Benchmark forecasting methods

EC 361–001

Prof. Santetti
Spring 2024

Materials

Required readings:

- `Hyndman & Athanasopoulos, ch. 2, section 2.7`
- `Hyndman & Athanasopoulos, ch. 6`
    - sections 6.2—6.3.

Motivation

Motivation

Now that we have been introduced to the most important **visualization** and **decomposition** techniques, it is time to (*finally*) start doing some forecasting exercises.

The best place to start is to **keep it simple**.

Thus, we begin with so-called **benchmark** forecasting methods.

The statistical perspective

The statistical perspective

Since any forecasting exercise involves estimating something we **do not know**, we may consider this (these) future observation(s) as **random variable(s)**.

Recall what a random variable is:

> A **random variable** is a variable whose possible values are numerical outcomes of a *random phenomenon*.

As one tries to forecast *further* in the future, the more **uncertain** the results are.

Given that, the usual procedure is to present the **average** value within the range of possible values the random variable we are trying to forecast could take.

In addition, a forecast is often accompanied by a **prediction interval** giving a range of values the random variable could take with relatively *high probability*.

The statistical perspective

If we want to forecast the future values of a variable y_t, we will usually do so based on some set of **information**.

This set of information may include *personal beliefs*, *past observations* of this variable, and/or values of *other variables* we assume may influence y_t.

Let us call this information the **information set**, $\mathcal{I}$.

Thus writing $y_t|\mathcal{I}$ means "*the random variable y_t given our information set $\mathcal{I}$.*"

Finally, the set of values that this random variable could take, along with their relative probabilities, is known as the probability distribution of $y_t|\mathcal{I}$.

In forecasting, the above is called the **forecast distribution**.

The statistical perspective

And we will establish our notation for forecasting by using a "*hat*" symbol ($\hat{}$) denoting a **fitted/estimated point forecast** value (usually the average).

So $\hat{y}_t$ is the (average) point forecast value of the forecast distribution of $y_t|\mathcal{I}$, meaning the average of the possible values that y_t could take given everything we know.

In case we want to be more **explicit** about our information set $\mathcal{I}$, we may write, for example,

$$\hat{y}_{t|t-1}$$

to denote that we are forecasting variable *y* using information from its past observations ($y_{t-j,\dots,t-1}$).

In a similar way,

$$\hat{y}_{T+h|T}$$

means that we are forecasting variable *y* *h* steps (periods) ahead, based on its observations up to time *T*.

Benchmark forecasting methods

Benchmark forecasting methods

When producing forecasts, we may test several different models and approaches.

However, one way to test the **efficiency** of these models is to *compare* them to very simple methods.

We will explore four of these "**benchmark**" forecasting models, which will provide a basic reference upon which we should improve our more sophisticated models.

These are:

1. The **mean** method;
2. The **naïve** method;
3. The **seasonal naïve** method;
4. The **drift** method.

Benchmark forecasting methods

- **The mean method**:

The **mean** method, as the name suggests, will produce forecasts which are equal to the **average** value of the historical data.

