Statistical Data Processing

Time series analysis

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

Outline of the lecture

- Time series
- Decomposition of the time series into

  the trend, seasonal, cyclic, and random component
- The trend component
- The seasonal component
- Autocorrelation of the random component: Durbin-Watson test
- Moving average: simple, centred, and weighted moving average

Time series

A time series is a sequence of data points listed in time order.

A time series is usually a sequence of data variable (Y) taken at equidistant points of time (such as one hour, one day, one week, one year, and so on).

Without loss of generality, we can assume that the time points are $t = 0, 1, 2, 3, \ldots$

In other words, we usually consider discrete-time data.

We distinguish:

- instantaneous value type time series
- step accumulated value type time series

Time series

- **Instantaneous value type time series**

  — the measured value is taken at a particular time.

  Examples: the air temperature, the wind speed

- **Step accumulated value type time series**

  — the measured value is take over an entire interval of time

  Example: the rainfalls during the past hour

Time series

It is usually assumed that

- **the main factor** that affects the observed value Y mainly **is the time**,
- the time points at which the data are observed are equidistant; that is, the time intervals are of the same length.

The goal is to formulate a mathematical model of the time series and to use it for **time series forecasting** (prediction).

We distinguish two types of prediction:

- point prediction

Time series

We usually assume that the time series can be decomposed into the following four components:

- T_t — the trend component
- S_t — the seasonal component
- C_t — the cyclic component
- ε_t — the irregular component / the random error

Time series

We then can assume:

- either **the additive model** of the time series:

$$Y_t = T_t + S_t + C_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

- or **the multiplicative model** of the time series:

$$Y_t = T_t \times S_t \times C_t \times \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where

- T_t — the trend component
- S_t — the seasonal component
- C_t — the cyclical component

Time series

The **trend component** T_t describes the long-term progression of the time series.

It reflects the systematic and long-term effect of the main factors.

A special case is if the trend is identically zero ($T_t = 0$ for all t);

the time series has no trend then.

The **seasonal component** S_t reflects the seasonality, i.e. seasonal factors,

which repeat periodically during a fixed and known period of time, often a year

(sometimes a week). The seasonality is then observed over intervals shorter than

a year (or a week), such as the quarter of the year / month / week

Time series

The **cyclical component** C_t captures fluctuations (rises and falls) that are repeated but have not a fixed period, such as the "business cycle" (economic cycle / trade cycle). The period is not fixed and usually longer than one year.

The **random error** ε_t is often assumed to be distributed normally and homoskedastic (with the same variance), that is

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2) \qquad \text{for every} \quad t = 0, 1, 2, 3, \ldots$$

and the random variables ε_t are assumed to be **mutually independent**.

Time series: Special cases

The general (additive) model, which we assume is

$$Y_t = T_t + S_t + C_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

A special case is when the cyclical component C_t is zero:

$$Y_t = T_t + S_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

A yet more special case is when both the cyclical component C_t and the seasonal component S_t are zero:

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

Time series: Trend component

From now on, we assume the following simple (additive) model of the time series:

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where T_t is the trend component.

The trend component T_t often falls into one of the following cases:

- constant / linear / quadratic / … / polynomial trend
- exponential / logarithmic trend
- logistic trend
- Gompertz trend

Time series: Constant trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \beta_0 \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real number

$$\beta_0 \in \mathbb{R}$$

The parameter $\beta_0 \in \mathbb{R}$ can be estimated by using the method of

Time series: Linear trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \beta_0 + \beta t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\beta_0 \in \mathbb{R} \qquad \text{and} \qquad \beta \in \mathbb{R}$$

The parameters $\beta_0, \beta \in \mathbb{R}$ can be estimated by using the method of

Time series: Quadratic trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2 \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\beta_0, \beta_1, \beta_2 \in \mathbb{R}$$

The parameters $\beta_0, \beta_1, \beta_2 \in \mathbb{R}$ can be estimated by using the method of

