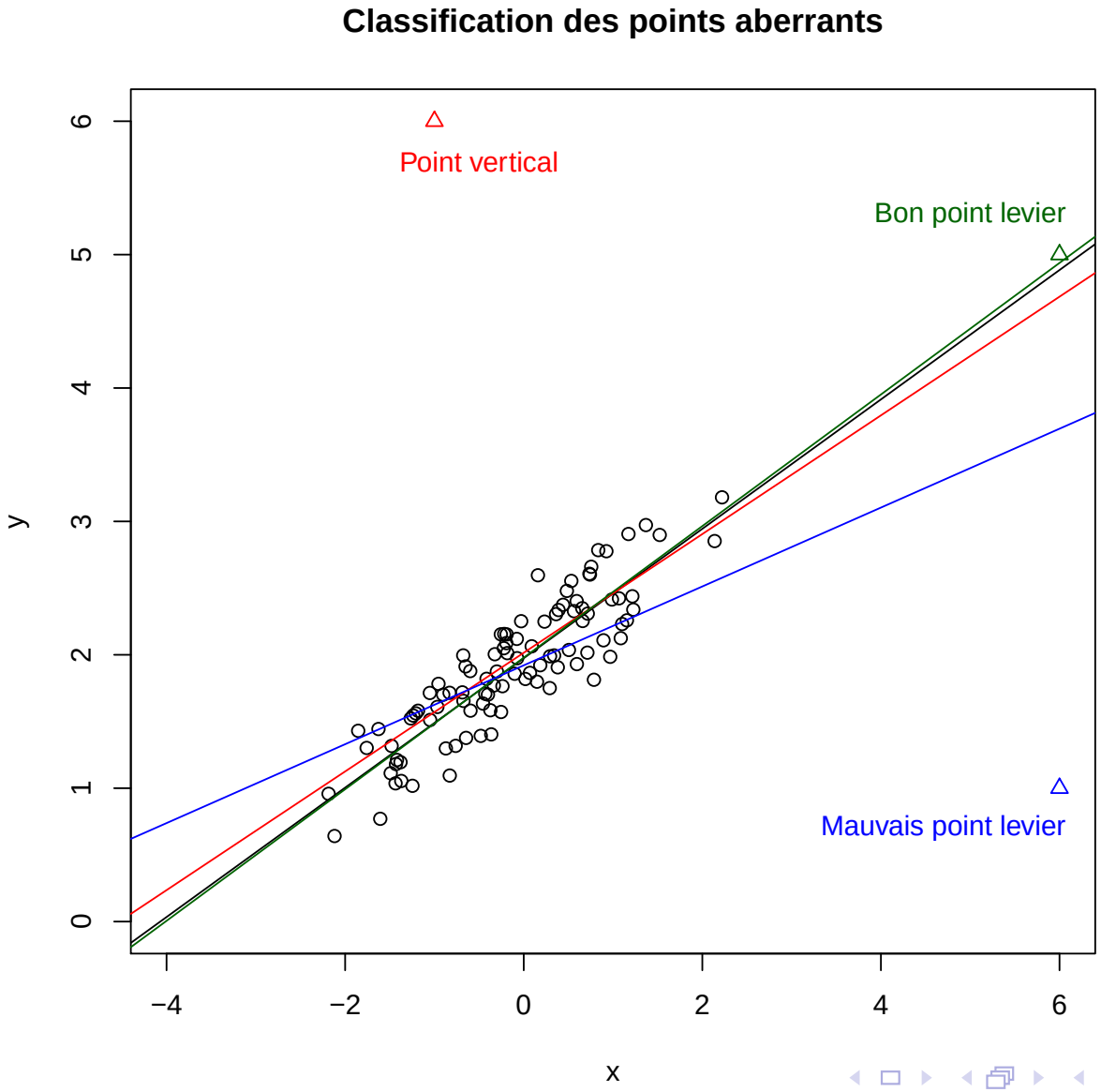
The maximum likelihood estimator $\hat{\beta}_{ML}$ minimizes

$$\sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}_{ML}}{\sigma} \right)^2 = \sum_{i=1}^{n} r_i^2,$$

or, alternatively, solves

$$\sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}_{ML}}{\sigma} \right) x_i = \sum_{i=1}^{n} r_i x_i = 0.$$

