Spurious multivariate regressions under fractionally integrated processes

Katarzyna Łasak[1]
Daniel Ventosa-Santaulària[2]
Ricardo Ramírez Vargas[2]
Eduardo Vera-Valdés[3]

Begin

(1) Tinbergen Institute (2) CIDE (3) University of Aalborg
WEDS Monterrey, México November 2018

This work...

Spurious regression under long memory was studied by Tsay and Chung (2000) for a univariate model. We extend their findings for the multivariate linear regression and find that inference drawn from the latter is also spurious. Our results hold for any finite number of independent fractionally integrated explanatory variables. We further extend Tsay and Chung's results by allowing for (some) explanatory variables to be correlated between them.

Spurious regression

Spurious regression in empirical econometrics is widely understood as the failure of conventional testing procedures when the series exhibit strong temporal properties. In economics, the levels of many macroeconomic time series are known to behave as nonstationary processes, which is fertile ground for spurious inference.

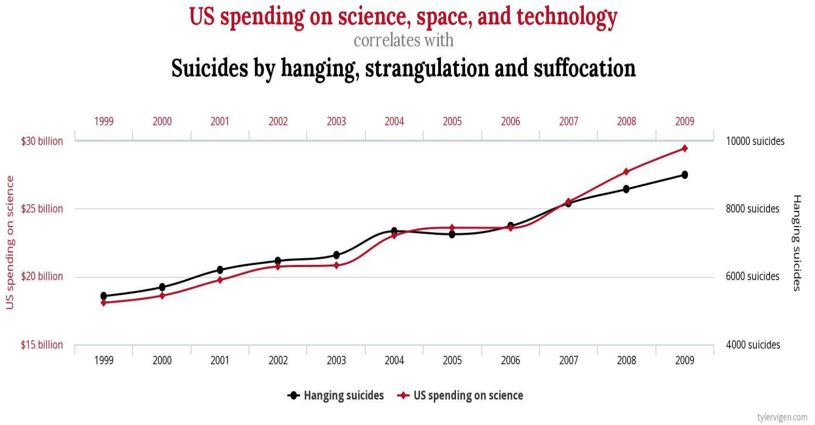


Figure: Source: http://www.tylervigen.com/spurious-correlations

Yule

Spurious regression

Several authors (McCallum, 2010, Kolev, 2011 and Agiakloglou, 2013, Agiakloglou & Agiropoulos, 2017) consider that non-resolved autocorrelation problems in the regression analysis lie at the origin of spurious regression. The aforementioned researchers supported their arguments with Monte Carlo evidence.

Spurious regression

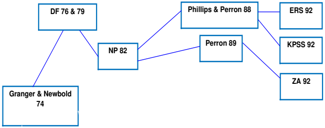
Nonetheless, it can be easily shown that the spurious regression phenomenon is more subtle (Sollis, 2011, Martínez et al, 2012, Ventosa-Santaulària et al, 2016, Tu, 2017, and Zhang, 2018). In particular, variables that exhibit long-range dependence can generate nonsense inference (see, for instance, Cappuccio and Lubian, 1997, Marmol. 1998, and Tsay and Chung, 2000). Moreover, long-range dependence can be easily confused with other time-dependence dynamics (Sibbertsen, 2014, Bertram et al, 2013, Haldrup and Kruse, 2014, and Jing and Zhang, 2017).

“Spuriosity” was mostly ignored in econometrics (i)

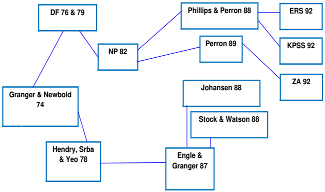
But was rediscovered in the seventies (ii)

Granger & Newbold
74

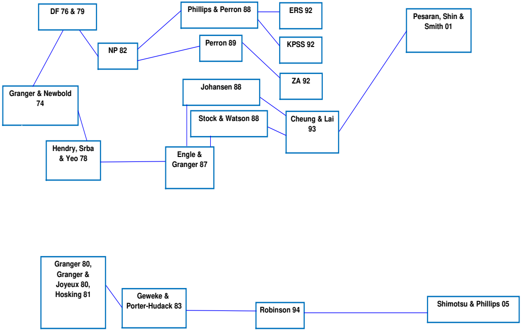
Tools to detect nonstationarity soon emerged (iii)



