Misleading Regressions with Constructed Variables

By

David Shepherd[★] and Robert Dixon[†]

[★] Imperial College, University of London, 53 Prince's Gate, London SW7 2PG
e-mail: david.shepherd@ic.ac.uk

[†] Department of Economics, University of Melbourne, Victoria 3010, Australia

Abstract

It is common practice to examine empirical models in which one of the regressors is constructed as the weighted average or sum of a set of series that includes the dependent variable. Examples include models relating money and wealth, consumption and income and regional and national unemployment. In this paper we show that biased results are likely to be generated by such models and that the identified bias is distinct from the more familiar simultaneous equation bias. The theoretical arguments are illustrated with simulation experiments and as a practical example we consider the relationship between regional and national unemployment in Australia.

Key Words: Misleading Regressions. Constructed Variables. Estimation Bias. Regional Unemployment.

JEL Classification Numbers: C30 E24 R23.

Misleading Regressions with Constructed Variables

1. Introduction

It is common practice to formulate empirical models that include a regressor which is constructed as the weighted average or sum of a set of variables that includes the chosen dependent variable. There are some cases in which the formulation is suggested by theory, such as the relationship between consumption and income, where income includes consumption, or the relationship between money and wealth, where wealth includes money balances. In other cases, the formulation of the model is chosen as much by convenience as by theory, such as models relating regional and national unemployment, where the national unemployment rate is a weighted average of the various regional unemployment rates. Whatever the motivation behind the formulation of such models, in this paper we show that they are likely to generate misleading results because of an inherent bias in the parameter estimates. The estimation bias arises from the formulation of the model and is caused by the inclusion of the dependent variable in the constructed regressor. In view of its origin, and to distinguish it from the more familiar simultaneous equation bias, we refer to the identified bias as "construction bias". We argue that, where possible, the model should be reformulated to remove any potential construction bias. To facilitate the development of the theoretical arguments, we concentrate on just one of the examples mentioned above, casting the analysis in terms of the relationship between regional and national unemployment. We therefore begin by explaining why this relationship is considered to be of interest.

An important issue in the analysis of regional unemployment is whether movements in regional unemployment reflect the impact of region-specific shocks or shocks affecting the entire economy. If the behaviour of regional unemployment is largely explained by national factors, it suggests that policies to reduce unemployment in the regions are indistinguishable from national macroeconomic policies designed to affect general demand and supply conditions across the economy. In contrast, if there are strong region-specific components explaining the behaviour of regional unemployment, the case for region-specific employment policies is stronger.

Previous studies examining the above problem have been based almost entirely on an examination of the relationship between each regional unemployment rate and the corresponding national rate, where the national rate effectively acts a proxy for the for the aggregate forces driving the economy. The view that insight can be gained by regressing regional unemployment rates on the national rate was developed by Thirlwall (1966) and by Brechling (1967) in a paper published in this journal. Since then, the notion that the behaviour of regional unemployment is best examined in relation to the national unemployment rate rather than the unemployment rates of other regions has become widespread. Relevant examples include Byers (1990), Chapman (1991), Groenewold (1991), Martin (1997) and Debelle and Vickery (1998). The typical conclusion from such studies is that a large proportion of the variation in regional unemployment is explained by the national unemployment rate, suggesting that national rather than region-specific forces are dominant[1]. In contrast, we argue that the procedure of comparing regional and national unemployment rates can give rise to serious statistical problems and that the results

[1] For example, in a recent study of Australian unemployment, Debelle and Vickery (1998) report that “at least three-quarters of the variation in a state's unemployment rate is attributable to variations in national unemployment”. Based on this finding, they suggest that “movements in the national unemployment rate explain most of the variation in state unemployment rates, suggesting that aggregate, rather than state specific factors, are most important in understanding Australia's high aggregate unemployment rate”

are likely to be biased towards the conclusion that national forces are dominant in explaining movements in regional unemployment. More generally, we argue that similar bias problems arise whenever the model incorporates regressors that are constructed to include the dependent variable.

The plan of the paper is as follows. In the next section we consider the nature of the statistical problems that arise when examining the regional-national unemployment relationship and illustrate the more general importance of these statistical problems via a series of monte carlo simulation experiments. Section 3 examines by way of example the behaviour of unemployment in the Australian states and suggests appropriate procedures to determine the relative importance of national and region-specific factors. The final section provides a brief summary and discussion of the points raised in the paper.

2. Analytical and Simulation Results

Continuing with the theme of the introduction, in this section we examine the nature of the construction bias problem in the context of a model that relates regional and national unemployment rates. The question we are concerned with is whether regional unemployment movements are explained by national factors, affecting the entire economy, or factors that are essentially region-specific. The basic idea is that time series movements in the rate of unemployment in each region may reflect the response both to innovations affecting the entire economy and innovations affecting only that region. This is summarised in equation (1)

$$U_{it} = \beta_i N_t + R_{it} + e_{it} \qquad (1)$$

where U_{it} is a time series of the unemployment rate in region i, N_t is some national stochastic process affecting unemployment in all regions, with an associated regional impact parameter β_i, R_{it} is a stochastic process specific only to region *i*, and e_{it} is some additional noise process.

The processes generating movements in the regional unemployment rates shown on equation (1) are of course not observed independently and the problem is to determine their relative importance when the only information available is the time series paths of the various U_{it} series. As discussed earlier, the standard approach to this problem is to take the national unemployment rate as a proxy for the national process N_t and then to examine the relationship between U_{it} and N_t using correlation or regression methods, with the importance of the national process assessed according to the proportion of the variation in U_{it} explained by N_t. As we shall see, a model which relates a component of an aggregate to the behaviour of the aggregate itself can give rise to important statistical problems that go beyond the simultaneity problem that most people would see as being potentially present in this situation.[2]

2.1 Correlation Analysis

In this section of the paper we demonstrate that it is quite possible to observe an apparently significant correlation between the regional rates and the national rate, even if the regional rates themselves are uncorrelated, leading to a situation in which a researcher may falsely conclude that national shocks are present even when the only innovations in the system are region-specific. This can be seen by considering the standard correlation coefficient.

For the sake of argument, consider two time series U_1 and U_2 generated by independent (uncorrelated) white noise processes with mean zero and with constant

[2] Johnston (1979) was one of the first to explore the difficulties which can arise when a component of a spatial aggregate is regressed on the aggregate value and the sensitivity of the outcome to size.

(and identical) variance σ_u^2. Now consider a third series Z which is a weighted sum of the two U series such that $Z = \alpha U_1 + (1-\alpha)U_2$, with 0 $<\alpha<$1. The correlation between U_1 and Z is defined as

$$r_{uz} = \frac{Cov\{U_1, Z\}}{\sqrt{Var\{U_1\}}\sqrt{Var\{Z\}}} \tag{2}$$

$$= \frac{Cov\{U_1, (\alpha U_1 + (1-\alpha)U_2\}}{\sqrt{Var\{U_1\}}\sqrt{Var\{(\alpha U_1 + (1-\alpha)U_2)\}}} \tag{3}$$

The covariance of U_1 with itself is simply its variance and, given that U_1 and U_2 are by assumption uncorrelated (zero covariance) and have a common variance σ_u^2, the correlation coefficient in this case reduces to

$$r_{uz} = \frac{\alpha}{\sqrt{1-2\alpha+2\alpha^2}} \quad > \alpha \quad \text{for} \quad 0 < \alpha < 1 \tag{4}$$

This shows that there is an inevitable correlation between any one of the series and the weighted average (or aggregate) of the two and that the degree of correlation is related to (and is actually higher than) the size of the α weighting. In cases where the constructed aggregate variable is used, any judgement about the significance of correlations between any one of the components of the aggregate and the aggregate itself, would therefore have to allow for the correlation generated by the weighting procedure used to construct the aggregate variable. For example, with α at say 0.3, equation (4) tells us that the implied correlation is 0.3939 and any correlation below this should certainly not be regarded as significant.

