OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 48, 3 (1986)
0305-9049 $3.00

AN APPLICATION OF THE GRANGER & ENGLE TWO-STEP ESTIMATION PROCEDURE TO UNITED KINGDOM AGGREGATE WAGE DATA*

S. G. Hall

INTRODUCTION

In a recent paper Granger and Engle (1985) establish a number of new results concerning cointegration and error correction models. In particular they show that if a vector of variables is cointegrated then there exists a valid error correction (see Davidson, Hendry, Srba and Yeo, 1978 or Hendry and Von Ungen-Sternberg, 1981) representation of the data which is not liable to the problems of 'spurious' regression (see Granger and Newbold, 1977). While this result establishes the validity of the traditional error correction model, Granger and Engle go on to suggest a two-step estimation procedure which allows explicit tests of the underlying assumption of cointegration. This procedure is shown to give coefficient estimates which converge on the true parameter values. Furthermore 'These estimates converge even faster to the true value than standard econometric estimates' (Granger and Engle, 1985, p. 14).

This two-step procedure is quite straightforward. First, a prior levels regression (see Hall and Brooks, 1985) is performed which allows the hypothesis of cointegration to be tested. Then the residuals from this regression are entered into the error correction model in place of the levels terms. This intuitively has the effect of imposing a set of parameter values on the levels terms which give minimum least squares errors in this part of the equation. Imposing this restriction is the intuitive explanation of the increased convergence speed of the two-stage estimates.

In this paper the Granger and Engle two-step procedure is used in the estimation of an aggregate wage equation for the UK which incorporates a term in the expected rate of price inflation. Typically when such a term is included and standard econometric techniques are used on UK data the value of the parameter on expected prices is quite a lot larger

* I am grateful for the helpful comments and suggestions of C. W. J. Granger, D. F. Hendry, S. G. B. Henry and S. J. Nickell.

than unity. The use of the two-stage estimation technique will be seen to produce a set of parameter estimates which are much closer to our *a priori* expectation. The long run from solution to the model is found to be quite different from the freely estimated error correction model.

COINTEGRATION AND THE TWO-STAGE PROCEDURE

The definition of cointegration given in Granger and Engle (1985) differs slightly in detail from that given in Granger and Weiss (1983), although its intuitive meaning is unchanged. A series, X_t, is said to be integrated of order d (denoted $X_t \sim I(d)$) if it is a series which has a stationary, invertible, non-deterministic ARMA representation after differencing d times. Cointegration is defined as follows: the components of the vector X_t are said to be cointegrated of order d, b (denoted $X_t \sim (d, b)$) if:

(i) all components of X_t are $I(d)$ and
(ii) there exists a vector $\alpha(\neq 0)$ so that $Z_t = \alpha' X_t \sim I(d-b), b > 0$

The vector α is called the cointegrating vector. This definition represents a slight generalization of that of Granger and Weiss to the case where Z_t is not itself stationary.

The importance of this definition to the error correction (ECM) model is that if the levels part of the ECM is not made up of a cointegrated vector then this part of the error term will be non-stationary. Unless there is an exactly offsetting non-stationarity coming from the difference terms in the equation the overall ECM error term will also be non-stationary. In that case the resulting estimates will be meaningless.

The advantage of the Granger and Engle two-stage procedure is that the Z_t errors may be tested for stationarity and α, the cointegrating vector, can be imposed on the ECM estimated equation. So not only do we know that X_t is a properly cointegrating vector but we also know that the final equation is based on a consistent estimate of α.

AN APPLICATION TO THE UK WAGE DATA

Using this concept of cointegration to construct an aggregate model of wage determination is particularly apt as the earliest examples of the application of ECM models in econometrics related to this sector (see Sargan, 1964). The early models involved only level terms in real wages and a trend representing target real wages. More recent models have also included elements from the Phillips curve literature with the level of unemployment also entering the formulation (see Hall, Henry and Trinder, 1983).

Before proceeding to test the sets of variables for cointegration it is sensible to establish the properties of the individual series. Much of the

theory of cointegration has been developed for the case where all the series are $I(1)$. Higher orders are of course possible and are allowed for under the general definition of cointegration given above. Complications arise, however, when the series are integrated of different orders (e.g. one series might be $I(1)$ and another $I(2)$); the two series cannot then be cointegrated. In this paper we will be concerned with five series; these are: LW: the log of wages; LP: the log of the consumer price index: LPROD: the log of aggregate productivity, UPC: the percentage unemployment rate; and LAVH: the log of average weekly hours worked. (Data definitions are supplied in an Appendix.) In order to test the level of integration of these variables the Dickey–Fuller (DF) and an Augmented Dickey–Fuller (ADF) test will be used. These are both 't' tests and rely on rejecting the hypothesis that the series is a random walk in favour of stationarity; this requires a negative and significant test statistic. Table 1 reports the DF and ADF statistics for the five series and their first differences.