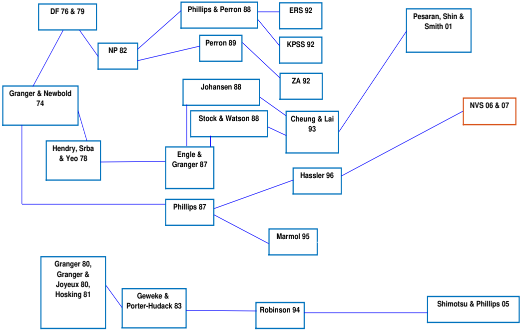
Tools to deal with nonstationarity too (iv)



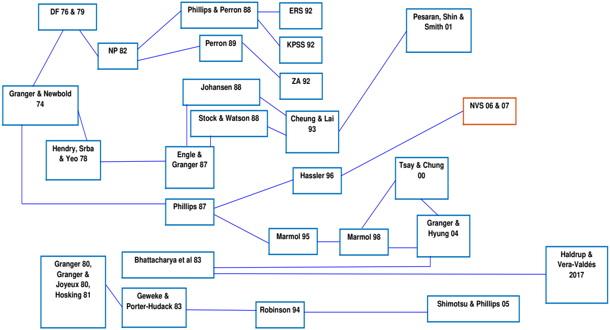
Long memory also became a relevant topic (v)



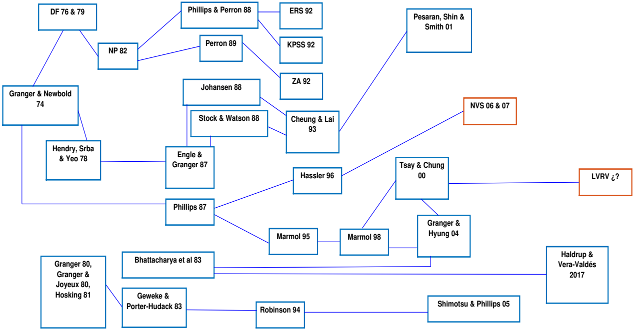
The “spuriosity” has been thoroughly studied (vi)



and was also uncovered under long memory (vii)



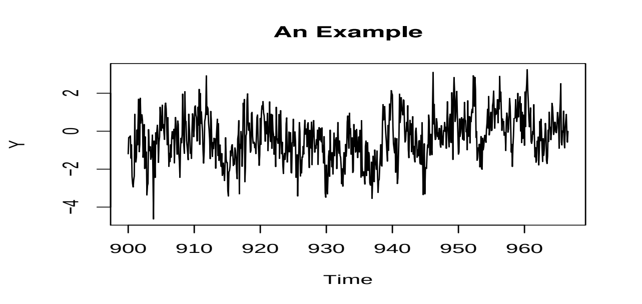
This paper uncovers a bit more (viii)



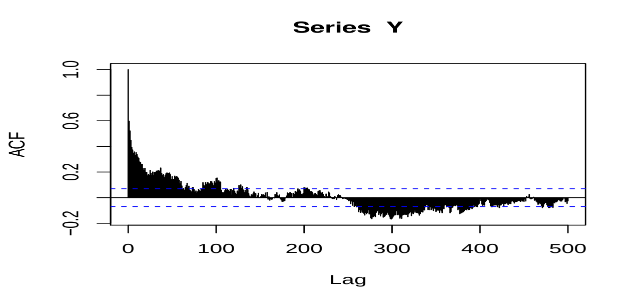
General definition of an $FI(d)$ process

- A discrete-time stochastic process z_t is said to be a fractionally process of order d if it satisfies $(1-L)^d z_t = a_{z,t}$.
- Here, L is the lag operator, d is the fractional differencing parameter, and $(1-L)^d$ is the fractional differencing operator, defined as $(1-L)^d = \sum_{j=0}^{\infty} \Psi_j L^j$, where $\Psi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}$ and $\Gamma(\cdot)$ is the gamma function.
- The innovations sequence $a_{z,t}$ is white noise with zero mean and finite variance $\sigma^2_{a_z}$.
- When (i) $d < 0$, z_t is anti-persistent, (ii) $0 < d < 1/2$, z_t is stationary, (ii) $1/2 < d < 1$, z_t is weakly nonstationary ($\sum_{j=-\infty}^{\infty} \gamma_z(j) = \infty$), $-1/2 < d < 1/2$, z_t is invertible.

