cemmap

centre for microdata methods and practice

A Note on Specification Testing in Some Structural Regression Models

Walter Beckert

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP22/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

A Note on Specification Testing in Some Structural Regression Models*

Walter Beckert†

May 10, 2019

Abstract

There exists a useful framework for jointly implementing Durbin-Wu-Hausman exogeneity and Sargan-Hansen overidentification tests, as a single artificial regression. This note sets out the framework for linear models and discusses its extension to non-linear models. It also provides an empirical example and some Monte Carlo results.

Word count: 3121.

JEL classification: C21, C26, C36.

Keywords: endogeneity, identification, testing, artificial regression.

*I thank the editor, Debopam Bhattacharya, and two anonymous referees, for insightful comments and suggestions that greatly improved the paper, as well as Haris Psaradakis, Ron Smith and Jouni Sohkanen for helpful comments and discussions.

†Department of Economics, Mathematics and Statistics, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK (w.beckert@bbk.ac.uk)

1 Introduction

Specification testing of structural linear simultaneous equations models with endogenous regressors is comprehensively surveyed in Hausman [1983]. A commonly applied test of the null hypothesis of exogenous regressors in linear regression models, under the maintained assumption of the exogeneity of a set of instruments, is due to Durbin [1954], Hausman [1978], Wu [1973]. If more instruments are available than necessary for identification, i.e. if the model is overidentified, again under the maintained assumption of the exogeneity (validity) of just identifying instruments, then a test of the validity of the imposed overidentifying restrictions, due to Sargan [1958, 1988], is another useful specification test.[1]

This note shows how, following a first-stage regression, in a *single* linear regression – on the independent variables, the first-stage residuals and a set of candidate overidentifying instruments – under the maintained assumption of the validity of just identifying instruments, (i) the coefficients of the structural regression equation can be consistently estimated, (ii) the null hypothesis of exogenous regressors can be tested and, in an overidentied model, (iii) the null hypothesis of the validity of overidentifying restrictions can be tested as well.

Importantly, the analysis of the linear regression model is interesting because the insights gained from it carry over to nonlinear models, such as nonlinear regression models and Generalized Linear Models [McCullagh and Nelder, 1983] in which there typically exist a variety of definitions for residuals – including Pearson, Anscombe, deviance residuals – and it is not a priori clear which one to use as the basis to construct test statistics and measure of fit. Such models can be estimated using an artificial or Gauss-Newton regression [Davidson and MacKinnon, 1990, 1993, 2001], and this algorithm provides the conceptual link to the analysis within the linear regression framework.

While the idea is straightforward it does not appear to be discussed in the literature, so I hope that this paper can assist practitioners, by alerting them to a tool that can be implemented easily and usefully in a variety of widely applied regression models.

2 Linear Model

2.1 Specification Testing

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon, \tag{1}$$

[1] See also Hansen [1982] for applications to nonlinear models.

where $\mathbf{y}$ is an $N \times 1$ vector, $\mathbf{X}_1$ and $\mathbf{X}_2$ are $N \times n_1$ and $N \times n_2$ matrices of regressors with full column rank, with β_1 and β_2 being commensurate n_1- and n_2-vectors of regression coefficients, and ϵ an N-vector of disturbances satisfying $\mathbb{E}[\mathbf{X}_2'\epsilon] = \mathbf{0}$ and $\mathbb{E}[\mathbf{X}_1'\epsilon] \neq 0$, i.e. the regressors $\mathbf{X}_1$ are endogenous.

Also, suppose that $\mathbf{Z}$ is an $N \times m$ matrix of instruments for $\mathbf{X}_1$, with $m > n_1$, full rank m, and $\mathbb{E}[\mathbf{Z}'(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{X}_1]$ having full rank n_1, where $\mathbf{P}_{X_2} = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$. i.e. the order and rank conditions for identification of equation (1) are satisfied.[2] The maintained assumption is that a subset of n_1 columns of $\mathbf{Z}$ is uncorrelated with the structural regression errors ϵ. Furthermore, it is assumed that the elements of ϵ are mean zero and homoskedastic, conditional on $\mathbf{X}$ and $\mathbf{Z}$. The case of conditionally heteroskedastic errors is discussed in section 2.4 below.