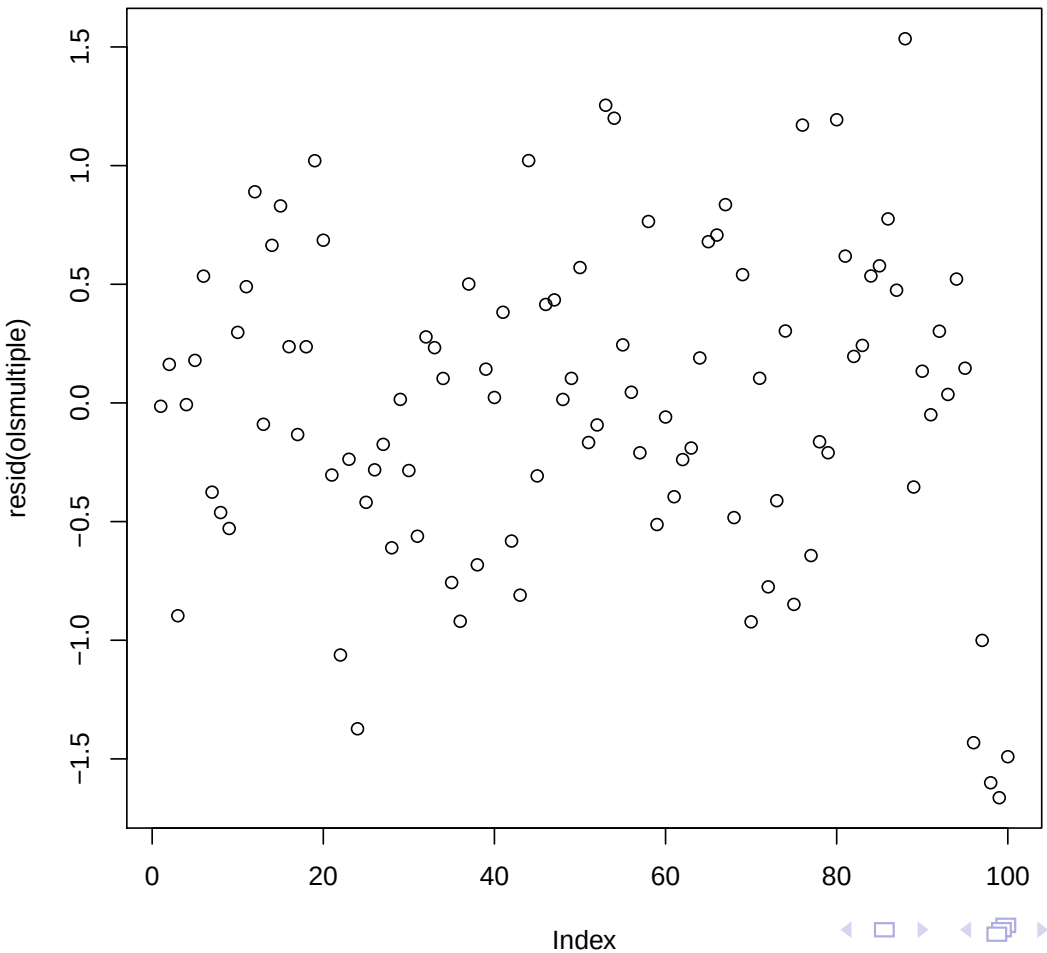
Outliers....


Classification des points aberrants
Point vertical
Bon point levier
Mauvais point levier
y
x
0
1
2
3
4
5
6
−4
−2
0
2
4
6

Robustness vs diagnostic

Masking.....

