MA Advanced Econometrics: Spurious Regressions and Cointegration

Karl Whelan

School of Economics, UCD

February 22, 2011

A More Serious Problem: Spurious Regressions

- When estimating a univariate time series process, we are often interested in calculating the value of ρ and whether this value equals one or something less can be of interest: We may be interested in whether shocks have permanent or temporary effects and, if temporary, how long they take to fade away. This is one reason to teach about the non-standard distributions that occur when a time series is nonstationary.
- However, there is a deeper problem when analysing nonstationary time series.
- Most of econometrics is concerned with assessing relationships between variables: Usually, we are asking the question "Does x have an effect on y?" But when two different unrelated nonstationary series are regressed on each other, the result is usually a so-called spurious regression, in which the OLS estimates and t statistics indicate that a relationship exists when, in reality, there is no such relationship.
- The modern literature on this dates from a famous paper by Granger and Newbold from 1974. However, the nature of the problem was known at least as far back as 1926.

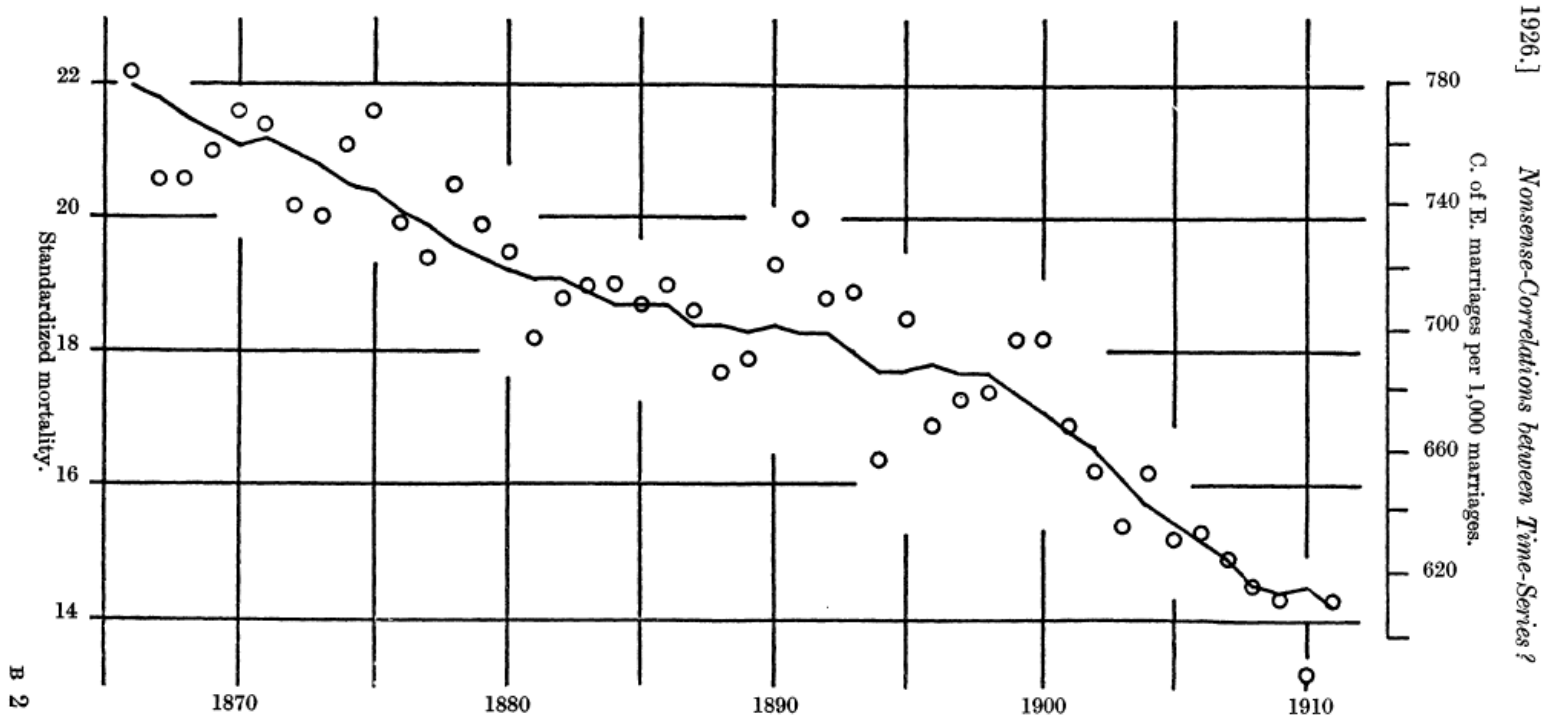
Yule (1926) on Nonsense Correlations

In 1926, Georges Udny Yule wrote a paper in the *Journal of the Royal Statistical Society* called "Why Do We Sometimes get Nonsense Correlations between Time-Series?"

SECTION I.—*The problem.*

It is fairly familiar knowledge that we sometimes obtain between quantities varying with the time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly "significant." As the occurrence of such "nonsense-correlations" makes one mistrust the serious arguments that are sometimes put forward on the basis of correlations between time-series—my readers can supply their own examples—it is important to clear up the problem how they arise and in what special cases.

George Udny Yule's Chart from 1926



1926.] *Nonsense-Correlations between Time-Series?* 3

B 2

FIG. 1.—Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0{\cdot}9512$.

Yule's Discussion of His Chart

Fig. 1 gives a very good illustration. The full line shows the proportion of Church of England marriages to all marriages for the years 1866–1911 inclusive: the small circles give the standardized mortality per 1,000 persons for the same years. Evidently there is a very high correlation between the two figures for the same year: the correlation coefficient actually works out at + 0·9512.

