Discussion Papers
Department of Economics
University of Copenhagen

No. 11-25

The Properties of Model Selection when Retaining Theory Variables

David F. Hendry, Søren Johansen

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

The Properties of Model Selection when Retaining Theory Variables

David F. Hendry[†] and Søren Johansen[⋆1]
[†]Economics Department and Institute for New Economic Thinking at the
Oxford Martin School, University of Oxford, UK
[⋆]Economics Department, University of Copenhagen and CREATES, Aarhus University, Denmark

JEL classifications: C521, C18.
KEYWORDS: Model selection, theory retention.

1. Introduction

Economic theories are often fitted directly to data to avoid possible model selection biases. This is an excellent strategy when the theory is complete and correct, but less successful otherwise. We show that embedding a theory model that specifies the correct set of m relevant exogenous variables, $\mathbf{x}_t$, within the larger set of $m+k$ candidate variables, $(\mathbf{x}_t, \mathbf{w}_t)$, then selection over the second set by their statistical significance can be undertaken without affecting the estimator distribution of the theory parameters. This strategy returns the theory-parameter estimates when the theory is correct, yet protects against the theory being under-specified because some $\mathbf{w}_t$ are relevant.

Section 2 shows that the distributions of the estimated coefficients of $\mathbf{x}_t$ are unaffected by model selection when the variables $\mathbf{w}_t$ are orthogonalized with respect to $\mathbf{x}_t$, for $(k+m) << T$, so the general model is estimable. Section 3 establishes that the same results apply even when $(k+m) > T$, provided $m << T$. Section 4 concludes. The appendix section 5 extends the analysis to a valid theory with endogenous variables and §5.1 notes how to assess the validity of the instrumental variables.

2. Selection when retaining a valid theory

Consider a theory model which correctly matches the data-generating process (DGP) by specifying over $t = 1, \ldots, T$ that:

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \epsilon_t \tag{1}$$

where $\epsilon_t \sim \mathsf{IID}[0, \sigma_\epsilon^2]$, and ϵ_t is independent of the m strongly exogenous variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$, assumed to satisfy:

$$T^{-1} \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \xrightarrow{P} \boldsymbol{\Sigma}_{xx}$$

which is positive definite, and:

$$T^{1/2} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) = \left(T^{-1} \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} T^{-1/2} \sum_{t=1}^{T} \mathbf{x}_t \epsilon_t \xrightarrow{D} \mathsf{N}_m \left[\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}_{xx}^{-1} \right] \tag{2}$$

where $\boldsymbol{\beta}_0$ is the constant population parameter.

However, an investigator may be willing to contemplate the possibility that an additional set of k exogenous variables $\mathbf{w}_t$ also influences y_t, so postulates the more general model:

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \epsilon_t \tag{3}$$

although in fact $\boldsymbol{\gamma}_0 = \mathbf{0}$. The $\mathbf{w}_t$ can be variables known to be exogenous, functions of those, lagged variables in time series, and indicators for outliers or breaks, and we assume the same assumptions as above for $\{\epsilon_t, \mathbf{x}_t, \mathbf{w}_t\}$. The investigator regards the theory in (1) as correct and complete, so wishes to ensure that the $\mathbf{x}_t$ are always retained and not selected over. The issue we address is the possible additional cost of searching over the candidate variables $\mathbf{w}_t$ in (3) when retaining the $\mathbf{x}_t$, rather than directly estimating (1) when $(k+m) << T$.

