Comparison With Other Implementations of Regression M- and GM-Estimators

Tobias Schoch

University of Applied Sciences Northwestern Switzerland FHNW
School of Business, Riggenbachstrasse 16, CH-4600 Olten
tobias.schoch@fhnw.ch

November 17, 2022

Abstract. In this report, we study the behavior of the methods `svyreg_huberM` and `svyreg_huberGM` in package `robsurvey` with other implementations. We restricted attention to studying the methods for 4 well-known datasets. For all datasets under study, our implementations are identical (in terms of floating point arithmetic) with results of the competing implementations. Although our comparisons provide only anecdotal evidence on the performance of the methods, we believe that the comparisons shed some light on the behavior of our implementations. We are fairly confident that the methods in package `robsurvey` behave the way they are supposed to.

1 Introduction

In this short report, we compare the behavior of the regression M- and GM-estimators in package `robsurvey` with the methods from other implementations. To this end, we study the estimated parameters for four well-known datasets/ cases studies. With regard to competing implementations, we consider the methods from the following R packages:

```
MASS, version: 7.3.57
robeth, version: 2.7.6
```

These packages are documented in, respectively, (Venables and Ripley, 2002) and Marazzi (2020). The datasets are from package

```
robustbase, version: 0.95.0
```

see Mächler, Rousseeuw, Croux, Todorov, Ruckstuhl, Salibian-Barrera, Verbeke, Koller, Conceicao, and Anna di Palma (2022). In all comparisons, we

- study M- or GM-estimators with the MAD (normalized median absolute deviation) as estimator of scale;
- use the robustness tuning constant `k = 1.345` of the Huber ψ-function;
- focus on sample data that do not contain sampling weights.

All studied methods compute the regression estimates by iteratively reweighted least squares (IRWLS) and the estimate of scale (more precisely, the trial value for the scale estimate) is updated at each iteration.

Remark. Our comparisons provide only anecdotal evidence on the performance of the methods. Nonetheless, we believe that the comparisons shed some light on the behavior of our implementations.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote two real-valued p-vectors. We define the absolute relative difference by

$$\text{abs\_rel\_DIFF}(\boldsymbol{x}, \boldsymbol{y}) = 100\% \cdot \max_{i=1,\ldots,p} \left\{ \left| \frac{x_i}{y_i} - 1 \right| \right\}.$$

The remainder of the paper is organized as follows. In Section 2, we compare several implementations of the Huber M-estimator of regression. Section 3 studies implementations of the Huber GM-estimator of regression. In Section 4, we summarize the findings.

2 Huber M-estimators of regression

In this section, we study the Huber M-estimator of regression. The parametrizations of the algorithms have been chosen to make them comparable; we use:

- `MASS::rlm`: `method = "M"`, `scale.est = "MAD"`, `acc = 1e-5`, `test.vec = "coef"`, and `maxit = 50`,
- `robeth::rywalg`: `tol = 1e-5`, `maxit = 50`, `itype = 1`, `isigma = 2`, `icnv = 1`, and `maxis = 1`; see Marazzi (1993) for more details.
- `robsurvey::svyreg_huberM`: `tol = 1e-5`, and `maxit = 50`.

The methods `MASS::rlm` and `robeth::rywalg` compute the regression scale estimate by the (normalized) median of the absolute deviations (MAD) *about zero*. The method `svyreg_huberM` (and `svyreg_tukeyM`) implements two variants of the MAD:

- `mad_center = FALSE`: MAD centered about zero,
- `mad_center = TRUE`: MAD centered about the (weighted) median. (This is the default).

For ease of reference, we denote the MAD centered about zero by `mad0`.

In practice, the estimate of regression and scale differ whether the MAD is centered about zero or the median because the median of the residuals is not exactly zero for empirical data. If the residuals have a skewed distribution, the two variants of the MAD can differ by a lot.