Time series: Polynomial trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k \in \mathbb{R}$$

The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k \in \mathbb{R}$ can be estimated by using the method of

Time series: Exponential trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \beta_0 \beta^t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\beta_0 \in \mathbb{R} \qquad \text{and} \qquad \beta \in \mathbb{R}$$

The parameters $\beta_0, \beta \in \mathbb{R}$ can be estimated by using the method of

Time series: Logarithmic trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \beta_0 + \beta \ln t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\beta_0 \in \mathbb{R} \qquad \text{and} \qquad \beta \in \mathbb{R}$$

The parameters $\beta_0, \beta \in \mathbb{R}$ can be estimated by using the method of

Time series: Logistic trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \frac{\kappa}{1 + \beta_0 \beta^t} \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\kappa > 0 \qquad \text{and} \qquad \beta_0 > 0 \qquad \text{and} \qquad 0 < \beta < 1$$

Time series: Gompertz trend

The time series is

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

where the trend component is of the form

$$T_t = \alpha \times \beta^{\gamma^t} \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

for some (unknown but) fixed real numbers

$$\alpha \quad (\text{usually } \alpha > 0) \qquad \text{and} \qquad \beta > 0 \qquad \text{and} \qquad 0 < \gamma < 1$$

Time series: Which trend to choose?

Rule	Suggested Trend
$\Delta^1 y_t \approx \text{const.}$	Linear
$\Delta^1 y_t \approx \text{linear} \;\&\; \Delta^2 y_t \approx \text{const.}$	Quadratic
$\Delta^1 y_t \approx \text{Gaussian curve}$	Logistic

where

$$\Delta^1 y_t = y_t - y_{t-1}$$

$$\Delta^2 y_t = \Delta^1 y_t - \Delta^1 y_{t-1}$$

Time series: Seasonal component

Assume that each period of time $t = 0, 1, 2, 3, \ldots$ consists of s_0 seasons.

For example, a year consists of $s_0 = 4$ quarters or $s_0 = 12$ months;

a week consists of $s_0 = 7$ days.

We thus assume that the time series

is of the form

$$Y_{ts} = T_t + S_s + \varepsilon_{ts} \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots \quad \text{and} \quad s = 1, 2, \ldots, s_0$$

where

- T_t is the trend component
- S_s is the seasonal component

Time series: Seasonal component

We here assume the **constant seasonality**.

That is, the time series is of the form

$$Y_{ts} = T_t + S_s + \varepsilon_{ts} \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots \quad \text{and} \quad s = 1, 2, \ldots, s_0$$

and the numbers $S_1, S_2, \ldots, S_{s_0} \in \mathbb{R}$ are such that

$$\sum_{s=1}^{s_0} S_s = 0$$

The trend component T_t is then assumed to be linear or polynomial, say.

Time series: Seasonal component

Assuming the constant seasonality, the model of the time series is written as

$$Y_{ts} = T_t + \gamma_0 + \gamma_2 x_2 + \gamma_3 x_3 + \cdots + \gamma_{s_0} x_{s_0}$$

where

$$\gamma_0 = S_1 \quad \text{and} \quad \gamma_2 = S_2 - S_1 \quad \gamma_3 = S_3 - S_1 \quad \ldots \quad \gamma_{s_0} = S_{s_0} - S_1$$

and

either

$$x_2 = x_3 = \cdots = x_{s_0} = 0 \quad \text{if} \quad s = 1$$

or

$$x_s = 1 \quad \text{and} \quad x_2 = \cdots = x_{s-1} = 0 = x_{s+1} = \cdots = x_{s_0} \quad \text{if} \quad s \in \{2, 3, \ldots, s_0\}$$

Time series: Seasonal component

If the trend is polynomial, say, we obtain:

$$Y_{ts} = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k + \gamma_2 x_2 + \gamma_3 x_3 + \cdots + \gamma_{s_0} x_{s_0}$$

The term $\gamma_0 = S_1$ is included in the constant (intercept) term β_0.

