ARIMA models: Introduction

EC 361–001

Prof. Santetti
Spring 2024

Materials

Required readings:

- `Hyndman & Athanasopoulos, ch. 9`
  - sections 9.1—9.4.

Motivation

Motivation

Along with *exponential smoothing*, **ARIMA models** are the most **widely used** approach for time series forecasting.

That said, these two techniques should not be seen as **competing**, but as **complementary** to each other.

One **key distinction** between these two approaches is that, while exponential smoothing focuses on time series **features** (*error*, *trend*, and *seasonality*), ARIMA models focus on the presence of **autocorrelation** in the data.

We will start studying ARIMA models by looking at time series **stationarity**.

Stationarity

Stationarity

In short:

> *A **stationary** time series is one whose statistical properties **do not** depend on the time at which the series is observed.*

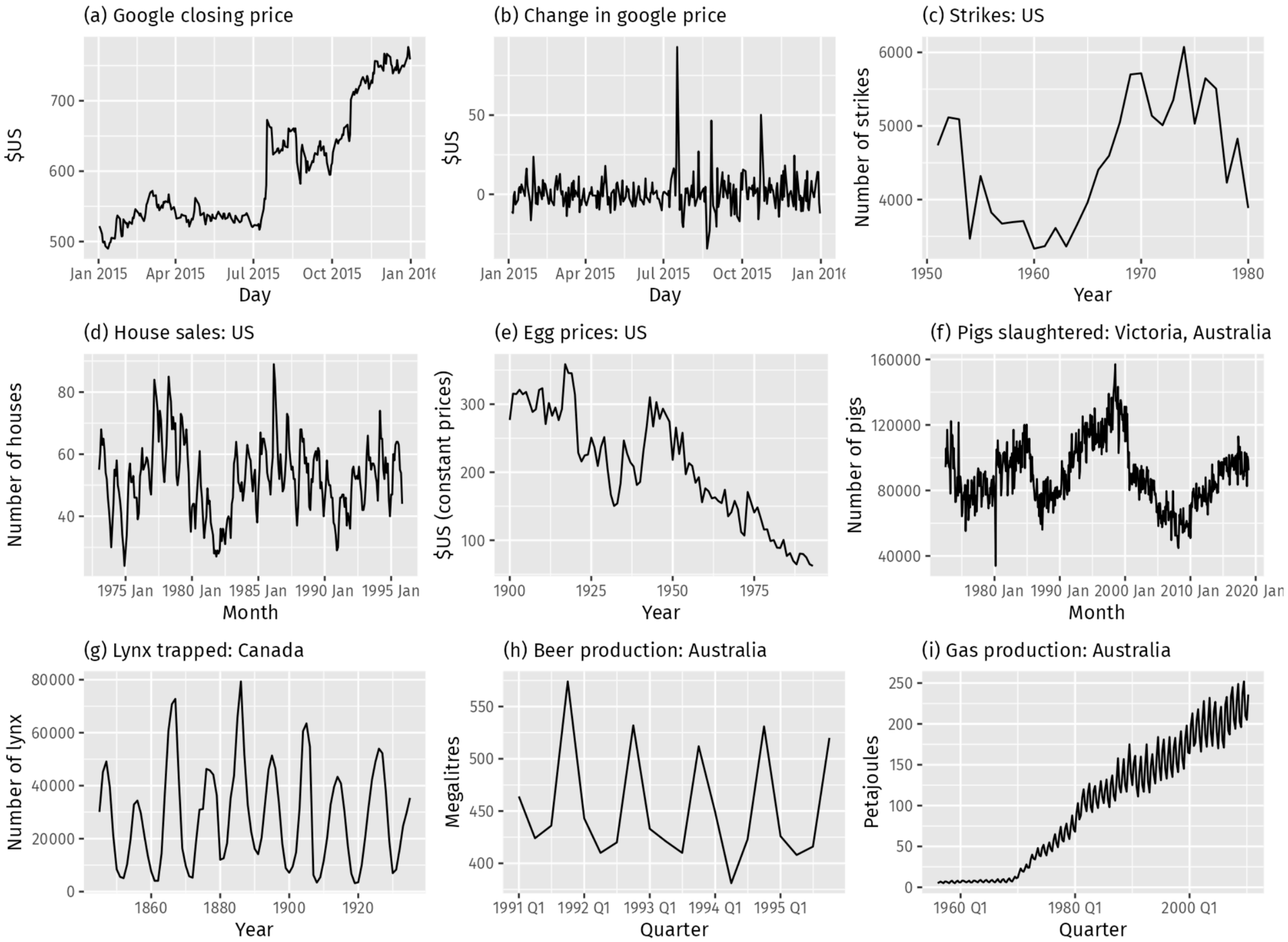
In other words, a time series that contains relevant features—such as *trend* and *seasonality*—will not be **stationary**, since its behavior changes over time.

On the other hand, a time series following a **white noise** process is stationary.

- Whenever you observe the series, it will look pretty much the same.

In **visual** terms, the *time plot* of a stationary time series will be roughly **horizontal** (which can include *cyclic* behavior) and with a *constant variance*.

Stationarity



Stationarity

Recall that one **key** characteristic of a stationary time series is **constant variance**.

A few weeks ago, we learned that **log-transformations** help to stabilize a series' variance over time.

But in order for a time series to be stationary, its **mean** should also be stable over time (i.e., with no trend).

A technique that helps to stabilize the **mean** of a time series is known as **differencing**.

A **differenced series** is the *change between consecutive observations* in the original series, and can be written as

$$y'_t = y_t - y_{t-1}$$

Stationarity

When the differenced series is **white noise**, the model for the original series can be written as:

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

where ε_t denotes a white noise process (i.e., with no autocorrelation).

Rearranging the above equation, we end up with

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

which, unsurprisingly, is a **random walk process**.

Stationarity

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

Random walk models are widely used for **non-stationary** data, as they typically show:

- long periods of apparent *trends* up or down;
- *sudden* and *unpredictable* changes in direction.

As future movements in a random walk process are **unpredictable**, forecasts from these processes are usually equal to the **last observation**.

- This fact generates the **naïve forecast method**.

Stationarity

Whenever the **differences** between y_t and y_{t-1} are, on average, different from zero, one can account for this in the following way:

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

where c accounts for the possibility of y_t **drifting** *upwards* or *downwards*.