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ denote the $N \times (n_1 + n_2)$ matrix of regressors, and $\mathbf{W} = [\mathbf{X}_2, \mathbf{Z}]$ the $N \times (n_2 + m)$ matrix of instruments. Also, let $\mathbf{P}_W = \mathbf{W}(\mathbf{W'W})^{-1}\mathbf{W}'$. For $\hat{\mathbf{X}}_1 = \mathbf{P}_W\mathbf{X}_1$ the fitted values of the first-stage regressions,

$$\begin{aligned} \mathbf{y} &= \hat{\mathbf{X}}_1\beta_1 + \mathbf{X}_2\beta_2 + \left(\mathbf{X}_1 - \hat{\mathbf{X}}_1\right)\beta_1 + \epsilon && (2) \\ &= \hat{\mathbf{X}}\beta + (\mathbf{I} - \mathbf{P}_W)\,\mathbf{X}_1\beta_1 + \epsilon && (3) \\ &= \hat{\mathbf{X}}\hat{\beta}_{2SLS} + \hat{\mathbf{X}}\left(\beta - \hat{\beta}_{2SLS}\right) + (\mathbf{I} - \mathbf{P}_W)\,\mathbf{X}_1\beta_1 + \epsilon && (4) \end{aligned}$$

where $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \mathbf{X}_2] = \mathbf{P}_W\mathbf{X}$ and $\hat{\beta}_{2SLS}$ denotes the two-stage least squares estimator for $\beta' = [\beta_1', \beta_2']$.

We follow the interpretation of Stock [2015] that, under the maintained assumption that at least n_1 valid instruments are available, one can test the null hypothesis of all instruments being valid, i.e. of the model being overidentified, against the alternative that up to $m - n_1$ instruments are invalid. Define the second-stage regression residuals

$$\begin{aligned} \hat{\epsilon} &= \mathbf{y} - \hat{\mathbf{X}}\hat{\beta}_{2SLS} && (5) \\ &= \hat{\mathbf{X}}\left(\beta - \hat{\beta}_{2SLS}\right) + (\mathbf{I} - \mathbf{P}_W)\,\mathbf{X}_1\beta_1 + \epsilon, && (6) \end{aligned}$$

and notice that

$$\begin{aligned} \hat{\epsilon} &= -\hat{\mathbf{X}}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_W\epsilon + (\mathbf{I} - \mathbf{P}_W)\,\mathbf{X}_1\beta_1 + \epsilon && (7) \\ &= \left(\mathbf{I} - \mathbf{P}_W\mathbf{X}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_W\right)\epsilon + (\mathbf{I} - \mathbf{P}_W)\,\mathbf{X}_1\beta_1. && (8) \end{aligned}$$

[2] The relevance condition implies that, in the reduced form system for $\mathbf{X}_1$, $\mathbf{X}_1 = \mathbf{X}_2\Pi_1 + \mathbf{Z}\Pi_2 + u$, Π_2 is identified, $\Pi_2 = \left(\mathbb{E}[\mathbf{Z}'(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{Z}]\right)^{-1}\mathbb{E}[\mathbf{Z}'(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{X}_1]$.

Therefore, a version of the Sargan test of the validity of the overidentifying restrictions in this model is based on the test statistic

$$S_N = \hat{\epsilon}'\mathbf{P}_W\hat{\epsilon} \tag{9}$$

$$= \epsilon'\left(\mathbf{P}_W - \mathbf{P}_W\mathbf{X}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_W\right)\epsilon. \tag{10}$$

Since the rank of the central matrix is equal to its trace, and its trace is equal to $m - n_1$, under the null hypothesis the statistic S_N is asymptotically distributed $\sigma_\epsilon^2\chi^2_{m-n_1}$, where σ_ϵ^2 is the conditional variance of the regression errors ϵ.

The Durbin-Wu-Hausman test examines the null hypothesis of the exogeneity of the regressors $\mathbf{X}_1$, against the alternative that these regressors are correlated with the regression error term. It is based on the OLS estimator of the n_1-vector γ in the regression

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \hat{\mathbf{U}}\gamma + \nu, \tag{11}$$

where $\hat{\mathbf{U}} = (\mathbf{I} - \mathbf{P}_W)\mathbf{X}_1$ are the residuals of the first-stage regressions, or so-called control functions. This regression can be interpreted as an "artificial regression" in the sense of Davidson and MacKinnon [1990, 1993, 2001][3] because under the null hypothesis of exogeneity the coefficient vector on the control functions $\gamma = \mathbf{0}$.[4] The Durbin-Wu-Hausman test therefore rejects the null hypothesis of exogeneity when $\hat{\gamma}$ is statistically significant. It is well known that the OLS estimator of β in this regression is identical to the two-stage least squares estimator $\hat{\beta}_{2SLS}$.

