Approximate Replication of High-Breakdown Robust Regression Techniques

Achim Zeileis
Wirtschaftsuniversität Wien

Christian Kleiber
Universität Basel

Abstract

We present a case study demonstrating that without data and code archives reproducibility is more the exception than the rule, especially if modern, complex algorithms are employed. Specifically, we show that stochastic extensions of OLS, as required in some combinatorial optimization problems arising in high-breakdown robust regression, can be difficult to replicate in the absence of detailed information on tuning parameters and further computational issues.

Keywords: combinatorial optimization, least squares, replication, robust regression, stochastic algorithm.

1. Introduction

Zaman, Rousseeuw, and Orhan (2001), in a paper aimed at popularizing robust regression techniques among economists, apply the least trimmed squares (LTS) and minimum covariance determinant (MCD) methods (Rousseeuw 1984) to three economic data sets. Specifically, they reanalyze an augmented Solow model applied to OECD countries (Nonneman and Vanhoudt 1996), a time series regression explaining US stock returns (Benderly and Zwick 1985), and a growth study for a cross section of 61 countries (De Long and Summers 1991).

Here "robust" means resistant to extreme (i.e., outlying or influential) observations; specifically, the methods used here can withstand up to 50% contamination in large samples. The LTS estimator is typically implemented via running a large number of OLS regressions (with certain adjustments) on random subsets of the data, thus it may be considered as a stochastic extension of the standard OLS method. Similarly, the MCD estimator is implemented via estimating covariances for a large number of random subsets, again with certain adjustments.

In the following, we attempt to replicate the results of Zaman *et al.* (2001), in the narrow sense of exact numerical replication using the same data and methodology. It emerges that, in the absence of the exact code and function calls used by the original authors, this seemingly simple task requires a substantial amount of reverse engineering. This ties in with the recent interest in reproducible research in economics and the suggestion of mandatory data and code archives (see, e.g. Anderson, Greene, McCullough, and Vinod 2008). In the case at hand, the absence of archived code virtually prevented the reproduction of published results.

We use the R system for statistical computing (R Development Core Team 2008), version 2.8.1, and the implementations of LTS and MCD in the R package **MASS** (Venables and Ripley 2002), version 7.2-45. Both are freely available from the Comprehensive R Archive Network

This is a preprint of an article published in *Journal of Economic and Social Measurement*, **34**(2–3), 191–203. Copyright © 2009 IOS Press and the authors. doi:10.3233/JEM-2009-0314

at http://CRAN.R-project.org/. The exact function calls for replicating our analysis are available in Appendix A.

2. Replication

The approach of Zaman *et al.* (2001) consists of running OLS on a subset of the data. This subset does not contain certain outlying observations and is determined utilizing two robust methods: First, the LTS estimator (Rousseeuw 1984) is used for flagging all observations with large residuals. In a second step, in order to not exclude too many such points (not all of which are dangerous), Zaman *et al.* (2001) suggest to also consider the leverages of the observations, determined via the robust minimum covariance determinant (MCD) method (Rousseeuw 1984, 1985). The final analysis then excludes only those observations with simultaneously large LTS residuals and high leverages (called *bad leverage* points) from a subsequent OLS regression, while those observations with large LTS residuals but low leverages (called *vertical outliers*) are retained in the regression. See Zaman *et al.* (2001) and the references therein for further details; a convenient recent survey of high-breakdown methodology is available in Hubert, Rousseeuw, and van Aelst (2008).

2.1. Computational issues

All the robust methods employed here are governed by hyperparameters, typically providing trimming parameters or cutoffs in certain algorithms. The end result of the computations often depends, to a considerable extent, on the selection of these parameters.

A first issue is the choice of trimming parameters in the optimization problems under consideration. More specifically, LTS minimizes the criterion

$$\sum_{i=1}^{q} (r^2)_{i:n}$$

where $(r^2)_{i:n}$ denotes the ith smallest out of n squared residuals. The parameter q determines the amount of trimming and thus the degree of robustness of the resulting estimator. Setting q to $\lfloor (n+k+1)/2 \rfloor$ yields maximal robustness (where k is the number of regressors including the constant term), but any value between $\lfloor (n+k)/2 \rfloor$ and n is admissible. For the MCD estimator, which minimizes the determinant of the covariance matrix for a subsample of size q, the choice $q = \lfloor (n+k)/2 \rfloor$ yields maximal robustness (Hubert *et al.* 2008). For assuring reproducibility, these trimming parameters need to be provided, but they are not available in the paper under investigation.

