

istat working papers

N.11
2016

Testing the validity of instruments in an exactly identified equation

Marco Ventura

istat working papers

N.11
2016

Testing the validity of instruments in an exactly identified equation

Marco Ventura

Comitato scientifico

Giorgio Alleva
Tommaso Di Fonzo
Fabrizio Onida

Emanuele Baldacci
Andrea Mancini
Linda Laura Sabbadini

Francesco Billari
Roberto Monducci
Antonio Schizzerotto

Comitato di redazione

Alessandro Brunetti
Romina Fraboni
Maria Pia Sorvillo

Patrizia Cacioli
Stefania Rossetti

Marco Fortini
Daniela Rossi

Segreteria tecnica

Daniela De Luca Laura Peci Marinella Pepe Gilda Sonetti

Istat Working Papers

Testing the validity of instruments in an exactly identified equation

N. 11/2016

ISBN 978-88-458-1902-5

© 2016

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Salvo diversa indicazione la riproduzione è libera,
a condizione che venga citata la fonte.

Immagini, loghi (compreso il logo dell'Istat),
marchi registrati e altri contenuti di proprietà di terzi
appartengono ai rispettivi proprietari e
non possono essere riprodotti senza il loro consenso.

Testing the validity of instruments in an exactly identified equation*

Marco Ventura †

Sommario

In econometria applicata, quando il numero di variabili strumentali é pari al numero delle variabili endogene non é possibile effettuare test di validit  degli strumenti, poich  questi non hanno una distribuzione di probabilit  nota. A tal riguardo, il presente lavoro si propone di spiegare ed implementare il metodo proposto da Lewbel (2012) per aumentare endogenamente il numero di strumenti validi ed effettuare cos  un test di validit  degli strumenti, agitando il vincolo della esatta identificazione.

Parole Chiave: variabili strumentali, esatta identificazione, test di sovra-identificazione, test di Hansen-Sargan, endogeneit 

Abstract

In applied econometrics, testing the validity of instruments is not a feasible task when the number of instruments equals the number of endogenous variables, because the statistics do not obey a well know probability distribution. Based on recent econometric developments (Lewbel, 2012), this article aims at illustrating an empirical strategy to overcome this setback, thus sidestepping the exact identification problem.

Keywords: instrumental variables, exact identification, overidentifying test, Hansen-Sargan test, endogeneity

* The views expressed in this paper are solely those of the author and do not involve the responsibility of Istat.

† mventura@istat.it

Contents

1. Introduction	8
2. Where the problem comes from	8
3. Implementing the solution	10
3.1 On the reliability of the estimates	10
4. The data and the estimates	10
5. Conclusion	11
References	12

1. Introduction

In many applied works OLS regressions suffer from endogeneity and/or measurement errors and/or omitted variable problems, which if ignored cause bias and inconsistency of the estimates. In such situations textbooks emphasize the usefulness of instrumental variables (IV) technique which can restore all the desired properties of the estimates. However, not any instrument can fix the problem, as they must simultaneously satisfy three conditions. Thus, finding appropriate instruments is quite difficult and this constitutes the greatest obstacle to the use of IV technique. Good examples of instruments are given by Angrist and Krueger (1991) and Angrist (1990). However, even when the researcher succeeds in finding appropriate instruments, those instruments may just be at a minimum in the sense that they are equal to the number of variables to be instrumented and in such situations overidentifying tests and orthogonality conditions cannot be performed (Hayashi, 2000). This instance is commonly referred to as the exact identification problem. Recently, Lewbel (2012) has shown that appropriate instruments can be generated on the basis of heteroskedastic errors, which is quite a plausible assumption easily testable. Hence, taking advantage of this fundamental contribution, this article presents a practical methodology enabling the researcher to overcome the exact identification problem.

The remainder of the article is as follows. Section 2. briefly states formally the exact identification problem, Section 3. details step by step the empirical strategy to be followed, Section 4. presents two examples and finally Section 5. draws some conclusions.

