Quantitative Social Research II

Workshop 5: Time-Series

Jose Pina-Sánchez

Workshop Aims

- Introduce key concepts used to describe time-series
  - trend, seasonality, irregular component
  - stationarity
  - auto-correlation and partial auto-correlation
- Introduce ARIMA models
  - autoregressive (AR)
  - moving average (MA)
  - autoregressive, integration, moving average (ARIMA)
- Learn to forecast and estimate the causal impact of discrete interventions

What Are Time-Series?

- Any metric that is measured over regular time intervals makes a time series
- Share the time dimension that we will see in longitudinal data
  - but the focus is on one (or a few) subjects/entities
  - across a much larger timespan
  - the focus tends to be on forecasting based on observed past patterns
  - the modelling strategies are very different

What Are Time-Series?

- Any metric that is measured over regular time intervals makes a time series
- Share the time dimension that we will see in longitudinal data
  - but the focus is on one (or a few) subjects/entities
  - across a much larger timespan
  - the focus tends to be on forecasting based on observed past patterns
  - the modelling strategies are very different
- Predominantly employed in Economics and by business analysts
  - to forecasts stock prices, changes in GDP
  - to understand economic cycles, or seasonal sales of certain products
- Although less frequent, also used in Sociology and Criminology
  - estimate the deterrent effect of new (more punitive) laws
  - estimate the extent to which public demonstrations are seasonally determined

Components of a Time-Series

- A time series can be broken down into its main components
  - this helps understand, analyse, and model it
  - and carry out forecasts

Components of a Time-Series

- A time series can be broken down into its main components
  - this helps understand, analyse, and model it
  - and carry out forecasts
- Each data point (Y_t) at time t in a time series can be expressed as the combination of
  - trend (T_t): the general tendency of a time series to increase, decrease or stagnate over a long period of time
  - seasonality (S_t): fluctuations within a regular period of time; Ciclicality (C_t) can also be included to capture wider fluctuations
  - irregular component (I_t): random variations, do not show a particular pattern, unpredictable

Components of a Time-Series

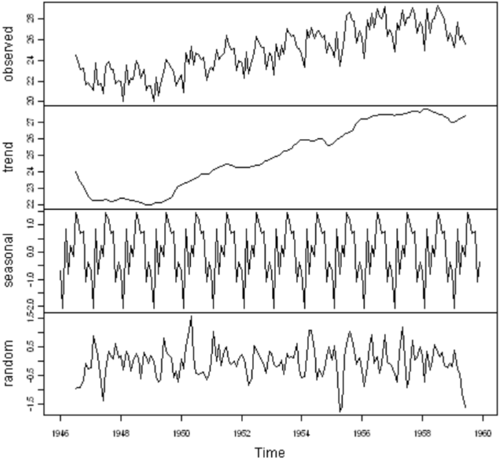
- A time series can be broken down into its main components
  - this helps understand, analyse, and model it
  - and carry out forecasts
- Each data point (Y_t) at time t in a time series can be expressed as the combination of
  - trend (T_t): the general tendency of a time series to increase, decrease or stagnate over a long period of time
  - seasonality (S_t): fluctuations within a regular period of time; Ciclicality (C_t) can also be included to capture wider fluctuations
  - irregular component (I_t): random variations, do not show a particular pattern, unpredictable
- Combined as either a sum or a product
  - additive model: $Y_t = T_t + S_t + I_t$

    assumes components are independent of each other
  - multiplicative model: $Y_t = T_t \times S_t \times I_t$

    assumes components can affect one another

Decomposition of Time-Series

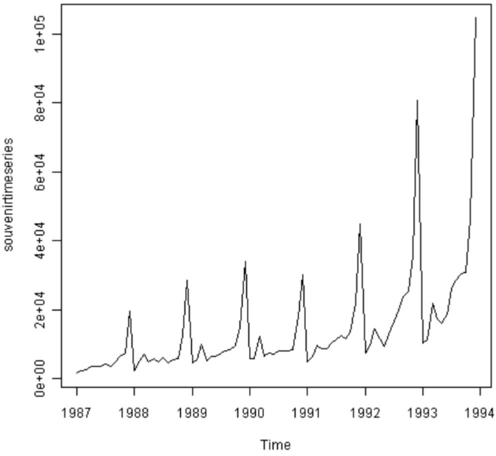
New York Birth Time-Series (decomposed)