Now consider the expanded artificial regression

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \bar{\mathbf{Z}}\delta + \hat{\mathbf{U}}\gamma + \xi \tag{12}$$

$$= \mathbf{X}\beta + \bar{\mathbf{Z}}\delta + \hat{\mathbf{U}}\gamma + \xi, \tag{13}$$

where $\bar{\mathbf{Z}}$ is an arbitrary subset of $m - n_1$ columns of $\mathbf{Z}$. Under the null hypothesis that all overidentifying restrictions are valid, the $m - n_1$-vector $\delta = \mathbf{0}$. And if and only if the null hypothesis is true, the OLS estimator of β is equal to the two-stage least squares estimator and the OLS estimator of γ permits a Durbin-Wu-Hausman exogeneity test.[5] Incidentally,

[3]See Davidson and MacKinnon [1993], chapters 3.6 and 6.

[4]This can be seen from the reduced form system $\mathbf{X}_1 = \mathbf{X}_2\Pi_1 + \mathbf{Z}\Pi_2 + u$, $\mathbb{E}\left[\mathbf{X}'u\right] = \mathbf{0}$ and $\mathbb{E}[\mathbf{Z}'u] = \mathbf{0}$, so that $\mathbf{X}_1$ is correlated with ϵ if, and only if, $\mathbb{E}\left[\mathbf{X}_1'\epsilon|\mathbf{X},\mathbf{Z}\right] = \mathbb{E}\left[\Pi_1'\mathbf{X}_2'\epsilon + \Pi_2'\mathbf{Z}'\epsilon + u'\epsilon|\mathbf{X},\mathbf{Z}\right] = \mathbb{E}[\epsilon'u|\mathbf{X},\mathbf{Z}] = \gamma' \neq \mathbf{0}'$ a.s.

[5]Analogous to the argument related to (11), this follows from the reduced form for $\mathbf{X}_1$, orthogonality of $\mathbf{Z}$ and u - and hence, in particular, of $\bar{\mathbf{Z}}$ and u -, so that under the null hypothesis, $\mathbb{E}[\epsilon'u|\mathbf{X},\mathbf{Z}] = \mathbb{E}[\delta'\bar{\mathbf{Z}}'u + \gamma'u'u|\mathbf{X},\mathbf{Z}] = \gamma' \neq \mathbf{0}'$ a.s.

these considerations show that the exogeneity test is not independent of the validity of all the instruments used to implement the test.

Since $\hat{\mathbf{U}}$ is orthogonal to $\mathbf{W}$,

$$\mathbf{P}_W\mathbf{y} = \hat{\mathbf{X}}\beta + \bar{\mathbf{Z}}\delta + \mathbf{P}_W\xi. \tag{14}$$

Here, $\mathbf{P}_W\xi$, captures the exogenous part of the disturbances under the hypothesis that all instruments are valid. Define $\mathbf{P}_{\hat{X}} = \hat{\mathbf{X}}\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'$. Then,

$$\begin{aligned}\hat{\delta} &= \delta + \left(\bar{\mathbf{Z}}'\left(\mathbf{I} - \mathbf{P}_{\hat{X}}\right)\bar{\mathbf{Z}}\right)^{-1}\bar{\mathbf{Z}}'\left(\mathbf{I} - \mathbf{P}_{\hat{X}}\right)\mathbf{P}_W\xi && (15)\\ &= \delta + \left(\bar{\mathbf{Z}}'\left(\mathbf{I} - \mathbf{P}_W\mathbf{X}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_W\right)\bar{\mathbf{Z}}\right)^{-1} \\ &\quad\times \bar{\mathbf{Z}}'\left(\mathbf{I} - \mathbf{P}_W\mathbf{X}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_W\right)\mathbf{P}_W\xi. && (16)\end{aligned}$$

Therefore, under the null hypothesis, the statistic

$$\begin{aligned}\tilde{S}_N &= \hat{\delta}'\left(\bar{\mathbf{Z}}'\left(\mathbf{I} - \mathbf{P}_W\mathbf{X}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_W\right)\bar{\mathbf{Z}}\right)\hat{\delta} && (17)\\ &= \xi'\left(\mathbf{P}_W - \mathbf{P}_W\mathbf{X}\left(\mathbf{X}'\mathbf{P}_W\mathbf{X}\right)^{-1}\mathbf{X}'P_W\right)\xi && (18)\end{aligned}$$

has a $\sigma_\xi^2\chi^2_{m-n_1}$ distribution and thus $\tilde{S}_N/\hat{\sigma}_\xi^2$ is equivalent to the test statistic $S_N/\hat{\sigma}_\epsilon^2$, where $\hat{\sigma}^2$ denotes the squared standard error of the respective regression.[6]

