MONASH University

MONASH BUSINESS SCHOOL

Forecasting using R

Rob J Hyndman

1.1 Time series graphics

Outline

Time series in R

- Time series consist of sequences of observations collected over time.
- We will assume the time periods are equally spaced.

Time series examples

- Daily IBM stock prices
- Monthly rainfall
- Annual Google profits
- Quarterly Australian beer production

Time series in R

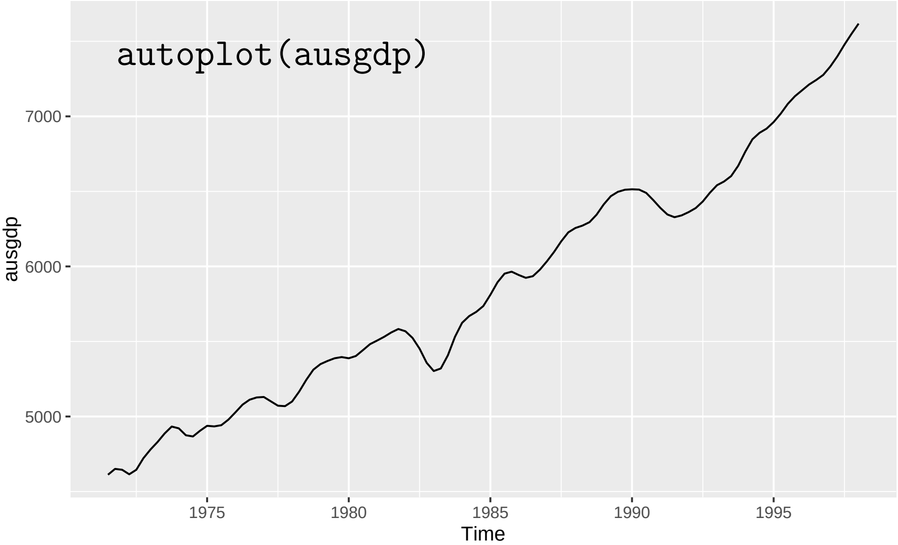
```
ausgdp <- ts(scan("gdp.dat"),frequency=4,
                              start=1971+2/4)
```

- Class: "ts"
- Print and plotting methods available.

```
ausgdp
```

```
##      Qtr1 Qtr2 Qtr3 Qtr4
## 1971           4612 4651
## 1972 4645 4615 4645 4722
## 1973 4780 4830 4887 4933
## 1974 4921 4875 4867 4905
```

Time series in R



Time series in R

Residential electricity sales

```
elecsales
```

```
## Time Series:
## Start = 1989
## End = 2008
## Frequency = 1
##  [1] 2354.34 2379.71 2318.52 2468.99 2386.09
##  [9] 2844.50 3000.70 3108.10 3357.50 3075.70
## [17] 3430.60 3527.48 3637.89 3655.00
```

Time series in R

Main package used in this course

```
> library(fpp)
```

This loads:

- some data for use in examples and exercises
- **forecast** package (for forecasting functions)
- **tseries** package (for a few time series functions)
- **fma** package (for lots of time series data)
- **expsmooth** package (for more time series data)
- **lmtest** package (for some regression functions)

Time series in R

Main package used in this course

```
> library(fpp)
```

This loads:

- some data for use in examples and exercises
- **forecast** package (for forecasting functions)
- **tseries** package (for a few time series functions)
- **fma** package (for lots of time series data)
- **expsmooth** package (for more time series data)
- **lmtest** package (for some regression functions)

Time series in R

Main package used in this course

```
> library(fpp)
```

This loads:

- some data for use in examples and exercises
- **forecast** package (for forecasting functions)
- **tseries** package (for a few time series functions)
- **fma** package (for lots of time series data)
- **expsmooth** package (for more time series data)
- **lmtest** package (for some regression functions)

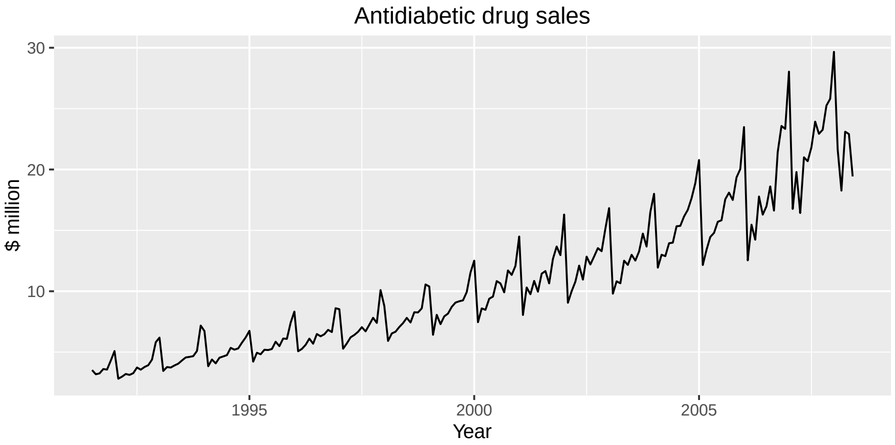
Outline

Time plots

```r
autoplot(melsyd[,"Economy.Class"])
```

Time plots

```r
autoplot(a10, ylab="$ million", xlab="Year",
        main="Antidiabetic drug sales")
```



