Lecture 5: Random walk and spurious correlation

UCSD, Winter 2017

Allan Timmermann[1]

[1]UC San Diego

Random walk model

- The random walk model is an AR(1) $y_t = \phi_1 y_{t-1} + \varepsilon_t$ with $\phi_1 = 1$:

$$y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

  This model implies that **the change** in y_t is unpredictable:

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t$$

- For example, the level of (log-) stock prices is easy to predict, but not its change (rate of return for log-prices)
- Shocks to the random walk have permanent effects: A one unit shock moves the series by one unit forever. This is in sharp contrast to a mean-reverting process such as $y_t = 0.8y_{t-1} + \varepsilon_t$

Random walk model (cont)

- The variance of a random walk increases over time so the distribution of y_t changes over time. Suppose that y_t started at zero, $y_0 = 0$:

$$
\begin{aligned}
y_1 &= y_0 + \varepsilon_1 = \varepsilon_1 \\
y_2 &= y_1 + \varepsilon_2 = \varepsilon_1 + \varepsilon_2 \\
&\vdots \\
y_t &= \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{t-1} + \varepsilon_t, \quad \text{so}
\end{aligned}
$$

$$
\begin{aligned}
E[y_t] &= 0 \\
var(y_t) &= var(\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t) = t\sigma^2 \Rightarrow \\
\lim_{t\to\infty} var(y_t) &= \infty
\end{aligned}
$$

- The variance of y grows proportionally with time
- A random walk does not revert back to the mean but wanders up and down at random

Forecasts from random walk model

- Recall that forecasts from the AR(1) process $y_t = \phi_1 y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$ are simply

$$f_{t+h|t} = \phi_1^h y_t$$

- For the random walk model $\phi_1 = 1$, so for all forecast horizons, h, the forecast is simply the current value:

$$f_{t+h|t} = y_t$$

Forecast of tomorrow = today's value

- The basic random walk model says that the value of the series next period (given the history of the series) equals its current value plus an unpredictable change. Random steps, ε_t, make y_t a "random walk"

Random walk with a drift

- Introduce a non-zero drift term, δ :

$$y_t = \delta + y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2).$$

- This is a popular model for the logarithm of stock prices
- The drift term, δ, plays the same role as a time trend. Assuming again that the series started at y_0, we have

$$\begin{aligned} y_t &= 2\delta + y_{t-2} + \varepsilon_t + \varepsilon_{t-1} \\ &= \delta t + y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{t-1} + \varepsilon_t, \quad \text{so} \end{aligned}$$

$$\begin{aligned} E[y_t] &= y_0 + \delta t \\ var(y_t) &= t\sigma^2 \\ \lim_{t \to \infty} var(y_t) &= \infty \end{aligned}$$

Summary of properties of random walk

- Changes in a random walk are unpredictable
- Shocks have permanent effects
- Variance grows in proportion with the forecast horizon
- These points are important for forecasting:
  - point forecasts never revert to a mean or a trend
  - since the variance goes to infinity, the width of interval forecasts increases without bound as the forecast horizon grows. Uncertainty grows without bounds.

Logs, levels and growth rates

- Certain transformations of economic variables such as their logarithm are often easier to model than the "raw" data
- If the standard deviation of a time series is proportional to its level, then the standard deviation of the logarithm of the series is approximately constant:

$$\begin{aligned} Y_t &= Y_{t-1}\exp(\varepsilon_t), \quad \varepsilon_t \sim (0,\sigma^2) \Leftrightarrow \\ \ln(Y_t) &= \ln(Y_{t-1}) + \varepsilon_t \end{aligned}$$

- The first difference of the log of Y_t is $\Delta\ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$
- The percentage change in Y_t between $t-1$ and t is approximately $100\Delta\ln(Y_t)$. This can be interpreted as a growth rate
- Example: US GDP follows an upward trend. Instead of studying the **level** of US GDP, we can study its **growth rate** which is not trending

Unit root processes

- Random walk is a special case of a unit root process which has a unit root in the AR polynomial, i.e.,

$$(1 - L)y_t = \theta(L)\varepsilon_t,$$

- We can test for a unit root using an Augmented Dickey Fuller (ADF) test:

$$\Delta y_t = \alpha + \beta y_{t-1} + \sum_{i=1}^{p} \lambda_i \Delta y_{t-i} + \varepsilon_t.$$

- Under the null of a unit root, $H_0 : \beta = 0$. Under the alternative of stationarity, $H_1 : \beta < 0$

Unit root processes (cont.)

