Estimation for ARMA(p,q) models[1],[2]

We have n observations, $x_1, \dots, x_n$, from a causal and invertible ARMA(p,q) process, where p and q are known. We want to estimate the parameters of the model $\phi_1, \dots, \phi_p$; $\theta_1, \dots, \theta_q$ and σ_w^2.

[1] Robert Shumway, David Stoffer. Time Series Analysis and its Applications with R Examples, 4th edition, Springer, 2017 (chapter 3)
[2] Peter Brockwell, Richard Davis. Time Series: Theory and Methods, Springer-Verlag, 1987 (chapter 8)

I. Method of moments

It consists of replacing population moments to sample moments and then determining the parameters in terms of the sample moments.

$E(X)$ replaced by $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$ - sample mean

$\gamma(h)$ replaced by $\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h}-\overline{x})(x_t-\overline{x})$

(good for $n \geq 50$, $h \leq \frac{n}{4}$)

I. Method of moments

For an AR(p) model, recall that

$$\left.\begin{aligned} \gamma(h) &= \phi_1 \gamma(h-1) + \dots + \phi_p \gamma(h-p), \quad h = 1, \dots, p \\ \sigma_w^2 &= \gamma(0) - \phi_1 \gamma(1) - \dots - \phi_p \gamma(p) \end{aligned}\right\} \text{Yule-Walker equations}$$

In matrix form, the Yule-Walker equations are:

$$\Gamma_p \phi = \gamma_p, \quad \sigma_w^2 = \gamma(0) - \phi' \gamma_p, \quad \text{where}$$

$$\Gamma_p = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(p-1) \\ \vdots & & & \\ \vdots & & & \\ \gamma(p-1) & & \dots & \gamma(0) \end{pmatrix} \qquad \phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \vdots \\ \phi_p \end{pmatrix} \qquad \gamma_p = \begin{pmatrix} \gamma(1) \\ \vdots \\ \vdots \\ \gamma(p) \end{pmatrix}$$

We replace $\gamma(h)$ by $\hat{\gamma}(h)$

I. Method of moments

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \cdot \hat{\gamma}_p$$

$$\hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\gamma}_p' \hat{\Gamma}_p^{-1} \hat{\gamma}_p$$

} the Yule - Walker estimators

or, equivalent

$$\hat{\phi} = \hat{R}_p^{-1} \hat{\rho}_p$$

$$\hat{\sigma}_w^2 = \hat{\gamma}(0)[1 - \hat{\rho}_p' \hat{R}_p^{-1} \hat{\rho}_p],$$

where $\hat{R}_p = \dfrac{\hat{\Gamma}_p}{\hat{\gamma}(0)}$ and $\hat{\rho}_p = \dfrac{\hat{\gamma}_p}{\hat{\gamma}(0)}$

I. Method of moments

<u>Proposition</u> For a causal AR process, the asymptotic behavior of the Yule-Walker estimators is normal:

$$\sqrt{n}\left(\hat{\phi}-\phi\right) \xrightarrow{n\to\infty} N\left(0, \sigma_w^2 \Gamma_p^{-1}\right)$$

$$\hat{\sigma}_w^2 \xrightarrow{n\to\infty} \sigma_w^2 \qquad (n \text{ is the sample size})$$

We can avoid calculating the inverse of $\hat{\Gamma}_p$ (or $\hat{R}_p$) by replacing $\gamma(h)$ with $\hat{\gamma}(h)$ in the Durbin-Levinson algorithm. We compute iteratively $\hat{\phi}_h = (\hat{\phi}_{h1}, \dots, \hat{\phi}_{hh})'$ for $h = 1, 2, \dots$

I. Method of moments

from C4

The Durbin-Levinson algorithm

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)} \qquad v_0 = \gamma(0)$$

$$\phi_{nn} = \frac{\gamma(n) - \sum_{k=1}^{n-1} \phi_{n-1,k}\, \gamma(n-k)}{v_{n-1}}, \quad n \geq 1$$

$$\phi_{nk} = \phi_{n-1,k} - \phi_{nn} \cdot \phi_{n-1,n-k} \qquad k = 1, \dots, n-1 \qquad n \geq 2$$

$$v_n = v_{n-1} \cdot \left[1 - \phi_{nn}^2\right]$$

Order of computation:

