# Time After Time I

## Statistical Issues in Univariate Time-Series Data

Jan Rovny

Sciences Po, Paris

February 7, 2020

# Time-Series 1

Time
: Time-series has a clear temporal ordering and is indexed by time

Sampling
: A time-series is obviously not a random sample but it is still the outcome of stochastic process

Population
: For time-series is the set of all possible realizations of the time series.

# Time-Series 2

- Given the time component to time-series, the individual observations are not truly independent of each other, they are ordered in time.
- This effectively removes the assumption of independence, central to cross-sectional sampling.
- Consequently, there are two major concerns in time-series analysis

1. Stationarity
2. Dependence
    - It is also called 'Persistence'

# Stationarity 1

- Stationarity is needed to assess meaningful relationships between variables
- Trends (kind of non-stationary TS) are always correlated (piracy and global warming)

**Strict Stationarity**

A time-series is stationary if the joint distribution of a collection of observations is the same as that of a collection down the series.

- That is, mean, variance, and all higher moments (skewness, kurtosis, etc.) are independent on $t$.
- This is a very strict definition

# Stationarity 2

- A weaker form of stationarity

**Covariance Stationary Processes**

A time-series is covariance stationary if (1) the first two moments (mean & variance) stay the same over time and (2) the covariance between $y_t$ and $y_{t+h}$ depends only on the distance between the two observations, not where in the series one started.

- There are no restrictions on how $y_t$ and $y_{t+1}$ are related to each other. They can be correlated.

# Dependence

**Weak Dependence**

A covariance stationary time-series is weakly dependent if the correlation between $y_t$ and $y_{t+h}$ goes to zero "sufficiently quickly" as h increases.

- Under weak dependence the covariance stationary series is asymptotically uncorrelated.
- Weak dependence plays the role of random sampling in time series analysis by ensuring that the law of large number and the central limit theorem hold

# Looking at trends

- Look at the time-series graph

  R: `ts(y)`, informs R that $y$ is a time-series variable

  R: `ts.plot(y)`, draws a line graph of $y$.
- Testing for a trend: Dicky-Fuller test for unit root (see below)

  R: `adf.test(y)` (in library `tseries`)

  Note that $H_o$ is non-stationarity
- Linear trend is easily modeled as

  $y_t = \alpha_0 + \alpha_1 t + e_t, t = 1, 2, ...$

  if $\alpha_1 \neq 0$, then have a trend

  Stata: `lm(y~t)`, then observe ttest on coefficient for $t$.

# Removing trends 1

- May include trend term ($t$) in regression,
  say: $y$ depends on observable $x_1$ and $x_2$,
  and unobserved trending factors:
  $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + e$
- Can 'detrend' the variables:
  regress $y, x_1$, and $x_2$ each on a trend term
  - For example, $x_1 = \gamma_0 + \gamma_1 t$
    R: `m1<-lm(x1~t)`
  - Save the residuals, $\ddot{y}_t, \ddot{x}_{t1}, \ddot{x}_{t2}$,
    R: `x1.r<-m1$residuals`
  - Regress $\ddot{y}_t$ on $\ddot{x}_{t1}, \ddot{x}_{t2}$
    R: `m2<-lm(y.r ~ x1.r + x2.r)`

# Removing trends 2

- Alternately (and most commonly), we can difference a time-series to remove a trend:

$$y_t = \alpha_0 + \alpha_1 t + e_t$$

- Here $y_t$ dependes on $t$
- Differencing implies subtracting previous values of $y$ from current $y$: $y_t - y_{t-1}$
  $y_t - y_{t-1} = (\alpha_0 + \alpha_1 t + e_t) - [\alpha_0 + \alpha_1(t-1) + e_{t-1}] = \alpha_1 + e_t - e_{t-1}$
- Notice that here $y_t$ does not depend on $t$.
- Depending on time-series, might need to difference twice, or difference a log of time-series etc. (See Becketti 2013: 230)
- R time operators:
  - Difference: `diff(x)`, if higher order is needed, specify `diff(x, difference=2)` for second difference etc.
  - Lag: `lag(x)` if longer lags are needed: `lag(x, k = 2)` etc., where $k$ is the lag length.

