Time Series Analysis and Forecasting with Application in R

Dr. Raju Maiti

Senior Research Fellow
Health Services and Systems Research
Duke-NUS Medical School, Singapore

July 31, 2020

What is time series?

- A time series is a sequence of data points $\{X_t : t = 1, 2, \ldots, n\}$ measured at successive time intervals. Here t indicates the time at which X_t is observed.
- Example 1: Number of births per month in a city.
- Example 2: Weekly number of deaths due to Acute Japanese Encephalitis Syndrome (AJEC) in North Bengal, India.
- Example 3: Monthly cases of dengue observed in Delhi, India.
- A time series generally reflects the fact that observations close together in time are more closely related than observations further apart.

DukeNUS Medical School

Example 1: Monthly number of births

Table 1: Monthly number of births in New York city during 1946-59

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1946	26.663	23.598	26.931	24.740	25.806	24.364	24.477	23.901	23.175	23.227	21.672	21.870
1947	21.439	21.089	23.709	21.669	21.752	20.761	23.479	23.824	23.105	23.110	21.759	22.073
1948	21.937	20.035	23.590	21.672	22.222	22.123	23.950	23.504	22.238	23.142	21.059	21.573
1949	21.548	20.000	22.424	20.615	21.761	22.874	24.104	23.748	23.262	22.907	21.519	22.025
1950	22.604	20.894	24.677	23.673	25.320	23.583	24.671	24.454	24.122	24.252	22.084	22.991
1951	23.287	23.049	25.076	24.037	24.430	24.667	26.451	25.618	25.014	25.110	22.964	23.981
1952	23.798	22.270	24.775	22.646	23.988	24.737	26.276	25.816	25.210	25.199	23.162	24.707
1953	24.364	22.644	25.565	24.062	25.431	24.635	27.009	26.606	26.268	26.462	25.246	25.180
1954	24.657	23.304	26.982	26.199	27.210	26.122	26.706	26.878	26.152	26.379	24.712	25.688
1955	24.990	24.239	26.721	23.475	24.767	26.219	28.361	28.599	27.914	27.784	25.693	26.881
1956	26.217	24.218	27.914	26.975	28.527	27.139	28.982	28.169	28.056	29.136	26.291	26.987
1957	26.589	24.848	27.543	26.896	28.878	27.390	28.065	28.141	29.048	28.484	26.634	27.735
1958	27.132	24.924	28.963	26.589	27.931	28.009	29.229	28.759	28.405	27.945	25.912	26.619
1959	26.076	25.286	27.660	25.951	26.398	25.565	28.865	30.000	29.261	29.012	26.992	27.897