The $\mathbf{x}_t$ and $\mathbf{w}_t$ can be orthogonalized by first computing:

$$\widehat{\boldsymbol{\Gamma}} = \left(\sum_{t=1}^{T} \mathbf{w}_t \mathbf{x}_t' \right) \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right)^{-1}$$

and defining the residuals $\widehat{\mathbf{u}}_t$ by:

$$\mathbf{w}_t = \widehat{\boldsymbol{\Gamma}} \mathbf{x}_t + \widehat{\mathbf{u}}_t \tag{4}$$

so that:

$$\sum_{t=1}^{T} \mathbf{x}_t \widehat{\mathbf{u}}_t' = \mathbf{0} \tag{5}$$

Using (4) in (3):

$$\begin{aligned} y_t &= \boldsymbol{\beta}' \mathbf{x}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \epsilon_t = \boldsymbol{\beta}' \mathbf{x}_t + \boldsymbol{\gamma}' \left(\widehat{\boldsymbol{\Gamma}} \mathbf{x}_t + \widehat{\mathbf{u}}_t \right) + \epsilon_t \\ &= \boldsymbol{\beta}_+' \mathbf{x}_t + \boldsymbol{\gamma}' \widehat{\mathbf{u}}_t + \epsilon_t, \end{aligned} \tag{6}$$

where $\boldsymbol{\beta}_+ = \boldsymbol{\beta} + \widehat{\boldsymbol{\Gamma}}' \boldsymbol{\gamma}$. Note that $\boldsymbol{\beta}_{0+} = \boldsymbol{\beta}_0$ because $\boldsymbol{\gamma}_0 = \mathbf{0}$.

Consequently, as (1) is the DGP, by orthogonality from (5):

$$
\begin{aligned}
&T^{1/2}\begin{pmatrix}\widetilde{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}_0\\ \widetilde{\boldsymbol{\gamma}}\end{pmatrix}\\
&= \begin{pmatrix} T^{-1}\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t' & T^{-1}\sum_{t=1}^T \mathbf{x}_t\widehat{\mathbf{u}}_t' \\ T^{-1}\sum_{t=1}^T \widehat{\mathbf{u}}_t\mathbf{x}_t' & T^{-1}\sum_{t=1}^T \widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_t' \end{pmatrix}^{-1}\\
&\times \begin{pmatrix} T^{-1/2}\sum_{t=1}^T \mathbf{x}_t\epsilon_t \\ T^{-1/2}\sum_{t=1}^T \widehat{\mathbf{u}}_t\epsilon_t \end{pmatrix}\\
&= \begin{pmatrix} \left(T^{-1}\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right)^{-1} T^{-1/2}\sum_{t=1}^T \mathbf{x}_t\epsilon_t \\ \left(T^{-1}\sum_{t=1}^T \widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_t'\right)^{-1} T^{-1/2}\sum_{t=1}^T \widehat{\mathbf{u}}_t\epsilon_t \end{pmatrix}\\
&\xrightarrow{D} \mathsf{N}_{m+k}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \sigma_\epsilon^2\begin{pmatrix}\boldsymbol{\Sigma}_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{ww|x}^{-1}\end{pmatrix}\right] \qquad (7)
\end{aligned}
$$

Thus, the estimator $\widetilde{\boldsymbol{\beta}}_+$ in (7) is identical to $\widehat{\boldsymbol{\beta}}$ in (2), independently of the inclusion or exclusion of any or all of the $\widehat{\mathbf{u}}_t$. Even after selection over the $\widehat{\mathbf{u}}_t$ at significance level α, and corresponding critical value c_α, say, by sequential t-tests on each $\widetilde{\gamma}_i$, the theory-parameter estimator is unaffected by retaining significant $\widehat{\mathbf{u}}_t$. For a Gaussian distribution and fixed regressors, the estimator $\widetilde{\boldsymbol{\beta}}_+ = \widehat{\boldsymbol{\beta}}$ is statistically independent of the test statistics used to select.