We have as above:

either

$$x_2 = x_3 = \cdots = x_{s_0} = 0 \qquad \text{if} \quad s = 1$$

or

$$x_s = 1 \quad \text{and} \quad x_2 = \cdots = x_{s-1} = 0 = x_{s+1} = \cdots = x_{s_0} \qquad \text{if} \quad s \in \{2, 3, \ldots, s_0\}$$

Autocorrelation of the random component

- Durbin-Watson test

Time series: The Classical Assumption

We here consider the time series of the form

$$Y_t = T_t + \varepsilon_t \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots$$

or

$$Y_{ts} = T_t + S_s + \varepsilon_{ts} \qquad \text{for} \quad t = 0, 1, 2, 3, \ldots \quad \text{and} \quad s = 1, 2, \ldots, s_0$$

with $S_1, S_2, \ldots, S_{s_0} \in \mathbb{R}$ such that $\sum_{s=1}^{s_0} S_s = 0$, where

- T_t is the trend component
- S_s is the seasonal component
- ε_t or ε_{ts} is the random component

Let us consider the first case $(Y_t = T_t + \varepsilon_t)$ only for simplicity.

Time series: The Classical Assumption

We then adopt the classical assumptions that

$$\boldsymbol{\varepsilon}: \Omega \to \mathbb{R}^n$$

is a random vector such that

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

so that

- $\mathrm{Var}(\varepsilon_t) = \sigma^2$ for $t = 1, 2, \ldots, n$ (homoskedasticity)
- $\mathrm{cov}(\varepsilon_s, \varepsilon_t) = 0$ if $s \neq t$ for $s, t = 1, 2, \ldots, n$ (no correlation)

The latter assumption is often violated in time series.

Time series: Autocorrelation

Consider the equation

$$\varepsilon_{t-1} = \rho\varepsilon_t + u_t \qquad \text{for} \quad t = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where $\rho \in \mathbb{R}$ and the random variables are $u_t \sim \mathcal{N}(0, \sigma^2)$ are independent.

If

- $\rho = 0$ then no autocorrelation is present
- $\rho > 0$ then positive autocorrelation AR(1) is present
- $\rho < 0$ then negative autocorrelation AR(1) is present

Our purpose is to test **the null hypothesis**

$$H_0: \quad \rho = 0$$

Time series: Durbin-Watson test

We have a sample

$$y_1, y_2, \dots, y_n$$

of observations of the variable Y for $t = 1, 2, \dots, n$.

By using the Linear Regression, we estimate the respective parameters $\beta_0, \beta_1, \dots, \beta_k$ and we calculate the corresponding theoretical values

$$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$$

and the residuals

$$e_t = y_t - \hat{y}_t \qquad \text{for} \quad t = 1, 2, \dots, n$$

Time series: Durbin-Watson test

We calculate the statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\text{RSS}} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

and formulate the null hypothesis

$$H_0: \quad \rho = 0$$

The alternative hypothesis is

— either

$$H_1: \quad \rho > 0$$

— or

$$H_1: \quad \rho < 0$$

Time series: Durbin-Watson test

Calculate the statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\mathrm{RSS}} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

and observe that

$$0 < d < 4$$

Since $\kappa_{n-t} = 4 - \kappa_t$, it holds for the quantile functions

$$d_{\mathrm{L}}(1-\alpha) = 4 - d_{\mathrm{U}}(\alpha) \qquad \text{and} \qquad d_{\mathrm{U}}(1-\alpha) = 4 - d_{\mathrm{L}}(\alpha)$$

Remark: The values of the quantile functions can be found

Time series: Durbin-Watson test

Calculate the statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\mathrm{RSS}} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

and calculate the estimate of the paired correlation coefficient

$$r = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

If $r > 0$ or $r < 0$, then the alternative hypothesis is $H_1: \rho > 0$ or $H_1: \rho < 0$,

Time series: Durbin-Watson test

Durbin-Watson test of the null hypothesis H_0: $\rho = 0$ against

the alternative hypothesis H_1: $\rho > 0$:

- Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.
- Calculate the statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\mathrm{RSS}} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- If $d \leq d_{\mathrm{L}}(\alpha)$, then reject the null hypothesis.
- If $d_{\mathrm{U}}(\alpha) \leq d$, then do not reject the null hypothesis.
- If $d_{\mathrm{L}}(\alpha) < d < d_{\mathrm{U}}(\alpha)$, then the test is indecisive.

