A Note on Missing Data Effects on the Hausman (1978) Simultaneity Test: Some Monte Carlo Results.

Dikaios Tserkezos and Konstantinos P. Tsagarakis

Department of Economics, University of Crete, University Campus, 74100, Rethymno, Greece

Abstract

This short paper demonstrates the effects of using missing data on the power of the well-known Hausman (1978) test for simultaneity in structural econometric models. This test is a reliable test and is widely used for testing simultaneity in linear and nonlinear structural models. Using Monte Carlo techniques, we find that the existence of missing data could affect seriously the power of the test. As their number is getting larger, the probability of rejecting simultaneity with Hausman test is increasing significantly especially in small samples. A Full Information Maximum Likelihood Missing Data correction technique is used to overcome the problem and then we find out that that the test is more effective when we retrieve these data and include them in the sample.

JEL Classification Numbers: C01,C12,C15.

Keywords: Hausman (1978) simultaneity test, structural econometric models, FIML, missing data, simulation

*Address for correspondence:
Dikaios Tserkezos
Department of Economics,
University of Crete,
University Campus,
74100 Rethymno,
Greece
Tel. +30 28310 77415
Fax. +30 28310 77406
E-mail: tserkez@econ.soc.uoc.gr

I. Introduction

It is a common practice for developments concerning most economic variables to be analyzed containing some missing data. Very often, however, using only the available data and ignoring the problem of missing data, frequently leads to erroneous results with respect to the time dependence in various economic variables. Similar effects should also be expected in the case of the application of various tests based on the available data each time.

Given the above, it is the aim of this short paper to study the effects of missing data in testing simultaneity in structural econometric models. More specifically, we report the results of a Monte Carlo experiment, which examines the effects of missing data on the power of the Hausman (1978) test for simultaneity. On the basis of these Monte Carlo results it appears that as the number of missing data increases, the probability of accepting the null hypothesis of simultaneity using the Hausman test is decreased, especially in small samples of data. As the number of the available data increases, the problem is getting less problematic and after a certain level of available data it disappears.

To overcome this problem we use a Full Information Maximum Likelihood missing data approach (Sargan and Drettakis, 1974) to 'estimate' the missing data and then use them for testing for simultaneity. According to our Monte Carlo result, the probability of accepting the null hypothesis of simultaneity increases significantly when we use simultaneously the available data and the estimated (retrieved) data for testing for simultaneity, especially in a small sample. The power of the test is more effective as compared to applying the Hausman test in the case where the missing data would have been ignored.

Finally the effects of missing data on the power of the Hausman 1978 simultaneity test disappear as the number of the total observations increases, independently of whether we ignore the missing data or we use the suggested missing observations technique to retrieve them.

The paper is organised as follows. Section II presents the Hausman 1978 structural systems simultaneity test and the Full Information Maximum Likelihood approach to estimate the missing data. Section III presents the simulation results. Section IV sums up and provides some concluding remarks.

II .The Hausman 1978 Simultaneity Test.

In order to apply the Hausman 1978 simultaneity test we follow a simple two-step procedure between two endogenous variables (y_{1t}, y_{2t}) of a structural system of equations. In the first step we regress the first endogenous variable with the exogenous and predetermined variables of the system and estimate the residuals, say v_t. In the second step we include the estimated residuals in the second structural equation with dependent variable being the y_{2t} and we perform[1] a t-test on the coefficient of the estimated residuals v_t. If the estimated coefficient is statistically significant, we accept the null hypothesis of simultaneity. The estimated 'missing' data were obtained using a Full Information Maximum approach[2] based on the Sargan-Drettakis approach (Drettakis, 1973), as follows:

$$\min_{missing\ Data\ Parameters} \sum_{t=1}^{T} g_t' V^{-1} g_t \qquad (1)$$

where g_t is the (mx1) vector of the residuals of the structural model with m endogenous variables and

$$Var(g_t) = \begin{bmatrix} \sigma^2_1 & \sigma_{12} & . & \sigma_{1m} \\ \sigma_{21} & \sigma^2_2 & & \sigma_{2m} \\ & & & \\ \sigma_{m1} & \sigma_{m2} & & \sigma^2_m \end{bmatrix} = \delta_{ts}\Omega \qquad (2)$$

$$\Omega^{-1} = V$$

and δ_{ts} is the Kronecker delta.