Source: Avril Coghlan

Multiplicative Model

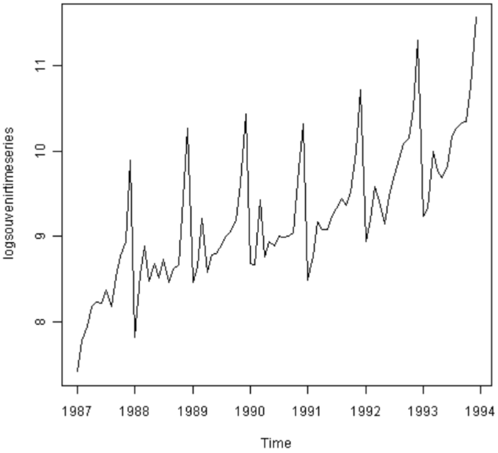
Monthly sales for a souvenir shop at a beach resort town in Queensland



Source: Avril Coghlan

Additive Model

Log of monthly sales for a souvenir shop at a beach resort town in Queensland



Source: Avril Coghlan

Stationarity

- A time-series is said to be stationary if it holds the following conditions true:
  - the mean value of time-series is constant over time, i.e. the trend component is nullified
  - the variance does not increase over time
  - seasonality effect is minimal
- It looks like random white noise irrespective of the observed time interval

Stationarity

- A time-series is said to be stationary if it holds the following conditions true:
  - the mean value of time-series is constant over time, i.e. the trend component is nullified
  - the variance does not increase over time
  - seasonality effect is minimal
- It looks like random white noise irrespective of the observed time interval
- In essence, stationarity means that the statistical properties of a process generating a time series do not change over time
  - to be able to model time-series we are going to seek to turn them stationary

ACF and PACF

- To help us decide the most appropriate model we are going to rely on two useful diagnostic functions
- Auto-correlation function (ACF)
  - gives us values of auto-correlation of any series with its lagged values
  - it describes how well the present value of the series is related to its past values
  - it can be used to inform the number of lags to be included in the model
  - ACF considers all the components in a time-series (T, S, C and I) in the calculation of its correlations
  - that is why it is know as a 'complete auto-correlation plot'

ACF and PACF

- To help us decide the most appropriate model we are going to rely on two useful diagnostic functions

- Auto-correlation function (ACF)
  - gives us values of auto-correlation of any series with its lagged values
  - it describes how well the present value of the series is related to its past values
  - it can be used to inform the number of lags to be included in the model
  - ACF considers all the components in a time-series (T, S, C and I) in the calculation of its correlations
  - that is why it is know as a 'complete auto-correlation plot'

- Partial auto-correlation function (PACF)
  - instead of finding correlations of present values with lags, PACF finds correlation of the residuals (remaining after removing the effects explained by the earlier lags) with the next lag value
  - 'partial' because we remove already found variations before calculating the correlation
  - can help us inform the model if there is any 'hidden information' in the residual which can be modelled by the next lag

ARIMA Models

- The most common approach to model time-series
- Composed of an autoregressive (AR) part
  - Associated with the ACF
- And a moving average (MA) part
  - associated with the PACF

ARIMA Models

- The most common approach to model time-series
- Composed of an autoregressive (AR) part
  - Associated with the ACF
- And a moving average (MA) part
  - associated with the PACF
- Requires the time-serie(s) to be stationary
  - to do so we can use differencing, aka integration (I)
  - which stands for the I in ARIMA

Differencing

- By subtracting each data point in the series from its successor we can often turn a non-stationary time series stationary
  - $Y_t^d = Y_t - Y_{t-1}$
  - useful to remove trends and cycles

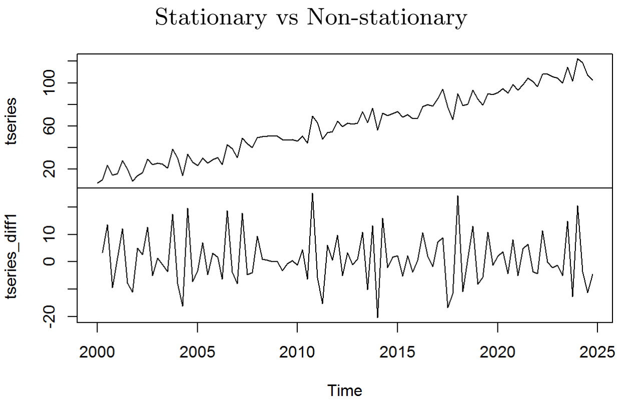
Differencing

- By subtracting each data point in the series from its successor we can often turn a non-stationary time series stationary
  - $Y_t^d = Y_t - Y_{t-1}$
  - useful to remove trends and cycles
- Sometimes higher order differences are necessary to achieve stationarity
  - often a second order difference is enough
  - $Y_t^{d2} = Y_t^d - Y_{t-1}^d = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$

