Approximate Replication of High-Breakdown Robust Regression Techniques

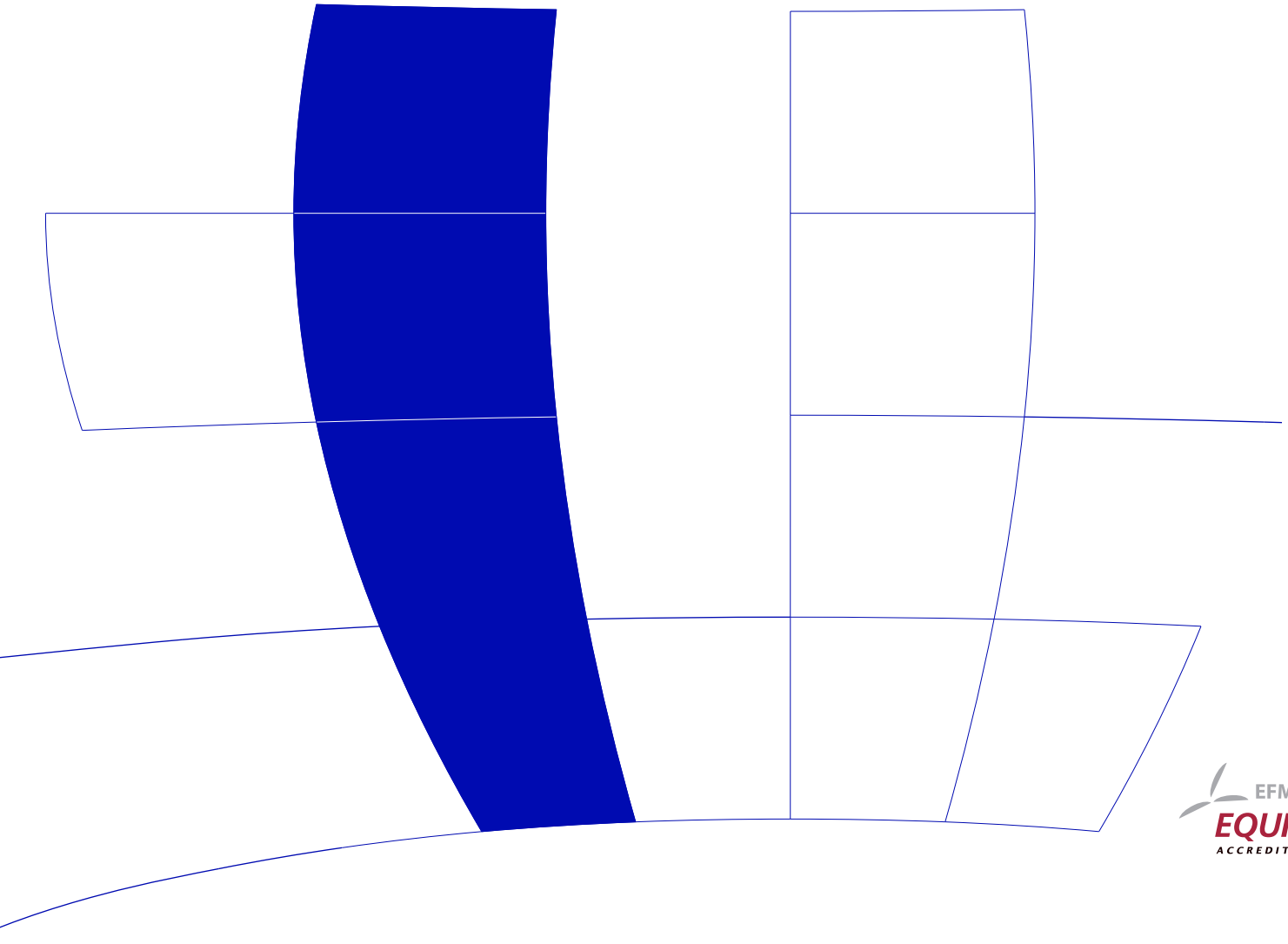
Achim Zeileis, Christian Kleiber

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 68
July 2008

http://statmath.wu-wien.ac.at/



Approximate Replication of High-Breakdown Robust Regression Techniques

Achim Zeileis
Wirtschaftsuniversität Wien

Christian Kleiber
Universität Basel

Abstract

This paper demonstrates that even regression results obtained by techniques close to the standard ordinary least squares (OLS) method can be difficult to replicate if a stochastic model fitting algorithm is employed.

Keywords: robust regression, least squares, replication, 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.

Here 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.

We use the R system for statistical computing (R Development Core Team 2008), version 2.7.1, and the implementation of LTS in the R package **MASS** (Venables and Ripley 2002), version 7.2-42. Both are freely available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`. The exact function calls for replicating our analysis are available as a supplement to this paper (see the appendix).

2. Replication

The approach of Zaman *et al.* (2001) consists of running OLS on a subset of the data. This

subset does not contain bad leverage points 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). See Zaman *et al.* (2001) and the references therein for further details. The final analysis then excludes only those observations with simultaneously large LTS residuals and high leverages from a subsequent OLS regression. Here large typically means larger than 2.5 in absolute size.

LTS minimizes the criterion

$$\sum_{i=1}^{q} r_{i:n}^2$$

where $r_{i:n}$ denotes the ith smallest out of n 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 $\lfloor (n+k+1)/2 \rfloor$ and n is admissible. For the MCD estimator, $q = \lfloor (n+k)/2 \rfloor$ yields maximal robustness.

