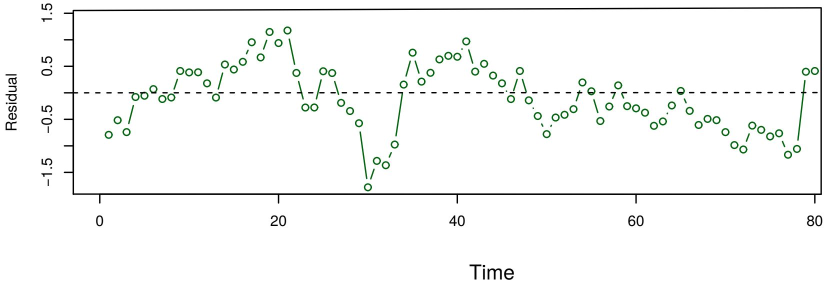
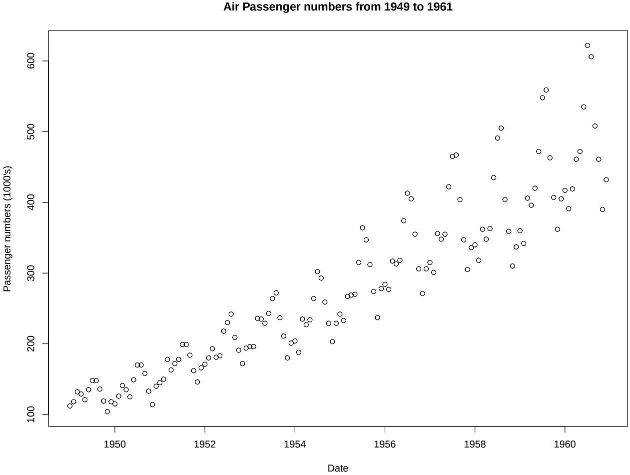
Time Series Analysis 1
DS351

Why can't we use linear regression



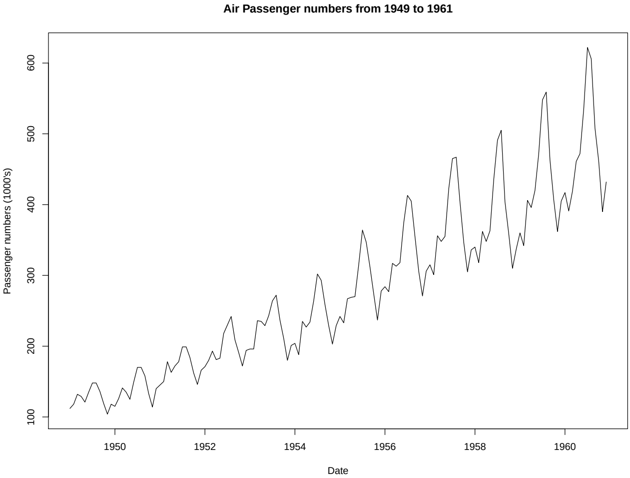
Error terms are correlated.

Why can't we use linear regression



Variance of the errors increases with time.

Why can't we use linear regression



Seasonality, which implies non-linearity!

Analyzing Time Series

Notations

Time series is often denoted by

$$\ldots \quad Y_{t-2} \quad Y_{t-1} \quad Y_t \quad Y_{t+1} \quad Y_{t+2} \quad \ldots$$

time index

Notations

Time series is often denoted by

$$\ldots \quad Y_{t-2} \quad Y_{t-1} \quad Y_t \quad Y_{t+1} \quad Y_{t+2} \quad \ldots$$

time index

Lag is an amount of time passed.

Example: lag 5 of Y_t is y_{t-5}.

Stationarity

A time series is **stationary** if **its statistical properties do not change over time**.

Stationarity

A time series is **stationary** if **its statistical properties do not change over time**.

What are statistical properties?

- mean
- variance
- covariance
- etc.

Stationarity

More precise definition:

A time series is stationary if for all lags k,
the distribution of $Y_t, Y_{t-1}, \ldots, Y_{t-k}$ does not depend on k.

