MW82: Time Series Analysis, Tutorial V

Dr. Giulia Sabbadini

DICE

WS2023/24

Recap: Tutorial IV

We did:

- Box-Jenkins approach
- Information Criteria
- Ljung–Box/Box-Pierce tests

Today we go through:

- Deterministic vs. stochastic trend
- Non-stationarity (Trends, Unit Roots, Seasonality)
- Unit root tests
- Seasonal ARIMA

Deterministic trends

- The slope of the trend is not going to change over time:

$$y_t = c + \alpha t + \varepsilon_t$$

- So far we have just differenced to remove trends (remove y_{t-1} from both sides):
- Sometimes not the best way to proceed:

$$\Delta y_t = \alpha + \varepsilon_t - \varepsilon_{t-1}$$

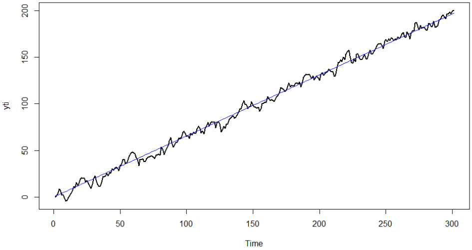
  - This is a MA(1) with $\theta_1 = -1$.
  - This is now a not invertible process!

Deterministic trends

- Better solution:
  - Fit a regression model modelling the trend component of the data and then remove the trend, i.e.,:
  - Run OLS of y_t on a constant and (a polynomial of) t and subtract the predictions from the original time series.
  - or: Include t as explanatory variable in `Arima`.

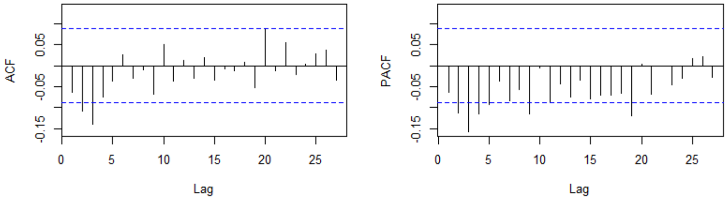
Overdifferencing Example

Example AR(1) with trend: $y_t = 0.7y_{t-1} + \varepsilon_t + 0.3 \cdot t$



Overdifferencing Example

Wrongly estimating ARIMA(1,1,0) yields $\hat{\phi} = -0.1494$. Residuals:



One solution (linear trend):

- `Arima(yt,c(1,0,0),xreg = 1:T)`
- `xreg = 1:T`: This includes an exogenous regressor (`xreg`) in the model. The exogenous regressor is specified as a trend, where `1:T` represents a sequence from 1 to T (the length of your time series).

```
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1    xreg
      0.7116  0.9940
s.e.  0.0248  0.0011
```

Stochastic Trend

- Trend that has a stochastic (= random) nature
- Stochastic trends can change, and the estimated growth is only assumed to be the average growth over the historical period, not necessarily the rate of growth that will be observed into the future
- Most important example is the random walk:

$$y_t = y_{t-1} + \varepsilon_t \qquad \text{means} \qquad \Delta y_t = \varepsilon_t$$

- similar: random walk model increases in each period by a given amount
- but: increase is stochastic and given by concrete realization

Random Walk

- Recall: Random Walk is an AR(1) with a unit root ($\phi = 1$).
- it is also the sum of all historical errors (shocks do not die out)

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i$$

- The mean is $E[y_t] = y_0 + 0 + 0 + \cdots = y_0$ fixed.
- The variance $Var[y_t] = t\sigma^2$ increases, a random walk is therefore non-stationary.
- First-difference will be a stationary process (special case: white noise).

Deterministic + Stochastic Trend

Random walk with drift

$$y_t = c + y_{t-1} + \varepsilon_t$$

- If $\phi = 1$, the constant gets added to each observation and also does not 'die out'.

$$y_t = y_0 + t \cdot c + \sum_{i=1}^{t} \varepsilon_i$$

- The first difference has a stochastic and a deterministic component:

$$\Delta y_t = c + \varepsilon_t$$

- Differencing is therefore good here (constant + white noise).

