MONASH University

MONASH BUSINESS SCHOOL

ETF3231/5231: Business forecasting

Week 2: Time series graphics

`https://bf.numbat.space/`

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

Outline

Outline

Included in week 1:

- `tsibble` objects
- The `tsibble` index

Show olympic_running

`dplyr` funtions

- `filter`: choose rows
- `select`: choose columns
- `mutate`: make new columns
- `group_by`: group rows
- `summarise`: summarise across groups
- `reframe`: summarise multiple rows across groups

You will practice these in your tutorials this week

Outline

Time series graphics

- Time plots: `autoplot()`
- Seasonal plots: `gg_season()`
- Seasonal subseries plots: `gg_subseries()`
- Lag plots: `gg_lag()`
- ACF plots: `ACF() |> autoplot()`

These are the tools you will use. Each provides a different view of your data.

Time plots

- First in any modelling/forecasting task should be to plot your data.
- Plots allow us to identify:
  - Patterns;
  - Unusual observations;
  - Changes over time;
  - Relationships between variables.

Time plots

- First in any modelling/forecasting task should be to plot your data.
- Plots allow us to identify:
  - Patterns;
  - Unusual observations;
  - Changes over time;
  - Relationships between variables.

Patterns:
- trend
- seasonal
- cycles

Outline

Time series patterns

Trend pattern exists when there is a long-term increase or decrease in the data.

Seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).

Cyclic pattern exists when data exhibit rises and falls that are not of fixed period (duration usually of at least 2 years).

Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.

Switch to R

Outline

Seasonal plots

- Data plotted against the individual "seasons" in which the data were observed. (In this case a “season” is a month.)
- Something like a time plot except that the data from each season are overlapped.
- Enables the underlying seasonal pattern to be seen more clearly, and also allows any substantial departures from the seasonal pattern to be easily identified.
- In R: `gg_season()`

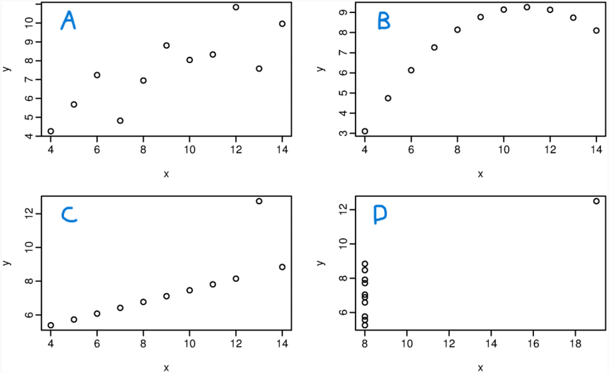
Seasonal subseries plots

- Data for each season collected together in time plot as separate time series.
- Enables the underlying seasonal pattern to be seen clearly, and changes in seasonality over time to be visualized.
- In R: `gg_subseries()`

Outline

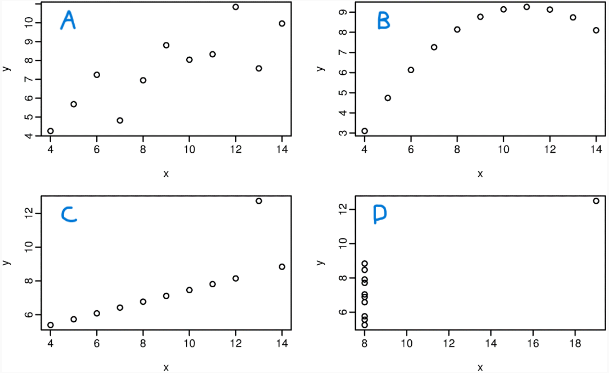
Correlation coefficient

- Which one has the highest correlation?



Correlation coefficient

- Which one has the highest correlation?



All these have r = 0.82. Hence importance of plots.

Autocorrelation

Autocovariance (c_k) and autocorrelation (r_k): measure linear relationship between lagged values of a time series y.

Autocorrelation

Autocovariance (c_k) and autocorrelation (r_k): measure linear relationship between lagged values of a time series y.

We measure the relationship between:

- y_t and y_{t-1}
- y_t and y_{t-2}
- y_t and y_{t-3}
- $\dots$
- y_t and y_{t-k}
- etc.

