The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 2, pp. 299–307

The S-estimator of multivariate location and scatter in Stata

Vincenzo Verardi
University of Namur (FUNDP)
Center for Research in the Economics of Development (CRED)
Namur, Belgium
and
Université Libre de Bruxelles (ULB)
European Center for Advanced Research in Economics and Statistics (ECARES)
Center for Knowledge Economics (CKE)
Bruxelles, Belgium
vverardi@fundp.ac.be

Alice McCathie
Université Libre de Bruxelles (ULB)
European Center for Advanced Research in Economics and Statistics (ECARES)
Bruxelles, Belgium
amccathi@ulb.ac.be

Abstract. In this article, we introduce a new Stata command, `smultiv`, that implements the S-estimator of multivariate location and scatter. Using simulated data, we show that `smultiv` outperforms `mcd`, an alternative robust estimator. Finally, we use `smultiv` to perform robust principal component analysis and least-squares regression on a real dataset.

Keywords: st0259, smultiv, S-estimator, robustness, outlier, robust principal component analysis, robust regression

1 Introduction

Robust methods for parameter estimation are rapidly becoming essential elements of the statistician's toolbox. A key parameter in most data analyses techniques is the parameter of dispersion, for which Stata currently offers methods of robust estimation. In Verardi and Dehon (2010), the authors propose such a method: the minimum covariance determinant (MCD) estimator. This estimator is based on the notion of generalized variance (Wilks 1932), a one-dimensional assessment of the multivariate spread, measured by the determinant of the sample covariance matrix. The MCD estimator looks for the 50% subsample with the smallest generalized variance. This subsample is assumed free of outliers and thus can be used to compute robust estimates of location and scatter.

Although the MCD estimator is appealing from a theoretical point of view, its practical implementation is problematic. Its estimation requires us to compute the determinant of the sample covariance matrix for all subsamples that contain 50% of the initial

data. Given that this is unfeasible, the code for the MCD estimator uses the p-subset algorithm (Rousseeuw and Leroy 1987) that allows for only a reduced number of subsets to be considered. Unfortunately, this algorithm leads to relatively unstable estimation results, especially for small datasets. Another weakness of the MCD estimator is its low Gaussian efficiency. In light of these shortcomings, we present in this article a Stata command that implements an alternative estimator of location and scatter in a multivariate setting, the S-estimator. The S-estimator is based on the minimization of the sum of a function of the deviations, where the choice of function determines the robustness properties of the estimator.

This article is structured as follows. First, we present the S-estimator in both a univariate and a multivariate setting. We then show that this estimator outperforms MCD in terms of both the stability of the results and its Gaussian efficiency. Finally, we use S-estimates of location and scatter to perform robust principal component analysis (PCA) and robust linear regression.

2 S-estimator of location and scatter

In this section, we introduce the S-estimator of location and scatter. Let us begin by recalling that a location parameter is a measure of the centrality of a distribution. To determine this parameter, one must look for the point in the distribution around which the dispersion of observations is the smallest. In the univariate case, if we choose the variance as a measure of this dispersion, the problem can be formalized as

$$\widehat{\theta} = \arg\min_{\widehat{\theta}} \sigma^2, \quad \text{where} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2,$$
$$\text{which can be rewritten as } 1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \theta}{\sigma}\right)^2$$

This optimization problem consists of the minimization of the sum of the squared distances between each point and the center of the distribution, subject to an equality constraint.

In a multivariate setting, the distance measure to the center of the data cloud is called the Mahalanobis distance and is defined as $\text{MD}_i = \sqrt{(X_i - \theta)\Sigma^{-1}(X_i - \theta)'}$, where X is a matrix of covariates, Σ is the scatter matrix, and θ is the location vector. To obtain a multivariate estimator of location, we minimize a univariate measure of the dispersion of the data cloud—the determinant of the scatter matrix, $\det(\Sigma)$. This is subject to a constraint similar to the one in the univariate case, that equalizes the sum of the squared (Mahalanobis) distances to the number of degrees of freedom, p (recall that the squared Mahalanobis distances follow a χ^2_p). The problem can be formally stated as

$$(\widehat{\theta}, \widehat{\Sigma}) = \arg\min_{\theta,\ \Sigma} \det(\Sigma), \text{ such that } \ p = \frac{1}{n}\sum_{i=1}^{n}\left\{\sqrt{(X_i - \theta)\Sigma^{-1}(X_i - \theta)'}\right\}^2$$

Because the distances enter the above equation squared, it is obvious that observations lying far away from the center of the distribution will have a large influence on its value. Thus in the presence of outlying values, the estimated $\widehat{\theta}$ may not reflect the actual centrality of a large part of the dataset.

