



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

THE STATA JOURNAL

Editor

H. Joseph Newton
Department of Statistics
Texas A&M University
College Station, Texas 77843
979-845-8817; fax 979-845-6077
jnewton@stata-journal.com

Editor

Nicholas J. Cox
Department of Geography
Durham University
South Road
Durham DH1 3LE UK
n.j.cox@stata-journal.com

Associate Editors

Christopher F. Baum
Boston College

Nathaniel Beck
New York University

Rino Bellocco
Karolinska Institutet, Sweden, and
University of Milano-Bicocca, Italy

Maarten L. Buis
Tübingen University, Germany

A. Colin Cameron
University of California–Davis

Mario A. Cleves
Univ. of Arkansas for Medical Sciences

William D. Dupont
Vanderbilt University

David Epstein
Columbia University

Allan Gregory
Queen's University

James Hardin
University of South Carolina

Ben Jann
University of Bern, Switzerland

Stephen Jenkins
London School of Economics and
Political Science

Ulrich Kohler
WZB, Berlin

Frauke Kreuter
University of Maryland–College Park

Peter A. Lachenbruch
Oregon State University

Jens Lauritsen
Odense University Hospital

Stanley Lemeshow
Ohio State University

J. Scott Long
Indiana University

Roger Newson
Imperial College, London

Austin Nichols
Urban Institute, Washington DC

Marcello Pagano
Harvard School of Public Health

Sophia Rabe-Hesketh
University of California–Berkeley

J. Patrick Royston
MRC Clinical Trials Unit, London

Philip Ryan
University of Adelaide

Mark E. Schaffer
Heriot-Watt University, Edinburgh

Jeroen Weesie
Utrecht University

Nicholas J. G. Winter
University of Virginia

Jeffrey Wooldridge
Michigan State University

Stata Press Editorial Manager
Stata Press Copy Editor

Lisa Gilmore
Deirdre Skaggs

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®

Copyright Statement: The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

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.

A robust instrumental-variables estimator

Rodolphe Desbordes
University of Strathclyde
Glasgow, UK
rodolphe.desbordes@strath.ac.uk

Vincenzo Verardi
University of Namur
(Centre for Research in the Economics of Development)
and Université Libre de Bruxelles
(European Center for Advanced Research in Economics and Statistics
and Center for Knowledge Economics)
Namur, Belgium
vverardi@ulb.ac.be

Abstract. The classical instrumental-variables estimator is extremely sensitive to the presence of outliers in the sample. This is a concern because outliers can strongly distort the estimated effect of a given regressor on the dependent variable. Although outlier diagnostics exist, they frequently fail to detect atypical observations because they are themselves based on nonrobust (to outliers) estimators. Furthermore, they do not take into account the combined influence of outliers in the first and second stages of the instrumental-variables estimator. In this article, we present a robust instrumental-variables estimator, initially proposed by [Cohen Freue, Ortiz-Molina, and Zamar \(2011, Working paper: <http://www.stat.ubc.ca/~ruben/website/cv/cohen-zamar.pdf>\)](http://www.stat.ubc.ca/~ruben/website/cv/cohen-zamar.pdf), that we have programmed in Stata and made available via the `robivreg` command. We have improved on their estimator in two different ways. First, we use a weighting scheme that makes our estimator more efficient and allows the computations of the usual identification and overidentifying restrictions tests. Second, we implement a generalized Hausman test for the presence of outliers.

Keywords: st0252, robivreg, multivariate outliers, robustness, S-estimator, instrumental variables

1 Theory

Assume a linear regression model given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{1}$$

where \mathbf{y} is the $n \times 1$ vector containing the value of the dependent variable, \mathbf{X} is the $n \times p$ matrix containing the values for the p regressors (constant included), and $\boldsymbol{\varepsilon}$ is the vector of the error term. Vector $\boldsymbol{\theta}$ of size $p \times 1$ contains the unknown regression parameters and needs to be estimated. On the basis of the estimated parameter $\hat{\boldsymbol{\theta}}$, it is then possible to fit the dependent variable by $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$ and estimate the residual vector