Autocorrelation

We denote the sample autocovariance at lag k by c_k and the sample autocorrelation at lag k by r_k. Then define

$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y}) / (T-1)}{\sum_{t=1}^{T}(y_t - \bar{y}) / (T-1)} = \frac{cov(y_t, y_{t-k})}{var(y_t)}$$

- r_1 indicates how successive values of y relate to each other
- r_2 indicates how y values two periods apart relate to each other
- r_k is *almost* the same as the sample correlation between y_t and y_{t-k}.

Trend and seasonality in ACF plots

- When data have a trend, the autocorrelations for small lags tend to be large and positive.
- When data are seasonal, the autocorrelations will be larger at the seasonal lags (i.e., at multiples of the seasonal frequency)
- When data are trended and seasonal, you see a combination of these effects.

Switch to R

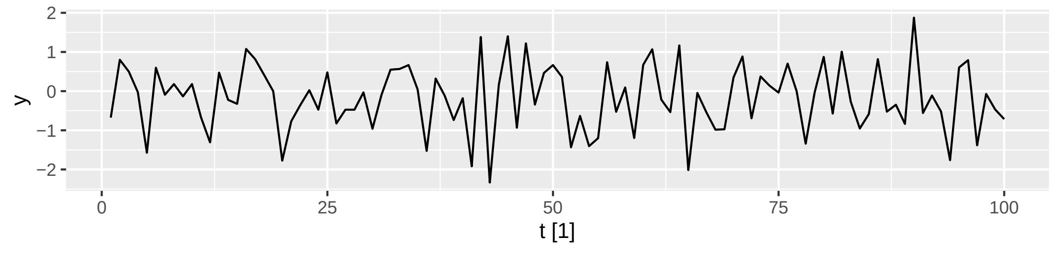
Outline

White noise

White noise data consists of purely random draws from the same distribution with mean zero and constant variance.

$$y_t = \varepsilon_t, \qquad \text{where } \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

```r
my_data <- tsibble(t = seq(100), y = rnorm(100), index = t)
my_data |> autoplot(y)
```

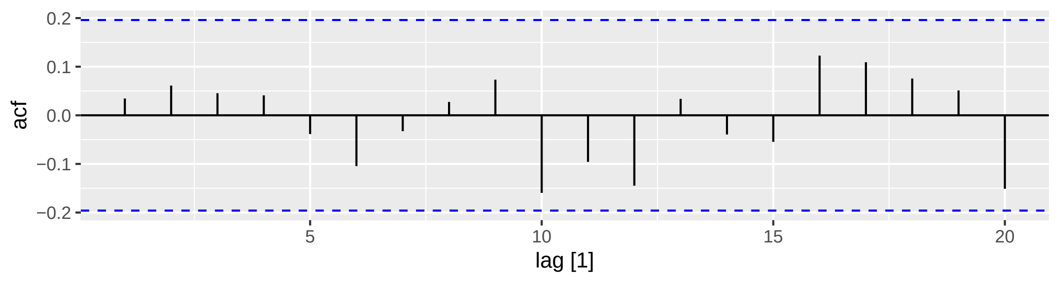


White noise

White noise data consists of purely random draws from the same distribution with mean zero and constant variance.

$$y_t = \varepsilon_t, \qquad \text{where } \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

```
my_data |> ACF(y) |> autoplot()
```



Sampling distribution of WN autocorrelations

Sampling distribution of r_k for white noise data is asymptotically N$(0,1/T)$.

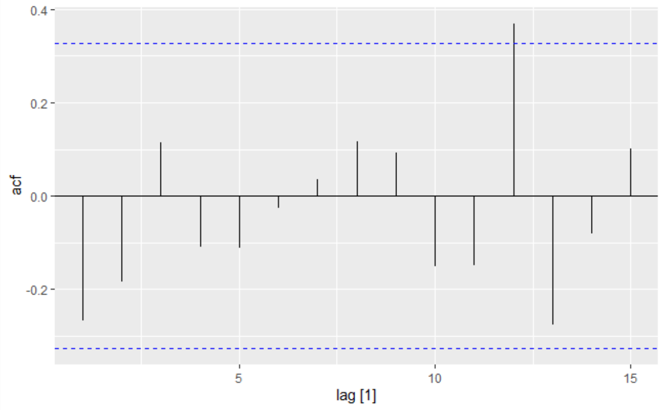
- 95% of all r_k for white noise must lie within $\pm 1.96/\sqrt{T}$.
- If this is not the case, the series is probably not WN.
- Common to plot lines at $\pm 1.96/\sqrt{T}$ when plotting ACF. These are the **critical values**.

