Discrete Time Series

Lecture 07

Dr. Colin Rundel

Random variable review

Mean and variance of RVs

- Expected Value

$$E(X) = \begin{cases} \sum_x x\, P(X = x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x\, f(x)\, dx & X \text{ is continuous} \end{cases}$$

- Variance

$$Var(X) = E\Big(\big(X - E(X)\big)^2\Big) = E(X^2) - E(X)^2$$

$$= \begin{cases} \sum_x \big(x - E(X)\big)^2\, P(X = x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} \big(x - E(X)\big)^2\, f(x)\, dx & X \text{ is continuous} \end{cases}$$

Covariance of RVs

$$\text{Cov}(X, Y) = \text{E}\Big(\big(X - \text{E}(X)\big)\big(Y - \text{E}(Y)\big)\Big) = \text{E}(XY) - \text{E}(X)\text{E}(Y)$$

$$= \begin{cases} \sum_x \big(x - \text{E}(X)\big)\big(y - \text{E}(Y)\big)\ \text{P}(X = x, Y = y) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \big(x - \text{E}(X)\big)\big(y - \text{E}(Y)\big)\ f(x, y)\ dx\ dy & X \text{ is continuous} \end{cases}$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Properties of Expected Value

- *Constant*

  $\mathrm{E}(c) = c$ if c is constant

- *Constant Multiplication*

  $\mathrm{E}(cX) = c\mathrm{E}(X)$

- *Constant Addition*

  $\mathrm{E}(X + c) = \mathrm{E}(X) + c$

- *Addition*

  $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$

- *Subtraction*

  $\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y)$

- *Multiplication*

  $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$

  if X and Y are independent

Properties of Variance

- *Constant*

  $\mathrm{Var}(c) = 0$ if c is constant

- *Constant Multiplication*

  $\mathrm{Var}(cX) = c^2\ \mathrm{Var}(x)$

- *Constant Addition*

  $\mathrm{Var}(X + c) = \mathrm{Var}(X)$

- *Addition*

  $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

  if X and Y are independent.

- *Subtraction*

  $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

  if X and Y are independent.

Properties of Covariance

- *Constant*

  $\text{Cov}(X, c) = 0$ if c is constant

- *Identity*

  $\text{Cov}(X, X) = \text{Var}(X)$

- *Symmetric*

  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

- *Constant Multiplication*

  $\text{Cov}(aX, bY) = ab\ \text{Cov}(X, Y)$

- *Constant Addition*

  $\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$

- *Distribution*

  $\text{Cov}(aX + bY, cV + dW) = ac\ \text{Cov}(X, V) + ad\ \text{Cov}(X, W) + bc\ \text{Cov}(Y, V) + bd$

Discrete Time Series

Stationary Processes

A stocastic process (i.e. a time series) is considered to be *strictly stationary* if the properties of the process are not changed by a shift in origin.

In the time series context this means that the joint distribution of $\{y_{t_1}, \ldots, y_{t_n}\}$ must be identical to the distribution of $\{y_{t_1+k}, \ldots, y_{t_n+k}\}$ for any value of n and k.

Weakly Stationary

Strict stationary is unnecessarily strong / restrictive for many applications, so instead we often opt for *weak stationary* which requires the following,

1. The process must have finite variance / second moment

$$E(y_t^2) < \infty \text{ for all } t$$

2. The mean of the process must be constant

$$E(y_t) = \mu \text{ for all } t$$

3. The cross moment (covariance) may only depends on the lag (i.e. $t - s$ for y_t and y_s)

$$Cov(y_t, y_s) = Cov(y_{t+k}, y_{s+k}) \text{ for all } t, s, k$$

When we say stationary in class we will almost always mean *weakly stationary*.

Autocorrelation

For a stationary time series, where $E(y_t) = \mu$ and $Var(y_t) = \sigma^2$ for all t, we define the autocorrelation at lag k as

$$\rho_k = Cor(y_t, y_{t+k}) = \frac{Cov(y_t, y_{t+k})}{\sqrt{Var(y_t)Var(y_{t+k})}}$$
$$= \frac{E((y_t - \mu)(y_{t+k} - \mu))}{\sigma^2}$$

this can be written in terms of the autocovariance function (γ_k) as

$$\gamma_k = \gamma(t, t+k) = Cov(y_t, y_{t+k})$$
$$\rho_k = \frac{\gamma(t, t+k)}{\sqrt{\gamma(t,t)\gamma(t+k, t+k)}} = \frac{\gamma(k)}{\gamma(0)}$$

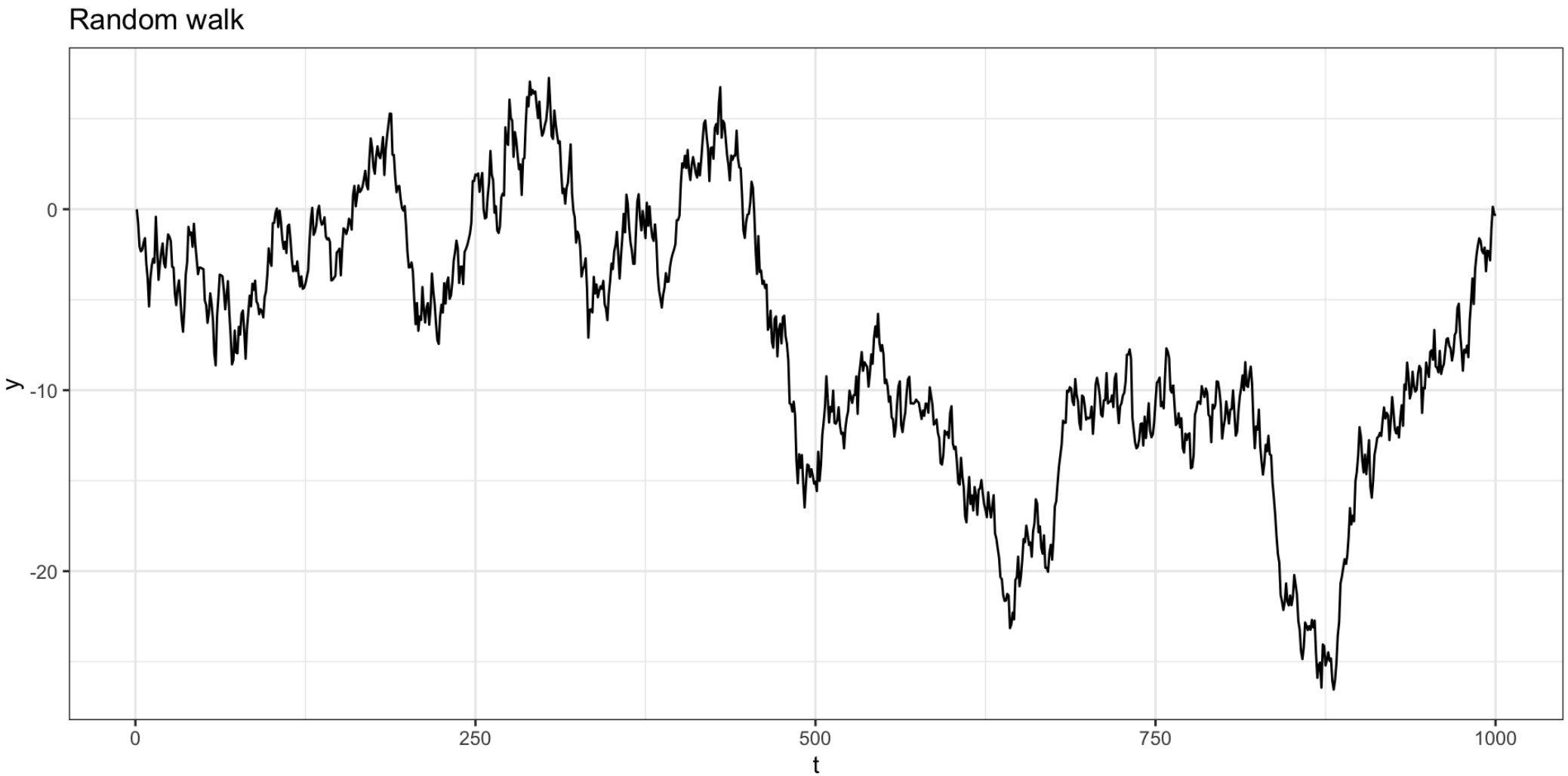
Covariance Structure

Based on our definition of a (weakly) stationary process, it implies a covariance of the following structure,

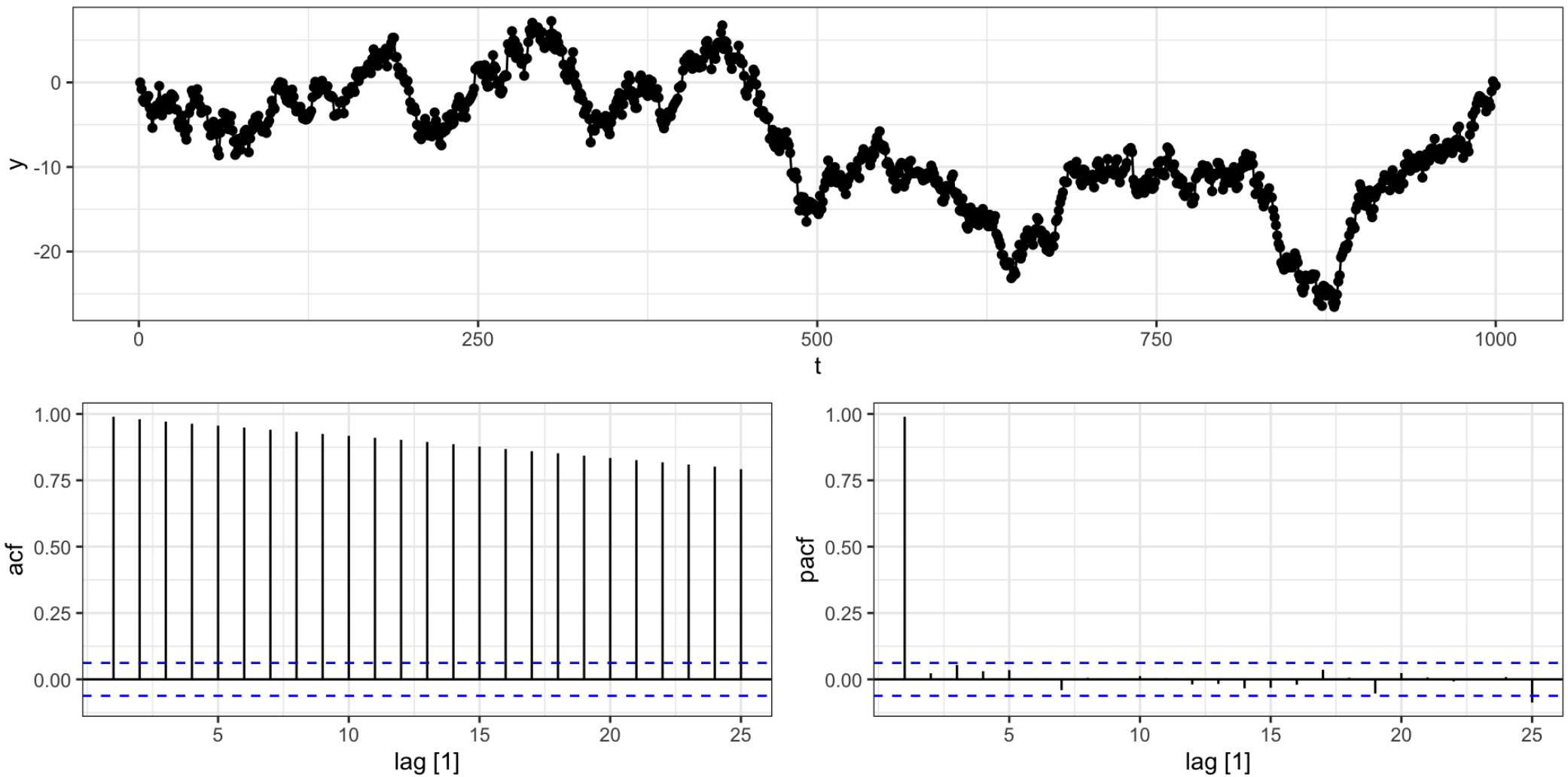
$$\boldsymbol{\Sigma} = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \gamma(3) & \cdots & \gamma(n-1) & \gamma(n) \\ \gamma(1) & \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(n-2) & \gamma(n-1) \\ \gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(n-3) & \gamma(n-2) \\ \gamma(3) & \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(n-4) & \gamma(n-3) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \gamma(n-3) & \gamma(n-4) & \cdots & \gamma(0) & \gamma(1) \\ \gamma(n) & \gamma(n-1) & \gamma(n-2) & \gamma(n-3) & \cdots & \gamma(1) & \gamma(0) \end{pmatrix}$$