The possible costs of selection are:
(a) chance retention by selection of some $\widehat{u}_{i,t}$, which may mislead on the validity of the theory model; and
(b) their impact on the estimated *distribution* of $\widehat{\boldsymbol{\beta}}$, through misestimation of σ_ϵ^2 in (7).
Against these, possible benefits are:
(c) the theory-model is tested against a wide range of alternatives; and
(d) when the theory is incomplete, the selected model will be less mis-specified.
For (a), if all $\widehat{u}_{i,t}$ are irrelevant, then on average αk of the $\widehat{u}_{i,t}$ will be retained by chance, with estimated coefficient $\widetilde{\gamma}_i$, where:

$$
\left|\mathsf{t}_{\gamma_i=0}\right| = \frac{|\widetilde{\gamma}_i|}{\mathsf{SE}\left[\widetilde{\gamma}_i\right]} \geq c_\alpha \qquad (8)
$$

Setting $\alpha = \min\left[1/k, 1/T, 1\%\right]$ is an appealing rule. When $T = 100$ and $k = T/4 = 25$, say, then because $k\alpha = 0.25$, the probability of retaining more than one irrelevant variable is:

$$
\mathsf{p}_1 = 1 - \sum_{i=0}^{1} \frac{(0.25)^i}{i!} e^{-0.25} \simeq 2.6\%.
$$

Moreover, under normality and letting $h > 2/c_\alpha$ then:

$$
\Pr\left(\left|\mathsf{t}_{\gamma_i=0}\right| \geq h c_\alpha \mid \mathsf{H}_0\right) \leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{h^2}{2}c_\alpha^2\right)
$$

which is 0.01% at $h = 1.5$ and $c_{0.01} = 2.65$. Thus, it is unlikely any $\left|\mathsf{t}_{\gamma_i=0}\right|$ will be larger than $1.5c_\alpha$. Problem (a) can be resolved by rejecting a theory when more than one of the $\widehat{u}_{i,t}$ are retained, or when one is more significant than $1.5c_\alpha$.

Addressing (b), an unbiased estimated error variance under the null that $\boldsymbol{\gamma}_0 = \mathbf{0}$, so that (2) is correctly estimated, is:

$$
\widetilde{\sigma}_\epsilon^2 = (T-m)^{-1}\sum_{t=1}^T \left(y_t - \widetilde{\boldsymbol{\beta}}_+'\mathbf{x}_t\right)^2 \qquad (9)
$$

although under the alternative, (9) will be an overestimate. Estimates of γ_i can be approximately bias corrected if desired after their chance retention, as in Hendry and Krolzig (2005).

The converse to (a) is (c), as the theory-model is tested simultaneously against all $\mathbf{w}_t$, and if incomplete as in (d), selection will reduce mis-specification relative to direct estimation.

2.1. Retaining an incomplete or invalid theory

Under the alternative $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, directly estimating (1) will result in biased outcomes. However, when (3) nests the DGP, from (6) the coefficient of $\mathbf{x}_t$ is $\boldsymbol{\beta}_0 + \widehat{\boldsymbol{\Gamma}}'\boldsymbol{\gamma}_0$, which will also be estimated if (1) is directly fitted to the data. When (3) nests the DGP, selection can improve the final model relative to (1), as in Castle, Doornik and Hendry (2011). While retaining $\mathbf{x}_t$ when selecting from (6) will then deliver an incorrect estimate of $\boldsymbol{\beta}_0$, some of the $\widehat{u}_{i,t}$ will also be retained, this time correctly, but an estimate of $\boldsymbol{\beta}_0$ can be derived from $\widetilde{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Gamma}}'\widetilde{\boldsymbol{\gamma}}$, $\widetilde{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$.

If the theory is completely incorrect in that $\boldsymbol{\beta}_0 = \mathbf{0}$, the estimated coefficient $\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Gamma}}'\widehat{\boldsymbol{\gamma}}$ of $\mathbf{x}_t$ in (6) will generally not be zero, so it may be worth also selecting without orthogonalization when estimates of $\boldsymbol{\beta}_0$ do not conform to theory expectations.