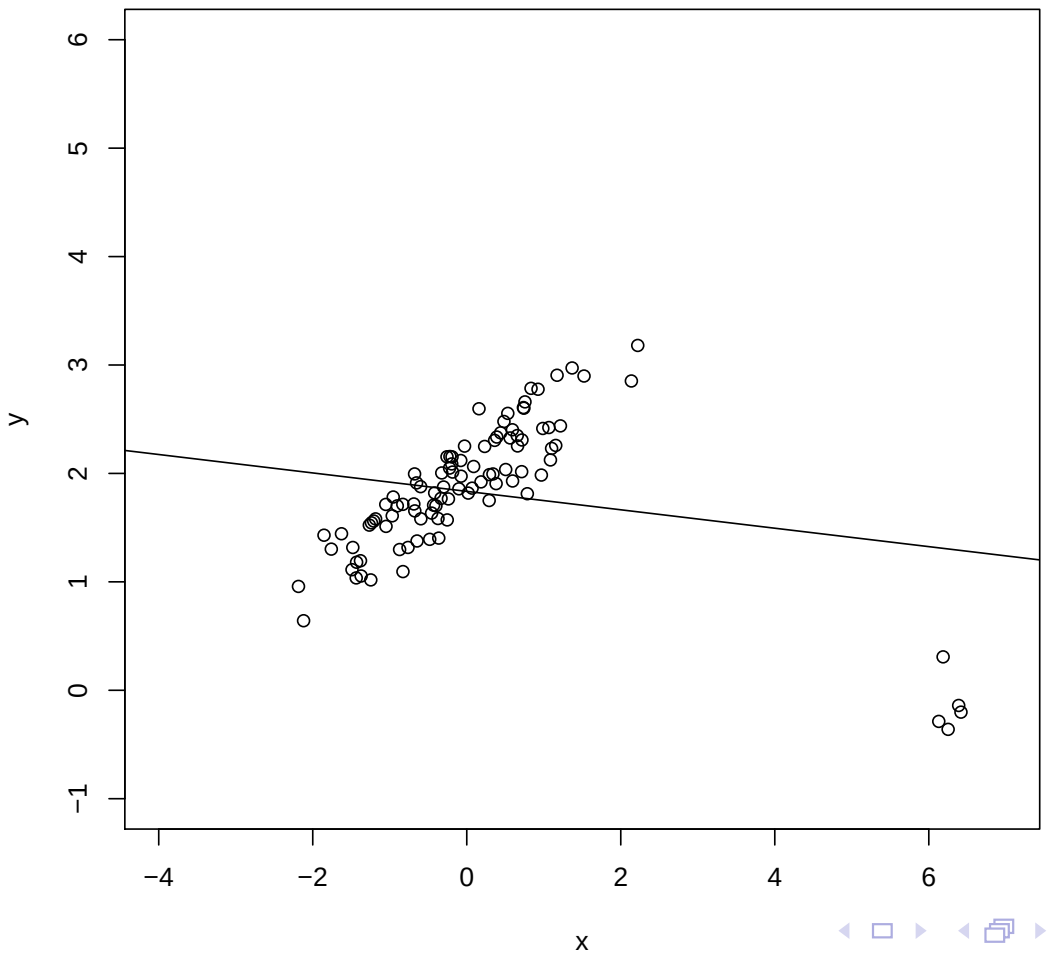
ML residuals



Robustness vs diagnostic

Masking.....

Data and ML fit



M and GM-estimation

The M-estimator $\hat{\beta}_M$ solves

$$\sum_{i=1}^{n} \psi_c \left(\frac{y_i - x_i^T \hat{\beta}_M}{\sigma} \right) x_i = \sum_{i=1}^{n} \psi_c(r_i) x_i = \sum_{i=1}^{n} \tilde{w}_c(r_i) \; r_i x_i = 0,$$

where $\tilde{w}_c(r_i) = \psi_c(r_i)/r_i$, or minimizes

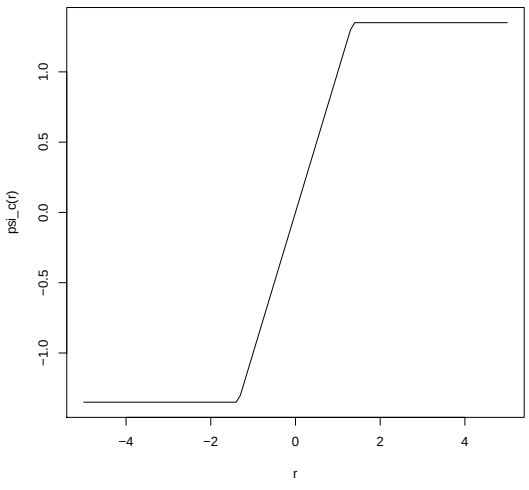
$$\sum_{i=1}^{n} \rho_c \left(\frac{y_i - x_i^T \hat{\beta}_M}{\sigma} \right).$$

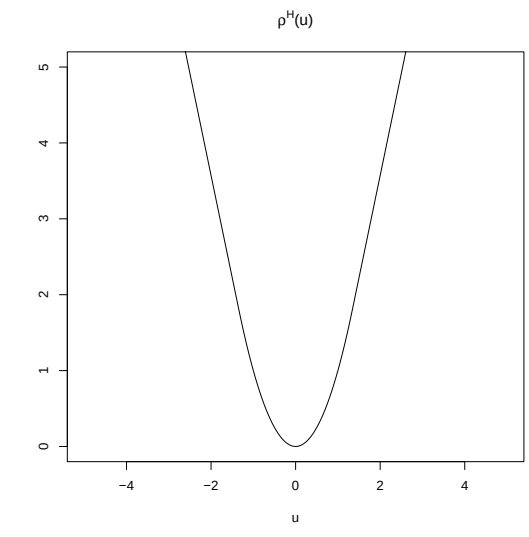
The Mallows GM-estimator $\hat{\beta}_{GM}$ is an alternative that solves

$$\sum_{i=1}^{n} \psi_c \left(\frac{y_i - x_i^T \hat{\beta}_{GM}}{\sigma} \right) w(x_i) x_i = \sum_{i=1}^{n} \tilde{w}(r_i) w(x_i) r_i x_i = 0.$$