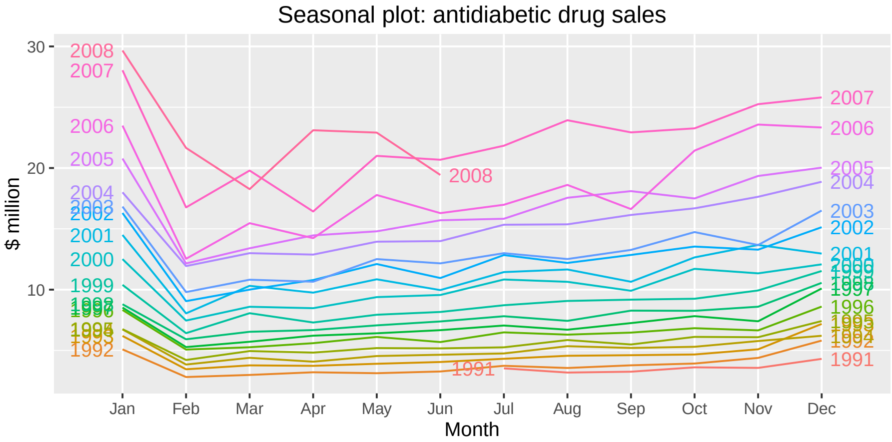
Outline

Lab Session 1

Outline

Seasonal plots

```r
ggseasonplot(a10, ylab="$ million",
  year.labels=TRUE, year.labels.left=TRUE) +
  ggtitle("Seasonal plot: antidiabetic drug sales")
```

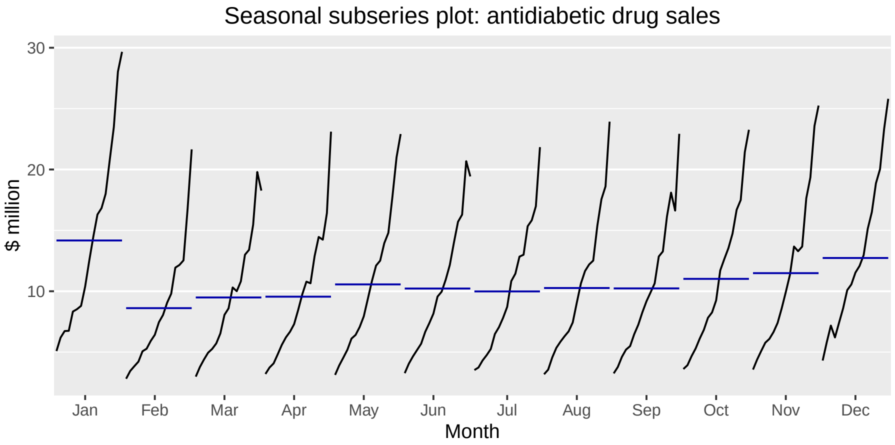


Seasonal plots

- Data plotted against the individual “seasons” in which the data were observed. (In this case a “season” is a month.)
- Something like a time plot except that the data from each season are overlapped.
- Enables the underlying seasonal pattern to be seen more clearly, and also allows any substantial departures from the seasonal pattern to be easily identified.
- In R: `ggseasonplot` or `seasonplot`.

Seasonal subseries plots

```
ggmonthplot(a10) + ylab("$ million") +
  ggtitle("Seasonal subseries plot: antidiabetic drug sal
```



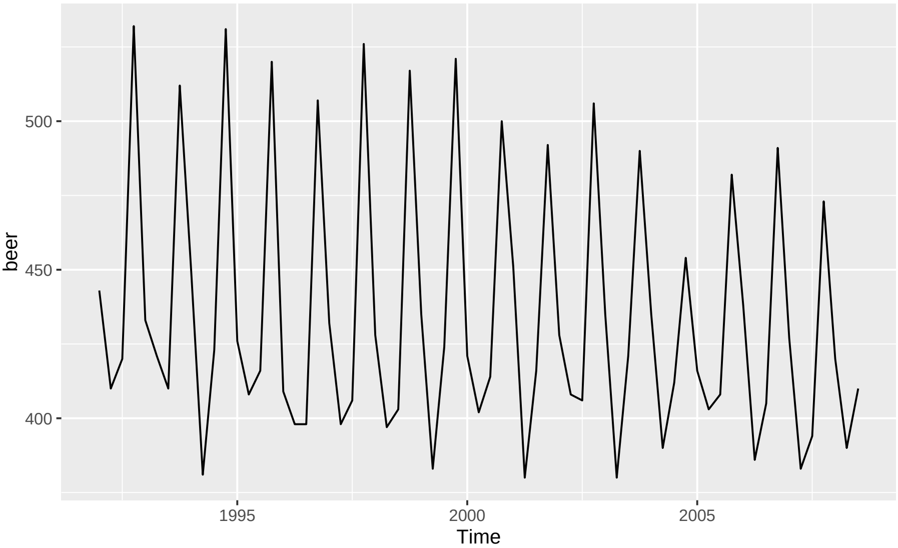
Seasonal subseries plots

- Data for each season collected together in time plot as separate time series.
- Enables the underlying seasonal pattern to be seen clearly, and changes in seasonality over time to be visualized.
- In R: `ggmonthplot` or `monthplot`