To illustrate the impact of the weighting procedure, we undertook a series of monte carlo experiments, calculating the correlation between independently generated white noise series and a third series constructed as a weighted average of the two. Based on independent random draws from the standard normal distribution, we constructed two uncorrelated series U_1 and U_2 for sample sizes ranging from 50 to 1000. The U series were then used to construct a third series $Z = \alpha U_1 + (1-\alpha)U_2$, with α varying from 0.0 to 0.9 in steps of 0.1. We then calculated the correlation coefficient between U_1 and Z for each sample size. This process was repeated 5000 times for each sample size.[3] The results are summarised in Tables 1 and 2.

[TABLE 1 NEAR HERE]

Table 1 reports the mean realisation of the correlation coefficient for each sample size, together with the theoretical correlation (shown as N=∝) suggested by equation (4). The mean realisations from each of the 5000 replications are very much in line with the theoretical correlation, indicating that the weighting process induces a correlation which is actually higher than the α weighting. The results are also consistent across the various sample sizes.

In addition to the correlation induced by the weighting procedure, we need to allow also for any chance correlation that might be present between the series. Following a well-known result from Bartlett (1946), to rule out any chance correlation at say the 5% significance level, for a sample size of N we would normally regard any calculated correlation as significantly different from zero only if its absolute value exceeded approximately $2/\sqrt{N}$. In the case of the constructed variable, the series are related via the weighting procedure and it is necessary to add the chance correlation to the correlation implied by the weighting before reaching any judgement about significance. To identify the importance of the additional chance factor in the present context, we used the full distribution of the simulation results to determine the

[3] The outcome for $N = 84$ is included to cover the sample size of the empirical examples reported later.

appropriate 5% and 1% critical values for the correlation coefficient. The results are reported in Table 2.

[TABLE 2 NEAR HERE]

The critical values shown in Table 2 are measured as the cut-off points for the 5% and 1% tails of the empirical distributions of the correlations for each sample size. In this case, as one would expect, the results do vary significantly with the sample size and it is only in very large samples that the 5% and 1% critical values approach the mean expected values of the correlation coefficient. Relating these results to our discussion of regional unemployment, for a sample size of say 100, using the standard formula, the 5% significance level would be approximately 0.2. In contrast, our results suggest that for a region accounting for say 30% of the national labour force, the correlation between the regional and national unemployment rates would have to exceed 0.53 before it should be regarded as significantly different from zero at the 5% significance level.

2.2 The Regression Model

An alternative way to examine the relationship between regional and national unemployment is via the standard regression model. Let us suppose that unemployment rates in the regions are driven by a common factor or process and that the true relationship can be expressed in terms of the regression model shown in equation (5)

$$U_{1t} = U_{2t}\theta + e_t \tag{5}$$

where U_{1t} and U_{2t} are column vectors, θ is a constant system parameter and e_t is a column vector representing some noise process that might incorporate any region-specific factors. The standard least squares solution for this model is

$$(U_2^T U_{2t})^{-1}(U_{2t}^T U_{1t}) = \hat{\theta} \tag{6}$$

Now, for the sake of argument, let us suppose $e_t = 0$ so that we have a true error-free model

$$U_{1t} = U_{2t}\theta \tag{7}$$

In this case, θ is a constant parameter and its value can easily be determined simply by dividing any of the U_{1t} values by the corresponding U_{2t} value. However, for the sake of comparison with the regression model, it is helpful to solve for θ in a slightly roundabout way. Premultiplying both sides of (7) by the transpose of U_{2t} and solving by matrix inversion gives

$$(U_{2t}^T U_{2t})^{-1}(U_{2t}^T U_{1t}) = \theta \tag{8}$$

Equation (8) is equivalent to the standard least squares solution for the regression model, except that there is no error present and we have an exact solution for θ.

Now let us consider the case in which the model relates one of the U_t series (one of the regions) to the constructed variable Z_t

$$Z_t = \alpha U_{1t} + (1-\alpha)U_{2t} \tag{9}$$

where Z_t is equivalent to the national unemployment rate, constructed as a weighted average of the regional rates (with the weightings given by α). The regression model relates the regional rate to the national rate, with some error term u_t

$$U_{1t} = Z_t\beta + u_t \tag{10}$$

Again thinking of (10) for the moment as a model with no error term, we have

$$U_{1t} = Z_t \beta \tag{11}$$

And the solution for β can be expressed as

$$(Z_t^T Z_t)^{-1}(Z_t^T U_{1t}) = \beta \tag{12}$$

which is again equivalent to the least squares solution with no error influence. Now, the relationship between θ and β can be seen if we re-write (12) more explicitly as

$$U_{1t} = [\alpha U_{1t} + (1-\alpha)U_{2t}]\beta \tag{13}$$

Given the true relationship between U_{1t} and U_{2t} shown by equation (7), we can write (13) as

$$U_{1t} = [\{\alpha\theta + (1-\alpha)\}U_{2t}]\beta \tag{14}$$

And the solution for β is

$$(U_{2t}^T U_{2t})^{-1}(U_{2t}^T U_{1t}) = \{\alpha\theta + (1-\alpha)\}\beta \tag{15}$$

We know from (8) that the left-hand side of (15) is the solution for θ and hence

$$\beta = \theta / \{\alpha\theta + (1-\alpha)\} \tag{16}$$

Equation (16) shows that the model with the constructed Z_t variable yields a solution for β which is a scaled version of θ, where the scaling depends on the size of the α weighting used to construct Z_t. For example, with θ at say 0.7, the values of β for α at 0.2 and 0.5 would be 0.7447 and 0.8235 respectively. In practical terms, this means that the inclusion of the constructed aggregate variable introduces a degree of bias to the model solution, with the bias increasing the size of the model parameter by some factor related to the size of the α weight. The special case in which $\beta = \theta$ occurs only at $\alpha = 0$, which is equivalent to the original model described by equation (7).

The point to note at this stage is that the models with the constructed Z_t variable (equations 10 and 11) contain no more information than the U_{1t}-U_{2t}

models (equations 5 and 7). Since β is a consistently biased estimate of θ, with the degree of bias rising with the weighting factor α, it might appear that the value of the true system parameter θ could be recovered from β if the α value is known. While this is true for the model with no errors, the presence of the error term makes it difficult to determine the true value of the system parameter and we argue later that it is preferable to concentrate on the U_{1t}-U_{2t} relationship directly. An additional reason for avoiding the model with the constructed variable is that the weightings on the regions (the α values) actually vary over time. Most of the system parameters in economic models are probably time-varying and the assumption of constant parameters is a simplification that hopefully does not distort the results. As a practical matter it is preferable to avoid introducing unnecessary inaccuracies, which again suggests that we should if possible avoid the model with the constructed variable.