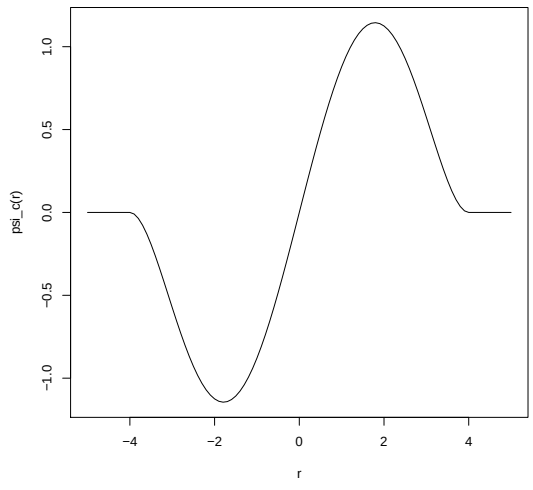
ψ_c and ρ_c functions

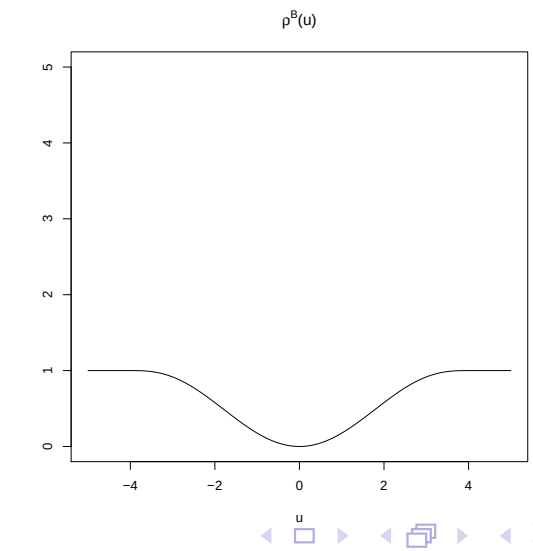
Huber



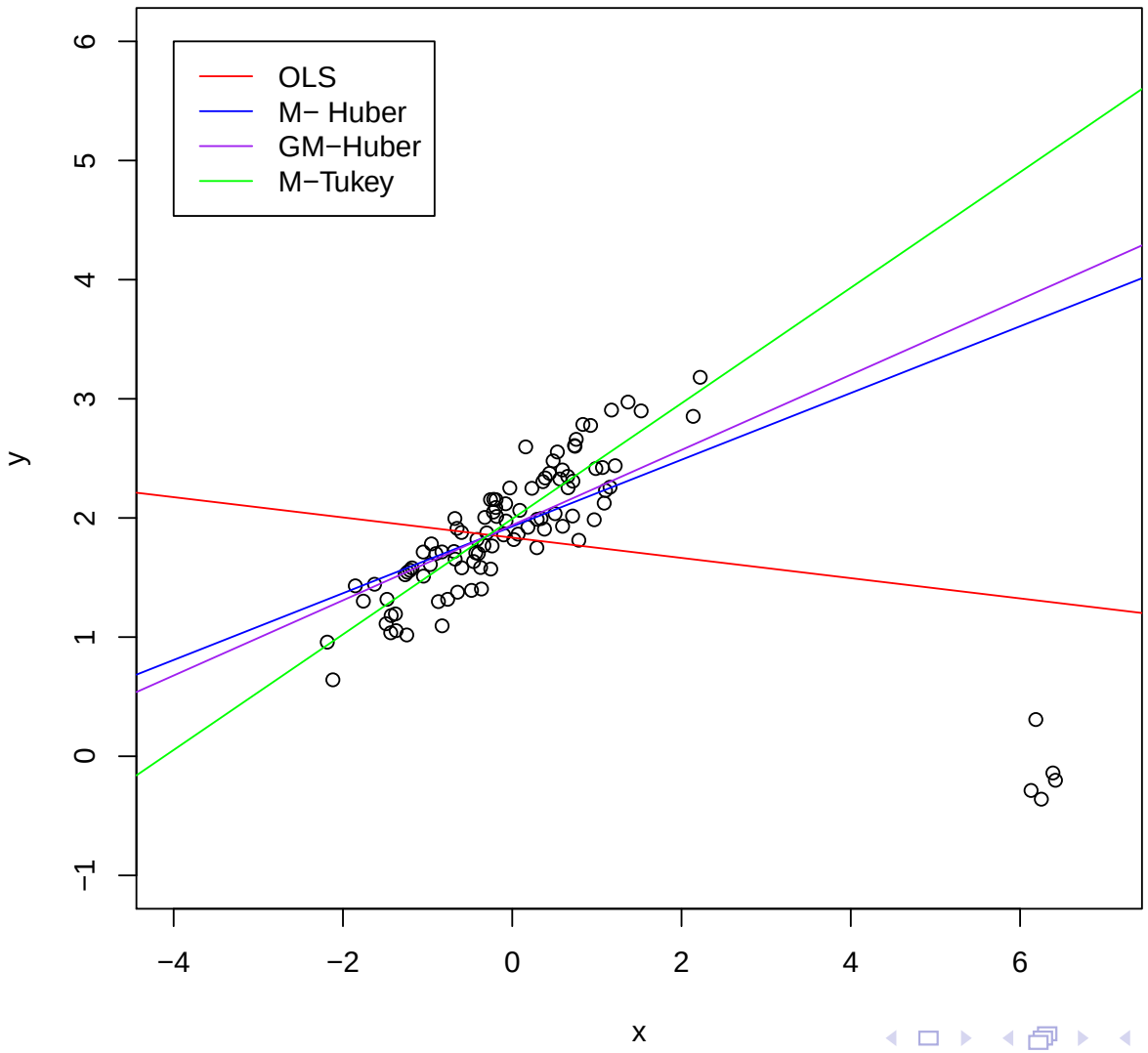


Tukey or bisquare





Comparison



S-estimation

The S-estimator $\hat{\beta}_S$ is an alternative that minimizes

$$\sum_{i=1}^{n} \rho_c \left(\frac{y_i - x_i^T \hat{\beta}_S}{s} \right),$$

where s is a scale M-estimator that solves

$$\frac{1}{n} \sum_{i=1}^{n} \rho_c^{(1)} \left(\frac{y_i - x_i^T \beta}{s} \right) = b.$$