Both problems essentially represent combinatorial optimization problems, and thus are somewhat distinct from the optimization problems typically encountered in econometrics. One way of solving the LTS optimization problem consists of running all $\binom{n}{q}$ OLS regressions utilizing q observations. Similarly, the MCD estimate can, in principle, be obtained by an exhaustive search over all subsets of size q. Unfortunately, this is rarely feasible in real-world applications as it would require to consider a vast number of subsamples. Instead, stochastic algorithms exploring large numbers of OLS regressions or sample covariances for random samples of size $p \leq q$ are used, with certain refinements. Specifically, the algorithms FastLTS and FastMCD (see Rousseeuw and van Driessen 1999; Hubert *et al.* 2008) have been suggested which start out from random samples of size $p = k$. Note that this does not guarantee that the global minima are found. From a replication perspective, one would at least want to assure that the same solutions are found (before, potentially, investigating whether these are good, or even optimal, solutions). Thus, a thorough description of the computational strategy is needed. Typically, a good summary of this is the code employed for the analyses (including starting values and, in the case of stochastic searches, preferably also random seeds). Below, we use implementations of LTS and MCD that are available in the **MASS** package (Venables and Ripley 2002): function `lqs()` includes an implementation of LTS and function `cov.rob()` implements MCD by means of the FastMCD algorithm. It is worth noting that here these algorithms find the global optima in at least two out of three applications (where the sample sizes are small enough to check by exhaustive searches).

A final but important issue consists in the choice of cutoffs that determine the subset on which to run OLS, namely the "good" observations. As noted above, the observations with simultaneously large LTS residuals and high leverages are excluded, where "large" is typically defined in terms of certain χ^2 quantiles (Hubert *et al.* 2008), e.g., $\sqrt{\chi^2_{1;0.975}} = 2.24$ for the LTS residuals and $\sqrt{\chi^2_{k-1;0.975}}$ for the leverages. Obviously, providing both cutoffs (at least in the underlying code, but preferably also in the printed paper) is essential to assure reproducibility of analyses employing these methods. However, Zaman *et al.* (2001) just provide the cutoff used for the LTS residuals. It was chosen as 2.5, a common rule of thumb replacing $\sqrt{\chi^2_{1;0.975}}$. In the absence of code, we can just guess that the $\sqrt{\chi^2_{k-1;0.975}}$ quantile was used for the robust distances.

We now discuss all three examples in turn, proceeding by increasing sample size of the original data sets.

2.2. Nonneman and Vanhoudt regression

We begin with the Solow model for OECD countries originally considered by Nonneman and Vanhoudt (1996), a regression of per capita (of working age) GDP growth on per capita GDP in 1960 (Y_0), the average annual ratio of domestic investment to real GDP (S_k) and annual population growth plus 5% (N), for a cross section of 22 OECD countries. As for all other data sets, we are able to successfully replicate the plain OLS regression coefficients and associated standard errors as well as the OLS regression after omitting those observations indicated by Zaman *et al.* (2001).

However, we encountered problems with the LTS residuals and the robust leverages given in their paper. First, their robust leverages appear to have arisen from a local optimum. We are able to reproduce their results by setting a suitable random seed (found by reverse engineering) and just taking a single solution. For these leverages the value of the criterion (i.e., the determinant of the covariance matrix) equals -12.64 (on a log scale), while an exhaustive search over all $\binom{22}{13}$ possible subsets yields a global minimum at -13.21.

Second, we could not reproduce the LTS residuals for the usual recommendation of $q = \lfloor (n + k + 1)/2 \rfloor = \lfloor (22 + 4 + 1)/2 \rfloor = 13$. Fortunately, in view of the modest sample size of 22 observations it is feasible to run all $\binom{22}{q}$ OLS regressions for any trimming parameter q, and thus solve the problem exactly. Our computations suggest that $q = 16$ was used: running all $\binom{22}{16} = 74613$ OLS regressions employing samples of size 16 yields exactly the results described by Zaman *et al.* (2001). Thus Canada, Turkey and New Zealand are the bad leverage points

Table 1: Robust regression coefficients (and standard errors) for Nonneman and Vanhoudt growth regression with $q = 22$ (OLS without omitting observations), $q = 16$ (omitting Canada, Turkey, New Zealand) and $q = 13$ (omitting Canada, USA, Turkey, Australia).