2. Where the problem comes from

Instrumental variables are employed in linear regression models, e.g.

$$y = \mathbf{X}\beta + \varepsilon \quad (1)$$

where y is an $N \times 1$ vector, \mathbf{X} is a $N \times k$ matrix containing k covariates and ε is an $N \times 1$ error term, where the zero conditional mean assumption $E[\varepsilon|\mathbf{X}] = 0$ does not apply. Reliance on IV methods requires that appropriate instruments are available to identify the model: often via exclusion restrictions. Those instruments, included in \mathbf{Z} , a $N \times p$ matrix with $p \geq k$, must satisfy three conditions: (i) they must themselves satisfy orthogonality conditions ($E[\mathbf{Z}'\varepsilon] = 0$); (ii) they must exhibit meaningful correlations with \mathbf{X} , i.e. ($E[\mathbf{Z}'\mathbf{X}] \neq 0$); and (iii) they must be properly excluded from the model, so that their effect on the response variable is only indirect. The contemporaneous fulfillment of all the three conditions is quite difficult, and when p equals k the Hansen-Sargan overidentifying test cannot be performed. Under the joint null hypothesis of this test that the instruments are valid, i.e. uncorrelated with the error term, and that the excluded instruments are correctly excluded from the estimated equation, the test statistic is distributed as chi-squared in the number of overidentifying restrictions, i.e. $\chi^2_{(p-k)}$.¹ Another important test is the so-called C test, or "difference-in-Sargan" which turns quite useful when the researcher suspects a subset of the instruments to be invalid, and wishes to test them. The statistic is computed as the difference between two Sargan statistics (or, for efficient GMM, two J statistics): that for the regression using the entire set of overidentifying restrictions, referred to as the restricted equation, versus that for the regression in which some instruments are removed from the set, referred to as the unrestricted equation. For excluded instruments, this is equivalent to dropping them from the instrument list. For included instruments, the C test places them in the list of included endogenous variables; in essence, treats them as endogenous regressors. The C test, distributed as a χ^2 with degrees of freedom equal to the number of suspect instruments being tested, has the null hypothesis that the specified variables are proper instruments. All these tests cannot be performed when $p = k$, namely when the number of instruments equals the number of regressors, or put another way when the number of excluded instruments equals the number of endogenous variables, which is quite a common situation.

¹ A rejection casts doubt on the validity of the instruments. For the efficient GMM estimator, the test statistic is Hansen's J statistic, the minimized value of the GMM criterion function. For the 2SLS estimator, the test statistic is Sargan's statistic. Just as IV is a special case of GMM, Sargan's statistic is a special case of Hansen's J under the assumption of conditional homoskedasticity. For further discussion see Hayashi (2000, 227-8, 407, 417).

Recently, Lewbel (2012) has proposed a method for constructing instruments as functions of the model's data. His approach may be applied when no standard external instruments are available, or, alternatively used to support the available instruments improving efficiency. In addition, supplementing external instruments can also allow to carry out a Hansen-Sargan tests of the orthogonality conditions, which would not be available in the case of exact identification by external instruments. Lewbel's approach may be applied to cross-sections, as well as time series and short panel data.

Consider y_1 and y_2 as observed endogenous variables, \mathbf{X} again a matrix of observed exogenous regressors, and $\varepsilon = (\varepsilon_1, \varepsilon_2)$ as unobserved error processes. Consider a structural model of the form:

$$\begin{aligned} y_1 &= \mathbf{X}\beta_1 + y_2\gamma_1 + \varepsilon_1 \\ y_2 &= \mathbf{X}\beta_2 + y_1\gamma_2 + \varepsilon_2, \end{aligned} \tag{2}$$

This system is triangular when $\gamma_2 = 0$, otherwise, it is fully simultaneous. The errors $(\varepsilon_1, \varepsilon_2)$ may be correlated with each other. If the exogeneity assumption, $E(X'\varepsilon_i) = 0$, for $i = 1, 2$ holds, the reduced form is identified, but in the absence of identifying restrictions, the structural parameters are not identified. These restrictions often involve setting certain elements of β_1 or β_2 to zero, which makes instruments available. In many applied contexts, condition (iii) is difficult to establish, the zero restriction on its coefficient may not be plausible and if it does not hold, IV estimates will be inconsistent. Identification in Lewbel's approach is achieved by restricting correlations of $\varepsilon\varepsilon'$ with X . This relies upon higher moments, and is likely to be less reliable than identification based on coefficient zero restrictions. However, in the absence of plausible identifying restrictions, this approach may be the only reasonable strategy. Therefore, in presence of heteroskedasticity related to, at least, some elements of X , identification can be achieved. In a fully simultaneous system, assuming that $cov(\mathbf{X}, \varepsilon_i^2) \neq 0$, $i = 1, 2$ and $cov(\mathbf{Z}, \varepsilon_1\varepsilon_2) = 0$ for observed \mathbf{Z} will identify the structural parameters. Note that \mathbf{Z} may be a subset of \mathbf{X} , so no information outside the model specified above is required. The key assumption that $cov(\mathbf{Z}, \varepsilon_1\varepsilon_2) = 0$ will automatically be satisfied if the mean zero error processes are conditionally independent: $\varepsilon_1 \perp \varepsilon_2 | \mathbf{Z}$. However, this independence is not always strictly necessary because in most of the cases only one equation is being estimated. The first-stage regression may be used to provide the necessary components. Indeed, generated instruments can be constructed from the first-stage equations residuals, multiplied by each of the included exogenous variables in mean-centered form:

$$Z_j = (X_j - \bar{X}_j)\hat{\varepsilon} \tag{3}$$

where $\hat{\varepsilon}$ is the residual vector from the first-stage regression of each endogenous regressor on all exogenous regressors, including a constant term. These first-step regression residuals have zero covariance with each of the regressors used to construct them, implying that the means of the generated instruments will be zero by construction. However, their element-wise products with the centered regressors will not be zero, and will contain sizable elements if there is clear evidence of scale heteroskedasticity with respect to the regressors. The greater the degree of scale heteroskedasticity in the error process, the higher will be the correlation of the generated instruments with the included endogenous variables which are the regressands in the first-step regressions. Hence, Lewbel's method can be used to estimate: a) a traditionally identified single equation, or b) a single equation that fails the order condition for identification: either by having no excluded instruments, or by having fewer excluded instruments than needed for identification. Notice that the exactly identified case falls into case a), it follows that by taking advantage of the greater number of instruments increased by generated instruments one can sidestep the impossibility to test the instruments by running a C test. As an example, suppose to have three included instruments, contained in \mathbf{X} , one endogenous variable, Y_1 , and one excluded instrument, Z . A Hansen-Sargan test on Z can be performed by increasing the number of instruments up to four, obtaining the Hansen-Sargan J statistics of the restricted equation (one standard plus three generated instruments), a $\chi^2_{(3)}$, then the Hansen-Sargan J of the unrestricted model (only generated instruments), a $\chi^2_{(2)}$. Finally the C statistics, testing the validity of the external instrument, is obtained as the difference of the two J 's statistics, and it will be distributed as a $\chi^2_{(1)}$.

3. Implementing the solution

Given the instruments provided by Lewbel's method, before running the IV estimates, all one needs to do to overcome the exact identification problem, is to make sure of the presence of heteroskedasticity in the first-step residuals. Hence, the whole procedure can be sketched in the following points:

1. run the first-step regression, one for each endogenous variable;
2. test for heteroskedasticity. If you cannot accept the null of homoskedasticity go to the next point;
3. generate the instruments as in eq. (3);
4. estimate the restricted equation and take the Hansen-Sargan J statistics;
5. estimate the unrestricted equation and take the Hansen-Sargan J statistics;
6. work out the J statistics (or more precisely the C) as the difference between the two J 's.

3.1 On the reliability of the estimates

To a certain extent, the idea proposed in this work can be regarded as a sort of corollary of Lewbel's identification strategy, not explicitly mentioned in his seminal work and a crucial issue is whether Lewbel's identification strategy is truly capable of replicating the results validated by the literature as being "the correct ones". It follows that a major challenge is whether the generated instruments are capable of producing estimates quite close to those that were obtained using outside instruments validated by the literature. At this purpose, the author himself provides evidence and a considerable number of published works can be found in the literature. See, for instance Mishra and Smyth (2015) Block (2007), Sabia (2007), Kevin and Oppedisano (2013), Brown (2014), Chowdhury et al (2014), Mishra and Smyth (2015), just to cite a few.

The next section is going to provide two examples applied on cross-section and panel data, respectively.

4. The data and the estimates

In order to make the estimates replicable we use two publicly available datasets.² The first one is taken from Mroz's (1987) article consisting of 428 married women between ages of 30 and 60 in 1975 working at some time during the year. A simple wage equation is estimated, in which the dependent variable is the (log) of hourly wage regressed on years of experience and its squared value, plus education (in years), which represents the endogenous variable because of self-selection. As a unique instrument for education we use the number of children between 6 and 18.

According to a Breusch-Pagan test the null of homoskedasticity cannot be accepted ($Pval = .058$). In all the three estimates education is never significant, however generated instruments help standard IV to increase precision. To our purpose, the important fact to notice is that the standard IV column reports the value of the test of the validity of the instruments, in spite of the exact identification problem. In particular, the difference-in-Sargan test does not reject the validity of the standard instrument ($Pval = .957$). Similarly, one cannot reject the null of validity of the instruments for both the unrestricted and restricted models.

