Working Paper 99-78
Statistics and Econometrics Series 30
October 1999

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 624-9849

NONSENSE REGRESSIONS DUE TO TIME-VARYING MEANS.

Uwe Hassler.*

Abstract

Regressions of two independent time series are considered. The variables are covariance stationary but display time-varying although not trending means. Two prominent examples are mean shifts due to structural breaks and seasonally varying means. If the variation of the means is not taken into account, this induces nonsense correlation. The asymptotic treatment is supplemented by experimental evidence.

Keywords: Structural breaks; deterministic seasonality; spurious correlation.

*Free University of Berlin. Institute of Statistics and Econometrics. Boltzmannstr. 20 D-14195 Berlín Germany, e-mail: uwe@wiwiss.fu-berlin.de. This work was carried out while visiting Universidad Carlos III de Madrid. Financial support from the European Commision through the Training and Mobility of Researchers programme is gratefully acknowledged.

1 Introduction

Since the early paper by Yule (1926) statisticians are aware of the danger of nonsense correlation between unrelated random walks. Granger and Newbold (1974) followed his work and established experimentally that stochastically independent random walks give rise to spurious regressions.[1] They coined the latter term to describe the fact that testing for the true null of no correlation one observes a much higher rejection rate than the nominal level. Phillips (1986) provided an asymptotic treatment of spurious regressions that arise whenever random walks are not cointegrated, see also the discussion e.g. in Banerjee, Dolado, Galbraith and Hendry (1993). His results were extended to the cases of cointegrated regressors and of stationary covariates by Choi (1994) and Hassler (1996), respectively. Similar findings were established in the presence of trending variables integrated of order two or higher orders, see Haldrup (1994) and Marmol (1996), respectively, or in the presence of linear time trends, cf. Hassler (1996a).

This note demonstrates that nonsense correlation, or spurious regressions, may arise even if the series are not trending. It is motivated by Perron (1990) who showed that the distinction between trending series integrated of order one and stationary series with a mean shift may be difficult as long as the shift is not taken into account. His finding suggests that independent covariance stationary time series with mean shifts may give rise to nonsense regression. Indeed, this will be proven here. Hence this paper contributes to a field of growing interest recently surveyed in Maddala and Kim (1998).

[1] I use the terms nonsense or spurious regressions interchangeably and speak as well of nonsense or spurious correlation.

In fact, I consider the more general case that the means of two independent series are time-varying although not trending. This includes e.g. series with deterministic seasonality where the mean varies from season to season.

Section 2 becomes precise on the model and establishes the general result which is illustrated by the example of quarterly varying means. Section 3 turns to the case of mean shifts due to structural breaks that is more relevant in practice. The asymptotic formulae are confronted with Monte Carlo evidence. The final section contains a more detailled summary.

2 The general case

The simple model considered here consists of two covariance stationary series u_{1t} and u_{2t} that are stochastically independent of each other and that are added to a deterministic mean function,

$$x_{it} = d_{it} + u_{it},\ t = 1, 2, \ldots, T,\ i = 1, 2. \tag{1}$$

The stochastic components are assumed to be zero mean processes with variance σ_i^2 independent of each other. The deterministic components are supposed to be not trending. They are bounded and square summable, more precisely,

$$|d_{it}| \leq D_i < \infty,\ t = 1, 2, \ldots, T,\ i = 1, 2, \tag{2}$$

$$(\overline{d_i},\ \overline{d_i^2},\ \overline{d_1 d_2}) = \frac{1}{T}\sum_{t=1}^{T}(d_{it},\ d_{it}^2,\ d_{1t}d_{2t}) \xrightarrow{p} (\delta_i,\ \gamma_i^2,\ \gamma_{12}), \tag{3}$$

where $\xrightarrow{p}$ denotes convergence in probability. For the stochastic series I assume in contrast