Now I suppose it is possible, given a little ingenuity and good-will, to rationalize very nearly anything. And I can imagine some enthusiast arguing that the fall in the proportion of Church of England marriages is simply due to the Spread of Scientific Thinking since 1866, and the fall in mortality is also clearly to be ascribed to the Progress of Science; hence both variables are largely or mainly influenced by a common factor and consequently ought to be highly correlated. But most people would, I think, agree with me that the correlation is simply sheer nonsense; that it has no meaning whatever; that it is absurd to suppose that the two variables in question are in any sort of way, however indirect, causally related to one another.

Spurious Regressions: Unit Roots with Drifts

- When discussing spurious regressions, econometric textbooks tend to focus on what happens when we take processes that are unit roots without drift (i.e. $y_t = y_{t-1} + \epsilon_t$ with no constant term) and regress them on each other.
- In applied econometric work, however, unit root without drift processes are not very common. Generally, we work with series that tend to be stationary or else with series that have a clear upward trend and which may be unit root processes with drift (i.e. take the form $y_t = \alpha + y_{t-1} + \epsilon_t$.)
- While explanations of how the spurious regression problem works for non-drifting unit root processes are quite complex, the spurious regression problem is far more relevant in the case where the processes have drift. It also turns out that the problem is easier to explain in this case.
- A property of drifting unit root processes that we will use is the following

$$
\begin{aligned}
y_t &= \alpha + y_{t-1} + \epsilon_t && (1)\\
&= \alpha + \alpha + y_{t-2} + \epsilon_t + \epsilon_{t-1} && (2)\\
&= \alpha t + \sum_{k=1}^{t} \epsilon_k + y_0 && (3)
\end{aligned}
$$

Useful Results About Infinite Sums

- Establishing properties about regressions involving drifting unit root series will require figuring out properties of sums of the form $\sum_{t=1}^{T} t$ and $\sum_{t=1}^{T} t^2$.
- Note that $1+2+3=6=\frac{(3)(4)}{2}$ and $1+2+3+4=10=\frac{(4)(5)}{2}$. The general rule is

$$\sum_{t=1}^{T} t = \frac{T(T+1)}{2} = \frac{1}{2}\left(T^2+T\right) \tag{4}$$

- For sums of squares, we have

$$\sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6} = \frac{1}{6}\left(2T^3+3T^2+T\right) \tag{5}$$