$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$. In the case of the ordinary least-squares (LS) method, the vector of estimated parameters is

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \arg \min_{\boldsymbol{\theta}} \mathbf{r}' \mathbf{r}$$

The solution to this minimization leads to the well-known formula

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \underbrace{(\mathbf{X}^t \mathbf{X})^{-1}}_{n \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{X}}} \underbrace{\mathbf{X}^t \mathbf{y}}_{n \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}}}$$

which is simply, after centering the data, the product of the $p \times p$ covariance matrix of the explanatory variables $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}}$ and the $p \times 1$ vector of the covariances of the explanatory variables and the dependent variable $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$ (the n simplify).¹

The unbiasedness and consistency of the LS estimates crucially depend on the absence of correlation between \mathbf{X} and $\boldsymbol{\varepsilon}$. When this assumption is violated, instrumental-variables (IV) estimators are generally used. The logic underlying this approach is to find some variables, known as instruments, that are strongly correlated with the troublesome explanatory variables, known as endogenous variables, but independent of the error term. This is equivalent to estimating the relationship between the response variable and the covariates by using only the part of the variability of the endogenous covariates that is uncorrelated with the error term.

More precisely, define \mathbf{Z} as the $n \times m$ matrix (where $m \geq p$) containing the instruments. The IV estimator (generally called two-stage least squares when $m > p$) can be conceptualized as a two-stage estimator. In the first stage, each endogenous variable is regressed on the instruments and on the variables in \mathbf{X} that are not correlated with the error term. In the second stage, the predicted value for each variable is then fit (denoted $\hat{\mathbf{X}}$ here). In this way, each variable is purged of the correlation with the error term. Exogenous explanatory variables are used as their own instruments. These new variables are then replaced in (1), and the model is fit by LS.

The final estimator is (again centering the data and recalculating the intercept term)

$$\hat{\boldsymbol{\theta}}_{\text{IV}} = \left\{ \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Z}} \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{Z}} \right)^{-1} \hat{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{X}} \right\}^{-1} \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Z}} \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{Z}} \right)^{-1} \hat{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{y}} \quad (2)$$

where $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Z}}$ is the covariance matrix of the original right-hand-side variables and the instruments, $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{Z}}$ is the covariance matrix of the instruments, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{y}}$ is the vector of covariances of the instruments with the dependent variable.

A drawback of the IV method is that if outliers are present, all the estimated covariances are distorted, even asymptotically. [Cohen Freue, Ortiz-Molina, and Zamar \(2011\)](#) therefore suggest replacing classical estimated covariance matrices in (2) with some robust counterparts that withstand the contamination. These could be minimum covariance determinant scatter matrices as presented in [Verardi and Dehon \(2010\)](#) or S-estimators of location and scatter as described by [Verardi and McCathie \(2012\)](#). We use the latter, and the superscript ^S is used to indicate it.

1. The constant term has to be recalculated.

The robust IV estimator can therefore be written as

$$\hat{\theta}_{\text{RIV}}^S = \left\{ \hat{\Sigma}_{\text{XZ}}^S \left(\hat{\Sigma}_{\text{ZZ}}^S \right)^{-1} \hat{\Sigma}_{\text{ZX}}^S \right\}^{-1} \hat{\Sigma}_{\text{XZ}}^S \left(\hat{\Sigma}_{\text{ZZ}}^S \right)^{-1} \hat{\Sigma}_{\text{ZY}}^S$$

As shown by [Cohen Freue, Ortiz-Molina, and Zamar \(2011\)](#), this estimator inherits the consistency properties of the underlying multivariate S-estimator and remains consistent even when the distribution of the carriers is not elliptical or symmetrical. They also demonstrate that under certain regularity conditions, this estimator is asymptotically normal, regression and carrier equivariant. Finally, they provide a simple formula for its asymptotic variance.