$$(\overline{u_i},\ \overline{u_i^2},\ \overline{u_1 u_2}) = \frac{1}{T}\sum_{t=1}^{T}(u_{it},\ u_{it}^2,\ u_{1t}u_{2t}) \xrightarrow{p} (0,\ \sigma_i^2,\ 0),$$

and that they satisfy the central limit theorem, which implies

$$\sum_{t=1}^{T} u_{it} = O_p(T^{0.5}).$$

Given (1), (2) and (3) it is straightforward to prove as $T \to \infty$

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}(x_{it} - \overline{x_i})^2 &= \overline{u_i^2} + \overline{d_i^2} - \overline{d_i}^2 + O_p(T^{-0.5}) \\ &\xrightarrow{p} \sigma_i^2 + \gamma_i^2 - \delta_i^2,\ i = 1, 2, \end{aligned} \tag{4}$$

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}(x_{1t} - \overline{x_1})(x_{2t} - \overline{x_2}) &= \overline{u_1 u_2} + \overline{d_1 d_2} - \overline{d_1}\,\overline{d_2} + O_p(T^{-0.5}) \\ &\xrightarrow{p} \gamma_{12} - \delta_1\delta_2. \end{aligned} \tag{5}$$

Please note that (2) and (3) allow for the special case of constant means, $d_{it} = d_i$. In this case $\delta_i = d_i$, $\gamma_i^2 = d_i^2 = \delta_i^2$ and $\gamma_{12} = d_1 d_2 = \delta_1\delta_2$ so that (4) and (5) reproduce the standard case of independent series with constant means.

The limits from (4) and (5) render themselves to determine the probability limit of the ordinary least squares (OLS) estimator from

$$x_{1t} = \hat{\alpha} + \hat{\beta}x_{2t} + \hat{\epsilon}_t,\ t = 1, 2, \ldots, T, \tag{6}$$

where the variable nature of d_{it} is not taken into account. We obtain under the assumptions made so far

$$\hat{\beta} \xrightarrow{p} \frac{\gamma_{12} - \delta_1\delta_2}{\sigma_2^2 + \gamma_2^2 - \delta_2^2} =: \breve{\beta}. \tag{7}$$

With (7) the behaviour of the standard error of regression (6),

$$\hat{s}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_t^2,\ \hat{\epsilon}_t = x_{1t} - \overline{x_1} - \hat{\beta}(x_{2t} - \overline{x_2}),$$

becomes obvious. It is needed to construct the t-statistic t_β testing for $\beta = 0$. To establish the divergence of t_β it is convenient to normalize as

$$T^{-0.5}t_\beta = \frac{\hat{\beta}\sqrt{T^{-1}\sum(x_{2t} - \overline{x_2})^2}}{\hat{s}}.$$

Moreover, the limit of the coefficient of determination can be derived from

$$R^2 = 1 - \frac{\hat{s}^2}{T^{-1}\sum(x_1 - \overline{x_1})^2}.$$

With (4), (5) and (7) the following results are easily established.

Proposition 1 (Nonsense regression) *Let x_{1t} and x_{2t} from (1) be stochastically independent of each other. Under the assumptions of this section it then holds for the OLS statistics from (6) as $T \to \infty$:*

$$\hat{\beta} \xrightarrow{p} \frac{\gamma_{12} - \delta_1\delta_2}{\sigma_2^2 + \gamma_2^2 - \delta_2^2} =: \check{\beta},$$

$$\hat{s}^2 \xrightarrow{p} \sigma_1^2 + \gamma_1^2 - \delta_1^2 - \check{\beta}(\gamma_{12} - \delta_1\delta_2) =: \check{s}^2,$$

$$T^{-0.5}t_\beta \xrightarrow{p} \frac{\check{\beta}\sqrt{\sigma_2^2 + \gamma_2^2 - \delta_2^2}}{\check{s}},$$

$$R^2 \xrightarrow{p} 1 - \frac{\check{s}^2}{\sigma_1^2 + \gamma_1^2 - \delta_1^2}.$$

Just as in the case of random walks, see Phillips (1986), we observe that in general $\hat{\beta}$, $T^{-0.5}t_\beta$ and R^2 have non-zero limits, only that the limits are deterministic in the present setup of deterministic time-varying means. Such nonsense correlation of course only occurs if the righthand side in (5) and hence $\check{\beta}$ is different from zero. Consider e.g. the case of additive outliers where d_{it} is zero except for a finite number of times. This clearly entails $\gamma_{12} = \delta_i = 0$, and therefore does not lead to spurious regression results.