The problem of determining the value of the system parameter θ is more complicated when the error term is present. Returning to equation (5)

$$U_{1t} = U_{2t}\theta + e_t$$

it can be shown that

$$\hat{\theta} = (U_{2t}^T U_{2t})^{-1}(U_{2t}^T U_{1t}) + (U_{2t}^T U_{2t})^{-1}(U_{2t}^T e_t) \qquad (17)$$

and $E[\hat{\theta}] = \theta$ so long as $E[(U_{2t}^T U_{2t})^{-1}(U_{2t}^T e_t)] = 0$, which we assume is the case for the model described by equation (5). Now consider the model with the constructed Z variable. The assumed model, including the error term, is shown by equation (10). Given the definition of Z in equation (9) and the true relationship described by (5) we can see that equation (10) is equivalent to the implicit regression

$$U_{1t} = \{\alpha\theta U_{2t} + \alpha e_t + (1-\alpha)U_{2t}\}\beta + (1-\alpha)e_t \qquad (18)$$

where in (10),

$$Z_t = \{\alpha\theta U_{2t} + \alpha e_t + (1-\alpha)U_{2t}\} \qquad \text{and} \qquad u_t = (1-\alpha)e_t$$

The least squares solution of (10) is given by

$$(Z_t^T Z_t)^{-1}(Z_t^T U_{1t}) = \hat{\beta} \tag{19}$$

and it can be shown that

$$\hat{\beta} = (Z_t^T Z_t)^{-1}(Z_t^T U_{1t}) + (Z_t^T Z_t)^{-1}(Z_t^T u_t) \tag{20}$$

and taking expectations we have

$$E[\hat{\beta}] = \beta + E[(Z_t^T Z_t)^{-1}(Z_t^T u_t)] \tag{21}$$

The statistical properties of the least squares estimator $\hat{\beta}$ hinge on the relationship between Z_t and u_t and $\hat{\beta}$ is an unbiased estimator of β only if $E[Z_t^T u_t] = 0$. Given that (10) is equivalent to the implicit regression (18), the expectation of $Z_t^T u_t$ can be written as

$$E[Z_t^T u_t] = E[\overset{n}{\underset{t=1}{}} \alpha\theta U_{2t} e_t(1-\alpha) + \alpha(1-\alpha)e_t^2 + U_{2t}(1-\alpha)e_t(1-\alpha)] \tag{22}$$

Given the assumption that there is no correlation between U_{2t} and e_t in the true model given in equation (5), the first and third multiplicative terms on the right-hand side of (22) are equal to zero and the expectation $E[Z_t^T u_t]$ reduces to

$$E[Z_t^T u_t] = \alpha(1-\alpha)\sigma_e^2 \tag{23}$$

where σ_e^2 is the variance of the error term in the true model. Using this result, the expression for the least squares estimator $\hat{\beta}$ shown by equation (21) can be written as

$$E[\hat{\beta}] = \beta + E[(Z_t^T Z_t)^{-1}\{\alpha(1-\alpha)\sigma_e^2\}] \tag{24}$$

and $E[\hat{\beta}] = \beta$ only if $\alpha = 0$, which means that the estimated model is equivalent to the true model described by (5). Another way of putting this is to say that the inclusion of a fraction of the dependent variable on the right-hand side of the model induces a degree of correlation between the Z_t regressor and the error term and the resulting estimate of the system parameter is likely to be biased and inconsistent, with the degree of bias related to the size of α. The two points to note here are, first, that the error-induced bias is in addition to the systematic model-bias discussed earlier and, secondly, that (24) suggests that the error-induced bias is greatest in the region of $\alpha = 0.5$.

The implication of our analysis is that the use of the constructed aggregate variable generates a degree of bias in the parameter estimates which is directly related to the weight that the dependent variable carries in the constructed regressor. In so far as this "construction bias" arises partly from an induced correlation between the regressor and the error term, it is similar to the more familiar simultaneous equation bias present in reduced form models of consumption, money demand and the like. The two forms of bias are however different. Simultaneous equation bias arises when the variables are related in a stochastic behavioural system that implies a correlation between the regressor and the error term. In contrast, the bias we have identified arises because the variables are related by a construction process (via an identity) that leads to a non-stochastic (deterministic) bias that is unrelated to the error-regressor problem, and an additional degree of bias arising from the error-regressor correlation induced by the construction of the regressor. The latter component of the construction bias (arising form the error-regressor correlation) is equivalent to a form of simultaneous equation bias, but it arises from a different source, while the former component (the deterministic bias) is completely separate from any simultaneity bias[4].

[4] Another way of putting this is to say that the deterministic component of construction bias is always associated with something equivalent to simultaneous equation bias, but simultaneous equation bias need not be associated with anything equivalent to the deterministic component of construction bias.

We come back to this matter in the next section, when we discuss the implications of the construction bias for cointegrations tests.

In cases where simultaneous equation bias is present, or there is a similar bias induced by measurement error, the recommended procedure (to obtain consistent parameter estimates) is to use instrumental variables. If this procedure were applied in the present context, the obvious instrument for Z_t would be the original U_{2t} variable (or variables). This suggests that instrumental variable estimation is in fact unnecessary and that the whole problem can be avoided simply by estimating (5) directly[5] rather than the Z_t model of equation (10). More generally, our analysis suggests that the entire bias problem can be avoided by reformulating the model so as to avoid the use of the dependent variable in the constructed regressor[6].

To illustrate the impact of the weighting procedure in the regression model, we consider two sets of simulations. In the first set we consider the impact of including the constructed variable when the component parts are completely uncorrelated. We first generated two independent (uncorrrelated) white noise series U_1 and U_2, based on random draws from the standard normal distribution and a sample size of 100. We then constructed a third series as a weighted average of the two, $Z = \alpha U_1 + (1-\alpha)U_2$, with α varying from 0.0 to 0.9 in steps of 0.1, and then estimated the regression model $U_{1t} = Z_t\beta + u_t$ for each α value. This procedure was repeated 5000 times. The results are reported on Table 3. The first three columns of the table show the mean realisation of $\hat{\beta}$, together with the mean realisation of the

[5] An additional reason for avoiding the Z_t model variable is that the estimates derived from the instrumental variable procedure are likely to remain biased in small samples even though they are consistent.

[6] In the case of the consumption-income and money-wealth examples mentioned in the introduction, our analysis would favour the use of models with the consumption/income and money/wealth ratios as the dependent variables, in preference to the more usual formulations.

associated t statistic and the proportion of t values that were in excess of the 5% critical value (CV). The final two columns show respectively the mean realisation of the R^2 and the 5% critical value implied by the full distribution of R^2 values.

[TABLE 3 NEAR HERE]

The results for the uncorrelated series are quite straightforward, indicating that $\hat{\beta}$ and the R^2 values rise in line with the α weighting, with a particularly pronounced increase in the 0.2 - 0.5 range for α as suggested by equation (24).