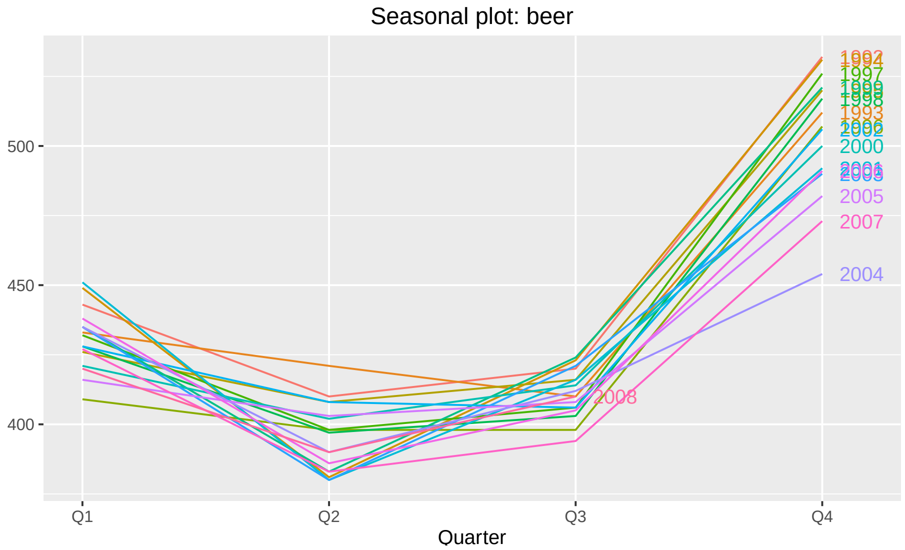
Quarterly Australian Beer Production

```
beer <- window(ausbeer,start=1992)

autoplot(beer)

ggseasonplot(beer,year.labels=TRUE)

ggmonthplot(beer)
```

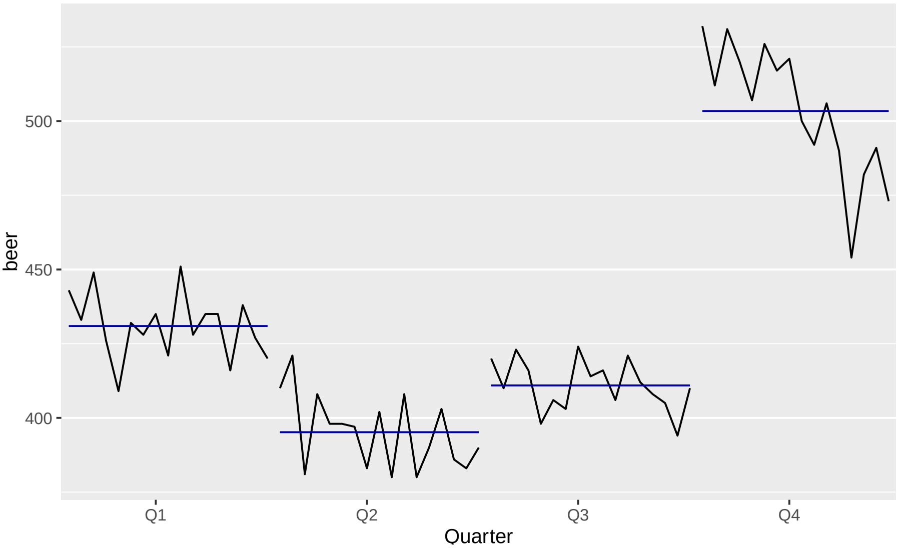
Quarterly Australian Beer Production



Quarterly Australian Beer Production



Quarterly Australian Beer Production



Outline

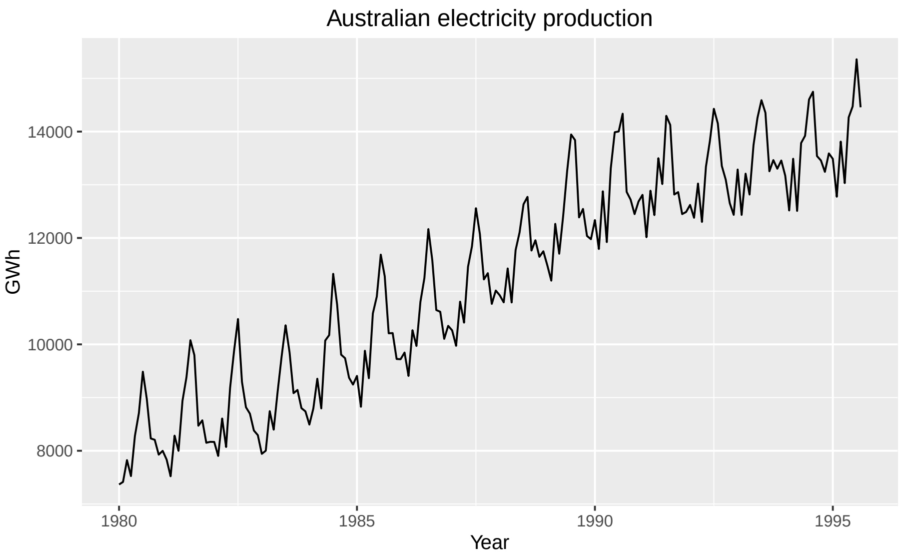
Time series patterns

Trend pattern exists when there is a long-term increase or decrease in the data.

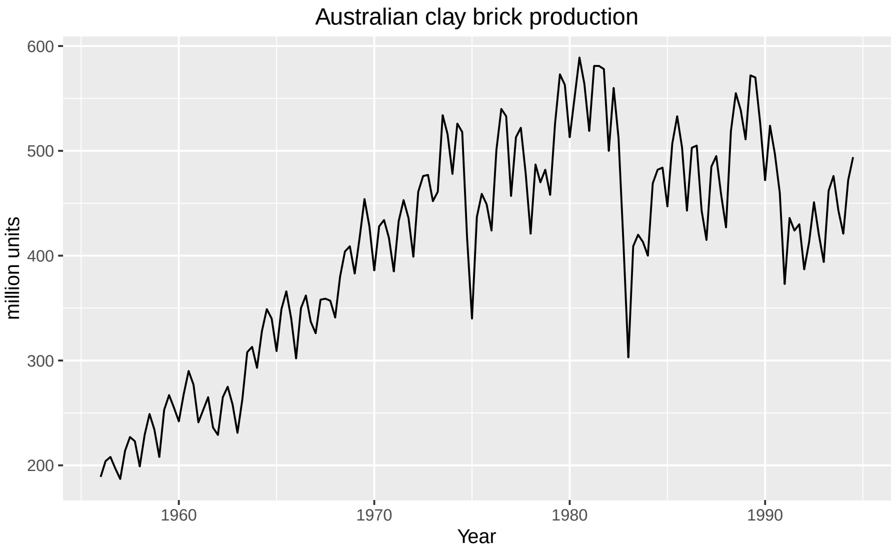
Seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).

