Robust Statistics Collaborative Package Development: robustbase

Martin Maechler[1] Andreas Ruckstuhl[2]

[1]ETH Zurich

[2]ZHW Winterthur
Switzerland
maechler@R-project.org , rks@zhwin.ch

useR! 2006, Wien
June 16, 2006

Outline

- Robust Statistics with R : History
  - Robust Statistics with S (R) — JWT, etc
  - RsR — MASS
  - RsR — past: “miscellaneous”
- Robust Statistics with R — reloaded
  - Workshop in Treviso
  - new books on Robust Statistics
- The package `robustbase`
  - `robustbase`: current status overview
  - `robustbase`: lmrob and glmrob examples
  - `robustbase`: plans
  - `robustbase`: relation to other R packages
- Package writing collaboration
  - The people
  - The functions / classes
  - The other free software
- Conclusions

Robust Statistics with S (R) — JWT

- The father of EDA and early robustness:
  John W. Tukey
- @ Bell Labs: heavily influenced development of S. Hence basic robust tools have been part of S forever.
  - `median()`, `mad()` (also: `mean(*, trim=` α`)`)
  - `stem()`, `fivenum()` → `boxplot()` etc
  - `medianpolish()`, `smooth()`, `line(x,y)` (“Tukey line”!)
- Robust nonparametric regression: `lowess()` (but it has been known that `lowess()` is not really robust. . . . Because it starts from least-squares *instead* of robust smooth.
  `loess()` and `locfit` from package ‘locfit’ do about the same.
- For a better start, I had added `runmed()` to R 1.7.0, in early 2003 (package `modreg`, now part of `stats`).

Robust Statistics with R— the past II

- Venables and Ripley had added robust functionality to S and R with their “**MASS**” book and package
  - `huber()` and `hubers()` M-estimator for location
  - `cov.rob()` (with MVE and MCD) and `cov.trob()` for “*Resistant Estimation of Multivariate Location and Scatter*”
  - `lqs()` incl. LQS, LTS, LMS, and S estimator for high-breakdown point (=: HBP)
  - `rlm()` for more efficient HBP **r**obust fitting of **l**inear **m**odels (MM– or M-estimation).

Robust Statistics with R– "Miscellaneous"

Additionally, there have been miscellaneous R packages providing robust (or at least "resistant") methods:

quantreg "Quantile regression and related methods" by Roger Koenker . . . of course L_1, but has unbounded influence of $\mathbf{x}$.

sfsmisc (**SfS** = Seminar für Statistik, ETH Zurich): `rnls()`: robust **non**linear regression (robust 'nls') `f.robftest()`: "Robust F-test, i.e., Wald test for multiple coefficients of rlm() B"; further `rrange()` and `huberM()`.

forward: "Forward search approach to robust analysis in LM and GLM" by Kjell Konis and Marco Riani (for S+)

wle "Robustness via Weighted Likelihood" by Claudio Agostinelli

rrcov "Functions for Robust Location and Scatter Estimation and Robust Regression with High Breakdown Point" by Valentin Todorov; originally: new *fast* MCD and LTS.

Robust Statistics with R– reloaded

Reload of "R s R":
"Organized" effort to provide more R functionality for robustness
. . .

Robust Statistics with R– more "Miscellaneous"

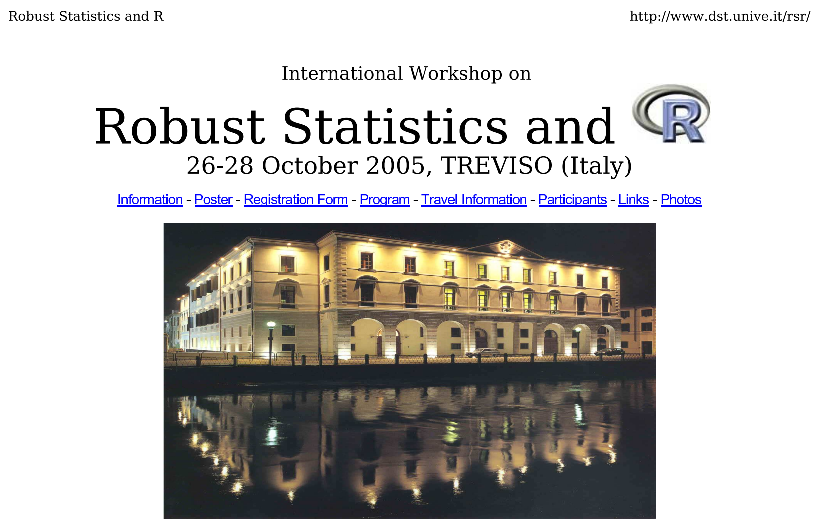
fields robust variograms etc by Doug Nychka