Example - Random walk

Let $y_t = y_{t-1} + w_t$ with $y_0 = 0$ and $w_t \sim N(0, 1)$.



ACF + PACF


y
0
-10
-20
0
250
500
750
1000
t
acf
1.00
0.75
0.50
0.25
0.00
5
10
15
20
25
lag [1]
pacf
1.00
0.75
0.50
0.25
0.00
5
10
15
20
25
lag [1]

Stationary?

Is y_t stationary?

Partial Autocorrelation - pACF

Given these type of patterns in the autocorrelation we often want to examine the relationship between y_t and y_{t+k} with the (linear) dependence of y_t on y_{t+1} through y_{t+k-1} removed.

This is done through the calculation of a partial autocorrelation ($\alpha(k)$), which is defined as follows:

$$\begin{aligned}
\alpha(0) &= 1 \\
\alpha(1) &= \rho(1) = \text{Cor}(y_t, y_{t+1}) \\
&\vdots \\
\alpha(k) &= \text{Cor}(y_t - P_{t,k}(y_t),\ y_{t+k} - P_{t,k}(y_{t+k}))
\end{aligned}$$

where $P_{t,k}(y)$ is the projection of y onto the space spanned by $y_{t+1}, \cdots, y_{t+k-1}$.

pACF - Calculation

Let $\rho(k)$ be the autocorrelation for the process at lag k then the partial autocorrelation at lag k will be $\phi(k,k)$ given by the Durbin-Levinson algorithm,

$$\phi(k,k) = \frac{\rho(k) - \sum_{t=1}^{k-1} \phi(k-1,t)\,\rho(k-t)}{1 - \sum_{t=1}^{k-1} \phi(k-1,t)\,\rho(t)}$$

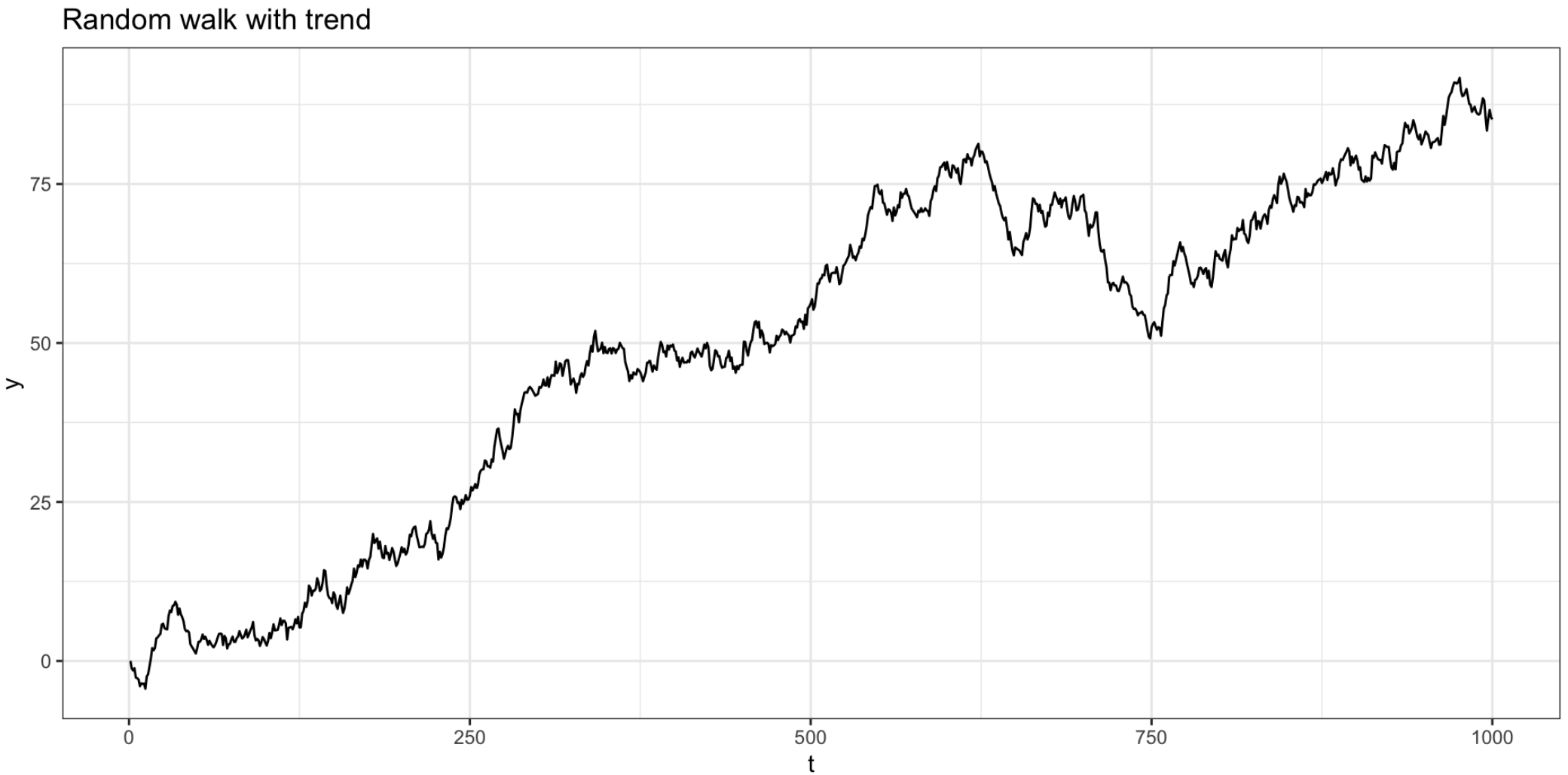
where

$$\phi(k,t) = \phi(k-1,t) - \phi(k,k)\,\phi(k-1,k-t)$$

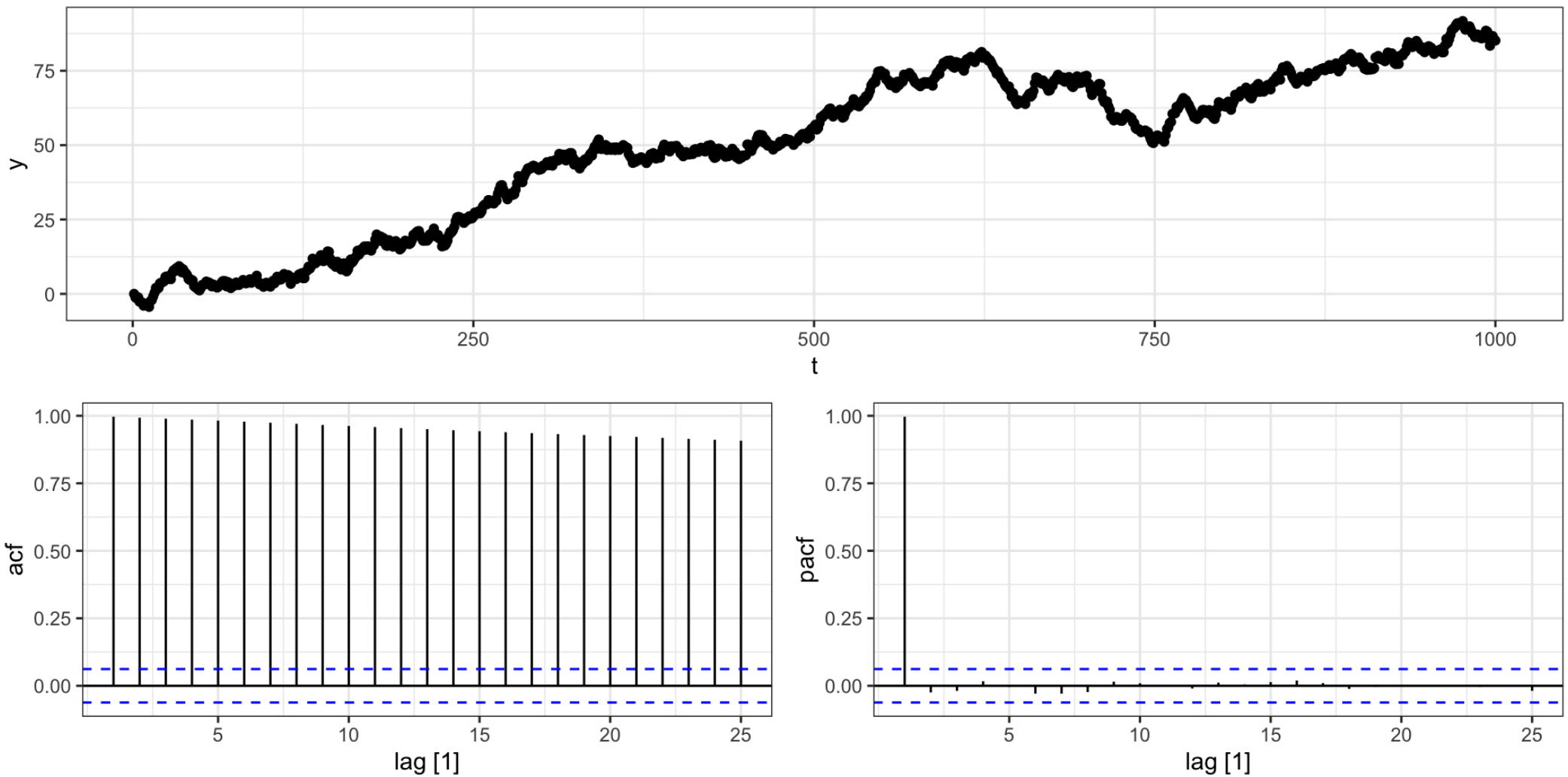
Starting with $\phi(1,1) = \rho(1)$ we can solve iteratively for $\phi(2,2), \ldots, \phi(k,k)$.