Variable	$q = 22$	$q = 16$	$q = 13$
Constant	2.976	4.715	3.776
	(1.022)	(1.166)	(1.282)
$\log(Y_0)$	-0.343	-0.412	-0.451
	(0.056)	(0.054)	(0.057)
$\log(S_k)$	0.650	0.518	0.703
	(0.202)	(0.179)	(0.191)
$\log(N)$	-0.573	-0.124	-0.650
	(0.290)	(0.352)	(0.419)

with LTS residuals equaling 4.21, -6.14, and -3.17, and corresponding suboptimal robust distances of 7.25, 9.36, and 5.98.

To complement these findings, we compared the above results to those obtained from utilizing the exact MCD estimator (i.e., the estimator based on an exhaustive search). Fortunately, the results are essentially identical: the same observations are selected as bad leverage points (now with robust distances of 5.14, 4.50, and 7.20), and hence the final robust OLS regression is the same.

It is also of interest to check how these results are affected if we use $q = 13$ in the LTS regression, the value of the trimming parameter yielding maximal robustness. It turns out there are slight changes, in that the bad leverage points are now Canada, USA, Turkey and Australia. Thus Canada and Turkey are still excluded; in addition, USA and Australia are flagged while this is no longer true for New Zealand. The final regression exhibits the same regressors as statistically significant as the regression based on LTS using 16 data points, but the coefficients are somewhat different (see Table 1). The largest change is associated with the coefficient on population growth which is, however, insignificant as before.

In summary, our reanalysis detects a local optimum in the MCD and apparent use of a non-standard trimming parameter in the LTS optimization problem.

2.3. Benderly and Zwick regression

In the Benderly and Zwick time series regression explaining US stock returns from 1954 to 1981, it is again feasible to run all $\binom{28}{16} = 30421755$ OLS regressions and thus solve the LTS problem exactly. We note that the authors of the original nonrobust OLS analysis (Benderly and Zwick 1985) already described some form of model instability in the sample period, suggesting that the stable period is 1956–1976.

Using the default trimming parameters $q = \lfloor(n + k + 1)/2\rfloor = \lfloor(28 + 3 + 1)/2\rfloor = 16$ and $q = \lfloor(n+k)/2\rfloor = \lfloor(28+3)/2\rfloor = 15$ for the LTS and MCD problems, we are able to replicate the bad leverage points. However, neither the robust residuals nor the robust distances agree. Specifically, the LTS residuals for the bad leverage points (the years 1979 and 1980) are 2.69

and 2.68 and thus fairly close, but not identical, to the values 2.60 and 2.82 given by Zaman *et al.* (2001), potentially pointing to a slightly inferior LTS fit. (Note that a differing q, as was the case in the preceding regression, cannot explain these deviations—we tried all qs!) Regarding the robust distances, we obtain 4.24 and 4.2 in contrast to the values 3.65 and 3.55 given by Zaman *et al.* (2001). Interestingly, the latter are perfectly reproducible for the trimming $q = \lfloor (n + k + 1)/2 \rfloor = \lfloor (28 + 3 + 1)/2 \rfloor = 16$, the value of q used for the LTS estimator. Fortunately, all conclusions drawn from this and, in particular, the resulting OLS regression (after omitting the observations for 1979 and 1980) are unaffected.

However, it is worth noting that with these data, there is the only deviation with respect to the OLS results for the full data set, in that we obtain a different R^2 and F statistic. Specifically, Zaman *et al.* (2001) report $R^2 = 0.56$ and $F = 10.5$, while we obtain $R^2 = 0.496$ and $F = 12.306$. We have been unable to identify the source of these discrepancies. In contrast, the corresponding results for the OLS regression after robust preprocessing, namely $R^2 = 0.65$ and $F = 21.04$, are in perfect agreement with those reported by Zaman *et al.* (2001).

In summary, our reanalysis cannot reproduce the LTS residuals for the two bad leverage points and the leverages only if we employ a slightly larger trimming parameter; also, the R^2 and F statistic of an OLS regression appear to be in error.