$\phi_{11},\ v_1,\ \phi_{22},\ \phi_{21},\ v_2,\ \phi_{33},\ \phi_{31},\ \phi_{32},\ v_3 \ \dots$

the predictor $\hat{X}_{n+1} = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \dots + \phi_{nn} X_1$

I. Method of moments

<u>Proposition</u> For a causal AR(p) process, we have

$$\sqrt{n} \cdot \hat{\phi}_{hh} \xrightarrow{n \to \infty} N(0,1) \quad \text{for all } h > p.$$

<u>Obs</u>. The Yule-Walker estimators, computed iteratively in the Durbin-Levinson algorithm, are:

$$\hat{\phi}_p = (\hat{\phi}_{p1}, \dots, \hat{\phi}_{pp})$$

$$\sqrt{n}\left(\hat{\phi}_p - \phi\right) \xrightarrow{n \to \infty} N(0, \sigma_w^2 \Gamma_p^{-1})$$

We say that the estimator $\hat{\phi}_p$ is consistent for ϕ.

I. Method of moments

The estimates $\hat{\phi}_{hh}$ are extremely valuable, for two reasons:

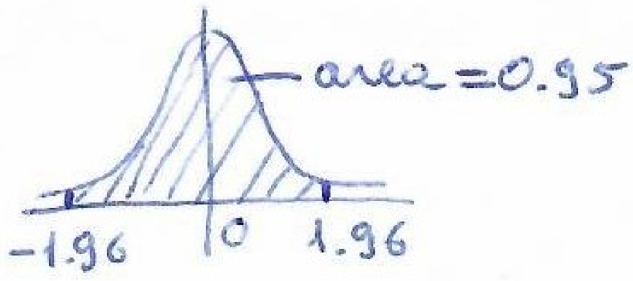
- to decide on the appropriateness of an AR model;
- then, to choose an appropriate order for the model to be fitted.

For an AR(p) process, the partial autocorrelations $\phi_{mm} = 0$ for $m > p$.

The estimator $\hat{\phi}_{mm} \xrightarrow{n \to \infty} N(0, \frac{1}{n})$ for $m > p$

I. Method of moments

if $X \sim N(\mu, \sigma^2)$ then $-1.96 < \frac{X-\mu}{\sigma} < 1.96$ with probability 0.95



This suggests a way to choose a preliminary value of p:

set $p = n$ initially

repeat
  if $|\hat{\phi}_{pp}| < 1.96 \cdot n^{-\frac{1}{2}}$ then it means that $\phi_{pp} = 0$
  $p = p - 1$
until $|\hat{\phi}_{pp}| \geq 1.96 \cdot n^{-\frac{1}{2}}$

I. Method of moments

The Yule-Walker estimators for AR(p) models are optimal as AR(p) models are linear models and Y-W estimators are essentially least squares estimators.

```
x=arima.sim(list(order=c(2,0,0), ar=c(1,-.9)), n=500)

x.yw = ar.yw(x, order=2)

x.yw$ar #coefficient estimates

[1] 0.9921305 -0.9003378

x.yw$var.pred  #wn variance estimate

[1] 1.076749
```

I. Method of moments

The method of moments for MA or ARMA models will not give optimal estimators because these models are nonlinear in the parameters.

- for an invertible MA(1) model,

$$X_t = \sum_{j=1}^{\infty} (-\theta)^j X_{t-j} + w_t \quad \text{– is nonlinear in } \theta$$

I. Method of moments

The method of moments for MA or ARMA models will not give optimal estimators because these models are nonlinear in the parameters.