If we first consider the levels of the five variables it is quite obvious that none of them are stationary processes. Four of the variables actually have positive tests statistics and the one negative one (LAVH) is insignificant. Of the first differences, ΔLPROD, ΔLAVH and ΔUPC yield negative and significant values on both tests. As differencing once produces stationarity, we may conclude that these series are $I(1)$. The two remaining series ΔLW and ΔLP are not significant on both tests so it is not possible to reject the hypothesis that they are a random

TABLE 1
The Time Series Properties of the Variables

	TEST	
Variable	*DF*	*ADF*
LW	10.9	2.6
LP	14.5	1.9
LPROD	3.8	3.3
LAVH	−0.3	−0.5
UPC	5.2	1.8
ΔLW	−3.5	−1.4
ΔLP	−1.4	−0.9
ΔLPROD	−0.8	−2.4
ΔLAVH	−11.3	−4.6
ΔUPC	−2.4	−2.5
LW-LP	2.6	2.2
Δ(LW-LP)	−8.5	−3.6

walk in first difference. This indicates that both LW and LP are probably $I(2)$. In the case of three or more variables it is possible to have a subset of the variables which are integrated at a higher order than the remaining variables and still have a valid cointegrating vector if the subset of variables together is integrated at the same order as the remaining variables. The remaining two rows of Table 1 show that the real wage (LW–LP) is $I(1)$ even though both LW and LP separately are $I(2)$. It is therefore possible that all the variables could form a cointegrating set.

The original Sargan wage bargaining model suggested that real wages would grow steadily along a simple trend, which was interpreted as the desired or target real wage growth of the union sector. There are two problems with this original formulation from the point of view of this paper. The first is simply that as the final wage equation is explaining nominal wages then in order to set up a full cointegrating vector of variables we should relax the unit coefficient on prices. The second problem arises from the definition of cointegration, given above, that the variables must be non-deterministic. A time trend is clearly deterministic and must strictly fall outside the definition of cointegration. It is, however, worth noting that this is also true of the constant, which is invariably included in the cointegrating regression. There are other reasons also for abandoning the use of a trend in this equation; in particular the existence of the long-term rise in real wages is widely associated with the long-term growth in productivity. So it may be preferable to use aggregate productivity rather than a time trend for this reason also.

The basic Sargan model using smoothed productivity instead of a time trend may be tested as a cointegrating vector in the following regression

$$LW = -5.49 + 0.99\ LP + 1.1\ LPRODS^{1} \qquad (1)$$

$$CRDW = 0.24;\ DF = -1.7;\ ADF = -2.6;\ R^2 = 0.9972$$

RCO:	0.86	0.72	0.52	0.35	0.18	0.04	0.08
			−0.20	−0.27	−0.29	−0.32	−0.34

Sample 1963Q4-1984Q4.

R^2 is the standard squared multiple correlation statistic associated with the regression and RCO is the residual correlogram.

Three tests of the cointegrating regression are reported here; these are: CRDW which is the cointegrating regression Durbin–Watson statistic (derived from Sargan and Bhargava, 1983), the Dickey and Fuller test (DF) and the Augmented Dickey and Fuller tests (ADF)). All three tests are used in Granger and Engle (1985) which derives a set of critical values for the tests on the basis of a Monte Carlo study. Some care must be taken in the use of these tests as the Granger and Engle paper only reports results for the two-variable case. In correspondence Professor Granger has kindly provided some critical values from a three-variable Monte Carlo study; these critical values are given below.

[1] No t statistics or other summary statistics will be reported for the cointegrating regressions as these estimates may be biased, as pointed out by Granger and Engle. The parameter estimates are not affected by this bias, of course, see Stock (1985); however, it should be remembered that if cointegration is rejected then the parameter estimates may be biased. All the regressions use a smoothed version of productivity to remove the effect of a very short term cyclical pattern; the results are not altered in any substantial way by the use of current productivity.

	1%	5%	10%
		2 variable case	
CRDW	0.511	0.386	0.322
DF	−4.07	−3.37	−3.03
ADF	−3.77	−3.17	−2.84
		3 variable case	
CRDW	0.488	0.367	0.308
ADF	−3.89	−3.13	−2.82

Source: two variable case: Granger and Engle (1985). Three variable case: my thanks are due to Professor Granger for his permission to report these.