2.4. De Long and Summers regression

We conclude with the computationally most demanding example. Specifically, in the growth study using the De Long and Summers (1991) data it is no longer feasible to determine the exact solution via an exhaustive search, as this would require running no fewer than $\binom{61}{33} = 191724747789809255$ (i.e., some 200 quadrillion) regressions in total. Hence, we must confine ourselves to an approximate LTS estimator in this example. We use one million random samples of size q (we tried larger values up to one billion samples, with virtually identical results).

A problem with this example is that here Zaman *et al.* (2001) appear to employ a different strategy, in that only LTS residuals are considered but not the leverages. This is implicit from the terminology used there because only a "vertical outlier" is excluded from the final regression. However, more casual reading of the paper might suggest a uniform approach (i.e., excluding only "bad leverage points") was employed throughout. To illustrate, Figure 1 provides a plot of LTS residuals vs. robust MCD distances for this regression. This plot reveals that, according to the strategy of the preceding examples, no bad leverage points exist. With a value of -5.20, Zambia by far has the largest LTS residual in absolute size while its robust distance, 2.20, is not unusually large, thus suggesting not to exclude this observation according to the strategy followed in the preceding two examples. On the other hand, a mechanical analysis excluding all vertical outliers would also omit Cameroon in view of its absolute LTS residual of 2.95 exceeding 2.5. In addition, there are two borderline cases with residuals in the vicinity of 2.5 and moderately large robust distances, namely Chile and Spain. Both would not be excluded under the previously followed strategy; however, they would be flagged as bad leverage points, again in a fairly mechanical analysis, if one used the cutoff $\sqrt{\chi^2_{1;0.975}} = 2.24$ (solid lines) for the LTS residuals, as recommended by Hubert *et al.* (2008). In spite of these robust diagnostics, Zaman *et al.* (2001) prefer to run OLS excluding only the largest vertical outlier Zambia.

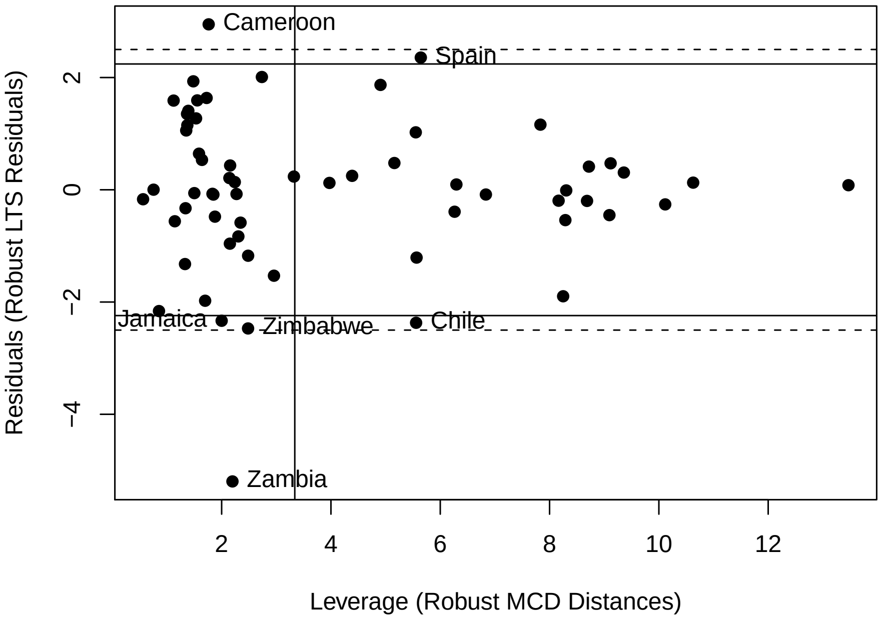

Figure 1: Robust LTS residuals vs. robust MCD distances for De Long and Summers regression, with highlighted outliers. Dashed lines at ±2.5, solid lines at $\pm\sqrt{\chi^2_1; 0.975}$ (horizontal) and $\sqrt{\chi^2_4; 0.975}$ (vertical).

We add that Huber-style M estimators, which are robust to vertical outliers but not to bad leverage points, also yield large residuals for Zambia and Cameroon and virtually identical results as an OLS regression excluding these observations. This suggests that, in this example, high-breakdown methods are not required – of course, the whole point of high-breakdown techniques is that one does not have to know this in advance! We note that Zambia would also be flagged by classical, but non-robust leave-one-out regression diagnostics, among them Cook's distances.

3. Conclusions

The preceding section revealed substantial difficulties with replicating the robust regression results of Zaman *et al.* (2001), partly due to insufficient computational detail in the original analysis.