The second set of experiments is designed to show the potential bias in the parameter estimates and R^2 when there is a true relationship between the series used to construct Z_t. In this case, the true Model is $U_{1t} = c + \theta U_{2t} + e_t$ and the estimated model is $U_{1t} = c + \beta Z_t + u_t$ where $Z_t = \alpha U_{1t} + (1-\alpha)U_{2t}$. The data for U_{1t} was derived as discussed earlier and U_{2t} was constructed with a θ value of 0.7. We then estimated the β parameter for the different values of α, with the procedure repeated 5000 times. Table 4 reports the $\hat{\beta}$ estimates for the range of α values, together with the mean realisations of the R^2 for each equation. The table also shows the difference between the $\hat{\beta}$ estimates and the true (error-free) values of θ and β. Note that for α=0 the estimated model is equivalent to the true model and $\hat{\beta}$ is the same as $\hat{\theta}$.

[TABLE 4 NEAR HERE]

The simulation results show quite clearly the impact of the two forms of bias discussed earlier. The first row, for $\alpha = 0$, shows the results for the true model, with an accurate estimate of θ and, based on the chosen signal-noise ratio of the simulations, an R^2 of 0.33. The results for the other α values show quite clearly the impact of both the systematic model bias and the error-regressor correlation bias generated by the use of the constructed Z_t variable. The results also confirm what was suggested earlier, by equation (24), that the error-induced bias is greatest in the region of $\alpha = 0.5$.

Taken together, the simulation results suggest that regressions of regional unemployment rates against the national rate are likely to generate biased results that make it difficult to identify the true relationship between the regions and, in particular, the extent to which they are driven by national or region-specific processes. More generally, the results suggest that the identified bias is potentially significant in any model that includes the dependent variable in the constructed regressor.

2.3 Cointegration Tests

The preceding analysis is implicitly based on the assumption that the variables in question are stationary, or have been rendered stationary by an appropriate transformation, so that the regression results do not suffer from any spurious correlation problem of the kind discussed by Granger and Newbold (1974) and Phillips (1986). For completeness, to cover the case of non-stationary variables, we need to consider whether the inclusion of the constructed variable in the regression model has any consequences for cointegration tests. In particular, what we need to know is whether the inclusion of the constructed variable on the right-hand side leads to any increase in the number of cases for which cointegration is incorrectly suggested, when neither of the independent series are in fact cointegrated.

Let us suppose that the regional unemployment rates U_{1t} and U_{2t} are driven by stochastic trend-generating processes T_{1t} and T_{2t} and additional stationary processes e_{1t} and e_{2t}

$$\begin{aligned} U_{1t} &= T_{1t} + e_{1t} \qquad & T_{1t} &= T_{1t-1} + v_{1t} \\ U_{2t} &= T_{2t} + e_{2t} \qquad & T_{2t} &= T_{2t-1} + v_{2t} \end{aligned} \tag{25}$$

where v_{1t} and v_{2t} are white noise innovations driving the (random walk) trend processes. If the series are cointegrated, it means that they are driven by a common trend and it should be possible to identify a linear combination of the series with no identifiable trend component. In the case of the series described by (25), the linear combination is

$$U_{1t} - \lambda U_{2t} = T_{1t} - \lambda T_{2t} + e_{1t} - \lambda e_{2t} \tag{26}$$

The series are regarded as cointegrated if there is a common factor of proportionality ϕ in the trend such that

$$T_{2t} = \phi T_{1t} \tag{27}$$

In this case, the linear combination (26) can be written as

$$U_{1t} - \lambda U_{2t} = T_{1t} - \lambda\phi T_{1t} + e_{1t} - \lambda e_{2t} \tag{28}$$

and with $\lambda = 1/\phi$ equation (28) reduces to the stationary series

$$U_{1t} - \lambda U_{2t} = e_{1t} - \lambda e_{2t}$$

In contrast, if (27) does not hold, there is no value of λ that removes the trend from the linear combination (28). The test for cointegration is thus a test of whether there is some cointegrating vector $[1 \quad \lambda]$ for which a linear combination of the U_{it} series is stationary.

Our concern is whether the use of the constructed Z_t is likely to introduce any bias into the cointegration test described above. Noting the descriptions of U_{1t} and

U_{2t} given by (25) and using the definition of Z_t given by equation (9) in place of U_{2t} in equation (26), the linear combination of the series is

$$U_{1t} - \lambda Z_t = T_{1t} - \lambda\alpha T_{1t} - \lambda(1-\alpha)T_{2t} + e_{1t} - \alpha e_{1t} - \lambda(1-\alpha)e_{2t} \tag{29}$$

If the trends are common as in (27) the linear combination in (29) can be written as

$$U_{1t} - \lambda Z_t = T_{1t} - \lambda\{\alpha + (1-\alpha)\phi\}T_{1t} + e_{1t} - \alpha e_{1t} - \lambda(1-\alpha)e_{2t} \tag{30}$$

and with $\lambda = 1/\{\alpha + (1-\alpha)\phi\}$ equation (30) reduces to the stationary series

$$U_{1t} - \lambda Z_t = (1-\alpha)e_{1t} - \lambda(1-\alpha)e_{2t}$$

In contrast, if (27) does not hold, so that the trends are uncommon, there is again no value of λ that renders the series stationary.

The implication of the above is that the use of the constructed Z_t series leads to a re-scaling of the cointegrating vector, which depends on the size of the weighting factor α. This is equivalent to the systematic bias discussed earlier in relation to equation (16). While this re-scaling in itself should have no impact on the on the direction of the cointegration test result, in practice we need to recognise that the power of the test in any practical setting depends on the extent to which the trend "signals" (the T_1 and T_2 terms) are overlain by any additional noise (the e_1 and e_2 terms). Generally speaking, the lower the signal-noise ratio, the lesser is the power of the test to determine whether or not the series share common trends (Cochrane, 1991).

In their original exposition of the cointegration test, Engle and Granger (1987) demonstrate that the usual simultaneous equation bias problem should have no impact on the results of the OLS cointegrating regression and that the parameter estimates remain consistent and actually approach the true asymptotic values more rapidly than in the normal case. We have already seen that the use of the constructed variable, as opposed to the independent series, leads to a re-scaling of the cointegrating vector. What we are most concerned with, however, is whether the use of the constructed variable has any impact on the power of the cointegration test[7]. Bearing in mind the power problem discussed by Cochrane (1991), it seems likely that the use of the constructed variable in the cointegration test, as opposed to the independent series, will have no impact on the results so long as the signal-noise ratios are similar for the two independent series, because the overall SNR of the model with the constructed variable will then be similar to the SNR of the model with the two independent series. If the SNRs of the two series are very different at the outset, however, the re-weighting implied by the use of the constructed variable will alter the SNR of the model and may consequently affect the result of the cointegration test.

To check the above reasoning, we undertook a series of simulation experiments based on two series U_1 and U_2 with a sample size of 100, generated by independent (unrelated) random walks. We first examined models with no additive noise, initially with stochastic trends of equal variance, and then with the variance of T_1 approximately four times the variance of T_2. We then examined models as above, but with added white noise (with noise variances broadly the same in each series). We then examined two further models, with the U_1 and U_2 series constructed first with similar trend variances and different noise variances and secondly with both different noise variances and different trend variances. In all cases, the white noise components were constructed with a zero mean. For the cases in which the trend and/or noise variances differ, we structured the differences arbitrarily in the ratio of 4/1. In summary, the simulation experiments are based on models of the following structure:

[7] If it does, it provides a further indication of the difference between construction bias and simultaneous equation bias.