Differencing

- By subtracting each data point in the series from its successor we can often turn a non-stationary time series stationary
  - $Y_t^d = Y_t - Y_{t-1}$
  - useful to remove trends and cycles
- Sometimes higher order differences are necessary to achieve stationarity
  - often a second order difference is enough
  - $Y_t^{d2} = Y_t^d - Y_{t-1}^d = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$
- If the time series appears to be seasonal, a better approach is to difference with respective season's data points
  - $Y_t^d = (Y_t - Y_{t-s}) - (Y_{t-1} - Y_{t-s-1})$
  - this can help to remove the seasonal effect

Differencing



Source: Troy Walters

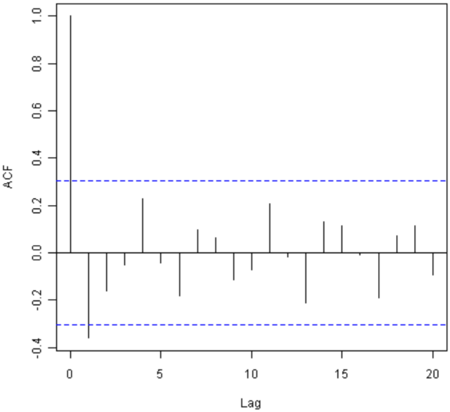
Auto-Regressive Models

- AR are linear models where the outcome variable is regressed on its own lagged values
  - $Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t$
  - if a lag up to p is included in the model, the AR process is said to be of order p

Auto-Regressive Models

- AR are linear models where the outcome variable is regressed on its own lagged values
  - $Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t$
  - if a lag up to p is included in the model, the AR process is said to be of order p
- To decide the order of the AR the model we can use the ACF
  - which plots the level of auto-correlation at each lag
  - and the 95% confidence interval to determine their statistical significance

Auto-Correlation Function

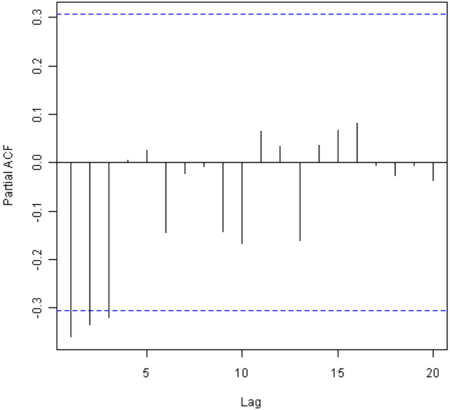


Source: Avril Coghlan

Moving Average Models

- MA are also linear models where the outcome variables is regressed on its own imperfectly predicted lagged values
  - $Y_t = \theta_0 + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} + e_t$
  - if a lag up to q is included in the model, the MA process is said to be of order q
- To decide the order of the MA the model we can use the PACF
  - which plots the level of auto-correlation at each lag
  - and the 95% confidence interval to determine their statistical significance

Partial Auto-Correlation Function



Source: Avril Coghlan

ARIMA Models

- An ARIMA(p, d, q) model is then defined as the combination of
  - d integrations (differentiations)
  - an AR model of order p
  - and a MA model of order q
- Once the series is turn stationary following the integration process, we have an ARMA(p, q) model
  - $Y_t = \delta + \{\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p}\} + \{\theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q}\} + e_t$
  - this is the model to be estimated
  - which can be used use to make predictions

Predictions

- Once we have estimated our ARIMA model we can use it to make predictions
  - let's assume a simple ARMA(1,1) model
  - $Y_t = \delta + \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$

Predictions

- Once we have estimated our ARIMA model we can use it to make predictions
  - let's assume a simple ARMA(1,1) model
  - $Y_t = \delta + \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$
- We can estimate future values of Y
  - this is done sequentially