Example - Random walk with drift

Let $y_t = \delta + y_{t-1} + w_t$ with $y_0 = 0$ and $w_t \sim N(0, 1)$.



ACF + PACF

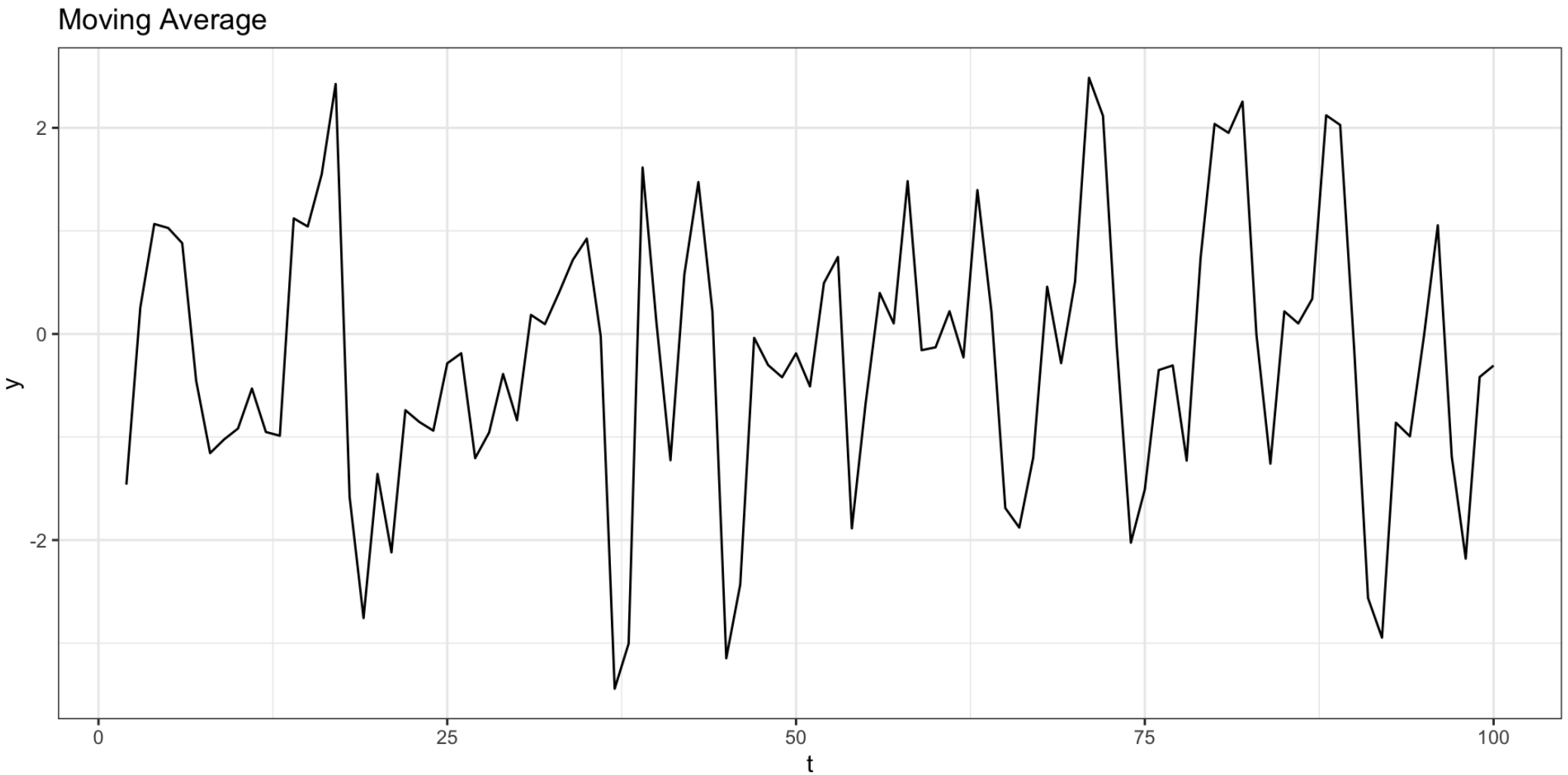

y
75
50
25
0
0
250
500
750
1000
t
acf
1.00
0.75
0.50
0.25
0.00
5
10
15
20
25
lag [1]
pacf
1.00
0.75
0.50
0.25
0.00
5
10
15
20
25
lag [1]

Stationary?

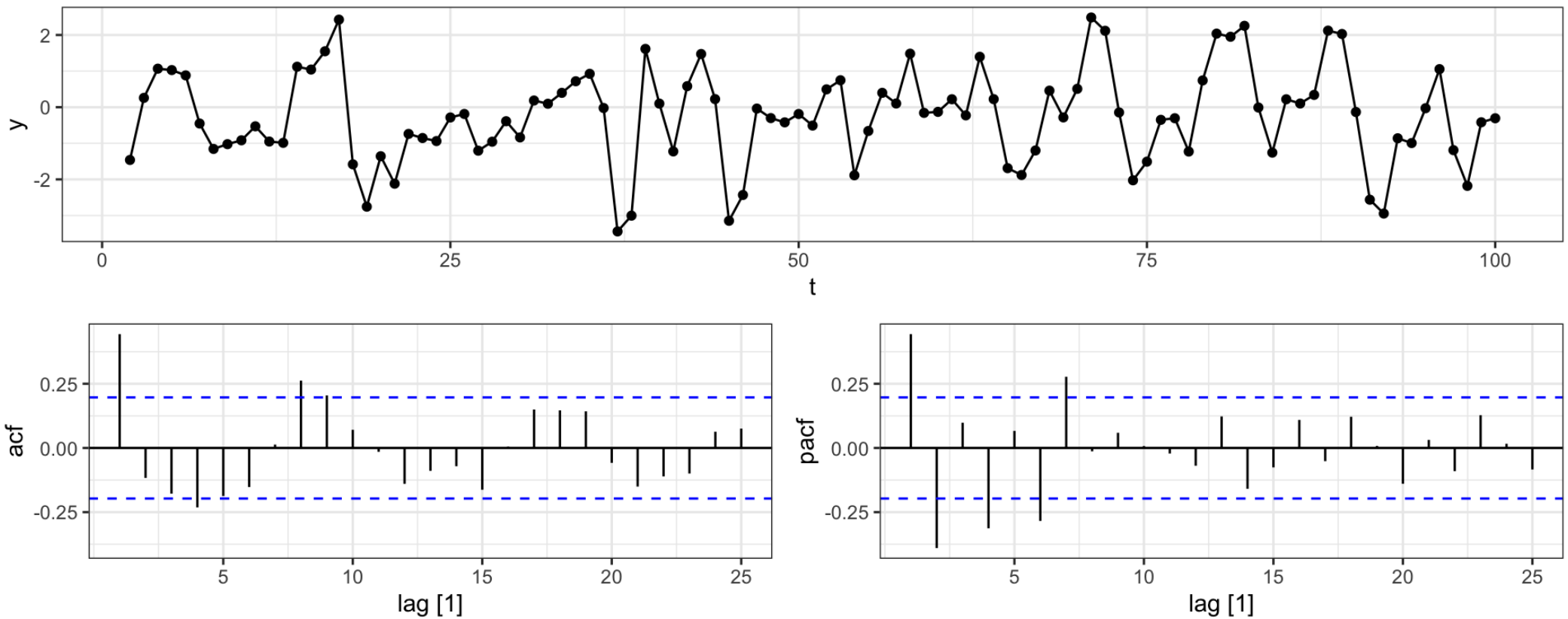
Is y_t stationary?

Example - Moving Average

Let $w_t \sim N(0, 1)$ and $y_t = w_{t-1} + w_t$.



ACF + PACF

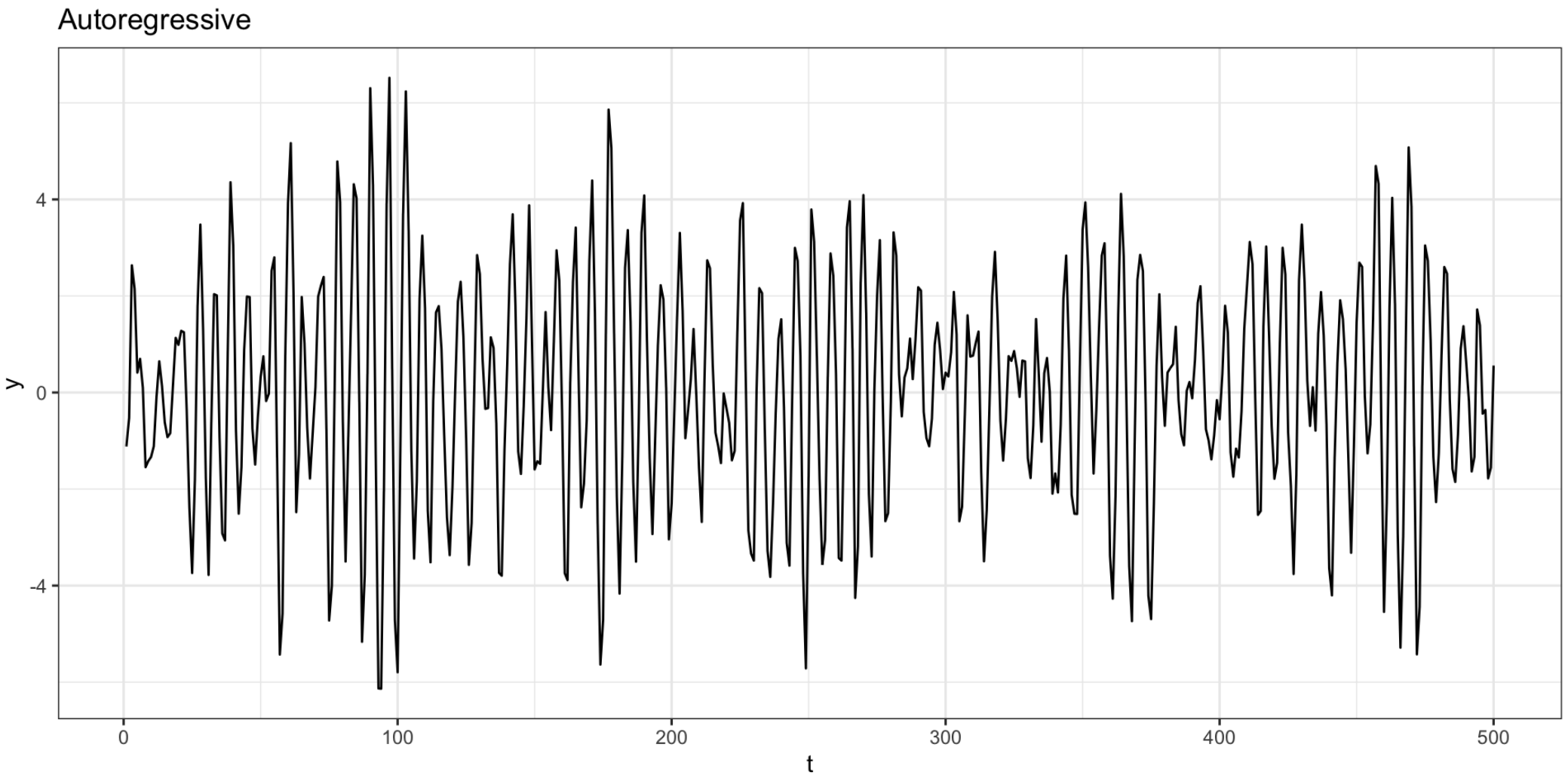

y
2
0
-2
0
25
50
75
100
t
acf
0.25
0.00
-0.25
5
10
15
20
25
lag [1]
pacf
0.25
0.00
-0.25
5
10
15
20
25
lag [1]

Stationary?