1. $Var(T_1)/Var(T_2) = 1/1$, $e_{1t} = e_{2t} = 0$

2. $Var(T_1)/Var(T_2) = 4/1$, $e_{1t} = e_{2t} = 0$

3. $Var(T_1)/Var(T_2) = 1/1$, $Var(e_1)/Var(e_2) = 1/1$

4. $Var(T_1)/Var(T_2) = 4/1$, $Var(e_1)/Var(e_2) = 1/1$

5. $Var(T_1)/Var(T_2) = 1/1$, $Var(e_1)/Var(e_2) = 1/4$

6. $Var(T_1)/Var(T_2) = 4/1$, $Var(e_1)/Var(e_2) = 1/4$

For each of the above models, we tested for cointegration between U_1 and U_2 using the residual-based approach[8] of Engle and Granger (1987). We then constructed a third series Z as a weighted average of U_1 and U_2 (with the α weights as before ranging from 0.1 to 0.9 in steps of 0.1) and tested for cointegration between U_1 and Z. This procedure was repeated 5000 times for each model. The null of no cointegrating relationship is rejected or not rejected according to the value of the t statistic derived from a unit root test on the residuals from an OLS regression, based on the procedure suggested by Dickey and Fuller (1979), using the 5% critical value reported by Davidson and Mackinnon (1993). From what we said earlier, the prior expectation is that the cointegration results should be similar for the first two models with no additive noise and the third model in which the series have similar trend variances and similar noise strengths[9]. For the remaining three models, the use of the constructed Z variable implies that the SNR will vary with the size of the α and this may well influence the cointegration tests.

The results of the experiments are summarised on table 5. The first row of each block of the table shows the α weights and the second row reports the mean

[8] We used the Engle-Granger approach rather than the more efficient VAR procedure of Johansen (1995) on the grounds of computational simplicity.

realisations ($\bar{t}$) of the cointegration tests for each value of α. The third row of each block shows the proportion of the full distribution of the t statistics for which cointegration is (incorrectly) suggested at the 5% significance level ($t < t_{0.05}$).The first column of each block (α=0.0) shows the results of the test for cointegration between the underlying series U_1 and U_2. The remaining columns report the results of the tests for cointegration between U_1 and the constructed Z variable, over the range α=0.1….0.9.

[TABLE 5 NEAR HERE]

For the models with no added noise (models 1 and 2) the results show quite clearly that the cointegration test is extremely robust with respect to the change in model specification, with little or no variation in the percentage of invalid rejections of the null, even when the α weighting is as high as 0.9. In the case of model 3, the presence of additive noise reduces the power of the test, in the sense that there is an increase in the percentage of cases for which the null of cointegrating relationship is incorrectly rejected at the conventional 5% significance level. However, the important point is that the results are again invariant to the change in specification from the U_1-U_2 model to the various U_1-Z models. This suggests that the cointegration test is robust with respect to the change in model specification so long as the noise in the system is evenly spread across the two underlying series, such that the change in specification does not significantly alter the SNR. For the remaining models (4, 5 and 6) the unequal trend and noise variances of the underlying series do imply significant

[9] These first three cases are the ones for which the overall SNR is unlikely to change significantly

changes in the SNRs and in these cases the results are sensitive to the specification of the model. In particular, it appears that the null of no cointegrating relationship is more likely to be incorrectly rejected in the models with the constructed Z variable and that the probability of an invalid rejection increases directly with the α weighting. The implication is that the test for cointegration should where possible be conducted in terms of the underlying series rather the constructed Z variable, particularly in cases where significant differences in trend or noise variances are suspected.

3. The Case of Australian Regional Unemployment

To illustrate the arguments of the previous section, we consider the behaviour of regional and national unemployment in the main states of the Australian Commonwealth (AUS). The states are: New South Wales (NSW), Victoria (VIC), Queensland (QLD), South Australia (SA), Western Australia (WA) and Tasmania (TAS). Our approach is to apply correlation and regression methods to analyse the relationship between the national unemployment rate and the unemployment rate of each state. We discuss the problems in interpreting the results in the light of the arguments of the previous section and then discuss appropriate procedures to determine the relative importance of national and regional shocks. The series in question are seasonally adjusted quarterly unemployment rates for persons covering the period 1978Q2 to 1999Q1. The proportions of the national labour force accounted for by each state vary over time and in 1988, the middle year of our study, they were: NSW, 34.0%; VIC, 26.0%; QLD, 16.4%; SA, 8.5%; WA, 9.6%; TAS, 2.6%. There is thus considerable variation in the weightings across the states.

between the U_1-U_2 and U_1-Z models.

We begin by examining the contemporaneous correlations between the states and between the states and the national rate. These are reported in Table 6. The correlations are all relatively high, indicating significant positive co-movement in unemployment across the states and in relation to the national unemployment rate. The preliminary points to note are, first, that the correlations for each state with the national unemployment rate are all higher than the cross-state correlations and, secondly, that the magnitudes of the correlations follow the same order as the weightings of the states with respect to the national unemployment rate, with the highest and lowest correlations with AUS recorded for NSW and TAS.

[TABLE 6 NEAR HERE]

The problem with interpreting the information in Table 6 is that the correlations are strictly meaningful only if the series are stationary. Previous studies of Australian unemployment suggest that both the national and state series should for practical purposes be regarded as following a unit root process, or at least a near-unit root process.[10] Preliminary data testing using the Dickey-Fuller and Phillips-Perron tests indicated the presence of unit roots in the series and we do therefore have the possibility that the correlations in Table 6 are essentially spurious. This immediately raises the question of whether the series are cointegrated and to determine this we tested for common trends in the series, following the Engle-Granger procedure discussed earlier.

[10] Numerous studies of aggregate Australian unemployment suggest a unit root or near unit root process (for example: Mitchell, 1993; Crosby and Olekans, 1998; Groenewold and Hagger, 1998; Gruen, Pagan and Thompson, 1999) whilst Groenewold & Hagger (1995) and Debelle & Vickery (1998 & 1999) and Dixon and Shepherd (2000) find that all state unemployment rates are non-stationary.

According to our analytical discussion in the previous section, and the simulation results, the test for cointegration is robust with respect to whether the model is cast in terms of the relationships between the states or between each of the states and the national rate only if the change in specification does not significantly alter the SNR of the model. While the SNR is not observable, prior examination of the data indicated that the autoregressive parameters and variances of the individual series (which measure their persistence and volatility characteristics) are similar, suggesting that they have similar structures and that the change in specification may perhaps not alter the SNRs dramatically. In any event, table 7 reports the results of a series of bivariate cointegration tests for each state with respect to both the other individual states and the national unemployment rate.

[TABLE 7 NEAR HERE]

The first column of results shows the cointegration tests for each state in relation to the national unemployment rate. These suggest that there are no cointegrating relationships with the possible exception of the SA-AUS relationship, which is near the test borderline. The interesting point to note is that the results for the state-national tests are confirmed by the state-state tests, which show no cointegrating relationships between the states, except (again) for some borderline results with respect to South Australia. This suggests that long-run unemployment movements have not followed a common trend path across the states[11].

[11] We should perhaps emphasise that, strictly speaking, the results imply only that we can't reject the null of no cointegrating relationship. This does not rule out the possibility that the series are cointegrated and that the failure to reject the null is a consequence of the relatively low power of the test, particularly in small samples.

Given the general absence of any cointegrating relationships across the states, we do indeed have the possibility that the correlations in Table 6 are spurious and so the appropriate procedure is to look instead at the relationships between the (stationary) first difference of the series. These are reported in Table 8.