Cyclic pattern exists when data exhibit rises and falls that are *not of fixed period* (duration usually of at least 2 years).

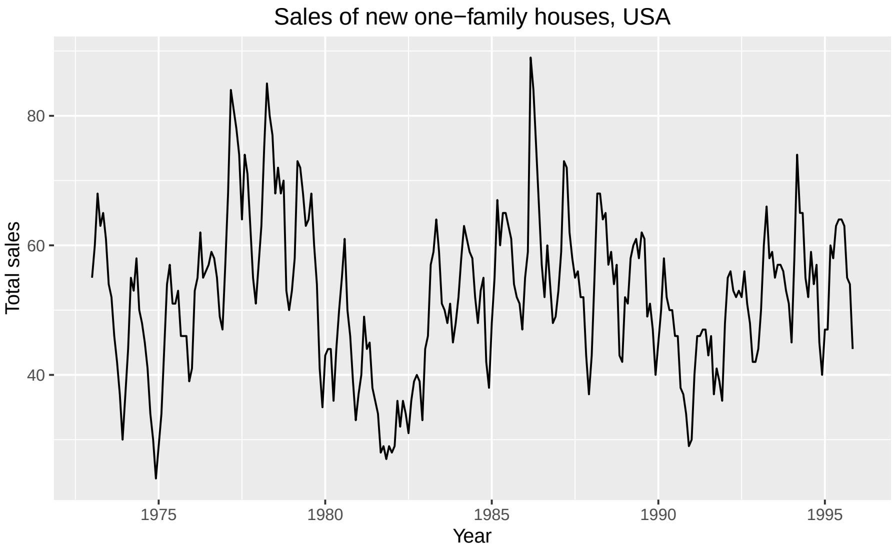
Time series patterns



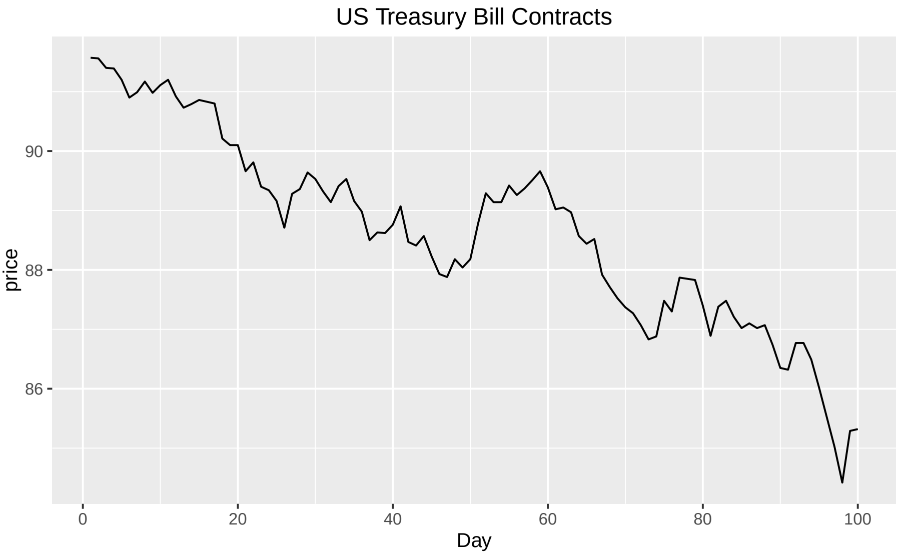
Time series patterns



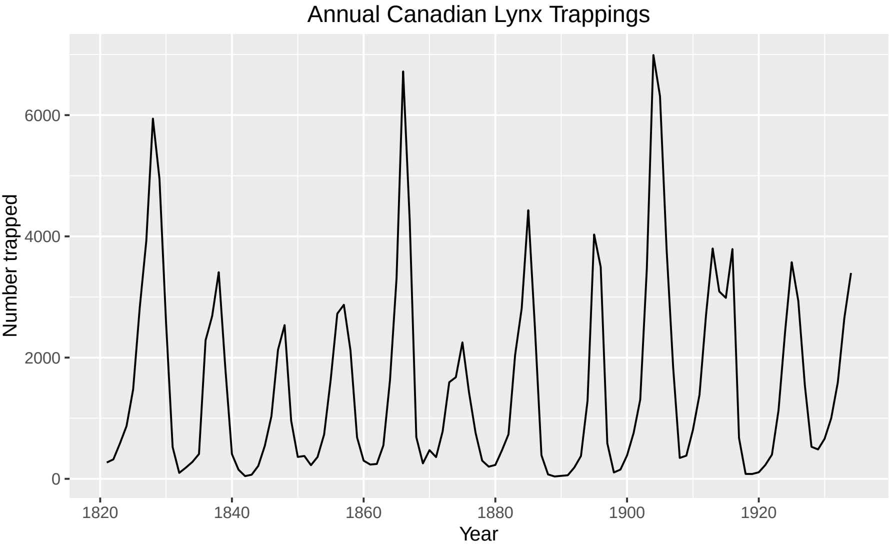
Time series patterns



Time series patterns



Time series patterns



Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.

Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

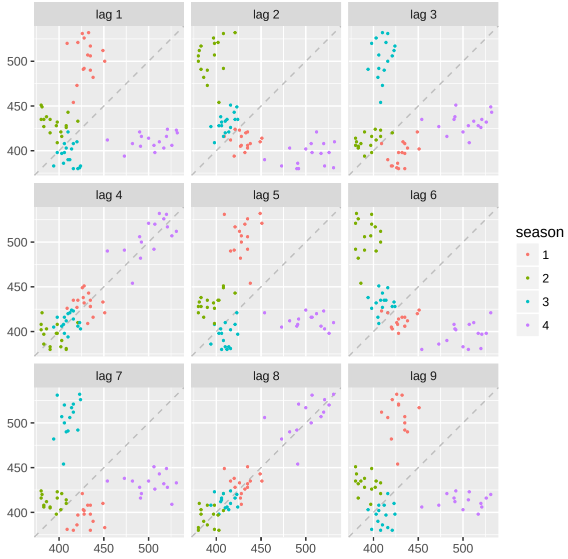
The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.

Outline

Example: Beer production

```
beer <- window(ausbeer, start=1992)
gglagplot(beer, lags=9, do.lines=FALSE, continu
```

Example: Beer production



Lagged scatterplots

- Each graph shows y_t plotted against y_{t-k} for different values of k.
- The autocorrelations are the correlations associated with these scatterplots.

Autocorrelation

Covariance and **correlation**: measure extent of **linear relationship** between two variables (y and X).

Autocovariance and **autocorrelation**: measure linear relationship between **lagged values** of a time series y.

We measure the relationship between: y_t and y_{t-1}
y_t and y_{t-2}
y_t and y_{t-3}
etc.

Autocorrelation

Covariance and **correlation**: measure extent of **linear relationship** between two variables (y and X).

Autocovariance and **autocorrelation**: measure linear relationship between **lagged values** of a time series y.

We measure the relationship between: y_t and y_{t-1}
y_t and y_{t-2}
y_t and y_{t-3}
etc.

Autocorrelation

Covariance and **correlation**: measure extent of **linear relationship** between two variables (y and X).

Autocovariance and **autocorrelation**: measure linear relationship between **lagged values** of a time series y.

We measure the relationship between: y_t and y_{t-1}
y_t and y_{t-2}
y_t and y_{t-3}
etc.

Autocorrelation

We denote the sample autocovariance at lag k by c_k and the sample autocorrelation at lag k by r_k. Then define

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

and $$r_k = c_k/c_0$$

- r_1 indicates how successive values of y relate to each other
- r_2 indicates how y values two periods apart relate to each other
- r_k is *almost* the same as the sample correlation between y_t and y_{t-k}.

Autocorrelation

We denote the sample autocovariance at lag k by c_k and the sample autocorrelation at lag k by r_k. Then define

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

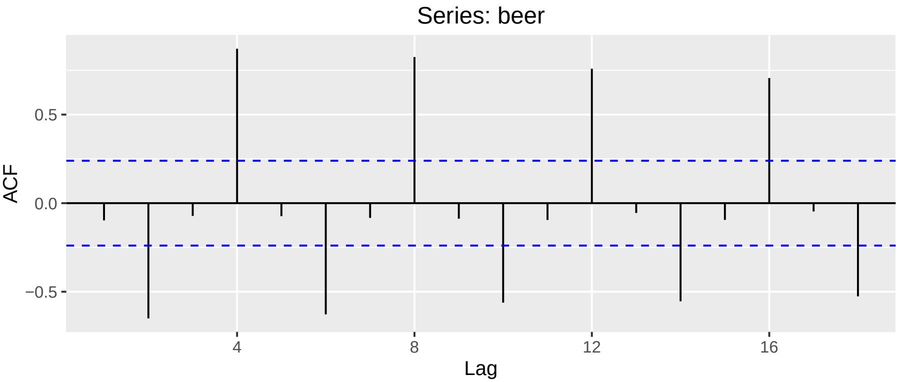
and $$r_k = c_k / c_0$$

- r_1 indicates how successive values of y relate to each other
- r_2 indicates how y values two periods apart relate to each other
- r_k is *almost* the same as the sample correlation between y_t and y_{t-k}.

Autocorrelation

Results for first 9 lags for beer `data:/footnotesize`

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r
-0.097	-0.651	-0.072	0.872	-0.074	-0.628	-0.083	0.825	-0.