An alternative estimator that would allow a substantial gain in efficiency is

$$\hat{\theta}_{\text{RIV}}^W = \left\{ \hat{\Sigma}_{\text{XZ}}^W \left(\hat{\Sigma}_{\text{ZZ}}^W \right)^{-1} \hat{\Sigma}_{\text{ZX}}^W \right\}^{-1} \hat{\Sigma}_{\text{XZ}}^W \left(\hat{\Sigma}_{\text{ZZ}}^W \right)^{-1} \hat{\Sigma}_{\text{ZY}}^W$$

where W stands for weights. First estimated are the robust covariance $\hat{\Sigma}_{\text{XZY}}^S$ and the robust Mahalanobis distances—that is, $\hat{d}_i = \sqrt{(\mathbf{M}_i - \hat{\boldsymbol{\mu}}_{\mathbf{M}})' \hat{\Sigma}_{\mathbf{M}}^{-1} (\mathbf{M}_i - \hat{\boldsymbol{\mu}}_{\mathbf{M}})}$, where $\mathbf{M} = (\mathbf{X}, \mathbf{Z}, \mathbf{y})$, $\hat{\Sigma}_{\mathbf{M}} = \hat{\Sigma}_{\text{XZY}}^S$ is the scatter matrix of explanatory variables, and $\hat{\boldsymbol{\mu}}_{\mathbf{M}}$ is the location vector. Outliers are then identified as the observations that have a robust Mahalanobis distance \hat{d}_i larger than $\sqrt{\chi_{p+m+1,q}^2}$, where q is a confidence level (for example, 99%), given that Mahalanobis distances are distributed as the square root of a chi-squared with degrees of freedom equal to the length of vector $\hat{\boldsymbol{\mu}}_{\mathbf{M}}$. Finally, observations that are associated with a \hat{d}_i larger than the cutoff point are downweighted, and the classical covariance matrix is estimated. The weighting that we adopt is simply to award a weight of 1 to observations associated with a \hat{d}_i smaller than the cutoff value and to award a weight of 0 otherwise.

The advantage of this last estimator is that standard overidentification, underidentification, and weak instruments tests can easily be obtained, because this weighting scheme amounts to running a standard IV estimation on a sample free of outliers and the asymptotic variance of the estimator is also readily available. We use the user-written `ivreg2` command ([Baum, Schaffer, and Stillman 2007](#)) to compute the final estimates; the reported tests and standard errors are those provided by this command.² Finally, a substantial gain in efficiency with respect to the standard robust IV estimator proposed by [Cohen Freue, Ortiz-Molina, and Zamar \(2011\)](#) can be attained. We illustrate this efficiency gain by running 1,000 simulations using a setup similar to that of [Cohen Freue, Ortiz-Molina, and Zamar \(2011\)](#) but with no outliers: 1,000 observations for five random variables (x, u, v, w, Z) drawn from a multivariate normal distribution with mean $\boldsymbol{\mu} = (0, 0, 0, 0, 0)$ and covariance

2. The `robivreg` command is not a full wrapper for the `ivreg2` command. However, a sample free of outliers can easily be obtained by using the `generate(varname)` option that we describe in the next section.

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 0.3 & 0.2 & 0 & 0 \\ 0 & 0.2 & 0.3 & 0 & 0 \\ 0.5 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The data-generating process is $Y = 1 + 2x + Z + u$, where x is measured with error and only variable $X = x + v$ is assumed observable. To remedy this endogeneity bias, X is instrumented by Z . For this setup, the simulated efficiency of the two estimators is 46.7% for the raw θ_{RIV}^S estimator and 95.5% for the reweighted θ_{RIV}^W estimator. The efficiency is calculated as follows: Assume θ_{RIV}^S (θ_{RIV}^W) is asymptotically normal with covariance matrix V , and assume V_0 is the asymptotic covariance matrix of the classical θ_{IV} estimator. The efficiency of θ_{RIV}^S (θ_{RIV}^W) is calculated as $\text{eff}(\theta_{\text{RIV}}^S) = \lambda_1(V^{-1}V_0)$, where $\lambda_1(E)$ denotes the largest eigenvalue of the matrix E , and V_0 and V are the simulated covariances.