Hence, the expanded artificial regression (13) implements the Durbin-Wu-Hausman exogeneity and Sargan overidentification tests as a single regression. In this regression, the null hypothesis that this paper focusses on is that $\delta = \mathbf{0}$, given that the hypothesis $\gamma = \mathbf{0}$ is rejected. As the Monte Carlo simulations in subsection 2.3 show, the first test does not significantly affect the size or power of the second test.

2.2 An Empirical Illustration

Table 1 provides an empirical example. It uses data provided by the statistical software Stata for the purpose of illustrating the Sargan test.[7] For the fifty US states, the data comprises rental rates for apartments (*rent*), next to housing values (*hsngval*) and the percentage of the state's population living in urban areas (*pcturban*). The housing values regressor is treated

[6]The test of the null hypothesis that $\delta = \mathbf{0}$ is typically implemented as an $F_{m-n_1,N-(n_2+m+1)}$ test. For large N, the squared standard error of the regression $\hat{\sigma}_\xi^2$ converges in probability to σ_ξ^2, so that this F-test is asymptotically equivalent to a $\chi^2_{m-n_1}$ test.

[7]The data can be downloaded from within Stata, using `webuse hsng2`.

as potentially endogenous in the regression of rents on housing values and the percentage of urban population at the state level. Median family income and 3 regional dummies - for the state's central, southern and western areas - are considered as instruments so that there are three over-identifying restrictions. The example shows that both the Sargan test and the test of the joint significance of $\bar{Z}$, the three regional dummies, reject the null hypothesis of the validity of the over-identifying restrictions.

2.3 Monte Carlo Simulation

The design of the Monte Carlo study follows Hahn and Hausman [2002] and Lee and Okui [2009]. The data is generated as

$$\begin{aligned} y_{1i} &= \beta \mathbf{z}_i' \pi + \mathbf{z}_i' \gamma + v_{1i} \\ y_{2i} &= \mathbf{z}_i' \pi + v_{2i}, \qquad\qquad i = 1, \cdots, n \end{aligned}$$

where

$$\begin{aligned} \mathbf{z}_i &\overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I}_K), \quad K = 5 \\ \begin{pmatrix} v_{1i} \\ v_{2i} \end{pmatrix} &\overset{i.i.d.}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \end{aligned}$$

Here, $\pi_k = \phi$, $k = 1, \cdots, K$, and ϕ is chosen such that the theoretical R^2 of the first-stage regression, $R^2 = \frac{K\phi^2}{K\phi^2+1}$, equals 0.01 and 0.2, respectively; these two cases capture situations with weak and strong instruments, respectively.[8] The parameter ρ is set to 0.5 and 0.9, to simulate cases of moderate and strong endogeneity. The parameter β is set such that the errors in the structural equation $y_{1i} = \beta y_{2i} + \epsilon_i$, i.e. $\epsilon_i = v_{1i} - \beta v_{2i}$, have unit variance. We investigate the size of specification tests when $\gamma = \mathbf{0}$, and their power when $\gamma_1 = 0.1$ and $\gamma_k = 0$, $k = 2, \cdots, K$. We run 2000 simulations, for sample sizes of $n = 250$ and $n = 1000$.

Table 2 summarises the results of our simulations. The size of the proposed test is generally slightly larger than the size of the conventional Sargan test, although the difference diminishes the stronger the simulated endogeneity is. The power of the proposed test, on the other hand, exceeds the one of the conventional Sargan test.[9] The results also show that both, the size and power of the proposed test, do not hinge on this being a second-stage test, following a first-stage Durbin-Wu-Hausman exogeneity test.

[8] The values of ϕ are 0.04495 and 0.22361.

[9] The reason for the strong power is that in the simulations the additional instruments $\bar{\mathbf{Z}}$ improve the fit of the model.

2.4 Extension to Heteroskedastic Models

Davidson et al. [1985] and Wooldridge [1995] discuss diagnostic tests based on the score or Lagrange multiplier principle that are heteroskedasticity robust.[10] Their approach can be adapted to the test proposed in this paper.