covRobust : `cov.nnve()` by Naisyin Wang and Adrian Raftery

amap : robust PCA `acprob()` and `varrob()`

multinomRob : overdispered multinomials

"Robust Statistics and R", Oct.2005, Treviso



Robust Statistics and R http://www.dst.unive.it/rsr/

International Workshop on

Robust Statistics and R

26-28 October 2005, TREVISO (Italy)

Information - Poster - Registration Form - Program - Travel Information - Participants - Links - Photos

“R s R”, Oct.2005, Treviso

Several working groups, notably

- Regression (incl. GLM)
- “Multivariate”

with the goal to unite efforts in providing more modern, coherent R functionality for robust statistics.

The package `robustbase`

“The” new package for robustness . . .
How to chose the package name ?

Had fun with a vote on chosing the package name. Every voter was allowed to allocate 3 votes; 20 “contestants” casting votes within a time limit. . . the final votes naming a new “basic robust statistics” R package were

robustbase	45
robustats	9
robusta	5
robustat	1

where I had voted (0,1,2,0) . . .

New books on robust statistics

Several classical books have had re-editions in 2005. . .

> *Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). Robust Statistics, Theory and Methods, Wiley.*

Ricardo Maronna and Victor Yohai — very reknown in robust statistics — took part in Treviso and agreed to support the idea of taking their book as a *target*:
→ Collaborate to provide “basic robust statistics” functionality in R, via *one* package:

`robustbase`: current status

1. Many data sets, particularly from the book of Rousseeuw and Leroy, mostly thanks to Valentin Todorov; all with full help pages:
   24 datasets, to be used in other packages, by, e.g., `data(wood, package = "robustbase")`. Data sets from Maronna, Martin and Yohai (2006) are also being added to the `robustbase` package.
2. `covMcd()` and `ltsReg()` by Valentin Todorov; originally in his `rrcov` package — now using shared code and notably using R’s random number generator (and seed).
   There have been `cov.mcd()` and `ltsreg()` in MASS. However, Valentin’s routines use the fast algorithms of Peter Rousseeuw and Katrien van Driessen (1999).

   (⟶ useR! talk by Valentin in Friday’s focus “robustness”)

robustbase: current status – 2 –

3. New functionality that hasn't been available in "public" R packages till now :
   - glmrob() by Andreas Ruckstuhl, based on Eva Cantoni's work for S-plus (and MM's for R) for robust Binomial GLMs, including model selection based on quasi deviance differences.

     E. Cantoni and E. Ronchetti (2001)
Robust Inference for Generalized Linear Models;
JASA **96**, *1022 ff*
   - lmrob() by Matias Salibian-Barrera, MM-estimate based on S.-B. & Yohai (2006) "*fast algorithm for S-regression*" (JCGS)
   - anova() model selection for both 'lmrob' and 'glmrob'. anova.lmrob() with option to choose between "Wald" and "Deviance" tests.
   - Qn() and Sn() scale estimates by Rousseeuw and Croux [50% breakdown but considerably more efficient than MAD]; based on their S-plus + Fortran code; ported to R by M.

robustbase: current status – 3 –

3. (..continued..)
   - covOGK(): The **o**rthogonalized **G**nanadesikan-**K**ettenring estimate for "fast" "high-dimensional" cov-estimation, by Maronna and Zamar (2002); based on code from Kjell Konis. This includes their univariate tau-estimate, I've called 'scaleTau2()' (since there's a different scaleTau() in other places), however amended with a consistency correction factor.
   - nlrob() for robust non-linear regression; this a slightly enhanced version of what has been available as 'rnls()' from package 'sfsmisc'. Also based mainly on Andreas Ruckstuhl's work.
   - huberM() — "a robust" version of MASS::huber()
4. Somewhat experimental code for an S4 class of "psi-function" (ψ, ρ, ψ', etc) objects.