DukeNUS Medical School

Example 1: Monthly number of births

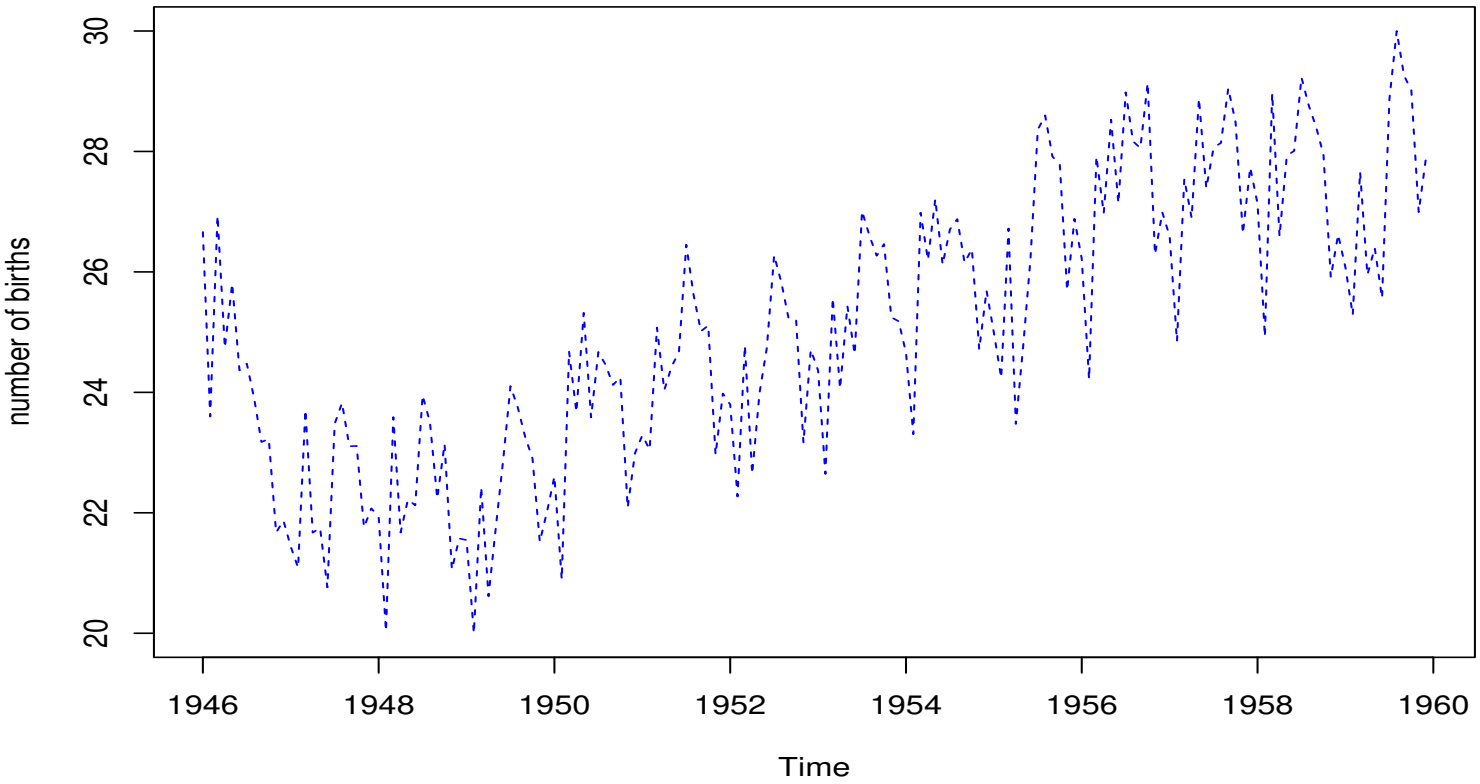


Figure 1: Monthly number of births.

Simple descriptive techniques to model a time series

- In general, a time series can be decomposed into four components: trend (T), seasonal (S), cyclical (C) and Residual or random (R), i.e.,

