Economics Letters 205 (2021) 109935

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Instrument approval by the Sargan test and its consequences for coefficient estimation

Jan F. Kiviet [a,b,*], Sebastian Kripfganz [c]

[a] Amsterdam School of Economics, University of Amsterdam, PO Box 15867, 1001 NJ, Amsterdam, The Netherlands
[b] Department of Economics, University of Stellenbosch, South Africa
[c] University of Exeter Business School, Streatham Court, Rennes Drive, Exeter EX4 4PU, UK

ARTICLE INFO

Article history:
Received 11 May 2021
Received in revised form 22 May 2021
Accepted 22 May 2021
Available online 29 May 2021

JEL classification:
C12
C15
C26

Keywords:
Endogeneity testing
Instrument validity
Instrument weakness
Size control
Simulation
Test power

ABSTRACT

Empirical econometric findings are often vindicated by supplementing them with the p-values of Sargan–Hansen tests for overidentifying restrictions, provided these exceed a chosen small nominal significance level. It is illustrated here that the probability that such tests reject instrument validity may often barely exceed small levels, even when instruments are seriously invalid, whereas even minor invalidity of instruments can severely undermine inference on regression coefficients by instrumental variable estimators. These uncomfortable patterns may be aggravated when particular valid or invalid instruments are relatively weak or strong.

© 2021 University of Amsterdam. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Many economic relationships are possibly characterized by instantaneous feedbacks. In that case not only the dependent variable is endogenous, but some explanatory variables as well. The standard methods to find out whether explanatory variables are endogenous, and – if they are – to cope with them when estimating reaction coefficients, exploit external instrumental variables. In order to qualify as an effective external instrument, a variable should not be explanatory for the dependent variable indeed (its exclusion from the model should be valid), and it should be – preferably strongly – correlated with the potentially endogenous regressors. Verification of the latter property is relatively straightforward, whereas the former is usually tested by the Sargan (1958) test – in relatively simple linear models – or by its generalization, the Hansen (1982) test. Various studies – see, for instance, Hayashi (2000, p. 218), Parente and Santos Silva (2012) and Kiviet (2017) – have spelled out a fundamental methodological limitation of these tests, being that they can only verify over-identifying restrictions, while implicitly adopting untestable just-identifying restrictions.

This limitation is usually not seriously addressed in empirical work. One reason for this may be that in the literature very little attention has been paid to illustrating its actual consequences. Here we provide numerical findings obtained from simulating a simple illustrative model. They show how deceiving high p-values of Sargan tests can be, simply because these are very likely to be found, even when instruments are patently invalid. In addition, it is shown that mildly invalid instruments may already seriously bias instrumental variables based coefficient estimates, especially when the invalid instruments are strong and any valid instruments are weak. A possible way out is indicated in the concluding section.

2. Design of the experiments

By simulation experiments, we examine Sargan test and coefficient estimation outcomes for a simple linear regression model under a range of practically relevant circumstances. This model is given by

$$y = c + \beta x + u, \tag{2.1}$$

* Corresponding author at: Amsterdam School of Economics, University of Amsterdam, PO Box 15867, 1001 NJ, Amsterdam, The Netherlands.
E-mail addresses: J.F.Kiviet@uva.nl (J.F. Kiviet), S.Kripfganz@exeter.ac.uk (S. Kripfganz).

https://doi.org/10.1016/j.econlet.2021.109935
0165-1765/© 2021 University of Amsterdam. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

where u is a disturbance, x a possibly endogenous explanatory variable with constant numerical coefficient β, c a constant intercept, and y is the dependent variable. Regressor x is generated such that its correlation with u, indicated by ρ_{xu}, can be varied in the experiments. Next to the internal instrument established by the constant, also two external variables z_1 and z_2 are generated. Their correlations with u can be controlled by ρ_{z_1u} and ρ_{z_2u} respectively. When $\rho_{z_ju} = 0$ $(j = 0, 1)$, then z_j qualifies as a valid instrument. Moreover, the correlations ρ_{z_1x} and ρ_{z_2x}, determining instrument strength/weakness, can be varied. Samples of size n will be generated for $\{y, x, u, z_1, z_2\}$ which are i.i.d. (independent and identically distributed). In this study, we have restricted ourselves to examining Gaussian data sets.