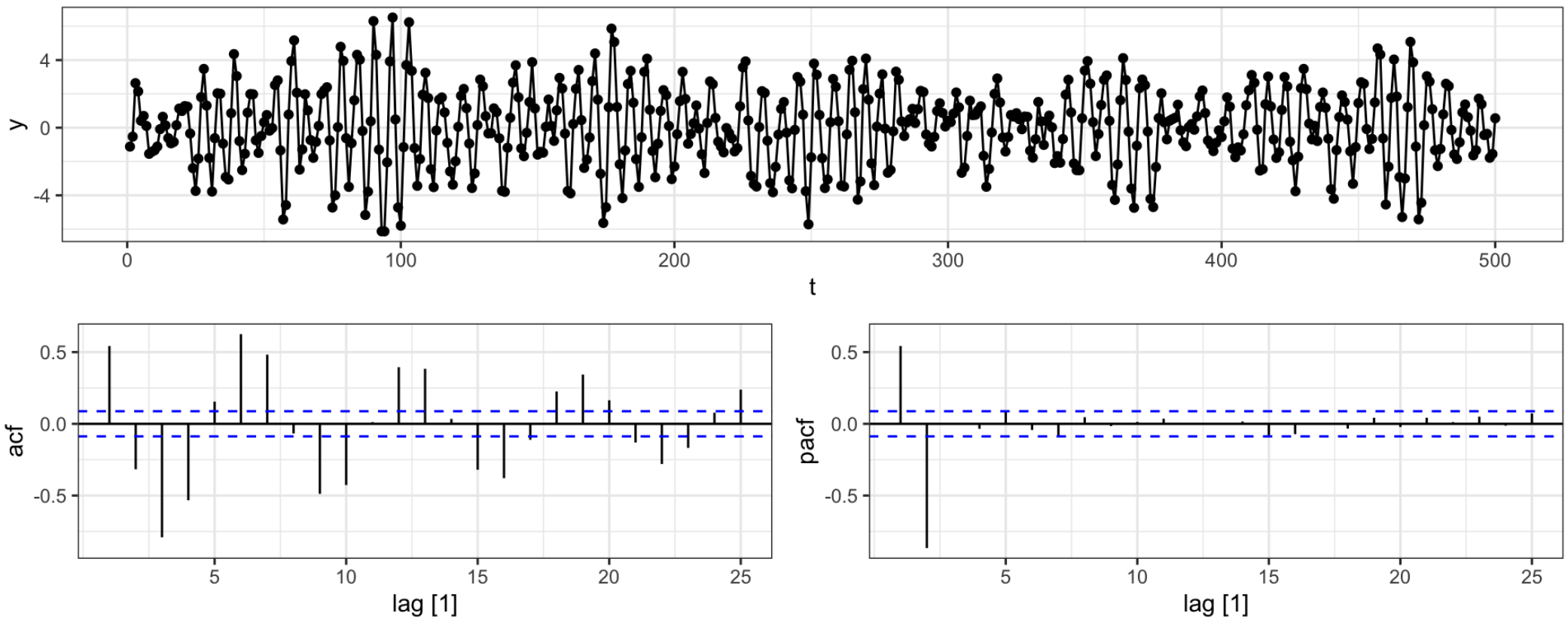
Is y_t stationary?

Autoregressive

Let $w_t \sim N(0, 1)$ and $y_t = y_{t-1} - 0.9y_{t-2} + w_t$ with $y_t = 0$ for $t < 1$.



ACF + PACF


y
4
0
-4
0
100
200
300
400
500
t
acf
0.5
0.0
-0.5
5
10
15
20
25
lag [1]
pacf
0.5
0.0
-0.5
5
10
15
20
25
lag [1]

Tidy time series

`ts` objects

In base R, time series are usually encoded using the `ts` S3 class,

tidyverts

This is an effort headed by Rob Hyndman (of forecast and fpp3 fame) and others to provide a consistent tidydata based framework for working with time series data and models.

Core packages:

- `tsibble` - temporal data frames and related tools
- `fable` - tidy forecasting / modeling
- `feasts` - feature extraction and statistics
- `tsibbledata` - sample tsibble data sets

tsibble

A tsibble is a tibble with additional infrastructure for encoding temporal data - specifically a tsibble is a tidy data frame with an *index* and *key* where

- the *index* is the variable that describes the inherent ordering of the data (from past to present)
- and the *key* is one or more variables that define the unit of observation over time
- each observation should be uniquely identified by the *index* and *key*

global_economy

```
1 tsibbledata::global_economy
```

```
# A tsibble: 15,150 x 9 [1Y]
# Key:       Country [263]
   Country  Code   Year     GDP Growth   CPI Imports Exports Population
   <fct>    <fct> <dbl>   <dbl>  <dbl> <dbl>   <dbl>   <dbl>      <dbl>
 1 Afghani… AFG    1960 5.38e8     NA    NA    7.02    4.13    8996351
 2 Afghani… AFG    1961 5.49e8     NA    NA    8.10    4.45    9166764
 3 Afghani… AFG    1962 5.47e8     NA    NA    9.35    4.88    9345868
 4 Afghani… AFG    1963 7.51e8     NA    NA   16.9     9.17    9533954
 5 Afghani… AFG    1964 8.00e8     NA    NA   18.1     8.89    9731361
 6 Afghani… AFG    1965 1.01e9     NA    NA   21.4    11.3     9938414
 7 Afghani… AFG    1966 1.40e9     NA    NA   18.6     8.57   10152331
 8 Afghani… AFG    1967 1.67e9     NA    NA   14.2     6.77   10372630
 9 Afghani… AFG    1968 1.37e9     NA    NA   15.2     8.90   10604346
10 Afghani… AFG    1969 1.41e9     NA    NA   15.0    10.1    10854428
# ℹ 15,140 more rows
```

vic_elec

```
1 tsibbledata::vic_elec
```

```
# A tsibble: 52,608 x 5 [30m] <Australia/Melbourne>
   Time                Demand Temperature Date       Holiday
   <dttm>               <dbl>       <dbl> <date>     <lgl>
 1 2012-01-01 00:00:00  4383.        21.4 2012-01-01 TRUE
 2 2012-01-01 00:30:00  4263.        21.0 2012-01-01 TRUE
 3 2012-01-01 01:00:00  4049.        20.7 2012-01-01 TRUE
 4 2012-01-01 01:30:00  3878.        20.6 2012-01-01 TRUE
 5 2012-01-01 02:00:00  4036.        20.4 2012-01-01 TRUE
 6 2012-01-01 02:30:00  3866.        20.2 2012-01-01 TRUE
 7 2012-01-01 03:00:00  3694.        20.1 2012-01-01 TRUE
 8 2012-01-01 03:30:00  3562.        19.6 2012-01-01 TRUE
 9 2012-01-01 04:00:00  3433.        19.1 2012-01-01 TRUE
10 2012-01-01 04:30:00  3359.        19.0 2012-01-01 TRUE
# ℹ 52,598 more rows
```

aus_retail

```
1 tsibbledata::aus_retail
```

```
# A tsibble: 64,532 x 5 [1M]
# Key:       State, Industry [152]
   State                       Industry `Series ID`    Month Turnover
   <chr>                       <chr>    <chr>          <mth>    <dbl>
 1 Australian Capital Territory Cafes, … A3349849A   1982 Apr      4.4
 2 Australian Capital Territory Cafes, … A3349849A   1982 May      3.4
 3 Australian Capital Territory Cafes, … A3349849A   1982 Jun      3.6
 4 Australian Capital Territory Cafes, … A3349849A   1982 Jul      4
 5 Australian Capital Territory Cafes, … A3349849A   1982 Aug      3.6
 6 Australian Capital Territory Cafes, … A3349849A   1982 Sep      4.2
 7 Australian Capital Territory Cafes, … A3349849A   1982 Oct      4.8
 8 Australian Capital Territory Cafes, … A3349849A   1982 Nov      5.4
 9 Australian Capital Territory Cafes, … A3349849A   1982 Dec      6.9
10 Australian Capital Territory Cafes, … A3349849A   1983 Jan      3.8
# ℹ 64,522 more rows
```

as_tsibble()

Existing ts objects or data frames can be converted to a tsibbles easily,

```
tsibble::as_tsibble(co2)
```

```
# A tsibble: 468 x 2 [1M]
      index value
      <mth> <dbl>
 1 1959 Jan  315.
 2 1959 Feb  316.
 3 1959 Mar  316.
 4 1959 Apr  318.
 5 1959 May  318.
 6 1959 Jun  318
 7 1959 Jul  316.
 8 1959 Aug  315.
 9 1959 Sep  314.
10 1959 Oct  313.
# ℹ 458 more rows
```

