Testing Endogeneity in a Regression Model: An Application of Instrumental Variable Estimation

Berta Rivera Castiñeira[1] *Luis Currais Nunes*[2]

[1]Department of Applied Economics
University of La Coruña
Campus da Zapateira, 15071, La Coruña, Spain
berta@udc.es

[2]Department of Economic Analysis
University of La Coruña
Campus da Zapateira, 15071, La Coruña, Spain
currais@udc.es

Abstract
A goal of economic research is to determine the causal relationship among economic variables. This paper introduces the Hausman specification test of endogeneity, for testing the independence between the stochastic regressor and the disturbances. We describe alternative approaches to estimating simultaneous-equation systems and we present an empirical example of economic growth and health that finds health expenditure endogenous. We use different sets of instruments as exogenous determinants of health in an instrumental variables estimation.

Keywords: Endogeneity, Instrumental Variables, Health.

1 Introduction

Econometric models divide variables into ”endogenous” variables - those determined by the current workings of the model being studied - and ”exogenous” variables. The latter are variables whose values are determined outside the model. From a formal standpoint the exogenous variables are assumed to be statistically independent of all stochastic disturbance terms of the model, while the endogenous variables are not statistically independent of those terms. Specifically, the exogeneity of a variable depends on the parameters of interest of the researcher and on the purpose of the model, whether for statistical inference, forecasting or policy analysis. Exogeneity thus plays a key role throughout economic and econometric analysis, both theoretical and applied.

Following different methodologies several authors confirm the positive relationship between health status and income. Although the positive effect of income on health was studied in depth by different works, there was little empirical evidence to support the causal effect of health on income. A healthier work-force would be related to the human capital accumulation process as a logical assumption that good health raises the economic productivity of individuals and countries economic growth rates.

Nevertheless, if the causal relationship between health and income runs in both directions the estimation by the Ordinary Least Squares (OLS) would yield biased and inconsistent estimates of the structural parameters. Contrasting the results obtained using OLS estimation we carry out the Hausman test [7] to check the existence of endogeneity and afterwards we use instrumental variables to estimate the effect of health on income growth. Results demonstrate the causal effect of health on productivity.

This paper is divided into five sections. In the following section we examine different estimators available for simultaneous-equation systems. In section three we describe the method of instrumental variables and the Hausman specification test. In next section we analyze the existence of causality in order to identify the links between health and income growth. Section five concludes with the main findings and their interpretation.

2 Estimation of Simultaneous-equation Systems

The division between endogenous variables and predetermined variables is crucial for consistent estimation whether the model consists of a single equation and is to be estimated by OLS or related techniques or whether a simultaneous equation model is involved. A simultaneous-equation model determines the values of one set of variables, the endogenous variables, in terms of another set of variables, the predetermined variables.

There are several different estimators available for simultaneous-equations systems.[1] We can classify them depending on the approach to estimating the system. They are the naive approach, the limited-information approach and the full-information approach. The first estimates each equation of the system as a single equation and it is called ordinary least squares (OLS). This approach has a number of desirable properties when appropiate assumptions are satisfied. Briefly, if the explanatory variables in the equation to be estimated are either nonstochastic or distributed independently of the disturbance term in that equation, if the disturbance

[1] See Brundy and Jorgenson [1][2], Fisher [6] and Madansky [11].

term is serially uncorrelated and homoskedastic, and if there are no a priori restrictions on the parameters to be estimated, OLS is the best linear unbiased estimator.

These assumptions can be weakened in several ways. First, if the explanatory variables are not independent of the disturbance term but are uncorrelated with it in the propability limit, OLS to be unbiased but is consistent. If the disturbances are serially correlated, OLS loses efficiency but retains consistency provided such serial correlation does not affect the validity of assumptions concerning the correlation of the current disturbance term and the explanatory variables. Finally, OLS leads to biased and inconsistent estimators if the equation to be estimated is one of a system of simultaneous structural equations. This approach applies least squares to each equation of the model separately, ignoring the distinction between explanatory endogenous and included exogenous variables. It also ignores all information available concerning variables not included in the equation being estimated.

The limited-information approach, estimates one equation at a time, estimating the system as does OLS, buy unlike OLS it distinguishes between explanatory endogenous variables and included exogenous variables. Thus it utilizes all identifying restrictions belonging to the equation. This approach does not require information on the specification of the other equtions of the system, in particular, the indentifying restrictions on these other equations. This approach includes several specific estimators as indirect least squares (ILS), two-stage least squares (2SLS) and k-class estimators, including limited-information maximun likelihood (LIML). These estimators can be expressed as instrumental variable estimators (IV), discussed in section 3, for particular choices of instrumental variables.