The missing observations $\hat{y}_i$ will be obtained solving the normal equations:

$$\partial\left(\frac{\sum_{t=1}^{T} g' V^{-1} g}{\partial \hat{y}_i}\right) = 0 \qquad (3)$$

III. The Monte Carlo Experiment

Our simulation experiment is based on the following structural dynamic model:

$$y_{1t} = \beta_1 y_{2t} + \beta_2 x_{1t} + \varepsilon_{1t} \qquad (4)$$

$$y_{2t} = \gamma_1 y_{1t} + \gamma_2 x_{2t} + \varepsilon_{2t} \qquad (5)$$

$$x_1^{*}{}_t = \tau_1 x_1^{*}{}_{t-1} + (\sqrt{(1-\tau_1)^2}) w_{1t} \qquad (6)$$

$$x_2^{*}{}_t = \tau_2 x_{2,t-1}^{*} + (\sqrt{(1-\tau_2)^2}) w_{2t} \qquad (7)$$

$$Cov\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} = \begin{bmatrix} \sigma^2{}_1 & \sigma_{12} \\ \sigma_{21} & \sigma^2{}_2 \end{bmatrix} = \begin{bmatrix} 2.1 & \\ .76 & 2.3 \end{bmatrix} \qquad (8)$$

$$w_{1t} \approx NID(.25) \quad w_{2t} \approx NID(.25) \qquad (9)$$

$$\tau_1 = .90 \quad \tau_2 = .25 \qquad (10)$$

For different numbers of available data between 60 and 700 we generate the dependent variables through the relations of the system taking into account the relations (6)-(10). Five thousand

[1] If we are concerned with the endogeneity of more than one variable, the analysis become somewhat more complicated, but a similar test can applied (Pindyck and Rubinfeld p. 254).

replications of the two dependent variables y_{1t} and y_{2t} are generated. For each application we assumed different numbers of missing data for the variable y_{1t}. The number of missing data is a percentage ranging from 6% to 60% on the number of the available data. The missing data for the dependent variable y_{1t} were estimated using the FIML approach which, in the case of the system described by (4)-(10) can be obtained using the relations:

$$\hat{y}_{1t} = \frac{P_2}{P_1} x_{2t} + \frac{P_3}{P_1} y_{2t} + \frac{P_4}{P_1} x_{1t} \tag{11}$$

with:

$$P_1 = \hat{\omega}_{11} + \hat{\gamma}_1^2 \hat{\omega}_{22} - 2\hat{\gamma}_1 \omega_{21} \tag{12}$$

$$P_2 = (\hat{\beta}_2 \hat{\omega}_{11} - \hat{\gamma}_1 \hat{\gamma}_2 \hat{\omega}_{21}) \tag{13}$$

$$P_3 = -(-\hat{\beta}_1 \hat{\omega}_{11} - \hat{\gamma}_1 \hat{\omega}_{22} + \hat{\omega}_{21} + \hat{\beta}_1 \hat{\beta}_2 \hat{\omega}_{21}) \tag{14}$$

$$P_4 = -(\hat{\gamma}_1 \hat{\gamma}_2 \hat{\omega}_{22} - \hat{\gamma}_2 \hat{\omega}_{21}) \tag{15}$$

A two-step iterative procedure is pursued to obtain the missing data:

In the first step we apply a FIML procedure to obtain estimates of the parameters of the system using only the set of complete observations. In the second step using the estimates of the parameters and the variance covariance matrix of the estimated residuals, we obtained the missing data using the relation (11) with (12)-(15). Then the missing data were included in the full set of data to apply the Hausman 1978 simultaneity test. The results are summarized in Table 1 and 2. In Table 1 we present the effects missing data have on the power of the Hausman 1978 simultaneity test for different numbers of available and missing data. In Table 2 we present the power of the Hausman 1978 simultaneity test having taken into account the suggested method of

[2] For a singe equation missing data approach applied to simultaneous systems see: Dagenais (1975).

recovering missing data and applying the test to the full set.

[TABLE 1 ABOUT HERE]

[TABLE 2 ABOUT HERE]

The following conclusions regarding the effects of time aggregation on the power of the Hausman 1978 simultaneity test can be drawn from an analysis of the data in Table 1 and 2.

On the basis of our Monte Carlo results in Table 1, it emerges that as the number of missing data increase, the probability of accepting simultaneity using the Hausman 1978 test is decreasing especially in small samples of data. As the number of the available data is increasing, the situation is getting less problematic and after a level of available (N=700) data it seems to disappear.

To overcome this problem we use a Full Information Maximum Likelihood approach to "estimate" the missing data and then use them, together with the initial available data, for testing r simultaneity. According to Monte Carlo results, the probability of accepting the true hypothesis increases significantly when we use simultaneously the available data and the estimated missing data for testing for simultaneity, especially in small samples. The power of the test is more effective compared to the case of leaving out the missing data completely.

The effects of missing data on the power of the Hausman 1978 simultaneity test disappear as the number of the total observations becomes large independently of whether we ignore the missing data or we use them (through the suggested missing observations technique to retrieve them).

IV. Conclusions

The results of this note show the implications of missing data in applied time series work. Using

Monte Carlo techniques, we found that the existence of missing data could sometimes lead to erroneous conclusions about the existence of simultaneity in the variables of a structural econometric system.

Using Hausman 1978 structural systems simultaneity test, we show that, as the number of missing data increases, the power of the test declines, leading to wrong conclusions about the existence of causal interaction between the endogenous variables of the structural system.

