A Robust Instrumental Variables Estimator

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

Verardi Vincenzo
University of Namur (CRED) and ULB (ECARES and CKE)
Namur, Belgium
vverardi@ulb.ac.be

Abstract.
The classical instrumental variables (IV) estimator is extremely sensitive to the presence of outliers in the sample. This is a concern as 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 since they are themselves based on non-robust (to outliers) estimators. Furthermore, they do not take into account the combined influence of outliers in the first and second stages of the IV estimator. In this paper we present a robust IV estimator, initially proposed by Cohen-Freue and Zamar (2006), that we have programmed in Stata and made available via the robivreg command. We have improved on Cohen-Freue and Zamar (2006)’s 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 generalised Hausman test for the presence of outliers.

Keywords: st0001, Multivariate outliers, Robustness, S-estimator, Instrumental variables

1 Theory
Assume a linear regression model given by:

\[ y = X\theta + \varepsilon \]  

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

$$\hat{\theta}_{LS} = \arg\min_{\theta} r' r$$  \hspace{1cm} (2)$$

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

$$\hat{\theta}_{LS} = \left( X'X \right)_{n \Sigma_{XX}}^{-1} X'y_{n \Sigma_{XY}}$$  \hspace{1cm} (3)$$

which is simply the product of the \((p \times p)\) covariance matrix of the explanatory variables \(\Sigma_{XX}\) and the \((p \times 1)\) vector of the covariances of the explanatory variables and the dependent variable \(\Sigma_{XY}\) (the \(n\) simplify).

The unbiasedness and consistency of the LS estimates crucially depend on the absence of correlation between \(X\) and \(\varepsilon\). 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, which 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 \(Z\) as the \((n \times m)\) matrix (where \(m \geq p\) containing the instruments. The IV estimator (generally called two stages least squares when \(m > p\)) can be conceptualised as a two stage estimator. In the first stage, each endogenous variable is regressed on the instruments and on the variables in \(X\) that are not correlated with the error term. The predicted value for each variable is then fitted (denoted \(\hat{X}\) 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 estimated by \(LS\). The final estimator is

$$\hat{\theta}_{IV} = \left( \Sigma_{XZ} \left( \Sigma_{ZZ} \right)^{-1} \Sigma_{ZX} \right)^{-1} \Sigma_{XZ} \left( \Sigma_{ZZ} \right)^{-1} \Sigma_{ZY}$$  \hspace{1cm} (4)$$

where \(\Sigma_{XZ}\) is the covariance matrix of the original right-hand side variables and the instruments, \(\Sigma_{ZZ}\) is the covariance matrix of the instruments and \(\Sigma_{ZY}\) 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 biased, even asymptotically. Cohen-Freue and Zamar (2006) therefore suggest to replace the classical estimated covariance matrices in (4) by some robust counterparts that withstands the contamination. These could be Minimum Covariance Determinant (MCD) scatter matrices as presented in Verardi and Dehon (2010) or S-estimators of location and scatter as described by Verardi and McCathie (2011). We use the latter and the superscript \(S\) is employed to indicate it.
The robust IV estimator can therefore be written as:

$$\hat{\theta}_{RIV}^S = \left( \Sigma_{XX}^S \left( \Sigma_{ZZ}^S \right)^{-1} \Sigma_{XZ}^S \right)^{-1}$$

$$\left( \Sigma_{ZZ}^S \left( \Sigma_{XX}^S \right)^{-1} \Sigma_{XZ}^S \right)^{-1} \Sigma_{Zy}^S$$

(5)

As shown by Cohen-Freue and Zamar (2006), 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 and/or symmetric. 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:

$$\hat{\theta}_{RIV}^W = \left( \Sigma_{XX}^W \left( \Sigma_{ZZ}^W \right)^{-1} \Sigma_{XZ}^W \right)^{-1}$$

$$\left( \Sigma_{ZZ}^W \left( \Sigma_{XX}^W \right)^{-1} \Sigma_{XZ}^W \right)^{-1} \Sigma_{Zy}^W$$

(6)

where $W$ stands for weights. Robust covariance $\hat{\Sigma}_{XX}^S$ and robust Mahalanobis distances, i.e. $d_i = \sqrt{(M_i - \hat{\mu}_M)^T \hat{\Sigma}_M^{-1} (M_i - \hat{\mu}_M)}$ where $M = (X, Z, Y)$, $\hat{\Sigma}_M = \hat{\Sigma}_{XX}^S$ is the scatter matrix of explanatory variables, and $\hat{\mu}_M$ is the location vector, are first estimated. Outliers are then identified as the observations that have a robust Mahalanobis distance $\hat{d}_i$ larger than $\sqrt{\chi^2_{p+m+1,q}}$ where $q$ is a confidence level (e.g. 99%), given that Mahalanobis distances are distributed as the square root of a Chi-square with degrees of freedom equal to the length of vector $\hat{\mu}_M$. Finally, observations that are associated to a $\hat{d}_i$ larger than the cut-off point are downweighted and the classical covariance matrix is estimated. The weighting that we adopt is simply to award a weight one for observations associated to a $\hat{d}_i$ smaller than the cut-off value and zero otherwise.