,

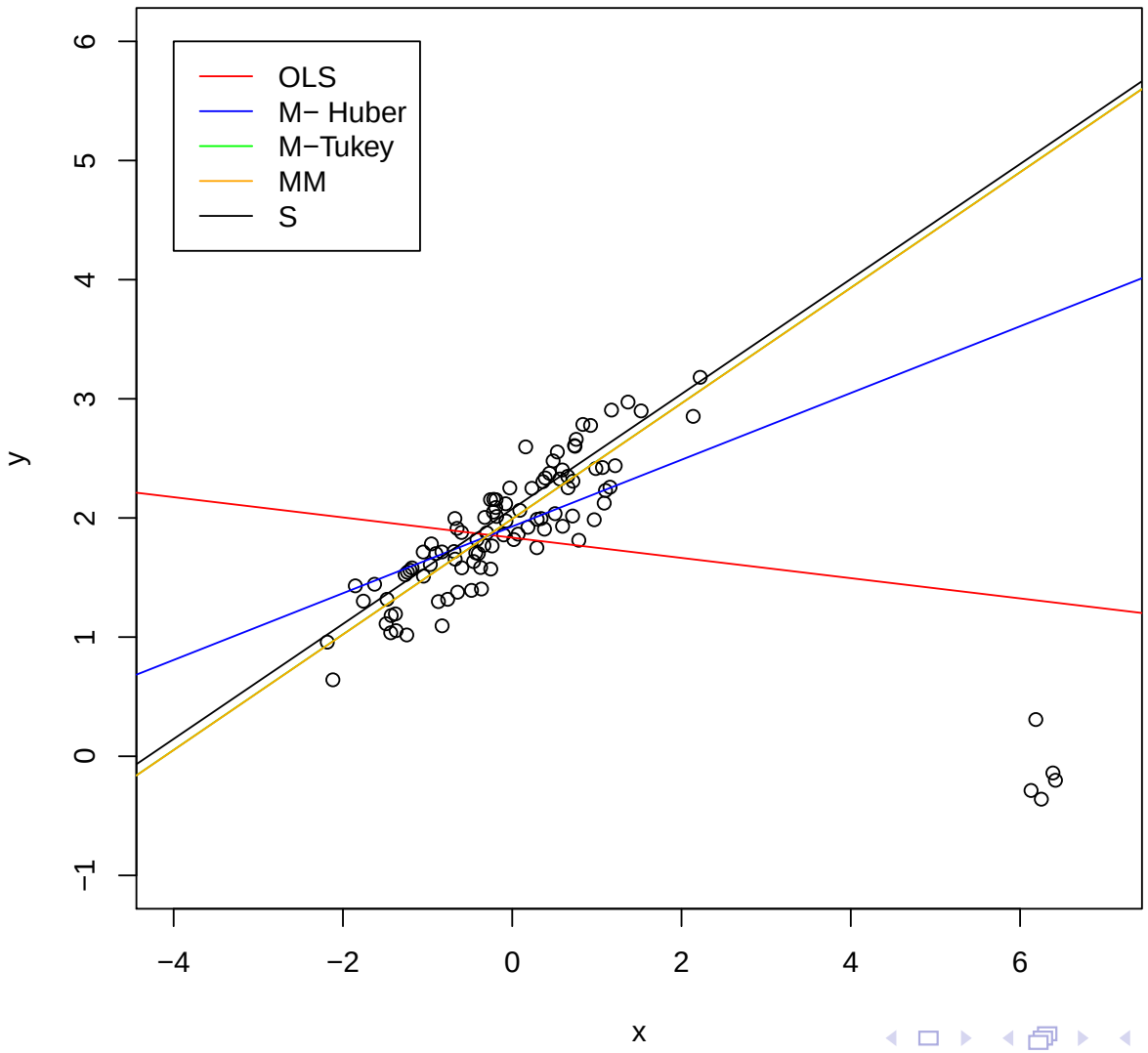
MM-estimation

The MM-estimator is a two-step estimator constructed as follow:

1. Let s_n be the scale estimate from an initial S-estimator.
2. With $\rho_c^{(2)}(\cdot) \leq \rho_c^{(1)}(\cdot)$, the MM-estimator $\hat{\beta}_{MM}$ minimizes

$$\sum_{i=1}^{n} \rho_c^{(2)} \left(\frac{y_i - x_i^T \hat{\beta}_{MM}}{s_n} \right).$$

Comparison



Robust GLM (GM-estimator)