As a first application of Propostion 1 I consider series with seasonally varying means. Abeysinghe (1991) observed spurious regression effects if seasonally integrated series are regressed on each other, where the regression includes seasonal dummies. Here the series are not integrated but display deterministic seasonality that is ignored in the regression. I restrict the treatment to quarterly varying means with observations over n years,

$$d_{it} = \begin{cases} s_{1i}, & t = 4j+1 \\ s_{2i}, & t = 4j+2 \\ s_{3i}, & t = 4j+3 \\ s_{4i}, & t = 4j+4 \end{cases}, \; j = 0, 1, \ldots, n-1, \tag{8}$$

where $T = 4n$ is assumed. It is again simple to check that

$$(\overline{d_i}, \overline{d_i^2}, \overline{d_1 d_2}) \xrightarrow{p} \frac{1}{4}\left(\sum_{k=1}^{4} s_{ki}, \sum_{k=1}^{4} s_{ki}^2, \sum_{k=1}^{4} s_{k1}s_{k2}\right).$$

Hence Proposition 1 can be applied to the situation of quarterly series with deterministic seasonality.

Corollary 2 (Deterministic seasonality) *Let x_{1t} and x_{2t} from (1) and (8) satisfy the assumptions of this section. Then the results from Proposition*

1 hold with

$$(\delta_i,\, \gamma_i^2,\, \gamma_{12}) = \frac{1}{4}\left(\sum_{k=1}^{4} s_{ki},\, \sum_{k=1}^{4} s_{ki}^2,\, \sum_{k=1}^{4} s_{k1}s_{k2}\right),\ i = 1, 2.$$

Please note that quarterly varying means like in (8) do not necessarily imply spurious regressions. For instance

$$\begin{aligned}(s_{11},\, s_{21},\, s_{31},\, s_{41}) &= (s_{11},\, 0,\, 0,\, -s_{11}),\\ (s_{12},\, s_{22},\, s_{32},\, s_{42}) &= (0,\, s_{22},\, s_{32},\, 0)\end{aligned}$$

yield $\gamma_{12} = 0$, $\delta_1 = 0$ and hence do not give rise to nonsense correlation, $\check{\beta} = 0$. Moreover, regressing seasonal series on each other empirical researchers typically include seasonal dummies accounting for seasonally varying means. That is why the nonsense regression situation underlying Corollary 2 will not be encountered often in practice. Therefore, I provide a more thorough discussion only for the more relevant example of mean shifts due to structural breaks.

3 Mean shifts

As a second example of model (1) I consider mean shifts,

$$d_{it} = \begin{cases} 0, & t \leq \lambda_i T \\ b_i, & t > \lambda_i T \end{cases}, \tag{9}$$

where $E(x_{it}) = 0$ is assumed without loss of generality for $t \leq \lambda_i T$. Both observed series x_{1t} and x_{2t} are subject to a break in the mean at possibly different times $\lambda_i T$ where $\lambda_i \in [0, 1]$ is the proportion when this shift occurs.

It is easily verified that

$$(\overline{d_i},\ \overline{d_i^2},\ \overline{d_1 d_2}) \xrightarrow{p} (b_i(1-\lambda_i),\ b_i^2(1-\lambda_i),\ b_1 b_2(1-\lambda_{max})),$$

where $\lambda_{max} = \max(\lambda_1, \lambda_2)$. Propositon 1 hence provides the following corollary.