The advantage of this last estimator is that standard overidentification, underidentification and weak instruments tests can easily be obtained, since 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 \texttt{ivreg2} command (Baum et al. 2007) to compute the final estimates; the reported tests and standard errors are those provided by this command.\footnote{The \texttt{robivreg} command is not a full wrapper for the \texttt{ivreg2} command. However, a sample free of outliers can easily be obtained by using the \texttt{gen(varname)} option that we describe in the next section.} Finally, a substantial gain in efficiency with respect to the standard robust IV estimator proposed by Cohen-Freue and Zamar (2006) can be attained. We illustrate this efficiency gain by running 1000 simulations using a setup similar to that of Cohen-Freue and Zamar (2006) with no outliers: 1000 observations for 5 random variables ($x, u, v, w, Z$) drawn from a multivariate normal distribution with mean $\mu = (0, 0, 0, 0, 0)$ and covariance
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 to this endogeneity bias, $X$ is instrumented by $Z$. For this set-up, the simulated efficiency of the two estimators is 46.7% for the raw $\theta_{RIV}^S$ estimator and 95.5% for the reweighted $\theta_{RIV}^W$ estimator, respectively. The efficiency is calculated as follows: Assume $\theta_{RIV}^S (\theta_{RIV}^W)$ is asymptotically normal with covariance matrix $V$, and assume $V_0$ is the asymptotic covariance matrix of the classical $\theta_{IV}$ estimator. The efficiency of $\theta_{RIV}^S (\theta_{RIV}^W)$ is calculated as $eff(\theta_{RIV}^S) = 1 - \frac{1}{\lambda_1(V^{-1}V_0)}$, 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) gen(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

The robivreg command has twelve options.

- `first` reports various first-stage results and identification statistics.

- `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 intra-group correlation, where `varname` identifies the group.
\texttt{gen(varname)} generates a dummy named \textit{varname}, which takes the value of one for observations which are flagged as outliers.

\texttt{raw} specifies that Cohen-Freue and Zamar (2006)'s estimator should be returned. Note that the standard errors reported are different from the ones that they proposed as they are robust to heteroskedasticity and asymmetry. The asymptotic variance of the raw estimator is described in Verardi and Croux (2009).

\texttt{cutoff(#)} allows the user to change the percentile above which an individual is considered as an outlier. By default, the cut-off point is set to 0.99.

\texttt{mcd} specifies that a Minimum Covariance Determinant (MCD) estimator of location and scatter is used to estimate the robust covariance matrices. By default, a S-estimator of location and scatter is used.

\texttt{graph label(varname)} generates a graphic in which outliers are identified according to their type, and labelled using the variable \textit{varname}. Vertical lines identify vertical outliers (observations with a large residual) and the horizontal line identifies leverage points.

\texttt{test} specifies that a test for the presence of outliers in the sample should be reported. 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 et al. (2009) and Desbordes and Verardi (2011) where

\begin{equation}
W = (\hat{\theta}_{IV} - \hat{\theta}_{RIV}^S)^T[Var(\hat{\theta}_{IV}) + Var(\hat{\theta}_{RIV}^S) - 2Cov(\hat{\theta}_{IV}, \hat{\theta}_{RIV}^S)]^{-1}(\hat{\theta}_{IV} - \hat{\theta}_{RIV}^S) \tag{7}
\end{equation}

Bearing in mind that this statistic is asymptotically distributed as a \( \chi^2_p \), where \( p \) is the number of covariates, it is possible to set an upper bound above which the estimated parameters can be considered as statistically different and hence the robust IV estimator should be preferred to the standard IV estimator. When the option \texttt{cluster(varname)} is requested, a cluster bootstrap is used to calculate the \( W \)-statistic. By default, the number of bootstrap replications is set to 50, but the option \texttt{nreps(#)} allows the user to modify that number. The option \texttt{nodots} suppresses the replication dots.