The full-information approach estimates the entire system of simultaneous equations simultaneously using all information available on each of the equations of the system. It estimates all structural parameters of the system, given the model and all identifying restrictions on each equation of the system. This approach includes two specific estimators, three-stage least squares (3SLS) and full-information maximum likelihood (FIML). However desirable the properties of full-information methods may be in principle when all assumptions are met, such estimators suffer realtively heavily from a lack of robusteness in the presence of common practical difficulties. Thus full-information maximum likelihood is more sensitive to multicollinearity than limited-information estimators. Further, it is evident that all full-information methods are rather sensitive to specification errors of the econometrics models. In particular, such estimators have the defects of their merits in that by using information of the entire system to estimate any single equation they carry the effects of specification error in any part of the system to estimate of any other part.

All the above estimators are extensions of the two basic techniques of single-equations estimation: least squares and maximum likelihood. As indicated by their names, ordinary least squares, two-stage least squares, and three-stage least squares are extensions of the least-squares technique to simultaneous-equations estimation. Similarly, limited-information maximum likelihood and full-information maximum likelihood are extensions of the maximum-likelihood technique to simultaneous-equations estimation.

3 Instrumental Variables and the Hausman Specification Test

The method of instrumental variables (IV) is a general approach to estimate a single equation in a system of equations, and all of the estimators introduced so far can be interpreted as IV estimators for particular choices of instrumental variables. Consider the following structural equation, which can be expressed as

$$y_1 = Z_1\delta_1 + \varepsilon_1 = \begin{pmatrix} Y_1 & X_1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \beta_1 \end{pmatrix} + \varepsilon_1, \tag{1}$$

where Z_1lumps together data on all included explanatory variables, whether endogenous or exogenous, and summarizes all coefficients to be estimated:

$$Z_1 = \begin{pmatrix} Y_1 & X_1 \end{pmatrix}, \tag{2}$$

$$\delta_1 = \begin{pmatrix} \gamma_1 \\ \beta_1 \end{pmatrix}. \tag{3}$$

A heuristic explanation of the OLS estimator, is based on premultiplying (1) by $Z_1^{'}$to make the matrix multiplying $\delta_1$square, yielding

$$Z_1^{'} = Z_1^{'} Z_1 \delta_1 + Z_1^{'}\varepsilon_1. \tag{4}$$

Dropping the term $Z_1^{'}\varepsilon_1$and solving the resulting "normal equations" yields the OLS estimator

$$\widehat{\delta}_{1OLS} = \begin{pmatrix} Z_1^{'} & Z_1 \end{pmatrix}^{-1} Z_1^{'} y_1. \tag{5}$$

In a single-equation context, dropping the term corresponding to $Z_1^{'}\varepsilon_1$was justified because the explanatory variables were exogenous and hence uncorrelated with the stochastic disturbance term. In the simultaneous-equation context, however, dropping this term cannot be so justified, since the explanatory endogenous variables in Z$_1$are not statistically independent of ε_1. They are correlated with the ε's even in the probability limit. Hence estimation by OLS would yield biased and inconsistent estimates of the structural parameters.

Suppose, however, that there exists a set of $g_1 - 1 + k_1$ variables (the same number as in Z_1) that there are uncorrelated with Z_1. Such variables are instrumental varialbes, and data on them are summarized by the $nx\,(g_1 - 1 + k_1)$ matrix W_1, where the subscript again refers to the first equation. Then premultiplying (1) by $W_1^{'}$ yields

$$W_1^{'}y_1 = W_1^{'}Z_1\delta_1 + W_1^{'}\varepsilon_1. \tag{6}$$

Dropping $W_1^{'}\varepsilon_1$, since the variables in W_1 were assumed uncorrelated with $\varepsilon_1$and solving for δ_1 yields

$$\widehat{\delta}_{1IV} = \widehat{\delta}_1\left(W_1^{'}W_1\right)^{-1}W_1^{'}y_1. \tag{7}$$

This is the instrumental variables (IV) estimator, which, as indicated by the functional relationship $\widehat{\delta}_1\,(W_1)$, depends on the choice of instruments and the data on these instruments. The IV estimator is extremely useful, since it represents a whole class of estimators, each defined by W_1, the matrix of data on the instrumental variables. As already noted, all the estimators introduced so far are members of this class and can be interpreded as IV estimators for particular choices of W_1.