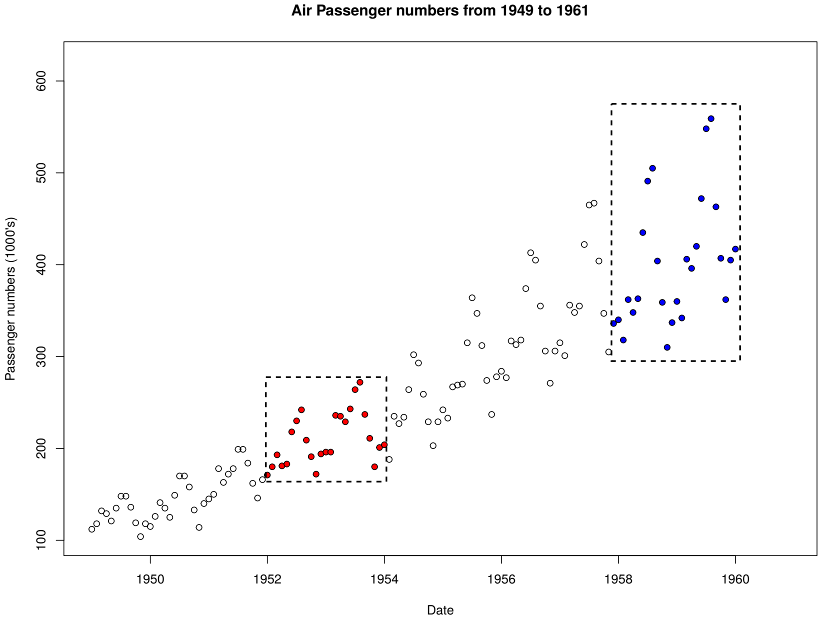
Stationarity

More precise definition:

$\boxed{\text{A time series is stationary if for all lags } k, \text{ the distribution of } Y_t, Y_{t-1}, \ldots, Y_{t-k} \text{ does not depend on } k.}$

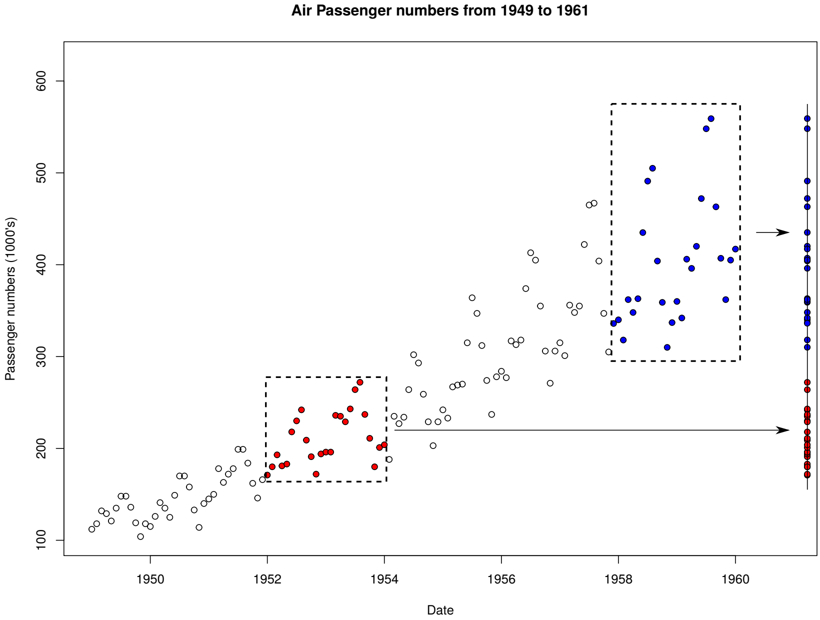
- we usually don't know the distribution of these variables.
- It is usually easier to detect that a time series is not stationary by looking at its plot.

Example



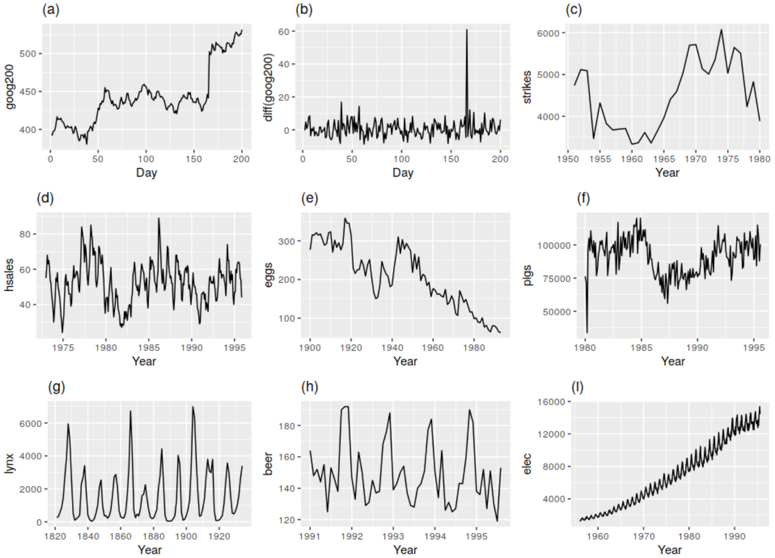
Air Passenger numbers from 1949 to 1961
Passenger numbers (1000's)
100
200
300
400
500
600
1950
1952
1954
1956
1958
1960
Date