2.1 Case 1: education data

The `education` data are on public education expenditures (at the level of US states), and are from Chatterjee and Price (1977) [see Chatterjee and Hadi (2012) for a newer edition]; see also Rousseeuw and Leroy (1987). The dataset contains 4 variables: the response variable (`Y`: per capita expenditure on public education in a state, projected for 1975) and the three explanatory variables

- `X1`: Number of residents per thousand residing in urban areas in 1970,
- `X2`: Per capita personal income in 1973,
- `X3`: Number of residents per thousand under 18 years of age in 1974.

The following tabular output shows the estimated coefficients (and the estimated scale; last column) under the model `Y ~ X1 + X2 + X3` for 4 different implementations/ methods.

```
R> data(education, package = "robustbase")
R> M_compare(Y ~ X1 + X2 + X3, education)
```

```
                     (Intercept)    X1    X2    X3  scale
svyreg_huberM           -434.837 0.030 0.061 1.270 40.379
svyreg_huberM (mad0)    -434.396 0.031 0.061 1.269 40.120
rywalg (ROBETH)         -434.465 0.031 0.061 1.269 40.161
rlm (MASS)              -434.395 0.031 0.061 1.269 40.120
```

The estimates of the 4 methods differ only slightly. We have the following findings:

- `svyreg_huberM (mad0)` is based on the MAD centered about zero. In methodological terms, it is identical with the implementations `rlm (MASS)` and `rywalg (ROBETH)`. The estimates of `svyreg_huberM (mad0)` are virtually identical with the ones of `rlm (MASS)`. The estimates of `rywalg (ROBETH)` deviate more from the other methods.
- `svyreg_huberM` is based on the MAD centered about the (weighted) median. The estimates differ slightly from `svyreg_huberM (mad0)`.

The discrepancies are mainly due to the normalization constant to make the MAD an unbiased estimator of the scale at the Gaussian core model. In `rlm (MASS)`, the MAD about zero is computed by `median(abs(resid)) / 0.6745`. The constant $1/0.6745$ is equal to 1.482580 (with a precision of 6 decimal places), which differs slightly from $1/\Phi^{-1}(0.75) = 1.482602$, where Φ denotes the cumulative distribution function of the standard Gaussian. The implementation of `svyreg_huberM` uses 1.482602 (see file `src/constants.h`). Now, if we replace $1/0.6745$ in the above code snippet by 1.482602 in the function body of `rlm.default`, then the regression coefficients of the so modified code and `svyreg_huberM` are (in terms of floating point arithmetic) almost identical. The absolute relative difference is

```
R> design <- svydesign(id = ~1, weights = rep(1, nrow(education)),
+                     data = education)
R> m1 <- svyreg_huberM(Y ~ X1 + X2 + X3, design, k = 1.345,
+                     mad_center = FALSE, tol = 1e-5,
+                     maxit = 50)
```

```
R> rlm_mod <- MASS:::rlm.default
R> body(rlm_mod)[[22]][[4]][[3]][[3]][[2]][[3]][[3]][[3]] <-
+       substitute(median(abs(resid)) * 1.482602)
R> m2 <- rlm_mod(m1$model$x, m1$model$y, k = 1.345,
+                  method = "M", scale.est = "MAD", acc = 1e-5,
+                  maxit = 50, test.vec = "coef")
R> cat("\nabs_rel_DIFF: ", 100 * max(abs(coef(m1) / coef(m2) - 1)),
+       "%\n")
```

```
abs_rel_DIFF:  1.054712e-12 %
```

Next, we consider comparing the estimated (asymptotic) covariance matrix of the estimated regression coefficients. To this end, we computed the diagonal elements of the estimated covariance matrix for the methods `svyreg_huberM (mad0)` and `rlm (MASS)`; see below. In addition, we computed the absolute relative difference between the two methods.

```
R> M_compare_cov(Y ~ X1 + X2 + X3, education)
```

```
 (Intercept)           X1           X2           X3
1.548342e+04 2.694538e-03 1.373338e-04 1.010169e-01
 (Intercept)           X1           X2           X3
1.548341e+04 2.694537e-03 1.373338e-04 1.010168e-01

abs_rel_DIFF:  5.187249e-05 %
```

The diagonal elements of the estimated covariance matrix differ only slightly between the two methods. The discrepancies can be explained by the differences in terms of the estimated coefficients.