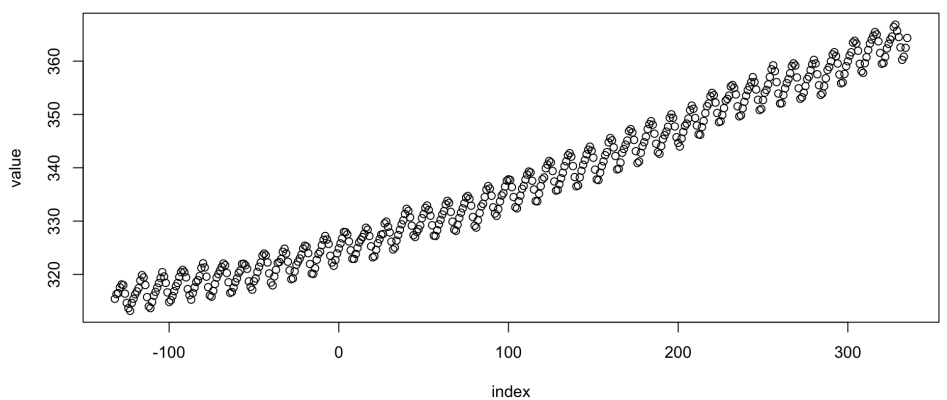
```
tibble(
  co2 = c(co2),
  t = c(time(co2)) |> tsibble::yearmonth()
) |>
  tsibble::as_tsibble(index = t)
```

```
# A tsibble: 468 x 2 [1M]
     co2        t
   <dbl>    <mth>
 1  315. 1970 Feb
 2  316. 1970 Mar
 3  316. 1970 Apr
 4  318. 1970 May
 5  318. 1970 Jun
 6  318  1970 Jul
 7  316. 1970 Aug
 8  315. 1970 Sep
 9  314. 1970 Oct
10  313. 1970 Nov
# ℹ 458 more rows
```

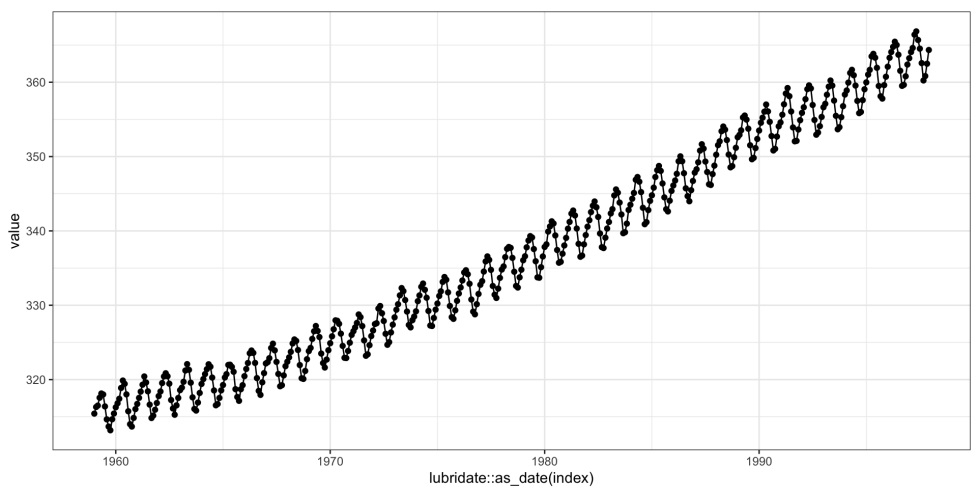
plotting tsibbles

As the tsibble is basically just a tibble which is just a data frame both base and ggplot plotting methods will work with tsibbles.

```
1 tsibble::as_tsibble(co2) |>
2   plot()
```

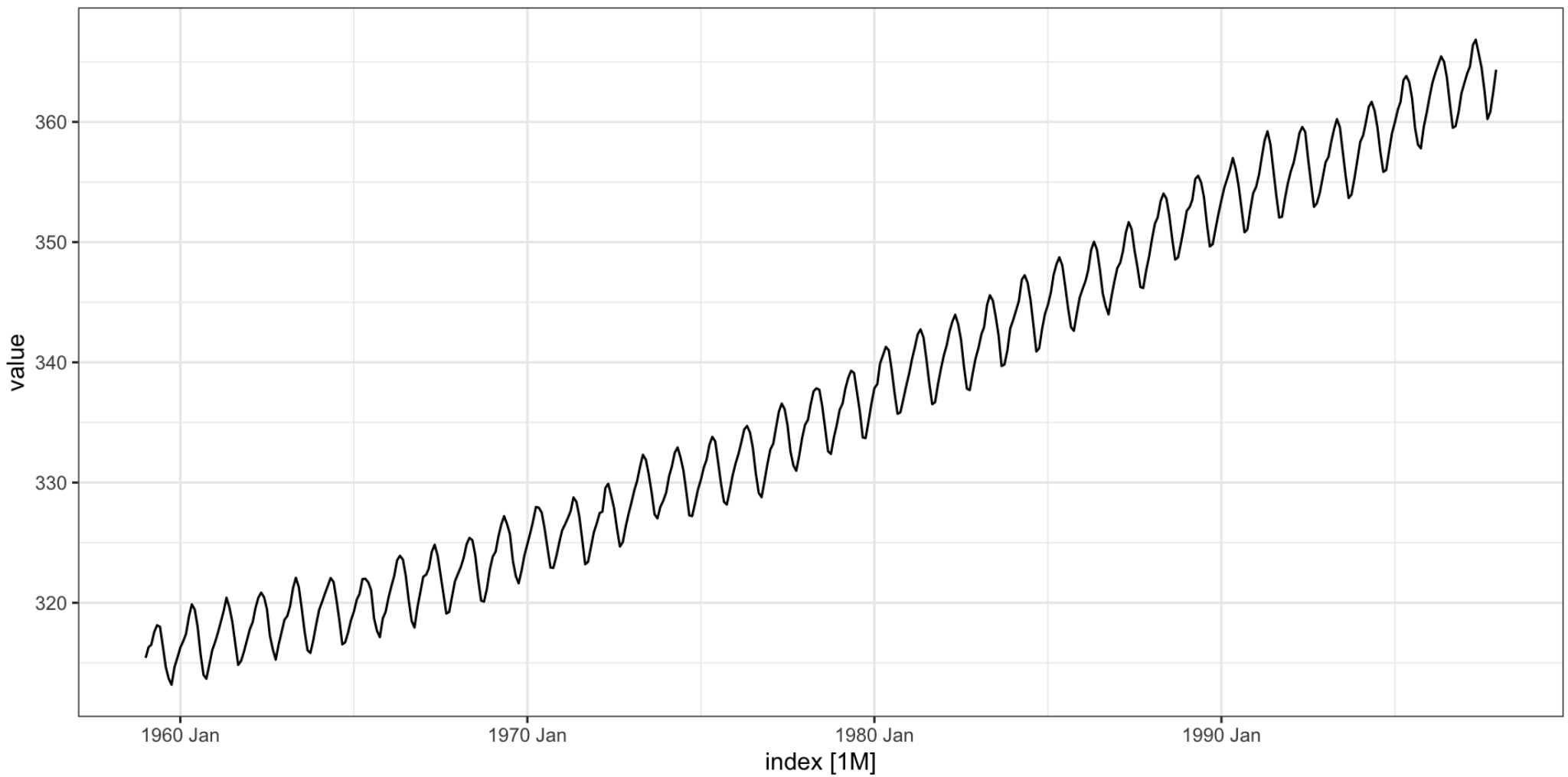



```
1 tsibble::as_tsibble(co2) |>
2   ggplot(
3     aes(x=lubridate::as_date(index), y=value)
4   ) +
5     geom_point() +
6     geom_line()
```



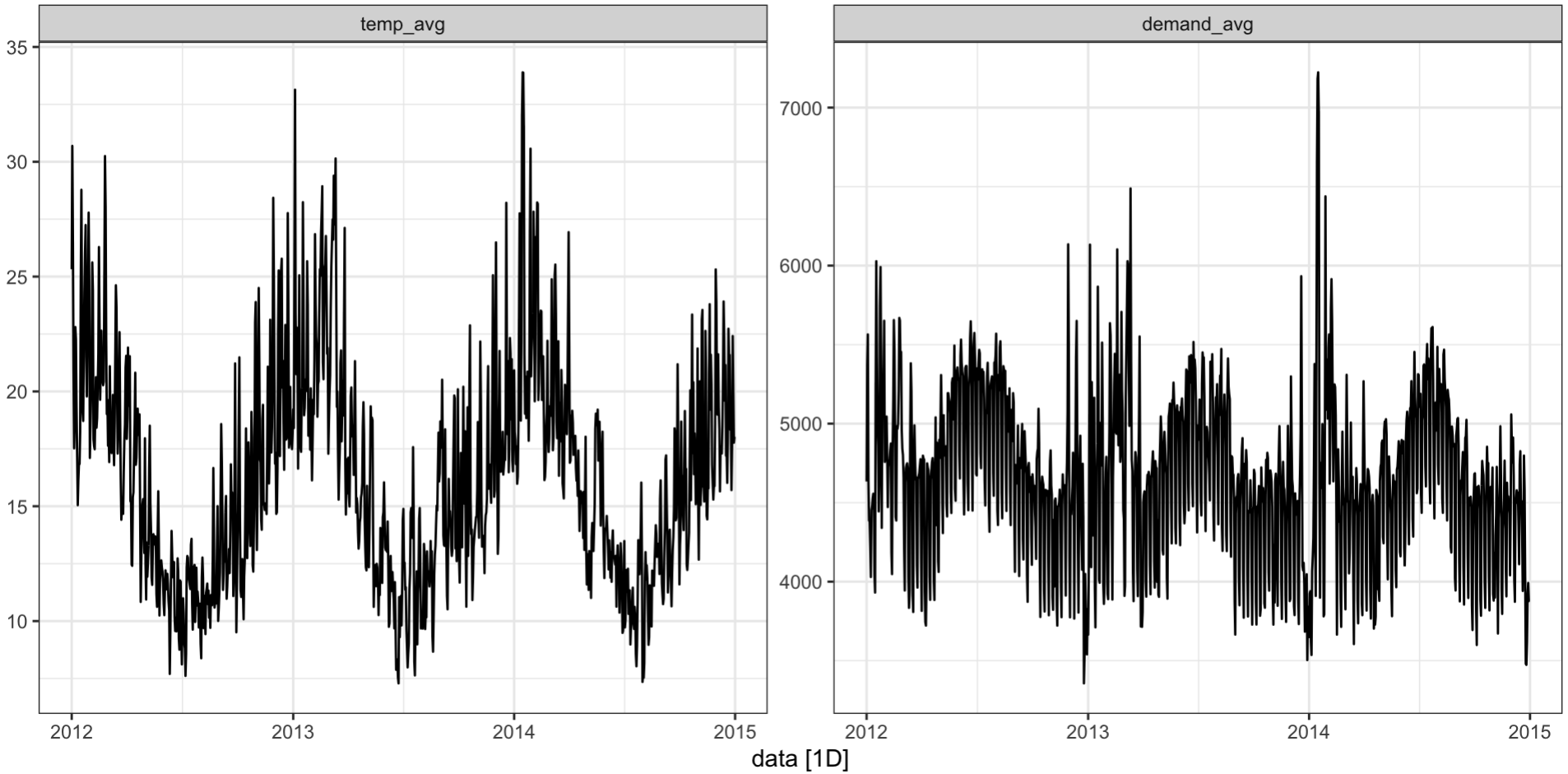
autoplot

```
1 tsibble::as_tsibble(co2) |>
2   autoplot(.vars = vars(value))
```