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 considering large numbers of OLS regressions or sample covariances for random samples of size $p \leq q$ are used, with certain refinements. For the MCD, we use an implementation of the FastMCD algorithm (Rousseeuw and van Driessen 1999) which starts out from random samples of size $p = k$. Note that this does not guarantee that the global minimum is found (even though it is found by this algorithm in at least two of the three applications considered).

2.1. 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 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

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)

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 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 now bad leverage points 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.

2.2. 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 same trimming parameter $q = \lfloor (n + k + 1)/2 \rfloor = \lfloor (28 + 3 + 1)/2 \rfloor = 16$ for the LTS and MCD problems, we are able to exactly reproduce the robust leverages. The LTS

residuals are very close, but not identical, to the values indicated 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!) However, all conclusions drawn from this and, in particular, the resulting OLS regression (omitting the observations for 1979 and 1980) are identical.

In addition, it is worth noting that with these data, there is the only deviation with respect to the original OLS results, in that we obtain a different R^2 and F statistic. We have been unable to identify the source of these discrepancies.

2.3. De Long and Summers regression

We conclude with the 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 $\binom{61}{33} = 191724747789809248$ regressions in total. Hence, we must confine ourselves to an approximate LTS estimator using one million random samples of size q (we tried larger values up to one billion samples, with virtually identical results).

It seems that in this example Zaman *et al.* (2001) have only looked at the LTS residuals but not the leverages. With a value of -5.20, Zambia by far has the largest residual in absolute size but leverage smaller than 2.5, thus suggesting not to exclude this observation according to the strategy followed in the preceding two examples. Furthermore, Cameroon and Zimbabwe also have fairly large residuals but their leverages do not exceed 2.5. In addition, many other countries have large leverages but are associated with smaller LTS residuals.

3. Conclusions

The preceding analysis broadly confirms the analysis of Zaman *et al.* (2001), although the exact numerical results are only reproducible with considerable effort in some cases. 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. Second, they underline that data archives alone are not sufficient to enable validation of published research, only the exact code will enable replicators to fully reproduce earlier results.

References

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

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

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.

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, critval = c(2.5, 2.5),
  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(quantile)) quantile <- c(floor((n + k + 1)/2),
    floor((n + k)/2))
  quantile <- rep(quantile, length.out = 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) > critval[1]

  ## robust leverage via MCD (or MVE)
  X <- model.matrix(fm_ols)[,-1]
  cv <- cov.rob(X, method = method,
    quantile = quantile[2], nsamp = dist_nsamp)
  rd <- sqrt(mahalanobis(X, cv$center, cv$cov))
  rd_nok <- rd > critval[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,
    cov.rob = cv, robresid = rr, robdist = rd,
    high_residuals = rr[rr_nok], high_leverage = rd[rd_nok],
    bad_leverage = nok, psamp = psamp, method = method,
    nsamp = list(lts = nsamp, dist = dist_nsamp),
    quantile = list(lts = quantile[1], dist = quantile[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** (Venables and Ripley 2002). 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 `critval[1]` *and* robust leverages greater than `critval[2]` (both defaulting to 2.5) are then flagged as bad leverage points and excluded in a final OLS regression. The function allows for different trimming parameters `quantile` and different cut-offs `critval` in the LTS and MCD results because this is relevant in some of the examples. 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 Zaman *et al.* (2001):

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

     Canada      Turkey New Zealand
   4.205574   -6.144400   -3.167203

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

```
     Canada      Turkey New Zealand
   7.250783    9.360896    5.976188
```

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 model: there, the value of the objective function is

```
R> nv_fit$cov.rob$crit

[1] -12.64370
```

while with an exhaustive search we obtain

```
R> nv_fit2$cov.rob$crit

[1] -13.20634
```

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):

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

     Canada      Turkey New Zealand
   4.205574   -6.144400   -3.167203

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

     Canada      Turkey New Zealand
  9.0730161  -4.0269118  -0.1778532
```

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
 9.073016  6.236138 -4.026912  4.518340

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

```
   Canada       USA    Turkey Australia
 5.143551  4.502653  7.203009  4.503551
```

A.4. Benderly and Zwick regression

For these data, we are able to exactly reproduce the robust leverages and obtain similar LTS residuals using the code chunk

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

The same observations are flagged as bad leverage points so that the robust regression results are identical:

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

    1979     1980
2.687650 2.678557

R> bz_fit$robdist[bz_fit$bad_leverage]

    1979     1980
3.651306 3.550658
```

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, critval = c(3.5, 0))
```

The critical values are modified here because it seems that Zaman *et al.* (2001) have only looked at the LTS residuals but not the leverages. With these settings we obtain

```
R> dls_fit$robresid[abs(dls_fit$robresid) > 2.5]

 Cameroon    Zambia
 2.948266 -5.196273

R> dls_fit$robdist[abs(dls_fit$robresid) > 2.5]

Cameroon   Zambia
1.762717 2.197666
```

Affiliation:

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

Christian Kleiber
Wirtschaftswissenschaftliches Zentrum
Universität Basel
Petersgraben 51
CH-4051 Basel, Switzerland
E-mail: Christian.Kleiber@unibas.ch
URL: http://www.wwz.unibas.ch/stat/team/kleiber/kleiber.htm