- Example: suppose $p = 0$ (no autoregressive terms for Δy_t) and $\beta = -0.2$. Then

$$\begin{aligned}\Delta y_t &= y_t - y_{t-1} = \alpha - 0.2 y_{t-1} + \varepsilon_t \Leftrightarrow \\ y_t &= 0.8 y_{t-1} + \varepsilon_t \quad \text{(which is stationary)}\end{aligned}$$

- If instead $\beta = 0.2$, we have

$$\begin{aligned}y_t - y_{t-1} &= \alpha + 0.2 y_{t-1} + \varepsilon_t \Leftrightarrow \\ y_t &= 1.2 y_{t-1} + \varepsilon_t \quad \text{(which is explosive)}\end{aligned}$$

- Test is based on the t-stat of β. Test statistic follows a non-standard distribution with wider tails than the normal distribution

Unit root test in matlab

- In matlab: *adftest*
- [h,pValue,stat,cValue,reg] = adftest(y)
- [h,pvalue,stat,cvalue] = adftest(logprice,'lags',1,'model','AR');

Critical values for Dickey-Fuller test

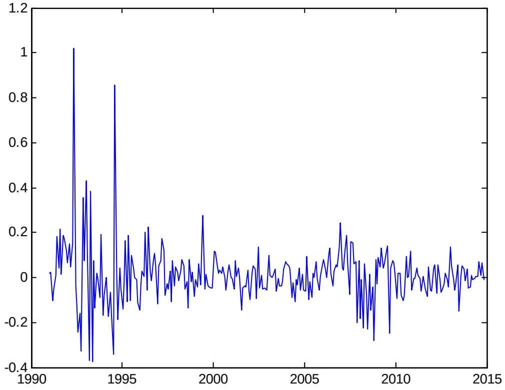
Critical values for Dickey–Fuller *t*-distribution.				
	Without trend		With trend	
Sample size	1%	5%	1%	5%
T = 25	−3.75	−3.00	−4.38	−3.60
T = 50	−3.58	−2.93	−4.15	−3.50
T = 100	−3.51	−2.89	−4.04	−3.45
T = 250	−3.46	−2.88	−3.99	−3.43
T = 500	−3.44	−2.87	−3.98	−3.42
T = ∞	−3.43	−2.86	−3.96	−3.41

Shanghai SE stock price (monthly, 1991-2014)

t-statistic: $1,0362$. p-value:0.92. Fail to reject null of a unit root.

Changes in Shanghai SE stock price

t-statistic: -11.15. $p-$value: 0.001. Reject null of a unit root.



Spurious correlation

- Time series that are trending systematically up or down may appear to be significantly correlated even though they are completely independent
- Correlation between a city's ice cream sales and the number of drownings in city swimming pools: Both peak at the same time even though there is no causal relationship between the two. In fact, a heat wave may drive both variables
- Dutch statistics reveal a positive correlation between the number of storks nesting in the spring and the number of human babies born at that time. Any causal relation?
- Cumulative rainfall in Brazil and US stock prices

Spurious correlation

- Two series with a random walk (unit root) component may appear to be related even when they are not. Consider an example:

$$\begin{aligned} y_{1t} &= y_{1t-1} + \varepsilon_{1t} \\ y_{2t} &= y_{2t-1} + \varepsilon_{2t}, \\ cov(\varepsilon_{1t}, \varepsilon_{2t}) &= 0 \end{aligned}$$