Corollary 3 (Mean shifts) *Let x_{1t} and x_{2t} from (1) and (9) satisfy the assumptions of the previous section. Then the results from Proposition 1 hold with*

$$(\delta_i,\ \gamma_i^2,\ \gamma_{12}) = (b_i(1-\lambda_i),\ b_i^2(1-\lambda_i),\ b_1 b_2(1-\lambda_{max})),\ \ i = 1, 2.$$

It is not difficult to supplement Corollary 3 with a statement referring to the Durbin-Watson statistic. To simplify matters I assume now that u_{1t} and u_{2t} are white noise series. With $\check{\beta}$ and $\check{s}$ defined in Proposition 1 it then follows

$$dw = \frac{T^{-1}\sum_{t=2}^{T} \Delta\hat{\epsilon}_t}{\hat{s}^2} \xrightarrow{p} \frac{2(\sigma_1^2 + \check{\beta}^2\sigma_2^2)}{\check{s}^2}.$$

We observe that dw does not tend to zero as with spurious regressions of random walks. However, its probability limit will in general differ from 2 even if the stochastic components are white noise.

Please note that nonsense correlation according to Corollary 3 arises if and only if both series are subject to a structural break, $b_i \neq 0$ and $0 < \lambda_i < 1$ for $i = 1, 2$. This is true because for $\lambda_{max} = \max(\lambda_1, \lambda_2)$

$$(1-\lambda_1)(1-\lambda_2) \neq (1-\lambda_{max})$$

implies $\check{\beta} \neq 0$ with

$$\check{\beta} = \frac{b_1 b_2(1-\lambda_{max}) - b_1 b_2(1-\lambda_1)(1-\lambda_2)}{\sigma_2^2 + b_2^2(1-\lambda_2)\lambda_2}.$$

Furthermore, certain symmetries appear. Looking closely at the limits in Proposition 1, Corollary 3 reveals the following symmetries about the line $\lambda = 0.5$.

Corollary 4 (Symmetries) *Let x_{1t} and x_{2t} from (1) and (9) satisfy the assumptions of Corollary 3 with $b_i \neq 0$ and $0 < \lambda_i < 1$ for $i = 1, 2$. Then it holds for the limits in Proposition 1:*

- *if $\lambda_1 = 1 - \lambda_2$ then all limits are symmetric about $\lambda_2 = 0.5$;*
- *if $\lambda_1 = 0.5$ then all limits are symmetric about $\lambda_2 = 0.5$;*
- *if $\lambda_2 = 0.5$ then all limits are symmetric about $\lambda_1 = 0.5$;*
- *if $\lambda_1 = \lambda_2 = \bar{\lambda}$ then all limits are symmetric about $\bar{\lambda} = 0.5$.*

Bearing in mind the German unification in 1990 the last case in Corollary 4 of a common breakpoint $\bar{\lambda}$ may be of particular interest. It is straightforward to show that $|\check{\beta}|$ has a maximum at $\bar{\lambda} = 0.5$, i.e. the spurious effect is strongest if the breakpoint is in the middle of the sample.

To verify the relevance of the asymptotic results a Monte Carlo experiment was performed. Two independent white noise series with means shift were simulated according to (1) and (9) with $u_{it} \sim N(0, 1)$. The number of observations was $T = 100$. From 5000 replications done with GAUSS32 the mean values of the OLS statistics were computed. In Tables 1 and 2 they are compared with the probability limits according to Corollary 3. Moreover, I report the frequency of rejection when testing with t_β for the true null hypothesis $\beta = 0$ at the 5% level, $|t_\beta| > 1.96$. As one would expect the nonsense correlation becomes more severe as the breaks $b_i > 0$ are increasing.

Throughout we observe that the asymptotic values well explain the experimental means already for $T = 100$. The asymptotic symmetry according to Corollary 4 was also found to be well reproduced experimentally. That's why I do not report tables for symmetric cases.