lm–robust lmrob

An example of using lmrob():

```
> data(table.b13, package = "MPV")
> Jet <- table.b13
> Jet.r1 <- lmrob(y ~ ., data = Jet)
> summary(Jet.r1)

Call:
lmrob(formula = y ~ ., data = Jet)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-49.530 -17.897  -1.110  18.744  54.023
```

lm–robust lmrob – 2 –

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.023e+03  2.820e+03  -1.426 0.163150
x1           1.209e+00  3.060e-01   3.952 0.000385 ***
x2          -3.325e-02  6.895e-02  -0.482 0.632875
x3           2.022e-01  1.279e-01   1.581 0.123449
x4           3.525e+00  3.748e+00   0.941 0.353771
x5           8.291e-01  3.111e-01   2.665 0.011812 *
x6          -1.629e+01  3.461e+00  -4.706 4.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 23.77
Convergence in 33 IRWLS iterations
```

lm–robust lmrob – 3 –

```
Robustness weights:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5846  0.8970  0.9413  0.9116  0.9932  0.9999

Algorithmic parameters:

tuning.chi         bb tuning.psi refine.tol
 1.5476400  0.5000000  4.6850610  0.0000001
 nResample     max.it     groups    n.group    best.r.s    k.fast.
       500         50          5        400           2
     k.max compute.rd
       200          0
seed : int(0)
```

lm robust model comparison → anova.lmrob

Robust model comparison for robustly fit models:

```
> anova(Jet.r1, y ~ x1 + x5 + x6, test = "Wald")

Robust Wald Test Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6
Model 2: y ~ x1 + x5 + x6
Largest model fitted by lmrob(), i.e. MM

  pseudoDf Test.Stat Df Pr(>chisq)
1       33
2       36    4.4289  3     0.2187
```

lm: robust model comparison → anova.lmrob – 2 –

```
> try(anova(Jet.r1, y ~ x1 + x5 + x6, test = "Deviance"))

Error in anovaLmrobPair(obj0, ........) :
Please fit the nested models by lmrob

> Jet.r2 <- lmrob(y ~ x1 + x5 + x6, data = Jet)
> anova(Jet.r1, Jet.r2, test = "Deviance")

Robust Deviance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6
Model 2: y ~ x1 + x5 + x6
Largest model fitted by lmrob(), i.e. MM

  pseudoDf Test.Stat Df Pr(>chisq)
1       33
2       36     5.544  3     0.1360
```

GLM - "binomial" – robust: glmrob

An example of using glmrob() for robust GLM estimation:

```
> data(carrots)
> Cfit1 <- glm(cbind(success, total - success) ~ logdose +
+     block, data = carrots, family = binomial)
> summary(Cfit1)

Call:
glm(formula = cbind(success, total - success) ~ logdose + block,
    family = binomial, data = carrots)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9200  -1.0215  -0.3239   1.0602   3.4324

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0226     0.6501   3.111  0.00186 **
logdose      -1.8174     0.3439  -5.285 1.26e-07 ***
blockB2       0.3009     0.1991   1.511  0.13073
blockB3      -0.5424     0.2318  -2.340  0.01929 *
---
```

GLM - "binomial" – robust: glmrob – 2 –

```
> Cfit2 <- glmrob(cbind(success, total - success) ~ logdose +
+     block, family = binomial, data = carrots, method = "Mqle",
+     control = glmrobMqle.control(tcc = 1.2))
> summary(Cfit2)

Call:  glmrob(formula = cbind(success, total - success) ~ logdos


Coefficients:
            Estimate Std. Error z-value Pr(>|z|)
(Intercept)   2.3883     0.6923   3.450 0.000561 ***
logdose      -2.0491     0.3685  -5.561 2.68e-08 ***
blockB2       0.2351     0.2122   1.108 0.267828
blockB3      -0.4496     0.2409  -1.866 0.061989 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of observations: 24
Fitted by method ‘Mqle’  (in 9 iterations)