This is a typical lm time series (i)



And this its ACF (ii)



Assumption 1: for the "spurious case" we assume that the regressand is independent from the regressors.

- Let the fundamental noises satisfy $E\left[\epsilon_{z,t}\right]^{q_z} < \infty$ with $q_z \geq \max\left\{4, -\frac{8d_z}{1+2d_z}\right\}$ for all $z = y, x_1, \cdots, x_k$.
- Let y_t be an independent stationary fractionally integrated process of order d_y, $y_t = (1-L)^{-d_y}\epsilon_{y,t}$,
- Nonetheless, we allow for a certain degree of linear correlation between some regressors. Let $x_{i,t}$ for $i = 1, 2, \ldots, k$ and $k < T$, be signal plus noise processes, $x_i = (1-L)^{-d_{x_i}}\epsilon_{i,t} + w_{i,t}$, where:
  - $\epsilon_{z,t}$ are i.i.d. white noise with zero mean and finite variance $\sigma^2_{\epsilon,z}$, and $d_z \in (0, \frac{1}{2})$ for $z = y, x_1, \cdots, x_k$.
  - $w_t = (w_{1,t}, w_{2,t}, \cdots, w_{k,t})'$ is a zero mean random vector with variance matrix Σ, $\forall t$, given by...

Assumption 1 [...]

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,m} & 0 & 0 & \cdots & 0 \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,m} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,m} & \sigma_{2,m} & \cdots & \sigma_m^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \sigma_{m+1}^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \sigma_{m+2}^2 & \ddots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_k^2 \end{bmatrix}; \quad (1)$$

that is, $\sigma_{i,j} \neq 0$ for $1 \leq i, j \leq m$, $i \neq j$, and $\sigma_{i,j} = 0$ otherwise. The only restriction imposed on w_t is that it has a moving average representation of the form $w_t = \sum_{i=0}^{\infty} \phi_i \nu_{t-i}$ where $\sum_{i=0}^{\infty} i|\phi_i|$ with ν_t a white noise process. Note that this restriction ensures that the noise process is less persistent than the signal. Moreover, it is satisfied by all stationary ARMA processes.

Theorem 1

Let Assumption 1 hold. Suppose that the linear specification $y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t$ is estimated by OLS. Then, as $T \to \infty$:

i) $\hat{\beta}_0 = O_p(T^{d_y - \frac{1}{2}})$;

ii) $\hat{\beta}_i =$
$$\begin{cases} O_p(T^{-\frac{1}{2}}) & \text{for } \bar{d}_{x_{1:m}} + d_y < \frac{1}{2}, \\ O_p\left[(T^{-1}\log T)^{\frac{1}{2}}\right] & \text{for } \bar{d}_{x_{1:m}} + d_y = \frac{1}{2}, \\ O_p(T^{\bar{d}_{x_{1:m}} + d_y - 1}) & \text{for } \frac{1}{2} < \bar{d}_{x_{1:m}} + d_y, \end{cases}$$
for $i = 1, \ldots, m$;

iii) $\hat{\beta}_i =$
$$\begin{cases} O_p(T^{-\frac{1}{2}}) & \text{for } d_{x_i} + d_y < \frac{1}{2}, \\ O_p\left[(T^{-1}\log T)^{\frac{1}{2}}\right] & \text{for } d_{x_i} + d_y = \frac{1}{2}, \\ O_p(T^{d_{x_i} + d_y - 1}) & \text{for } \frac{1}{2} < d_{x_i} + d_y, \end{cases}$$
for $i = m+1, \ldots, k$;

iv) $t_{\beta_0} = O_p(T^{d_y})$;

Theorem 1

Let Assumption 1 hold. Suppose that the linear specification $y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t$ is estimated by OLS. Then, as $T \to \infty$:
Furthermore,

i) $$t_{\beta_i} = \begin{cases} O_p(1) & \text{for } \bar{d}_{x_{1:m}} + d_y < \frac{1}{2}, \\ O_p\left[(\log T)^{\frac{1}{2}}\right] & \text{for } \bar{d}_{x_{1:m}} + d_y = \frac{1}{2}, \\ O_p(T^{\bar{d}_{x_{1:m}} + d_y - \frac{1}{2}}) & \text{for } \frac{1}{2} < \bar{d}_{x_{1:m}} + d_y, \end{cases}$$
for $i = 1, \ldots, m$;

ii) $$t_{\beta_i} = \begin{cases} O_p(1) & \text{for } d_{x_i} + d_y < \frac{1}{2}, \\ O_p\left[(\log T)^{\frac{1}{2}}\right] & \text{for } d_{x_i} + d_y = \frac{1}{2}, \\ O_p(T^{d_{x_i} + d_y - \frac{1}{2}}) & \text{for } \frac{1}{2} < d_{x_i} + d_y, \end{cases}$$
for $i = m + 1, \ldots, k$.