2 The robivreg command

2.1 Syntax

The **robivreg** command implements an IV estimator robust to outliers.

```
robivreg depvar [varlist1] (varlist2 = instlist) [if] [in] [, first robust
      cluster(varname) generate(varname) raw cutoff(#) mcd graph
      label(varname) test nreps(#) nodots]
```

where *depvar* is the dependent variable, *varlist1* contains the exogenous regressors, *varlist2* contains the endogenous regressors, and *instlist* contains the excluded instruments.

2.2 Options

first reports various first-stage results and identification statistics. May not be used with **raw**.

robust produces standard errors and statistics that are robust to arbitrary heteroskedasticity.

cluster(varname) produces standard errors and statistics that are robust to both arbitrary heteroskedasticity and intragroup correlation, where *varname* identifies the group.

`generate(varname)` generates a dummy variable named *varname*, which takes the value of 1 for observations that are flagged as outliers.

`raw` specifies that [Cohen Freue, Ortiz-Molina, and Zamar's estimator \(2011\)](#) should be returned. Note that the standard errors reported are different from the ones that they proposed because these are robust to heteroskedasticity and asymmetry. The asymptotic variance of the `raw` estimator is described in [Verardi and Croux \(2009\)](#).

`cutoff(#)` allows the user to change the percentile above which an individual is considered to be an outlier. The default is `cutoff(0.99)`.

`mcd` specifies that a minimum covariance determinant estimator of location and scatter be used to estimate the robust covariance matrices. By default, an S-estimator of location and scatter is used.

`graph` generates a graphic in which outliers are identified according to their type, and labeled using the variable *varname*. Vertical lines identify vertical outliers (observations with a large residual), and the horizontal line identifies leverage points.

`label(varname)` labels the outliers as *varname*. `label()` only has an effect if specified with `graph`.

`test` specifies to report a test for the presence of outliers in the sample. To test for the appropriateness of a robust IV procedure relative to the classical IV estimator, we rely on the *W* statistic proposed by [Dehon, Gassner, and Verardi \(2009\)](#) and [Desbordes and Verardi \(2011\)](#), where

$$W = \left(\hat{\theta}^{\text{IV}} - \hat{\theta}_{\text{RIV}}^{\text{S}} \right)^t \left\{ \widehat{\text{Var}} \left(\hat{\theta}^{\text{IV}} \right) + \widehat{\text{Var}} \left(\hat{\theta}_{\text{RIV}}^{\text{S}} \right) - 2\widehat{\text{Cov}} \left(\hat{\theta}^{\text{IV}}, \hat{\theta}_{\text{RIV}}^{\text{S}} \right) \right\}^{-1} \left(\hat{\theta}^{\text{IV}} - \hat{\theta}_{\text{RIV}}^{\text{S}} \right)$$

Bearing in mind that this statistic is asymptotically distributed as a χ_p^2 , where *p* is the number of covariates, it is possible to set an upper bound above which the estimated parameters can be considered to be statistically different and hence the robust IV estimator should be preferred to the standard IV estimator. When the `cluster()` option is specified, a cluster-bootstrap is used to calculate the *W* statistic.

`nreps(#)` specifies the number of bootstrap replicates performed when the `test` and `cluster()` options are both specified. The default is `nreps(50)`.

`nodots` suppresses the replication dots.

3 Empirical example

In a seminal article, [Romer \(1993\)](#) convincingly shows that more open economies tend to have lower inflation rates. Worried that a simultaneity bias may affect the estimates, he instruments the trade openness variable—the share of imports in gross domestic product—by the logarithm of a country's land area.

From a pedagogical perspective, it is useful to start with the dependent variable (which is the average annual inflation rates since 1973), in levels, as in example 16.6 of Wooldridge (2009, 558).

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/OPENNESS
. merge 1:1 _n using http://www.rodolphedesbordes.com/web_documents/names.dta
```

Result	# of obs.
not matched	0
matched	114

```
( _merge==3)

. ivregress 2sls inf (opendec = lland)
Instrumental variables (2SLS) regression
```

Number of obs =	114
Wald chi2(1) =	5.73
Prob > chi2 =	0.0167
R-squared =	0.0316
Root MSE =	23.511

inf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
opendec	-33.28737	13.91101	-2.39	0.017	-60.55245	-6.022284
_cons	29.60664	5.608412	5.28	0.000	18.61435	40.59893

```
Instrumented: opendec
Instruments: lland
```

The coefficient on **opendec** is significant at the 5% level and suggests that a country with a 50% import share had an average inflation rate about 8.3 percentage points lower than a country with a 25% import share.

We may be worried that outliers distort these estimates. For instance, it is well known that countries in Latin America have experienced extremely high inflation rates in the 1980s. Hence, we refit the model with the **robivreg** command.