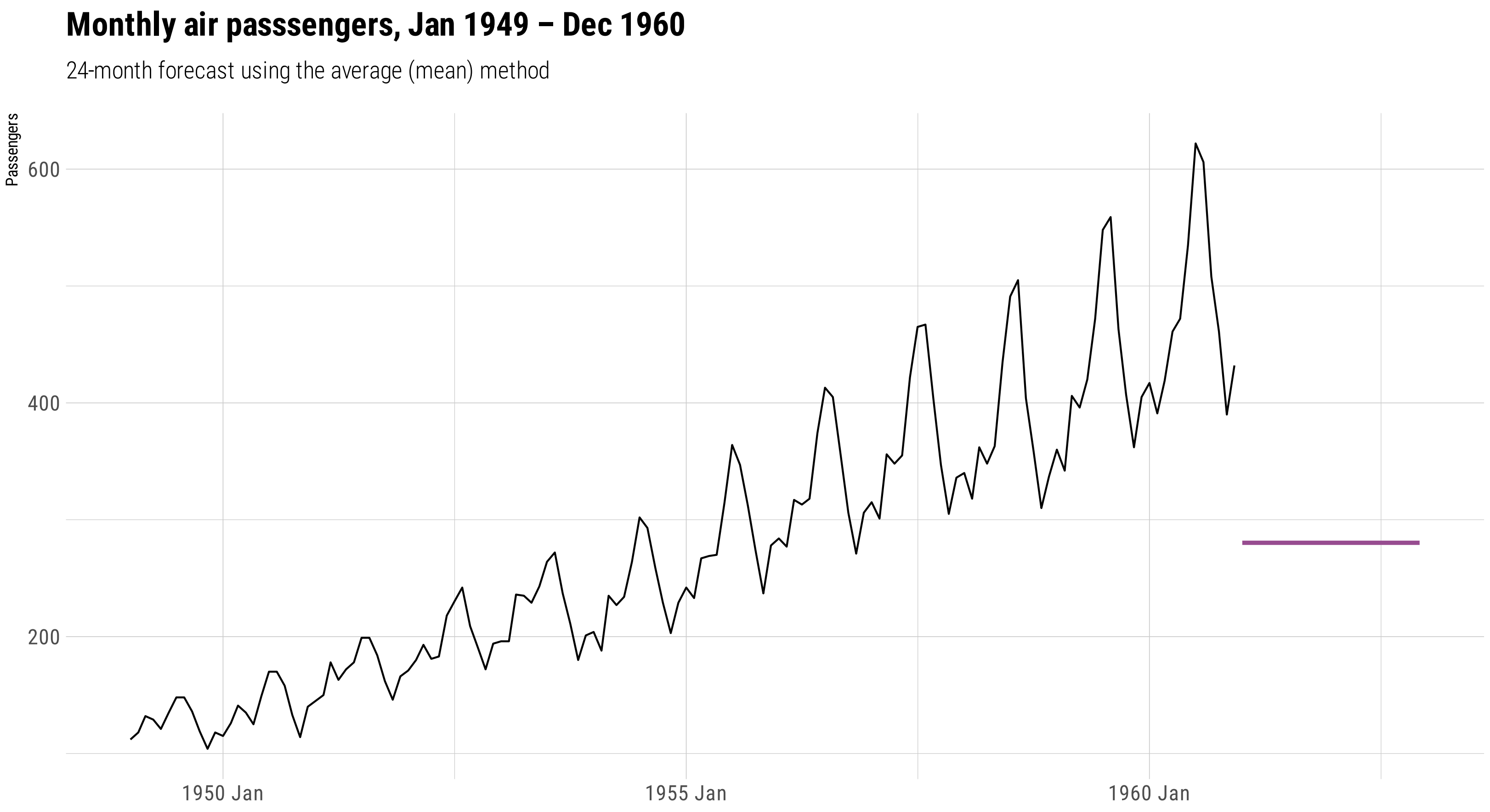
Let the historical data (i.e., the information set $\mathcal{I}$) be denoted by $(y_1, \ldots, y_T)$.

Then, the forecasts may be written as

$$\hat{y}_{T+h|T} = \bar{y} = \frac{\sum_{j=1}^{T} y_j}{T}$$

Benchmark forecasting methods

- **The mean method**:



Benchmark forecasting methods

- **The naïve method**:

The **naïve** method is based on producing forecasts whose values are simply equal to the **last** observation:

$$\hat{y}_{T+h|T} = y_T$$

Despite its *simplicity*, it may be well-suited for when the data follows a **random walk** process.

Benchmark forecasting methods

- **The naïve method**:

A **random walk** process can be defined as follows:

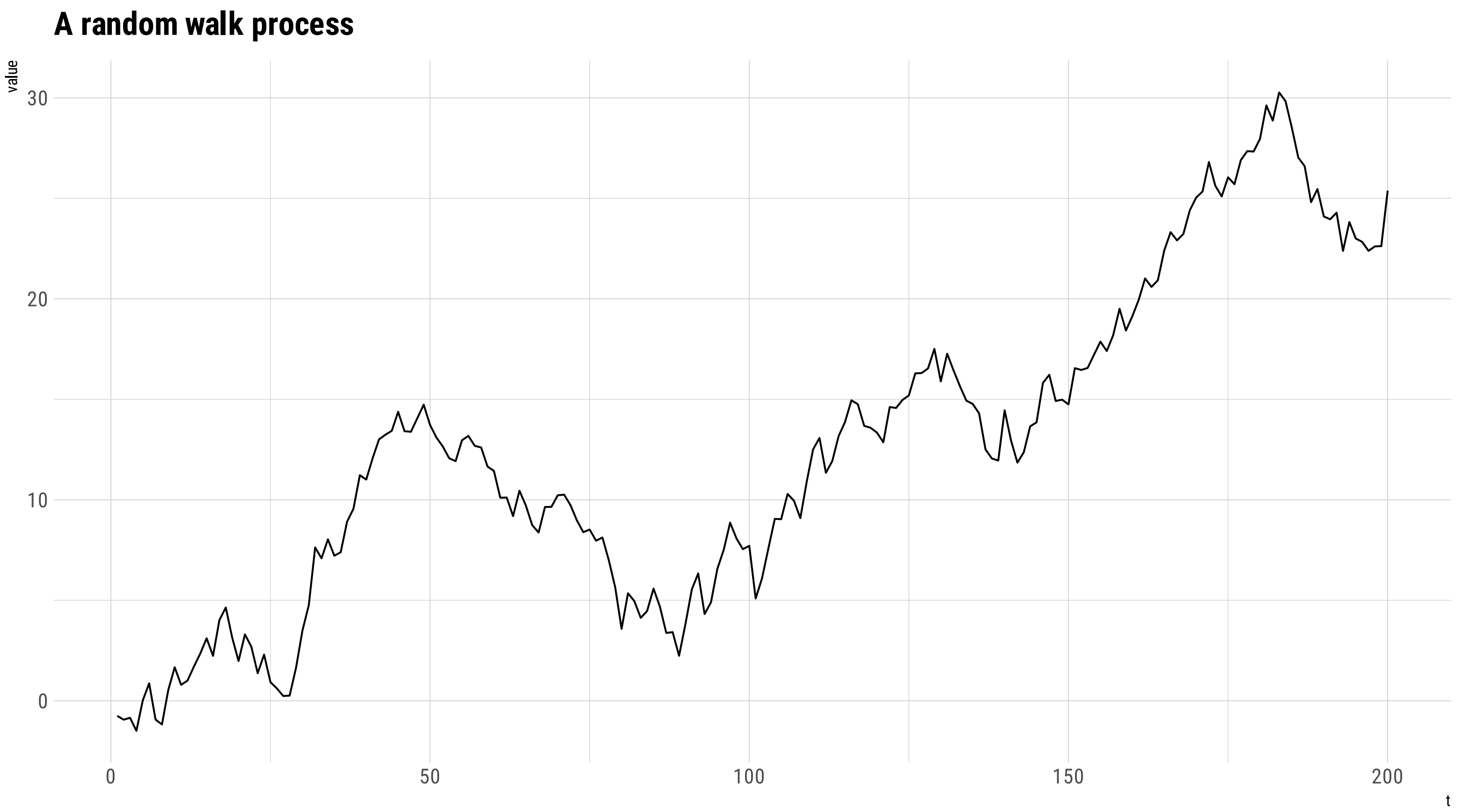
$$y_t = y_{t-1} + \varepsilon_t$$

where ε_t is a **random** component following a white noise process (i.e., not **autocorrelated**).

The term **random walk** comes from the fact that the value of the time series at time *t* is the value of the series at time *t-1* plus a completely **random movement** determined by ε_t.