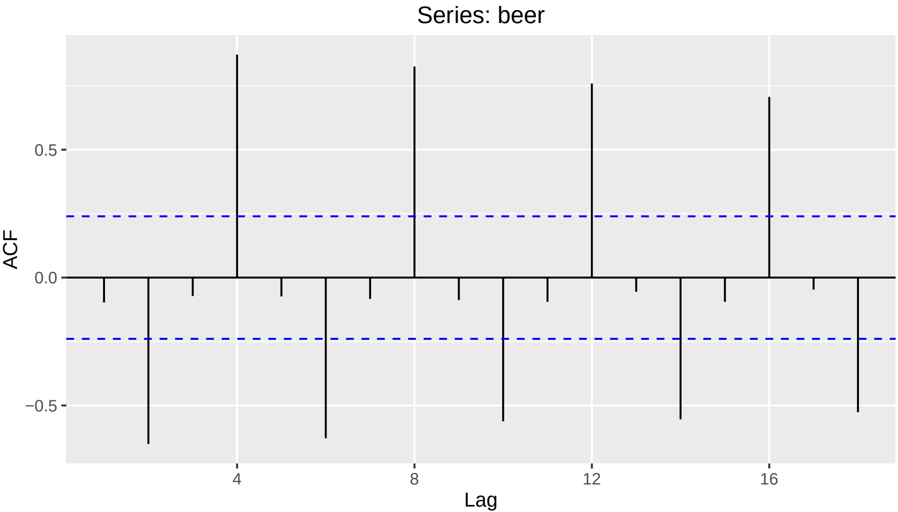


Autocorrelation

- r_4 higher than for the other lags. This is due to **the seasonal pattern in the data**: the peaks tend to be **4 quarters** apart and the troughs tend to be **2 quarters** apart.
- r_2 is more negative than for the other lags because troughs tend to be 2 quarters behind peaks.
- Together, the autocorrelations at lags 1, 2, $\dots$, make up the *autocorrelation* or ACF.
- The plot is known as a **correlogram**

ACF

```r
ggAcf(beer)
```

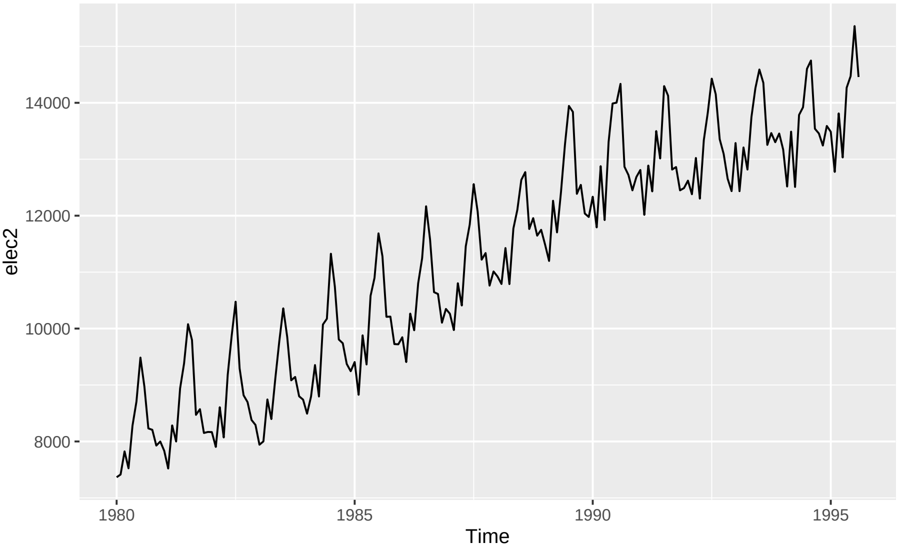


Recognizing seasonality in a time series

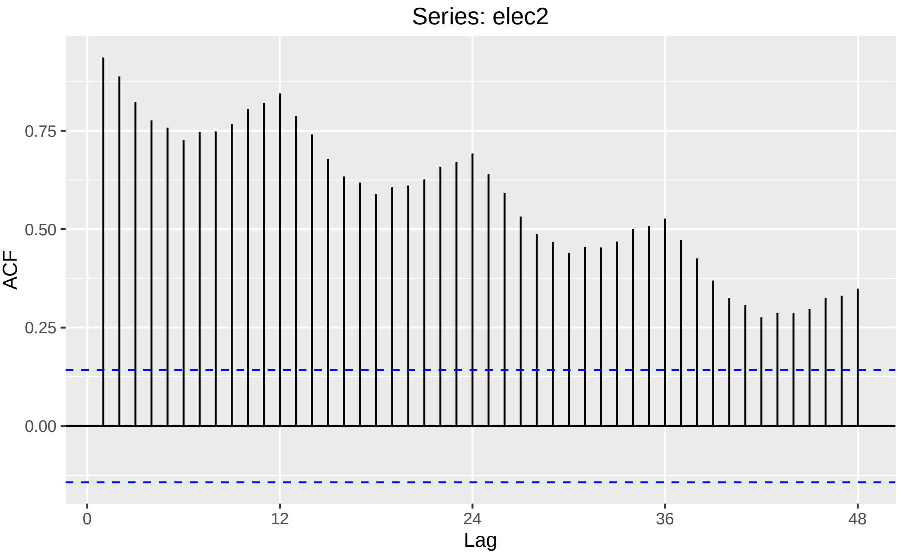
If there is seasonality, the ACF at the seasonal lag (e.g., 12 for monthly data) will be **large and positive**.

- For seasonal monthly data, a large ACF value will be seen at lag 12 and possibly also at lags 24, 36, …
- For seasonal quarterly data, a large ACF value will be seen at lag 4 and possibly also at lags 8, 12, …

Aus monthly electricity production



Aus monthly electricity production



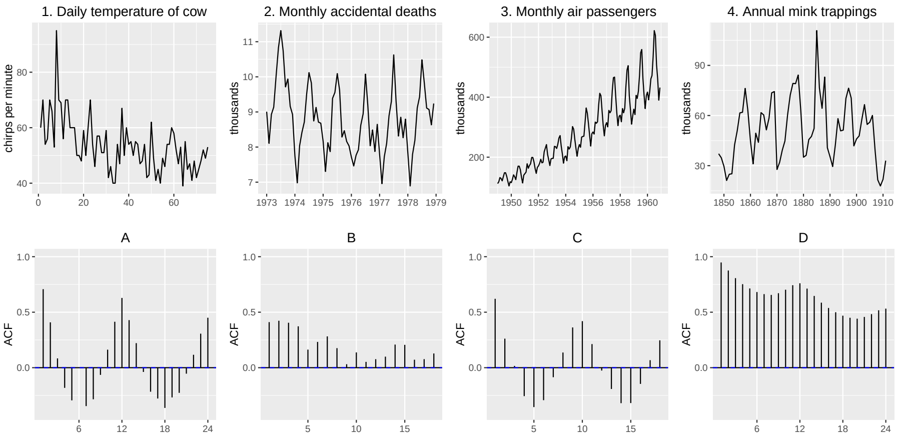
Aus monthly electricity production

Time plot shows clear trend and seasonality.