This model is behind the **drift forecast method**.

Stationarity

When a time series shows **seasonality**, one option is to work with **seasonal differencing**.

What this implies is that we can take the difference between an observation y_t and the *previous* observation from the **same season**:

$$y'_t = y_t - y_{t-m}$$

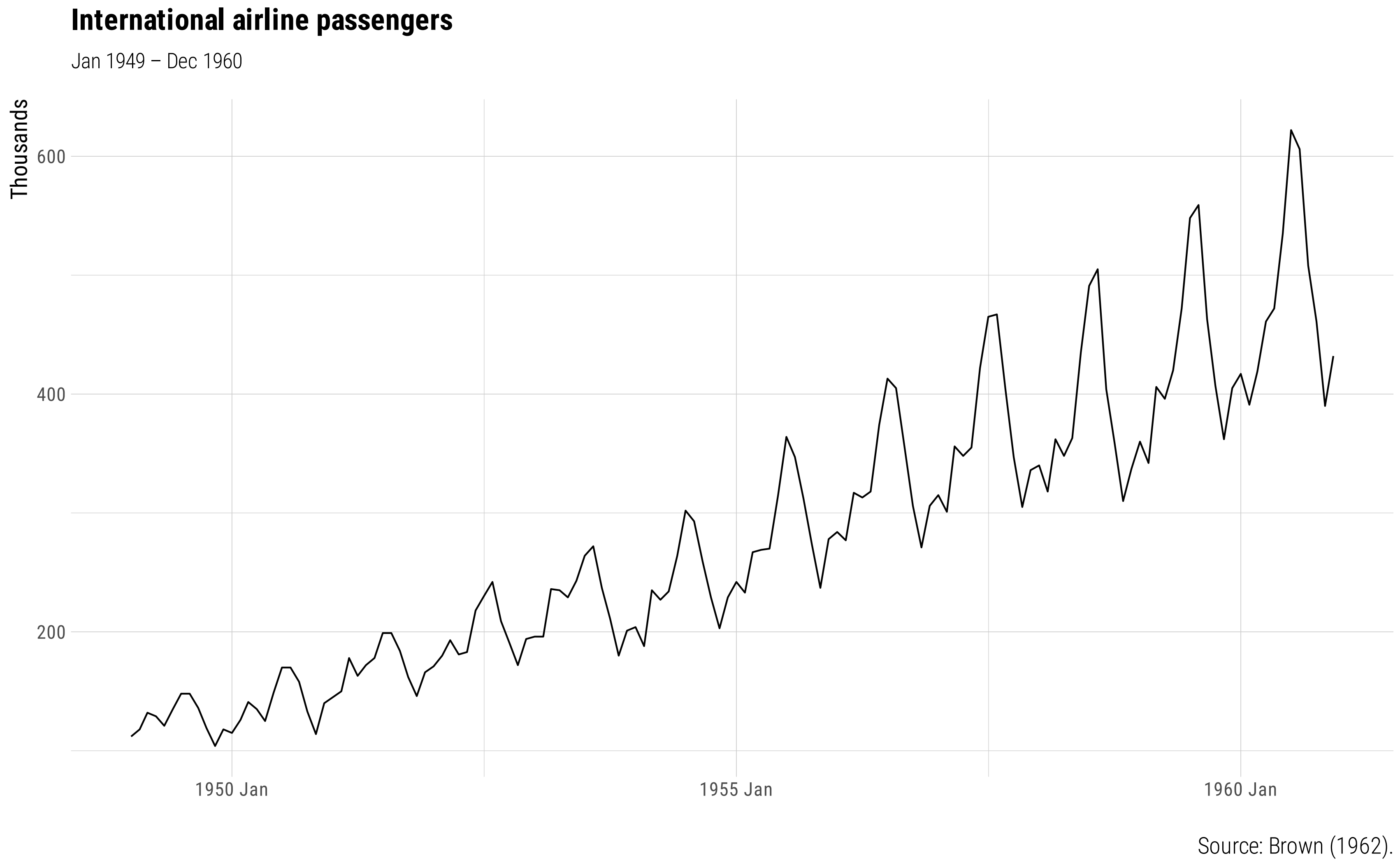
where m denotes the number of seasonal periods.

In case the differenced data from seasonal differencing is **white noise**, then one can write:

$$y_t = y_{t-m} + \varepsilon_t$$

And this model gives origin to the **seasonal naïve forecast method**.