The first step toward robustifying this estimator is to consider replacing the square function with an alternative function we call ρ. This function ρ should be nondecreasing in positive values of the argument but less increasing in these values than the square function we wish to replace. Given such a function ρ, the problem becomes

$$(\widehat{\theta}, \widehat{\Sigma}) = \underset{\theta,\ \Sigma}{\arg\min} \det(\Sigma), \text{ such that } b = \frac{1}{n}\sum_{i=1}^{n} \rho\left\{\sqrt{(X_i-\theta)\Sigma^{-1}(X_i-\theta)'}\right\}$$

where $b = E\{\rho(u)\}$[1] to guarantee Gaussian consistency of the estimator. The parameter $\widehat{\theta}$ for which we can find the smallest $\det(\widehat{\Sigma})$ satisfying this equality is called an M-estimator of location, where $\widehat{\Sigma}$ is a multivariate M-estimator of dispersion. If these parameters are estimated simultaneously, they are called S-estimators. From this point on, we will consider only S-estimators because they are highly resistant to outliers for an appropriately chosen function ρ.

The quality of the S-estimator depends on the function ρ. This function should be chosen such that it maximizes both the robustness of the estimator and its Gaussian efficiency. Although there exist a variety of candidate functions for ρ, here we will consider only the Tukey biweight function, which is known to possess good robustness properties. This function is defined as

$$\rho(\mathrm{MD}_i) = \begin{cases} 1 - \left\{1 - \left(\frac{\mathrm{MD}_i}{k}\right)^2\right\}^3 & \text{if } |\mathrm{MD}_i| \le k \\ 1 & \text{if } |\mathrm{MD}_i| > k \end{cases}$$

The breakdown point of the estimator (that is, the maximum contamination the estimator can withstand) is determined by the tuning constant k. One first selects a cutoff value for a constant c on the univariate scale as the number of standard deviations from the mean beyond which $\rho'(\mathrm{MD}_i)$ becomes zero (Campbell, Lopuhaä, and Rousseeuw 1998). This constant c can then be converted to a value of k on a chi-squared scale of MD^2 using the Wilson–Hilferty (1931) transformation (Campbell 1984):

$$k = \sqrt{p\left[\sqrt{\left(\frac{1}{9}\right)\left(\frac{2}{p}\right)}c + \left\{1 - \left(\frac{1}{9}\right)\left(\frac{2}{p}\right)\right\}\right]^3}$$

The S-estimator has a breakdown point of 50% when $c = 1.548$, which for $p = 3$ implies $k = 2.707$. Figure 1 depicts the Tukey biweight function ρ and its derivative ρ'.

1. Under the assumption that u is a standard normal distribution.

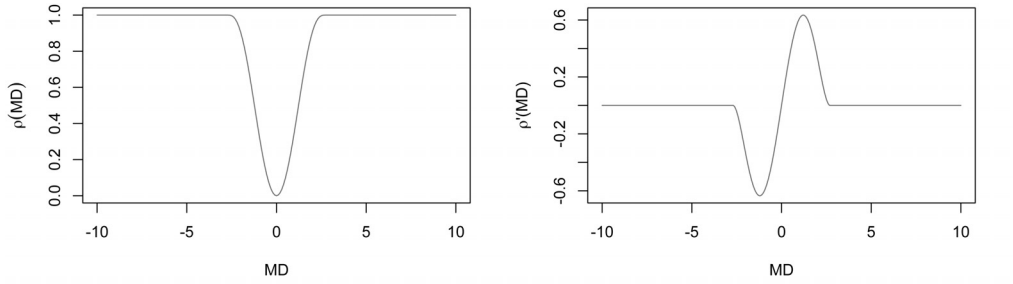

Figure 1. Tukey biweight

The Gaussian consistency parameter ($E\{\rho(u)\}$ if we assume a p-variate standard normal distribution) is