The same features are reflected in the ACF.

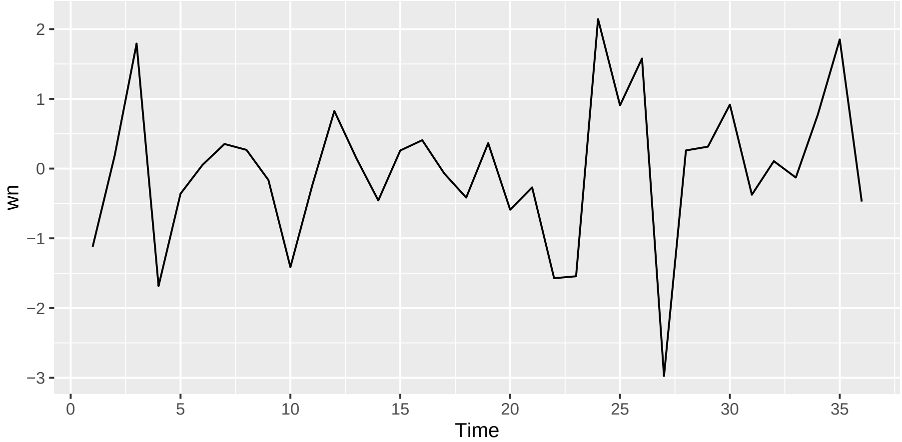
- The slowly decaying ACF indicates trend.
- The ACF peaks at lags 12, 24, 36, . . . , indicate seasonality of length 12.

Which is which?



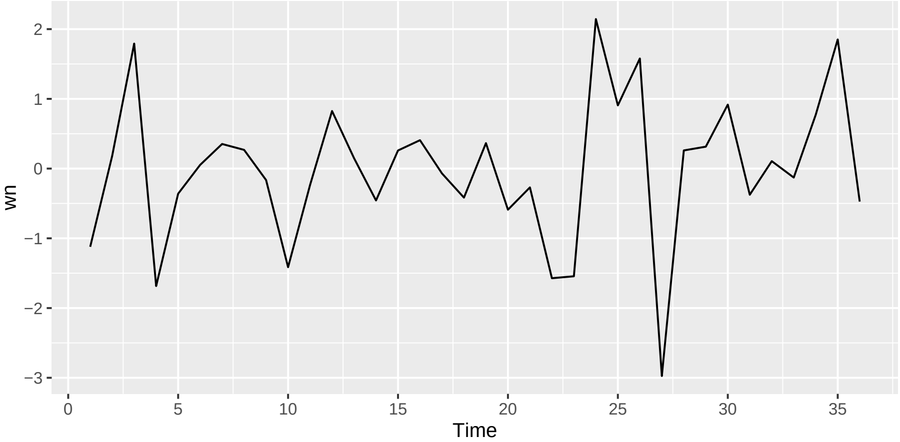
Outline

Example: White noise



White noise data is uncorrelated across time with zero mean and constant variance.
(Technically, we require independence as well.)

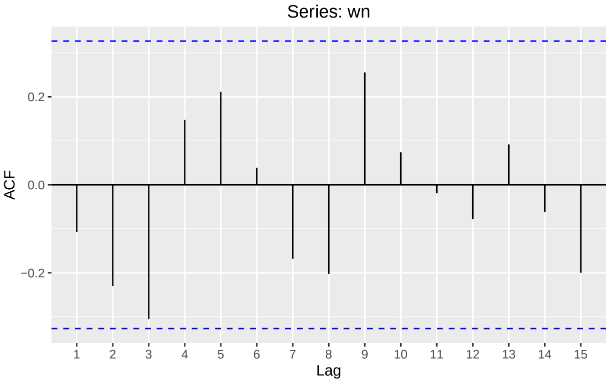
Example: White noise



White noise data is uncorrelated across time with zero mean and constant variance.
(Technically, we require independence as well.)

Example: White noise

r_1	-0.11
r_2	-0.23
r_3	-0.31
r_4	0.15
r_5	0.21
r_6	0.04
r_7	-0.17
r_8	-0.20
r_9	0.26
r_{10}	0.07



Sample autocorrelations for white noise series.

For uncorrelated data, we would expect each autocorrelation to be close to zero.

Sampling distribution of autocorrelations

Sampling distribution of r_k for white noise data is asymptotically $\text{N}(0,1/T)$.

- 95% of all r_k for white noise must lie within $\pm 1.96/\sqrt{T}$.
- If this is not the case, the series is probably not WN.
- Common to plot lines at $\pm 1.96/\sqrt{T}$ when plotting ACF. These are the ***critical values***.

Sampling distribution of autocorrelations

Sampling distribution of r_k for white noise data is asymptotically $\mathrm{N}(0,1/T)$.