[TABLE 8 NEAR HERE]

The results suggest that changes in unemployment in the states are positively related to changes in the national unemployment rate, with correlations that at first sight appear highly significant. While this is probably the case, we should note that the highest correlations are again recorded for the states which account for the largest proportion of the national unemployment rate and, in view of the simulation results reported earlier in Tables 1 and 2, it is by no means clear whether NSW, VIC and QLD should be regarded as being more strongly driven by national forces than are SA and WA. Looking at the results for the inter-state relationships, which do not suffer from a weighting bias, the correlations suggest a much more even pattern of unemployment movement, with the exception of TAS, which has generally lower correlations than any other state.

Following the direction of our discussion in the last section, we consider next the unemployment relationships in terms of the regression model. In view of the fact that the series are non-stationary and are not cointegrated, the appropriate procedure is to examine the relationships between the first differences of the series. Table 9 reports the results of a series of regressions, equivalent to equation (10) of the last section, in which changes in the unemployment rates for each state in turn are regressed on a constant *c* and changes in the national unemployment rate ΔZ_t

$$\Delta U_{it} = c + \beta_i \Delta Z_t + u_{it}$$

As we discussed in the first section of the paper, the idea is to determine whether or not unemployment movements in the states are driven by national rather than region-specific forces.

The second and third columns of Table 9 show the constants and the β parameters for each state, with *t* statistics in parenthesis. The fourth and fifth columns show respectively the conventional R^2 for each regression and the LM test for first-order serial correlation in the residuals. We also tested for the significance of lags in the regional and national rates, but none were significant, with the exception of a marginally significant *t* value on the lagged national unemployment rate in the VIC equation. In all other cases there was no evidence to support the inclusion of a dynamic structure in the model. This confirms the impression given by the serial correlation tests.

[TABLE 9 NEAR HERE]

At first sight the regression results suggest that changes in unemployment in all of the states are driven partly by national forces, reflected in movements in the national unemployment rate, but that the importance of national forces varies considerably across the states. For example, changes in the national unemployment rate appear to account for 73% of the variation in unemployment changes in NSW, compared to less than 10% for TAS. In general, it appears that the larger states have the closest association with the national unemployment rate and the parameter values and R^2 figures are ranked in the same order as the weightings attached to the states in the calculation of the national unemployment rate. While this may reflect the fact that the larger states have a genuinely closer association with the national forces generating unemployment movements, it may also reflect in part the bias problems discussed in previous sections of the paper.

Our earlier analysis suggests that the weighting procedure used to construct the regressor leads to results that consistently over-estimate the strength of the relationship between each state and the national rate and that the degree of bias rises as the weighting increases and is particularly acute in the middle ranges. Given the relatively large weights attached to NSW and VIC, and to a lesser extent QLD, it is pertinent to ask whether the relatively high parameter and R^2 values for those states, particularly in comparison with SA and WA, do indicate a closer relationship with the national shocks affecting unemployment or whether they mainly reflect the bias problems arising from the use of the constructed national unemployment variable. For example, the α weightings for VIC and NSW in the middle of our sample period are approximately 0.26 and 0.34 respectively and the simulation results reported in Tables 3 and 4 suggest that the upward bias in the R^2 for those states could be as high as 0.2 and 0.3 respectively. Figures such as these would suggest that the true relationship between state unemployment movements and national forces is more even across the states than Table 9 suggests.

The thrust of our argument is that the inclusion of the national unemployment rate in the regression model leads to severe statistical problems and that it would be preferable to adopt an estimation strategy that utilises available information about the direct interactions between the states rather than their relationship with the constructed national rate. One possibility would be to use instrumental variables estimation, using the state unemployment rates as instruments for the national rate, with the state on the left-hand side of the regression excluded from the instrument set. While this would help to improve the accuracy of the parameter estimates and the associated R^2, there would remain a potential small-sample bias problem and in the present context there is no compelling reason for choosing the instrumental variables procedure in preference to the obvious alternative, which is simply to estimate the inter-state relationships directly, in the context of a series of single-equation multivariate regression models or perhaps a vector autoregressive framework.

Table 10 reports the results of a series of single-equation regressions relating unemployment changes in each state to unemployment changes in all of the others. The rows of Table 10 show the parameter estimates and associated *t* statistics (in parenthesis) together with the overall R^2 for each regression. The final column shows the LM test for first-order serial correlation in the residuals. The constants were negligible and insignificant in each regression and are not reported. The high degree of correlation across the states reported earlier suggests that some of the *t* values on the regression parameters may be unreliable, but the presence of multicollinearity shouldn't affect the overall R^2 values of the regressions, which is what we are interested in here[12].

[TABLE 10 NEAR HERE]

The results for the inter-state regressions provide strong confirmation that the previously reported R^2 values from the state-national regressions (Table 9) are biased upwards for the larger states, over-emphasising the importance of national as opposed

[12] The muticollinearity problem would also be present in a VAR model. One way to overcome the problem is to estimate the model using principal component-common factor methods. As a check on our results we estimated a series of models in which each state was in turn regressed on the principal components extracted from the data set. This method avoids the muticollinearity problem since the principal components are by construction orthogonal to each other (see for example Jackson, 1991). The principal component regressions confirmed the results of the standard multivariate regressions discussed in the text, suggesting that, with the exception of Tasmania, national as opposed to regional forces explain somewhere between 40% and 50% of the changes in state unemployment.

to region-specific shocks. It is interesting to note that the differences between the R^2 values from the state-national and state-state regressions for NSW, VIC and QLD are almost exactly equal to the mean (spurious) R^2 values suggested for those states by table 3. For example, the difference between the state-national and state-state R^2 values for NSW is 0.2, which is approximately equal to the spurious R^2 suggested by interpolation for a state with a weighting of 0.34, based on the results reported in table 3. In contrast, the weights attached to SA, WA and TAS in the national unemployment rate are relatively small and, as expected, the R^2 values from the state-national regressions are similar to those reported in Table 10. Overall, the R^2 results from the state-state regressions suggest that, if there is a national force affecting state unemployment movements, it is associated with not much more than 50% of the unemployment variation in the larger states and somewhere in the region of 40% in the smaller states. The exception is Tasmania, which does not appear to be closely related to the behaviour of the other states.

4. Summary and Conclusion

In this paper we have argued that there are significant statistical problems connected with the estimation of models which include regressors that are constructed as the weighted average or sum of a set of variables that includes the chosen dependent variable. Our theoretical analysis shows that the parameter estimates of such models are likely to be affected by a deterministic bias and a stochastic bias, both of which arise from the procedure used to construct the regressor. In view of their origin, we have referred to these two forms of bias collectively as "construction bias". The stochastic component of the construction bias is equivalent to a form of simultaneous equation bias, but arises form a different source, while the deterministic component is independent of any stochastic simultaneity problem. Furthermore, our analysis suggests that the identified bias problem is of significance in the interpretation of cointegration tests as well as models of stationary series.