Stationarity

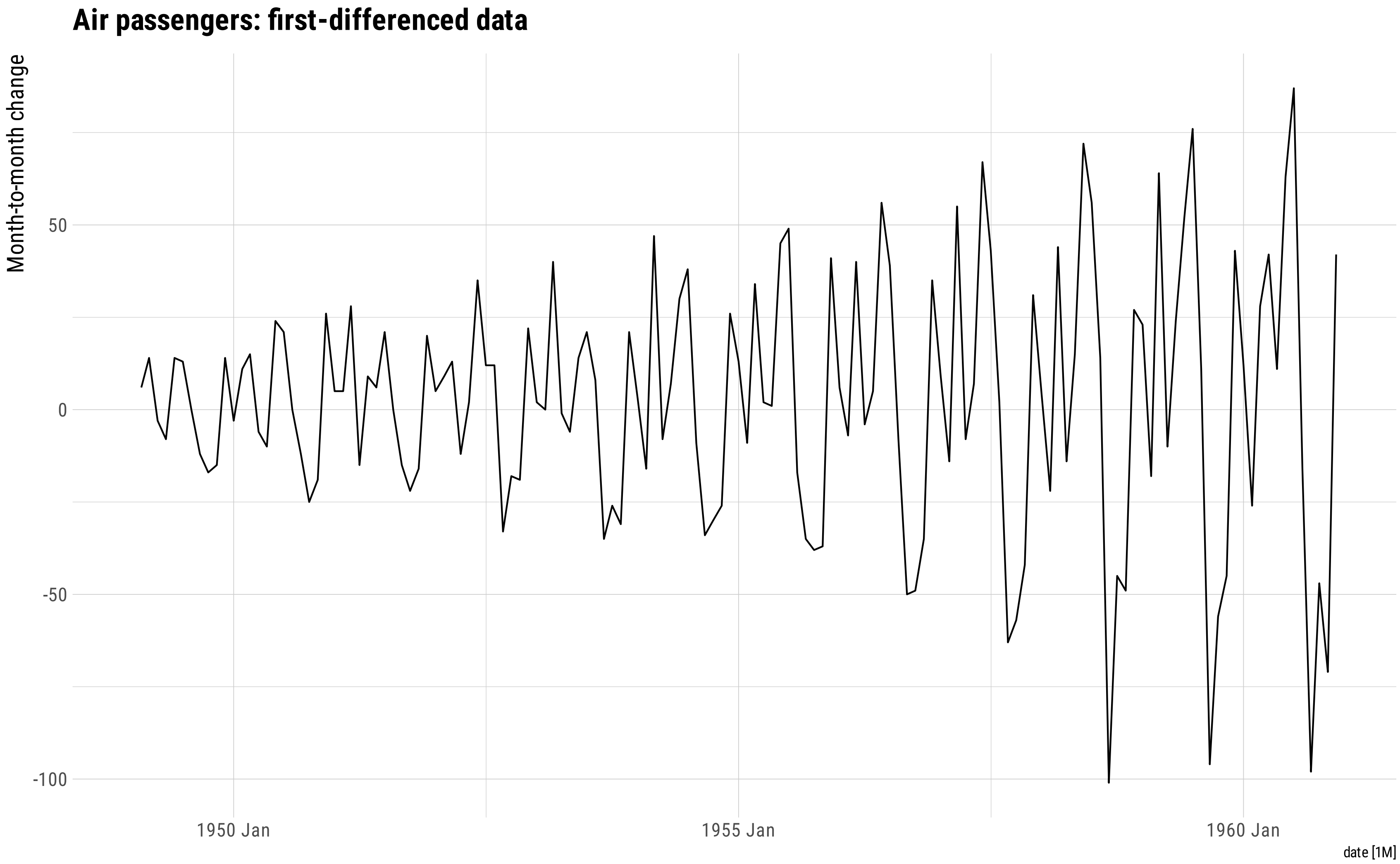


Stationarity

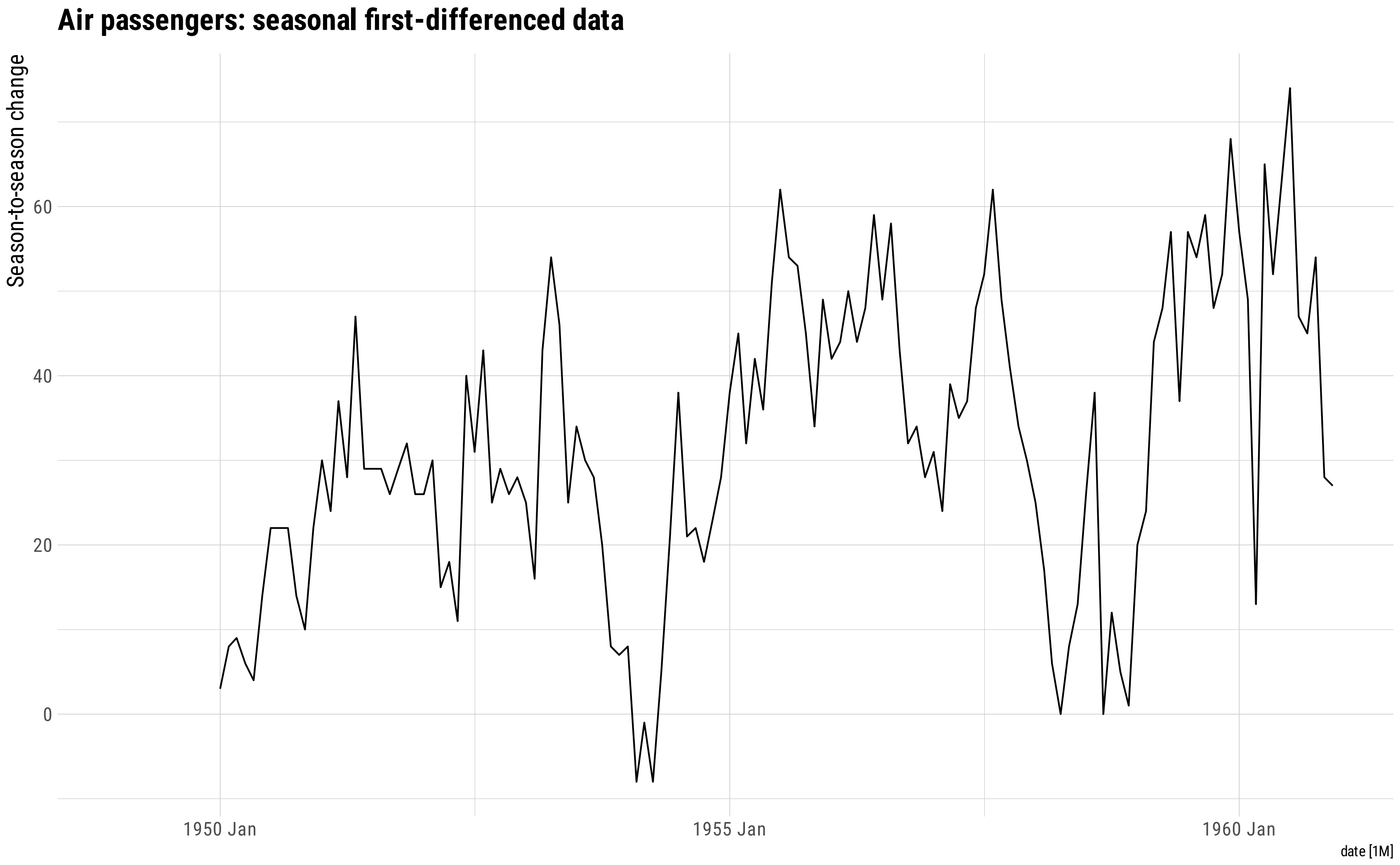
```
air_ts ▷
  mutate(diff_passengers = difference(passengers, lag = 1),
         seas_diff_passengers = difference(passengers, lag = 12),
         diff_log_passengers = difference(log(passengers), lag = 1),
         seas_diff_log_passengers = difference(log(passengers), lag = 12))
```

```
#> # A tsibble: 144 x 6 [1M]
#>        date passengers diff_passengers seas_diff_passengers diff_log_passengers seas_diff_log_passengers
#>       <mth>      <dbl>           <dbl>                <dbl>               <dbl>                    <dbl>
#>   1 1949 Jan        112              NA                   NA             NA                             NA
#>   2 1949 Feb        118               6                   NA              0.0522                        NA
#>   3 1949 Mar        132              14                   NA              0.112                         NA
#>   4 1949 Apr        129              -3                   NA             -0.0230                        NA
#>   5 1949 May        121              -8                   NA             -0.0640                        NA
#>   6 1949 Jun        135              14                   NA              0.109                         NA
#>   7 1949 Jul        148              13                   NA              0.0919                        NA
#>   8 1949 Aug        148               0                   NA              0                             NA
#>   9 1949 Sep        136             -12                   NA             -0.0846                        NA
#> 10 1949 Oct        119             -17                   NA             -0.134                         NA
#> # ℹ 134 more rows
```