For various interesting combinations of numerical values for ρ_{xu}, ρ_{z_1x}, ρ_{z_2x}, ρ_{z_1u}, ρ_{z_2u} and n, we investigate: (i) the rejection probability of the Sargan test at nominal significance level α, where we shall consider $0.01 \leq \alpha \leq 0.5$; and (ii) the distribution of the estimation errors $\hat{\beta} - \beta$ for various estimators of the slope coefficient, namely ordinary least squares (OLS), instrumental variable (IV, just using the external instrument z_1) and two-stage least squares (TSLS, using both z_1 and z_2) estimation.

Not all values smaller than one in absolute value for the five correlation coefficients are compatible. For instance, it is self-evidently impossible to have $\rho_{z_1u} = 0$, whereas both ρ_{xu} and ρ_{z_1x} are close to unity. Close to boundary values, and to notoriously problematic cases such as $\rho_{z_1x} \to 0$, $\rho_{xu} \to 1$, or n very small, instrumental variable estimators may show pathological behavior. It is not our intention here to demonstrate that such cases exist and are also problematic for the Sargan test.[1] Our primary aim is here to demonstrate that serious problems occur, too, for parameter combinations which seem pretty harmless at first sight. Therefore, we start to examine a reasonably large sample ($n = 250$) and rather middle of the road combinations of the correlations, viz.:

$$\rho_{xu} \in \{0.2, 0.4\}, \qquad \rho_{z_1u} \in \{0.0, 0.1\}, \qquad \rho_{z_1x} \in \{0.3, 0.6\}, \qquad \rho_{z_2u} \in \{0.0, 0.2\}, \qquad \rho_{z_2x} \in \{0.1, 0.4\}. \tag{2.2}$$

Hence, the instruments will not be chosen ultra-weak, nor extremely invalid. The estimation errors will in fact be examined by presenting graphs of their quartiles for all values $0 \leq \rho_{xu} \leq 0.9$ that are compatible with the other correlations.

All findings to be presented are invariant with respect to the actual values of the intercept c, the slope β, the means of the variables x, z_1 and z_2, and regarding $\sigma^2_{z_1}$ and $\sigma^2_{z_2}$, the variances of z_1 and z_2. The rejection frequencies of the Sargan test are invariant with respect to both σ_u and σ_x, whereas the quartiles of the various estimation errors will be presented for $\sigma_u/\sigma_x = 1$. Outcomes for different σ_u/σ_x ratios can be obtained simply by adapting the scale of the vertical axis of the graphs, as we shall show below. Hence, all findings will have wide relevance, especially for cross-sectional applications from which any additional uncontested exogenous regressors have been partialled out from the instruments and from the model, so that just one potentially endogenous regressor remains.[2]

3. Findings

Figs. 1 and 2 have two rows of two panels. Each row combines particular ρ_{z_1u} and ρ_{z_2u} values. The left-hand panels present rejection frequencies of the Sargan test (vertical axis) at nominal significance α (horizontal axis, $0.01 \leq \alpha \leq 0.5$) for eight combinations of ρ_{xu}, ρ_{z_1x}, ρ_{z_2x} values. For seven different estimator/instrument combinations, the right-hand panels present three similarly colored/marked lines, being the quartiles of the estimation error distribution (vertical axis), at endogeneity correlation ρ_{xu} (horizontal axis, for feasible $\rho_{xu} \in [0, 0.9]$). For each triple of lines the central one is the median. The other two provide an impression of the dispersion of the distribution of the estimation errors around the median. Their vertical distance represents the interquartile range at ρ_{xu} (50% of the generated estimation errors landed between these two lines).[3]

In the top-row of panels in Fig. 1, both instruments are valid. Its left-hand panel shows that for the examined eight cases (mentioned in the legend) the Sargan test shows no size problems: the actual probability of type I errors is extremely close to the nominal significance level for all α values examined. In the top right-hand panel, for all estimators/cases represented, except OLS, the three lines are found to be almost horizontal. Thus, these distributions are hardly determined by endogeneity of x, and they suggest median unbiasedness, especially for moderate values of ρ_{xu}. On the other hand, the bias of OLS seems proportional to the degree of endogeneity. The graphs for IV and TSLS show the decreasing effects on the dispersion of using stronger or (one) extra instruments. Note that the dispersion of OLS improves for higher ρ_{xu} and is not beaten by any of the much less biased instrumental variable estimators.