Let $\hat{\nu}$ denote the residuals of the regression (11). Note that this regression equation is simply (13), subject to the restriction that $\delta = 0$, i.e. subject to the null hypothesis of valid overidentifying restrictions. And define $\hat{\mathbf{R}}$ as the $N \times (n_2 + m)$ matrix of residuals of the regression of $\bar{\mathbf{Z}}$ onto $[\mathbf{X}, \hat{\mathbf{U}}]$. Then, regress an n-vector of ones onto $\hat{\nu} \cdot \hat{\mathbf{R}} = [\hat{\nu}_i \hat{R}_{ik}]_{i=1,\cdots,n;k=1,\cdots,n_2+m}$, without intercept, and retrieve the sum of squared residuals, SSR. Then, the test statistic $N - SSR$ is distributed $\chi^2_{m-n_1}$ under the null hypothesis that the overidentifying restrictions are valid. Appendix A provides a formal derivation.

3 Extension to Nonlinear Models

A nonlinear version of model (1) is given by

$$\mathbf{y} = \mathbf{x}(\beta) + \epsilon, \tag{19}$$

where $\mathbf{x}(\cdot)$ is a known, differentiable function of $\beta \in \mathbb{R}^{n_1+n_2}$. This function is the inverse link function in the class of Generalized Linear Models discussed in McCullagh and Nelder [1983] who also propose an estimation algorithm which amounts to an iterative weighted least squares procedure, a variant of the Newton-Raphson algorithm.

Endogeneity in the nonlinear model amounts to n_1 elements of $\mathbb{E}\left[\nabla_\beta \mathbf{x}(\beta)'\epsilon\right]$ being nonzero.[11]

Davidson and MacKinnon [1990, 1993, 2001] have shown how an "artificial regression", or Gauss-Newton regression, can be used to test the null hypothesis of exogeneity, i.e. the consistency of the nonlinear least squares (NLS) estimator $\hat{\beta}$, under the maintained hypothesis of a set of valid instruments $\mathbf{Z}$.

The NLS estimator solves

$$\mathbf{X}\left(\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{x}\left(\hat{\beta}\right)\right) = \mathbf{0}, \tag{20}$$

where $\mathbf{X}(\beta) = \nabla_\beta \mathbf{x}(\beta)$ is assumed to have full column rank in a neighborhood about the

[10] Wooldridge [2010] discusses a heteroskedastic robust version of the Durbin-Wu-Hausman test and the Sargan test as in 9. See also Windmeijer et al. [2018] and Chao et al. [2014].

[11] This can be thought of as $\beta' = (\beta_1', \beta_2')$, where $\beta_1 \in \mathbb{R}^{n_1}$ and $\beta_2 \in \mathbb{R}^{n_2}$, and $\mathbf{X}_1 = \nabla_{\beta_1}\mathbf{x}(\beta)$ satisfying $\mathbb{E}[\mathbf{X}_1'\epsilon] \neq \mathbf{0}$ at the true parameter vector β.

true population β.

As an analogue to the residual based exogeneity test in the linear model as implemented in (11), Davidson and MacKinnon [1993] propose the test of the null hypothesis of $\tau = \mathbf{0}$ in the regression

$$\mathbf{y} - \mathbf{x}\left(\hat{\beta}\right) = \mathbf{X}\left(\hat{\beta}\right)\alpha + (I - \mathbf{P}_W)\mathbf{X}^*\left(\hat{\beta}\right)\tau + \zeta, \tag{21}$$

where $\mathbf{X}^*$ are the $m - n_1$-columns of $\mathbf{X}$ that are not annihilated by the orthogonal projector $(I - \mathbf{P}_W)$ and $\mathbf{W} = [\mathbf{X}_2, \mathbf{Z}]$ is a set of $m + n_2$ instruments.[12] The contribution of $(I - \mathbf{P}_W)\mathbf{X}^*\left(\hat{\beta}\right)$ can again be viewed as a set of control functions. This is an artificial or Gauss-Newton regression because under the null hypothesis one would expect the least squares estimator of τ to be statistically insignificant.[13] The regressand in this Gauss-Newton regression is $\hat{\epsilon} = \mathbf{y} - \mathbf{x}\left(\hat{\beta}\right)$.