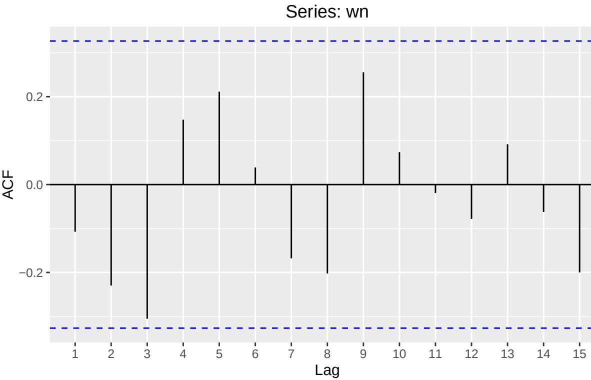
- 95% of all r_k for white noise must lie within $\pm 1.96/\sqrt{T}$.
- If this is not the case, the series is probably not WN.
- Common to plot lines at $\pm 1.96/\sqrt{T}$ when plotting ACF. These are the ***critical values***.

Autocorrelation

Example:

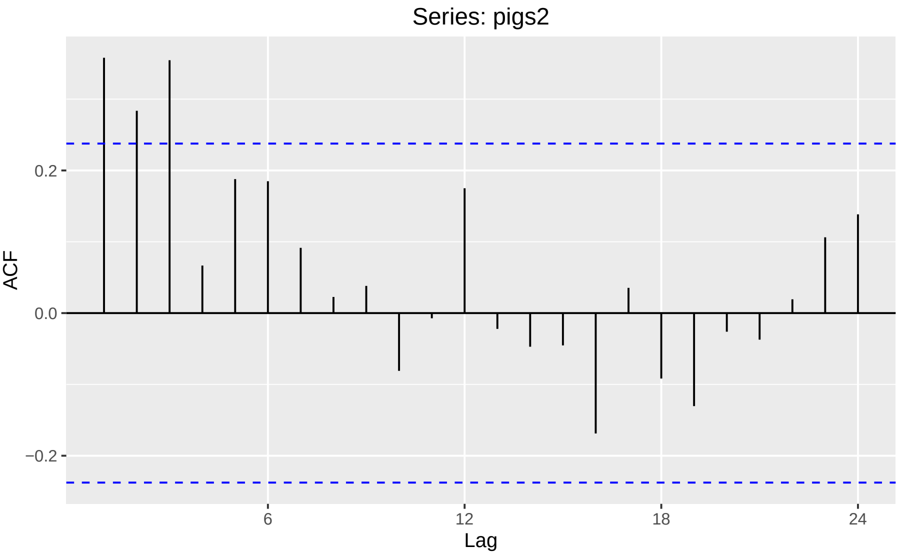
$T = 36$ and so critical values at $\pm 1.96/\sqrt{36} = \pm 0.327$.

All autocorrelation coefficients lie within these limits, confirming that the data are white noise. (More precisely, the data cannot be distinguished from white noise.)



Example: Pigs slaughtered

Example: Pigs slaughtered



Example: Pigs slaughtered

Monthly total number of pigs slaughtered in the state of Victoria, Australia, from January 1990 through August 1995. (Source: Australian Bureau of Statistics.)

- Difficult to detect pattern in time plot.
- ACF shows some significant autocorrelation at lags 1, 2, and 3.
- r_{12} relatively large although not significant. This may indicate some slight seasonality.

These show the series is **not a white noise series.**

Example: Pigs slaughtered

Monthly total number of pigs slaughtered in the state of Victoria, Australia, from January 1990 through August 1995. (Source: Australian Bureau of Statistics.)

- Difficult to detect pattern in time plot.
- ACF shows some significant autocorrelation at lags 1, 2, and 3.
- r_{12} relatively large although not significant. This may indicate some slight seasonality.

These show the series is **not a white noise series.**

Example: Pigs slaughtered

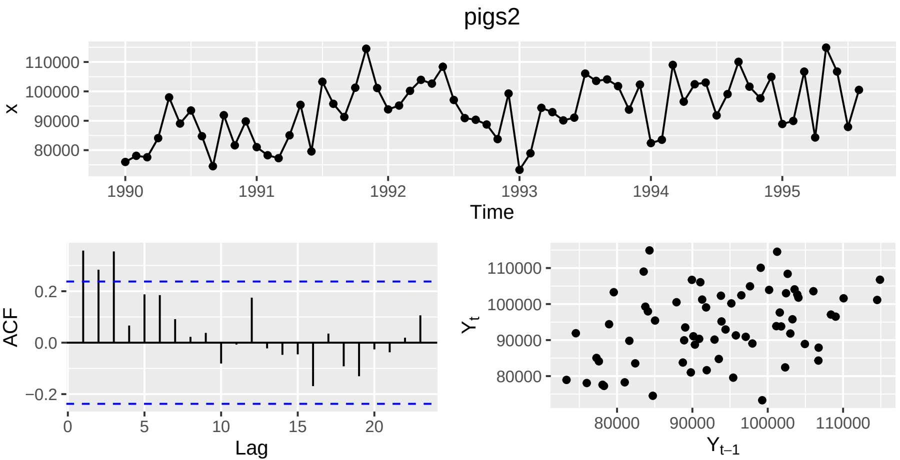
Monthly total number of pigs slaughtered in the state of Victoria, Australia, from January 1990 through August 1995. (Source: Australian Bureau of Statistics.)

- Difficult to detect pattern in time plot.
- ACF shows some significant autocorrelation at lags 1, 2, and 3.
- r_{12} relatively large although not significant. This may indicate some slight seasonality.

These show the series is **not a white noise series**.

Combination graph

```
ggtsdisplay(pigs2, plot.type='scatter')
```



Outline

Lab Session 2