3. More candidate variables than observations

The analytic approach in Johansen and Nielsen (2009) to understanding impulse-indicator saturation (IIS) also applies for $k = T$ IID mutually-orthogonal candidate regressors under the null. Add the first $k/2$ of the variables and select at significance level $\alpha = 1/T = 1/k$. Record which are significant, then drop them. Now add the second block of $k/2$, again selecting at significance level $\alpha = 1/k$, and record which are significant in that subset. Finally, combine the recorded variables from the two stages (if any), and select again at significance level $\alpha = 1/k$. At both sub-steps, on average $\alpha k/2 = 1/2$ a variable will be retained by chance, so on average $\alpha k = 1$ will be retained from the combined stage. Under the null, one degree of freedom is lost on average. A combination of expanding and contracting block searches is implemented in (e.g.) *Autometrics* (see Doornik, 2009, and Doornik and Hendry, 2009)

If the model also has relevant variables to be retained, so $k + m = N > T$, orthogonalize the relevant variables with respect to the other candidates as above, but in blocks: under the null, doing so has no impact on the coefficients of the relevant variables, or the estimates. When $N > T$, divide the k variables into sub-blocks of smaller than $T/4$ (say), setting $\alpha = 1/N$ overall. The selected model retains the desired sub-set of m theory-based variables at every stage, and only selects over the putative irrelevant variables at a stringent significance level.

4. Conclusion

Model selection has had numerous critics from 'data mining' in Lovell (1983) through Leeb and Pötscher (2005). Yet the key implication of the above analysis is that it is almost costless to check large numbers of candidate exogenous variables when retaining a theory-based specification. The retention of the theory variables ensures that there is no selection over the parameters of interest, so that the distribution of their estimates is unaffected by selection over the orthogonalized set of candidates. Under the null that all those candidates are irrelevant, the parameters of interest are unaffected by the reparametrization and therefore by selection.

Conversely, there are substantial benefits if the initial specification is incorrect, but the enlarged model nests the data generation process. Thus, this variant of model selection is either costless or beneficial, even with endogenous variables and when there are more potential variables than observations.

Acknowledgments

The first author was supported in part by grants from the Open Society Institute and the Oxford Martin School, and the second author by the Center for Research in Econometric Analysis of Time Series (CREATES, funded by the Danish National Research Foundation). We are grateful to Jennifer L. Castle, David Cox, Jurgen Doornik, Neil R. Ericsson, Grayham E. Mizon and Bent Nielsen for helpful discussions.

References

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, **3 (1)**, DOI: 10.2202/1941–1928.1097.

Castle, J. L., and Shephard, N.(eds.)(2009). *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

Doornik, J. A. (2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

Doornik, J. A., and Hendry, D. F. (2009). *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.

Durbin, J. (1954). Errors in variables. *Review of the Institute of International Statistics*, **22**, 23–54.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46**, 1251–1271.

Hendry, D. F. (2011). On adding over-identifying instrumental variables to simultaneous equations. *Economics Letters*, **111**, 68–70.

Hendry, D. F., and Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.

Hendry, D. F., and Santos, C. (2010). An automatic test of super exogeneity. In Watson, M. W., Bollerslev, T., and Russell, J.(eds.), *Volatility and Time Series Econometrics*, pp. 164–193. Oxford: Oxford University Press.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, and Shephard (2009), pp. 1–36.

Leeb, H., and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21**, 21–59.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.

Wu, D. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, **41**, 733–750.

5. Appendix: Retaining a valid theory with endogenous variables

When some of the right-hand side variables are potentially endogenous, the theory model is still:

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \epsilon_t \tag{10}$$

where $\mathbf{x}_t$ is $m \times 1$, and $\epsilon_t \sim \mathsf{IID}[0, \sigma_\epsilon^2]$, but now ϵ_t is independent of the $n \geq m$ instrumental variables $\mathbf{z}_1, \ldots, \mathbf{z}_t$ where $(m+n) < T$. The partial DGP for the variables $(y_t, \mathbf{x}_t)$ given $\mathbf{z}_t$ has the form:

$$\begin{aligned} y_t &= \boldsymbol{\beta}' \boldsymbol{\Pi} \mathbf{z}_t + \eta_t \\ \mathbf{x}_t &= \boldsymbol{\Pi} \mathbf{z}_t + \boldsymbol{\xi}_t \end{aligned}$$

where $(\eta_t, \boldsymbol{\xi}_t)$ are $\mathsf{IID}[\mathbf{0}, \boldsymbol{\Omega}]$ with:

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_\eta^2 & \boldsymbol{\sigma}'_{\eta\xi} \\ \boldsymbol{\sigma}_{\xi\eta} & \boldsymbol{\Omega}_\xi \end{pmatrix}$$

and $(\eta_t, \boldsymbol{\xi}_t)$ is independent of $\mathbf{z}_1, \ldots, \mathbf{z}_t$, but $\epsilon_t = y_t - \boldsymbol{\beta}' \mathbf{x}_t = \eta_t - \boldsymbol{\beta}' \boldsymbol{\xi}_t$ is correlated with $\mathbf{x}_t$ as $\mathsf{Cov}[\mathbf{x}_t \epsilon_t] = \boldsymbol{\sigma}_{\xi\eta} - \boldsymbol{\Omega}_\xi \boldsymbol{\beta}$.

Then instrumental variables estimation of (10) coincides with two-stage least squares (2SLS) and delivers:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \left[\left(\sum_{t=1}^T \mathbf{x}_t \mathbf{z}'_t \right) \left(\sum_{t=1}^T \mathbf{z}_t \mathbf{z}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{z}_t \mathbf{x}'_t \right) \right]^{-1} \\ \times \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{z}'_t \right) \left(\sum_{t=1}^T \mathbf{z}_t \mathbf{z}'_t \right)^{-1} \sum_{t=1}^T \mathbf{z}_t \epsilon_t \end{aligned} \tag{11}$$

so that:

$$T^{1/2} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) \xrightarrow{D} \mathsf{N}_m \left[\mathbf{0}, \sigma_\epsilon^2 \mathbf{Q}^{-1} \right] \tag{12}$$

where we assume:

$$\mathbf{Q} = \operatorname*{plim}_{T \to \infty} \left[\left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{z}'_t \right) \left(\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{z}'_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}'_t \right) \right]$$

is positive definite. Let:

$$\widehat{\boldsymbol{\Pi}} = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{z}'_t \right) \left(\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{z}'_t \right)^{-1}$$

and define:

$$\widehat{\mathbf{x}}_t = \widehat{\boldsymbol{\Pi}} \mathbf{z}_t \text{ with } \widehat{\boldsymbol{\xi}}_t = \mathbf{x}_t - \widehat{\mathbf{x}}_t = \left(\boldsymbol{\Pi} - \widehat{\boldsymbol{\Pi}} \right) \mathbf{z}_t + \boldsymbol{\xi}_t,$$

then a 2SLS reformulation that is algebraically convenient is:

$$y_t = \boldsymbol{\beta}' \widehat{\mathbf{x}}_t + e_t \tag{13}$$

where:

$$e_t = \epsilon_t + \boldsymbol{\beta}' \widehat{\boldsymbol{\xi}}_t = \eta_t + \boldsymbol{\beta}' (\boldsymbol{\xi}_t - \widehat{\boldsymbol{\xi}}_t)$$

so that:

$$\operatorname*{plim}_{T \to \infty} \frac{1}{T} \sum_{t=1}^T \widehat{\mathbf{x}}_t e_t = \operatorname*{plim}_{T \to \infty} \widehat{\boldsymbol{\Pi}} \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \left(\eta_t + \boldsymbol{\beta}' \left(\boldsymbol{\xi}_t - \widehat{\boldsymbol{\xi}}_t \right) \right) = \mathbf{0}$$