Now consider the instrumental variable estimator $\tilde{\beta}$ which satisfies

$$\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\left(\mathbf{y} - \mathbf{x}\left(\tilde{\beta}\right)\right) = \mathbf{0}. \tag{22}$$

The residuals induced by the IV estimator are $\tilde{\epsilon} = \mathbf{y} - \mathbf{x}\left(\tilde{\beta}\right)$. The Sargan test of the validity of over-identifying restrictions is[14]

$$\begin{aligned}
T_N &= \tilde{\epsilon}'\mathbf{P}_W\tilde{\epsilon} && (23)\\
&\approx \left(\mathbf{y} - \mathbf{x}(\beta) - \mathbf{X}(\beta)\left(\tilde{\beta} - \beta\right)\right)'\mathbf{P}_W\left(\mathbf{y} - \mathbf{x}(\beta) - \mathbf{X}(\beta)\left(\tilde{\beta} - \beta\right)\right) && (24)\\
&= \left(\left(I - \mathbf{X}(\beta)\left(\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\tilde{\beta}\right)\right)^{-1}\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\right)\epsilon\right)'\mathbf{P}_W\\
&\quad \times \left(\left(I - \mathbf{X}(\beta)\left(\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\tilde{\beta}\right)\right)^{-1}\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\right)\epsilon\right) && (25)\\
&= \epsilon'\left(\mathbf{P}_W - \mathbf{P}_W\mathbf{X}(\beta)\left(\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\tilde{\beta}\right)\right)^{-1}\mathbf{X}\left(\tilde{\beta}\right)'\mathbf{P}_W\right)\epsilon. && (26)
\end{aligned}$$

Under the null hypothesis, $\tilde{\beta}$ is consistent for β, and provided $\mathbf{X}(\cdot)$ is continuous, $\mathbf{X}(\tilde{\beta})$ tends to $\mathbf{X}(\beta)$ in large samples. Then, under the null hypothesis, T_N is asymptotically distributed $\chi^2_{m-n_1}$.

[12] Here, $\mathbf{X}_2 = \nabla_{\beta_2}\mathbf{x}(\beta)$, satisfying $\mathbb{E}[\mathbf{X}_2'\epsilon] = \mathbf{0}$.

[13] As in the linear case, the reduced form for $\mathbf{X}_1 = \nabla_{\beta_1}\mathbf{x}(\beta) = \nabla_{\beta_2}\mathbf{x}(\beta)\Pi_1 + \mathbf{Z}\Pi_2 + u$, $\mathbb{E}[\nabla_{\beta_1}\mathbf{x}(\beta)'\epsilon] = \mathbf{0}$ and $\mathbb{E}[\mathbf{Z}'\epsilon] = \mathbf{0}$, implies $\nabla_{\beta_1}\mathbf{x}(\beta)$ is endogenous if, and only if, $\mathbb{E}[\epsilon' u|\mathbf{X}(\beta), \mathbf{Z}] = \tau' \neq \mathbf{0}'$.

[14] In the approximation following the definition of T_N, we ignore higher-order terms.

Now consider an expanded Gauss-Newton regression,

$$\hat{\epsilon} = \mathbf{X}\left(\hat{\beta}\right)\alpha + \bar{\mathbf{Z}}\pi + (I - \mathbf{P}_W)\mathbf{X}^*\left(\hat{\beta}\right)\tau + \zeta, \tag{27}$$

where $\bar{Z}$ is an arbitrary subset of $m - n_1$ columns of $\mathbf{Z}$. Under the null hypothesis of exogeneity, just as in (13), one would expect the least squares estimates $\hat{\tau}$ to be statistically insignificant.[15] Also, analogous to (13), since

$$\mathbf{P}_W\hat{\epsilon} = \mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right) + \bar{\mathbf{Z}}\pi + \mathbf{P}_W\zeta, \tag{28}$$

it follows that

$$\begin{aligned}\hat{\pi} &= \pi + \left(\bar{\mathbf{Z}}'\left(I - \mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\left(\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\right)^{-1}\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\right)\bar{\mathbf{Z}}\right)^{-1} \\ &\quad \times\bar{\mathbf{Z}}'\left(I - \mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\left(\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\right)^{-1}\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\right)\zeta, \end{aligned} \tag{29}$$

a test statistic based on $\hat{\pi}$ satisfies

$$\tilde{T}_N = \hat{\pi}'\left(\bar{\mathbf{Z}}'\left(I - \mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\left(\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\right)^{-1}\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\right)\bar{\mathbf{Z}}\right)\hat{\pi} \tag{30}$$

$$= \zeta'\left(\mathbf{P}_W - \mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\left(\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\mathbf{X}\left(\hat{\beta}\right)\right)^{-1}\mathbf{X}\left(\hat{\beta}\right)'\mathbf{P}_W\right)\zeta. \tag{31}$$

Under the null hypothesis, $\hat{\beta}$ is consistent for β, and $\tilde{T}_N$ is distributed asymptotically $\sigma_\zeta^2\chi^2_{m-n_1}$.