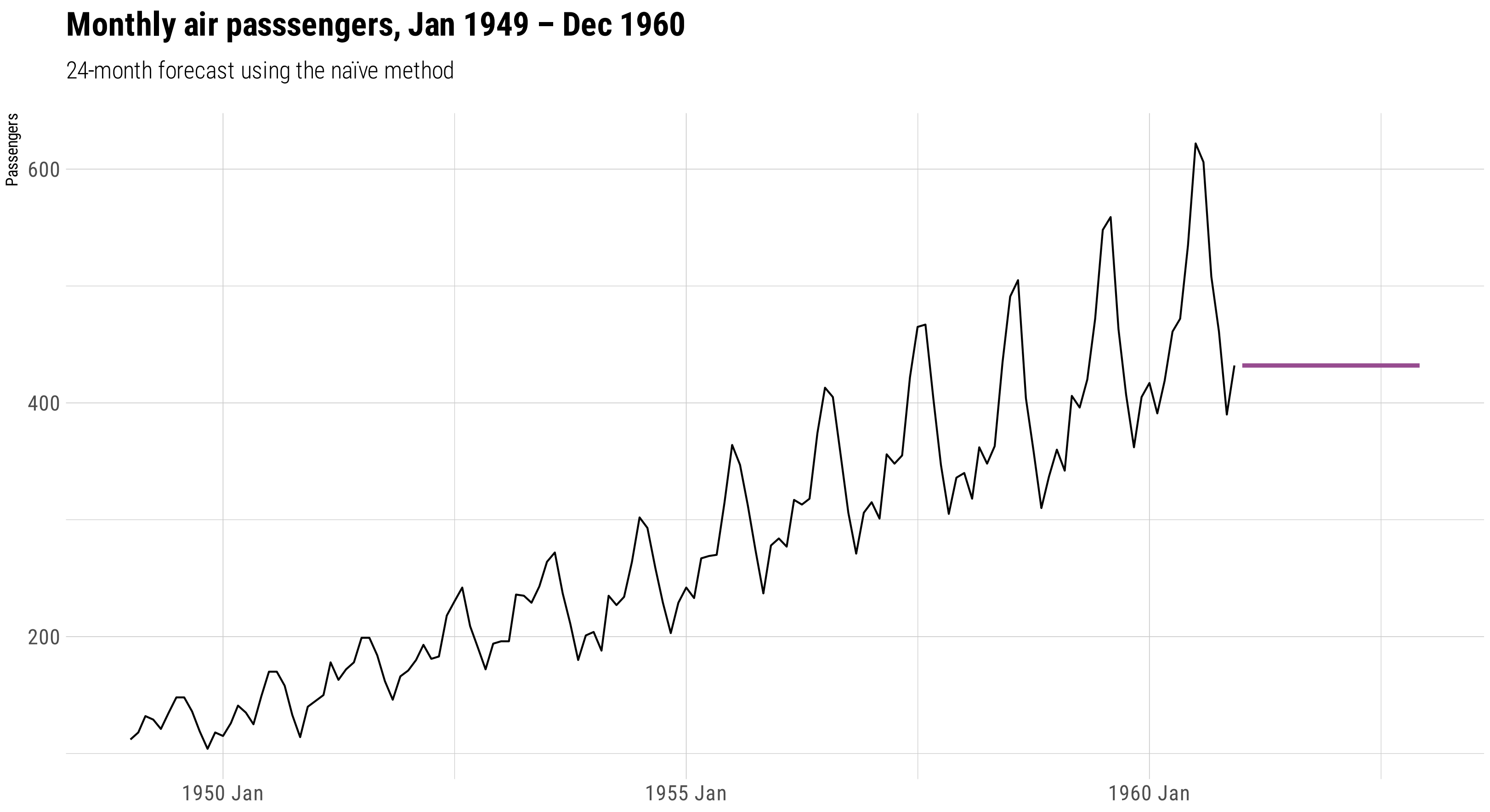
Benchmark forecasting methods

- **The naïve method**:



Benchmark forecasting methods

- **The naïve method**:



Benchmark forecasting methods

- **The seasonal naïve method**:

When the data are **highly seasonal**, the *naïve* method can be improved by including a **seasonal** component.

This way, each forecast will be *equal* to the **last observed value** from the **same season**.

- e.g., the same *month* of the *previous year*, the same *quarter* of the *previous year*, and so on.