# Time-series Processes

# Time-series processes

- The key issue in time-series is how their values are related to each other
- The main concern is that past values are (somehow) reflected in current values (we are not free of our past)
- If this is the case, our observations are not independent of each other
- This then violates the OLS assumption of error independence $Cov(e_t, e_{t-j}) = 0, \forall j$

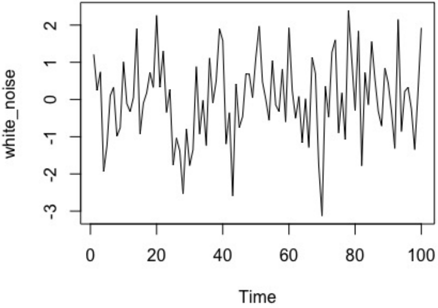
# Time-series processes

Typical time-series error aggregation processes:

1. White Noise
2. Moving Average
3. Autoregressive
4. Integration
   - Random Walk

# 1) White Noise

- The initial building block of a time-series
- White noise, $e_t$, is truly random error: $E(e_t) = 0$; $V(e_t) = \sigma^2$; $Cov(e_t, e_{t-j}) = 0, \forall j$
- White noise is our friend, because it is by definition random. We like seeing it.



# 2) Moving Average MA

- A combination of current and past white noise produces a Moving Average process:

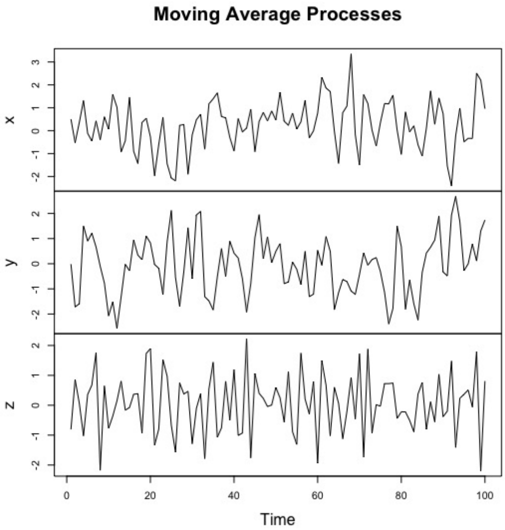
$$y_t = e_t + \psi_1 e_{t-1}, t = 1, 2, ...,$$

- Here $y_t$ is a weighted average of $e_t$ and $e_{t-1}$, a moving average of order 1, MA(1)
- We could have an MA of higher orders, say 3:

$$y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \psi_3 e_{t-3}$$

- $y_t$ is weakly dependent because once the observations are two – MA(1) – or four – MA(3) – periods apart, they are independent.

# 2) Moving Average MA




x: $\psi = 0.5$, y: $\psi = 0.9$, z: $\psi = -0.5$

# 3) Autoregression AR

- A more common time-series process is autoregression:

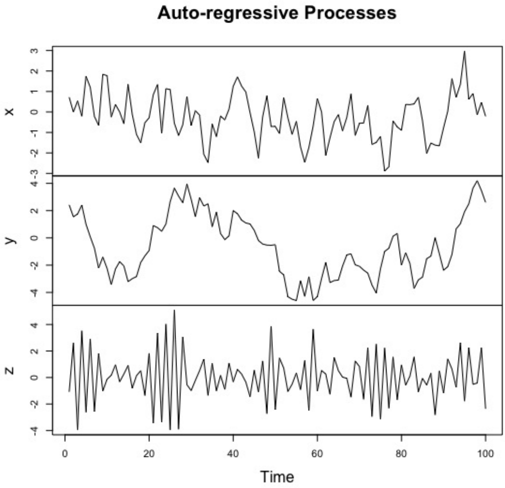
$$y_t = e_t + \phi_1 y_{t-1}, t = 1, 2, ...,$$

- Here $y_t$ is a discounted function of its past value: AR(1)
- A higher order autoregression, say AR(3), is also possible:

$$y_t = e_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3}$$

- As long as $|\phi_i| < 1$ (the stability condition), the influence of $e_{t-j}$ on $y_t$ diminishes as $j$ increases, and $y_t$ is weakly dependent.
- Under the stability condition, AR is a very reasonable model of history, as the impact on past events decays exponentially with time.