- for an invertible MA(1) model,

$$X_t = \sum_{j=1}^{\infty} (-\theta)^j X_{t-j} + w_t \quad \text{– is nonlinear in } \theta$$

Obs. Analogically to the estimation for AR models using the Durbin-Levinson algorithm, the innovation algorithm can be used to estimate the parameters of an MA(q).

$$\hat{\theta}_q = (\hat{\theta}_{q1}, \ldots, \hat{\theta}_{qq})' \text{ estimator for } \boldsymbol{\theta}_q = (\theta_1, \ldots, \theta_q)'$$

But there is one important difference – for MA(q), the estimator $\hat{\theta}_q$ is not consistent for the true parameter vector $\boldsymbol{\theta}_q$.

II. Maximum likelihood and Least squares estimations for ARMA models (MLE and LSE)

Let $\mathbf{X}_n = (X_1, \dots, X_n)'$

$\hat{\mathbf{X}}_n = (\hat{X}_1, \dots, \hat{X}_n)'$ where $\hat{X}_1 = 0$ and

$\hat{X}_j = P_{\overline{sp}\{X_1, \dots, X_{j-1}\}} X_j, \quad j \geq 2$

$\Gamma_n = E(\mathbf{X}_n \cdot \mathbf{X}_n')$ - assume that is non-singular

if $\{X_t\}$ is Gaussian (i.e. the distribution functions are all multivariate normal) with mean zero, the likelihood of $\mathbf{X}_n$ is:

$$L(\Gamma_n) = (2\pi)^{-\frac{n}{2}} \cdot (\det \Gamma_n)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2} \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n\right)$$

II. MLE and LSE

If $\{X_t\}$ is not Gaussian, $L(\Gamma_n)$ is still a measure of the goodness of fit of the covariance matrix Γ_n to the data.

$$\Gamma_n = \Gamma_n(\phi, \theta, \sigma_w^2)$$

$$\phi = (\phi_1, \dots, \phi_p)'$$

$$\theta = (\theta_1, \dots, \theta_q)'$$

We want to maximize $L(\Gamma_n)$ with respect to $(\phi, \theta, \sigma_w^2)$.

II. MLE and LSE

The direct calculation of $\det \Gamma_n$ and Γ_n^{-1} in $L(\Gamma_n)$ can be avoided, using instead the recursive formulas from the innovation algorithm.

The mean squared error $E\left[(X_{n+1} - \hat{X}_{n+1})^2\right] = V_n$

$$r_n \stackrel{def}{=} \frac{V_n}{\sigma_w^2}$$

II. MLE and LSE

The likelihood function can be written as (for details, see Brockwell & Davis, pages 247 - 250):

$$L(\Gamma_n) = L(\phi, \theta, \sigma_w^2) = (2\pi\sigma_w^2)^{-\frac{n}{2}} \cdot (r_0 \cdot r_1 \cdots r_{n-1})^{-\frac{1}{2}}$$
$$\cdot \exp\Big(-\frac{1}{2}\sigma_w^{-2} \underbrace{\sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / r_{j-1}}_{\overset{\text{not}}{=}\, S(\phi, \theta)}\Big)$$

$$\ln L(\phi, \theta, \sigma_w^2) = -\frac{n}{2}\ln(2\pi\sigma_w^2) - \frac{1}{2}\sum_{j=1}^{n} \ln r_{j-1} - \frac{1}{2}\sigma_w^{-2} S(\phi, \theta)$$

$\hat{X}_j$ and r_j are independent of σ_w^2

II. MLE and LSE

$$\frac{\partial \ln L(\phi, \theta, \sigma_w^2)}{\partial \sigma_w^2} = 0$$

$$-\frac{n}{2} \cdot \frac{1}{\sigma_w^2} + \frac{1}{2} \cdot \frac{1}{(\sigma_w^2)^2} \cdot S(\phi, \theta) = 0$$

$\hat{\sigma}_w^2 = \frac{1}{n} \cdot S(\hat{\phi}, \hat{\theta})$, where $\hat{\phi}$ and $\hat{\theta}$ are the values of ϕ, θ which minimize

$$\ell(\phi, \theta) = \ln\left(n^{-1} S(\phi, \theta)\right) + n^{-1} \sum_{j=1}^{n} \ln r_{j-1}$$

$\ell(\phi, \theta)$ is called the concentrated (or reduced) likelihood.