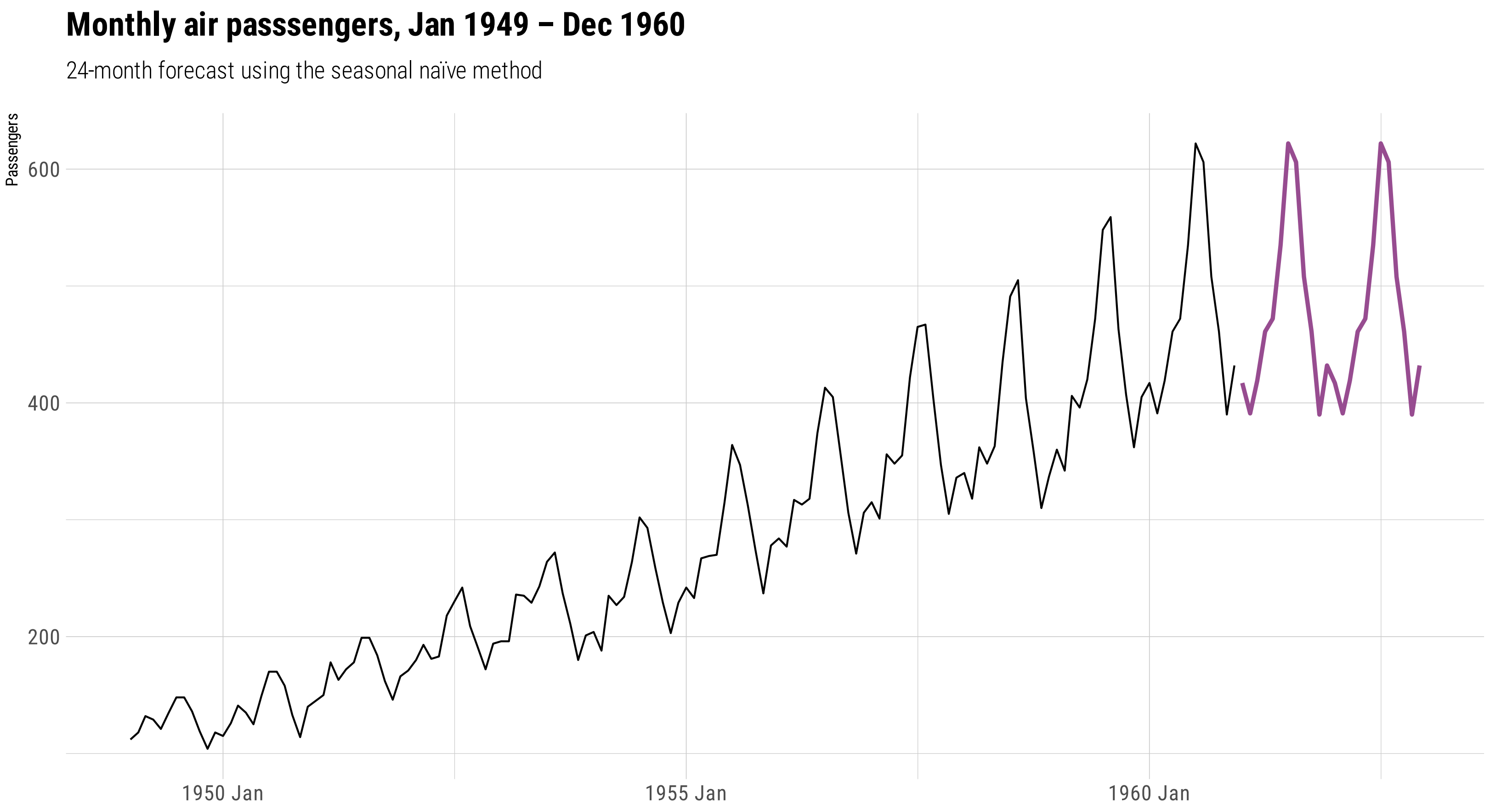
More formally,

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

where m is the seasonal period, and $k = (h-1)/m$.

Benchmark forecasting methods

- **The seasonal naïve method**:



Benchmark forecasting methods

- **The drift method**:

The **drift** method expands on the naïve model by allowing the forecasts to **increase or decrease** over time.

This *change* over time is called the **drift**.

The drift term is set to be the *average change seen in the historical data*.