For the GLM model (e.g. logistic, Poisson)

$$g(\mu_i) = x_i^T \beta$$

where $E(Y_i) = \mu_i$, $Var(Y_i) = v(\mu_i)$ and $r_i = \frac{(y_i - \mu_i)}{\sqrt{\phi v_{\mu_i}}}$, the robust estimator is defined by

$$\sum_{i=1}^{n} \left[\psi_c(r_i) w(x_i) \frac{1}{\sqrt{\phi v_{\mu_i}}} \mu_i' - a(\beta) \right] = 0, \qquad (2)$$

where $\mu_i' = \partial \mu_i / \partial \boldsymbol{\beta} = \partial \mu_i / \partial \eta_i \ \mathbf{x}_i$ and $a(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} E[\psi(r_i; c)] w(\mathbf{x}_i) / \sqrt{\phi v_{\mu_i}} \ \mu_i'$. The constant $a(\boldsymbol{\beta})$ is a correction term to ensure Fisher consistency.

R functions for robust linear regression

(G)M-estimation

- MASS: `rlm()` with `method=''M''` (Huber, Tukey, Hampel)
  Choice for the scale estimator: MAD, Huber Proposal 2

S-estimation

- robust: `lmRob` with `estim=''Initial''`
- robustbase: `lmrob.S`

MM-estimation

- MASS: `rlm()` with `method=''MM''`
- robust: `lmRob` (with `estim=''Final''`, default)
- `robustbase`: `lmrob()`

R functions for other models

- robustbase: `glmrob` GM-estimation, Huber (include Gaussian)

- Negative binomial model: `glmrob.nb` from `https://github.com/williamaeberhard/`
- From our book webpage: `http://www.unige.ch/gsem/rcs/members2/profs/eva-cantoni/books/robust-methods-in-biostatistics/`

Coleman

Dataset `coleman` from package `robustbase`.

A data frame with 20 observations on the following 6 variables.

- `salaryP`: staff salaries per pupil
- `fatherWc`: percent of white-collar fathers
- `sstatus`: socioeconomic status composite deviation: means for family size, family intactness, father's education, mother's education, and home items
- `teacherSc`: mean teacher's verbal test score
- `motherLev`: mean mother's educational level, one unit is equal to two school years
- `Y`: verbal mean test score (y, all sixth graders)

Coleman

```
> summary(m.lm)

Call:
lm(formula = Y ~ ., data = coleman)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9497 -0.6174  0.0623  0.7343  5.0018

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.94857   13.62755   1.464   0.1653
salaryP     -1.79333    1.23340  -1.454   0.1680
fatherWc     0.04360    0.05326   0.819   0.4267
sstatus      0.55576    0.09296   5.979 3.38e-05 ***
teacherSc    1.11017    0.43377   2.559   0.0227 *
motherLev   -1.81092    2.02739  -0.893   0.3868
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.074 on 14 degrees of freedom
Multiple R-squared:  0.9063,    Adjusted R-squared:  0.8728
F-statistic: 27.08 on 5 and 14 DF,  p-value: 9.927e-07
```

Coleman

```
> require(robustbase)
> summary(m.lmrob, setting = "KS2011")

Call:
lmrob(formula = Y ~ ., data = coleman)
 \--> method = "MM"
Residuals:
     Min       1Q   Median       3Q      Max
-4.16181 -0.39226  0.01611  0.55619  7.22766

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.50232    6.71260   4.544 0.000459 ***
salaryP     -1.66615    0.43129  -3.863 0.001722 **
fatherWc     0.08425    0.01467   5.741 5.10e-05 ***
sstatus      0.66774    0.03385  19.726 1.30e-11 ***
teacherSc    1.16778    0.10983  10.632 4.35e-08 ***
motherLev   -4.13657    0.92084  -4.492 0.000507 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 1.134
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.9747
Convergence in 11 IRWLS iterations

Robustness weights:
 observation 18 is an outlier with |weight| = 0 ( < 0.005);
 The remaining 19 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1491  0.9412  0.9847  0.9279  0.9947  0.9982
```