```

. robivreg inf (opendec = lland), test graph label(countryname)
(sum of wgt is      8.3000e+01)
IV (2SLS) estimation

```

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

		Number of obs =	83
		F(1, 81) =	2.50
		Prob > F =	0.1181
Total (centered) SS	=	Centered R2 =	0.0514
Total (uncentered) SS	=	Uncentered R2 =	0.8844
Residual SS	=	Root MSE =	3.935

inf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
opendec	-12.07379	7.642656	-1.58	0.118	-27.28027 3.1327
_cons	14.60224	2.500814	5.84	0.000	9.626404 19.57808

Underidentification test (Anderson canon. corr. LM statistic): 16.073
Chi-sq(1) P-val = 0.0001

Weak identification test (Cragg-Donald Wald F statistic): 19.453
Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38
15% maximal IV size 8.96
20% maximal IV size 6.66
25% maximal IV size 5.53

Source: Stock-Yogo (2005). Reproduced by permission.

Sargan statistic (overidentification test of all instruments): 0.000
(equation exactly identified)

Instrumented: opendec
Excluded instruments: lland

H0: Outliers do not distort 2SLS classical estimation

chi2(2)=10.53
Prob > chi2 = .005

Once the influence of outliers is downweighted, the value of the coefficient on `opendec` becomes much smaller and loses statistical significance. Our test for outliers, requested using the option `test`, confirms that outliers distort enough the original estimates such that robustness should be favored at the expense of efficiency.

The outliers can be easily identified using the `graph` option. We facilitate the identification of each type of outlier by setting vertical and horizontal cutoff points in the reported graph. The vertical cutoff points are 2.25 and -2.25. If the residuals were normally distributed, values above or below these cutoff points would be strongly atypical because they would be 2.25 standard deviations away from the mean (which is 0 by construction), with a probability of occurrence of 0.025. The reported residuals are said to be robust and standardized because the residuals are based on a robust-to-outliers estimation and have been divided by the standard deviation of the residuals associated with nonoutlying observations. In line with our downweighting scheme, the horizontal

cutoff point is, by default, $\sqrt{\chi^2_{p+m+1,0.99}}$. Vertical outliers are observations above or below the vertical lines, while leverage points are to the right of the horizontal line.

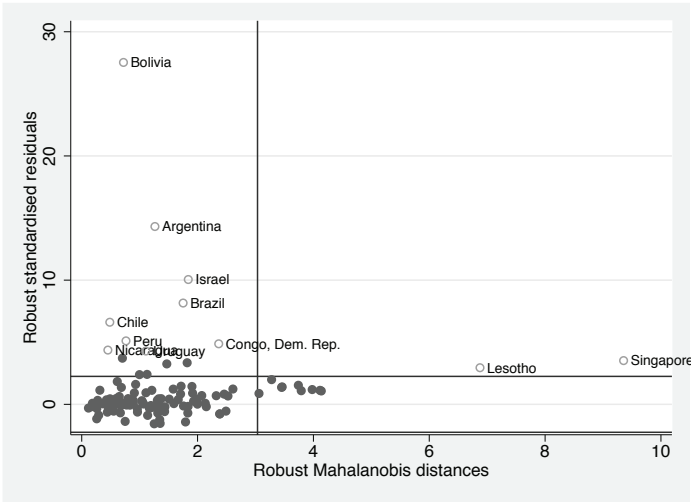


Figure 1. Identification of outliers when `inf` is used

Romer (1993) was fully aware that his results could be sensitive to outliers. This is why he decided to use as a dependent variable the log of average inflation.