Additive case

$$X_t = T_t + S_t + Y_t$$

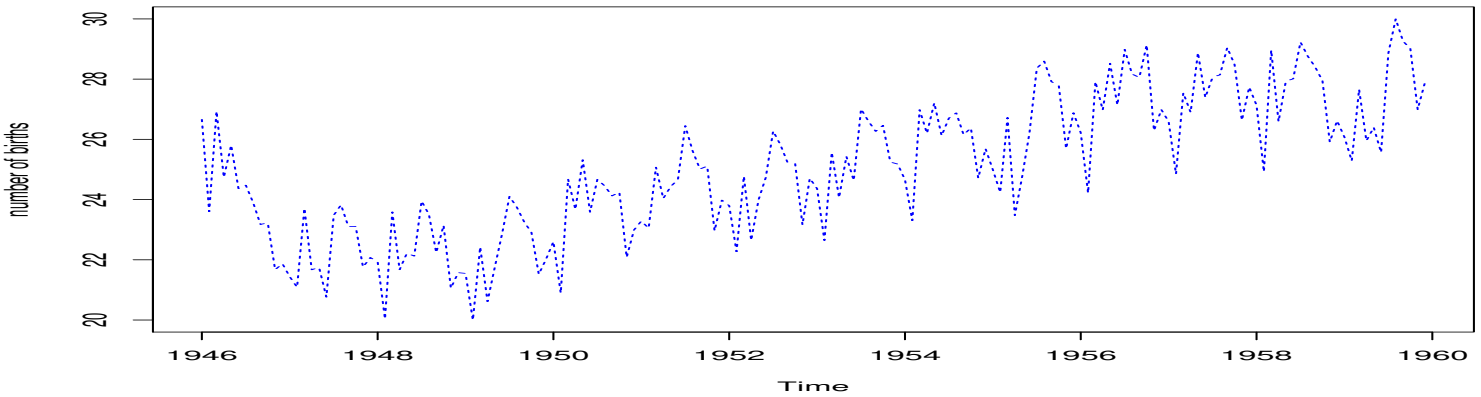
Multiplicative case

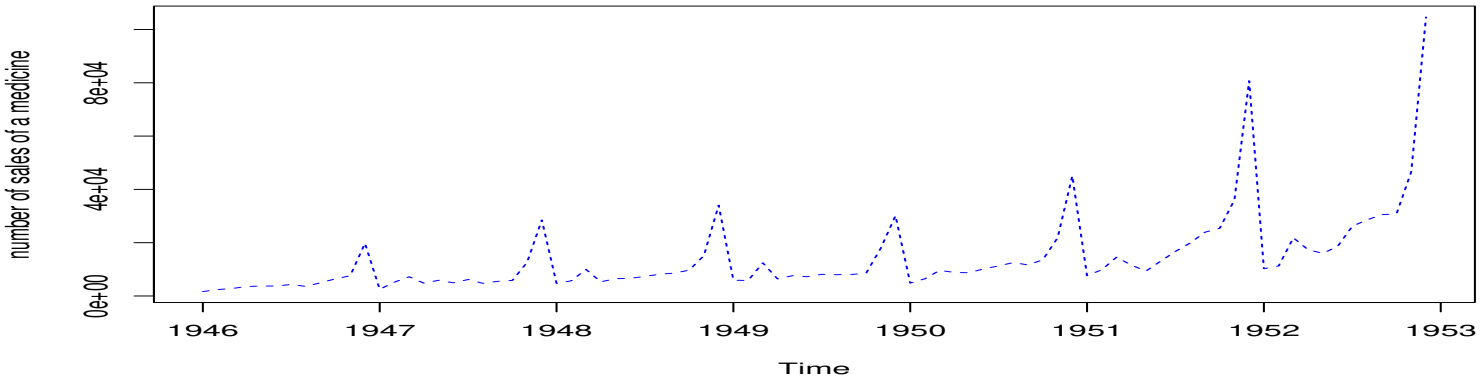
$$X_t = T_t \times S_t \times Y_t$$

$$\log X_t = \log T_t + \log S_t + \log Y_t$$

DukeNUS Medical School

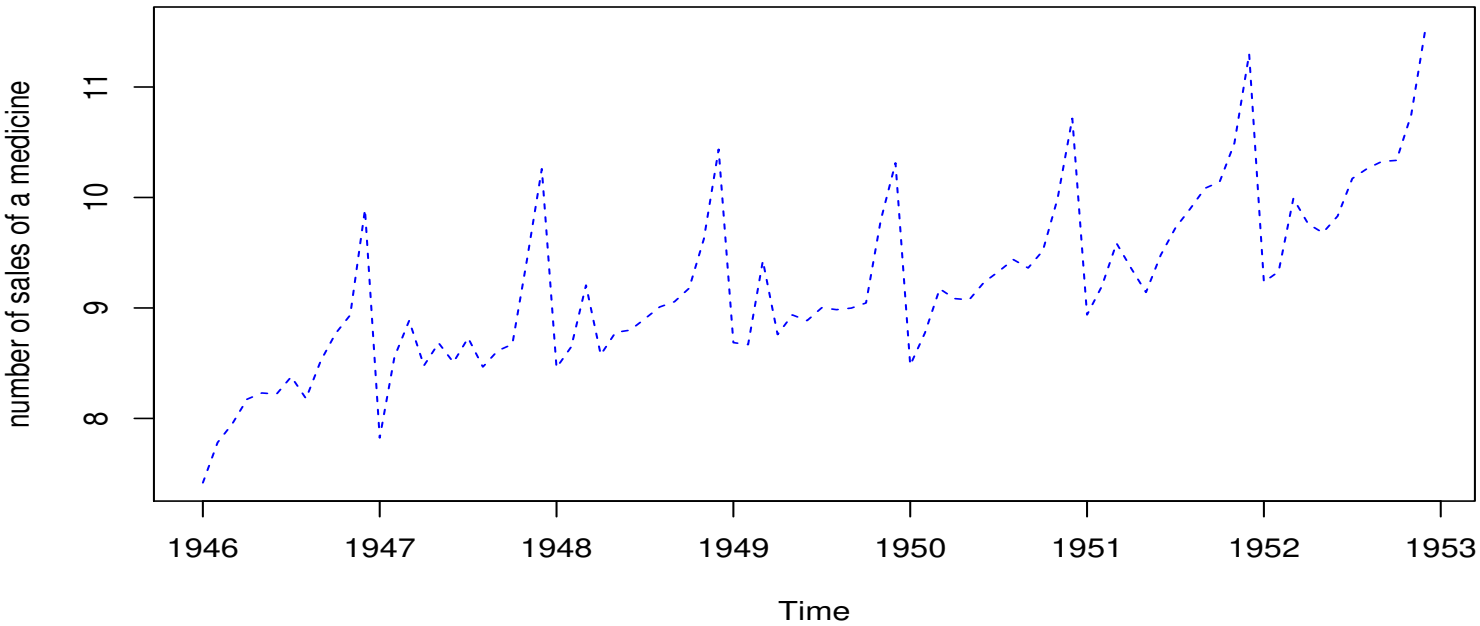