$$b = \frac{p}{2}\chi^2_{p+2}(k^2) - \frac{p(p+2)}{2k^2}\chi^2_{p+4}(k^2) + \frac{p(p+2)(p+4)}{6k^4}\chi^2_{p+6}(k^2) + \frac{k^2}{6}\left\{1 - \chi^2_p(k^2)\right\}$$

3 Relative stability of MCD and S-estimation

Both MCD and S-estimators give robust estimates of multivariate location and scatter. However, implementation of MCD is more problematic because the fast MCD algorithm[2] that performs the estimation yields results that are both unstable over multiple replications and sensitive to the chosen starting value. This is not a weakness of the S-estimator, whose estimation procedure is based on an iterative reweighting algorithm; see Campbell, Lopuhaä, and Rousseeuw (1998).

To illustrate the comparative performances of both estimators, we create a dataset of 1,000 observations with values drawn from three independent $N(0,1)$ random variables and replace 10% of the x_1 values by an arbitrarily large number (100 in this example). We then compute the MCD and S-estimator of scatter of the data and repeat the procedure 1,000 times. We report in table 1 the estimates obtained for the elements of the main diagonal of the scatter matrix. The stability of the estimates refers to the proportion of appearances made by the most frequently observed value for each estimate.

2. Developed by Rousseeuw and van Driessen (1999) and implemented in Stata by Verardi and Croux (2009, 2010).

Table 1. MCD and S-estimator of scale

	S-estimator			MCD-estimator		
	$\widehat{\sigma}_1^2$	$\widehat{\sigma}_2^2$	$\widehat{\sigma}_3^2$	$\widehat{\sigma}_1^2$	$\widehat{\sigma}_2^2$	$\widehat{\sigma}_3^2$
minimum	1.049	1.171	1.118	0.761	0.874	0.815
maximum	1.051	1.171	1.120	0.972	1.016	1.050
stability	99.6%	99.6%	99.6%	0%	0%	0%

We see here that the range of values observed is much smaller with the S-estimator than with MCD. In fact, the difference between the minimum and maximum S-estimates is virtually zero (between 0 and 0.002), whereas with MCD this difference is small but not negligible (between 0.142 and 0.235). Perhaps the most striking result is that the most frequently observed S-estimate occurred in 99.6% of the trials. This is in sharp contrast with the MCD estimates, none of which reappeared more than once.

4 The multivariate S-estimator of location and scatter in other statistical applications

Many statistical applications require the prior estimation of a scatter matrix. In this section, we demonstrate how to use the S-estimator of location and scatter to 1) identify outliers in a dataset, 2) obtain robust estimates of linear regression parameters, and 3) perform robust PCA. Again we create a dataset of 1,000 observations, generated here from the following data-generating process: $y = x_1 + x_2 + x_3 + \varepsilon$. We then contaminate the data by replacing 10% of the observations in x_1 by draws from a $N(100, 1)$. In Stata language, this is done as follows:

```
set seed 123
set obs 1000
matrix C0=(1,0.3,0.2 \ 0.3,1,0.4 \ 0.2,0.4,1)
drawnorm x1-x3, corr(C0)
gen e=invnormal(runiform())
gen y=x1+x2+x3+e
gen bx1=x1
replace x1=invnormal(runiform())+100 in 1/100
```

To identify the outliers present in our dataset, we need to estimate the robust Mahalanobis distance for each observation in the sample. Computing these distances requires robust estimates of both the location and the scale parameters of our sample. We thus run the `smultiv` command to obtain S-estimators of these parameters and request that the outliers be flagged and the robust Mahalanobis distances be reported. This is done in Stata by typing

```
smultiv x*, generate(robust_outlier robust_distance)
summarize robust_distance in 101/l
summarize robust_distance in 1/100
```

Observations are identified as outliers if their robust Mahalanobis distance is greater than the critical value at 95% of a $\sqrt{\chi^2_p}$. (We know that if our data are Gaussian, the Mahalanobis distances will follow a $\sqrt{\chi^2_p}$ distribution, where p is equal to the number of variables in our dataset.) Our results show that the robust distances of the outliers range from 96.82 to 102.85 with an average of 99.58. These values are consistent with our contamination scheme. The average distance for observations in the noncontaminated dataset is 1.493, which is significantly smaller than the critical value of $\sqrt{\chi^2_{3,0.95}} = 2.80$.

