# A Note on Specification Testing in Some Structural Regression Models*

Walter Beckert†

August 19, 2018

**Abstract**

There is a useful but not widely known 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.

Word count: 2130.
JEL classification: C21, C26, C36.
Keywords: endogeneity, identification, testing, artificial regression.

*I thank the editor, an anonymous referee, as well as Haris Psaradakis, Ron Smith and Jouni Sohkanen for insightful 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, in a *single* linear regression and under the maintained assumption of the validity of just identifying instruments, following a first-stage regression (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.

# 2 Linear Model

Consider the linear regression model

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

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 $\beta_1$ and $\beta_2$ being commensurate $n_1$- and $n_2$-vectors of regression coefficients, and $\epsilon$ an $N$-vector of mean zero and homoskedastic 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

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

$m$, and $\mathbb{E}[\mathbf{Z}'\mathbf{X}_1]$ having full rank $n_1$, i.e. the order and rank conditions for identification of equation (1) are satisfied. The maintained assumption is that a subset of $n_1$ columns of $\mathbf{Z}$ is uncorrelated with the structural regression errors $\epsilon$.

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 $P_W = \mathbf{W}(\mathbf{W'W})^{-1}\mathbf{W}'$. For $\hat{\mathbf{X}}_1 = P_W\mathbf{X}_1$ the fitted values of the first-stage regressions,

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

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

Define the second-stage regression residuals

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

and notice that

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

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

$$\begin{aligned}
S_N &= \hat{\epsilon}'P_W\hat{\epsilon} && (9)\\
&= \epsilon'\left(P_W - P_W\mathbf{X}\left(\mathbf{X}'P_W\mathbf{X}\right)^{-1}\mathbf{X}'P_W\right)\epsilon. && (10)
\end{aligned}$$

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 $\sigma_\epsilon^2$ is the variance of the regression errors $\epsilon$.

A version of the Durbin-Wu-Hausman test of the exogeneity of the regressors $\mathbf{X}_1$ is based on the OLS estimator of the $n_1$-vector $\gamma$ in the regression

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

where $\hat{\mathbf{U}} = (\mathbf{I} - P_W)\,\mathbf{X}_1$ are the residuals of the first-stage regressions, or so-called control functions. It is well known that the OLS estimator of $\beta$ in this regression is identical to the two-stage least squares estimator $\hat{\beta}_{2SLS}$. This regression can be interpreted as an "artificial regression" in the sense of Davidson and MacKinnon [1990, 1993, 2001] because under the null hypothesis of exogeneity we expect the estimator of $\gamma$, the coefficient vector on the control functions, to be indistinguishable from the zero vector.

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 $\beta$ is equal to the two-stage least squares estimator and the OLS estimator of $\gamma$ permits a Durbin-Wu-Hausman exogeneity test. Incidentally, 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 $\mathbf{W}$,

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

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

$$\hat{\delta} = \delta + \left(\bar{\mathbf{Z}}'\,(\mathbf{I} - P_{\hat{X}})\,\bar{\mathbf{Z}}\right)^{-1}\bar{\mathbf{Z}}'\,(\mathbf{I} - P_{\hat{X}})\,P_W\xi \tag{15}$$

$$\begin{aligned} &= \delta + \left(\bar{\mathbf{Z}}'\left(\mathbf{I} - P_W\mathbf{X}\,(\mathbf{X}'P_W\mathbf{X})^{-1}\,\mathbf{X}'P_W\right)\bar{\mathbf{Z}}\right)^{-1} \\ &\quad \times\ \bar{\mathbf{Z}}'\left(\mathbf{I} - P_W\mathbf{X}\,(\mathbf{X}'P_W\mathbf{X})^{-1}\,\mathbf{X}'P_W\right)P_W\xi. \end{aligned} \tag{16}$$

Therefore, under the null hypothesis, the statistic

$$\tilde{S}_N = \hat{\delta}'\left(\bar{\mathbf{Z}}'\left(\mathbf{I} - P_W\mathbf{X}\,(\mathbf{X}'P_W\mathbf{X})^{-1}\,\mathbf{X}'P_W\right)\bar{\mathbf{Z}}\right)\hat{\delta} \tag{17}$$

$$= \xi'\left(P_W - P_W\mathbf{X}\,(\mathbf{X}'P_W\mathbf{X})^{-1}\,\mathbf{X}'P_W\right)\xi \tag{18}$$

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.[2]

Hence, the expanded artificial regression (13) implements the Durbin-Wu-Hausman exogeneity and Sargan overidentification tests as a single regression.

Table 1 provides an empirical example. It uses data provided by the statistical software Stata for the purpose of illustrating the Sargan test.[3] 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 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.

# 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 $\mathbb{E}[\nabla_\beta \mathbf{x}(\beta)'\epsilon]$ being nonzero.[4]

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 $\mathbf{Z}$.

[2] 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 $\hat{\sigma}^2_\xi$ converges in probability to $\sigma^2_\xi$, so that this $F$-test is asymptotically equivalent to a $\chi^2_{m-n_1}$ test.

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

[4] 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 $\beta$.

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 true population $\beta$.

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

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

where $\mathbf{X}^*$ are the $m-n_1$-columns of $\mathbf{X}$ that are not annihilated by the orthogonal projector $(I-P_W)$ and $\mathbf{W}=[\mathbf{X}_2,\mathbf{Z}]$ is a set of $m+n_2$ instruments.[5] The contribution of $(I-P_W)\mathbf{X}^*\left(\hat{\beta}\right)$ 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 $\tau$ to be statistically insignificant. 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

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

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[6]

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

Under the null hypothesis, $\tilde{\beta}$ is consistent for $\beta$, and provided $\mathbf{X}(\cdot)$ is continuous, $\mathbf{X}(\tilde{\beta})$ tends

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

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

to $\mathbf{X}(\beta)$ in large samples. Then, under the null hypothesis, $T_N$ is asymptotically distributed $\chi^2_{m-n_1}$.

Now consider an expanded Gauss-Newton regression,

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

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

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

it follows that

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

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

$$\begin{aligned}
\tilde{T}_N &= \hat{\pi}'\left(\bar{\mathbf{Z}}'\left(I - P_W\mathbf{X}\left(\hat{\beta}\right)\left(\mathbf{X}\left(\hat{\beta}\right)'P_W\mathbf{X}\left(\hat{\beta}\right)\right)^{-1}\mathbf{X}\left(\hat{\beta}\right)'P_W\right)\bar{\mathbf{Z}}\right)\hat{\pi} \qquad (30) \\
&= \zeta'\left(P_W - P_W\mathbf{X}\left(\hat{\beta}\right)\left(\mathbf{X}\left(\hat{\beta}\right)'P_W\mathbf{X}\left(\hat{\beta}\right)\right)^{-1}\mathbf{X}\left(\hat{\beta}\right)'P_W\right)\zeta. \qquad (31)
\end{aligned}$$

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

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 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.

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.

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

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.

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.

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

# A Tables

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 $\chi^2_3$ distribution.

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