As up to five variables will be considered below these figures should be taken as only an approximate guide.

On the basis of the CRDW, the DF, and the ADF test statistics we are unable to reject the assumption that equation (1), the simple Sargan model, represents a non cointegrating vector of level terms.

If we go on to add the percentage level of unemployment to the vector of variables we can test whether incorporating this element of the Phillips curve literature produces a set of cointegrating variables. The relevant cointegrating equation is then

$$LW = -5.6 + 1.03\ LP + 1.07\ LPRODS - 0.72\ UPC \qquad (2)$$

$$CRDW = 0.28;\ DF = -2.12;\ ADF = -3.0;\ R^2 = 0.9974$$

RCO:	0.85	0.70	0.49	0.29	0.10	−0.06	−0.18
			−0.31	−0.37	−0.39	−0.41	−0.34

All the parameter values of this regression have reasonable values and are correctly signed. However, the CRDW and the DF statistic are well below their critical value, although they have risen considerably from equation (1). Again, we cannot reject the hypothesis that these variables are not a cointegrated vector.

There is, however, another term which often appears in the specification of aggregate wage equations in the UK. This term is the log of average hours worked. The reason for its inclusion is due to the way in

which the aggregate wage data are generated. This is often done by taking total wages and salaries for the UK as a whole from the National Accounts and dividing this number by the product of employment and hours to give the average hourly wage. This means that a change in hours worked will have a direct effect on the measured wage if total wages and salaries do not move enough to offset it. As many workers are salaried rather than paid hourly this may well be the case. Another effect is that if overtime hours are paid at a different rate than basic hours, then as marginal changes in hours will occur mainly in the over-time section the weighting pattern of basic and over-time wage rates will vary with hours worked. Some researchers have tried to remove this effect by making an *ad hoc* adjustment to wages for hours worked. A more successful practice is simply to include hours as one of the explanatory variables. The following cointegrating regression includes a term in hours:

$$LW = 2.88 + 1.02\ LP + 0.93\ LPRODS - 0.61\ UPC - 1.79\ LAVH \tag{3}$$

$$CRDW = 0.74;\ DF = -4.07;\ ADF = -2.88;\ R^2 = 0.9993$$

RCO: 0.63 0.39 0.09 −0.1 −0.03 −0.06 −0.05
−0.04 −0.06 −0.05 −0.06 −0.02

where LAVH is the log of average hours.

The CRDW test now rejects the hypothesis of non cointegration decisively, as does the DF test; the ADF test statistic has actually fallen slightly compared with (2): it is still fairly high although it is not able to reject non cointegration. The residual correlogram also would strongly suggest a stationary error process.

It would seem reasonable to conclude that the five variables in (3) constitute a cointegrating vector. By comparing (3) with (2) we know that the inclusion of LAVH is necessary, but any of the others might be excluded at this stage and cointegration still retained. In order to test this, each of the three variables (LP, LPRODS and UPC) were dropped, one at a time, and cointegration tests were performed. These are reported in Table 2. In all cases the test statistics are considerably lower than in equation (3) and the residual correlograms do not strongly suggest stationarity. However the exclusion of both LPRODS and UPC are both passed by the CRDW test (at the 5 per cent level). Given the uncertainty surrounding the Granger and Engle critical values for this model, there may be a strong argument for relying more heavily on the informal evidence of the correlogram.

In order to estimate a valid ECM model of UK wage determination we must therefore include the full cointegrating vector in the level part of the model. That is to say, we must include the level of wages, prices, unemployment, productivity and average hours to achieve a stationary error process.

TABLE 2
Testing for the Exclusion of Three of the Variables

	Excluded variable		
	LP	*LPRODS*	*UPC*
CRDW	0.05	0.339	0.64
DF	−0.68	−2.648	−3.66
ADF	−1.43	−1.378	−2.14
R^2	0.9502	0.9968	0.9990
RCO 1	0.96	0.82	0.68
2	0.92	0.73	0.47
3	0.86	0.64	0.22
4	0.78	0.55	0.06
5	0.72	0.57	0.14
6	0.65	0.52	0.13
7	0.58	0.46	0.14
8	0.50	0.40	0.16
9	0.41	0.31	0.13
10	0.32	0.25	0.11
11	0.23	0.23	0.20
12	0.13	0.17	0.16

Before going on to look at the second stage equation there is a further complication which needs to be considered. Equation (3) is a valid cointegration regression involving five variables. In general, however, we would not expect it to be unique. It would have been quite in order to have used any of the four independent variables in (3) as the dependent variable in a regression. However, given the properties of OLS, the resulting equilibrium relationship implied by the regression would not normally be identical to (3). It is important therefore to know just how different the implied equilibrium relationship given by the different inversions of (3) would be. This question is investigated in Table 3 below, which shows the various inversions of equation (3); the table reports the various regressions rearranged so that LW is on the LHS for ease of comparison.