Next we look at PCA. There are two different methods for performing robust PCA. One can run the `pca` command on a cleaned-up version of the initial sample after removing the outlying values identified. Alternatively, one can use the `pcamat` command in Stata to extract the eigenvalues and eigenvectors of the robust correlation matrix estimated. In our study, we perform four PCA. The first `pca` command is run on the noncontaminated sample and serves as our benchmark (PCA_clean). We then perform the `pca` command on the contaminated sample (PCA). Next we run the `pca` command on the cleaned-up sample obtained by removing the outliers identified previously (PCA_cleaned). Finally, we run the `pcamat` command on the robust correlation matrix (PCA_mat). These commands are typed in Stata as follows:

```
pca bx1 x2 x3
pca x*
pca x* if robust_outlier==0
matrix C=e(C)
pcamat C, n(1000)
```

For each PCA performed, we report in table 2 the first principal component obtained, its corresponding eigenvalue, and the proportion of variance it explains. We see clearly that both robust PCA methods yield results similar to those obtained from the standard PCA performed on the uncontaminated sample, whereas the classical PCA run on the contaminated dataset leads to uninformative results. A similar robustification of the factor model can be performed either by downweighting the outliers or by calling on the `factor` command.

Table 2. Principal component analysis

	PCA_clean	PCA	PCA_cleaned	PCA_mat
x1	0.5339	−0.0225	0.5243	0.5243
x2	0.6313	0.7078	0.6331	0.6331
x3	0.5626	0.7060	0.5695	0.5695
eigenvalue	1.63	1.38	1.62	1.62
	(54%)	(46%)	(54%)	(54%)

S-estimation can also be used to perform robust linear regression analysis. Consider the following regression model: $y = \alpha + X\beta + \varepsilon$, where y is the dependent variable,

α is a constant, X is the matrix of covariates, and ε is the error term vector. Least-squares estimation gives us parameter estimates $\widehat{\beta} = \left(\widehat{\Sigma}_{xx}\right)^{-1} \widehat{\Sigma}_{xy} = (X'X)^{-1} X'Y$ and $\widehat{\alpha} = \widehat{\mu}_y - \widehat{\mu}_x\widehat{\beta}$ (where $\widehat{\mu}_x$ is a vector and $\widehat{\mu}_y$ is a scalar). To obtain robust estimators for β and α, one can replace μ and Σ in the previous expressions by their robust S-estimation counterparts $\widehat{\mu}^S$ and $\widehat{\Sigma}^S$. Robust estimates of the regression parameters are thus $\widehat{\beta} = \left(\widehat{\Sigma}^S_{xx}\right)^{-1} \widehat{\Sigma}^S_{xy}$ and $\widehat{\alpha} = \widehat{\mu}^S_y - \widehat{\mu}^S_x\widehat{\beta}$.

We conclude this section by performing regression analysis on a real dataset that contains information about 120 U.S. universities.[3] The dependent variable is `score`, a weighted average of five of the six factors used by Shanghai University to compute their *Academic Ranking of World Universities*. We consider five explanatory variables: `cost` (logarithm of the out-of-state cost per year), `enrol` (proportion of admitted students that choose to enroll), `math` (SAT math score first quartile), `undergrad` (logarithm of the total undergraduate population), and `private` (a dummy variable that indicates whether a university is private).