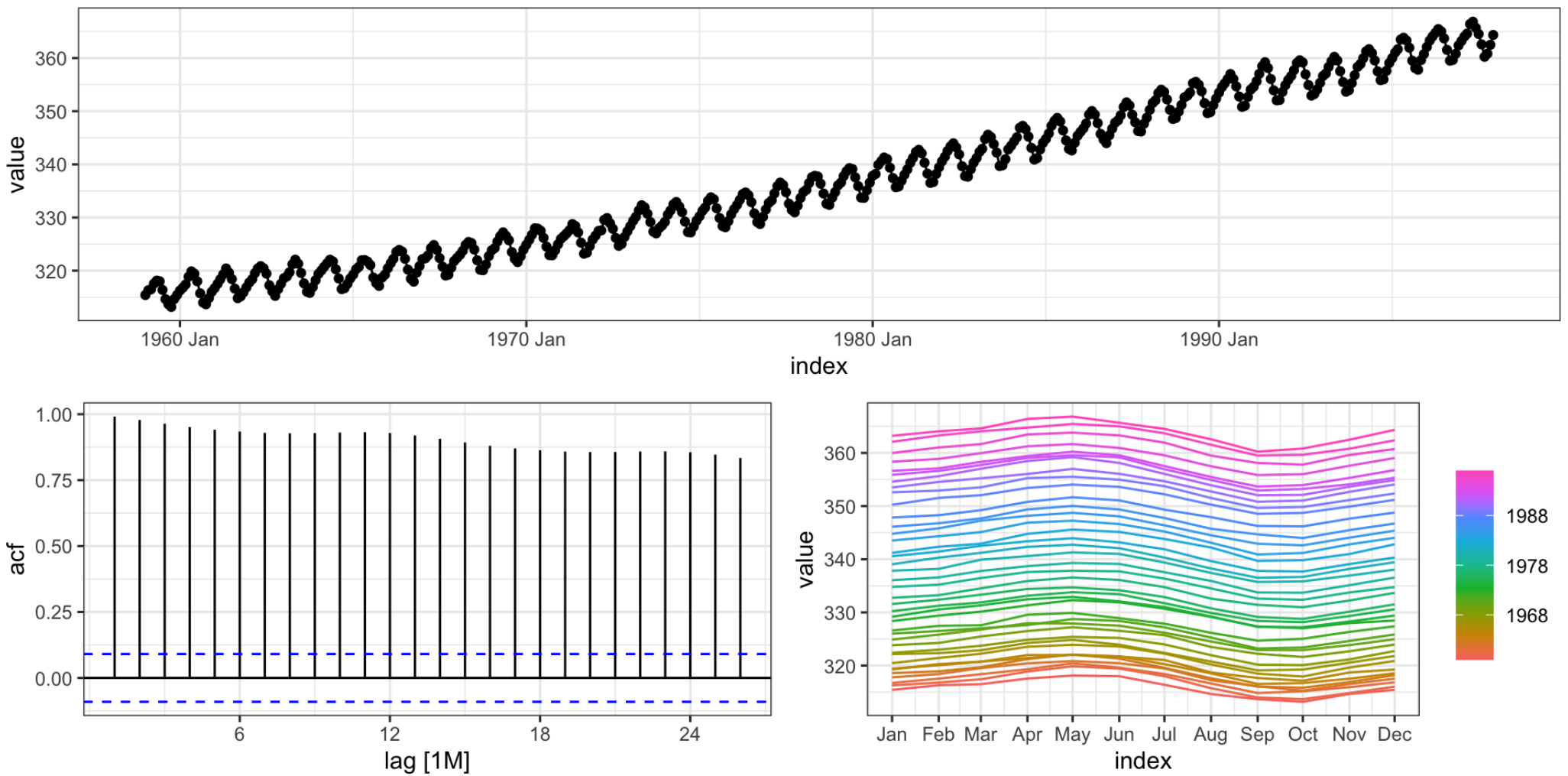
Multiple variables

```
tsibbledata::vic_elec |>
  tsibble::index_by(data = ~ lubridate::as_date(.)) |>
  summarize(
    demand_avg = mean(Demand, na.rm=TRUE),
    temp_avg = mean(Temperature, na.rm=TRUE)
  ) |>
  autoplot(.vars = vars(temp_avg, demand_avg))
```


temp_avg
demand_avg
35
30
25
20
15
10
7000
6000
5000
4000
2012
2013
2014
2015
data [1D]

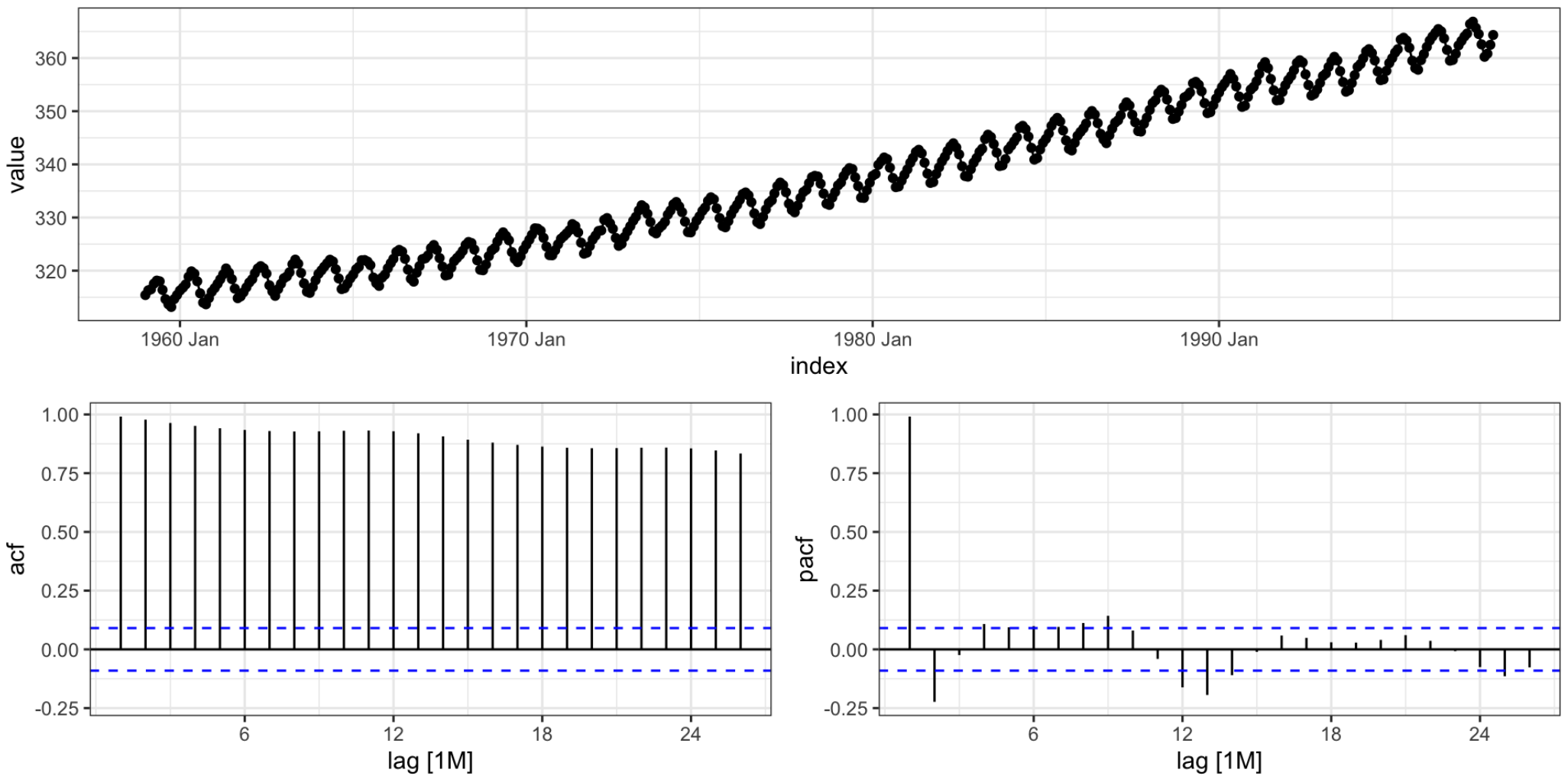
gg_tsdisplay()

```
1 tsibble::as_tsibble(co2) |>
2   feasts::gg_tsdisplay(y = value)
```



gg_tsdisplay() w/ pACF

```
1 tsibble::as_tsibble(co2) |>
2   feasts::gg_tsdisplay(y = value, plot_type = "partial")
```