Time series: Durbin-Watson test

Durbin-Watson test of the null hypothesis H_0: $\rho = 0$ against

the alternative hypothesis H_1: $\rho < 0$:

- Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.
- Calculate the statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\mathrm{RSS}} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- If $d_U(1-\alpha) \leq d$, then reject the null hypothesis.
- If $d \leq d_L(1-\alpha)$, then do not reject the null hypothesis.
- If $d_L(1-\alpha) < d < d_U(1-\alpha)$, then the test is indecisive.

Moving average

- Simple moving average
- Weighted moving average

Moving average

A **moving average** is a method of the synthetic approach to the trend analysis.

It consists in averaging a moving sample of consecutive observations

of the random variable Y.

The values smoothed in this way describe the sole trend contained in the

time series, i.e. the trend without the external factors.

The new series of the averages can be analysed then.

Simple moving average

Let a sample

$$y_1, y_2, \dots, y_n$$

of observations of the random variable Y for $t = 1, 2, \dots, n$ be given.

Choose the length m of the moving part of the time series.

We usually choose the length

$$m = 2p + 1 \qquad \text{for some} \quad p \in \left\{1, 2, \dots, \left\lfloor \frac{n-1}{2} \right\rfloor\right\}$$

i.e. an odd number.

Remark: If the seasonal component is assumed, then the length m is chosen

Simple moving average

Having chosen the length

$$m = 2p + 1$$

and having the sample

$$y_1, y_2, \dots, y_n$$

we consider the new time series

$$\bar{y}_{p+1}, \bar{y}_{p+2}, \dots, \bar{y}_{n-p}$$

of the moving averages

$$\bar{y}_\tau = \frac{1}{2p+1} \sum_{t=\tau-p}^{\tau+p} y_t \qquad \text{for} \quad \tau = p+1, p+2, \dots, n-p$$