Estimating the equation in its different inversions produces different estimates of the equilibrium parameters, as we would expect. The interpretation of this result is not completely satisfactory at present; one approach might be to let these estimates define the limits of an equilibrium sub-space, so the true long-run equilibrium might be anywhere within the area defined by these points. An alternative inter-

TABLE 3

The Effects on the Equilibrium Relationship of Changing the Dependent Variable

	Coefficients					
Dependent variable	*Constant*	*LP*	*LPRODS*	*UPC*	*LAVH*	R^2
LW	2.88	1.02	0.93	−0.61	−1.79	0.9993
LP	2.79	1.03	0.88	−0.73	−1.78	0.9988
UPC	1.74	1.20	0.85	−3.52	−1.65	0.8508
LAVH	6.89	1.01	0.86	−0.57	−2.64	0.8096
LPRODS	2.28	0.966	1.21	−0.56	−1.66	0.9746

pretation rests on Stocks' (1985) theorem 3 which establishes that the estimates of the cointegrating regression are consistent but subject to a finite sample bias. This bias seems to be related to the overall goodness of fit of the regression, and so the regression with the highest R^2 should be subject to the smallest bias. Neither of these interpretations has so far been given a satisfactorily rigorous foundation, however, and so the two-stage procedure must be *ad hoc* at this stage. The estimation is continued on the basis of the equation normalized on LW which gave the highest R^2.

Having achieved a suitable specification of the cointegrating equation we can proceed to the second stage of the Granger and Engle procedure. If we define Z to be the derived residual from equation (3) we may then include these residuals in a standard ECM model. A fairly simple search procedure produced the following equation

$$\begin{aligned}\Delta LW = &-0.007 + 1.04\ EDP - 1.18\ \Delta^2 UPC_{-1} - 0.98\ \Delta LAVH \\ &(1.4)\quad (6.0)\qquad (1.4)\qquad\qquad (8.6) \\ &+ 0.22\ \Delta LW_{-2} - 0.26\ Z \\ &(2.9)\qquad\quad (3.3)\end{aligned} \tag{4}$$

IV estimates

DW = 1.99; BP(16) = 23.7; SEE = 0.0129

CHISQ(12) = 2.3; CHOW(64,12) = 0.22

Data period: 1965Q3-1984Q3.

where instrumental variable estimation has been used, following the suggestion of McCallum (1976), to allow for the simultaneity in expected future inflation (EDP). Some noteworthy features of this equation are the near unit coefficient on prices and the good out of

sample forecasting performance described by the CHOW and CHISQ statistics (reported from this model, estimated without the last 12 observations). BP(16) is the Box–Pierce test for a random correlogram.

In order to get some idea of how influential the two-step estimation procedure has been it seems sensible to relax the restriction implied by the cointegration regression and estimate a free ECM equation. Exactly the same dynamic specification as equation (4) has been used, to give

$$\begin{aligned}\Delta LW = &\ 1.02 + 1.1\ EDP - 1.3\ \Delta^2 UPC_{-1} - 1.01\ \Delta LAVH \\ &\ (1.5)\quad (2.2)\qquad (1.5)\qquad\qquad (7.3) \\ &+ 0.23\ \Delta LW_{-2} - 0.28\ LW_{-1} \\ &\ (2.6)\qquad\quad (2.1) \\ &+ 0.29\ LP_{-1} - 0.14\ UPC_{-1} - 0.55\ LAVH_{-1} \\ &\ (2.2)\qquad (0.6)\qquad\quad (2.0) \\ &+ 0.21\ LPRODS_{-1} \\ &\ (2.6)\end{aligned} \tag{5}$$

IV estimates

DW = 2.03 BP(16) = 24.9 SEE = 0.0130

CHISQ(12) = 65.5 CHOW(60,12) = 6.5

Data period: 1965Q3-1984Q3.