Additive and Multiplicative cases





DukeNUS Medical School

log transformation to make it into additive model



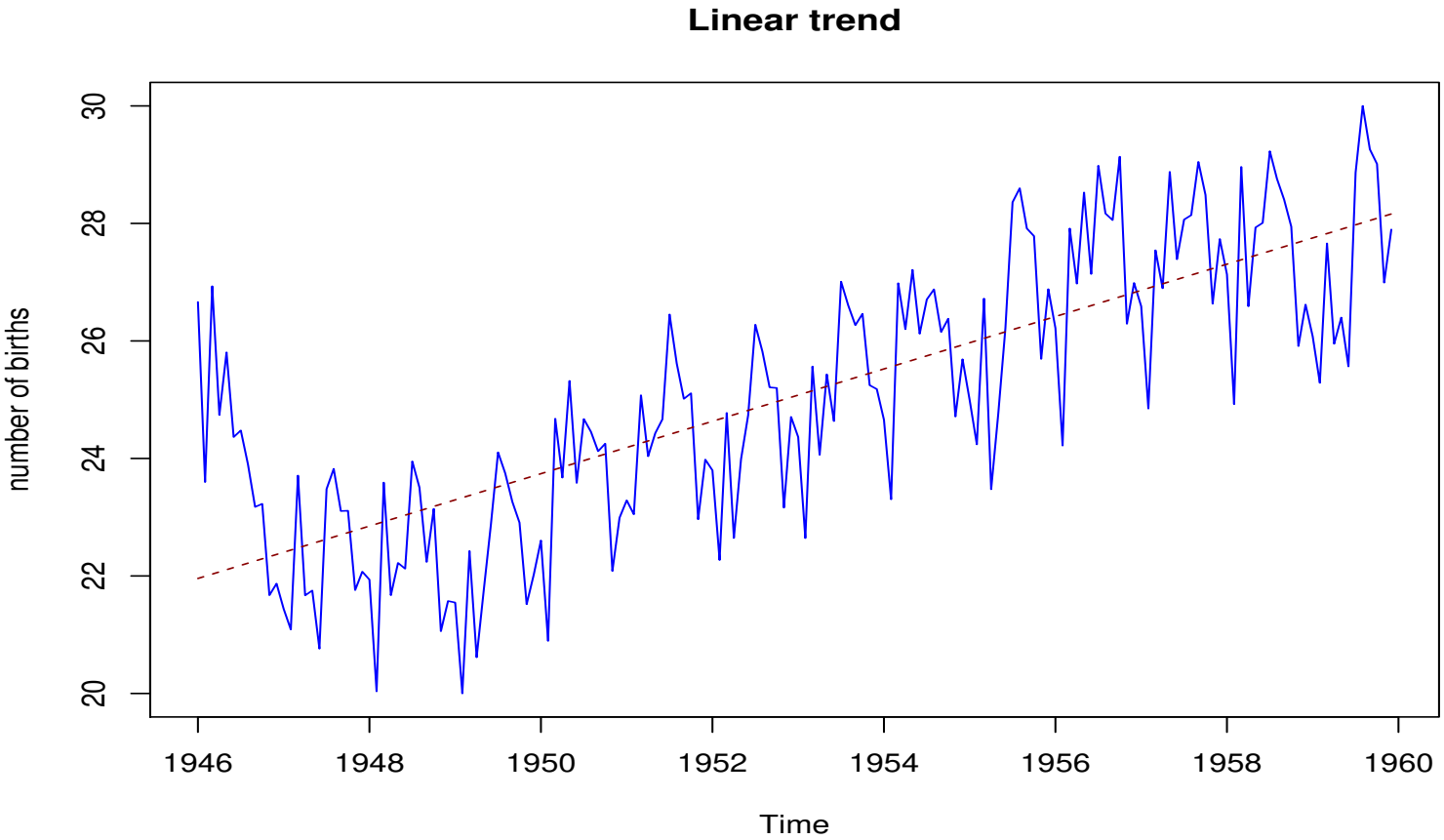
DukeNUS Medical School

Estimating trend component

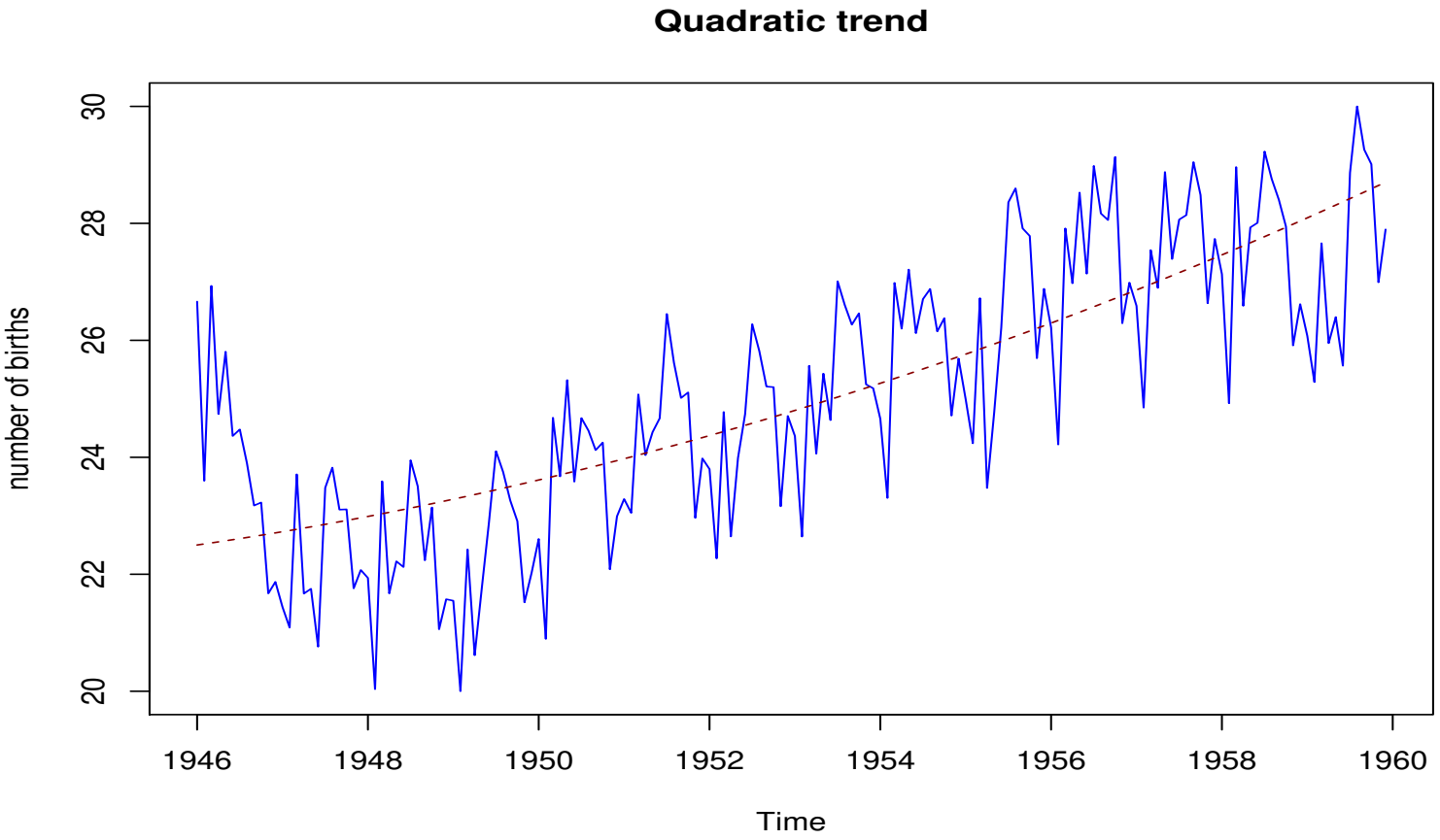
To estimate trend, several methods are there