```
. ivregress 2sls linf (opendec = lland)
Instrumental variables (2SLS) regression

Number of obs =      114
Wald chi2(1)    =    11.06
Prob > chi2     =    0.0009
R-squared       =    0.1028
Root MSE      =    .66881
```

linf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
opendec	-1.315804	.3957235	-3.33	0.001	-2.091408	-.5401999
_cons	2.98983	.1595413	18.74	0.000	2.677135	3.302525

```
Instrumented:  opendec
Instruments:   lland
```

The coefficient on `opendec` is now significant at the 1% level and suggests, using the Duan smearing estimate, that a country with a 50% import share had an average inflation rate about 5.4 percentage points lower than a country with a 25% import share.

However, even though taking the log of average inflation has certainly reduced the influence of extreme values of the dependent variable, outliers may still be an issue. Hence, we refit the model again with the `robivreg` command.

```
. robivreg linf (opendec = lland), test graph label(countryname)
(sum of wgt is      8.9000e+01)
IV (2SLS) estimation
```

Estimates efficient for homoskedasticity only
 Statistics consistent for homoskedasticity only

		Number of obs =	89
		F(1, 87) =	3.03
		Prob > F =	0.0851
Total (centered) SS	=	18.9763507	
Total (uncentered) SS	=	533.2560195	
Residual SS	=	16.17036311	
		Centered R2 =	0.1479
		Uncentered R2 =	0.9697
		Root MSE =	.4311

linf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
opendec	-1.225348	.7034456	-1.74	0.085	-2.623523	.1728262
_cons	2.796497	.2300043	12.16	0.000	2.339339	3.253656

Underidentification test (Anderson canon. corr. LM statistic): 20.963
 Chi-sq(1) P-val = 0.0000

Weak identification test (Cragg-Donald Wald F statistic): 26.806
 Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38
 15% maximal IV size 8.96
 20% maximal IV size 6.66
 25% maximal IV size 5.53

Source: Stock-Yogo (2005). Reproduced by permission.

Sargan statistic (overidentification test of all instruments): 0.000
 (equation exactly identified)

Instrumented: opendec
 Excluded instruments: lland

H0: Outliers do not distort 2SLS classical estimation

chi2(2)=4.86
 Prob > chi2 = .088

In that case, the magnitude of the coefficient is preserved, but its statistical significance sharply decreases. Once again, we can identify outliers by using the `graph` option.

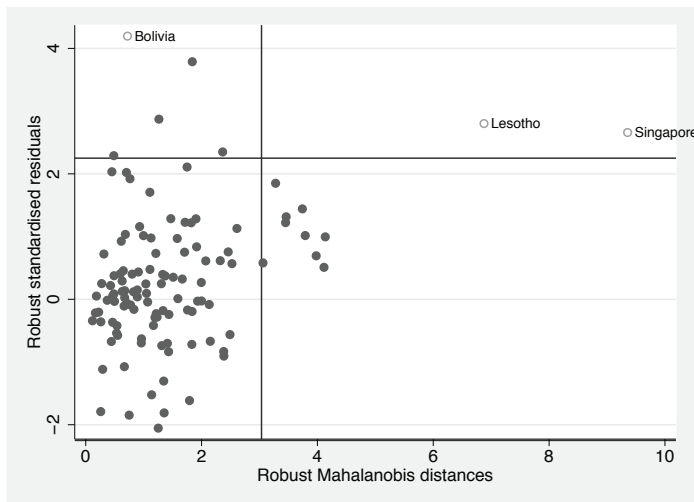


Figure 2. Identification of outliers when `ln(inf)` is used

In figure 1, we can see that taking the log of `inf` has been insufficient to deal with all outliers in the dependent variable because **Bolivia** remains an outlier. Furthermore, Romer was right to be worried that **Lesotho** or **Singapore** may “have an excessive influence on the results” [Romer \(1993, 877\)](#). The remoteness of these two observations from the rest of the data led to an inflation of the total sample variation in trade openness, resulting in undersized standard errors and spuriously high statistical significance.

For the final example, we illustrate the use of the `test` option with the `cluster()` option. The clustering variable is `idcode`.