The major problem in using the instrumental-variables technique is simply that of obtaining a suitable set of instrumental variables that are both sufficiently uncorrelated with the stochastic disturbance terms and sufficiently correlated with the relevant explanatory variables.

The first requirement of a good instrumental variable is that it be predetermined, more precisely, that it be asymptotically uncorrelated with the disturbance in the equation to be estimated. There must be no simultaneous feedback loops connecting the equation to be estimated and the equation or equations explaining the potential instrument; further, the disturbance from the equation to be estimated must not be correlated with that of the equation explaining the potential instrumental variables.

The other requirement of a good instrumental variable is that it push the endogenous variables in the equation to be estimated. In the limit, it does no good at all to regress Y_1on variables unrelated to the model, so that $\widehat{Y_1}$ turns out to be asymptotically constant. The model itself, however, provides what is supposed to be a complete theory of the determination of the endogenous variables and therefore a complete list of instruments eligible from this point of view. Any other variable can affect the endogenous variables only by affecting one of these; otherwise, it should be in the model.

Having decided on a list of eligible instrumental variables, the problem arises of how to choose among them. Either some can be dropped from the list or, more generally, certain linear combinations of them can be selected for use. In this sense, Kloek and Mennes [8] suggest to use the first m principal components of the instrumental variables together with the l included predetermined variables. This has two advantages: It reduces multicollinearity, since principal components and mutually orthogonal and, in some sense, the use of principal components summarizes the information in the list of instrumental variables. The problem is, however, that the sense in which that information is summarized is not ovbiously the right one. The same principal components will be chosen for use in the estimation of each equation despite the fact that different endogenous variables appear in the different equations.

Fisher[6] suggested one particular way of utilizing prior structural information to select instruments. The method suggested, called SOIV (structurally ordered instrumental variables) essentially proceeds in three steps. The first of these establishes a preference ordering of the instruments relative to a particular right-hand-side endogenous variable. Secondly, given that preference ordering, the endogenous variable in question is regressed on the insstruments in differing combinations to determine whether an instrument far down in the ordering has an independent effect on the endogenous variable in the presence of more preferred instruments or whether it is just using up a degree or freedom. The end result of this stage is a final list of instruments relative to a particualr right-hand-side endogenous variable. Finally, that list is used to construct elements of $\widehat{Y_1}$.

Since "exogenity" is fundamental to most empirical econometric modelling, is conceptualization, its role in reference, and the testing of its validity have been subject of extensive discussion.[2] Hausman [7] has suggested a method for testing exogeneity specification. The fundamental idea of the Hausman specification test is as follows: If there are two estimators $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ that converge to the true value β under the null hypothesis but converge to different values under the alternative, the null hypothesis can be tested by testing whether the probability limit of the difference of the two estimators, $\widehat{q} = \widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$, is zero. Suppose, based on a sample of T observations that $\widehat{\beta}^{(1)}$ attains the asymptotic Cramer-Rao lower bound and both $\sqrt{T}\left(\widehat{\beta}_1 - \beta\right)$ and $\sqrt{T}\left(\widehat{\beta}_2 - \beta\right)$ are asymptotically normally distributed with mean zero and covariance matrix V_1 and V_2 respectively; then $Var\left(\sqrt{T}\widehat{q}\right) = V_2 - V_1$ and $T\widehat{q}(v_2 - v_1)^{-1}\widehat{q}$ is asympticially chi-square distributed with k degrees of freedom, where k is the dimension of the vector β.

4 Empirical Results. The Use of Instrumental Variables

In order to establish the effect of health status on income growth we run a log-linear equation that implies that the growth of per capita income is a function of the determinants of the steady state and the initial level of income as follows:

[2] See Ericsson and Irons [5].

$$\ln\left(\frac{y(t)}{y(0)}\right) = \ln y(t) - \ln y(0) = \left(1 - e^{-\lambda t}\right)\frac{\alpha}{\mu+\beta}\ln s_k + \left(1 - e^{-\lambda t}\right)\frac{\beta}{\mu+\beta}\ln e^* +$$

$$\left(1 - e^{-\lambda t}\right)\frac{\eta}{\mu+\beta}\ln s_h - \left(1 - e^{-\lambda t}\right)\frac{1-\mu-\beta}{\mu+\beta}\ln(n+g+\delta) -$$

$$\left(1 - e^{-\lambda t}\right)\ln y(0)\,. \tag{8}$$

We analyze the effect of heatlh on income growth using cross-country data on income per worker in two periods of time y(t) and y(0), investment rate (s_k), education (e^*), health expenditure (s_h)and population growth (n). The variables g and δ are the technical progress and the rate of physical capital depreciation. This equation is estimated using OLS with White´s heterocedasticity-consistent covariance estimation method. The analysis was carried out for OECD countries and covers the period 1960-1990 (see Currais and Rivera [3].Available on request).