The bottom-row of graphs shows what the effects are when one of the two instruments is mildly invalid. When the valid instrument is relatively weak, the rejection probability of the Sargan test barely exceeds the significance level, especially when the invalid instrument is relatively strong. At $\alpha = 0.05$, instrument invalidity will be detected with probability 0.3 at most (for the sample size and correlation combinations examined). Thus, the type II error probability (wrongly approving the instruments) is high, exceeding 0.7. The adjacent panel shows that the often undetected instrument invalidity (of just $\rho_{z_1u} = 0.1$) is devastating for the TSLS estimators based on a valid and an invalid instrument, especially when both instruments are relatively weak. For the just-identified IV estimator – for which the Sargan test is not available – just using a mildly invalid instrument yields substantial bias, especially when this instrument is relatively weak. For all IV and TSLS estimators presented, the probability of a positive estimation error exceeds 0.75. For ρ_{xu} small, OLS yields smaller estimation errors than IV and TSLS. Note that in both rows of panels the OLS results are similar, because they are invariant to the properties of the two instruments. For judging cases where $\sigma_u/\sigma_x = \phi > 0$, one should simply scale the figures along the vertical axis by the factor ϕ.

In the top-row of panels in Fig. 2, again one instrument is valid and the other one invalid, but more seriously invalid than in row two. Self-evidently, the Sargan test rejects more frequently now, but at $\alpha = 0.05$ with a probability still below roughly 0.6 when the valid instrument is relatively weak and the invalid one relatively strong. On the other hand, using $\alpha = 0.5$ leads for all cases examined to a detection of the invalidity with a probability of 0.9 or larger. For the TSLS estimation errors, the closeness of the median to zero deteriorates when the valid instrument gets weaker and the invalid one stronger. Note that the IV results (using valid z_1) are similar to those in the top right-hand panel of Fig. 1.

In the bottom-row both instruments are invalid. This case highlights the perils of the Sargan test not being consistent for

[1] This is one of the main objectives in Davidson and MacKinnon (2015).

[2] Technical details on the generation of the simulated data series, derivation of invariance properties, and some extra simulation results are available as Supplementary Material.

[3] Taking a 95% interpercentile range leads to comparable relative differences between the various estimators. The presentation of means and standard errors has been avoided, because their population equivalent does not exist for some of the instrumental variable estimators.