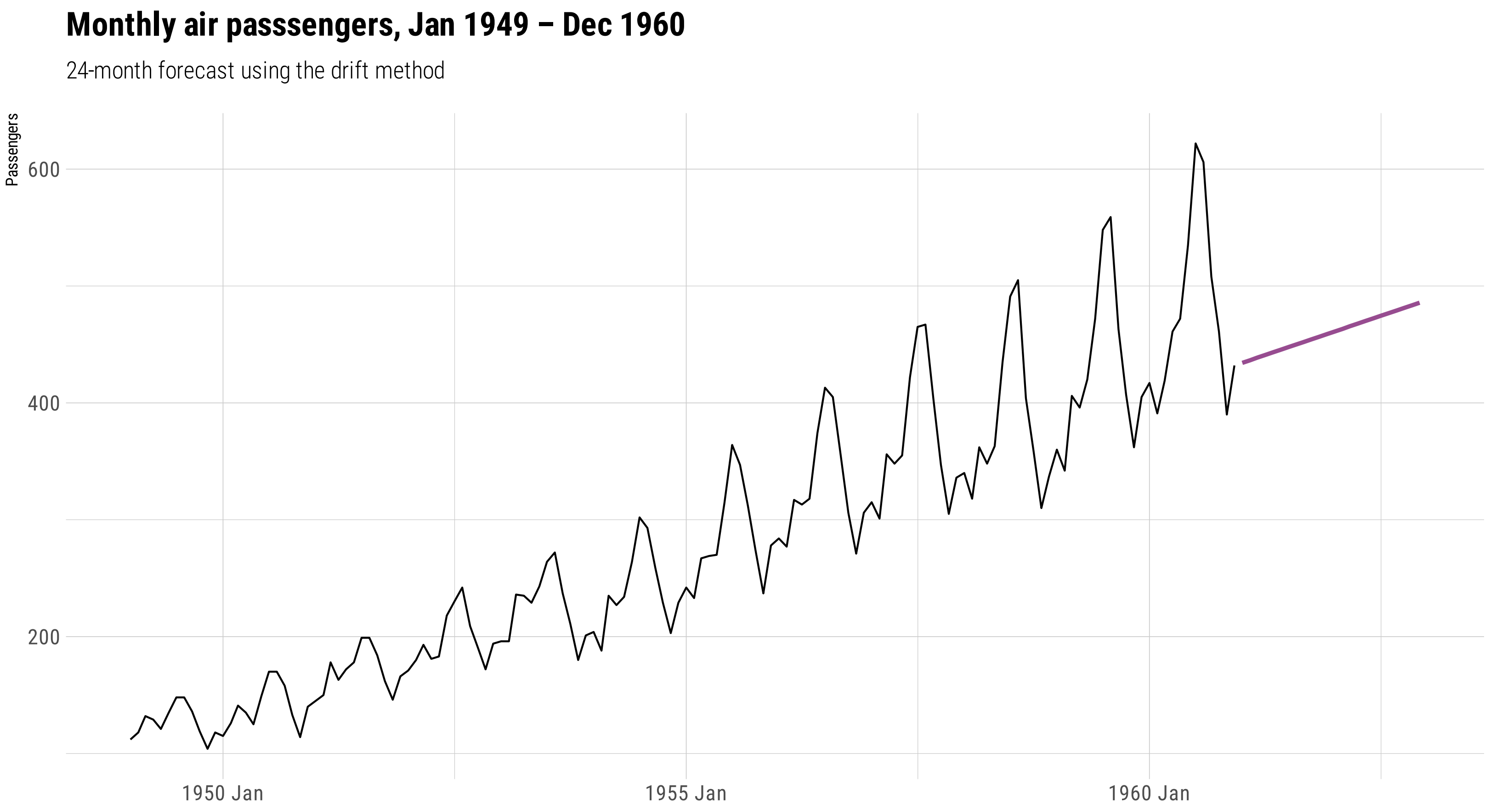
Formally,

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) = y_T + h\left(\frac{y_T - y_1}{T-1}\right)$$

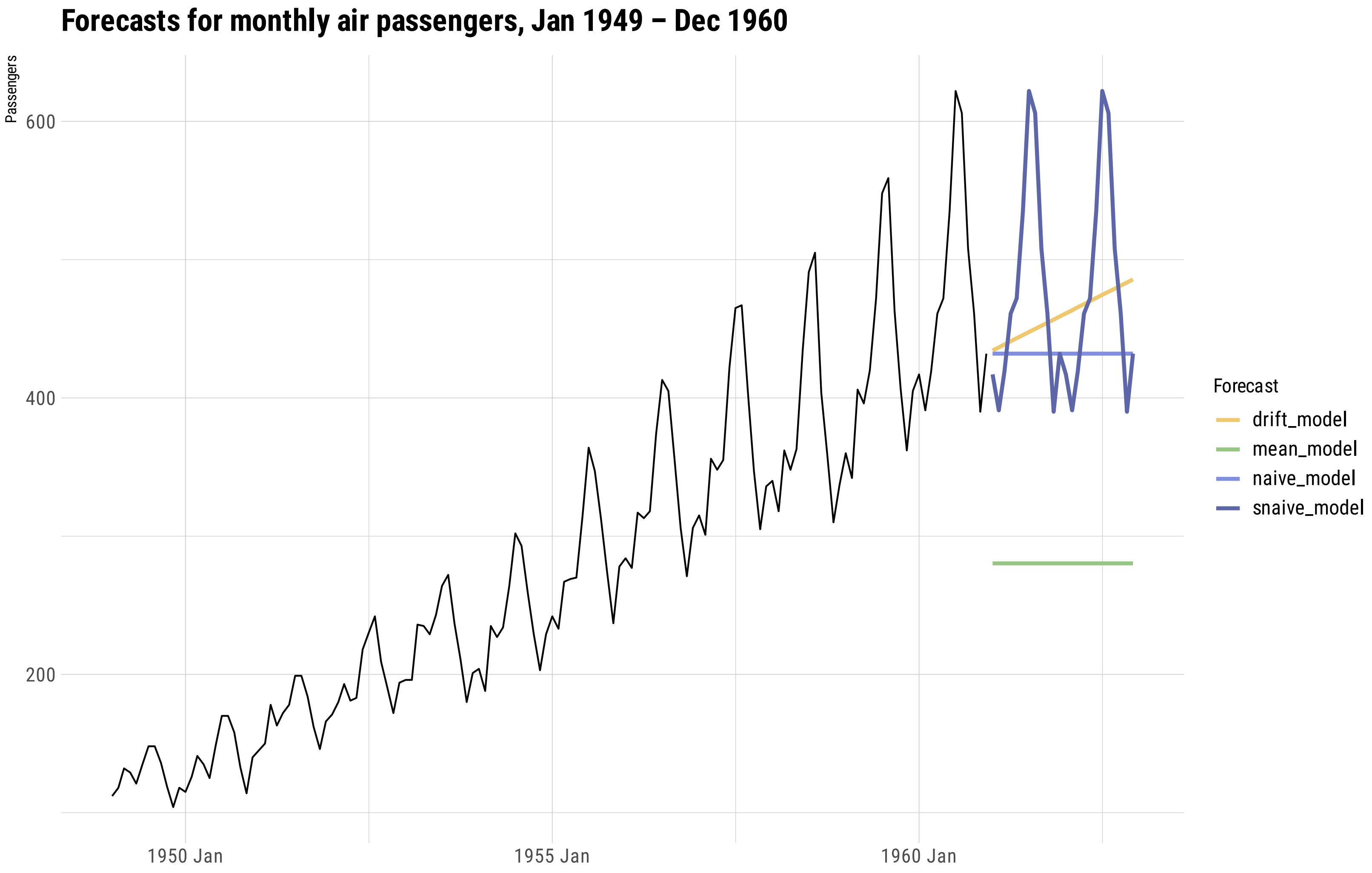
This formula is equivalent to drawing a **line** between the first and last observations, and *extrapolating* it into the future.