To illustrate the nature of the statistical problems, we framed our discussion in terms of an analysis of whether movements in regional unemployment rates reflect the impact of national or region-specific shocks. Previous studies have examined this problem by considering how the various regional rates move in relation to the national unemployment rate, where the national rate is constructed as a weighted average of the regional rates. Our theoretical analysis and simulation results suggest that the parameter estimates of such models are prone to construction bias and that the results may be misleading in that they are likely to over-estimate the impact of national as opposed to region-specific forces. As a practical example of the potential importance of the construction bias problem, we examined the behaviour of Australian national and state unemployment. The empirical results confirm that correlation and regression models relating state and national unemployment rates are indeed likely to over-state the importance of national as opposed to region-specific forces. We have argued that the issue is best considered via a direct examination of the inter-state relationships, rather than indirectly via the national unemployment rate. Based on this approach, our results suggest that the degree of integration between the Australian states is considerably less than the state-national regressions would imply and that the potential scope for regional unemployment policy is consequently greater than the state-national unemployment analysis would suggest.

Finally, we would emphasise again that the bias problem we have identified is quite general and is likely to be present in any model that incorporates a regressor which includes all or part of the dependent variable by construction. In such cases, the regression and correlation parameters are likely to be biased and, to avoid any misleading conclusions, we would suggest that the model should where possible be re-formulated so as to remove the influence of the dependent variable from the regressor set.

References

Bartlett, M. S. (1946). On the Theoretical Specification and Sampling Properties of Autocorrerlated Time-Series, *Journal of the Royal Statistical Society*, Series B, **8**, 27-41

Brechling, F. (1967). Trends and Cycles in British Regional Unemployment, *Oxford Economic Papers*, **19**, 1-21

Byers, J. (1990). The Cyclical Sensitivity of Unemployment: An Assessment, *Regional Studies*, **24**, 447-53

Chapman, P. (1991). The Dynamics of Regional Unemployment in the UK, 1974-89, *Applied Economics*, **23**, 1059-64

Cochrane, J. H. (1991). A Critique of the Application of Unit Root Tests, *Journal of Economic Dynamics and Control*, **15**, 275-284

Crosby, M. and Olekalns, N. (1998). Inflation, Unemployment and NAIRU in Australia, *Australian Economic Review*, **31**, 117-29

Davidson, R. and Mackinnon, J. (1993). *Estimation and Inference in Econometrics*, Oxford University Press, Oxford

Debelle, G. and Vickery, J. (1998). Labour Market Adjustment: Evidence on Interstate Labour Mobility, Research Discussion Paper 9801, Reserve Bank of Australia, Sydney

Debelle, G. and Vickery, J. (1999). Labour Market Adjustment: Evidence on Interstate Labour Mobility, *Australian Economic Review*, **32**, 249-63

Dickey, D. and Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, **74**, 427-31

Dixon, R. and Shepherd, D. (2000). Trends and Cycles in Australian State and Territory Unemployment Rates, Research Paper 730, Department of Economics, University of Melbourne

Engle, R. and Granger, C. (1987). Cointegration and Error Correction: Representation, Estimation and Testing, *Econometrica*, **55**, 251-76

Granger, C. and Newbold, P. (1974). Spurious Regressions in Econometrics, *Journal of Econometrics*, **2**, 111-20

Groenewold, N. (1991). Regional Unemployment Disparities and Cyclical Sensitivities: Some Australian Results, *Australian Journal of Regional Studies*, **6**, 15-28

Groenewold, N. and Hagger, A. (1995). Regional Unemployment Dynamics: The Big Neighbour Effect, *Australasian Journal of Regional Studies*, **1**, 197-214

Groenewold, N. and Hagger, A. (1998). The Natural Unemployment Rate in Australia since the Seventies, *Economic Record*, **74**, 24-35

Gruen, D., Pagan, A. and Thompson, C. (1999). The Phillips Curve in Australia, Research Discussion Paper 1999-01, Reserve Bank of Australia, Sydney

Isard, W. (1960). *Methods of Regional Analysis*, MIT Press, Cambridge, Mass

Jackson, J. (1991). *A User's Guide to Principal Components*, John Wiley, New York

Johansen, S. (1995). *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford

Johnston, R. (1979). On the Relationships Between Regional and National Unemployment Trends, *Regional Studies*, **13**, 453-64

Martin, R. (1997). Regional Unemployment Disparities and their Dynamics, *Regional Studies*, **31**, 237-52

Mitchell, W. (1993). Testing for Unit Roots and Persistence in OECD Employment Rates, *Applied Economics*, **25**, 1489-1501

Phillips, P. (1986). Understanding Spurious Regressions in Econometrics, *Journal of Econometrics*, **33**, 311-40

Thirlwall, A. (1966). Regional Unemployment as a Cyclical Phenomenon, *Scottish Journal of Political Economy*, **13**, 205-19

Table 1 Empirical Correlation Coefficients: Mean Realisations

Weight (α)	**r N = ∞**	**$\hat{r}$ N = 50**	**$\hat{r}$ N = 84**	**$\hat{r}$ N = 100**	**$\hat{r}$ N = 1000**
0.0	0.0000	0.1153	0.0880	0.0802	0.0251
0.1	0.1104	0.1475	0.1287	0.1245	0.1099
0.2	0.2450	0.2462	0.2423	0.2414	0.2419
0.3	0.3939	0.3913	0.3920	0.3917	0.3933
0.4	0.5547	0.5514	0.5526	0.5522	0.5542
0.5	0.7071	0.7104	0.7052	0.7049	0.7067
0.6	0.8321	0.8299	0.8307	0.8305	0.8318
0.7	0.9191	0.9179	0.9184	0.9183	0.9190
0.8	0.9701	0.9696	0.9699	0.9698	0.9701
0.9	0.9939	0.9938	0.9938	0.9938	0.9939

Table 2 Correlation Coefficient: Empirical Critical Values

Weight	**Critical Values**		**Critical Values**		**Critical Values**		**Critical Values**	
	N=50		**N=84**		**N=100**		**N=1000**	
α	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**
0.0	0.3652	0.2854	0.2856	0.2120	0.2596	0.1956	0.0824	0.0628
0.1	0.4297	0.3448	0.3574	0.2905	0.3334	0.2728	0.1818	0.1613
0.2	0.5375	0.4560	0.4699	0.4085	0.4533	0.3933	0.3095	0.2906
0.3	0.6478	0.5802	0.5887	0.5401	0.5754	0.5260	0.4544	0.4370
0.4	0.7535	0.7028	0.7069	0.6734	0.6992	0.6601	0.6031	0.5902
0.5	0.8451	0.8101	0.8156	0.7907	0.8067	0.7818	0.7415	0.7326
0.6	0.9136	0.8943	0.8974	0.8825	0.8925	0.8772	0.8534	0.8474
0.7	0.9596	0.9499	0.9518	0.9440	0.9492	0.9419	0.9300	0.9268
0.8	0.9852	0.9818	0.9823	0.9794	0.9813	0.9788	0.9742	0.9731
0.9	0.9970	0.9963	0.9964	0.9958	0.9962	0.9957	0.9947	0.9945

Table 3 Parameter Estimates and *t* Statistics for Unrelated Random Processes

α	**Mean $\hat{\beta}$**	**Mean *t* statistic**	**Proportion $t > t_{0.05}$**	**Mean R^2**	**R^2 5% C. V.**
0.0	0.0011	0.7981	0.0522	0.01	0.04
0.1	0.1230	1.2497	0.1984	0.02	0.07
0.2	0.2949	2.5110	0.6964	0.07	0.15
0.3	0.5174	4.2919	0.9870	0.16	0.27
0.4	0.7687	6.6710	1.0000	0.31	0.44
0.5	0.9989	10.0020	1.0000	0.50	0.61
0.6	1.1527	14.9982	1.0000	0.69	0.77
0.7	1.2061	23.3253	1.0000	0.84	0.89
0.8	1.1761	39.9795	1.0000	0.94	0.96
0.9	1.0974	89.9421	1.0000	0.99	0.99