Example



Air Passenger numbers from 1949 to 1961
Passenger numbers (1000's)
100
200
300
400
500
600
1950
1952
1954
1956
1958
1960
Date

Examples

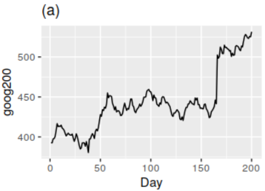


(a)
goog200
Day
(b)
diff(goog200)
Day
(c)
strikes
Year
(d)
hsales
Year
(e)
eggs
Year
(f)
pigs
Year
(g)
lynx
Year
(h)
beer
Year
(I)
elec
Year

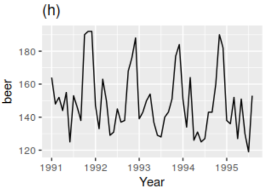
Find stationarity from the plot

In summary, a time series is **not** stationary if there is

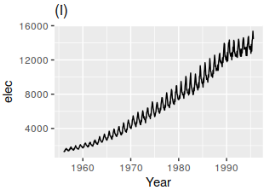
- trend
- seasonality
- increase/decrease in variance



has trend



has seasonality



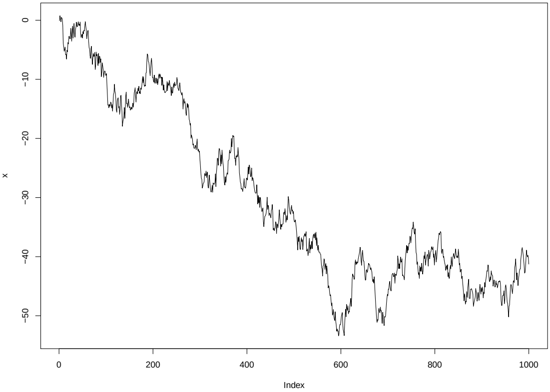
has both

Random walk

Random walk is a simple non-stationary process.

$$y_t = y_{t-1} + \epsilon_t.$$

where ϵ_t is a *white noise* with zero mean and variance σ^2 e.g. $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.



Differencing

From the random walk

$$y_t = y_{t-1} + \epsilon_t,$$

which is not stationary, we can transform it into

$$z_t = y_t - y_{t-1} = \epsilon_t.$$

Now, z_t is stationary.

Differencing

From the random walk

$$y_t = y_{t-1} + \epsilon_t,$$

which is not stationary, we can transform it into

$$z_t = y_t - y_{t-1} = \epsilon_t.$$

Now, z_t is stationary.

In general, we try to make a stationary time series by transforming $z_t = y_t - y_{t-1}$. This is known as **differencing**.

Second-order differencing

Second-order differencing
If differencing is not enough to make a time series stationary, we do it twice.

$$z_t' = z_t - z_{t-1} = z_t - 2z_{t-1} + z_{t-2}.$$

Second-order differencing

Second-order differencing
If differencing is not enough to make a time series stationary, we do it twice.

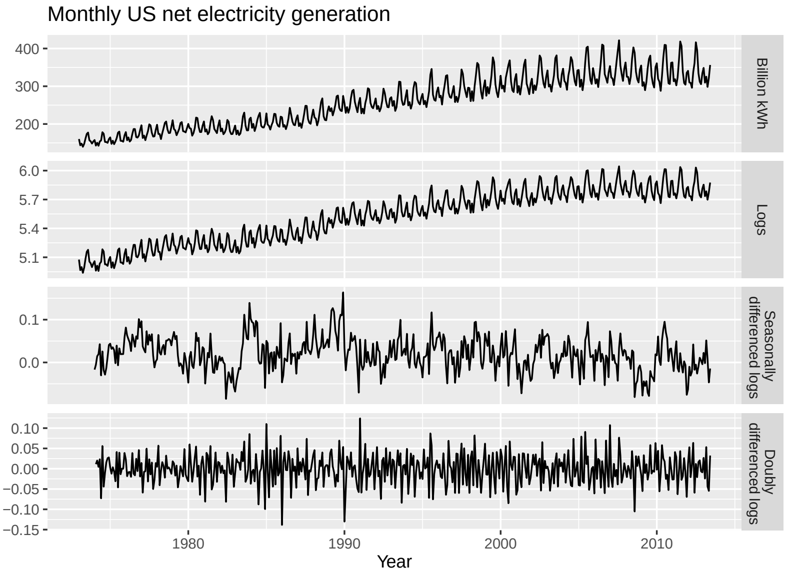
$$z'_t = z_t - z_{t-1} = z_t - 2z_{t-1} + z_{t-2}.$$