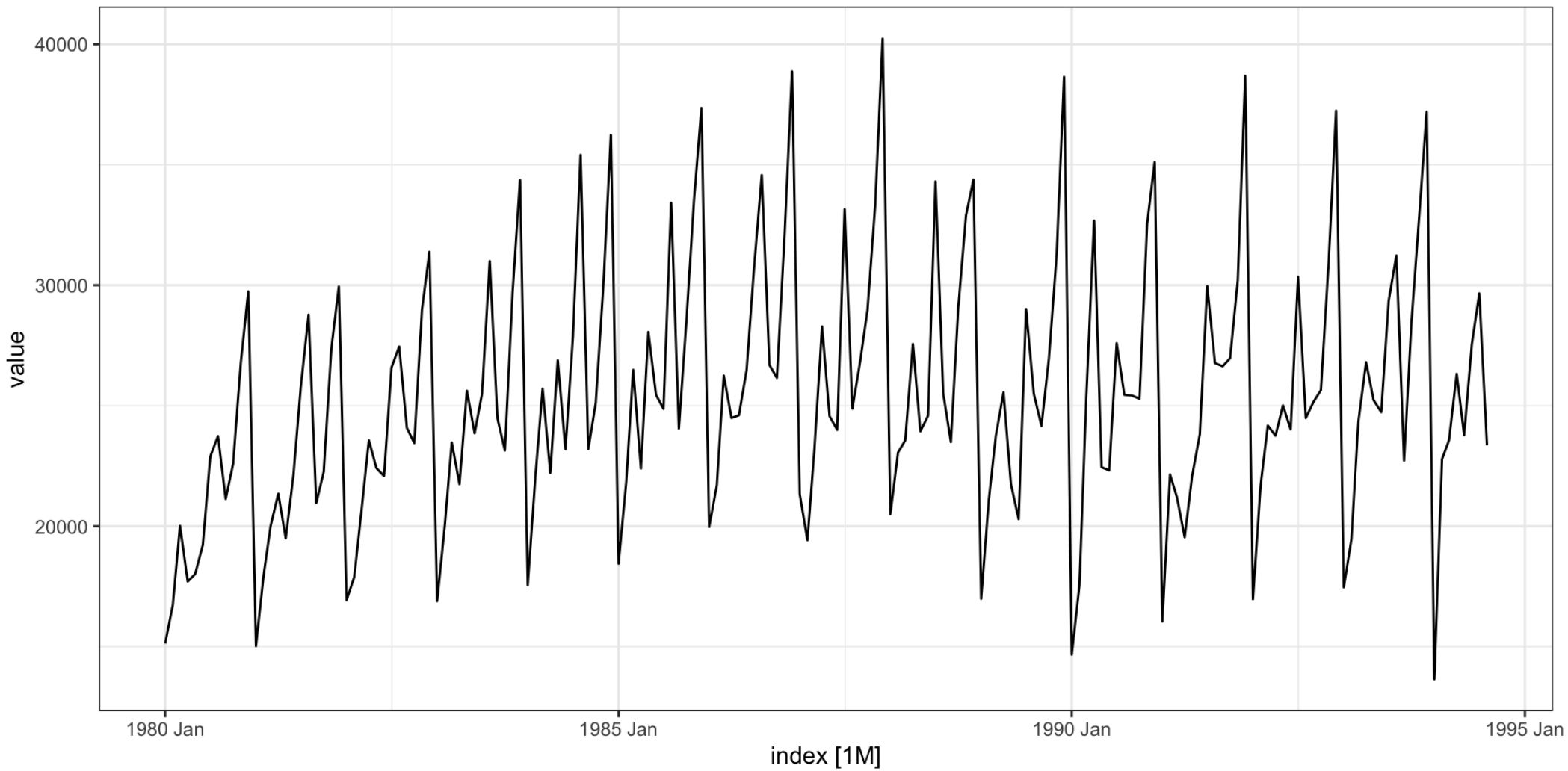
Example - Australian Wine Sales

Australian total wine sales by wine makers in bottles <= 1 litre. Jan 1980 – Aug 1994.

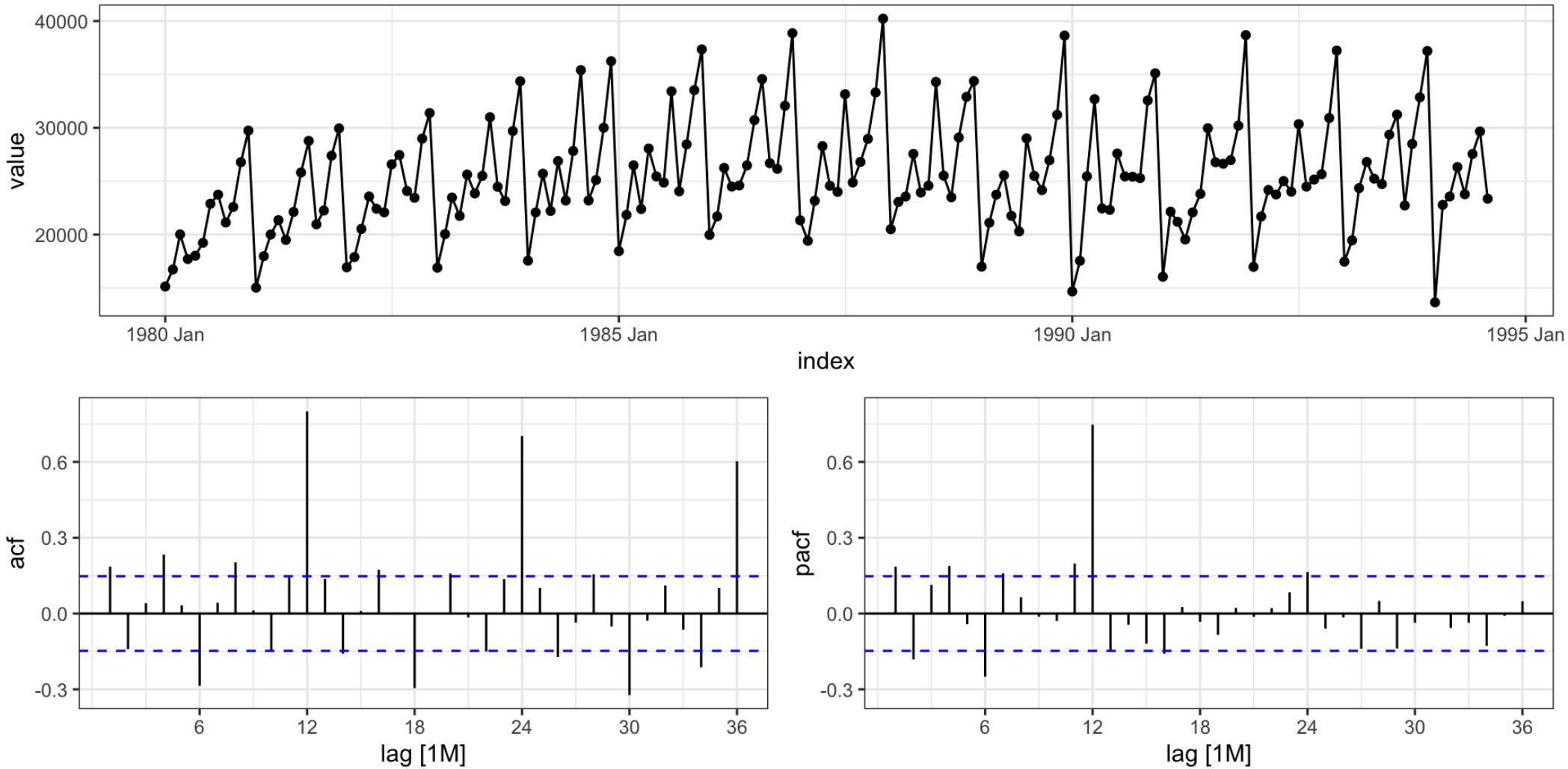
```
data(AustralianWine, package="Rssa")
( aus_wine = AustralianWine |>
    tsibble::as_tsibble() |>
    filter(key == "Total", !is.na(value))
)
```

```
# A tsibble: 176 x 3 [1M]
# Key:       key [1]
      index key   value
      <mth> <chr> <int>
 1 1980 Jan Total 15136
 2 1980 Feb Total 16733
 3 1980 Mar Total 20016
 4 1980 Apr Total 17708
 5 1980 May Total 18019
 6 1980 Jun Total 19227
 7 1980 Jul Total 22893
 8 1980 Aug Total 23739
 9 1980 Sep Total 21133
10 1980 Oct Total 22591
# ℹ 166 more rows
```

Time series

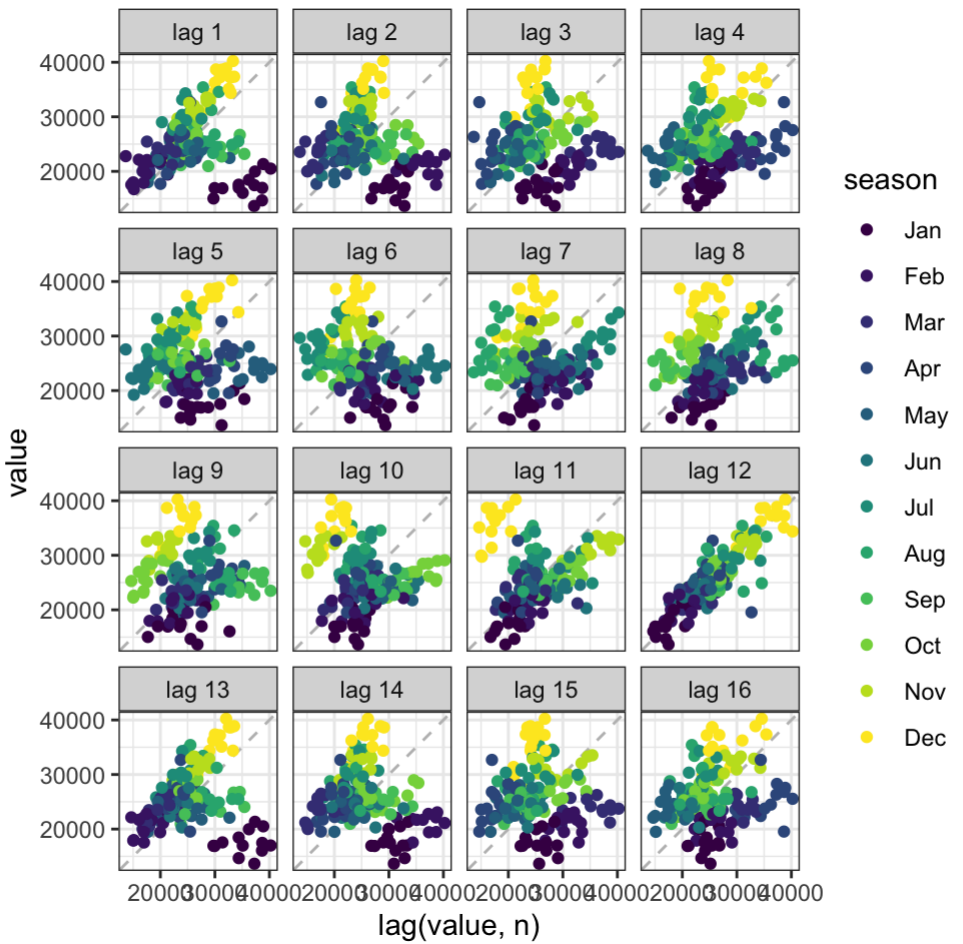

40000
30000
20000
value
1980 Jan
1985 Jan
1990 Jan
1995 Jan
index [1M]