Example: WN autocorrelation

Example:

$T = 36$ and so critical values at $\pm 1.96/\sqrt{36} = \pm 0.327$.

All autocorrelations lie within these limits, confirming that the data are white noise. (More precisely, the data cannot be distinguished from white noise.)



Note: 5% chance to be outside the critical values (Type I error). You want to see spikes a long way out or many of them. Don't get too excited for 1 just outside especially at longer lags.

Random walks

Random walks are a type of time series where the value at time t is equal to the previous value plus a random amount from a white noise process.

$$y_t = y_{t-1} + \varepsilon_t, \qquad \text{where } \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Equivalently, we can take the cumulative sum of a white noise process.

$$y_t = y_0 + \sum_{t=1}^{T} \varepsilon_t, \qquad \text{where } \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

```r
my_data <- tsibble(t = seq(100), y = cumsum(rnorm(100)), index = t)
```

$$y_t = y_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2) \qquad t = 1, \dots, T$$

$y_t = y_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2) \qquad t = 1, \ldots, T$

$t = 1 \qquad y_1 = y_0 + \varepsilon_1$

$$y_t = y_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2) \qquad t = 1, \ldots, T$$

$t = 1 \qquad y_1 = y_0 + \varepsilon_1$

$t = 2 \qquad y_2 = y_1 + \varepsilon_2 = y_0 + \varepsilon_1 + \varepsilon_2$

$$y_t = y_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2) \qquad t = 1, \dots, T$$

$t = 1 \qquad y_1 = y_0 + \varepsilon_1$

$t = 2 \qquad y_2 = y_1 + \varepsilon_2 = y_0 + \varepsilon_1 + \varepsilon_2$

$t = 3 \qquad y_3 = y_2 + \varepsilon_3 = y_0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3$

$$y_t = y_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2) \qquad t = 1, \dots, T$$

$t=1$ $\quad y_1 = y_0 + \varepsilon_1$

$t=2$ $\quad y_2 = y_1 + \varepsilon_2 = y_0 + \varepsilon_1 + \varepsilon_2$

$t=3$ $\quad y_3 = y_2 + \varepsilon_3 = y_0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3$

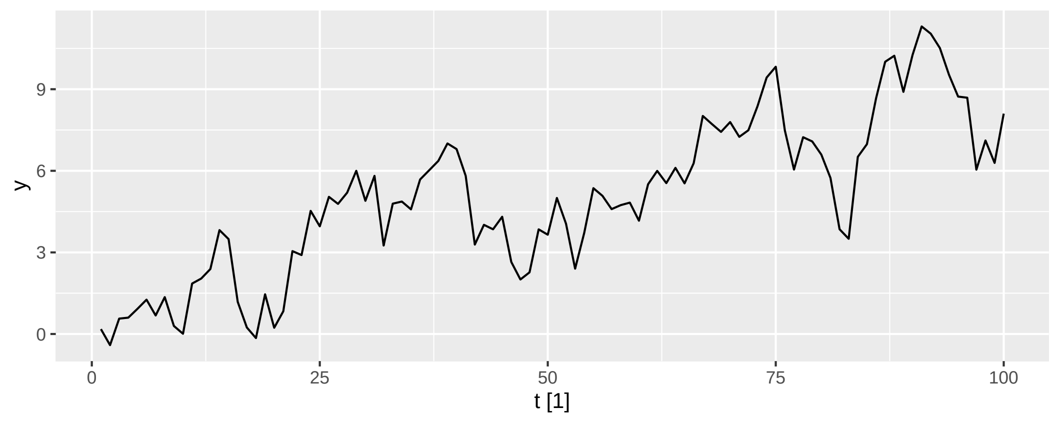
$\vdots$

$t=T$ $\quad y_T = y_{T-1} + \varepsilon_T = y_0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_T$

$$= y_0 + \sum_{t=1}^{T} \varepsilon_t$$

Random walks

```r
my_data |> autoplot(y)
```



Random walks

```
my_data |> ACF(y) |> autoplot()
```