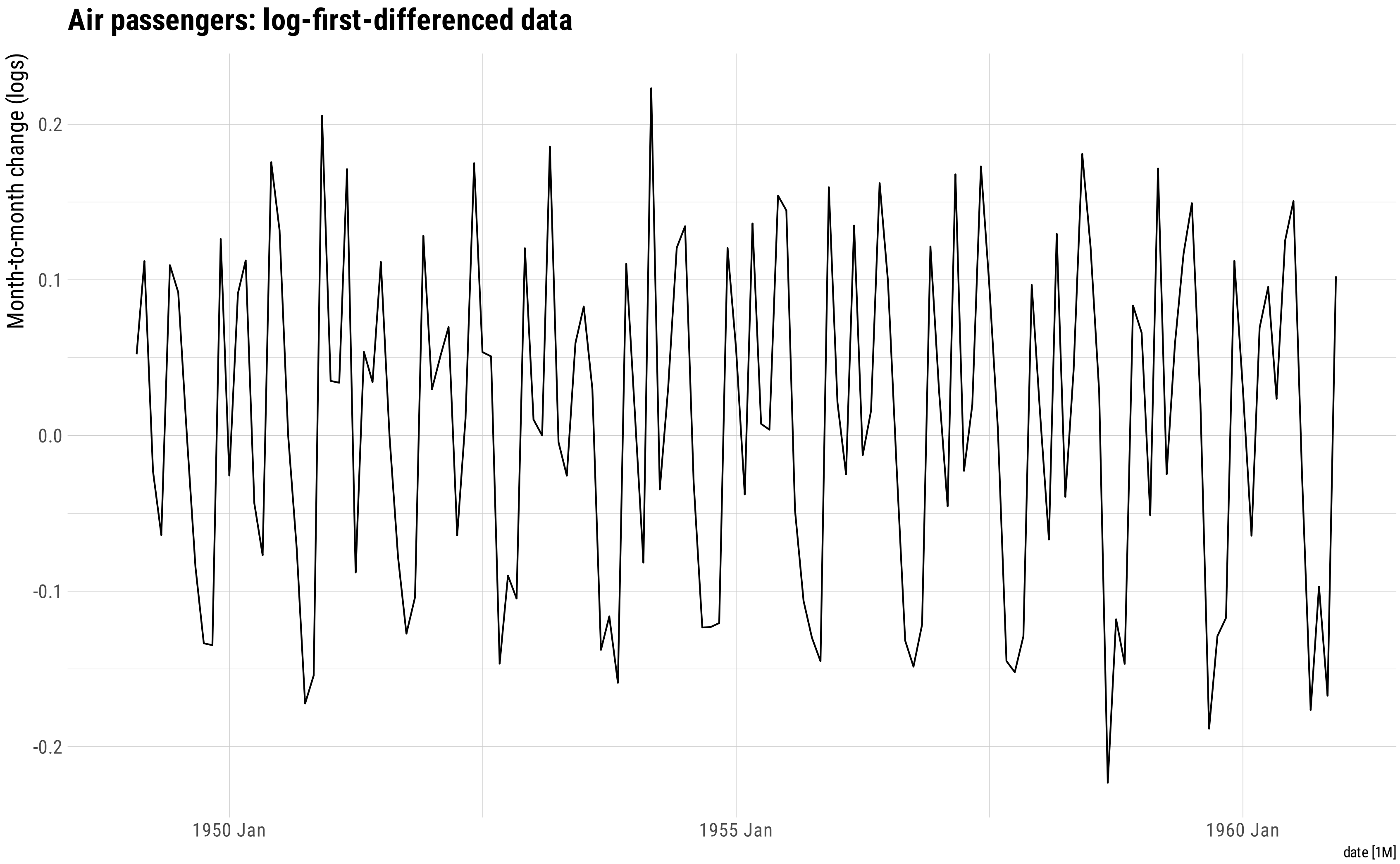
Stationarity



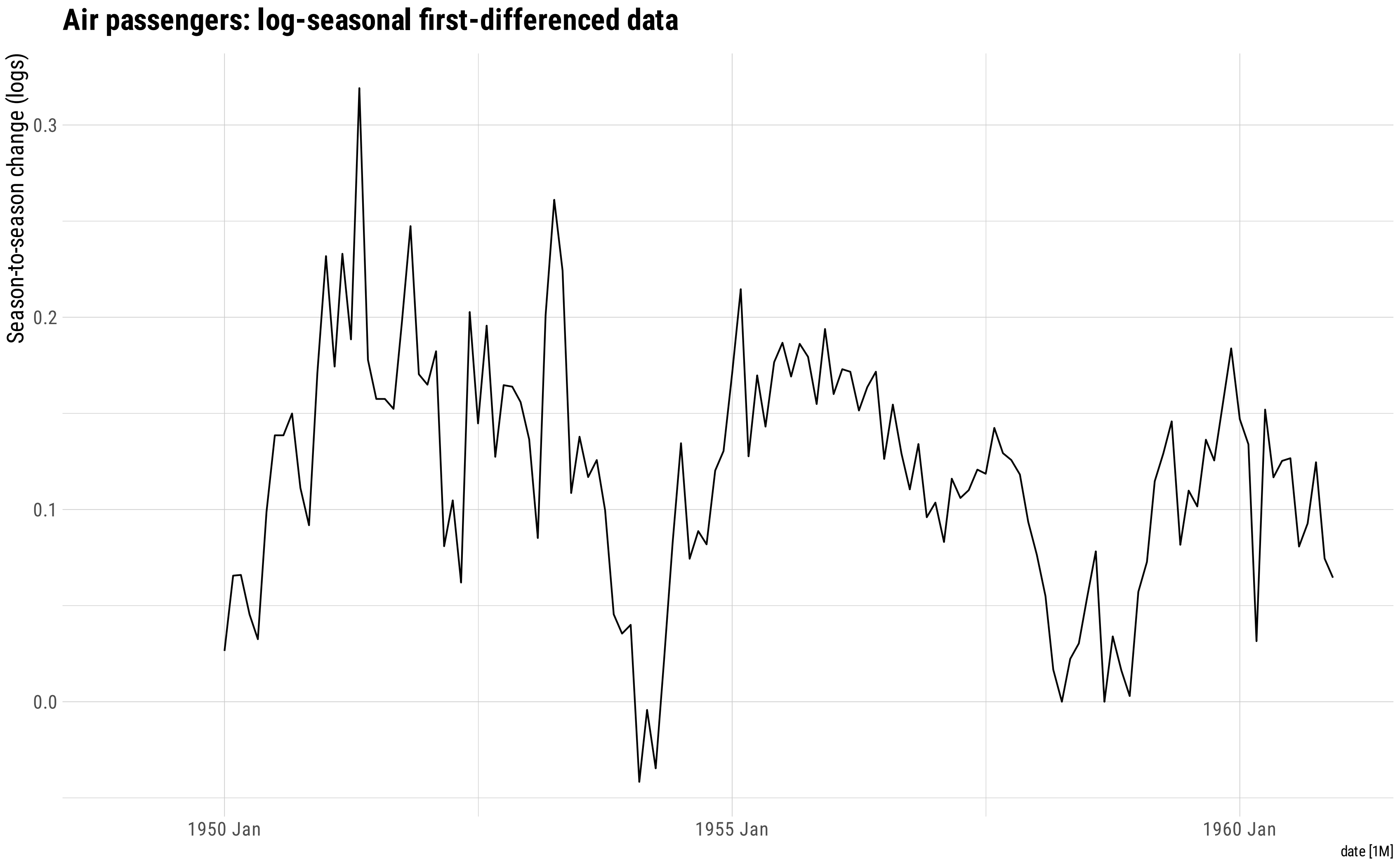
Stationarity



Stationarity



Stationarity



Stationarity

When applying **differencing** techniques to our data, there is **no definitive approach**.

There is always a **degree of subjectivity** when choosing the best way of dealing with the data we have at hand.

What has to be kept in mind, though, is **interpretability**.

- *First differences* are the change between one observation and the next;
- *Seasonal differences* are the change between one year to the next;
- Other lags are unlikely to make much *interpretable* sense and should be avoided.

Unit root tests

Unit root tests

Don't ever *undermine* your **subjectivity**/**gut feeling** when it comes to analyzing time series!

That said, a way to be more objective when it comes to determining whether a time series is stationary or not is by using **unit root testing**.

Unit root tests are statistical **hypothesis tests** of stationarity that are designed for determining *whether differencing is required*.

One of these tests is the **Kwiatkowski-Phillips-Schmidt-Shin (KPSS)** test.

Its **null hypothesis** states that the data are *stationary*, and in case the latter is rejected, we have evidence in favor of **differencing** our data.

Unit root tests

```
air_ts ▷
  features(passengers, unitroot_kpss)
```

```