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

robust GLM for counts: “poisson”

```
> data(epilepsy)
> Efit1 <- glm(Ysum ~ Age10 + Base4 * Trt, family = poisson,
+     data = epilepsy)
> summary(Efit1)

Call:
glm(formula = Ysum ~ Age10 + Base4 * Trt, family = poisson, data

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0032  -2.0744  -1.0803   0.8202  11.0386

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.968014   0.135929  14.478  < 2e-16 ***
Age10               0.243490   0.041297   5.896 3.72e-09 ***
Base4               0.085426   0.003666  23.305  < 2e-16 ***
Trtprogabide       -0.255257   0.076525  -3.336 0.000851 ***
Base4:Trtprogabide  0.007534   0.004409   1.709 0.087475 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

robust GLM “poisson” – 2 –

```
> Efit2 <- glmrob(Ysum ~ Age10 + Base4 * Trt, family = poisson,
+     data = epilepsy, method = "Mqle", control = glmrobMqle.con
+         maxit = 100))
> summary(Efit2)

Call:  glmrob(formula = Ysum ~ Age10 + Base4 * Trt, family = poi


Coefficients:
                    Estimate Std. Error z-value Pr(>|z|)
(Intercept)         2.036768   0.154168  13.211  < 2e-16 ***
Age10               0.158434   0.047444   3.339 0.000840 ***
Base4               0.085132   0.004174  20.395  < 2e-16 ***
Trtprogabide       -0.323886   0.087421  -3.705 0.000211 ***
Base4:Trtprogabide  0.011842   0.004967   2.384 0.017124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of observations: 59
Fitted by method ‘Mqle’  (in 14 iterations)

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

robust GLM model comparison → anova.glmrob

By Andreas Ruckstuhl, based on work by Eva Cantoni (2004) JSS, and E.C.& Ronchetti (2001) JASA:

Continuing the example:

```
> Efit3 <- glmrob(Ysum ~ Age10 + Base4 + Trt, family = poisson,
+     data = epilepsy, method = "Mqle", control = glmrobMqle.con
+         maxit = 100))
> anova(Efit3, Efit2, test = "Wald")

Robust Wald Test Table

Model 1: Ysum ~ Age10 + Base4 + Trt
Model 2: Ysum ~ Age10 + Base4 * Trt
Models fitted by method ’Mqle’

  pseudoDf Test.Stat Df Pr(>chisq)
1       55
2       54    5.6836  1    0.01712 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(Efit3, Efit2, test = "QD")
```

robustbase: plans for the future

The source package has a file named TODO. It's open to the public at `https://svn.r-project.org/R-packages/trunk/robustbase/`.

1. Add S4 classes for "Covariance-estimator" objects `Cov`, i.e., "location and scatter", based on proposals of the working group in Treviso, then by Peter Filzmoser and Heinrich Fritz, and currently implented by Valentin Todorov
   ⟶ useR! talk by Valentin in Friday's focus "robustness".
2. S4 classes for `"psi-function"` (ψ, ρ, ψ', etc) objects, see above. Make use them, and consequently allow others than only Tukey's biweight.

robustbase: relation to other R packages

- `robustbase` provides *basic* infrastructure for other R packages:
- Basic algorithms: R functions, sometimes also with C API.
- Basic classes and methods: Classes "Cov", "psi_function", see above.
  Methods for plotting; possibly in conjunction with modularizing `plot.lm` into separate functions

Package writing collaboration

Experiences from collaborating with a diverse group of (potential) co-authors . . .

Package writing collaboration: The people

The DESCRIPTION file has as authors

Author: Original code by many authors, notably Peter Rousseeuw, Christophe Croux, see file 'Copyrights';
Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Martin Maechler

- meeting each other some time at first was important
- "talking" by e-mail: on a public (archived, searchable) mailing list
- talking in person from time to time — necessary (? !) much better motivation to get things done

Package writing collaboration: The functions / classes

Integration code from four to five different partly unpublished packages needs work, but has been achieved relatively easily:

- ‘rrcov’ (Valentin),
- ‘sfsmisc’ (Andreas, Martin),
- ‘robGLM’ (Eva → Martin → Andreas),
- ‘RobFit’ (Andreas),
- ‘roblm’ (Matias).

Package writing collaboration: other software

- There’s the R-SIG-robust mailing list, run via “Mailman”. as R-help and quite a few other lists, → `http://stat.ethz.ch/mailman/listinfo`
- Subversion `svn`: Version control of files with history, backtracking, branching and merging for collaborative software development
- Emacs, gcc, etc.

Conclusions

- “`robustbase`” is there to be used and built upon
- It will be extended in several ways
- Collaborative package development is exciting!