3 Empirical example

In a seminal paper, 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 GDP, by the logarithm of a country's land area.
From a pedagogical perspective, it is useful to start with the dependent variable, the average annual inflation rates (since 1973), in levels, as in example 16.6 of Wooldridge (2009), p.558.

```stata
use http://fmwww.bc.edu/ec-p/data/wooldridge/OPENNESS,clear
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 to -6.02284 |
| _cons       | 29.60664 | 5.608412 | 5.28 | 0.000 | 18.61435 to 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 eighties. Hence, we reestimate the model, using 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 = 1322.080301 Centered R2 = 0.0514
Total (uncentered) SS = 10844.0201 Uncentered R2 = 0.8844
Residual SS = 1254.073829 Root MSE = 3.935
```

| inf         | Coef. | Std. Err. | t   | P>|t| | [95% Conf. Interval] |
|-------------|-------|-----------|-----|------|----------------------|
| opendec     | -12.0739 | 7.642656 | -1.58 | 0.118 | -27.28027 to 3.1327  |
| _cons       | 14.60224 | 2.500814 | 5.84 | 0.000 | 9.626404 to 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


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

\[ \chi^2(2) = 10.53 \]
Prob > \chi^2 = 0.005

Once the influence of outliers are downweighted, the value of the coefficient on \textit{opendec} becomes much smaller and loses statistical significance. Our test for outliers, requested using the option \texttt{test}, confirms that outliers distort enough the original estimates that robustness should be favoured at the expense of efficiency.

The outliers can be easily identified using the \texttt{graph} option. We facilitate the identification of each type of outliers by setting vertical and horizontal cut-off points in the reported graph. The vertical cut-off points are 2.25 and -2.25. If the residuals were normally distributed, values above or below these cut-off points would be strongly atypical since they would be 2.25 standard deviations away from the mean (which is zero by construction), with a probability of occurrence of 0.025. The reported residuals are said to be robust and standardised because the residuals are based on a robust-to-outliers estimation and have been divided by the standard deviation of the residuals associated with non-outlying observations. In line with our downweighting scheme, the horizontal cut-off point is, by default, \( \sqrt{\frac{2}{p+m+1}} \). Vertical outliers are observations above or below the vertical lines, while leverage points are to the right of the horizontal line.

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

\[ . \text{ivreg2 llnf ( opendec= lland),r small} \]

\begin{verbatim}
IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

Number of obs = 114
F( 1, 112) = 10.96
Prob > F = 0.0013

Total (centered) SS = 56.83528551 Centered R2 = 0.1028
Total (uncentered) SS = 770.4438311 Uncentered R2 = 0.9338
\end{verbatim}
Figure 1: Identification of outliers when \( inf \) is used

Residual SS = 50.99226107  Root MSE = 0.6748

|     | Robust Coef. | Std. Err. | t   | P>|t|  | [95% Conf. Interval] |
|-----|--------------|-----------|-----|------|---------------------|
| opendec | -1.315804 | 0.3974457 | -3.31 | 0.001 | -2.103292 to -0.528316 |
| _cons  | 2.98983    | 0.179767  | 16.63 | 0.000 | 2.633645 to 3.346016 |

Underidentification test (Kleibergen-Paap rk LM statistic): 14.090  
Chi-sq(1) P-val = 0.0002

Weak identification test (Cragg-Donald Wald F statistic): 90.900  
(Kleibergen-Paap rk Wald F statistic): 44.466

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

NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.

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

Instruments: opendec
Excluded instruments: lland

The coefficient on opendec is now significant at the 1% level and suggests, using 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 reestimate the model again, using the `robivreg` command.

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

IV (2SLS) estimation

|                | Coef. Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------------|-----------------|-------|-----|---------------------|
| linf           | -1.225348       | .7034456 | -1.74 | 0.085 | -2.623523 .1728262 |
| opendec        | 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

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 using the `graph` option.

In figure 2, we can see that taking the log of `inf` has been insufficient to deal with all outliers in the dependent variable as `Bolivia` remains an outlier. Furthermore, Romer was right to be worried that `Lesotho` or `Singapore` may “have an excessive influence on the results” (p.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.

In a last example, we illustrate the use of the option `test` in conjunction with the option `cluster`. 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

<table>
<thead>
<tr>
<th>Number of clusters (idcode)</th>
<th>209</th>
<th>Number of obs</th>
<th>847</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (centered) SS</td>
<td>126.4871762</td>
<td>Centered R2</td>
<td>-0.0831</td>
</tr>
<tr>
<td>Total (uncentered) SS</td>
<td>2988.447827</td>
<td>Uncentered R2</td>
<td>0.9542</td>
</tr>
<tr>
<td>Residual SS</td>
<td>136.9920776</td>
<td>Root MSE</td>
<td>0.4031</td>
</tr>
</tbody>
</table>

| ln_wage | 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


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) = 2.02
Prob > chi2 = .732

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 employed 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 generalised 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