#> # A tibble: 1 × 2
#>   kpss_stat kpss_pvalue
#>       <dbl>       <dbl>
#> 1      2.74        0.01
```

```
air_ts ▷
  features(difference(passengers, lag = 1), unitroot_kpss)
```

```
#> # A tibble: 1 × 2
#>   kpss_stat kpss_pvalue
#>       <dbl>       <dbl>
#> 1    0.0146         0.1
```

Backshift notation

Backshift notation

As we will progressively explore autocorrelations in our time series, working with **lagged values** will become more necessary.

To make this easier in terms of **notation**, we can introduce the **backshift operator**, *B*:

$$By_t = y_{t-1}$$

In words, *B* **shifts back** the data one period in time.

In case we want to shift back the series two periods, we may write:

$$B^2 y_t = y_{t-2}$$

Backshift notation

For **differencing** purposes, the backward shift operator can be used as follows:

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t$$

In case we use **second differencing**, we can write:

$$y''_t = (1 - B)^2 y_t$$

In general, a ***d*th-order difference** can be written as

$$y'_t = (1 - B)^d y_t$$

Autoregressive models

Autoregressive models

When we estimate a **regression model**, our aim is to explain the behavior of a *dependent* variable in terms of a **linear combination** of *independent* (*predictor*) variables.

When these predictor variables are **past values** of the variable of interest, we have the following:

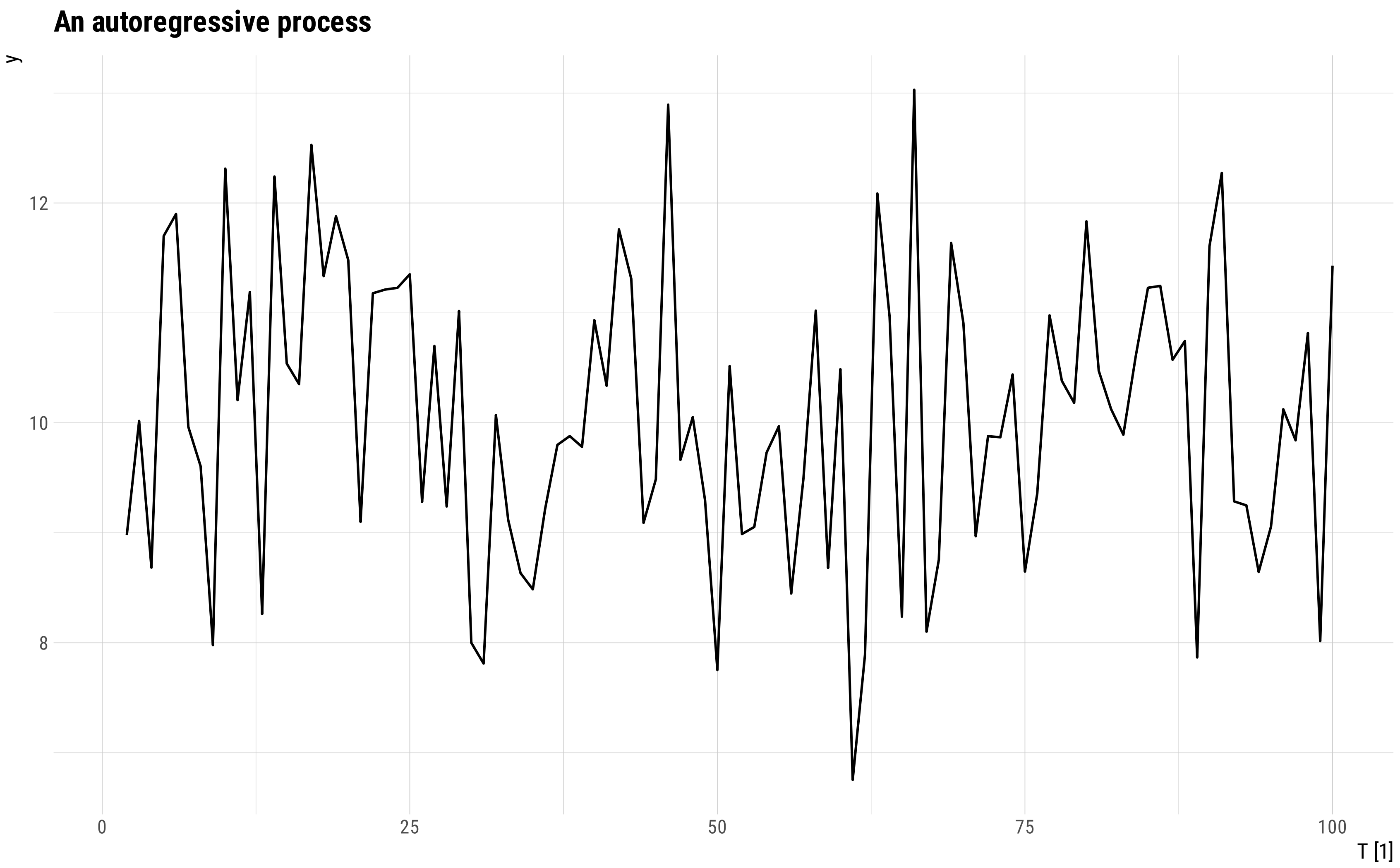
$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where ε_t is white noise.

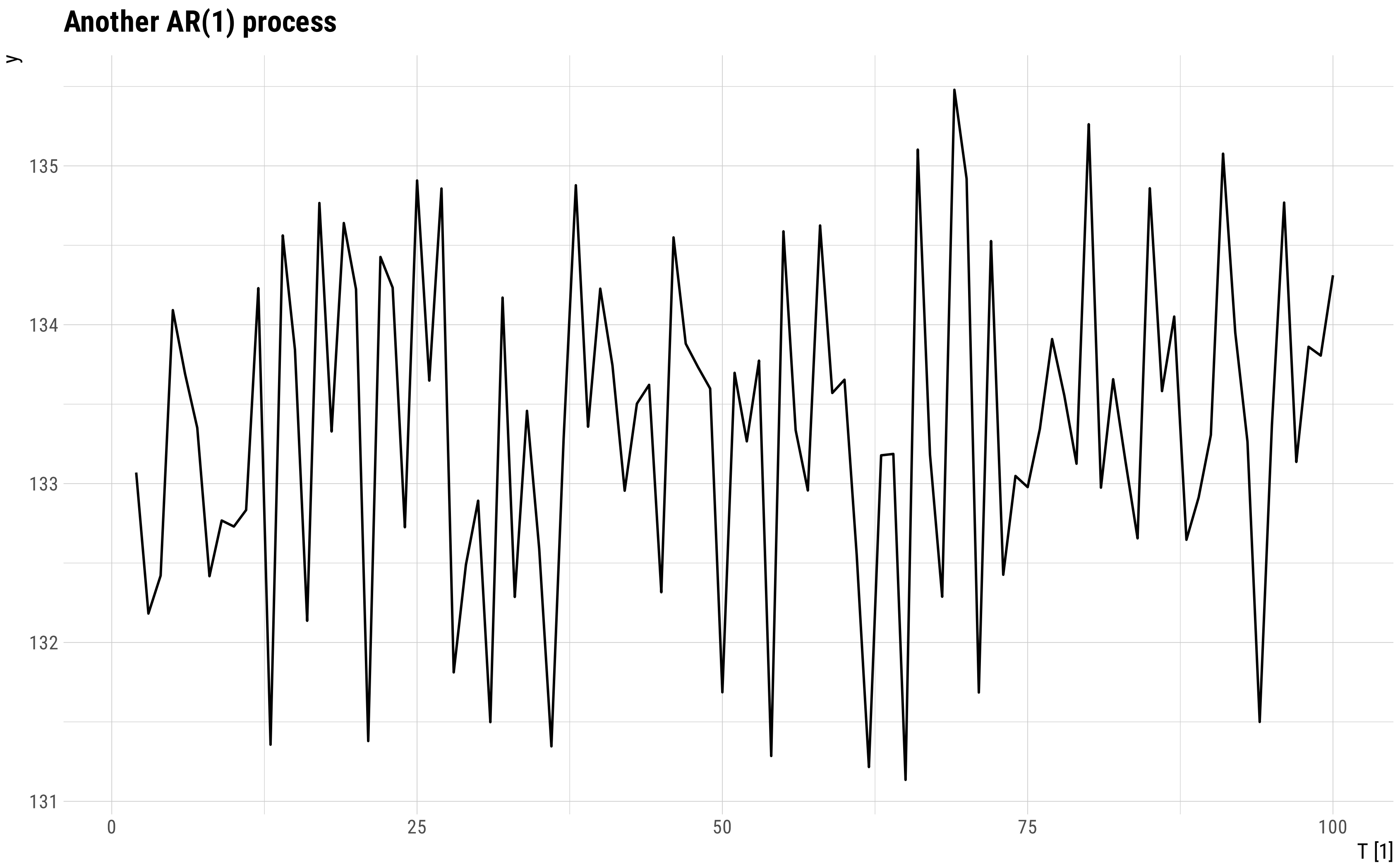
The above model is known as an **autoregressive model** of order p.

- In short, we may write this as an **AR(*p*)** model.