Theorem 1

Let Assumption 1 hold. Suppose that the linear specification $y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t$ is estimated by OLS. Then, as $T \to \infty$: Furthermore,

i) $R^2 = \begin{cases} O_p\left(T^{-1}\right) & \text{for } \bar{d}_{x_{1:k}} + d_y < \frac{1}{2}, \\ O_p\left(T^{-1}\log T\right) & \text{for } \bar{d}_{x_{1:k}} + d_y = \frac{1}{2}, \\ O_p\left[T^{2(\bar{d}_{x_{1:k}}+d_y-1)}\right] & \text{for } \frac{1}{2} < \bar{d}_{x_{1:k}} + d_y; \end{cases}$

ii) $\mathcal{F} = \begin{cases} O_p(1) & \text{for } \bar{d}_{x_{1:k}} + d_y < \frac{1}{2}, \\ O_p(\log T) & \text{for } \bar{d}_{x_{1:k}} + d_y = \frac{1}{2}, \\ O_p\left[T^{2(\bar{d}_{x_{1:k}}+d_y)-1}\right] & \text{for } \frac{1}{2} < \bar{d}_{x_{1:k}} + d_y; \end{cases}$

iii) $\mathcal{DW} \overset{P}{\to} 2 - 2\rho_y(1) = \frac{2(1-2d_y)}{1-d_y}$;

where $\bar{d}_{x_{r:s}} := \max\{d_{x_i} \mid r \le i \le s\}$; and $\overset{P}{\to}$, and $O_p(\cdot)$ are short for convergence in probability and order in probability, respectively.

Assumption 2: for the "cointegrated case", arguably the antipode of the previous one, the assumption is simpler.

- Let $x_{i,t}$, for $i = 1, 2, \ldots, k$ with $k < T$, and ϵ_t be independent stationary fractionally integrated processes that satisfy $(1 - L)^{d_z} z_t = a_{z,t}$, for $z = x_i, \epsilon$, where $a_{z,t}$ are i.i.d. white noises with finite variance $\sigma^2_{a,z}$, and d_x is the order of integration of processes $x_{i,t}$
- d_y is the order of integration of process ϵ_t, such that $0 \leq d_y < d_x < \frac{1}{2}$.
- Assume also that $E\left[a_{z,t}\right]^{q_z} < \infty$ with $q_z \geq \max\left\{4, -\frac{8d_z}{1+2d_z}\right\}$ for all $z = x_i, \epsilon$.
- Finally, let y_t be generated as a linear combination of the previous processes: $y_t = \alpha + \sum_{i=1}^{k} \beta_i x_{i,t} + \epsilon_t$.

Theorem 2

Let Assumption 2 hold. Suppose that the linear specification $y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t$ is estimated by OLS. Then, as $T \to \infty$,

i) $\hat{\beta}_0 \xrightarrow{P} \alpha$;

ii) $\hat{\beta}_i \xrightarrow{P} \beta_i$, for $i = 1, \dots, k$;

iii) $t_{\beta_0} = T \frac{1}{\gamma_\epsilon(0)} \alpha$;

iv) $t_{\beta_i} = T \frac{\gamma_{x_i}(0)}{\gamma_\epsilon(0)} \beta_i$, for $i = 1, \dots, k$;

v) $R^2 \xrightarrow{P} \left[1 - \frac{\gamma_\epsilon(0)}{\sum_{i=1}^{k} \beta_i^2 \gamma_{x_i}(0) + \gamma_\epsilon(0)} \right]$.

Assumption 3: for the nonstationary "spurious case", the assumption is also simpler.