The second example is taken from Grunfeld (1960), a panel sample of 10 firms observed over the years 1934 – 1954. The dependent variable is the firms' investment, regressed on three lags of the capital stock plus the market value of the firm, which is clearly endogenous. As a standard instrument we use the lag of the market value.

A Hausman test provides evidence against the random effect hypothesis and the Breusch-Pagan test on the first-stage equation does not accept the null of homoskedasticity at 1%. The market value

² The dataset are available at <http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.dta> and <http://www.stata-press.com/data/r9/grunfeld.dta>.

Table 1 - Standard IV, generated instruments, standard plus generated instruments

	standard IV	generated inst. (unrestricted)	standard plus generated (restricted)
educ	.172 (.136)	.227 (.916)	.180 (.134)
exper	.379** (.015)	.0348** (.054)	.037 (.016)
expersq	-6.88E-04 (4.76E-04)	-5.83E-04 (.002)	-6.72E-04 (4.80E-04)
cons	-1.322 (1.682)	-2.0 (11.3)	-1.429 (1.66)
N	428	428	428
J	.003	.359	.362
Pval(J)	.957	.549	.835
df	1	1	2

*, ** and *** significant at 10%, 5% and 1% levels, respectively, robust standard errors in parenthesis. Dependent variable (log) wage, standard IV number of children between 6 and 18.

Table 2 - Standard IV, generated instruments, standard plus generated instruments

	standard IV	generated inst. (unrestricted)	standard plus generated (restricted)
mkt value	.333*** (.122)	.198*** (.059)	.204*** (.058)
K stock(t-1)	.836*** (.161)	.826 (.192)	.826*** (.188)
K stock(t-2)	-.857*** (.307)	-.888 (.24)	-.887*** (.239)
K stock(t-3)	.1 (.238)	.287 (.19)	.279 (.188)
cons	-13.38 (11.1)	-3.742 (8.8)	-4.159 (8.67)
N	170	170	170
J	2.11	7.629	9.738
Pval(J)	.146	.022	.021
df	1	2	3

*, ** and *** significant at 10%, 5% and 1% levels, respectively, standard Errors in parentheses. Dependent variable firms' investment, standard IV lag of firms' market value. Fixed effects estimates.

coefficient is always positive and significant, it ranges from .33 in the standard IV case, to a lower value of .20 of the restricted equation, estimated with higher precision. Again, the relevant point to our purpose is that it is still possible to test the validity of the unique external instrument which turns valid ($Pval = .146$). Differently, the unrestricted restricted set of instruments turn not valid.

5. Conclusion

This paper has put forth an empirical strategy to test for the validity of the instruments in an exactly identified equation. In particular, it consists in incrementing the number of instruments by means of Lewbel's instruments generation which takes advantage of heteroskedastic errors, which is quite a common condition in applied works. In doing so one can work out the difference-in-Sargan statistics as the difference between restricted and unrestricted equation.

Bibliography

- Angrist, J.D., 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review*, 80, 313-336.
- Angrist, J.D., Krueger, A.B., 1991. Does Compulsory School Attendance Affect Schooling and Earnings?. *Quarterly Journal of Economics*, 106, 979-1014.
- Block S.A., 2007. Maternal nutrition knowledge versus schooling as determinants of child micronutrient status. *Oxford Economic Papers*, 59, 330-353.
- Brown T.T., 2014. How effective are public health departments at preventing mortality?. *Economics and Human Biology* 13, 34-45
- Chowdhury M.T.H., Bhattacharya P.S., Mallick D., Ulubasoglu M.A., 2014. An empirical inquiry into the role of sectoral diversification in exchange rate regime choice. *European Economic Review*, 67 (2014) 210-227
- Grunfeld, Y., 1960. The Determinants of Corporate Investment. In *The demand for durable goods*. Harberger eds. Chicago: University of Chicago Press.
- Hayashi, F., 2000. *Econometrics*. Princeton, NJ: Princeton University Press.
- Kevin D., Oppedisano V., 2013. The surprising effect of larger class sizes: Evidence using two identification strategies. *Labour Economics*, 23, 57-65
- Lewbel, A., 2012. Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models. *Journal of Business and Economic Statistics*, 30(1), 67-80.
- Mishra V., Smyth R. 2015. Estimating returns to schooling in urban China using conventional and heteroskedasticity-based instruments. *Economic Modelling* 47, 166-173
- Mroz, T.A., 1987. The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica* 55, 765-799
- Sabia J.J., 2007. Reading, writing, and sex: The effect of losing virginity on academic performance. *Economic Inquiry*, 45(4), 647-670.