- Linear trend : $T_t = a + bt$
- Quadratic trend : $T_t = a + bt + ct^2$
- Moving average : $T_t = \frac{1}{2k+1} \sum_{i=-k}^{k} X_{t+i}$

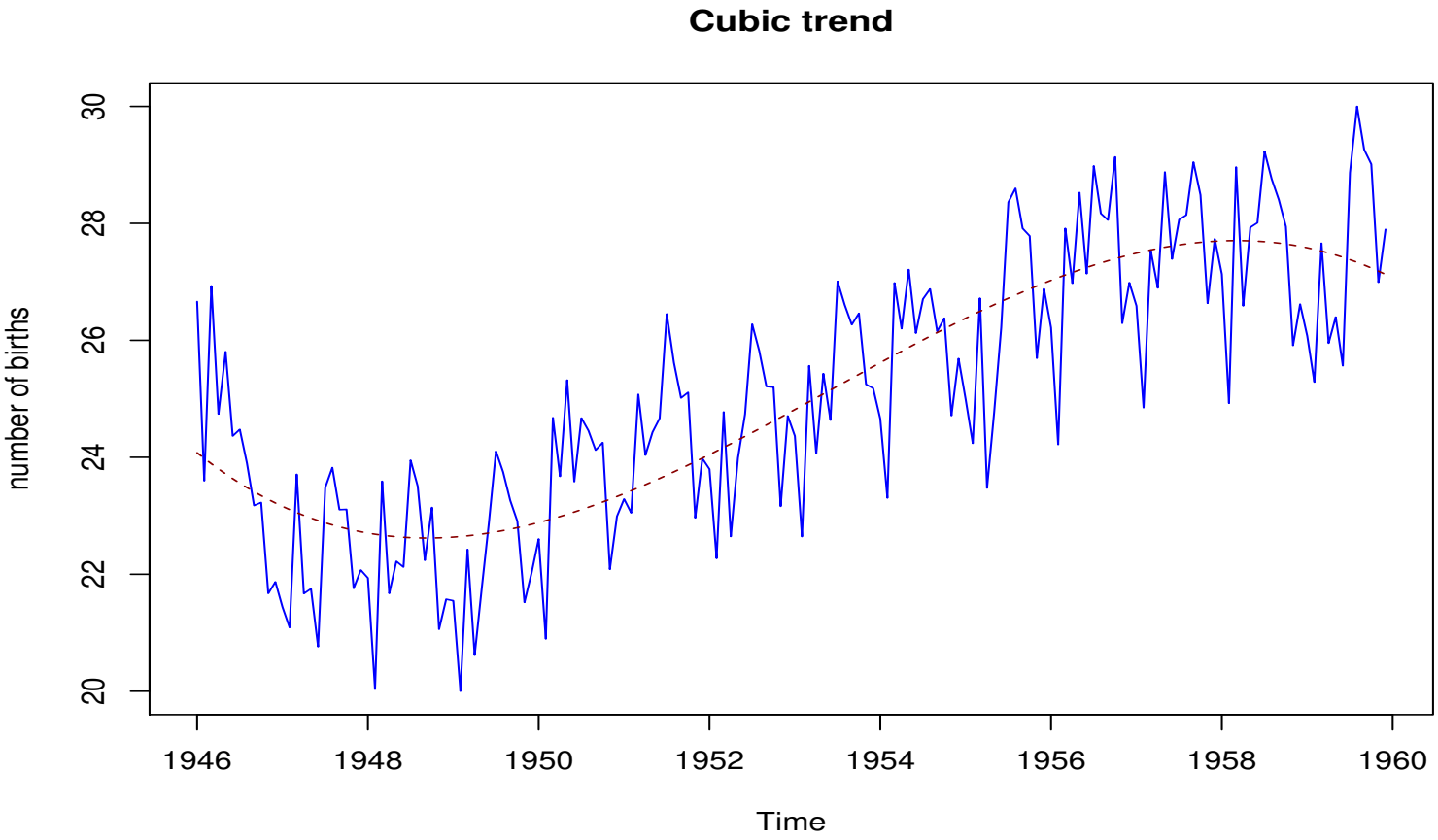
Linear trend



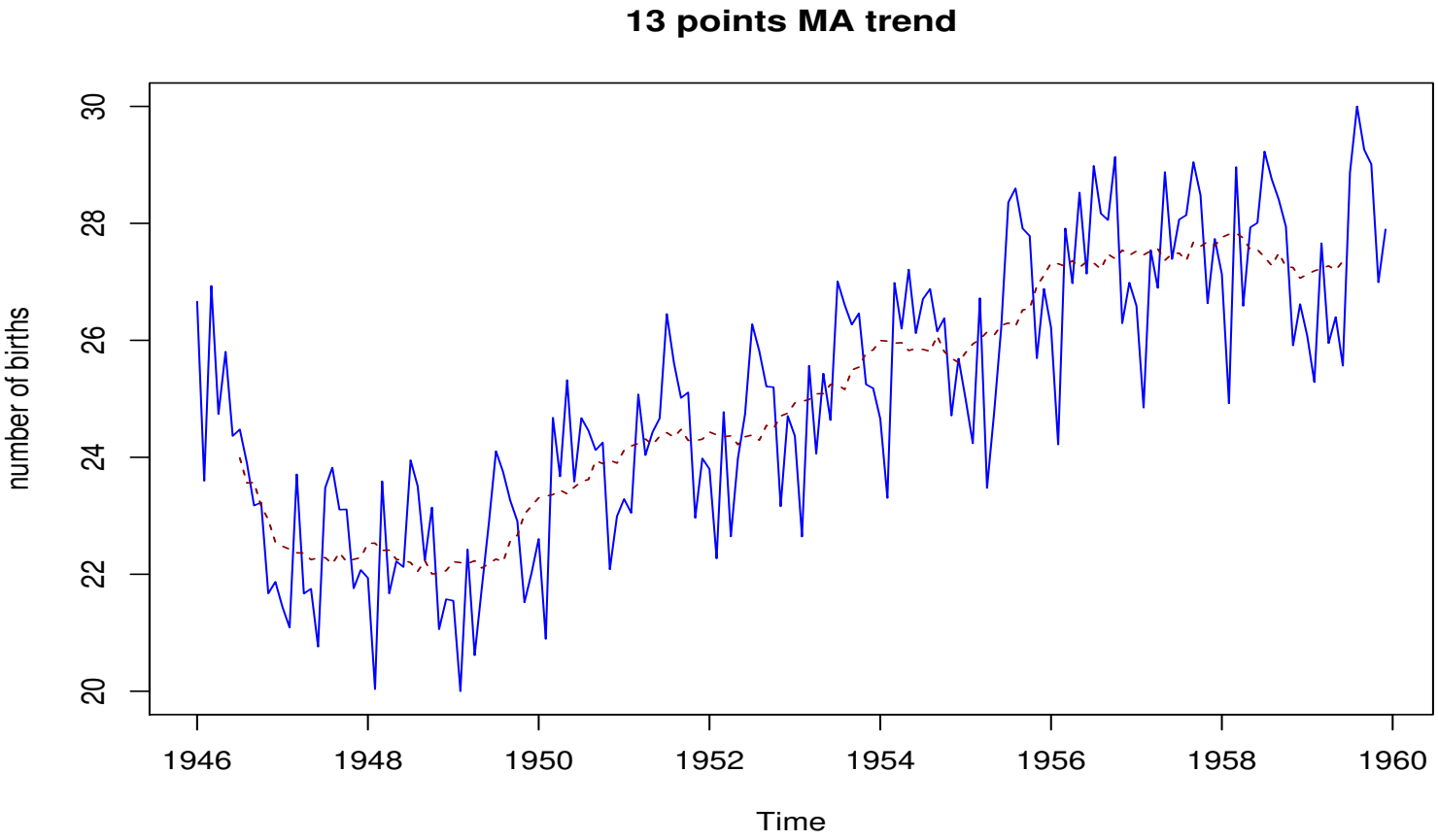
Quadratic trend



Cubic trend


Cubic trend
number of births
20
22
24
26
28
30
1946
1948
1950
1952
1954
1956
1958
1960