Hence, again, the expanded artificial regression implements the exogeneity and overidentification test is a single regression.

4 Conclusions

This note presents a useful but not widely known framework for jointly implementing Durbin-Wu-Hausam exogeneity and Sargan-Hansen overidentification tests, as a single artificial regression. It covers linear models and discusses its extension to a class of non-linear models.

Future research might explore how to adapt this methodology to semi-parametric single index models [Horowitz, 2009] and quantile regression models in which the control function

[15] Given the orthogonality of $\bar{\mathbf{Z}}$ and u in the reduced form system, under the null hypothesis, $\mathbb{E}[\epsilon' u|\mathbf{X}(\beta), \mathbf{Z}] = \pi'\mathbb{E}[\bar{\mathbf{Z}}'u + \gamma' u' u|\mathbf{X}(\beta), \mathbf{Z}] = \tau' \neq \mathbf{0}$.

approach is already widely employed [Blundell and Powell, 2004, Lee, 2007].

References

Richard W. Blundell and James L. Powell. Endogeneity in Semiparametric Binary Response Models. *The Review of Economics Studies*, 71(3):655–679, 2004.

John C Chao, Jerry A Hausman, Whitney K Newey, Norman R Swanson, and Tiemen Woutersen. Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of econometrics*, 178:15–21, 2014.

Russell Davidson and James G. MacKinnon. Specification Tests Based on Artificial Regressions. *Journal of the American Statistical Association*, 85(409):220–227, 1990.

Russell Davidson and James G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1993.

Russell Davidson and James G. MacKinnon. Artificial Regressions. Queen's Economics Department Working Paper No. 1038,, 2001.

Russell Davidson, James G MacKinnon, and Russel Davidson. Heteroskedasticity-robust tests in regressions directions. In *Annales de l'INSEE*, pages 183–218. JSTOR, 1985.

James Durbin. Errors in Variables. *Review of the International Statistical Institute*, 22(1/3): 23–32, 1954.

Jinyong Hahn and Jerry Hausman. A new specification test for the validity of instrumental variables. *Econometrica*, 70(1):163–189, 2002.

Lars P. Hansen. Large Sample Properties of Generalized Method of Moments Estimators. 1982.

Jerry A. Hausman. Specification Tests in Econometrics. *Econometrica*, 46(6):1251–1271, 1978.

Jerry A. Hausman. *Handbook of Econometrics, Vol.1*, chapter Specification and estimation of simultaneous equation models, pages 391–448. North Holland, 1983.

Joel Horowitz. *Semiparametric and Nonparametric Methods in Econometrics*. Springer Verlag, 2009.

Sokbae Lee. Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics*, 141(2):1131–1158, 2007.

Yoonseok Lee and Ryo Okui. A specification test for instrumental variables regression with many instruments. 2009.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1983.

John D. Sargan. The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica*, 26(3):393–415, 1958.

John D. Sargan. *Contributions to Econometrics: John Denis Sargan, vol. I*, chapter Testing for misspecification after estimating using instrumental variables, pages 213–235. Cambridge University Press, 1988.

James H Stock. Instrumental variables in statistics and econometrics. 2015.

Halbert White et al. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, 48(4):817–838, 1980.

Frank Windmeijer et al. Testing over-and underidentification in linear models, with applications to dynamic panel data and asset-pricing models. *University of Bristol Department of Economics Working Paper*, 2018.

Jeffrey M Wooldridge. Score diagnostics for linear models estimated by two stage least squares. *Advances in econometrics and quantitative economics: Essays in honor of Professor CR Rao*, pages 66–87, 1995.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

De-Min Wu. Alternative Tests of Independence between Stochastic Regressors and Disturbances. *Econometrica*, 41(4):733–750, 1973.