$\hat{\phi}$ and $\hat{\theta}$ that minimize $\ell(\phi, \theta)$ are determined using non-linear numerical optimization algorithms (e.g. Newton-Raphson algorithm).

II. MLE and LSE

An alternative estimation procedure is to minimize $S(\phi, \theta)$ with respect to ϕ and θ. The estimators obtained will be referred to as the "least squares" estimators $\tilde{\phi}$ and $\tilde{\theta}$ of ϕ and θ.

Proposition For causal and invertible ARMA processes, the maximum likelihood and the least squares estimators provide optimal estimators of σ_w^2, ϕ and θ.

II. MLE and LSE

```
x=arima.sim(list(order=c(2,0,1), ar=c(1,-.9),ma=c(0.9)), n=100)

v=arima(x, order = c(2,0,1))

v$coef ar1 ar2 ma1 intercept
       1.0061145 -0.8754776 0.9847590 -0.3537682

v$sigma2
[1] 0.8888034
```

```
x=arima.sim(list(order=c(2,0,1), ar=c(1,-.9),ma=c(1.7)), n=100)

v=arima(x, order = c(2,0,1))

v$coef ar1 ar2 ma1 intercept
       0.9809513 -0.8705035 0.6825593 0.4325465

v$sigma2
[1] 2.80407
```

Dealing with non-stationary time series

Two approaches were developed in order to make a non-stationary time series stationary (Brockwell and Davis, pages 14-25):

- estimate and extract the deterministic components m_t and s_t, with the hope that the residual Y_t will be stationary;

$$X_t = m_t + s_t + Y_t$$

- apply difference operators repeatedly to the data x_t until we get a realization of a stationary process

$$\nabla X_t = X_t - X_{t-1}$$

Dealing with non-stationary time series

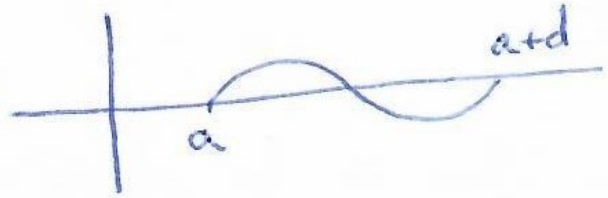
We assume that the non-stationary $\{X_t\}$ is of the form:

$$X_t = m_t + s_t + Y_t,$$

where m_t is a slowly changing function called "trend component"

s_t is the "seasonal component" - is a function with known period d and $\sum_{t=a}^{a+d} s_t = 0$