The implications of this regression are somewhat different from those of (3) and (4); this equation would suggest dropping the level of unemployment altogether, even though Table 1 showed that this had a major effect on the properties of the cointegrating regression. It is also interesting to note that the out-of-sample stability tests indicate considerable parameter instability for this equation. The coefficient on expected price inflation is also quite a lot larger than unity, suggesting that this equation does not exhibit derivative homogeneity in prices. If the object of this exercise were simply to carry out a normal estimation process, an obvious move at this stage would be to combine the levels terms in wages and prices into a real wage term. This restriction was easily accepted by the data and produced a large improvement in the parameter stability tests (CHISQ(12) = 7.9, CHOW(60,12) = 0.75). However, the coefficient on expected price inflation fell to 0.82 and the level of unemployment remained insignificant. Finally, let us consider the static long-run solution to the model (5):

$$LW = 3.64 + 1.03\ LP + 0.75\ LPRODS - 0.50\ UPC - 1.96\ LAVH \tag{6}$$

If we interpret the parameters in Table 3 as limiting bounds on the equilibrium sub-space then the coefficients on LP, LPRODS and

LAVH all lie within this space, and the coefficient on unemployment is just outside the range suggested by Table 3.

In conclusion, while the concept of cointegration is clearly an important theoretical underpinning to the error correction model there are still a number of problems surrounding its practical application; the critical values and small sample performance of many of the tests are unknown for a wide range of models; informed inspection of the correlogram may still be an important tool. The interpretation of the equilibrium relationship when it is not unique also presents some problems. Nevertheless in the example presented here the two-stage procedure seems to perform well and to offer a number of insights into the data in terms of the time series properties of the variables in isolation and in combination.

National Institute for Economic and Social Research

Date of Receipt of Final Manuscript: March 1986

REFERENCES

Davidson, J. E. H., Hendry, D. F., Srba, F. and Yeo, S. (1978). 'Econometric Modelling of the Aggregate Time Series Relationship Between Consumers' Expenditure and Income in the United Kingdom', *Economic Journal*, Vol. 88, pp. 661-92.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*, New York, Academic Press.

Granger, C. W. J. and Weiss, A. A. (1983). 'Time Series Analysis of Error-Correcting Models', in Karlin, S., Amemiya, T. and Goodman, L. A. (eds.), *Studies in Econometrics, Time Series and Multivariate Statistics*, New York, Academic Press.

Granger, C. W. J. and Engle, R. F. (1985). 'Dynamic Model Specification with Equilibrium Constraints: Cointegration and Error Correction', presented at the World Congress of the Econometric Society, Boston.

Hall, S. G., Henry, S. G. B. and Trinder, C. (1983). 'Wages and Prices' in Britton, A. J. C. (ed.), *Employment, Output and Inflation*, London, Heinemann.

Hall, S. G. and Brooks, S. J. (1985). 'The Use of Price Regression in the Estimation of Error Correction Models', forthcoming, *Economic Letters*.

Hendry, D. F. and Von Ungen-Sternberg, T. (1981). 'Liquidity and Inflation Effects on Consumers Expenditure'. In Deaton, A. S. (ed.), *Essays in the Theory and Measurement of Consumer's Behaviour*, Cambridge University Press.

Sargan, J. D. (1964). 'Wages and Prices in the UK: A Study in Econometric Methodology', in Hart, P. *et al. Econometric Analysis for National Economic Planning*, London, Butterworths.

Sargan, J. D. and Bhargava, A. (1983). 'Testing Residuals from Least Squares Regression for being Generated by the Gaussian Random Walk', *Econometrica*, Vol. 51, pp. 153-74.

Stock, J. H. (1985). *Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors*, Mimeo, Harvard University.

DATA APPENDIX

All the data are taken from the NIESR Database.

LW is the log of the wage space rate where

WAGERATE = WS/(EMP.AVHMF).B
WS ≡ total wages and salaries (£M Economic trends)
EMP ≡ total employment (UK) (Dept. of Employment Gazette)
AVHMF ≡ Average hours worked in manufacturing (Economic Trends)
B scales WAGERATE to be 1 in 1980.

LP is the log of the Consumer Price Index (Economic Trends)

LPROD is the log of Productivity defined as the ratio GDP (Economic Trends) to EMP.

LPRODS is a smoothed version of LPROD defined as

$$\text{LPRODS} = \frac{1}{8}\sum_{i=0}^{7}(\text{LPROD}_{-i})$$

LAVH is the log of average hours worked in manufacturing.

UPC is the percentage rate of unemployment ≡ UNEMP/(UNEMP + EMP) where UNEMP = registered unemployed (Dept. of Employment Gazette).