Time

Moving average trend



Estimating seasonal component

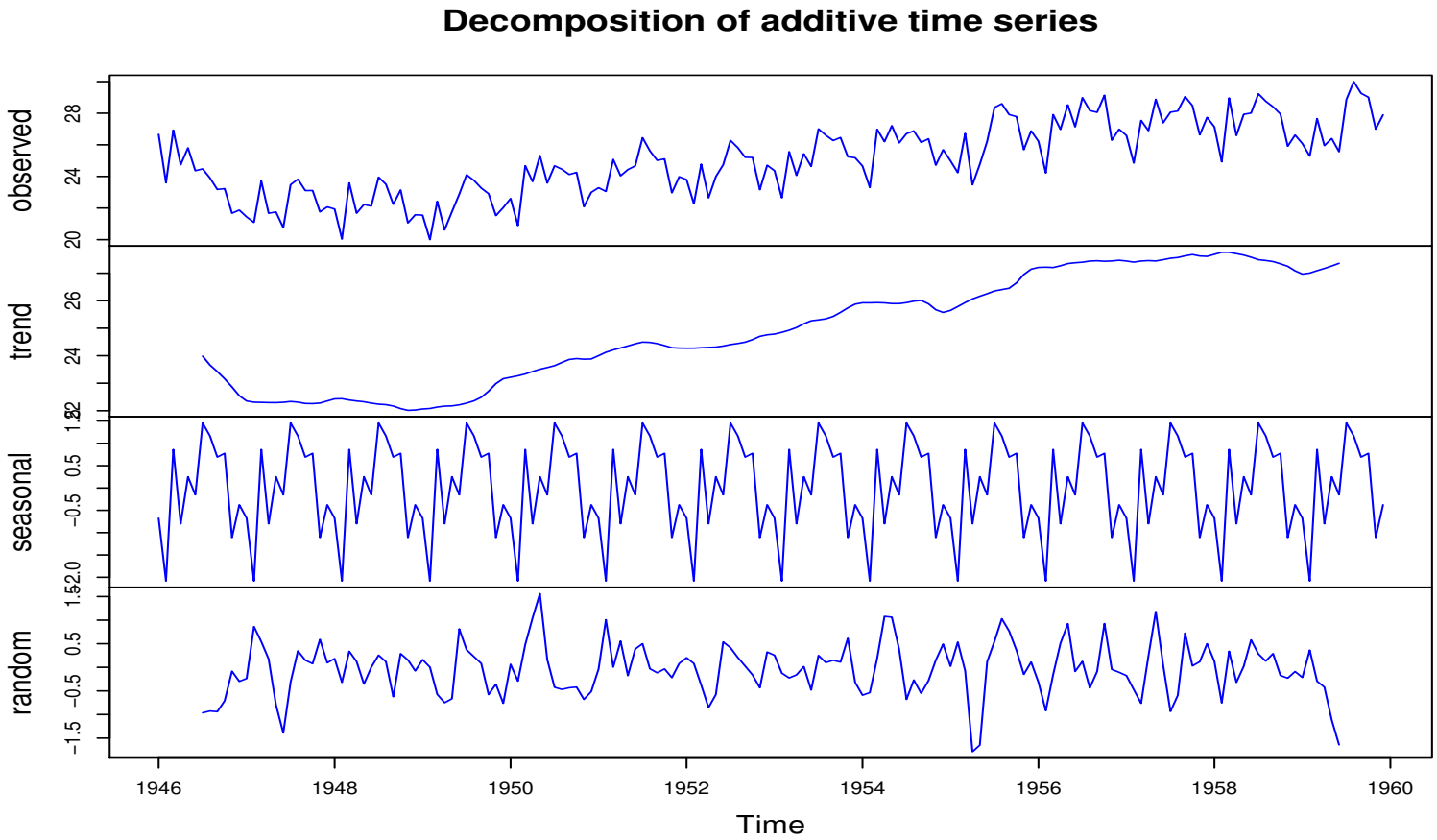
- $X_t - T_t = S_t + Y_t$
- To estimate seasonal component, it assumes that $S_t = S_{t-d}$, where $d = 4$ if the data is obtained quarterly; $d = 12$ if the data is obtained monthly. Under this assumption, possible method is
  - moving-average method