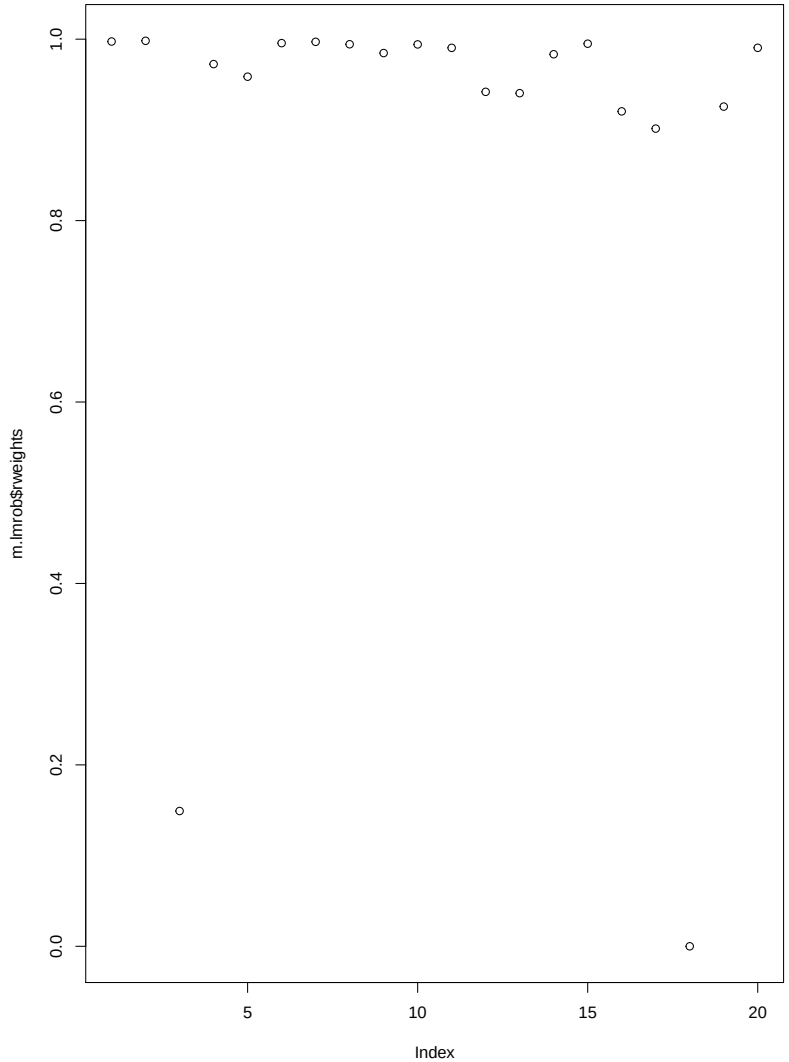
Coleman

Ctn'd:

```
Algorithmic parameters:
       tuning.chi              bb         tuning.psi        refine.tol              rel.tol
        1.548e+00        5.000e-01        4.685e+00        1.000e-07                1.000e-0
        solve.tol     eps.outlier            eps.x warn.limit.reject warn.limit.meanr
        1.000e-07        5.000e-03        1.569e-10        5.000e-01                5.000e-0
     nResample          max.it        best.r.s       k.fast.s            k.max     maxit.scal
           500              50               2              1              200
200
     trace.lev              mts      compute.rd fast.s.large.n
             0             1000               0           2000
                psi      subsampling                     cov compute.outlier.stats
         "bisquare"    "nonsingular"     ".vcov.avar1"                "SM"
seed : int(0)
```

Coleman

Robustness weights



Coleman

```
> summary(m.rlm)

Call: rlm(formula = Y ~ ., data = coleman)
Residuals:
     Min       1Q   Median       3Q      Max
-4.2059  -0.3886  -0.1092   0.4231   6.7054

Coefficients:
            Value   Std. Error  t value
(Intercept) 27.3497   7.6808     3.5608
salaryP     -1.6207   0.6952    -2.3314
fatherWc     0.0752   0.0300     2.5045
sstatus      0.6401   0.0524    12.2182
teacherSc    1.1557   0.2445     4.7271
motherLev   -3.5195   1.1427    -3.0801

Residual standard error: 0.7461 on 14 degrees of freedom
```

Breastfeeding

UK study on the decision of pregnant women to breastfeed. 135 expecting mothers asked on their feeding choice (`breast`= 1 if breastfeeding, try to breastfeed and mixed brest- and bottle-feeding, =0 if exclusive bottle-feeding).

Covariates: advancement of the pregnancy (`pregnancy`, end or beginning), how mothers fed as babies (`howfed`, some breastfeeding or only bottle-feeding), how mother's friend fed their babies (`howfedfriends`, some breastfeeding or only bottle-feeding), if had a partner (`partner`, no or yes), age (`age`), age at which left full time education (`educat`), ethnic group (`ethnic`, white or non white) and if ever smoked (`smokebf`, no or yes) or if stopped smoking (`smokenow`, no or yes). The first listed level of each factor is used as the reference (coded 0).

Breastfeeding

The sample characteristics are as follow:

out of the 135 observations, 99 were from mothers that have decided at least to try to breastfeed, 54 mothers were at the beginning of their pregnancy, 77 were themselves breastfed as baby, 85 of the mother's friend had breastfed their babies, 114 mothers had a partner, median age was 28.17 (with minimum equal 17 and maximum equal 40), median age at the end of education was 17 (minimum=14, maximum=38), 77 mothers were white and 32 mothers were smoking during the pregnancy, whereas 51 had smoked before.

Breastfeeding

$\texttt{breast}_i \sim \textit{Bernoulli}(\mu_i)$, so that $E(\texttt{breast}_i) = \mu_i$ and $\textit{Var}(\texttt{breast}_i) = \mu_i(1-\mu_i)$ (Binomial family).