2.2 Case 2: stackloss data

The `stackloss` data consist of 21 measurements on the oxidation of ammonia to nitric acid for an industrial process; see Brownlee (1965). The variables are:

- `Air Flow`: flow of cooling air,
- `Water Temp`: cooling water inlet temperature,
- `Acid Conc.`: concentration of acid [per 1000, minus 500],
- `stack.loss`: stack loss.

The variable `stack.loss` (stack loss of amonia) is regressed on the explanatory variables air flow, water temperature and the concentration of acid. The regression coefficients and the estimate of scale are tabulated for the 4 implementations/ methods under study.

```
R> data(stackloss, package = "datasets")
R> M_compare(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
+             stackloss)
```

```
                     (Intercept) Air.Flow Water.Temp Acid.Conc. scale
svyreg_huberM            -41.051    0.827      0.939     -0.129 2.530
svyreg_huberM (mad0)     -41.027    0.829      0.926     -0.128 2.441
rywalg (ROBETH)          -41.027    0.829      0.926     -0.128 2.442
rlm (MASS)               -41.027    0.829      0.926     -0.128 2.441
```

The estimates of the regression M-estimator which is based on the MAD centered about zero are virtually identical (see rows 2–4). The estimates of `svyreg_huberM` deviate slightly from the latter because it is based on the MAD centered about the (weighted) median.
We did not repeat the analysis on differences in the estimated covariance matrices because the results are qualitatively the same as in Case 1.

3 Huber GM-estimators of regression

In this section, we consider regression GM-estimators with Huber ψ-function (tuning constant fixed at $k = 1.345$). The scale is estimated by MAD. With regard to the MAD, we distinguish two cases: `svyreg_huberGM` and `svyreg_huberGM (mad0)`, where `mad0` refers to the MAD about zero.

We computed the weights to downweight leverage observations (`xwgt`) with the help of the methods in package `robeth`. The so computed weights were then stored to be utilized in all implementations of GM-estimators of regression. This approach ensures that the implementations do not differ in terms of the `xgwt`'s.

3.1 Case 3: delivery data

The `delivery` data consist of observations on servicing 25 soft drink vending machines. The data are from Montgomery and Peck (2006); see also Rousseeuw and Leroy (1987). The variables are:

- `n.prod`: number of products stocked in the vending machine,
- `distance`: distance walked by the route driver (ft),
- `delTime`: delivery time (minutes).

The goal is to model/ predict the amount of time required by the route driver to service the vending machines. The variable `delTime` is regressed on the variables `n.prod` and `distance`.

Mallows GM-estimator

The regression coefficients and the estimate of scale are tabulated for the 3 implementations/ methods under study.

```
R> data(delivery, package = "robustbase")
R> GM_mallows_compare(delTime ~ n.prod + distance, delivery)
```

```
                                    (Intercept) n.prod distance scale
svyreg_huberGM (Mallows)                  4.468  1.514     0.01 2.446
svyreg_huberGM (Mallows, mad0)            4.476  1.509     0.01 2.255
rywalg (ROBETH, Mallows)                  4.476  1.509     0.01 2.256
```

The estimates of `svyreg_huberGM (Mallows, mad0)` are almost identical with results of `rywalg (ROBETH, Mallows)`; see rows 2 and 3. The estimates of `svyreg_huberGM (Mallows)` (i.e., based on the MAD centered about the weighted median differ slightly as is to be expected.

Schweppe GM-estimator

```
R> GM_schweppe_compare(delTime ~ n.prod + distance, delivery)
```

```
                                     (Intercept) n.prod distance scale
svyreg_huberGM (Schweppe)                  4.011  1.429    0.014 1.398
svyreg_huberGM (Schweppe, mad0)            4.012  1.429    0.014 1.392
rywalg (ROBETH, Schweppe)                  3.964  1.430    0.014 1.434
```

The estimates of `svyreg_huberGM (Schweppe, mad0)` and `rywalg (ROBETH, Schweppe)` (see rows 2 and 3) are slightly different. We could not figure out the reasons for this discrepancy.