Benchmark forecasting methods

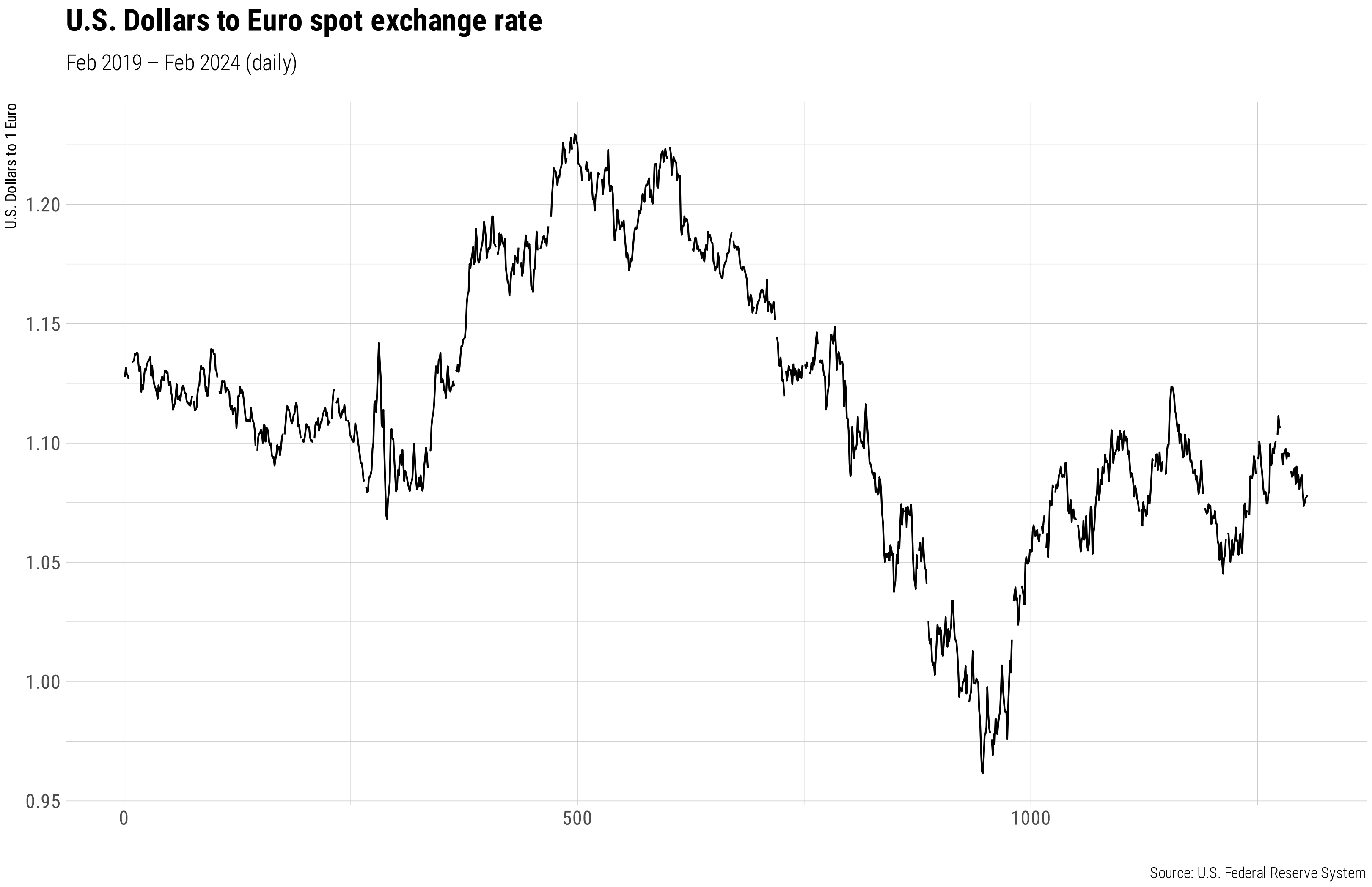
- **The drift method**:



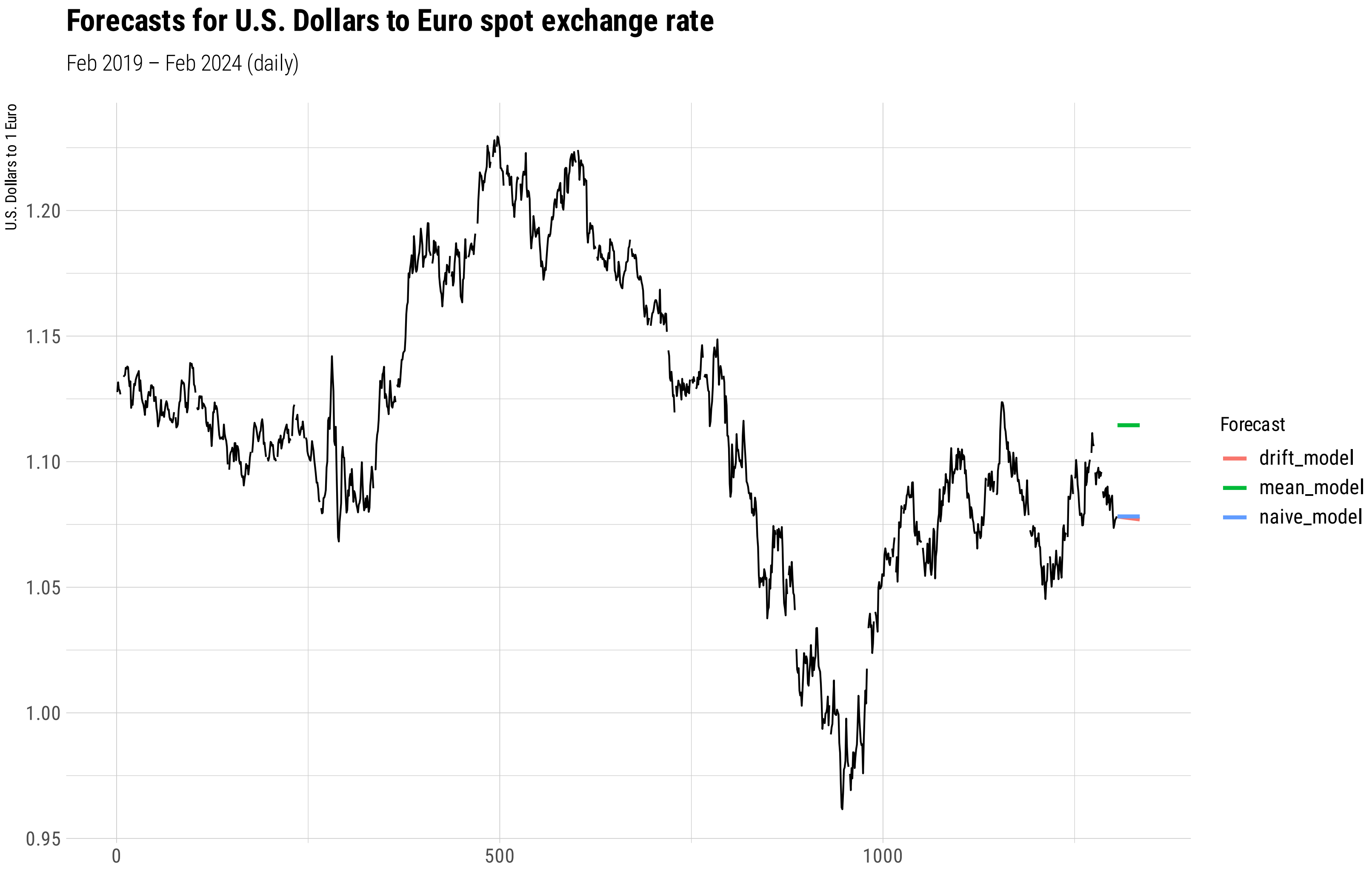
Benchmark forecasting methods



A second example



A second example



Next time: Residual analysis