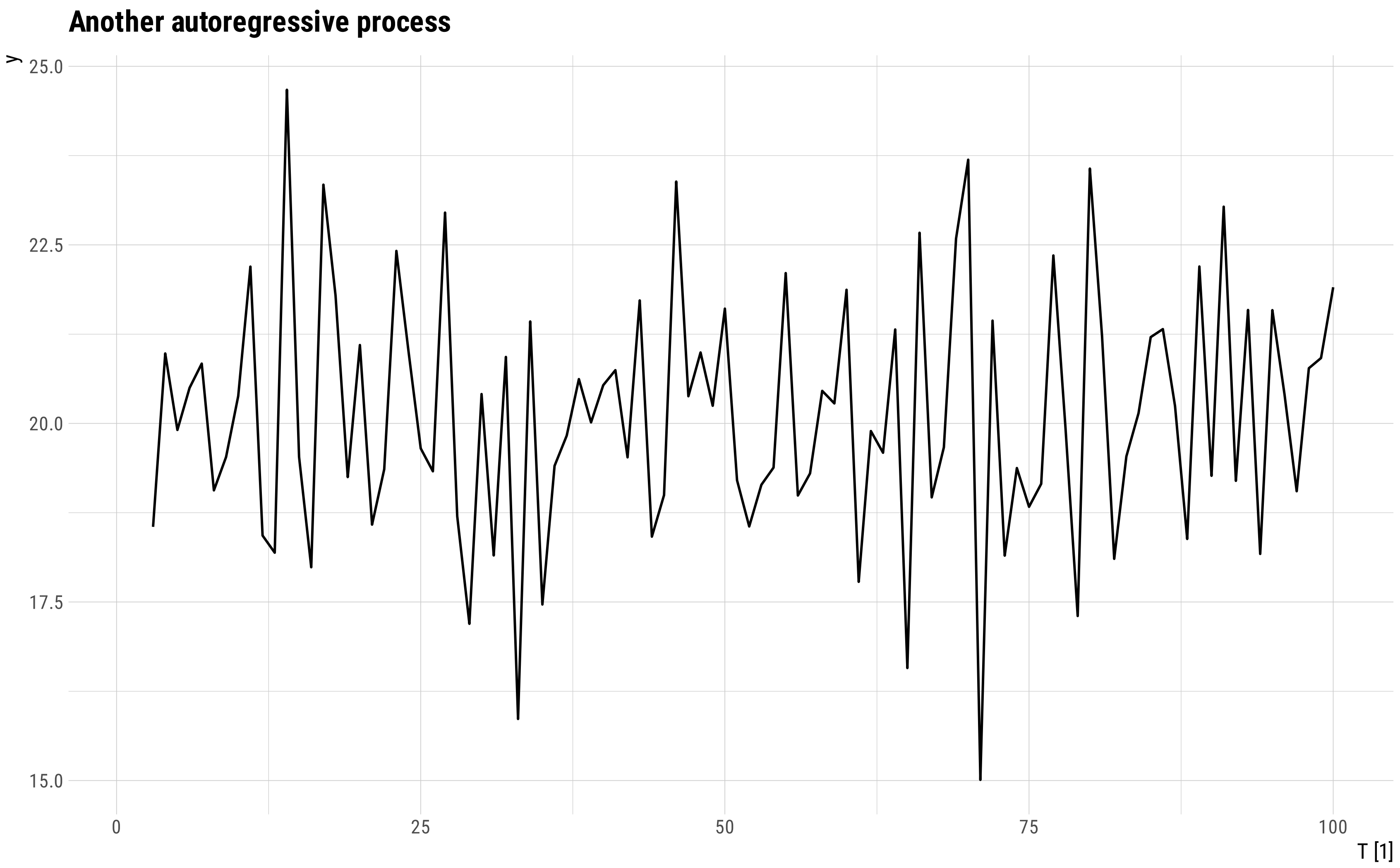
$$y_t = 18 - 0.8 y_{t-1} + \varepsilon_t$$



$$y_t = 100 + 0.25y_{t-1} + \varepsilon_t$$



$$y_t = 8 + 1.3y_{t-1} - 0.7y_{t-2} + \varepsilon_t$$



Autoregressive models

A few **remarks**:

- For an **AR(1)** model: $y_t = c + \phi_1 y_{t-1} + \epsilon_t$:
  - If $\phi_1 = 0$ and $c = 0$, y_t is *white noise*;
  - If $\phi_1 = 1$, y_t follows a *random walk* process;
  - If $\phi_1 = 1$ and $c \neq 0$, y_t is follows a *random walk with drift*.