On the basis of our Monte Carlo results, it emerges that as the number of missing data increases, the probability of rejecting simultaneity using the Hausman 1978 test is increasing especially in small samples of data. As the number of the available data increases, the situation is getting less problematic and as data increase beyond a certain number (N= 700), it disappears.

To overcome this problem we used a Full Information Maximum Likelihood approach to 'estimate' the missing data and then used them for testing for simultaneity. According to our Monte Carlo results, the probability of accepting the true hypothesis is increasing significantly, especially in small samples and the test is more effective compared with the case where the missing data are ignored.

Last, the conclusions of this paper are in line with the more general findings on the negative effects of missing data and specially the case of time aggregation (Tserkezos *et al.*, 1998) on the effectiveness of the statistical criteria for controlling the interdependence between economic magnitudes.

References

Anderson, T. W. (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association*, 52(278), pp. 200-203.

Drettakis, E. (1973) Missing data in econometric estimation, *Review of Economic Studies*, 40(4): 537-552.

Dagenais, M. (1975) Incomplete Observations and simultaneous equations models, *Journal of Econometrics*, 4(3), pp. 231-241.

Hausman, J.A. (1978) Specification Tests in Econometrics, *Econometrica* 46(6), pp. 1251-1271.

Gilbert, C.L. (1977) Regression using mixed annual and quarterly data, *Journal of Econometrics*, 5(2), pp. 221-239.

Pindyck R. S. & Rubinfeld D. L. (1997) Econometric Models and Economic Forecasts, Irwin/McGraw Hill Massachusetts.

Oguchi, N. and Fukuchi, T. (1990). On temporal aggregation of linear dynamic models, *International Economic Review,* 31(1), pp. 187-193.

Sargan, J.D. and Drettakis, E.S. (1974) Missing data in an autoregressive model, *International Economic Review*, 15(1): 39-58.

Tserkezos, D., Georgutsos, D. & Kouretas, G. (1998) Temporal Aggregation in Structure VAR Models, *Applied Stochastic Models and Data Analysis*, 14(1), pp.19-34.

Tserkezos, D. (1991) Simultaneous Use of Annual and Quarterly Data in Econometric Models, Working Paper 12,. Department of Economics, University of Piraeus.

Table 1. Probability of type I error (different numbers of missing data and available data).

Percentage of missing data	Number of available data							
	60	120	200	300	400	600	600	700
6%	34.7	93.6	96.4	97.6	97.9	98.1	98.4	98.9
12%	10.8	92.8	96.1	97.4	97.8	97.9	98.3	98.9
18%	10.7	91.3	95.5	96.7	97.6	97.7	98.2	98.8
24%	10.5	88.0	94.8	96.4	97.3	97.5	98.1	98.8
30%	10.4	79.2	93.7	95.9	96.9	97.2	98.1	98.7
36%	10.3	45.8	92.5	95.0	96.4	96.7	97.8	98.6
42%	10.7	10.2	89.2	94.1	95.9	96.3	97.4	98.3
48%	10.2	10.2	78.5	92.6	95.1	95.9	96.9	98.0
54%	10.3	9.8	18.0	89.1	93.7	95.0	96.4	97.6
60%	10.5	10.1	8.7	68.1	90.4	93.3	95.1	96.6

Source: Data entries are probabilities of rejecting the wrong hypothesis, i.e. the existence of measurement errors in the independent variable of model (2)-(4) The Hausman 1978 test was replicated 5000 times for the specification (4)-(10) The size of the test is a=0.05. Data entries are given by n/5000 where n is the number of times the null hypothesis is accepted.

Table 2. Probability of type I error (different numbers of missing data and total data).

Percentage of missing data	Number of available data							
	60	120	200	300	400	600	600	700
6%	36.7	92.5	95.4	96.8	97.0	97.0	97.3	97.8
12%	30.6	92.1	94.6	96.0	96.9	96.5	96.8	97.0
18%	31.5	91.7	94.5	95.6	96.0	96.4	96.4	96.9
24%	32.8	89.1	94.6	95.6	96.3	96.0	96.0	96.7
30%	32.9	82.6	94.2	95.7	96.3	96.1	97.1	97.5
36%	33.0	70.7	93.3	95.7	96.7	96.9	97.4	98.2
42%	33.1	63.3	91.4	95.2	96.8	96.8	97.7	98.2
48%	32.9	64.3	86.3	94.4	96.4	96.7	97.4	98.3
54%	31.4	65.4	75.0	92.5	95.6	96.1	97.1	98.0
60%	31.2	65.6	74.5	84.1	93.6	95.2	96.5	97.6

Source: Data entries are probabilities of rejecting the wrong hypothesis, i.e. the existence of measurement errors in the independent variable of the model (2)-(4) .The Hausman 1978 test was replicated 5000 times for the specification (4)-(10). The size of the test is a=0.05. Data entries are given by n/5000 where n is the number of times the null hypothesis is accepted.