Seasonal differencing
To remove seasonality, we take the difference between an observation and the previous observation from the same season.

$$z_t = y_t - y_{t-m},$$

where m is the length of seasonality e.g. $m = 12$ for annual seasonality.

Example

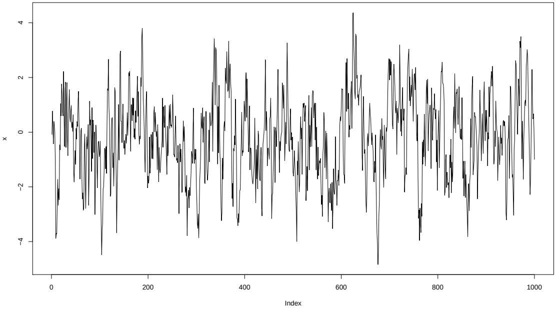


Monthly US net electricity generation
400
300
200
Billion kWh
6.0
5.7
5.4
5.1
Logs
0.1
0.0
Seasonally differenced logs
0.10
0.05
0.00
−0.05
−0.10
−0.15
Doubly differenced logs
1980
1990
2000
2010
Year

Unit root

The random walk is an example of a time series that has a **unit root**: the value of α in $y_t = \alpha y_{t-1} + \epsilon_t$ is $\alpha = 1$.

In general, **a time series is non-stationary if it has a unit root**.



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

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

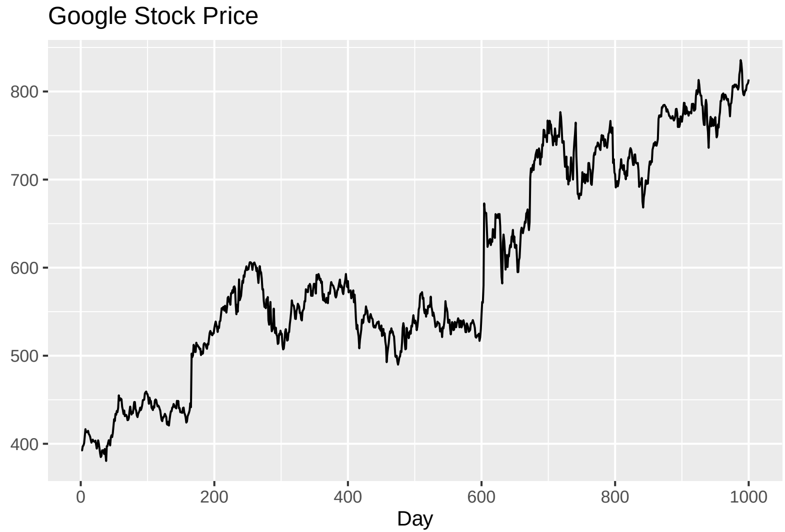
This is a test to see if we want differencing.

KPSS test

$$H_0 : y_t \text{ is stationary}$$
$$H_a : y_t \text{ has a unit root}$$

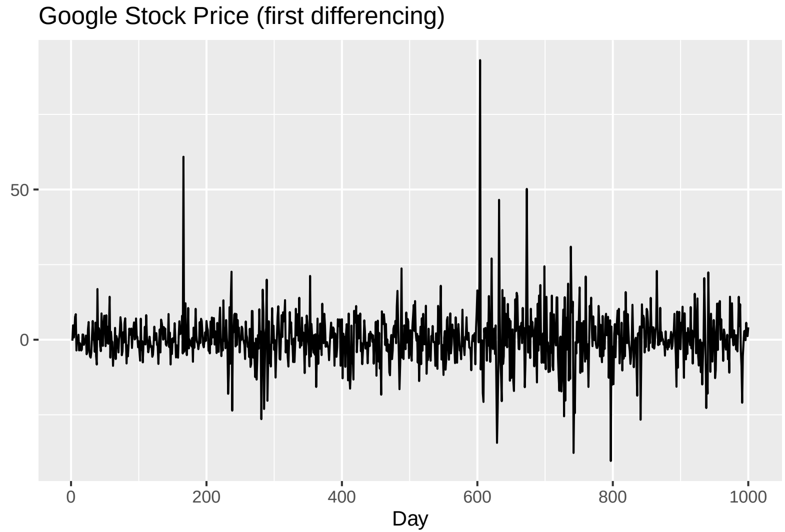
If H_0 is rejected ($p-$value < 0.05) then differencing y_t is required.