Table 1 is restricted to the situation of different breakpoints. With $\lambda_1 = 0.25$, $\lambda_2 = 0.75$ only moderate spurious results arise. As the breakpoints move closer, $\lambda_1 = 0.5$, $\lambda_2 = 0.25$ or $\lambda_1 = 0.25$, $\lambda_2 = 0.5$, the problem of nonsense regressions becomes more severe. In Table 2 results for common breakpoints are collected. If $\lambda_1 = \lambda_2$, the feature of nonsense correlation is stronger in comparison with Table 1, and it is strongest if the break occurs in the middle of the sample.

Very low values of the coefficient of determination are common to Tables 1 and 2. This suggests to detect spurious regressions due to mean shifts by means of the R^2. Alternatively one might of course avoid nonsense correlation from the beginning by testing the univariate series for a break and eventually removing it prior to regression.

4 Summary

It is widespread textbook knowledge that nonsense regressions may arise between independent time series that are trending. Moreover, there is a growing literature concerned with the effect of structural breaks. In this paper the two topics are related. It is shown that the danger of nonsense correlation, or spurious regression, is present even if the independent series are covariance stationary. This is true if their means are time-varying functions and

Table 1: Mean values and limits, $\lambda_1 \neq \lambda_2$

	$\hat{\beta}$	$t_\beta/T^{0.5}$	R^2	dw	5%
$\lambda_1 = 0.25,\ \lambda_2 = 0.75$ ($\lambda_1 = 0.75,\ \lambda_2 = 0.25$)					
$b_1 = 1$	0.050	0.050	0.013	1.709	8.06
$b_2 = 1$	(0.053)	(0.053)	(0.003)	(1.694)	
$b_1 = 1$	0.071	0.087	0.017	1.722	14.02
$b_2 = 2$	(0.071)	(0.087)	(0.008)	(1.706)	
$b_1 = 2$	0.107	0.089	0.017	1.194	13.02
$b_2 = 1$	(0.105)	(0.087)	(0.008)	(1.164)	
$b_1 = 2$	0.146	0.1.46	0.029	1.220	29.30
$b_2 = 2$	(0.143)	(0.144)	(0.020)	(1.190)	
$\lambda_1 = 0.5,\ \lambda_2 = 0.25$ ($\lambda_1 = 0.5,\ \lambda_2 = 0.75$)					
$b_1 = 1$	0.108	0.105	0.020	1.644	18.42
$b_2 = 1$	(0.105)	(0.103)	(0.011)	(1.635)	
$b_1 = 1$	0.144	0.173	0.037	1.698	39.28
$b_2 = 2$	(0.143)	(0.171)	(0.029)	(1.681)	
$b_1 = 2$	0.212	0.166	0.035	1.103	37.00
$b_2 = 1$	(0.211)	(0.164)	(0.026)	(1.073)	
$b_1 = 2$	0.287	0.279	0.079	1.195	82.22
$b_2 = 2$	(0.286)	(0.277)	(0.071)	(1.165)	
$\lambda_1 = 0.25,\ \lambda_2 = 0.5$ ($\lambda_1 = 0.75,\ \lambda_2 = 0.5$)					
$b_1 = 1$	0.103	0.106	0.021	1.738	17.84
$b_2 = 1$	(0.100)	(0.103)	(0.011)	(1.719)	
$b_1 = 1$	0.125	0.164	0.035	1.778	38.06
$b_2 = 2$	(0.125)	(0.164)	(0.026)	(1.757)	
$b_1 = 2$	0.202	0.173	0.038	1.256	41.90
$b_2 = 1$	(0.200)	(0.171)	(0.029)	(1.224)	
$b_1 = 2$	0.255	0.283	0.081	1.343	80.78
$b_2 = 2$	(0.250)	(0.277)	(0.071)	(1.308)	

For notes see Table 2.