Predictions

- Once we have estimated our ARIMA model we can use it to make predictions
  - let's assume a simple ARMA(1,1) model
  - $Y_t = \delta + \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$
- We can estimate future values of Y
  - this is done sequentially
  - we estimate the value for the next period (Y_{t+1}) as
    $\hat{Y}_{t+1} = \delta + \hat{\phi}_1 Y_t + \hat{\theta}_1 e_t$
    and then use that value to estimate the next period (Y_{t+2})
    $\hat{Y}_{t+2} = \delta + \hat{\phi}_1 \hat{Y}_{t+1} + \hat{\theta}_1 \hat{e}_{t+1}$
    and so on

Predictions

- Once we have estimated our ARIMA model we can use it to make predictions
  - let's assume a simple ARMA(1,1) model
  - $Y_t = \delta + \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$
- We can estimate future values of Y
  - this is done sequentially
  - we estimate the value for the next period (Y_{t+1}) as

    $\hat{Y}_{t+1} = \delta + \hat{\phi}_1 Y_t + \hat{\theta}_1 e_t$

    and then use that value to estimate the next period (Y_{t+2})

    $\hat{Y}_{t+2} = \delta + \hat{\phi}_1 \hat{Y}_{t+1} + \hat{\theta}_1 \hat{e}_{t+1}$

    and so on
  - the uncertainty of our predictions will grow as we move into the future, away from t

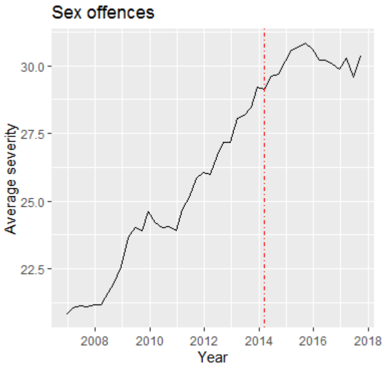
Detecting Causal Effects

- We can also use time-series analysis to detect the causal effect of discrete interventions
  - i.e. policies/events that take place at a specific date
  - we can assess whether the time-series changes its properties after the intervention took place
  - ex.1: Did the 2018 minimum wage increase in Spain had an impact on the unemployment rate?
  - ex.2: Did the new sentencing guidelines increased sentence severity in England and Wales?
  - often referred as interrupted time-series models

Detecting Causal Effects

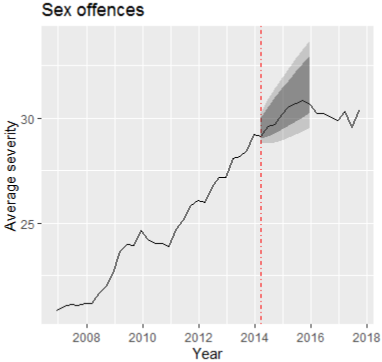
- We can also use time-series analysis to detect the causal effect of discrete interventions
  - i.e. policies/events that take place at a specific date
  - we can assess whether the time-series changes its properties after the intervention took place
  - ex.1: Did the 2018 minimum wage increase in Spain had an impact on the unemployment rate?
  - ex.2: Did the new sentencing guidelines increased sentence severity in England and Wales?
  - often referred as interrupted time-series models
- For these types of analyses we divide the time-series in two parts
  - we model the start of the time-series up to the last time point before the intervention took place
  - based in that model, we predict values for time periods following the intervention
  - then we compare the predicted against the observed values

Detecting Causal Effects



Source: Pina-Sánchez et al. 2019

Detecting Causal Effects



Source: Pina-Sánchez et al. 2019

Recap

- We have learnt about...
  - the main components defining time-series (T, S, C, and I)
  - the key statistics and properties to be considered in modelling time-series (stationarity, ACF, and PACF)
  - the main family of models for the analysis of time-series (ARIMA)

Recap

- We have learnt about...
  - the main components defining time-series (T, S, C, and I)
  - the key statistics and properties to be considered in modelling time-series (stationarity, ACF, and PACF)
  - the main family of models for the analysis of time-series (ARIMA)
- There is so much we have not covered though
  - exponential-smoothing methods
  - properly modelling seasonal and cyclical effects
  - and much more

Recap

- We have learnt about...
  - the main components defining time-series (T, S, C, and I)
  - the key statistics and properties to be considered in modelling time-series (stationarity, ACF, and PACF)
  - the main family of models for the analysis of time-series (ARIMA)
- There is so much we have not covered though
  - exponential-smoothing methods
  - properly modelling seasonal and cyclical effects
  - and much more
- Recommended readings
  - lots of free tutorials, short courses and handbooks covering time-series online
  - Hanck et al. (2019) Chapter 14 'Introduction to Time Series Regression and Forecasting'