- This means that as $T \to \infty$

$$\frac{1}{T^2}\sum_{t=1}^{T} t \quad \to \quad \frac{1}{2} \tag{6}$$

$$\frac{1}{T^3}\sum_{t=1}^{T} t^2 \quad \to \quad \frac{1}{3} \tag{7}$$

Regressions Featuring Unit Roots with Drifts

- Consider regressing y_t on the completely unrelated series x_t where

$$y_t = \alpha_y + y_{t-1} + \epsilon_t^y \tag{8}$$
$$x_t = \alpha_x + x_{t-1} + \epsilon_t^x \tag{9}$$

- The OLS estimator is

$$\hat{\beta} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2} \tag{10}$$
$$= \frac{\sum_{t=1}^{T} \left(\alpha_x t + \sum_{k=1}^{t} \epsilon_k^x + x_0\right)\left(\alpha_y t + \sum_{k=1}^{t} \epsilon_k^y + y_0\right)}{\sum_{t=1}^{T} \left(\alpha_x t + \sum_{k=1}^{t} \epsilon_k^x + x_0\right)^2} \tag{11}$$

- As T gets large, the terms in t^2 will dominate all other terms. Re-writing this as

$$\hat{\beta} = \frac{\frac{1}{T^3}\sum_{t=1}^{T} \left(\alpha_x t + \sum_{k=1}^{t} \epsilon_k^x + x_0\right)\left(\alpha_y t + \sum_{k=1}^{t} \epsilon_k^y + y_0\right)}{\frac{1}{T^3}\sum_{t=1}^{T} \left(\alpha_x t + \sum_{k=1}^{t} \epsilon_k^x + x_0\right)^2} \tag{12}$$

then all of the terms that are not of the form $\frac{1}{T^3}\sum_{t=1}^{T} t^2$ will go to zero.

Spurious Regression Results

- This means that as T gets large

$$\hat{\beta} \xrightarrow{p} \frac{\alpha_x \alpha_y}{\alpha_x^2} = \frac{\alpha_y}{\alpha_x} \tag{13}$$

- In other words, the OLS estimator will tend towards the ratio of the two drift terms. In addition, the t statistics will generally indicate that there is a highly statistically significant relationship.
- The next pages show $\hat{\beta}$'s and t-stats from regressing y_t on x_t where