Example



$p-$value < 0.01

Example



$$p\text{-value} > 0.1$$

Autocorrelation

Autocorrelation measures the linear relationship between a time series and its *lagged values*.

$\boldsymbol{Y}_t$	y_1	$\dots$	y_{k+1}	$\dots$	y_t	$\dots$	y_T		
$\boldsymbol{Y}_{t-k}$		$\rightarrow$	y_1	$\dots$	y_{t-k}	$\dots$	y_{T-k}	$\dots$	$\rightarrow$

Autocorrelation

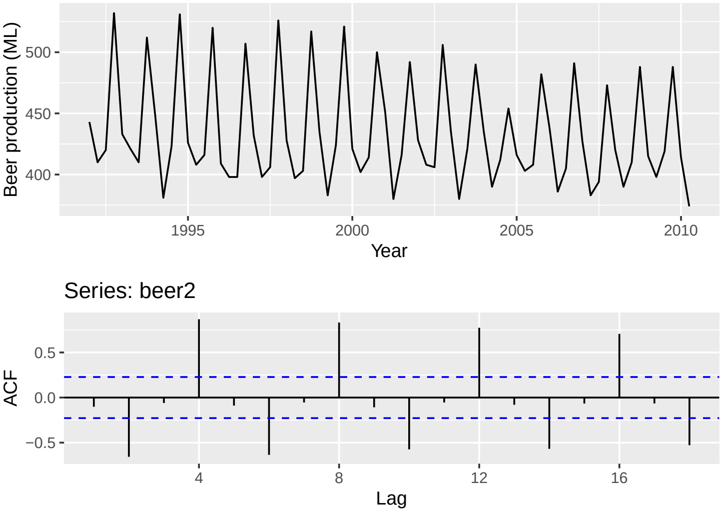
Autocorrelation measures the linear relationship between a time series and its *lagged values*.

$\boldsymbol{Y}_t$	y_1	$\dots$	y_{k+1}	$\dots$	y_t	$\dots$	y_T		
$\boldsymbol{Y}_{t-k}$		$\rightarrow$	y_1	$\dots$	y_{t-k}	$\dots$	y_{T-k}	$\dots$	$\rightarrow$

The autocorrelation at lag k, r_k can be written as

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y - \bar{y})^2}$$

Beer production

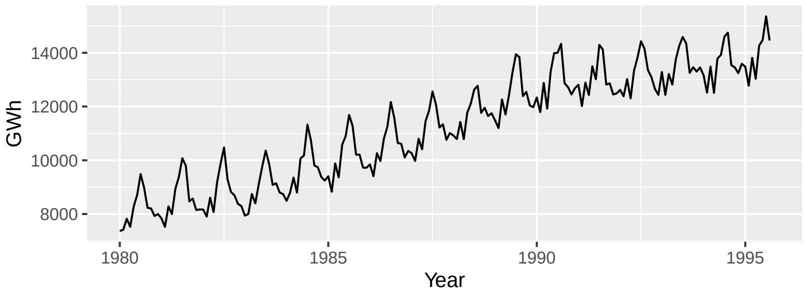


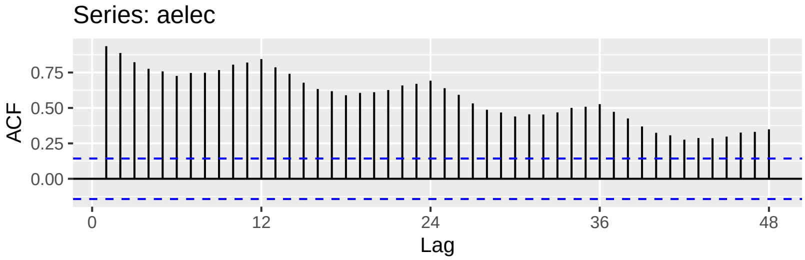
The values beyond blue lines are significantly different than zero.

Trend and seasonality in ACF plots

- When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.
- When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.

Australian electricity demand





Partial autocorrelation function (PACF)

Partial autocorrelation function measures the part of the correlation that has not been explained by the earlier lags.