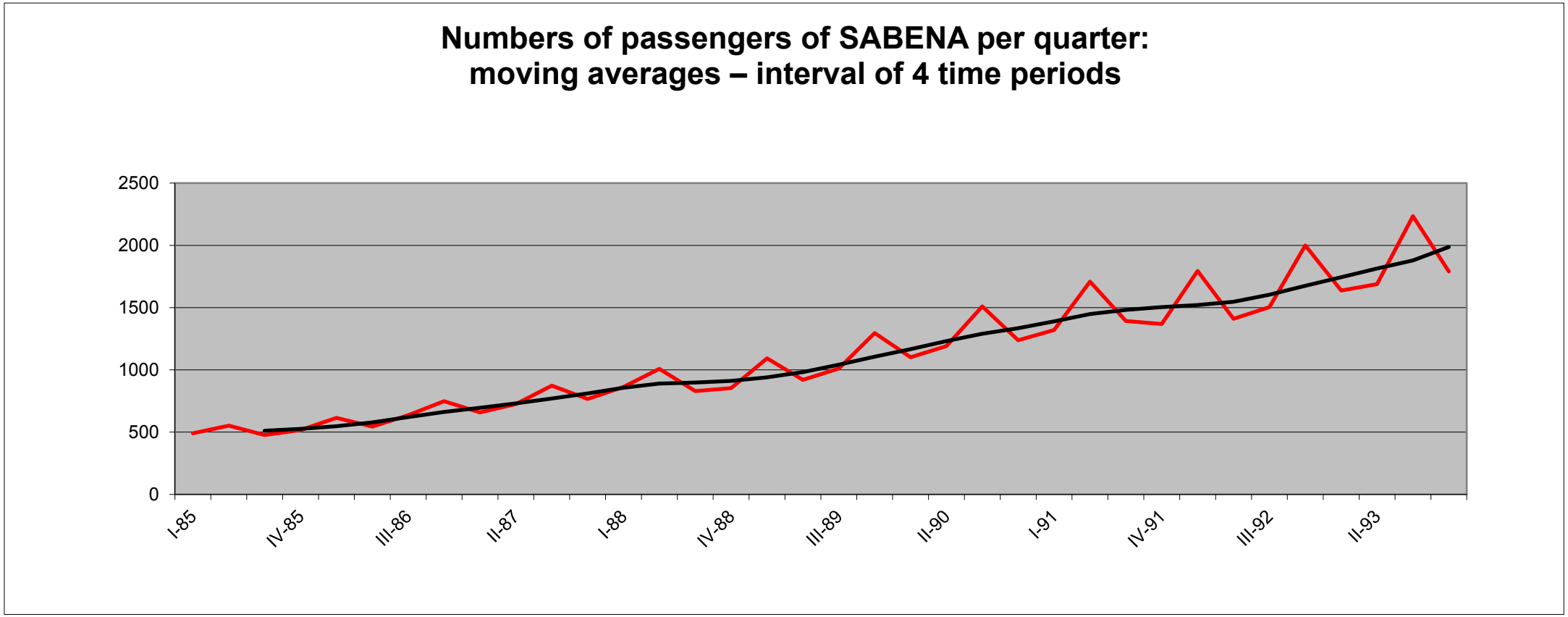
Simple moving average: Example

A time series and its moving averages of length $m = 5$:

t	1	2	3	4	5	6	7	8	9	10
y_t	34	40	37	42	45	47	44	51	52	58
average			39,6	42,2	43,0	45,8	47,8	50,4	52,0	56,0

t	11	12	13	14	15	16	17	18	19	20
y_t	55	64	59	66	68	62	72	75	72	77
average	57,6	60,4	62,4	63,8	65,4	68,6	69,8	71,6		

Simple moving average: Example



Moving average

In general, we choose:

- the length m of the moving window,
- the order $k \in \{1, 2, \ldots, m-1\}$ of the approximating polynomial.

We then approximate each segment

$$y_{\tau-p}, \cdots, y_{\tau}, \cdots, y_{\tau+p}$$

by a polynomial

$$y_{\tau+x} \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k \qquad \text{for} \quad x = -p, \ldots, 0, \ldots, +p$$

by using the Least Squares Method (i.e. Multiple Linear Regression)

Moving average

That is, we calculate

$$\sum_{x=-p}^{+p} \left(y_{\tau+x} - b_0 - b_1 x - b_2 x^2 - \cdots - b_k x^k\right)^2 \longrightarrow \min$$

and let

$$\bar{y}_\tau = b_0$$

for each $\tau = p+1, p+2, \ldots, n-p$.

- The choice $k = 1$ (approximation by a linear polynomial) yields the simple moving average, as above.
- The choice $k = 2, 3, \ldots$ (approximation by a quadratic, cubic, … polynomial) yields other centred moving averages.

Moving average: Weighted moving average

Remark: Instead of the simple moving average

$$\bar{y}_\tau = \frac{1}{2p+1} \sum_{t=\tau-p}^{\tau+p} y_t \qquad \text{for} \quad \tau = p+1, p+2, \ldots, n-p$$

we can also consider a weighted moving average

$$\bar{y}_\tau = \frac{1}{\sum_{x=-p}^{+p} w_x} \sum_{x=-p}^{+p} w_x y_{\tau+x} \qquad \text{for} \quad \tau = p+1, p+2, \ldots, n-p$$

where $w_{-p}, \ldots, w_0, \ldots, w_{+p} \geq 0$ are weights.