Example 1

Table 2: Computation of seasonal component

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1946	-0.64	-1.45	0.78	-0.86	-0.39	-1.21	0.49	0.24	-0.25	0.07	-1.19	-0.68
1947	-0.91	-1.22	1.41	-0.63	-0.54	-1.54	1.14	1.51	0.84	0.85	-0.52	-0.28
1948	-0.49	-2.40	1.20	-0.68	-0.10	-0.15	1.71	1.28	0.07	1.06	-0.95	-0.45
1949	-0.52	-2.08	0.29	-1.55	-0.41	0.66	1.83	1.39	0.77	0.20	-1.47	-1.14
1950	-0.61	-2.38	1.34	0.25	1.81	0.01	1.03	0.70	0.26	0.36	-1.79	-0.89
1951	-0.71	-1.07	0.87	-0.25	0.08	0.23	1.96	1.13	0.58	0.74	-1.33	-0.30
1952	-0.47	-2.00	0.49	-1.66	-0.33	0.39	1.87	1.37	0.72	0.61	-1.54	-0.05
1953	-0.42	-2.21	0.64	-0.96	0.27	-0.63	1.71	1.26	0.84	0.89	-0.49	-0.70
1954	-1.27	-2.62	1.05	0.28	1.31	0.23	0.78	0.90	0.14	0.49	-0.96	0.11
1955	-0.66	-1.55	0.79	-2.59	-1.40	-0.03	2.01	2.19	1.46	1.13	-1.26	-0.27
1956	-0.99	-3.00	0.71	-0.29	1.18	-0.24	1.58	0.73	0.60	1.70	-1.15	-0.48
1957	-0.85	-2.55	0.10	-0.56	1.43	-0.10	0.52	0.57	1.42	0.81	-0.99	0.12
1958	-0.55	-2.84	1.20	-1.12	0.27	0.43	1.74	1.30	0.98	0.60	-1.34	-0.47
1959	-0.89	-1.72	0.57	-1.22	-0.86	-1.80	1.85	0.88	1.51	0.89	-0.89	-0.15
	-0.73											

Seasonal component for monthly birth data



Residual analysis

- In all of the previous slides, several descriptive methods were discussed to identify the macroscopic components like trend and seasonality of a time series.
- Now we assume that this preliminary analysis has been completed and we focus on analyzing the residual part R_t for microscopic structure.
- To model the residual part, many well known time series models are available, e.g.,
  - Autoregressive of order p, AR(p)
  - Moving average of order q, MA(p)
  - Autoregressive and moving average, ARMA(p, q)
  - Autoregressive integrated moving average, ARIMA(p, d, q)

DukeNUS Medical School

AR(p) process

- A residual process $\{Y_t\}$ is said to follow an AR(p) process if it can be written as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

where $\varepsilon_t \sim WN(0, \sigma^2)$

- In particular, an AR(1) process can be written as

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

where $\varepsilon_t \sim WN(0, \sigma^2)$

DukeNUS Medical School

MA(q) process

- A residual process $\{Y_t\}$ is said to follow a MA(q) process if it can be written as

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where $\varepsilon_t \sim WN(0, \sigma^2)$

- In particular, an MA(1) process can be written as

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

where $\varepsilon_t \sim WN(0, \sigma^2)$

ARMA(p, q) process

- A residual process $\{Y_t\}$ is said to follow an ARMA(p, q) process if it can be written as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where $\varepsilon_t \sim WN(0, \sigma^2)$