Our findings are of interest for at least two reasons: First, they highlight that even methodology reasonably close to plain OLS, in our case a stochastic algorithm making use of a large number of OLS regressions, is not always easy to replicate. A detailed description of chosen settings is therefore needed. Naturally, this is in conflict with the scarcity of available journal space, but there would seem to be an easy way out: put this information into an online supplement, namely the data and code archive of the journal.

This leads to our second point: our results would seem to support the recent proposals of mandatory data and code archives, see, e.g., Anderson *et al.* (2008), McCullough, McGeary, and Harrison (2006), McCullough, McGeary, and Harrison (2008), and the references therein. With such an archive, our exercise would have been much easier: although code would have been in a different programming language and presumably been long obsolete, it would at least have been clear how hyperparameters were selected and thus to what extent resulting differences could be attributed to such settings. However, *Economics Letters* currently has no archive. Unfortunately, this is not the first incident regarding a paper from that journal that proved difficult to replicate. Recently, Davis (2007) encountered problems with a paper aimed at measuring pro-poor growth, apparently due to inconsistent growth spell classifications by the original author. Since the original data were available to Davis as well as to us, these incidents underline that the existence of a data archive alone would not have helped – only the exact code will enable researchers to fully replicate earlier results. We suggest *Economics Letters* introduces a data and code archive as soon as possible.

Acknowledgments

We are grateful to Bruce McCullough and Valentin Todorov for valuable comments on a previous version.

References

Anderson RD, Greene WH, McCullough BD, Vinod HD (2008). “The Role of Data/Code Archives in the Future of Economic Research.” *Journal of Economic Methodology*, **15**, 99–119.

Benderly J, Zwick B (1985). “Inflation, Real Balances, Output and Real Stock Returns.” *American Economic Review*, **75**, 1115–1123.

Davis GA (2007). “Measuring Unambiguously Pro-Poor Growth.” *Journal of Economic and Social Measurement*, **32**(4), 253–261.

De Long JB, Summers LH (1991). “Equipment Investment and Economic Growth.” *Quarterly Journal of Economics*, **106**, 445–501.

Hubert M, Rousseeuw PJ, van Aelst S (2008). “High-Breakdown Robust Multivariate Methods.” *Statistical Science*, **23**(1), 92–119.

McCullough BD, McGeary KA, Harrison TD (2006). “Lessons from the JMCB Archive.” *Journal of Money, Credit and Banking*, **38**(4), 1093–1107.

McCullough BD, McGeary KA, Harrison TD (2008). “Do Economics Journal Archives Promote Replicable Research?” *Canadian Journal of Economics*, **41**(4), 1406–1420.

Nonneman W, Vanhoudt P (1996). “A Further Augmentation of the Solow Model and the Empirics of Economic Growth for OECD Countries.” *Quarterly Journal of Economics*, **111**, 943–953.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Rousseeuw PJ (1984). "Least Median of Squares Regression." *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw PJ (1985). "Multivariate Estimation with High Breakdown Point." In W Grossmann, G Pflug, I Vincze, W Wertz (eds.), *Mathematical Statistics and Applications*, pp. 283–297. Reidel, Dordrecht.

Rousseeuw PJ, van Driessen K (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics*, **41**, 212–223.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Zaman A, Rousseeuw PJ, Orhan M (2001). "Econometric Applications of High-Breakdown Robust Regression Techniques." *Economics Letters*, **71**, 1–8.

A. R code

This appendix provides the full R code to replicate our replication study.

Initially, the **MASS** package (Venables and Ripley 2002) is loaded and the number of displayed digits is reduced:

```
R> library("MASS")
R> options(digits = 4)
```

A.1. Data

All three data sets are provided in space-separated plain text format (with column *and* row names). They can be easily read into R via

```
R> nv  <- read.table("NonnemanVanhoudt.dat")
R> bz  <- read.table("BenderlyZwick.dat")
R> dls <- read.table("DeLongSummers.dat")
```