Table 2: Mean values and limits, $\lambda_1 = \lambda_2 = \bar{\lambda}$

	$\hat{\beta}$	$t_\beta/T^{0.5}$	R^2	dw	5%
$\bar{\lambda} = 0.25$ ($\bar{\lambda} = 0.75$)					
$b_1 = 1$	0.159	0.162	0.034	1.785	36.22
$b_2 = 1$	(0.158)	(0.159)	(0.025)	(1.770)	
$b_1 = 1$	0.215	0.270	0.076	1.904	77.72
$b_2 = 2$	(0.214)	(0.269)	(0.067)	(1.889)	
$b_1 = 2$	0.319	0.272	0.077	1.374	76.62
$b_2 = 1$	(0.316)	(0.269)	(0.067)	(1.348)	
$b_1 = 2$	0.431	0.479	0.191	1.682	99.78
$b_2 = 2$	(0.429)	(0.474)	(0.184)	(1.657)	
$\bar{\lambda} = 0.5$					
$b_1 = 1$	0.202	0.206	0.049	1.750	54.76
$b_2 = 1$	(0.200)	(0.204)	(0.040)	(1.733)	
$b_1 = 1$	0.250	0.335	0.107	1.905	92.40
$b_2 = 2$	(0.250)	(0.333)	(0.100)	(1.888)	
$b_1 = 2$	0.405	0.337	0.109	1.310	92.68
$b_2 = 1$	(0.400)	(0.333)	(0.100)	(1.289)	
$b_1 = 2$	0.505	0.583	0.257	1.686	100
$b_2 = 2$	(0.500)	(0.577)	(0.250)	(1.666)	

Simulated white noise series with mean shifts as in (9). Given are mean values from 5000 replications of regression (6) with $T = 100$ observations (asymptotic values according to Corollary 3 in brackets). Moreover, the column denoted '5%' contains the frequency of rejecting the true null $\beta = 0$ when performing a two-sided test based on t_β at the nominal 5% level.

if this is not taken into account when regressing them on each other. The first example are series with deterministic seasonality where the means differ from season to season. Nonsense regressions are avoided by simply including seasonal dummies in the regression, which is of course standard practice.

The second example discussed at some length are mean shifts due to structural breaks. The following asymptotic results are well supported by finite sample evidence. The closer the breakpoints of the two series are, the more severe is the problem of nonsense correlation. The strongest spurious regression effects occur for a common breakpoint in the middle of the sample. Depending on the magnitude of the breaks the true null of no correlation will be rejected with very high probability. This clearly highlights the importance of testing for structural breaks and removing them before running regressions.

References

Abeysinghe, T. (1991): Inappropriate use of seasonal dummies in regressions; Economics Letters 36, 175-179.

Banerjee, A., J.J. Dolado, J.W. Galbraith, and D.F. Hendry (1993): Cointegration, error-correction, and the econometric analysis of non-stationary data; Oxford University Press.

Choi, I. (1994): Spurious regressions and residual-based tests for cointegration when regressors are cointegrated; Journal of Econometrics 60, 313-320.

Granger, C.W.J., and P. Newbold (1974): Spurious regressions in econometrics; Journal of Econometrics 2, 111-120.

Haldrup, N. (1994): The asymptotics of single-equation cointegration regressions with I(1) and I(2) variables; Journal of Econometrics 63, 153-181.

Hassler, U. (1996): Spurious regressions when stationary regressors are included; Economics Letters 50, 25-31.

Hassler, U. (1996a): Nonsense correlation between time series with linear trends; Allgemeines Statistisches Archiv 80, 227-235.

Maddala, G.S., and I.-M. Kim (1998): Unit roots, cointegration, and structural breaks; Cambridge University Press.

Marmol, F. (1996): Nonsense regressions between integrated processes of different orders; Oxford Bulletin of Economics and Statistics 58, 525-536.

Perron, P. (1990): Testing for a unit root in a time series with a changing mean; Journal of Business & Economic Statistics 8, 153-162.

Phillips, P.C.B. (1986): Understanding spurious regressions in econometrics; Journal of Econometrics 33, 311-340.

Yule, G.U. (1926): Why do we sometimes get nonsense correlation between time series? A study in sampling and the nature of time series; Journal of the Royal Statistical Society 89, 1-64.