- Let y_t and $x_{i,t}$, for $i = 1, 2, \ldots, k$ and $k < T$, be independent nonstationary fractionally integrated processes of orders d_y and d_{x_i}, respectively, that satisfy $(1-L)^{d_z} z_t = \epsilon_{z,t}$.
- $\epsilon_{z,t}$ are i.i.d. white noises with zero mean and finite variance $\sigma^2_{\epsilon,z}$, and $d_z \in \left(\frac{1}{2}, 1\right)$ for $z = y, x_i$.
- Suppose also that $E\left[\epsilon_{z,t}\right]^{q_z} < \infty$ with $q_z \geq \max\left\{4, -\frac{8d_z}{1+2d_z}\right\}$ for all $z = y, x_i$.

Theorem 3

Let Assumption 3 hold. Suppose that the linear specification $y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t$ is estimated by OLS. Then, as $T \to \infty$,

i) $\hat{\beta}_0 = O_p(T^{d_y - \frac{1}{2}})$;

ii) $\hat{\beta}_i = O_p\left(T^{d_y - d_{x_i}}\right)$, for $i = 1, \ldots, k$;

iii) $t_{\beta_0} = O_p(T^{\frac{1}{2}})$;

iv) $t_{\beta_i} = O_p\left(T^{\frac{1}{2}}\right)$, for $i = 1, \ldots, k$;

Furthermore,

v) $R^2 = O_p(1)$;

vi) $\mathcal{F} = O_p(T)$;

vii) $\mathcal{DW} \xrightarrow{P} 0$.

Interpretation

Under fractionally integration, the risk of drawing nonsense inference is non trivial. No matter whether: (i) there are a few or many regressors; (ii) the regressors are independent from each other or not; (iii) the variables are stationary or not. Nonsense inference remains much in line with Tsay and Chung's (2000) results. Our results show that inference drawn of on such a regression would be "dangerous", since some estimates would appear as statistically significant whilst others wouldn't. Nonetheless, the R^2 in any case should collapse (although sometimes rather slowly). This gave us the idea to compare it with the R^2 of a regression between genuinely (and linearly) related variables...

Monte Carlo Evidence

... A classical tool, the R^2 could be useful to disentangle genuine linear relationships from spurious one. Asymptotically, we found that a test based on the R^2 would depend on unknown parameters. Nonetheless, finite-sample simulations tend to show that the R^2 could still prove useful. Incidentally, they illustrate the theoretical results.

Table: Spurious regression, stationary variables, $\max_k\{d_{x_k}\} + d_y < \frac{1}{2}$.

	$y_t \sim FI(d_y)$		$x_{1,t} \sim FI(d_{x_1}) + w_1$			$x_{2,t} \sim FI(d_{x_2}) + w_2$			$x_{3,t} \sim FI(d_{x_3}) + w_3$		
	d_y	$\sigma^2_{\epsilon,y}$	d_{x_1}	σ^2_{ϵ,x_1}	σ^2_1	d_{x_2}	σ^2_{ϵ,x_2}	σ^2_2	d_{x_3}	σ^2_{ϵ,x_3}	σ^2_3
	0.25	2	0.20	1	1	0.15	0.75	0.75	0.10	0.75	0.40

	$\sigma_{1,2} = 0$; $\sigma_{1,3} = 0$; $\sigma_{2,3} = 0$			$\sigma_{1,2} = 0.4$; $\sigma_{1,3} = 0$; $\sigma_{2,3} = 0$			$\sigma_{1,2} = 0.4$; $\sigma_{1,3} = 0.6$; $\sigma_{2,3} = 0.3$		
T	100	1,000	10,000	100	1,000	10,000	100	1,000	10,000
$RR_{t\beta_0}$	0.4554	0.6736	0.8188	0.4514	0.6737	0.8176	0.4534	0.6855	0.8175
$RR_{t\beta_1}$	0.0618	0.0792	0.0914	0.0622	0.0861	0.0977	0.0707	0.0905	0.1036
$RR_{t\beta_2}$	0.0568	0.0656	0.0735	0.0569	0.0656	0.0791	0.0597	0.0688	0.0744
$RR_{t\beta_3}$	0.0583	0.0642	0.0637	0.0563	0.0567	0.0672	0.0600	0.0705	0.0759
R^2	0.0326	0.0035	0.0004	0.0326	0.0035	0.0004	0.0333	0.0036	0.0004
$RR_{\mathcal{F}}$	0.0647	0.0789	0.0927	0.0630	0.0811	0.1010	0.0686	0.0895	0.1047
$\mathcal{DW}$	1.4973	1.3797	1.3467	1.4977	1.3808	1.3465	1.5010	1.3789	1.3464

RR_t and $RR_{\mathcal{F}}$ account for rejection rate of the t-ratio and the $\mathcal{F}$ tests at a 5% nominal size, respectively. $R = 10,000$.