Table 4 Mean Experimental Realisations: True and Estimated parameters

α	θ	β	$\hat{\beta}$	$\hat{\beta}-\beta$	$\hat{\beta}-\theta$	**R^2**
0.0	0.7	0.7000	0.7007	0.007	0.007	0.33
0.1	0.7	0.7216	0.8208	0.0992	0.1208	0.43
0.2	0.7	0.7447	0.9302	0.1855	0.2302	0.53
0.3	0.7	0.7692	1.0213	0.2521	0.3213	0.64
0.4	0.7	0.7955	1.0877	0.2922	0.3877	0.74
0.5	0.7	0.8235	1.1260	0.3025	0.4260	0.83
0.6	0.7	0.8537	1.1371	0.2834	0.4371	0.89
0.7	0.7	0.8861	1.1245	0.2384	0.4245	0.94
0.8	0.7	0.9211	1.0938	0.1727	0.3938	0.98
0.9	0.7	0.9589	1.0506	0.0917	0.3506	0.99

Table 5. Cointegration Tests for Independently Generated Random Walks

1. $Var(T_1)/Var(T_2)=1/1,\ e_{1t}=e_{2t}=0$

α	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
$\bar{t}$	-2.04	-2.05	-2.05	-2.05	-2.06	-2.05	-2.05	-2.04	-2.04	-2.04
$t < t_{0.05}$	0.054	0.054	0.055	0.055	0.056	0.053	0.051	0.051	0.053	0.052

2. $Var(T_1)/Var(T_2)=4/1,\ e_{1t}=e_{2t}=0$

α	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
$\bar{t}$	-2.04	-2.05	-2.05	-2.05	-2.06	-2.05	-2.04	-2.04	-2.04	-2.04
$t < t_{0.05}$	0.053	0.056	0.055	0.0056	0.055	0.055	0.054	0.054	0.053	0.055

3. $Var(T_1)/Var(T_2)=1/1,\ Var(e_1)/Var(e_2)=1/1$

α	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
$\bar{t}$	-2.47	-2.46	-2.46	-2.46	-2.45	-2.46	-2.46	-2.46	-2.47	-2.47
$t < t_{0.05}$	0.158	0.157	0.157	0.158	0.155	0.158	0.161	0.158	0.172	0.172

4. $Var(T_1)/Var(T_2)=4/1,\ Var(e_1)/Var(e_2)=1/1$

α	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
$\bar{t}$	-2.21	-2.23	-2.25	-2.29	-2.34	-2.37	-2.40	-2.41	-2.42	-2.43
$t < t_{0.05}$	0.093	0.097	0.103	0.117	0.129	0.139	0.151	0.154	0.155	0.153

5. $Var(T_1)/Var(T_2)=1/1,\ \ Var(e_1)/Var(e_2)=1/4$

α	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
$\bar{t}$	-2.61	-2.62	-2.64	-2.69	-2.77	-2.89	-3.02	-3.12	-3.19	-3.32
$t < t_{0.05}$	0.213	0.216	0.219	0.235	0.271	0.304	0.351	0.389	0.419	0.432

6. $Var(T_1)/Var(T_2)=4/1,\ Var(e_1)/Var(e_2)=1/4$

α	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*
$\bar{t}$	-2.35	-2.31	-2.33	-2.46	-2.64	-2.81	-2.95	-3.05	-3.11	-3.16
$t < t_{0.05}$	0.144	0.128	0.138	0.181	0.230	0.297	0.344	0.374	0.400	0.416

Note: Residual Cointegration Test 5% Critical Value = -3.34

Table 6 Correlation Matrix: State Unemployment Rates

	AUS	NSW	VIC	QLD	SA	WA
AUS	1.00					
NSW	0.95	1.00				
VIC	0.92	0.79	1.00			
QLD	0.91	0.90	0.75	1.00		
SA	0.94	0.87	0.92	0.82	1.00	
WA	0.87	0.83	0.73	0.76	0.76	1.00
TAS	0.84	0.79	0.80	0.78	0.83	0.61

Table 7 Residual Cointegration Tests for Unemployment Rates

	AUS	NSW	VIC	QLD	SA	WA
AUS	*					
NSW	-2.06	*				
VIC	-1.64	-2.71	*			
QLD	-2.34	-2.54	-2.37	*		
SA	-3.37	-2.92	-3.80	-2.59	*	
WA	-2.02	-1.41	-1.35	-1.35	-3.35	*
TAS	-2.88	-2.61	-2.65	-2.62	-3.20	-2.01

Residual Unit Root Test 1% and 5% Critical Values = -3.90 and –3.34

Table 8 Correlation Matrix: First Differences of Unemployment Rates

	ΔAUS	ΔNSW	ΔVIC	ΔQLD	ΔSA	ΔWA
ΔAUS	1.00					
ΔNSW	0.85	1.00				
ΔVIC	0.81	0.63	1.00			
ΔQLD	0.72	0.63	0.60	1.00		
ΔSA	0.66	0.56	0.59	0.54	1.00	
ΔWA	0.64	0.56	0.57	0.48	0.36	1.00
ΔTAS	0.30	0.30	0.33	0.27	0.19	0.25

Table 9 State-National Unemployment Regressions

Regional ΔU	**Constant**	**National ΔU**	**R^2**	**LM(1)**
	(t value)	**(t value)**		
NSW	-0.010	1.08	0.73	0.00
	(-0.41)	(14.76)		
VIC	0.004	1.03	0.66	0.89
	(0.14)	(12.48)		
QLD	-0.005	0.89	0.52	1.02
	(-0.16)	(9.42)		
SA	0.016	0.77	0.44	0.45
	(0.45)	(8.00)		
WA	-0.004	0.74	0.41	0.10
	(-0.11)	(7.57)		
TAS	0.425	0.48	0.09	1.77
	(0.69)	(2.86)		

Table 10 State-State Unemployment Rate Regressions

State	**NSW**	**VIC**	**QLD**	**SA**	**WA**	**TAS**	**R^2**	**LM(1)**
NSW	*	0.212	0.206	0.257	0.277	0.037	0.53	0.14
		(1.84)	(2.01)	(2.38)	(2.38)	(0.56)		
VIC	0.198	*	0.188	0.314	0.274	0.080	0.57	1.91
	(1.84)		(1.89)	(3.08)	(2.67)	(1.25)		
QLD	0.242	0.236	*	0.157	0.107	0.076	0.43	0.12
	(2.01)	(1.89)		(1.31)	(0.90)	(1.06)		
SA	0.267	0.349	0.139	*	-0.080	-0.031	0.42	0.55
	(2.38)	(3.08)	(1.31)		(-0.71)	(-0.45)		
WA	0.293	0.310	0.097	-0.081	*	0.021	0.40	4.27
	(2.61)	(2.67)	(0.90)	(-0.71)		(0.30)		
TAS	0.110	0.250	0.191	-0.086	0.058	*	0.13	1.25
	(0.56)	(1.25)	(1.06)	(-0.45)	(0.30)			