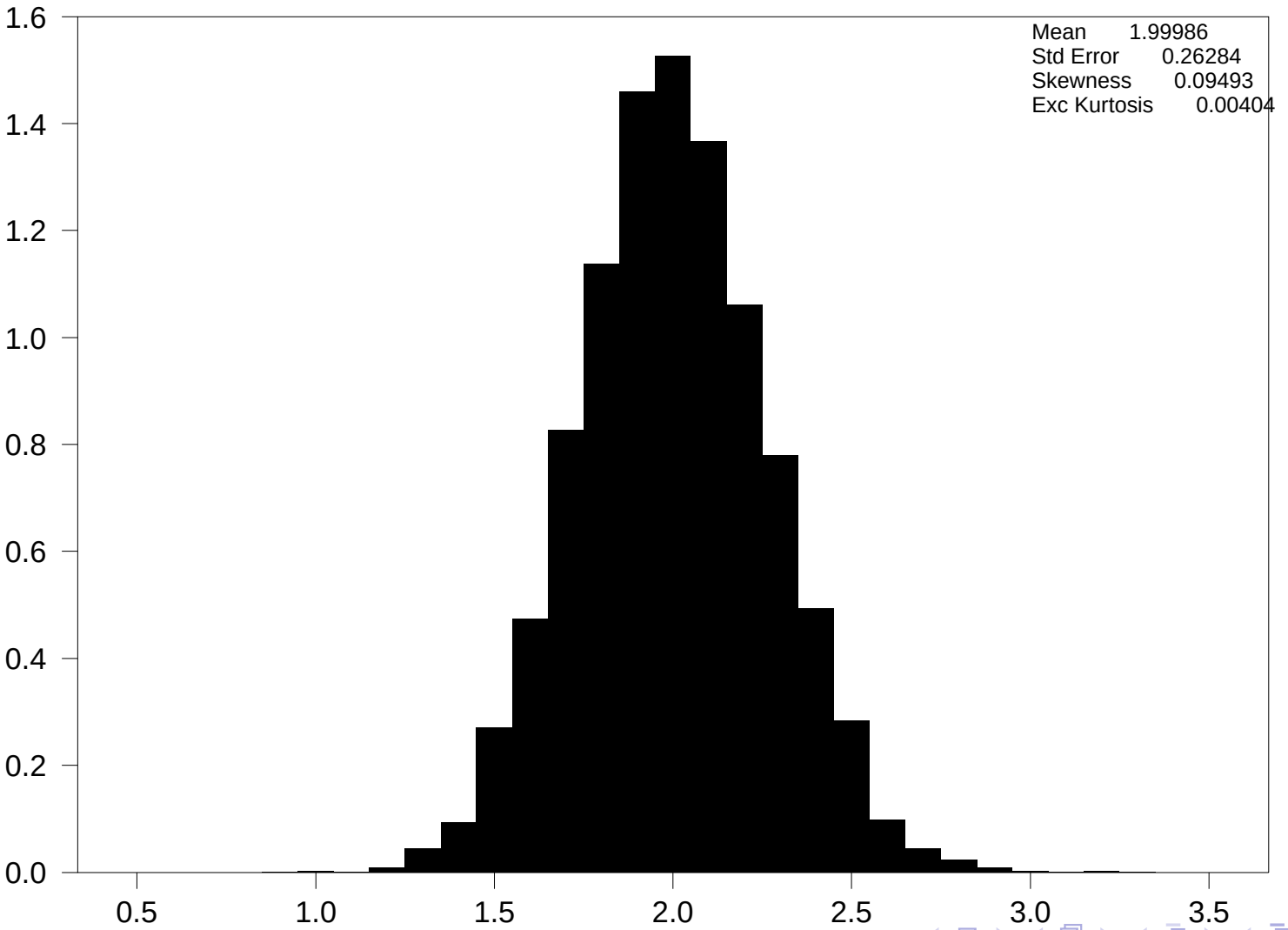
$$y_t = 0.2 + y_{t-1} + \epsilon_t^y \tag{14}$$