Use the logit link.

$$\begin{aligned}\text{logit}(E(\texttt{breast})) = \text{logit}(P(\texttt{breast})) = \\ = \quad \beta_0 + \beta_1\texttt{pregnancy} + \beta_2\texttt{howfed} + \beta_3\texttt{howfedfr} \\ + \quad \beta_4\texttt{partner} + \beta_5\texttt{ethnic} + \beta_6\texttt{smokebf} \\ + \quad \beta_7\texttt{smokenow} + \beta_8\texttt{age} + \beta_9\texttt{educat},\end{aligned}$$

where $\text{logit}(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$, with $\mu_i/(1-\mu_i)$ being the odds of a success, and $\mu_i = P(\texttt{breast})$ is the probability of at least try to breastfeed.

Breastfeeding

```
> require(robustbase)
> breast.glmrobWx=glmrob(decwhat~howfedfr+ethnic+educat+age+grp+howfed+partner+
  smokenow+smokebf, family=binomial, weights.on.x="hat", tcc=1.5, data=breast)
> summary(breast.glmrobWx)

Call:  glmrob(formula = decwhat ~ howfedfr + ethnic + educat + age + grp +
howfed + partner + smokenow + smokebf, family = binomial, data = breast,
weights.on.x = "hat", tcc = 1.5)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.77423    3.35952  -2.314  0.02066 *
howfedfrBreast    1.49177    0.68777   2.169  0.03008 *
ethnicNon-white   2.68758    1.11582   2.409  0.01601 *
educat            0.37372    0.18486   2.022  0.04321 *
age               0.03116    0.05955   0.523  0.60082
grpBeginning     -0.81318    0.69269  -1.174  0.24042
howfedBreast      0.52823    0.70456   0.750  0.45342
partnerPartner    0.78295    0.81448   0.961  0.33641
smokenowYes      -3.44560    1.12137  -3.073  0.00212 **
smokebfYes        1.50733    1.09900   1.372  0.17020
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
Robustness weights w.r * w.x:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04103 0.82460 0.86500 0.82890 0.89400 0.93790

Number of observations: 135
Fitted by method Mqle  (in 12 iterations)

(Dispersion parameter for binomial family taken to be 1)
```

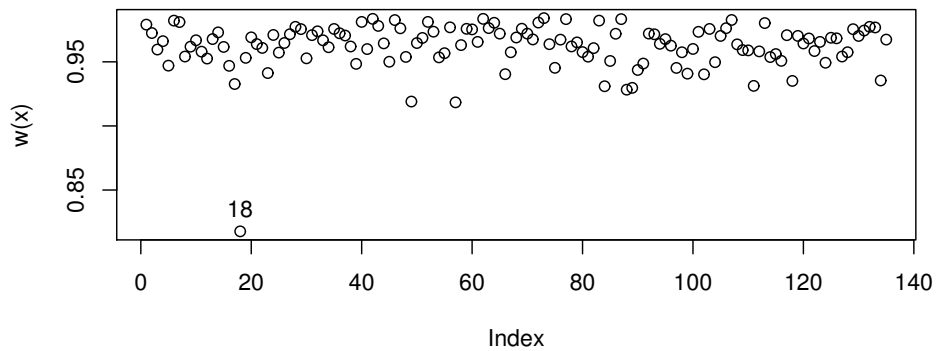
Breastfeeding

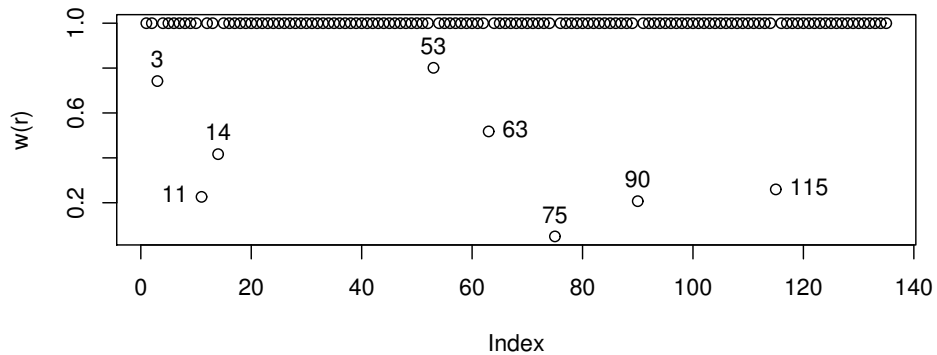
Ctn'd:

```
No deviance values available
Algorithmic parameters:
    acc     tcc
0.0001  1.5000
maxit
    50
test.acc
   "coef"
```

Breastfeeding

Robustness weights





Mailing list and conferences

Dedicated mailing list: `r-sig-robust@r-project.org`

ICORS 2017 in Wollongong, Australia:

ICORS 2016 in Geneva:

References

[1] Jean-Jacques Droesbeke, Gilbert Saporta, and Christine Thomas-Agnan. *Méthodes robustes en statistique*. Editions TECHNIP, 2015.

[2] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986.

[3] S. Heritier, E. Cantoni, S. Copt, and M.-P. Victoria-Feser. *Robust Methods in Biostatistics*. Wiley-Interscience, 2009.

[4] Peter J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[5] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, New York, 2009. Second Edition.

[6] R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust statistics*. Wiley New York, 2006.