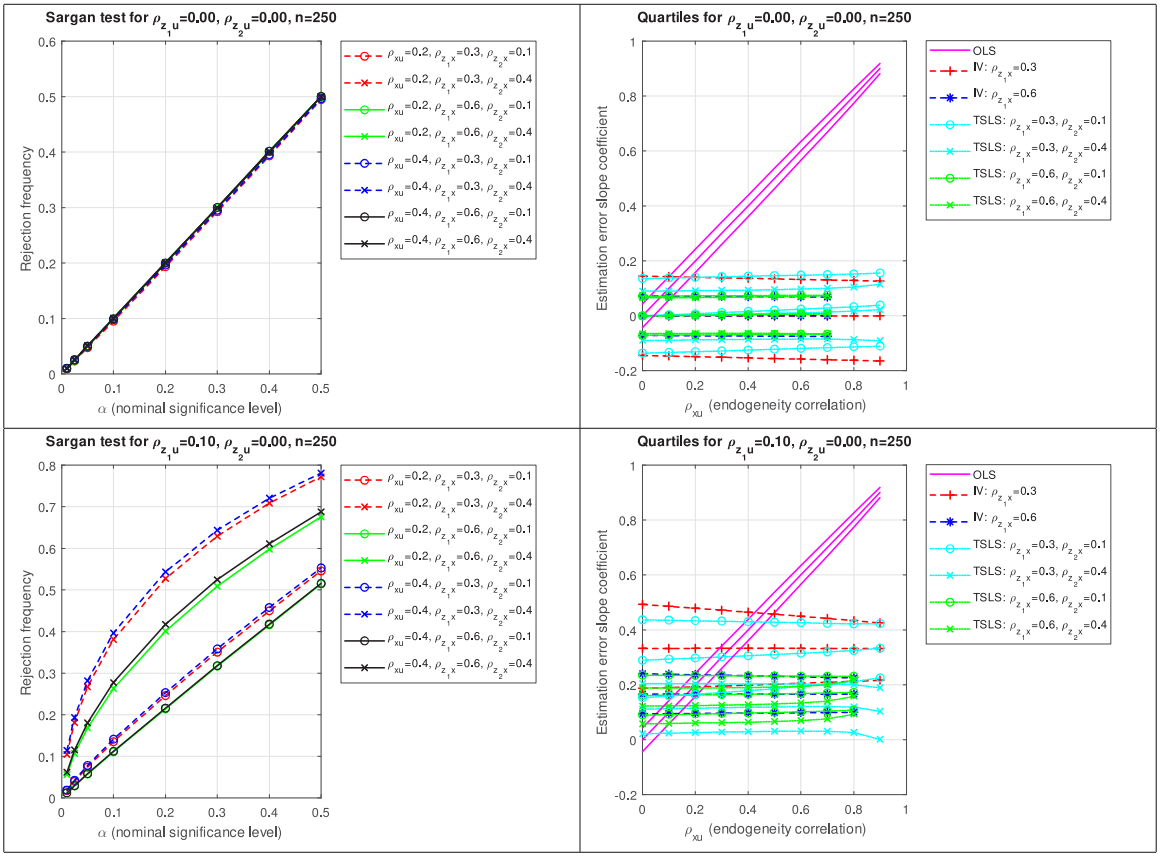

Fig. 1. Simulation results for $n = 250$, $\sigma_u/\sigma_x = 1$, and all correlation combinations with z_2 valid.