$$x_t = 0.1 + x_{t-1} + \epsilon_t^x \tag{15}$$

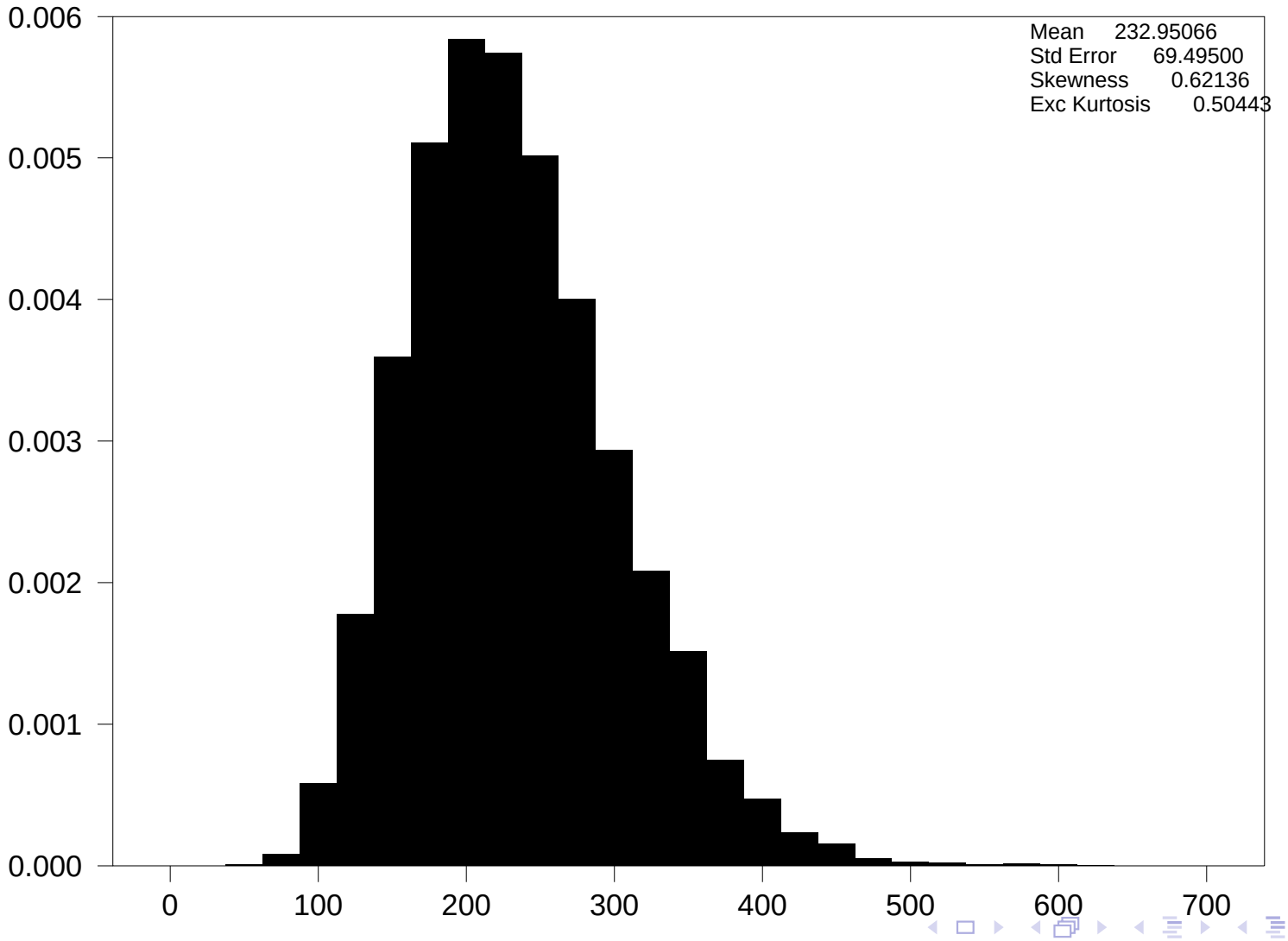
where the error terms are i.i.d. normally distributed errors. They show OLS coefficients averaging 2 and highly significant t-stats.

- Note that the key terms driving these results were the time trends. These results also apply to "trend stationary" series like $y_t = \alpha t + \rho y_{t-1} + \epsilon_t$, so the problem is not specific to the unit root series.
- Similar results apply to regressions featuring unit roots without drifts but deriving these results analytically is beyond the scope of this module.

Distribution of β from Regressions of Unrelated Unit Roots With Drift ($T = 500$)



Distribution of t Statistics from Regressions of Unrelated Unit Roots With Drift ($T = 500$)



The $I(k)$ Terminology and Cointegration

- Unit root series such as $y_t = \delta + y_{t-1} + \epsilon_t$ are nonstationary: Sums of y_t don't settle down at a stable mean and it's covariances change over time.
- However, once you calculate the first difference of this series $\Delta y_t = \delta + \epsilon_t$, it becomes a covariance stationary series.
- We say a series is **integrated of order** k (denoted $I(k)$) if it has to be differenced k times before it becomes stationary. Sometimes, one can can come across examples involving $I(2)$ series, but generally the time series in practical applications are either $I(1)$ or $I(0)$.
- The spurious regression problem can be stated as the fact that unrelated $I(1)$ series regressed upon each other tend to appear to be related according to the usual OLS diagnostics.
- However, what if there really is a relationship? For example, what if y_t and x_t are both $I(1)$ series but there existed a coefficient β such that $y_t - \beta x_t \sim I(0)$. In this case, there is a common trend across the series and we say that the series y_t and x_t are **cointegrated**.
- In this case, it turns out that OLS estimates of β are consistent.