A.2. High-breakdown robust regression

The following code chunk defines a convenience function `robreg()` implementing the strategy described by Zaman *et al.* (2001).

```
robreg <- function(formula, data, cutoff = NULL,
  quantile = NULL, psamp = NULL, nsamp = "exact",
  method = "mcd", dist_nsamp = "exact")
{
  ## OLS results
  fm_ols <- lm(formula, data)

  ## default: choose psamp = quantile
  n <- length(residuals(fm_ols))
  k <- length(coef(fm_ols))
  if(is.null(cutoff)) cutoff <- c(2.5, sqrt(qchisq(0.975, k-1)))
  if(is.null(quantile)) quantile <- floor((n + k + c(1, 0))/2)
  if(is.null(psamp)) psamp <- quantile[1]

  ## LTS results with robust residuals
  fm_lts <- lqs(formula, data,
    quantile = quantile[1], psamp = psamp, nsamp = nsamp)
  rr <- residuals(fm_lts)/fm_lts$scale[2]
  rr_nok <- abs(rr) > cutoff[1]

  ## robust leverage via MCD (or MVE)
  X <- model.matrix(fm_ols)[,-1]
  rc <- cov.rob(X, method = method,
    quantile = quantile[2], nsamp = dist_nsamp)
  rd <- sqrt(mahalanobis(X, rc$center, rc$cov))
  rd_nok <- rd > cutoff[2]

  ## ROBUST results
  nok <- rr_nok & rd_nok
  fm_rob <- lm(formula, data[!nok,])

  rval <- list(ols = fm_ols, lts = fm_lts, robust = fm_rob,
    robcov = rc, robresid = rr, robdist = rd,
    bad_leverage = nok, psamp = psamp, method = method,
    nsamp = list(lts = nsamp, dist = dist_nsamp),
    quantile = list(lts = quantile[1], dist = quantile[2]),
    cutoff = list(lts = cutoff[1], dist = cutoff[2]))
  return(rval)
}
```

Given a description of a regression model by a `formula` and `data`, it first fits the OLS regression. Then it fits the LTS regression minimizing the sum of squares of the `quantile[1]` smallest residuals (default: $\lfloor (n+k+1)/2 \rfloor$) using the function `lqs()` from package **MASS**. By default all possible samples (`nsamp = "exact"`) of size `psamp = quantile[1]` are searched assuring that the LTS minimization problem is solved exactly. Subsequently, it computes the robust leverages via `cov.rob()`; by default the MCD estimator is computed with `quantile[2]` set to $\lfloor (n+k)/2 \rfloor$. For `cov.rob()` the argument `nsamp = "exact"` means that all $\binom{n}{k}$ subsamples of size $p = k$ (often called "elemental sets") will be searched. Next, those observations with scaled LTS residuals greater than `cutoff[1]` (default: 2.5) *and* robust leverages greater than `cutoff[2]` (default: $\sqrt{\chi^2_{k-1;0.975}}$) are then flagged as bad leverage points and excluded in a final OLS regression. A list of all (intermediate and final) results is returned.

A.3. Nonneman and Vanhoudt regression

The Zaman *et al.* (2001) MCD covariance estimate appears to correspond to a local optimum. It can be reproduced by setting a suitable random seed and just taking a single solution. Furthermore, while the usual recommendation of $q = 13$ seems to have been used for the MCD estimate, $q = 16$ apparently has been employed in the LTS regression. The code chunk

```
R> set.seed(2)
R> nv_fit <- robreg(log(gdp85/gdp60) ~ log(gdp60) +  log(invest) +
+    log(popgrowth + .05), data = nv, quantile = c(16, 13), dist_nsamp = 1)
```

reproduces the results of Zaman *et al.* (2001):

```
R> nv_fit$robresid[nv_fit$bad_leverage]

     Canada      Turkey New Zealand
      4.206      -6.144      -3.167

R> nv_fit$robdist[nv_fit$bad_leverage]
```

```
     Canada      Turkey New Zealand
      7.251       9.361       5.976
```

However, it would have been more natural to take $q = 13$ (the default in `robreg()`) for both LTS and MCD and perform exhaustive searches for both problems:

```
R> nv_fit2 <- robreg(log(gdp85/gdp60) ~ log(gdp60) +  log(invest) +
+    log(popgrowth + .05), data = nv)
```

This confirms that MCD indeed did not find the optimum in the first setting: there, the value of the objective function is

```
R> nv_fit$robcov$crit
```

```
[1] -12.64
```

while with an exhaustive search we obtain

```
R> nv_fit2$robcov$crit
```

```
[1] -13.21
```

Fortunately, the suboptimal MCD estimate does not change the results qualitatively: Combining the exact LTS estimate for $q = 16$ and the exact MCD estimate for $q = 13$ identifies the same bad leverage points as indicated in Zaman *et al.* (2001), namely

```
R> nv_fit$robresid[abs(nv_fit$robresid) > 2.5 & abs(nv_fit2$robdist) > 3.06]
```

```
     Canada      Turkey New Zealand
      4.206      -6.144      -3.167
```

where $3.06 = \sqrt{\chi^2_{3;0975}}$.