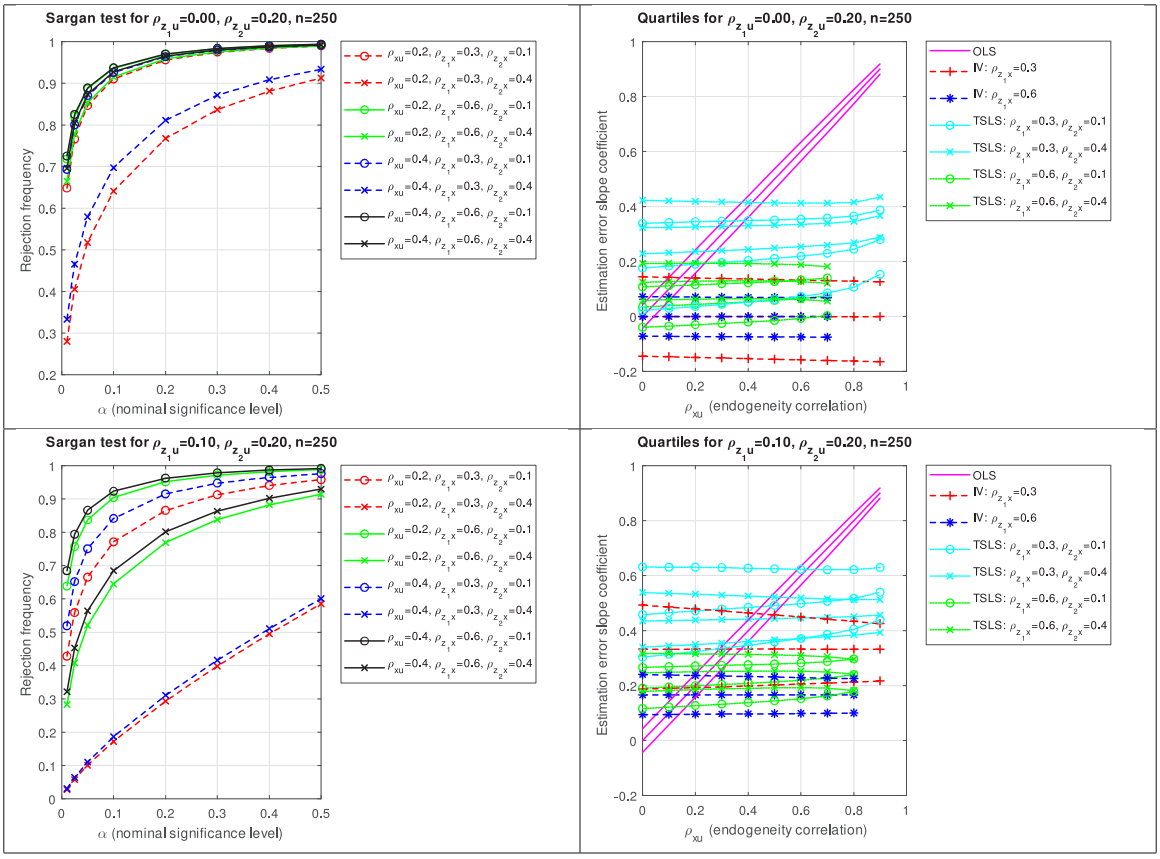

Fig. 2. Simulation results for $n = 250$; $\sigma_u/\sigma_x = 1$; and all correlation combinations with z_2 invalid.

inference on validity of both instruments. For some cases the rejection probability of instrument validity is quite high, but for two of them it hardly exceeds the nominal significance level. These are the two cases where $\rho_{z_1x} = 0.3$ and $\rho_{z_2x} = 0.4$, so the most seriously invalid instrument is also the strongest. The area in the parameter space where the Sargan test lacks any power in the present simple model is represented by (for proof see Supplementary Material):

$$\rho_{z_1u}/\rho_{z_1x} = \rho_{z_2u}/\rho_{z_2x}. \tag{3.1}$$

Hence, when for the two instruments their ratios of invalidity over strength are equal (or close), then the Sargan test is unable (or has great difficulty) to detect instrument invalidity, irrespective of the seriousness of this invalidity. This finding dramatically undermines trust in instrument-based methods, as this shows that the TSLS estimator for these two cases (where the ratios ρ_{z_1u}/ρ_{z_1x} and ρ_{z_2u}/ρ_{z_2x} are 0.33 and 0.5 respectively) is very badly biased over the whole range of ρ_{xu} values, whereas the Sargan test lacks power.

Results for $n = 50$ and $n = 2500$ are provided as Supplementary Material. The size control is still found to be close to perfect in the much smaller sample. As expected, the detection probabilities of instrument invalidity are generally lower/higher in smaller/larger samples, and the estimators have a smaller/larger dispersion, whereas the median varies little with n. For small n, more parameterizations show a futile power of the Sargan test. For $n = 2500$ the power of the Sargan test is often (almost) unity, except for cases that come close to satisfying (3.1), and these are also the cases where the TSLS estimation errors are furthest distributed away from zero.

4. Conclusions

Sargan/Hansen tests are only applicable when more candidate external instruments are available than the regression has potentially endogenous explanatories, and they just test the overidentifying restrictions. From our analytic and numerical results it follows that the Sargan test is not a trustful guide to decide on validity of all external instruments indeed, because its rejection probability can be close to the chosen significance level, even when instruments are seriously invalid and the sample size arbitrarily large. In a simple model, this occurs when for one endogenous regressor two external instruments are available, while these happen to have an almost similar ratio between their correlations regarding degree of invalidity and degree of strength. The test is shown to have poor power to detect instrument invalidity, too, when from the external instruments one (say, z_1) is valid and relatively weak, while the other one (z_2) is invalid and relatively strong, so that $\rho_{z_2u}\rho_{z_1x}/\rho_{z_2x}$ is close to $\rho_{z_1u} = 0$.

Regarding possible size distortions of Sargan–Hansen overidentifying restrictions tests, the literature provides mixed evidence.[4] In the linear static homoskedastic model examined here, we establish that size problems seem not a major issue, except perhaps for (not examined) pathological parameter combinations. To counter (putative) over-rejection problems, Hansen (2021, Ch. 12) advises practitioners to use the Sargan test at a very low nominal significance level. Given our findings, however, we would argue in favor of testing at a very high nominal significance level, because an insignificant value of the test is used in practice to approve validity of all instruments. Therefore, given the devastating effects that we established of even mildly invalid instruments on instrumental variable estimators, the primary worry should be to fail to reject invalid instruments (commit type II errors), and not so much to limit type I errors (wrongly rejecting valid instruments). Hence, one might decide to corroborate instruments and resulting TSLS findings only when the p-value of the Sargan test is really high; perhaps only when it is larger than 50% or even higher, instead of the habitual 5%, or just 1% as Hansen suggests!