- In particular, an ARMA(1, 1) process can be written as

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

where $\varepsilon_t \sim WN(0, \sigma^2)$

ARIMA(p, d, q) process

- An ARIMA model is characterized by three terms: p, d, and q
  - where d is the number of differencing required to make the time series stationary.
  - If $d = 1$, $Y_t^* = Y_t - Y_{t-1}$ follows ARMA(p, q)
  - If $d = 2$, $Y_t^{**} = Y_t^* - Y_{t-1}^* = Y_t - 2Y_{t-1} + Y_{t-2}$ follows ARMA(p, q)

DukeNUS Medical School

Auto-correlation function (ACF) and partial ACF (PACF)

Autocorrelation function (ACF) between Y_t and its lag value Y_{t-h} is defined as

$$\rho(h) = Cor(Y_t, Y_{t-h})$$

and partial ACF (PACF) is defined as

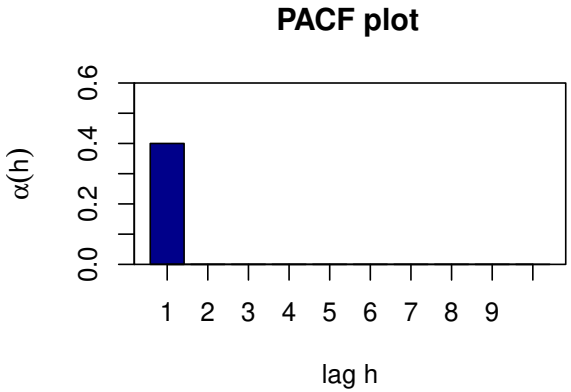
$$\alpha(h) = Cor(Y_t - f(Y_{t-1}, \cdots, Y_{t-h+1}), Y_{t-h} - g(Y_{t-1}, \cdots, Y_{t-h+1}))$$

where $f(\cdot)$ and $g(\cdot)$ are some suitable linear regression functions of Y_t and Y_{t-h} on $Y_{t-1}, Y_{t-2}, \cdots, Y_{t-h+1}$ respectively.

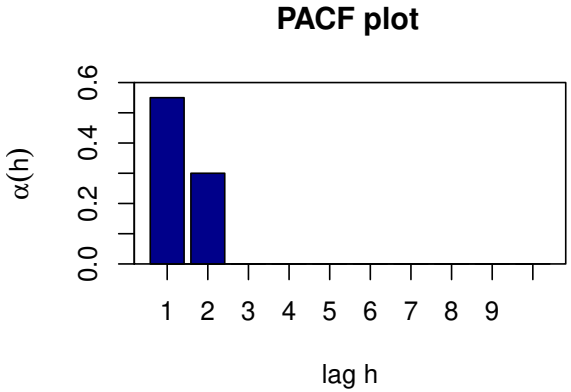
ACF and PACF

- ACF and PACF can be used to select the order of an ARMA(p, q) process.
  - For example, if the PACF of order one i.e., $\alpha(1) \neq 0$ and $\alpha(h) = 0$ for $h > 1$, then the time series process $\{Y_t\}$ might be an AR(1) process.
  - Similarly, if the ACF of order one i.e., $\rho(1) \neq 0$ and $\rho(h) = 0$ for $h > 1$, then the time series process $\{Y_t\}$ might be an MA(1) process.

ACF and PACF



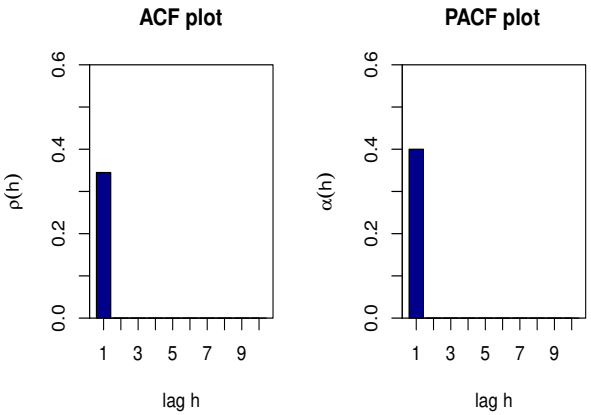
$\Longrightarrow \quad Y_t = \phi Y_{t-1} + \varepsilon_t$, possible indication of an AR(1) process



$\Longrightarrow \quad Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$, possible indication of an AR(2) process

DukeNUS Medical School

ACF and PACF



$$\Longrightarrow \quad Y_t = \phi Y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

possible indication of an ARMA(1,1) process

- Furthermore, ACF can be used to obtain the Yule-Walker estimates (a method of moments estimation) of an ARMA process.

Forecasting through AR(p) process: Iterative algorithm

- One-step ahead forecast:

$$\hat{Y}_{t+1|t} = \phi_1 Y_t + \phi_2 Y_{t-1} + \cdots + \phi_p Y_{t-p+1}$$

- Two-step ahead forecast:

$$\begin{aligned}
\hat{Y}_{t+2|t} &= \phi_1 Y_{t+1} + \phi_2 Y_t + \cdots + \phi_p Y_{t-p+2} \\
&= \phi_1 \hat{Y}_{t+1|t} + \phi_2 Y_t + \cdots + \phi_p Y_{t-p+2} \\
&= (\phi_1^2 + \phi_2) Y_t + (\phi_1\phi_2 + \phi_3) Y_{t-1} + \cdots \\
&\quad + (\phi_1\phi_{p-1} + \phi_p) Y_{t-p+2} + \phi_1\phi_p Y_{t-p+1}
\end{aligned}$$

Thank You