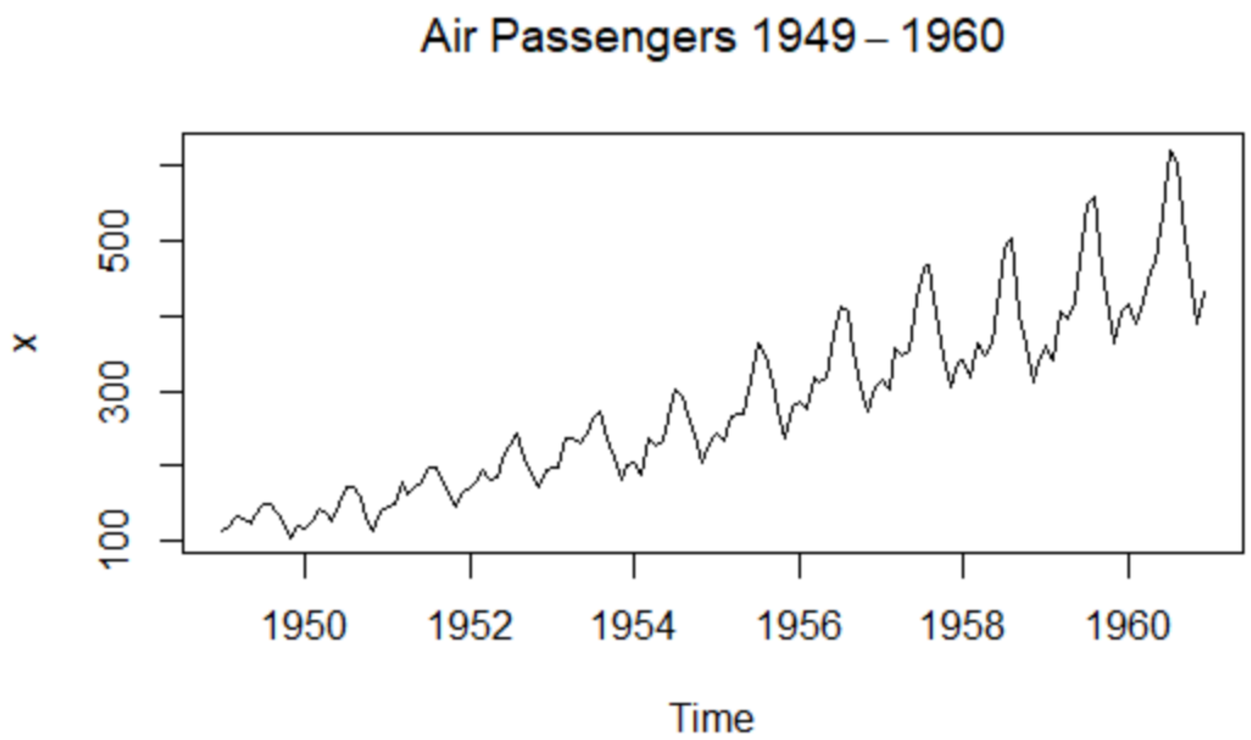
$s_t = s_{t+d}$



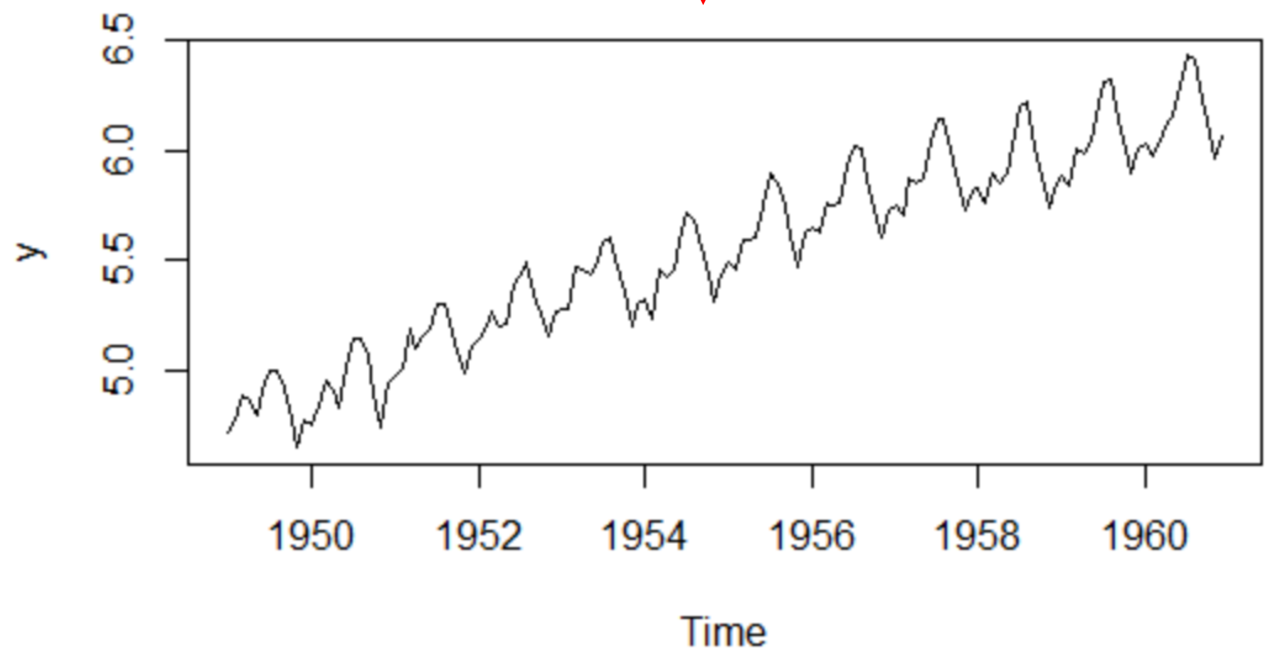
Y_t is the random noise component – is stationary.