In the simple static model, the results confirm that when the regressor is exogenous the OLS estimator is unbiased with the most attractive interquartile range, whereas for soaring endogeneity its bias sharply increases while its interquartile range slightly shrinks. Instrumental variable estimators are found to be almost (median) unbiased when the employed instruments are valid and not very weak. The findings on the relative width of their actual interquartile range when instruments are mildly weak already indicate that successful identification-robust[5] IV/TSLS inference must necessarily produce relatively uninformative inference. When instruments are invalid, just- and over-identified IV/TSLS are biased, and we find that this bias is largely invariant regarding the degree of endogeneity (unlike for OLS) and size of the sample. Currently, practitioners seem much more concerned about the misleading inference that will result from using supposedly valid though weak instruments than from invalid instruments, possibly because weakness – unlike invalidity – can directly be observed.

It is not self-evident how to examine in practice the sensitivity of IV/TSLS with respect to varying degrees of invalidity of instruments, whereas this is simpler and already feasible for OLS, because degree of endogeneity and instrument invalidity are the same thing for OLS, which uses its regressors as instruments. Kripfganz and Kiviet (2021) provide computer code and detailed instructions for a method, obtained in Kiviet (2020), which uses plausible assumptions on ρ_{xu} to correct OLS regarding its endogeneity bias while preserving its attractive dispersion. Next, inference can be obtained regarding the adequacy of the model specification (which may have an arbitrary number of endogenous regressors) and on its coefficients, including an alternative test for over-identification restrictions. This inference is endogeneity robust in the following sense. Specialized to the simple model of this study, these inferences are valid provided $\rho_{xu} \in [\rho_{xu}^L, \rho_{xu}^U] \subset (-1, 1)$. Choosing a narrow interval $[\rho_{xu}^L, \rho_{xu}^U]$ leads to narrow and thus attractive confidence intervals for the coefficients (with the risk that they are invalid if actually $\rho_{xu} \notin [\rho_{xu}^L, \rho_{xu}^U]$), and choosing $[\rho_{xu}^L, \rho_{xu}^U]$ wider yields more trustworthy wider (and ultimately unbounded) and thus less informative intervals.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2021.109935.

References

Andrews, I., Stock, J., Sun, L., 2019. Weak instruments in IV regression: Theory and practice. Annu. Rev. Econ. 11, 727–753.
Bowsher, C.G., 2002. On testing overidentifying restrictions in dynamic panel data models. Econ. Lett. 77, 211–220.
Davidson, R., MacKinnon, J.G., 2015. Bootstrap tests for overidentification in linear regression models. Econometrics 3, 825–863.
Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1053.
Hansen, B.E., 2021. Econometrics. Revised version: March 11, 2021. https://www.ssc.wisc.edu/~bhansen/econometrics/.

[4] Hayashi (2000, p. 218) suggests substantial over-rejection in finite samples, whereas Bowsher (2002) and Kiviet et al. (2017) report serious under-rejection in dynamic panel data models.

[5] This primarily aims to overcome the effects of poor estimation of the standard errors of IV/TSLS estimators when instruments are really weak, without coping with the estimator's bias and actual inefficiency. See Andrews et al. (2019) for a recent overview.

Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton, NJ, USA.
Kiviet, J.F., 2017. Discriminating between (in)valid external instruments and (in)valid exclusion restrictions. J. Econometric Methods 6, 1–9.
Kiviet, J.F., 2020. Testing the impossible: Identifying exclusion restrictions. J. Econometrics 218, 294–316.
Kiviet, J.F., Pleus, M., Poldermans, R.W., 2017. Accuracy and efficiency of various GMM inference techniques in dynamic micro panel data models. Econometrics 5 (1), 14.
Kripfganz, S., Kiviet, J.F., 2021. Kinkyreg: Instrument-free inference for linear regression models with endogenous regressors. Stata J. (in press).
Parente, P.M.D.C., Santos Silva, J.M.C., 2012. A cautionary note on tests of overidentifying restrictions. Econom. Lett. 115, 314–317.
Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. Econometrica 26, 393–415.