# 3) Auto-Regression AR




$$x: \ \phi = 0.5, \quad y: \ \phi = 0.9, \quad z: \ \phi = -0.75$$

# 4) Integration I

- Integrated time-series accumulate history – there is no decay, every update has an infinite impact.
- When we have an AR process and $|\phi_i| = 1$ (unit root), we have a particular integrated series, a drift called random walk

# Random Walk

- A random walk is a time series where the current observation is the previous observation with a random step up or down.
- It is thus the previous observation plus *white noise*

$$x_t = x_{t-1} + w_t$$

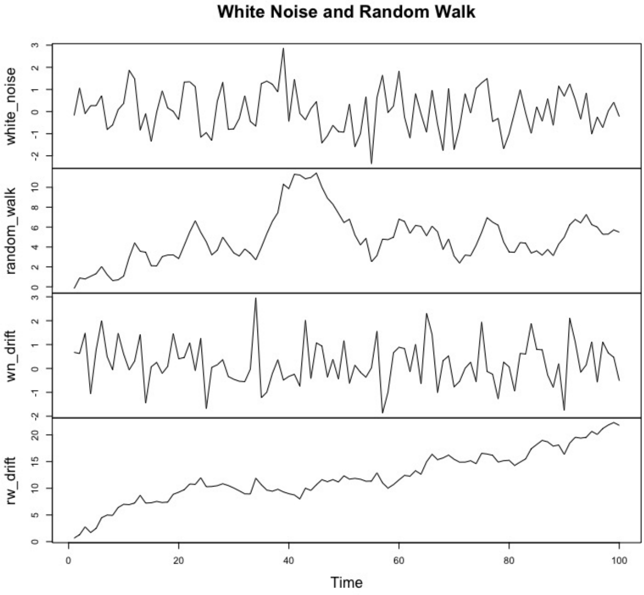
where $w_t$ is a random value = white noise

- A random walk can have a so-called *drift*, which adds a constant $\alpha$

$$x_t = \alpha + x_{t-1} + w_t$$

- The difference of random walk is white noise
- The difference of random walk with a drift is white noise with a drift

# White Noise and Random Walk



# Modelling a time-series

# Modeling a time-series

- To model a time-series means to estimate its MA and AR parameters ($\psi_i$ and $\phi_i$)
- We do this by fitting so-called ARMA($p, q$) models,
  where $p$ refers to the AR order, and $q$ to the MA order of the time-series.
- In case of non-stationary or integrated series, we should estimate an ARIMA($p, d, q$) model,
  where $d$ indicates the order of differencing.

- Box and Jenkins (1970) suggest the following procedure:
  1. Identification: Determine the order of the ARMA model. Is it ARMA(1,1) or ARMA(1,2) etc.?
  2. Estimation: Estimate the parameters $\psi_i$ and $\phi_i$
  3. Diagnostics: Test for the adequacy of the model. If we are successful, we should be left with white noise.

# Identification of a time-series 1

- This is a scarily subjective matter...
- Look at the autocorrelation function and partial autocorrelation
- R: `acf(y)` and `pacf(y, lag = #)`
  - `acf` shows the autocorrelations going back # of lags
  - `pacf` shows the $j$th partial autocorrelation between $y_t$ and $y_{t-j}$ after controlling for the correlations between $y_t$ and the previous lags.

# Identification of a time-series 2

Tips in time-series identification (Becketti 2013: 242)

| Process | Autocorrelation Function | Partial Autocorrelation function |
|---|---|---|
| Non-stationary | AC do not die out<br>they remain large<br>or diminish linearly | |
| Stationary | After first few lags<br>AC die out (collapse to 0)<br>in exponential decay or<br>dampened oscillation | |
| AR($p$) | AC die out | PAC cut off after the first $p$ lags |
| MA($q$) | AC cut off after the first $q$ lags | PAC die out |
| ARMA($p, q$) | AC die out<br>after the first $q - p$ lags | PAC die out<br>after the first $p - q$ lags |

# Estimation of time-series

- If you know the order of AR and MA and have a stationary series, this is the easy part.
- R `arima(y, order=c(#p, #d, #q))`
- R will fit the values for all the specified AR and MA parameters using MLE.

# Diagnostics

- Our aim in modeling a time-series is to arrive at truly random error – white noise.
- If we fit our time series correctly, we should be able to observe no structure in the AC and PAC functions.
- To get a general overview, use the white noise test on the residuals from the ARIMA model:
  R:
  ```
  model<-arima(y, order=c(#p, #d, #q))
     whitenoise.test(model$residuals)
  ```
- $H_0$ is white noise, so want to fail to reject $H_0$ (see higher values on the p statistic)

# Modeling a real time-series

Do example on U.S. GDP in Stata