Moving average models

Moving average models

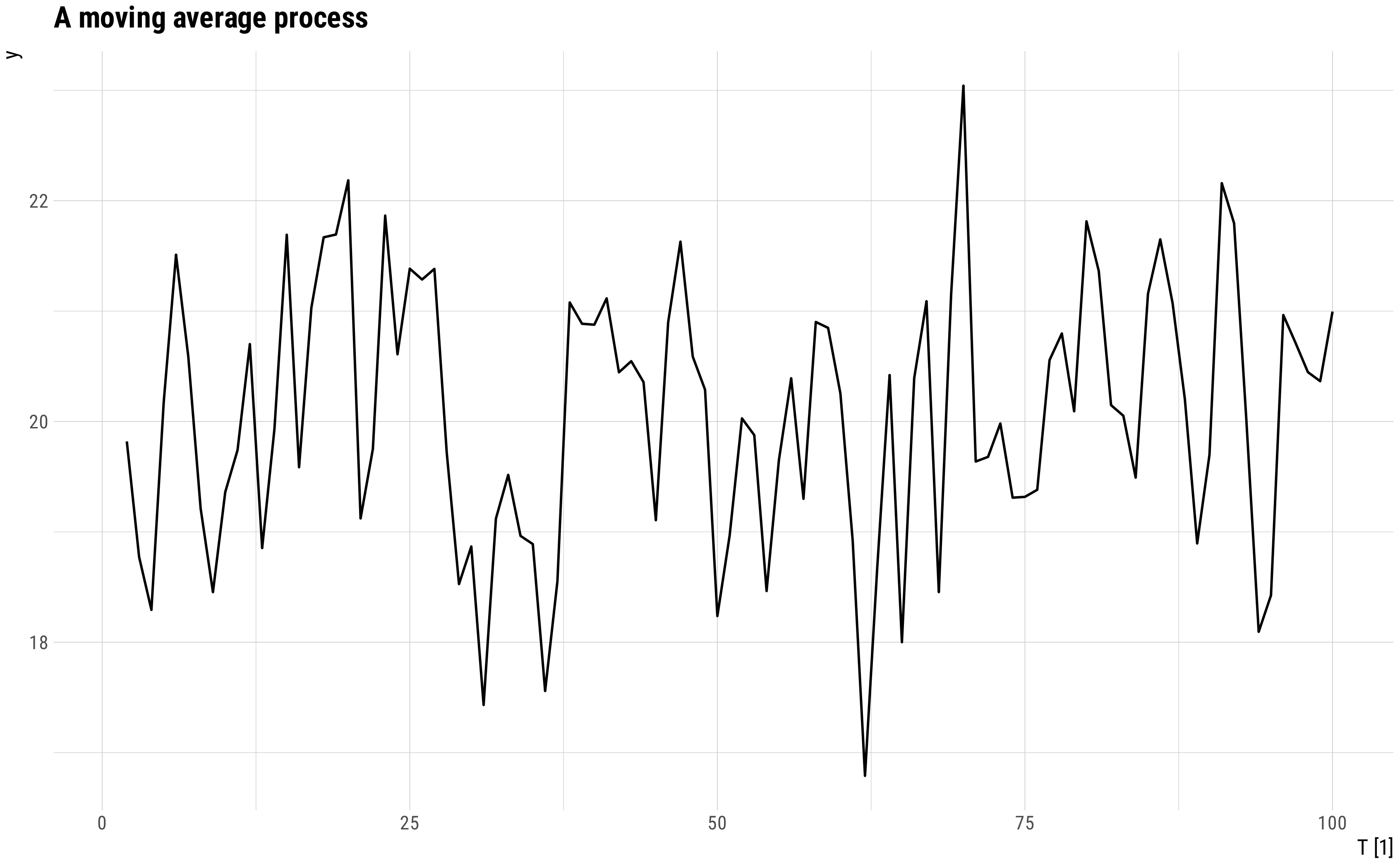
A **Moving Average (MA) process** of order q can be expressed as follows:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \cdots + \theta_q \varepsilon_{t-q}$$

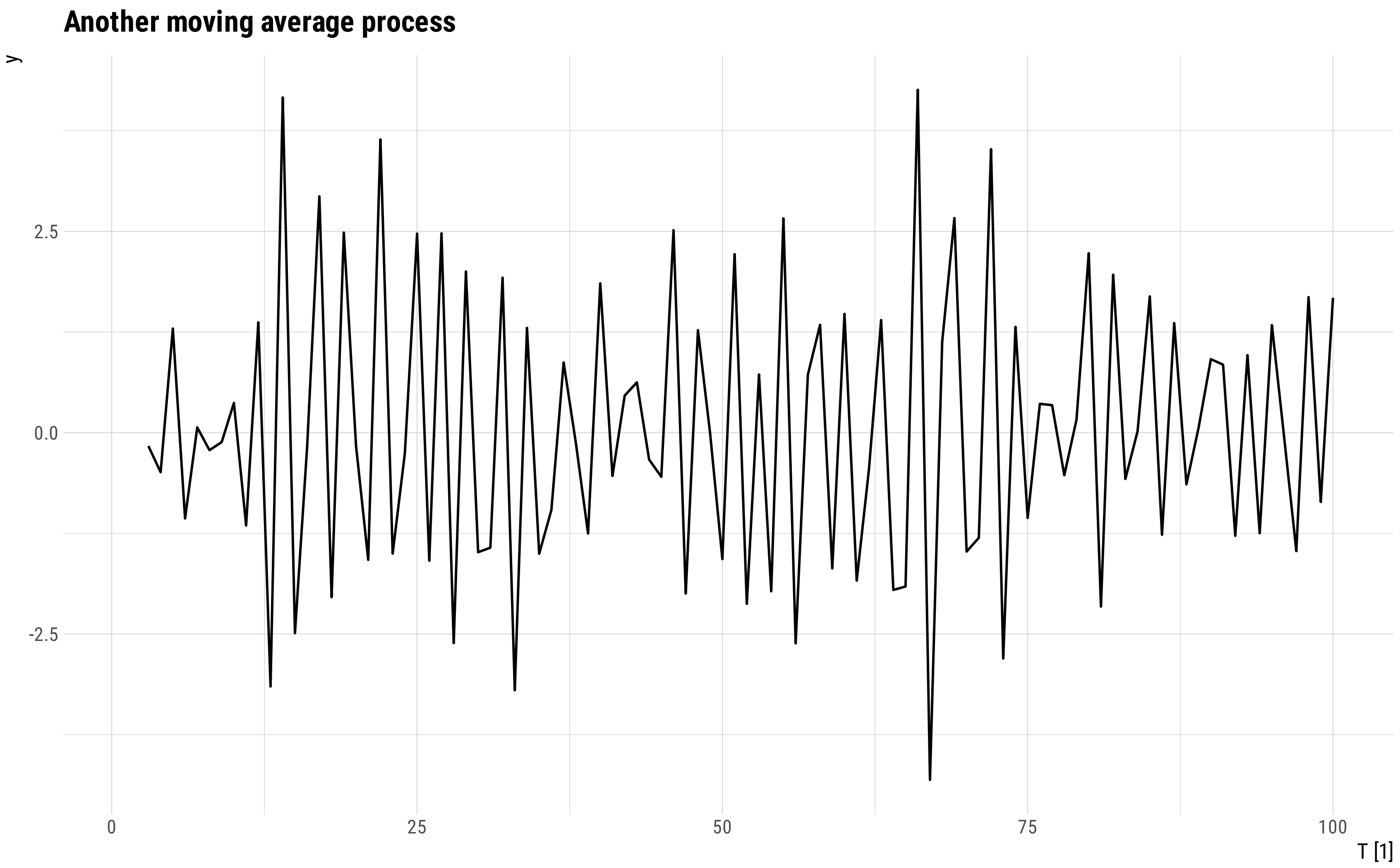
where ε_t is white noise.

In words, a moving average model uses a linear combination of **past forecast errors** to explain the current behavior of a time series y_t.

$$y_t = 20 + \varepsilon_t + 0.8\varepsilon_{t-1}$$



$$y_t = \varepsilon_t - \varepsilon_{t-1} + 0.8\varepsilon_{t-2}$$



Next time: More on ARIMA models