Summary: Trends

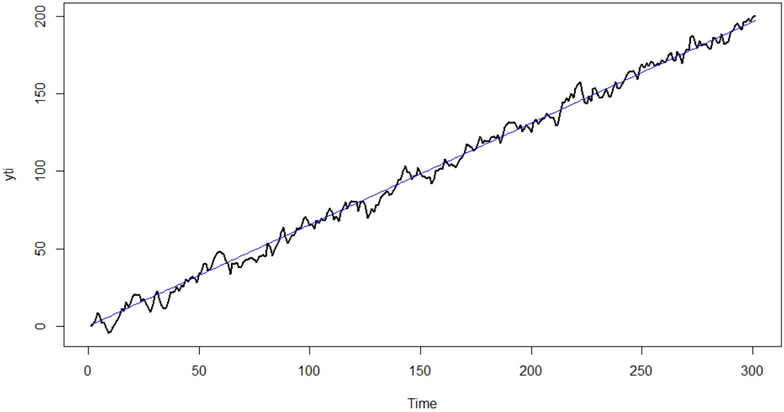
- Time series can contain deterministic, stochastic or both trends
- Trend makes a time series non-stationary
- Stochastic trends require differencing, deterministic require modelling (to remove or control for)

Definitions:

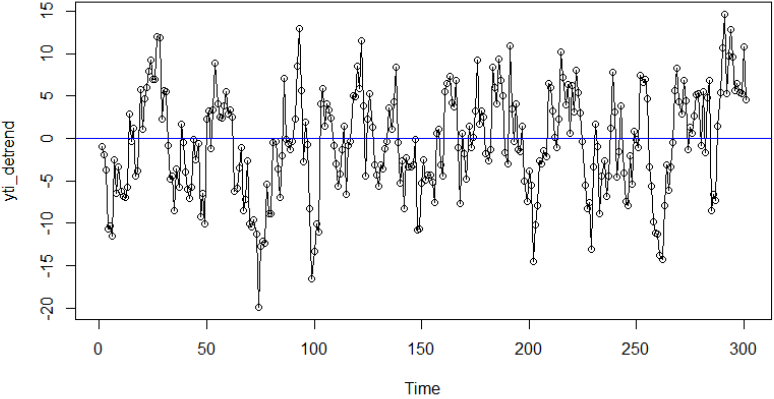
- If a time series is stationary around a trend, we call it **trend-stationary**
- If a time series is stationary after differencing, we call it **difference-stationary**, or integrated, for example I(1), as in ARIMA

Trend-stationary

AR(1) with trend: $y_t = 0.7y_{t-1} + \varepsilon_t + 0.3 \cdot t$

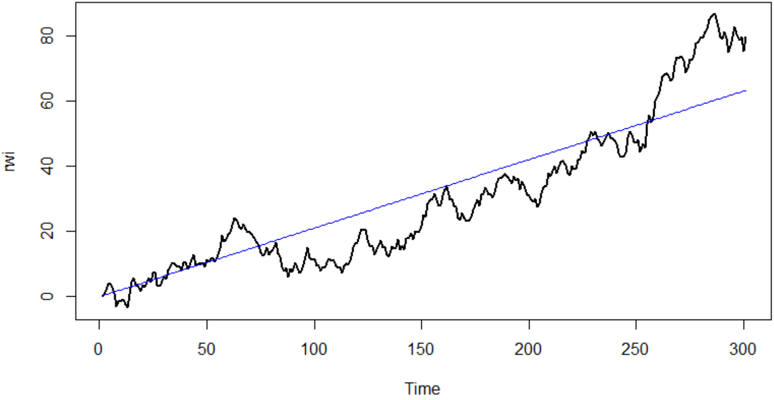


Trend-stationary: after removing linear trend

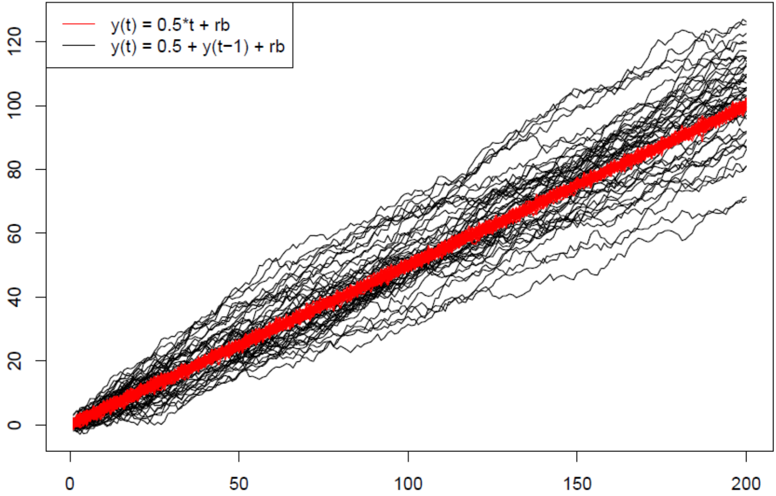


Random Walk with Drift

Random walk with drift: $y_t = 0.3 + y_{t-1} + \varepsilon_t$



Simulating Stochastic vs. Deterministic Trend



Intuition: Real GDP

Real GDP is trend stationary:

- (most) shocks to GDP are transitory, and GDP reverts to an already existing trend.