Income and investment rates data were acquiered from the Summers-Heston data set[18]. This data is expressed in real terms and at 1985 international prices. GDP appears in per worker terms and investment in percentage form. The information which is related to educational attainment was obtained from Kyriacou [9]. Health and labor force data are obtained from OECD [14]. We use a proxy for the rate of health accumulation that measures the total expenditure in health as a percentage of GDP. The conversion is made by using the overall purchasing power parity of the countries.

We use instrumental variables estimation to identify the strength of the association between health and income, by means of an analysis of the causal effect of health expenditure/GDP on per worker income. If the causation in the relationship between health and income runs in both directions it implies that the regressor and the disturbance term are correlated. Hence the estimation by OLS would yield biased and inconsistent estimates of the structural parameters. We used different sets of instruments as exogenous determinants of health expenditure.

The magnitudes cited in several countries suggest that the aging population and new medical technology add nearly half a percentage point per annum to the growth of the health bill in OECD countries (OECD, [13]). Posnett and Hitiris [16] among others, demonstrated that health expenditure is explained in part by the proportion of a population over 65. Evans [4] shows that increases in hospital costs are associated with increasing utilization rates and the intensity of services. Our metodology involves physician contacts per person, rate of dialyses treatments per million population and number of beds in-patient care per 1000 population. Leu [10] finds that a 10 per cent difference in hospital beds corresponds to a 4 per cent difference in medical care expenditure in OECD countries. We also use alcohol consumption per capita over 15 as a non-medical health determinant. The problems caused by abusive alcohol consumption make medical resources scarce and extend to beyond the damage that drinkers do to themselves.

In table 1 we estimate the effect of health on income growth using cross-country data on income per worker, health (health expenditure), education, investment rate, population growth and the various instruments. As the variables represent quite different information and have only a weak correlation among themselves we can trust them to be good instruments.

	1	2	3	4	5	6	7	8
	OLS	IV	IV	IV	IV	IV	IV	IV
ln s_h	0.22	0.71	0.21	-0.02	-0.38	0.28	0.29	0.28
	(2.40)	(1.42)	(1.29)	(-0.05)	(-0.56)	(1.26)	(1.99)	(1.96)
ln $y(0)$	-0.59	-0.73	-0.59	-0.53	-0.43	-0.61	-0.62	-0.61
	(-18.35)	(-7.15)	(-9.41)	(-5.31)	(-2.31)	(-7.90)	(-10.19)	(-10.27)
ln s_k	0.33	0.25	0.34	0.37	0.44	0.32	0.32	0.32
	(2.51)	(1.12)	(2.17)	(2.12)	(1.65)	(2.03)	(2.01)	(2.08)
ln$(n + g + \delta)$	-0.45	-0.51	-0.45	-0.43	-0.39	-0.46	-0.47	-0.46
	(-2.42)	(-1.55)	(-1.95)	(-1.64)	(-1.03)	(-1.96)	(-1.95)	(-1.96)
ln e^*	0.20	0.07	0.20	0.25	0.35	0.18	0.17	0.18
	(4.52)	(0.49)	(2.12)	(2.08)	(1.61)	(4.31)	(1.79)	(1.93)
Instrumental	-	Alcohol	Population	Beds/1000	Physicians	Dialyses	(a)	(b)
Variable	-	consump.	over 65	population	contacts			
First stage	-	0.36	0.27	0.28	0.33	0.35	0.49	0.51
R^2								
Hausman	-	0.002	0.37	0.37	0.14	0.23	0.02	0.06
(p-value)								
(a)Alcohol,	over 65,	beds,	contacts					
(b)Over 65,	beds,	contacts,	dialyses					

Table 1. Growth and health expenditure. Instrumental variable estimates, 1960-1990.

Notes: Value of t-statistics in parenthesis. "First stage" R^2 is the R^2 regressing health expenditure on the instrument set. The Hausman statistic tests equality of the IV and OLS estimates.

In column 1, only results obtained by the OLS estimation for comparison purposes are reproduced. From column 2 to 6 different variables are used as instruments, although each one considered individually is very imprecise. Columns 7 and 8 show the results when all four instruments are used together. In both cases the IV estimates are higher than OLS (from 0.29 to 0.28 versus 0.22), and each is statistically significant. Probable explanations for this effect might be the loss of efficiency due to the use of a small sample and that the linear specification used might no have been able to capture the asymptotic effect observable in higher levels of health status.[3] Nevertheless empirically results clearly demostrate the existence of the effect according to the significance of the t-ratio obtained.