Table: Spurious regression, stationary variables, $\max_k\{d_{x_k}\} + d_y \geq \frac{1}{2}$.

$y_t \sim FI(d_y)$		$x_{1,t} \sim FI(d_{x_1}) + w_1$			$x_{2,t} \sim FI(d_{x_2}) + w_2$			$x_{3,t} \sim FI(d_{x_3}) + w_3$		
d_y	$\sigma^2_{\epsilon,y}$	d_{x_1}	σ^2_{ϵ,x_1}	σ^2_1	d_{x_2}	σ^2_{ϵ,x_2}	σ^2_2	d_{x_3}	σ^2_{ϵ,x_3}	σ^2_3
0.35	2	0.25	1	1	0.15	0.75	0.75	0.10	0.75	0.40

	$\sigma_{1,2} = 0$; $\sigma_{1,3} = 0$; $\sigma_{2,3} = 0$			$\sigma_{1,2} = 0.4$; $\sigma_{1,3} = 0$; $\sigma_{2,3} = 0$			$\sigma_{1,2} = 0.4$; $\sigma_{1,3} = 0.6$; $\sigma_{2,3} = 0.3$		
T	100	1,000	10,000	100	1,000	10,000	100	1,000	10,000
$RR_{t\beta_0}$	0.6005	0.8076	0.9155	0.5929	0.8058	0.9151	0.5982	0.8135	0.9098
$RR_{t\beta_1}$	0.0800	0.1360	0.2182	0.0816	0.1458	0.2195	0.0922	0.1614	0.2499
$RR_{t\beta_2}$	0.0612	0.0768	0.0972	0.0620	0.0806	0.1155	0.0646	0.0867	0.1077
$RR_{t\beta_3}$	0.0602	0.0754	0.0817	0.0597	0.0665	0.0855	0.0687	0.0935	0.1192
R^2	0.0347	0.0042	0.0005	0.0349	0.0042	0.0005	0.0361	0.0044	0.0006
$RR_{\mathcal{F}}$	0.0781	0.1320	0.2109	0.0789	0.1365	0.2220	0.0888	0.1541	0.2391
$\mathcal{DW}$	1.2627	1.0599	0.9841	1.2628	1.0615	0.9839	1.2666	1.0595	0.9837

Table: Correctly specified fractional cointegration.

$$y_t = 0.70 + 0.2x_{1,t} + 0.3x_{2,t} + 0.4x_{3,t} + u_t;$$
$$z_t \sim FI(d_z);\ \sigma^2_{\epsilon,x_1} = 1;\ \sigma^2_{\epsilon,x_2} = 0.75;\ \sigma^2_{\epsilon,x_3} = 0.75;\ \sigma^2_u = 0.5;\ d_u = 0.2$$

	$d_z = 0.25$ $z = x_1, x_2, x_3$			$d_z = 0.35$ $z = x_1, x_2, x_3$			$d_z = 0.45$ $z = x_1, x_2, x_3$		
T	100	1,000	10,000	100	1,000	10,000	100	1,000	10,000
$RR_{t\beta_0}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$RR_{t\beta_1}$	0.958	1.000	1.000	0.966	1.000	1.000	0.974	1.000	1.000
$RR_{t\beta_2}$	0.984	1.000	1.000	0.987	1.000	1.000	0.990	1.000	1.000
$RR_{t\beta_3}$	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
R^2	0.443	0.434	0.436	0.468	0.480	0.492	0.506	0.554	0.595

Table: Spurious regression, nonstationary variables.

$$y_t \sim FI(d_y);\ x_{1,t} \sim FI(d_{x_1});\ x_{2,t} \sim FI(d_{x_2});\ x_{3,t} \sim FI(d_{x_3})$$