Real GDP is a random walk with drift:

- At least a portion of each time period's shock to GDP is permanent. Booms and recessions therefore permanently change the future path.

Unit-Root/Stationarity Tests

- There are different tests to check whether a process has a unit root/it is stationary
- The null hypothesis varies across tests

Dickey-Fuller Test

$$y_t = \phi_1 y_{t-1} + \varepsilon$$

- stationary $|\phi_1| < 1$, unit root if $\phi_1 = 1$
- Estimate with OLS, and use a one-sided t-test for $H_0 : \phi_1 = 1$
- The test statistic $T \cdot (\hat{\rho} - 1)$ is called the Dickey-Fuller, or DF, ρ statistic.
- Or DF t-statistics: $\frac{\hat{\rho}-1}{se_{\hat{\rho}}}$
- We will use the Augmented Dickey-Fuller Test (ADF), that works for AR(p)

Augmented Dickey-Fuller (ADF)

- The testing procedure for the ADF test is the same as for the Dickey–Fuller test but it is applied to the model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

- α is a constant, β the coefficient on a time trend, and p the lag order of the autoregressive process.
- Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modeling a random walk, and using the constraint $\beta = 0$ corresponds to modeling a random walk with a drift.
- Consequently, there are three main versions of the test.
  - no drift/constant, no trend
  - constant, no trend
  - both drift and trend

Augmented Dickey-Fuller (ADF)

- It includes lagged differences to control for autocorrelation
- By including lags of the order p, the ADF formulation allows for higher-order autoregressive processes.
- `aTSA::adf.test`

Phillips-Perron-Test

- Built on ADF test, with additional care for autocorrelation and heteroskedasticity of residuals
- same H_0: unit root
- `aTSA::pp.test`

KPSS test

- Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test
- technically completely different from ADF/PP tests
- most important: H_0 is trend/level-stationarity against alternative of a unit root
- `tseries::kpss.test(,.null = "T")` for trend-stationarity
- `tseries::kpss.test(,.null = "L")` for level-stationarity

Seasonality

Seasonal patterns:

- e.g. number of tourists, ski sales
- also: data collection or accounting practices
- at sub-yearly intervals (monthly, quarterly, ...)

1. Detect seasonality
2. Remove seasonality *or* estimate jointly with irregular pattern (SARIMA)

Remove Seasonality

- In practice, many time series are already *seasonally adjusted* with filters
- one simple way of removing seasonal patterns yourself: seasonal differencing
  - e.g. $\Delta_{12} y_t = y_t - y_{t-12}$
  - seasonal differencing with s will yield the time series with the lowest variance if s is the true seasonal pattern
  - in R: `diff(y,12)`
- usually removes most of the trend too
- but can be combined with 'regular' differencing if needed
- $\Delta\Delta_s y_t = (y_t - y_{t-s}) - (y_{t-1} - y_{t-s-1})$

Recognizing and Estimating seasonal pattern

Find seasonal patterns:

- intuition
- plot the time series
- plot data by season
- look at significant spikes in your ACF, PACFs

Remove or include:

- Similar to trends, it can be useful to incorporate the seasonality instead of removing it (e.g. forecasting business' daily sales)

Seasonal Arima (SARIMA)

ARIMA $(p, d, q)(P, D, Q)_s$

- Extension of ARIMA to seasonal data with season s
- can accommodate seasonal differencing of order D as well as seasonal AR(P) and MA(Q) terms
- identify P, Q similar to non-seasonal order but restrict attention to the seasonal lags.
- $d + D$ should never exceed 2 for almost all cases
- in R: `Arima(yt, order = c(p,d,q), seasonal = c(P,D,Q))`

Example Seasonal Lag Order

- ARIMA$(0,0,0)(1,0,0)_{12}$ will show:
  - exponential decay in the seasonal lags of the ACF (12, 24, ...)
  - a single significant spike at lag 12 in the PACF

- ARIMA$(0,0,0)(0,0,1)_{12}$ will show:
  - a single significant spike at lag 12 in the ACF
  - exponential decay in the seasonal lags of the PACF (12, 24, ...)
- Intuition: including both seasonal and non-seasonal terms allows for example last winter (Dec 20) as well as previous months (Nov 21) to affect this winter's sales (Dec 21).

Exercise I

- Load the *hotel.csv* dataset. It contains monthly hospitality employment (million) in the US from 2001.
- Discuss the stationarity of the time series.
- If needed, use seasonal differencing. Is the resulting time series stationary or not? Check with unit-root/stationarity tests.
- Apply the seasonal Box-Jenkins approach to this dataset.