3.2 Case 4: salinity data

The `salinity` data are a set of measurements of water salinity and river discharge taken in North Carolina's Pamlico Sound; Ruppert and Carroll (1980); see also Rousseeuw and Leroy (1987). The variables are

- `Y`: salinity,
- `X1`: salinity lagged two weeks,
- `X2`: linear time trend,
- `X3`: river discharge.

There a 28 observations. We consider fitting the model `Y ~ X1 + X2 + X3` by several implementations of the regression GM-estimators.

Mallows GM-estimator

```
R> data(salinity, package = "robustbase")
R> GM_mallows_compare(Y ~ X1 + X2 + X3, salinity)
```

```
                                    (Intercept)    X1     X2     X3 scale
svyreg_huberGM (Mallows)                 18.884 0.721 -0.174 -0.655 0.763
svyreg_huberGM (Mallows, mad0)           18.877 0.721 -0.174 -0.654 0.768
rywalg (ROBETH, Mallows)                 18.869 0.721 -0.174 -0.654 0.774
```

The differences between the estimates of `svyreg_huberGM (Mallows, mad0)` and `rywalg (ROBETH, Mallows)` are larger (see rows 2 and 3) than in Case 3. Still, the estimates are very similar.

Schweppe GM-estimator

```
R> GM_schweppe_compare(Y ~ X1 + X2 + X3, salinity)
```

```
                                  (Intercept)     X1      X2      X3 scale
svyreg_huberGM (Schweppe)              19.911 0.679 -0.173 -0.675 0.707
svyreg_huberGM (Schweppe, mad0)        19.916 0.679 -0.173 -0.675 0.682
rywalg (ROBETH, Schweppe)              19.974 0.680 -0.177 -0.679 0.732
```

The estimates of `svyreg_huberGM (Schweppe, mad0)` and `rywalg (ROBETH, Schweppe)` (see rows 2 and 3) are slightly different. But the differences are minor.

4 Summary

In this paper, we studied the behavior of the methods `svyreg_huberM` and `svyreg_huberGM` in package `robsurvey` with other implementations. We restricted attention to studying the methods for four well-known datasets. For all datasets under study, our implementations replicate (or are at least very close to) the results of the competing implementations. Although our comparisons provide only anecdotal evidence on the performance of the methods, we believe that the comparisons shed some light on the behavior of our implementations.

References

BROWNLEE, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, New York: John Wiley and Sons, 2nd ed.

CHATTERJEE, S. AND A. HADI (2012). *Regression Analysis by Example*, Hoboken (NJ): John Wiley and Sons, 5th ed.

CHATTERJEE, S. AND B. PRICE (1977). *Regression Analysis by Example*, New York: John Wiley and Sons.

MÄCHLER, M., P. ROUSSEEUW, C. CROUX, V. TODOROV, A. RUCKSTUHL, M. SALIBIAN-BARRERA, T. VERBEKE, M. KOLLER, E. L. T. CONCEICAO, AND M. ANNA DI PALMA (2022). *robustbase: Basic Robust Statistics*, r package version 0.95-0.

MARAZZI, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics: The FORTRAN Library ROBETH with an interface to S-PLUS*, New York: Chapman & Hall.

MARAZZI, A. (2020). *robeth: R Functions for Robust Statistics*, r package version 2.7-6.

MONTGOMERY, D. C. AND E. A. PECK (2006). *Introduction to Linear Regression Analysis*, Hoboken (NJ): John Wiley and Sons, 4th ed.

ROUSSEEUW, P. J. AND A. M. LEROY (1987). *Robust Regression and Outlier Detection*, (Wiley Series in Probability and Statistics), Hoboken (NJ): John Wiley and Sons.

RUPPERT, D. AND R. J. CARROLL (1980). Trimmed least squares estimation in the linear model, *Journal of the American Statistical Association* **75**, 828–838.

VENABLES, W. N. AND B. D. RIPLEY (2002). *Modern Applied Statistics with S*, New York: Springer, 4th ed.