The Hausman test indicates the existence of endogeneity. The association betweeen income growth and health status runs in both directions. The last row in table 1 displays the p-value of the Hausman test for each of the instrument sets. The test rejects that the OLS and IV estimates are equal when all instruments are used together in the two last columns of table 1. The evidence points to the existence of a relationship between health and income that occurs in both directions.

5 Conclusion

Due to the existence of reverse causation between two economic variables, the OLS estimation could bias the result observed. We use instrumental variables estimation as an limited-information approach for simultaneous equation-systems. Using a conditional convergence regression where the growth of per capita income is a function of the determinants of the steady state and considering health as an important determinant of an enhanced labor force, we obtain the result that health affects income growth both positively and significantly. Carrying out the Hausman test we confirm the existence of a feedback effect between health and income. When different variables related to health expenditure are used to estimate the eleasticity produces slightly higher results (with a $\widehat{\beta}_{IV}$ of roughly 0.29) than the OLS estimation. These results are acceptable and confirm the positive effect of health on income variation.

Acknowledgement

We are especially indebted to José Ramón Cancelo, of University of La Coruña, for his significant contribution to this study. We are grateful to Guillem López Casasnovas and Carlos Murillo, of Pompeu Fabra University, for helpful and valuable comments. Financial support from La Coruña University, as part of a research study, is gratefully acknowledged.

References

[1] Brundy, J.M., Jorgenson, D.W., "Efficient Estimation of Simultaneous Equations by Instrumental Variables", Review of Economic and Statistics, 53, 1971, pp. 207-224.

[2] Brundy, J.M., Jorgenson, D.W., "Consistent and Efficient Estimation of Systems of Simultaneous Equations", P. Zarembka, Ed., Frontiers in Econometrics, Academic Press, 1973.

[3] Results obtained carrying out different proofs, for instance, for smaller values of $\widehat{\beta}_{IV}$ such as 0.20 and 0.22 we obtain a t-ratio of 1.22 and 1.32 respectively.

[3] Currais, L., Rivera, B., "Human Capital and Growth: Does Health Affect Productividy?". La Coruña University Working Paper N 97.6, 1997.

[4] Evans, R.G., Strained Mercy: The Economics of Canadian Health Care. Butter-Worths, 3-21, 1984.

[5] Ericsson, N., Irons, J., Testing Exogeneity. Oxford University Press, 1994.

[6] Fisher, F., Econometrics. Essays in Theory and Applications. Harvester Wheatsheaf, 1991.

[7] Hausman, J.A., "Specification Test in Econometrics", Econometrica 46(6), 1978, pp. 1251-1271.

[8] Kloek, T., Mennes, L., "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables", Econometrica 28, 1960, pp. 45-61.

[9] Kyriacou, G., "Level and Growth Effects of Human Capital: A Cross-country Study of the Convergence Hypothesis", C.V. Starr Working Paper 91-26, 1991.

[10] Leu, R., "The Public-private Mix and International Health Care Cost". In: Culyer, A.J. Jonsson, B. (Eds.), Public and Private Health Services: Complementarities and Conflicts. Basil Blackwell, 1986.

[11] Madansky, A., Foundation of Econometrics, North Holland, 1976.

[12] Newhouse, J.P., Phelps, C.E., "New Estimates of Price and Income Elasticities". In: Rosset R.N. (Ed.), The Role of Health Insurance in the Health Services Sector, NBER, 1976.

[13] OECD., "Health Systems. Facts and Trends 1960-1991". Health Policy Studies 3. OECD, 1993.

[14] OECD., Health Data File. OECD, 1995.

[15] Parkin, D., McGuire, A, Yule, B., "Aggregate Health Care Expenditures and National Income: Is Health Care a Luxury Good?", Journal of Health Economics 6,1987, pp. 109-127.

[16] Posnett, J., Hitiris, T., "The Determinats and Effects of Health Expenditure in Developed Countries", Journal of Health Economics 11, 1992, pp. 173-181.

[17] Prichett, L., Summers, L.H., "Wealthier is Healthier". The Journal of Human Resources 31(4), 1996, pp. 841-868.

[18] Summers, R., Heston, A., "The Penn World Table (mark 5): An Expanded Set of International Comparisons", Quaterly Journal of Economics, 1991, pp. 327-368.