Autocorrelation plots



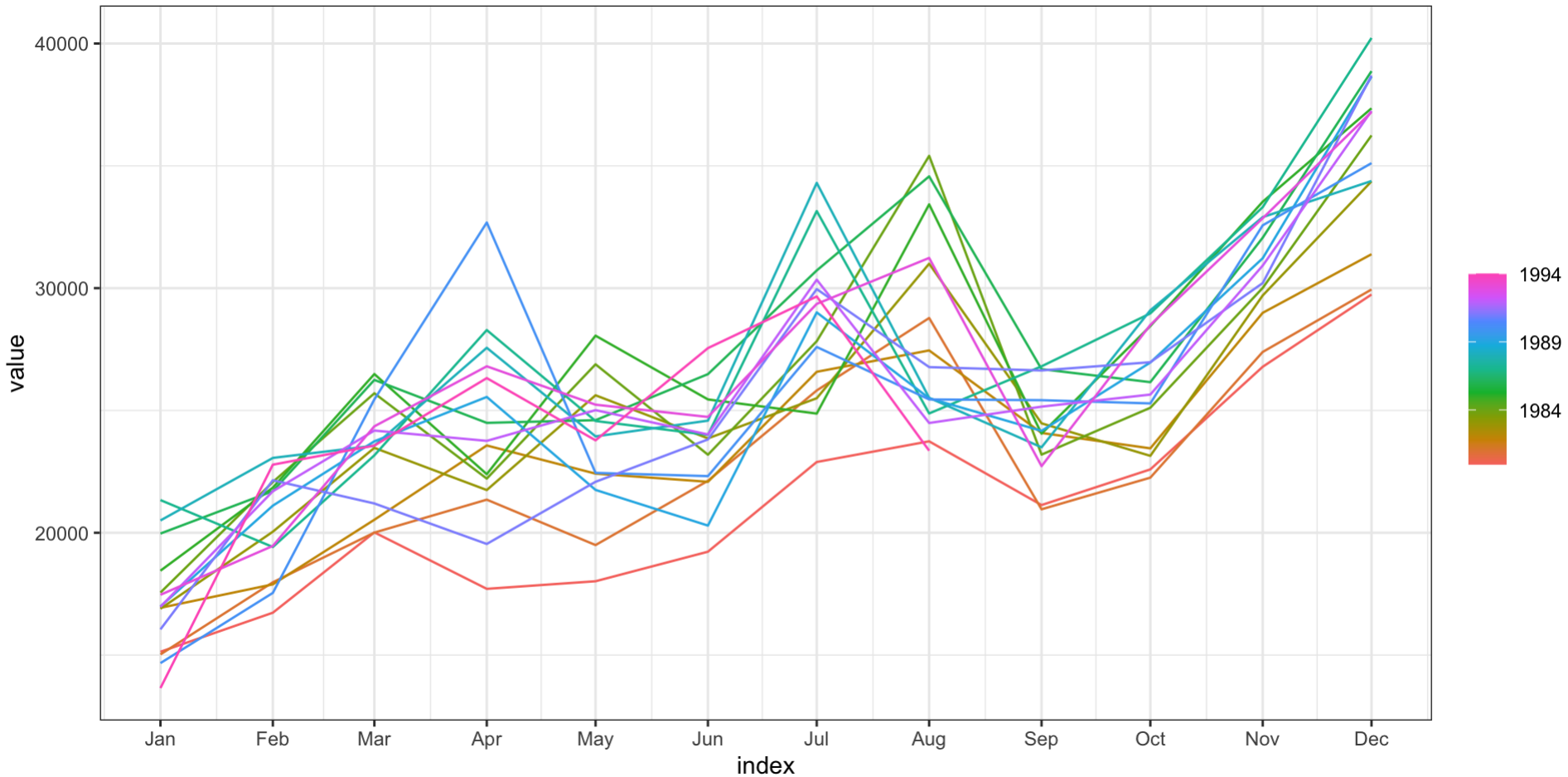
Lag plot

```
1  feasts::gg_lag(aus_wine, lags = 1:16, geom = "point")
```



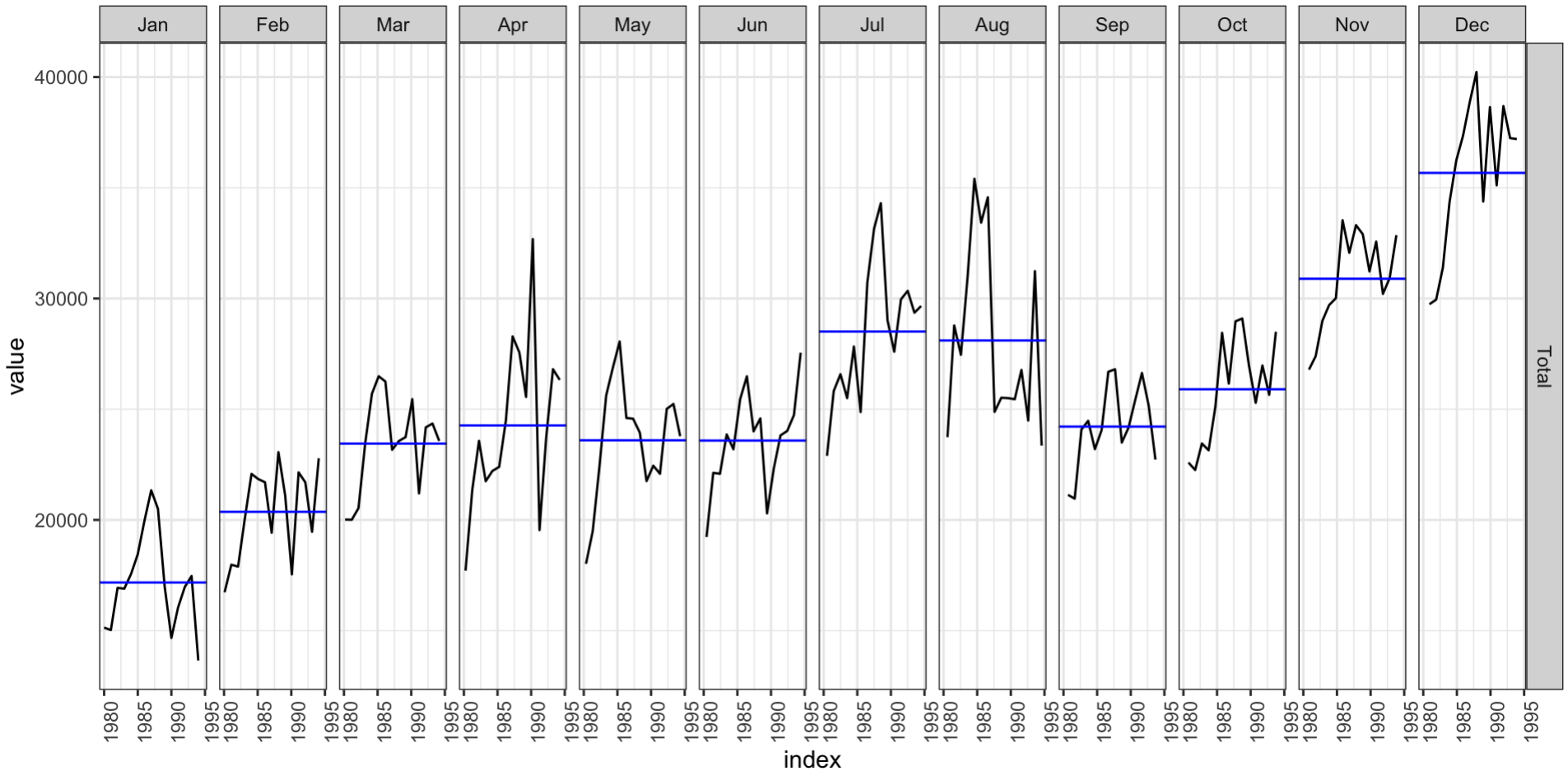
Seasonal plot

```
1  feasts::gg_season(aus_wine)
```



Subseries plot

```
1  feasts::gg_subseries(aus_wine)
```



A model?

```r
l = lm(value ~ lag(value,12), data = aus_wine)
summary(l)
```

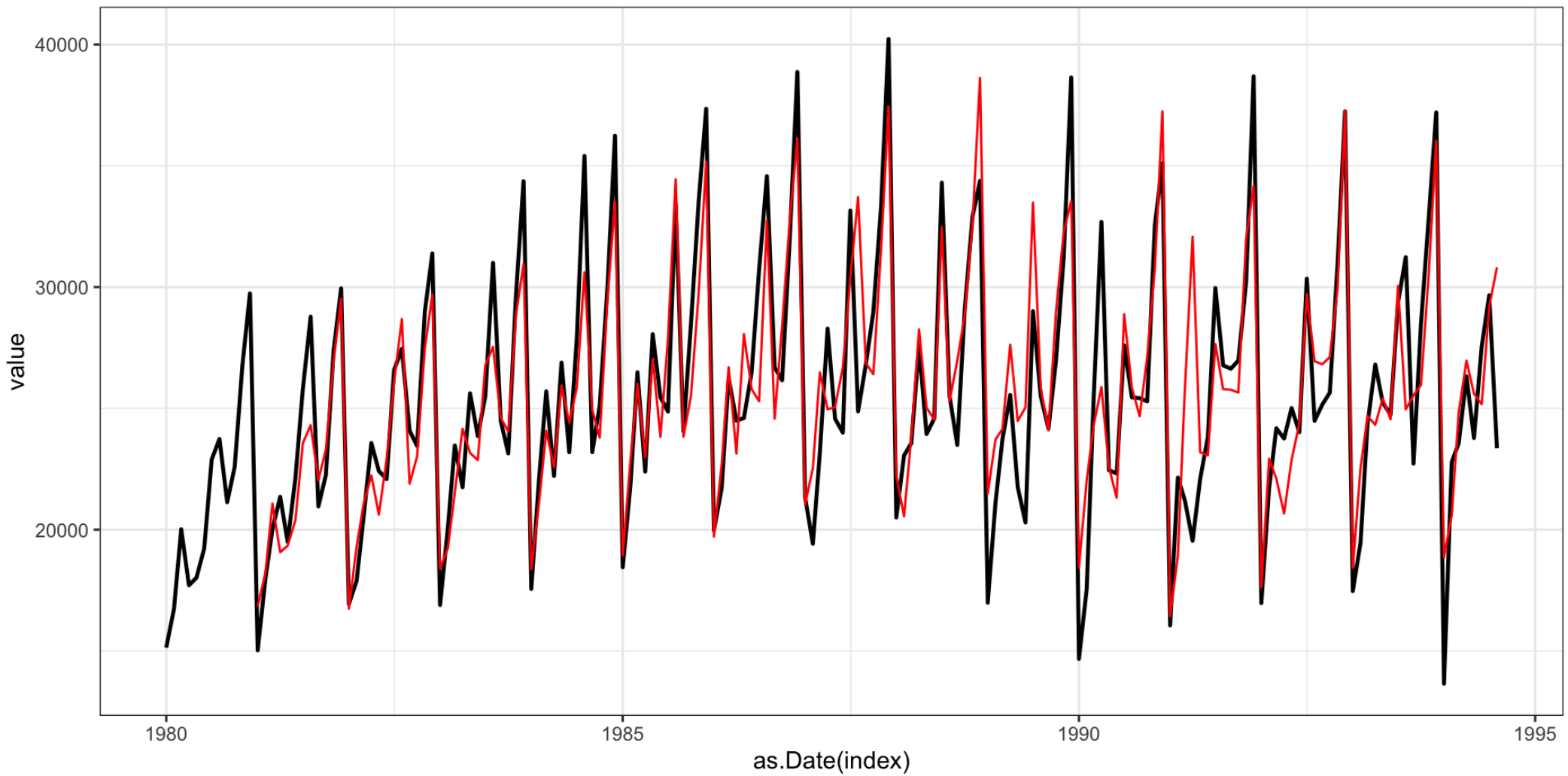
```
Call:
lm(formula = value ~ lag(value, 12), data = aus_wine)

Residuals:
     Min       1Q   Median       3Q      Max 
-12529.5  -1226.2    116.1   1744.0   6802.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    3.692e+03  9.903e+02   3.728 0.000267 ***
lag(value, 12) 8.684e-01  3.824e-02  22.707  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2594 on 162 degrees of freedom
```

Predictions



Observed vs predicted