A Heteroskedasticity Robust Test

Assume the errors in (13) are i.i.d. normal, with mean zero and heteroskedastic variances, conditional on $\mathbf{X}$ and $\mathbf{Z}$.[16] Then, under the null hypothesis $\delta = 0$,

$$\mathbf{y}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} = \nu'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}}, \tag{32}$$

with mean zero and variance-covariance matrix $\bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\Omega(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}}$, where Ω is a diagonal matrix when the errors are conditionally heteroskedastic. Therefore, under the assumption of normality, the statistic

$$L = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} \left[\bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\Omega(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}}\right]^{-1} \bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\mathbf{y} \tag{33}$$

is distributed $\chi^2_{m-n_1}$. Using results in Davidson et al. [1985] and White et al. [1980], it can be implemented by replacing Ω with $\hat{\Omega} = \text{diag}(\hat{\nu}_i)$, where $\hat{\nu}$ is the vector of residuals of regression (11). Let $\hat{L}$ denote this feasible test statistic.

Now consider the regression of an n-vector of ones onto $\hat{\nu} \cdot \hat{\mathbf{R}}$. Notice, first, that

$$\begin{aligned} \hat{\nu} \cdot \hat{\mathbf{R}} &= \text{diag}(\hat{\nu}_i)(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} && (34)\\ &= \hat{\nu}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} && (35)\\ &= \mathbf{y}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} && (36) \end{aligned}$$

Therefore, the resulting sum of squared residuals is

$$\begin{aligned} SSR &= N - \hat{\nu}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} \left[\bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\text{diag}(\hat{\nu}_i)(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}}\right]^{-1} \bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\hat{\nu} && (37)\\ &= N - \mathbf{y}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}} \left[\bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\text{diag}(\hat{\nu}_i)(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\bar{\mathbf{Z}}\right]^{-1} \bar{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_{X,\hat{U}})\mathbf{y} && (38)\\ &= N - \hat{L}. && (39) \end{aligned}$$

Therefore, $\hat{L} = N - SSR$.

B Tables

[16] Absent the assumption of normality, the derivation of the distribution of the test statistic holds asymptotically, under assumptions as in White et al. [1980].

Table 1: Example

	2SLS[a]	DWH[c]	Expanded[c]
	rent	rent	rent
hsngval	0.00224***	0.00224***	0.00387***
	(6.82)	(8.36)	(9.64)
pcturban	0.0815	0.0815	-0.498*
	(0.27)	(0.33)	(-2.15)
$\hat{U}$		-0.00159***	-0.00322***
		(-3.99)	(-6.86)
2.region			1.529
			(0.23)
3.region			7.743
			(1.14)
4.region			-40.61***
			(-4.62)
constant	120.7***	120.7***	88.27***
	(7.93)	(9.71)	(6.22)
Test	Sargan[d]		F-test[e]
p-value	0.00103		0.0002
N	50	50	50
R^2	0.599	0.754	0.845

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes:

[a] 2SLS: hsngval instrumented by family income and 3 region dummies.

[b] Durbin-Wu-Hausman regression.

[c] Expanded artificial regression, as in equations (12) and (13).

[d] The Sargan test statistic has a χ^2_3 distribution.

[e] The test statistic has an $F_{3,43}$ distribution.

Table 2: Monte Carlo Simulations

n	ρ	Instruments	Sargan	Test[a]	Test[a], given DWH[b] rejects H_0
			Size[c]		
250	0.5	strong	0.058	0.126	0.124
250	0.9	strong	0.051	0.062	0.068
250	0.5	weak	0.055	0.086	0.083
250	0.9	weak	0.041	0.058	0.061
1000	0.5	strong	0.061	0.146	0.146
1000	0.9	strong	0.051	0.066	0.068
1000	0.5	weak	0.076	0.121	0.106
1000	0.9	weak	0.046	0.068	0.059
			Power[d]		
250	0.5	strong	0.947	0.998	0.998
250	0.9	strong	1.000	1.000	1.000
250	0.5	weak	0.725	0.985	0.977
250	0.9	weak	0.789	0.977	0.997
1000	0.5	strong	1.000	1.000	1.000
1000	0.9	strong	1.000	1.000	1.000
1000	0.5	weak	0.888	1.000	1.000
1000	0.9	weak	0.925	1.000	1.000

Notes:

2000 Simulation sample draws. The nominal size of all tests is 0.05.

[a] Proposed test of null hypothesis that $\delta = 0$.

[b] Durbin-Wu-Hausman regression based exogeneity test.

[c] Here, $\gamma = \mathbf{0}$.

[d] Here, $\gamma_1 = 0.1$ and $\gamma_k = 0$, $k = 2, \cdots, 5$.