Consistency of OLS Under Cointegration

- Consider again the case where x_t is a unit root with drift

$$x_t = \alpha_x + x_{t-1} + \epsilon_t^x \tag{16}$$

but in this case the variable y_t is cointegrated with x_t so that

$$y_t = \beta x_t + u_t \tag{17}$$

where u_t is mean-zero $I(0)$ series.

- We can calculate the properties of the OLS estimator as follows:

$$\hat{\beta} = \beta + \frac{\sum_{t=1}^{T} x_t u_t}{\sum_{t=1}^{t} x_t^2} \tag{18}$$

$$= \beta + \frac{\sum_{t=1}^{T} \left(\alpha_x t + \sum_{k=1}^{t} \epsilon_k^x + x_0\right) u_t}{\sum_{t=1}^{T} \left(\alpha_x t + \sum_{k=1}^{t} \epsilon_k^x + x_0\right)^2} \tag{19}$$

- In this case, the terms in t^2 will dominate as $T \to \infty$ so that the denominator of the last term will grow faster than the numerator. This means that $\hat{\beta} \xrightarrow{p} \beta$. (One could show this more formally using the formulae for infinite sums derived earlier.)

Super-Consistency!

- Not only is the OLS estimator $\hat{\beta}$ of a cointegrating regression consistent, in the sense that it is likely to get ever-closer to the true value of β as samples get larger, it turns out it's **superconsistent**. What's that mean?
- Now multiply both sides of (19) by T but do this to the right hand side by dividing the numerator by T^2 and the denominator by T^3:

$$T\left(\hat{\beta}-\beta\right)=\frac{\frac{1}{T^2}\sum_{t=1}^{T}\left(\alpha_x t+\sum_{k=1}^{t}\epsilon_k^x+x_0\right)u_t}{\frac{1}{T^3}\sum_{t=1}^{T}\left(\alpha_x t+\sum_{k=1}^{t}\epsilon_k^x+x_0\right)^2} \tag{20}$$

- The numerator converges to zero (the sum $\frac{1}{T^2}\sum_{t=1}^{T}\alpha_x t\rightarrow\frac{\alpha_x}{2}$ but is multiplied by an uncorrelated mean zero series u_t) while the denominator converges to $\frac{\alpha_x^2}{3}$. So $T\left(\hat{\beta}-\beta\right)\xrightarrow{p}0$.
- We have seen cases before where $\hat{\beta}-\beta$ converges in distribution to a mean zero series when multiplied by $\sqrt{T}$. In this case, the gap between the estimator and the true value converges in probability to zero even when multiplied by T. This property is known as superconsistency.

The Error-Correction Representation

- Consider two $I(1)$ series, y_t and x_t. We would expect their first-differences to have stationary representations

$$\Delta y_t = \alpha^y + \gamma_1^y \Delta y_{t-1} + ... + \gamma_k^y \Delta y_{t-k} + \epsilon_t^y \quad (21)$$
$$\Delta x_t = \alpha^x + \gamma_1^x \Delta x_{t-1} + ... + \gamma_k^x \Delta x_{t-k} + \epsilon_t^x \quad (22)$$

- Now suppose that y_t and x_t are cointegrated. This means there exists a value β such that $y_t - \beta x_t \sim I(0)$. But if the processes are as described above, then there is nothing about the behaviour of either series that would see the two series tending to move together. So, additional terms are required to describe these processes.
- Specifically, we need additional **error-correction** terms of the form $y_t - \beta x_t$, to get a representation of the form

$$\Delta y_t = \alpha^y + \gamma_1^y \Delta y_{t-1} + ... + \gamma_k^y \Delta y_{t-k} + \theta_y (y_t - \beta x_t) + \epsilon_t^y \quad (23)$$
$$\Delta x_t = \alpha^x + \gamma_1^x \Delta x_{t-1} + ... + \gamma_k^x \Delta x_{t-k} + \theta_x (y_t - \beta x_t) + \epsilon_t^x \quad (24)$$

where we expect to have $\theta_y \leq 0$ and $\theta_x \geq 0$. In other words, when y_t rises above its long-run relationship with x_t it tends to fall back and/or x_t tends to increase.