However, if we follow the usual recommendation and use $q = 13$ also for LTS, the results change slightly, in that Canada, USA, Turkey, Australia are now selected as the bad leverage points:

```
R> nv_fit2$robresid[nv_fit2$bad_leverage]
```

```
   Canada        USA     Turkey Australia
    9.073      6.236     -4.027     4.518
```

```
R> nv_fit2$robdist[nv_fit2$bad_leverage]
```

```
   Canada        USA     Turkey Australia
    5.144      4.503      7.203     4.504
```

A.4. Benderly and Zwick regression

For these data, using the default settings

```
R> bz_fit <- robreg(returns ~ growth + inflation, data = bz)
```

reproduces the bad leverage points obtained by Zaman *et al.* (2001). However, neither the LTS residuals nor the MCD leverages match their published values:

```
R> bz_fit$robresid[bz_fit$bad_leverage]

 1979  1980
2.688 2.679

R> bz_fit$robdist[bz_fit$bad_leverage]

 1979  1980
4.237 4.203
```

The LTS residuals are quite close but the leverages are clearly different. However, using $q = 16$ instead of the default recommendation $q = \lfloor (n+k)/2 \rfloor = \lfloor (28+3)/2 \rfloor = 15$, we are able to exactly reproduce the robust leverages:

```
R> bz_fit2 <- robreg(returns ~ growth + inflation, data = bz,
+    quantile = c(16, 16))

R> bz_fit2$robdist[bz_fit2$bad_leverage]

 1979  1980
3.651 3.551
```

A.5. De Long and Summers regression

We employ an approximate LTS estimate using one million random samples of size q, setting a random seed for making the result reproducible:

```
R> set.seed(4003)
R> dls_fit <- robreg(gdp ~ lfg + gap + eqp + neq, data = dls,
+    nsamp = 1e6, cutoff = c(3.5, 0))
```

The cutoffs are modified here because it seems that Zaman *et al.* (2001) have only looked at the LTS residuals but not the leverages. For considering both, a scatterplot of LTS residuals vs. MCD distances is useful. The code chunk

```
R> dls_rob <- with(dls_fit, cbind(robdist, robresid))
R> plot(dls_rob, pch = 19, xlab = "Leverage (Robust MCD Distances)",
+    ylab = "Residuals (Robust LTS Residuals)")
```

```
R> abline(h = c(-1, 1) * sqrt(qchisq(0.975, df = 1)))
R> abline(v = sqrt(qchisq(0.975, df = dls_fit$ols$rank - 1)))
R> abline(h = c(-1, 1) * 2.5, lty = 2)
R> dls_out <- abs(dls_rob[,2]) > sqrt(qchisq(0.975, df = 1))
R> text(dls_rob[dls_out,], rownames(dls_rob)[dls_out],
+    pos = c(4, 4, 2, 4, 4, 4))
```

gives Figure 1. The coordinates of the outliers using the cutoff $\sqrt{\chi^2_{1;0.975}}$ (i.e., the coordinates of the points outside the region defined by the solid horizontal lines) are:

```
R> dls_rob[dls_out,]
```

```
         robdist robresid
Cameroon   1.763    2.948
Chile      5.558   -2.369
Jamaica    2.000   -2.334
Spain      5.644    2.356
Zambia     2.198   -5.196
Zimbabwe   2.483   -2.471
```

Affiliation:

Achim Zeileis
Department of Statistics and Mathematics
WU Wirtschaftsuniversität Wien
Augasse 2–6
1090 Wien, Austria
E-mail: Achim.Zeileis@R-project.org
URL: http://statmath.wu.ac.at/~zeileis/

Christian Kleiber
Wirtschaftswissenschaftliches Zentrum (WWZ)
Universität Basel
Peter Merian-Weg 6
4002 Basel, Switzerland
E-mail: Christian.Kleiber@unibas.ch
URL: http://www.wwz.unibas.ch/kleiber/