	$d_y = 0.60$; $d_{x_1} = 0.60$; $d_{x_2} = 0.60$; $d_{x_3} = 0.60$			$d_y = 0.80$; $d_{x_1} = 0.80$; $d_{x_2} = 0.80$; $d_{x_3} = 0.80$			$d_y = 0.75$; $d_{x_1} = 0.70$; $d_{x_2} = 0.65$; $d_{x_3} = 0.60$		
T	100	1,000	10,000	100	1,000	10,000	100	1,000	10,000
$RR_{t\beta_0}$	0.6931	0.8931	0.9643	0.7216	0.9081	0.9732	0.7617	0.9204	0.9760
$RR_{t\beta_1}$	0.3906	0.7395	0.9041	0.5701	0.8538	0.9530	0.5250	0.8354	0.9434
$RR_{t\beta_2}$	0.3841	0.7331	0.9024	0.5665	0.8482	0.9515	0.4854	0.8164	0.9374
$RR_{t\beta_3}$	0.3904	0.7395	0.9039	0.5644	0.8517	0.9522	0.4524	0.7860	0.9248
R^2	0.1537	0.1064	0.0867	0.3140	0.2955	0.2931	0.2257	0.1914	0.1780
$RR_{\mathcal{F}}$	0.7082	0.9762	0.9984	0.9096	0.9969	1.0000	0.8342	0.9917	1.0000
$\mathcal{DW}$	0.8174	0.3891	0.2112	0.5143	0.1244	0.0306	0.5726	0.1767	0.0597

Concluding remarks: i ii iii iv v

We studied the asymptotic and the finite-sample behavior of the OLS-estimated multivariate regression with an arbitrary finite number of regressors and a constant term, where all the variables are independent fractionally integrated processes.

Concluding remarks: i ii iii iv v

The *t*-statistics associated to the estimated parameters diverge if the processes underlying the dependent variable and a specific explanatory variable (or its correlated variables) are sufficiently persistent.

Concluding remarks: i ii iii iv v

Comprehensive finite sample evidence is consistent with our asymptotic results and shows that they hold even for small sample sizes such as 100 observations.

Concluding remarks: i ii iii iv v

The finite sample evidence also shows that, although imperfect, the R^2 could prove useful in disentangle spurious from genuine relationships.

Concluding remarks: i ii iii iv v

Tsay & Chung original results were extended not only by studying OLS-estimated multivariate regressions, but also by allowing for a certain degree of correlation between some explanatory variables, through a realistic signal-noise DGP.

And yet, we should: i ii iii iv

Allow for combinations of stationary and nonstationary variables in the specification, for the inclusion of deterministic trends and combinations thereof.

And yet, we should: i ii iii iv

Allow for correlation between the regressors under cointegration.

And yet, we should: i ii iii iv

Extend the search for a testing procedure to distinguish between spurious and genuine regressions.

And yet, we should: i ii iii iv

Provide theoretical and finite sample evidence about the usefulness of the (fractional-) differencing strategy.

The paper is available ... upon request to the authors:

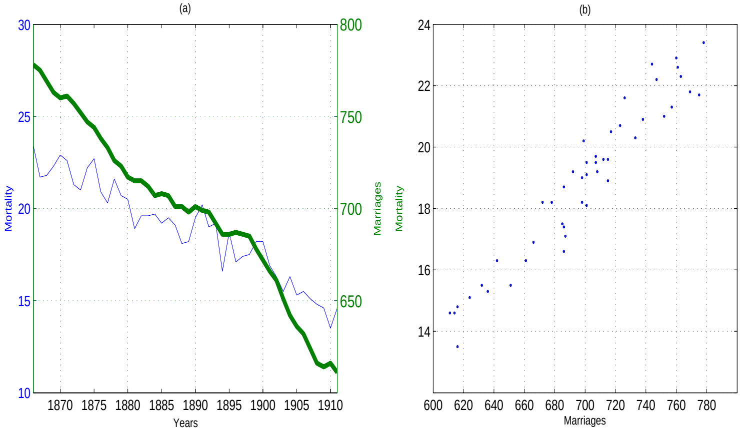
http://www.ventosa-santaularia.com

Thank you

Suppose the following data from England and Wales (1866-1911):

- Proportion of Church of England Marriages to all Marriages (per 1,000 persons)
- Mortality Rate (per 1,000 persons)

Is there a relationship?



Misleading LS inference

- If you estimate by LS the following specification:
- $Mortality_t = \alpha + \beta Marriages_t$,
- t-stat associated to β: $t_\beta = 23.54 > 1.96$,
- LS therefore produces nonsense inference.

Back