- Regressing one variable on the other $y_{2t} = \alpha + \beta y_{1t} + u_t$ often leads to apparently high values of R^2 and of the associated $t-$statistic for β. Both are unreliable! Solutions:
  - instead of regressing y_{1t} on y_{2t} in levels, regress Δy_{1t} on Δy_{2t}
  - use cointegration analysis

Simulations of stationary processes

- 1,000 simulations $(T = 500)$ of uncorrelated stationary AR(1) processes:

$$
\begin{aligned}
y_{1t} &= 0.5y_{1t-1} + \varepsilon_{1t} \\
y_{2t} &= 0.5y_{2t-1} + \varepsilon_{2t} \\
cov(\varepsilon_{1t}, \varepsilon_{2t}) &= 0
\end{aligned}
$$

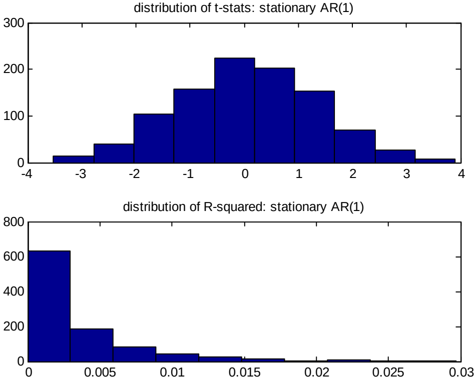
- The two time series y_{1t} and y_{2t} are independent **by construction**
- Next, estimate a regression

$$y_{1t} = \beta_0 + \beta_1 y_{2t} + u_t$$

- What do you expect to find?

Simulation from stationary AR(1) process

Average $t-$stat: 1.02. Rejection rate: 5.7%. Average R^2 : 0.003



Spurious correlation: simulations

- 1,000 simulations of uncorrelated **random walk** processes:

$$
\begin{aligned}
y_{1t} &= y_{1t-1} + \varepsilon_{1t} \\
y_{2t} &= y_{2t-1} + \varepsilon_{2t} \\
cov(\varepsilon_{1t}, \varepsilon_{2t}) &= 0
\end{aligned}
$$

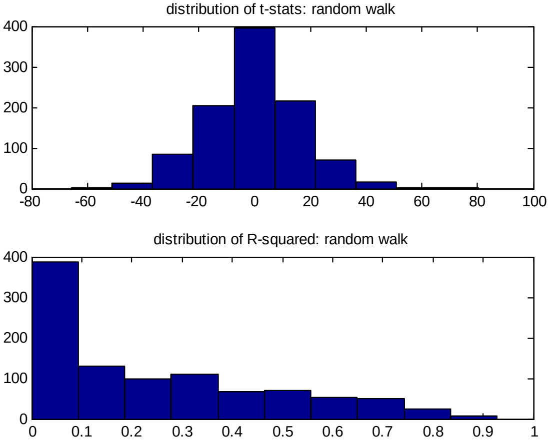
- Then estimate regression

$$y_{1t} = \beta_0 + \beta_1 y_{2t} + u_t$$

- What do we find now?

Spurious correlation: simulation from random walk

Average $t-$stat: 13.4. Rejection rate: 44%. Average $R^2: 0.25$



Spurious correlation: dealing with the problem

- 1,000 simulations of uncorrelated **random walk** processes:

$$
\begin{aligned}
y_{1t} &= y_{1t-1} + \varepsilon_{1t} \\
y_{2t} &= y_{2t-1} + \varepsilon_{2t} \\
cov(\varepsilon_{1t}, \varepsilon_{2t}) &= 0
\end{aligned}
$$

- Next, estimate regression on **first-differenced** series:

$$\Delta y_{1t} = \beta_0 + \beta \Delta y_{2t} + u_t$$

Average $t-$stat: 0.78. Rejection rate: 1.5%. Average R^2 : 0.002