We begin by running a basic ordinary least-squares (OLS) regression on the complete sample. We then perform S-estimation using the command `smultiv` and flag the outliers. We run an OLS regression on the subsample obtained by removing the outliers identified by `smultiv`. Finally, we repeat this two-step procedure with the `mcd` command (Verardi and Croux 2010) to identify the outliers in our sample. This is done in Stata by typing

```
use http://homepages.ulb.ac.be/~vverardi/unidata, clear

/* Classic OLS */
regress score cost enrol math undergrad private

/* Removal of outliers identified through S-estimation */
smultiv score cost enrol math undergrad, ///
     generate(robust_outlier robust_distance) dummies(private)
regress score cost enrol math undergrad private if robust_outlier==0

/* Removal of outliers identified through MCD estimation */
mcd score cost enrol math undergrad, generate(mcd_outlier mcd_distance)
regress score cost enrol math undergrad private if mcd_outlier==0
```

The results in table 3 show that although the OLS regression performed on the whole sample identifies all five explanatory variables as statistically significant, both robust regressions reveal that this is in fact not the case for the variable `enrol`.

3. Data for 2007, available at http://www.usuniversities.ca/.

Table 3. Linear regression analysis

	cost	enrol	math	undergrad	private	_cons
regress	27.690*	4.484*	84.321*	3.731*	−15.469*	−845.546*
	(5.96)	(0.57)	(12.55)	(1.87)	(3.61)	(78.85)
smultiv	25.080*	−0.506	65.753*	9.659*	−8.802*	−747.318*
	(6.38)	(1.08)	(12.11)	(1.93)	(3.95)	(68.41)
mcd	20.246*	−0.620	72.041*	7.446*	−11.986*	−714.984*
	(6.26)	(1.02)	(11.77)	(1.64)	(3.16)	(69.97)

5 The smultiv command

5.1 Syntax

The general syntax for the command is

`smultiv` *varlist* [*if*] [*in*] [`, generate(`*varname1 varname2*`) dummies(`*varlist*`) nreps(`#`)`]

5.2 Options

`generate(`*varname1 varname2*`)` creates two new variables: *varname1*, a dummy variable that flags the outliers, and *varname2*, a continuous variable that reports the robust Mahalanobis distances. The user must specify the name for each of the variables generated. These variables cannot be generated separately.

`dummies(`*varlist*`)` specifies the dummy variables in the dataset.

`nreps(`#`)` specifies the number of replications.

5.3 Saved results

`smultiv` saves the following in `e()`:

Matrices

`e(mu)`	location vector
`e(S)`	scatter matrix
`e(C)`	correlation matrix

6 Conclusion

In this article, we proposed a new Stata command to implement the S-estimator of location and scatter in a multivariate setting. Our simulated data examples show that this estimator yields more stable results than its closest competitor, the MCD estimator. Furthermore, the estimates obtained can be used to perform other statistical data analysis techniques, such as PCA and linear regression, and obtain robust results.

7 Acknowledgment

Vincenzo Verardi is an associate researcher at the FNRS and gratefully acknowledges their financial support.

8 References

Campbell, N. A. 1984. Mixture models and atypical values. *Mathematical Geology* 16: 465–477.

Campbell, N. A., H. P. Lopuhaä, and P. J. Rousseeuw. 1998. On the calculation of a robust S-estimator of a covariance matrix. *Statistics in Medicine* 17: 2685–2695.

Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. J., and K. van Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41: 212–223.

Verardi, V., and C. Croux. 2009. Robust regression in Stata. *Stata Journal* 9: 439–453.

———. 2010. Software update: st0173_1: Robust regression in Stata. *Stata Journal* 10: 313.

Verardi, V., and C. Dehon. 2010. Multivariate outlier detection in Stata. *Stata Journal* 10: 259–266.

Wilks, S. S. 1932. Certain generalizations in the analysis of variance. *Biometrika* 24: 471–494.

Wilson, E. B., and M. M. Hilferty. 1931. The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America* 17: 684–688.

About the authors

Vincenzo Verardi is a research fellow of the Belgian National Science Foundation (FNRS). He is a professor at the Faculté Notre Dame de la Paix of Namur and at the Université Libre de Bruxelles. His research interests include applied econometrics and development economics.

Alice McCathie is a PhD candidate in economics at the Université Libre de Bruxelles. Her main research interests are the economics of higher education and discrete choice modeling.