The Vector Error-Correction Representation

- When there are only two series, any potential cointegrating vector is unique up to multiplication by a scalar (e.g. we could say $y_t - \beta x_t \sim I(0)$ or that $x_t - \beta^{-1} y_t \sim I(0)$).
- However, when there are n different variables, then there may be multiple cointegrating vectors, e.g. for $Y_t = (y_{1t}, y_{2t}, y_{3t}, y_{4t})$, one could have $y_{1t} - \gamma_1 y_{3t} \sim I(0)$ and $y_{2t} - \gamma_1 y_{4t} \sim I(0)$.
- Consider the general case, in which there are r cointegrating relationships among n variables. Specifically, consider the case in which the $n \times 1$ vector of $I(1)$ series Y_t has the property that there exists an $r \times n$ matrix A such that the r series defined by $Z_t = AY_t$ are all $I(0)$. In this case, there exists an $n \times r$ matrix B such that Y_t is described by a Vector Error Correction Mechanism representation

$$\begin{aligned} \Delta Y_t &= \gamma_1^x \Delta Y_{t-1} + ... + \gamma_k^x \Delta Y_{t-k} + \alpha + BZ_{t-1} + \epsilon_t \quad (25) \\ &= \gamma_1^x \Delta Y_{t-1} + ... + \gamma_k^x \Delta Y_{t-k} + \alpha + BAY_{t-1} + \epsilon_t \quad (26) \end{aligned}$$

- This result is part of what is known as the Granger Representation Theorem.

Testing for Cointegration

- Suppose we have two $I(1)$ series, y_t and x_t. How do we test whether they are cointegrated or whether the relationship between them is spurious? Tests are based on the idea that if there is no underlying relationship than the OLS residuals, $\hat{u}_t = y_t - \hat{\beta}x_t$ will also have a unit root.
- One might be tempted to simply apply an augmented Dickey-Fuller test to $\hat{u}_t$. However, the OLS procedure produces residuals that may appear stationary, even when applying the DF critical values.
- This means that special critical values must be applied when testing for cointegration. These critical values differ depending on whether the underlying y_t and x_t series have drifts and on whether the potential cointegrating regression includes a constant.
- In the case where the y_t and x_t series are both unit roots with drift and the regression includes a contant, the critical values for testing for a unit root in $\hat{u}_t$ are the same as those presented in the previous notes for testing for a unit root against the alternative of trend stationarity (see next page).
- When testing for r different cointegrating vectors among n variables, testing procedures involve estimating a VAR process and assess whether the relevant VECM is the best fit for the data.

t Tests of $\rho = 1$ Applied to Residuals from Regressing Two Unrelated Random Walks with Drift On Each Other

Statistics on Series ADFS			
Observations	10000		
Sample Mean	-2.150141	Variance	0.612456
Standard Error	0.782596	of Sample Mean	0.007826
t-Statistic (Mean=0)	-274.744713	Signif Level	0.000000
Skewness	0.165626	Signif Level (Sk=0)	0.000000
Kurtosis (excess)	0.700088	Signif Level (Ku=0)	0.000000
Jarque-Bera	249.938203	Signif Level (JB=0)	0.000000
Minimum	-5.695853	Maximum	1.579588
01-%ile	-3.979664	99-%ile	-0.073867
05-%ile	-3.419176	95-%ile	-0.871311
10-%ile	-3.129077	90-%ile	-1.204794
25-%ile	-2.655159	75-%ile	-1.668208
Median	-2.144511		