```

. webuse nlswork, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
. keep if _n<1501
(27034 observations deleted)
. robivreg ln_w age not_smsa (tenure = union south), cluster(idcode) test
(sum of wgt is      8.4700e+02)
IV (2SLS) estimation
-----

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on idcode

Number of clusters (idcode) =      209      Number of obs =      847
                                F( 3, 208) =      5.16
                                Prob > F      =      0.0018
Total (centered) SS      = 126.4871762      Centered R2      = -0.0831
Total (uncentered) SS    = 2988.447827      Uncentered R2    = 0.9542
Residual SS              = 136.9920776      Root MSE        = .4031

```

ln_wage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
tenure	.1260126	.0439573	2.87	0.005	.0393536	.2126715
age	.0029424	.0032327	0.91	0.364	-.0034306	.0093155
not_smsa	-.2569617	.0921363	-2.79	0.006	-.4386024	-.075321
_cons	1.434409	.1009936	14.20	0.000	1.235307	1.633511

```

Underidentification test (Kleibergen-Paap rk LM statistic):      16.263
                                                                Chi-sq(2) P-val =    0.0003

Weak identification test (Cragg-Donald Wald F statistic):      27.689
(Kleibergen-Paap rk Wald F statistic):      11.305
Stock-Yogo weak ID test critical values: 10% maximal IV size    19.93
                                           15% maximal IV size    11.59
                                           20% maximal IV size     8.75
                                           25% maximal IV size     7.25

Source: Stock-Yogo (2005). Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.

```

Hansen J statistic (overidentification test of all instruments):	0.010
Chi-sq(1) P-val =	0.9207

```

Instrumented:      tenure
Included instruments: age not_smsa
Excluded instruments: union south

```

```

Test with clustered errors
-----

bootstrap replicates (50)
|-----| 1 |-----| 2 |-----| 3 |-----| 4 |-----| 5
.....
50

H0: Outliers do not distort 2SLS classical estimation
-----

chi2(4)=2.86
Prob > chi2 = .582

```

The robust-cluster variance estimator has been used to estimate the standard errors, and as previously explained, a cluster-bootstrap procedure (sampling is done from clusters with replacement to account for the correlations of observations within cluster) has been used to calculate the W statistic of the outlier test.

4 Conclusion

The `robivreg` command implements an IV estimator robust to outliers and allows their identification. In addition, a generalized Hausman test provides the means to evaluate whether the gain in robustness outweighs the loss in efficiency and thus justifies the use of a robust IV estimator.

5 References

- Baum, C. F., M. E. Schaffer, and S. Stillman. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal* 7: 465–506.
- Cohen Freue, G. V., H. Ortiz-Molina, and R. H. Zamar. 2011. A natural robustification of the ordinary instrumental variables estimator. Working paper. <http://www.stat.ubc.ca/~ruben/website/cv/cohen-zamar.pdf>.
- Dehon, C., M. Gassner, and V. Verardi. 2009. Extending the Hausman test to check for the presence of outliers. ECARES Working Paper 2011-036, Université Libre de Bruxelles. <http://ideas.repec.org/p/eca/wpaper/2013-102578.html>.
- Desbordes, R., and V. Verardi. 2011. The positive causal impact of foreign direct investment on productivity: A not so typical relationship. Discussion Paper No. 11-06, University of Strathclyde Business School, Department of Economics. http://www.strath.ac.uk/media/departments/economics/researchdiscussionpapers/2011/11-06_Final.pdf.
- Romer, D. 1993. Openness and inflation: Theory and evidence. *Quarterly Journal of Economics* 108: 869–903.
- Verardi, V., and C. Croux. 2009. Robust regression in Stata. *Stata Journal* 9: 439–453.
- Verardi, V., and C. Dehon. 2010. Multivariate outlier detection in Stata. *Stata Journal* 10: 259–266.
- Verardi, V., and A. McCathie. 2012. The S-estimator of multivariate location and scatter in Stata. *Stata Journal* 12: 299–307.
- Wooldridge, J. M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. Mason, OH: South-Western.

About the authors

Rodolphe Desbordes is a senior lecturer in economics at the University of Strathclyde in Glasgow, UK. His research interests include applied econometrics, international economics, and economic growth.

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