When an investigator includes an additional set of k candidate exogenous variables $\mathbf{w}_t$, consider the partial DGP:

$$\begin{aligned} y_t &= \boldsymbol{\beta}' \boldsymbol{\Pi} \mathbf{z}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \eta_t \\ \mathbf{x}_t &= \boldsymbol{\Pi} \mathbf{z}_t + \boldsymbol{\xi}_t \end{aligned} \tag{14}$$

where $\boldsymbol{\gamma}_0 = \mathbf{0}$, and the $\mathbf{x}_t$ are retained. Since $\boldsymbol{\gamma}_0 = \mathbf{0}$, when the $\widehat{\mathbf{x}}_t = \widehat{\boldsymbol{\Pi}} \mathbf{z}_t$ and $\mathbf{w}_t$ are orthogonalized as in (4), from (14):

$$\begin{aligned} y_t &= \boldsymbol{\beta}' \widehat{\mathbf{x}}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \eta_t + \boldsymbol{\beta}' \left(\boldsymbol{\xi}_t - \widehat{\boldsymbol{\xi}}_t \right) \\ &= \boldsymbol{\beta}' \widehat{\mathbf{x}}_t + \boldsymbol{\gamma}' \left(\widehat{\boldsymbol{\Gamma}} \mathbf{z}_t + \widehat{\mathbf{u}}_t \right) + e_t = \boldsymbol{\beta}'_+ \widehat{\mathbf{x}}_t + \boldsymbol{\gamma}' \widehat{\mathbf{u}}_t + e_t \end{aligned} \tag{15}$$

When (10) is the DGP, by orthogonality from (4):

$$\begin{aligned} & T^{1/2} \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}_0 \\ \widetilde{\boldsymbol{\gamma}} \end{pmatrix} \\ &= \begin{pmatrix} T^{-1} \sum_{t=1}^T \widehat{\mathbf{x}}_t \widehat{\mathbf{x}}'_t & T^{-1} \sum_{t=1}^T \widehat{\mathbf{x}}_t \widehat{\mathbf{u}}'_t \\ T^{-1} \sum_{t=1}^T \widehat{\mathbf{u}}_t \widehat{\mathbf{x}}'_t & T^{-1} \sum_{t=1}^T \widehat{\mathbf{u}}_t \widehat{\mathbf{u}}'_t \end{pmatrix}^{-1} \\ &\times \begin{pmatrix} T^{-1/2} \sum_{t=1}^T \widehat{\mathbf{x}}_t e_t \\ T^{-1/2} \sum_{t=1}^T \widehat{\mathbf{u}}_t e_t \end{pmatrix} \\ &= \begin{pmatrix} \left(T^{-1} \sum_{t=1}^T \widehat{\mathbf{x}}_t \widehat{\mathbf{x}}'_t \right)^{-1} T^{-1/2} \sum_{t=1}^T \widehat{\mathbf{x}}_t e_t \\ \left(T^{-1} \sum_{t=1}^T \widehat{\mathbf{u}}_t \widehat{\mathbf{u}}'_t \right)^{-1} T^{-1/2} \sum_{t=1}^T \widehat{\mathbf{u}}_t e_t \end{pmatrix} \\ &\xrightarrow{D} \mathsf{N}_{m+k} \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_\eta^2 \begin{pmatrix} \boldsymbol{\Sigma}_{\widehat{x}\widehat{x}}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ww|z}^{-1} \end{pmatrix} \right] \end{aligned} \tag{16}$$

Thus, the estimator $\widetilde{\boldsymbol{\beta}}_+$ in (16) is again identical to the estimator $\widehat{\boldsymbol{\beta}}$ in (12), independently of the inclusion or exclusion of any or all of the $\widehat{\mathbf{u}}_t$.

5.1. Assessing the validity of the instrumental variables

The validity of the instrumental variables and any additional candidate regressors can be checked by the usual Durbin–Wu–Hausman test when the equation is over-identified (see Durbin, 1954, Wu, 1973, and Hausman, 1978), testing against the most reliable instruments as the baseline. Alternatively, the least reliable instruments can be added to the theory-based equation (see Hendry, 2011), or the equation evaluated using the super-exogeneity test based on IIS in Hendry and Santos (2010).