In practice, preliminary transformation of the data (e.g. log, square root, cube root etc.) may be necessary to make the transformed data compatible with this model.

Dealing with non-stationary time series



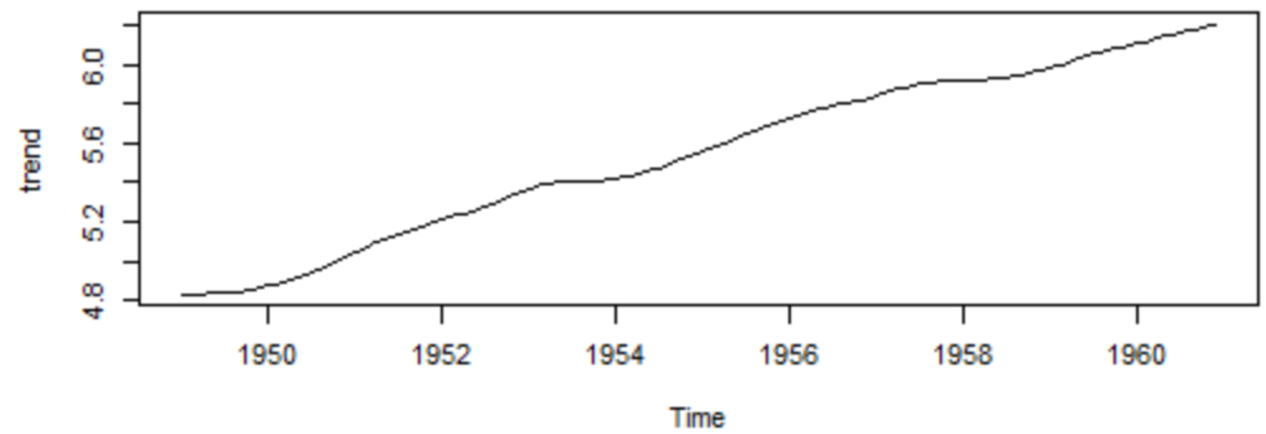
log



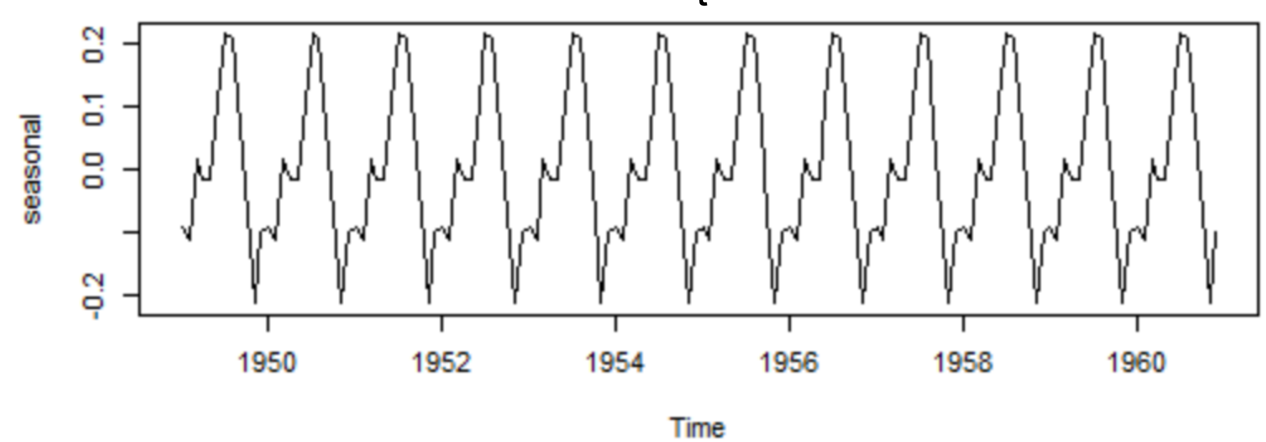
$m_t+s_t+Y_t$

Dealing with non-stationary time series

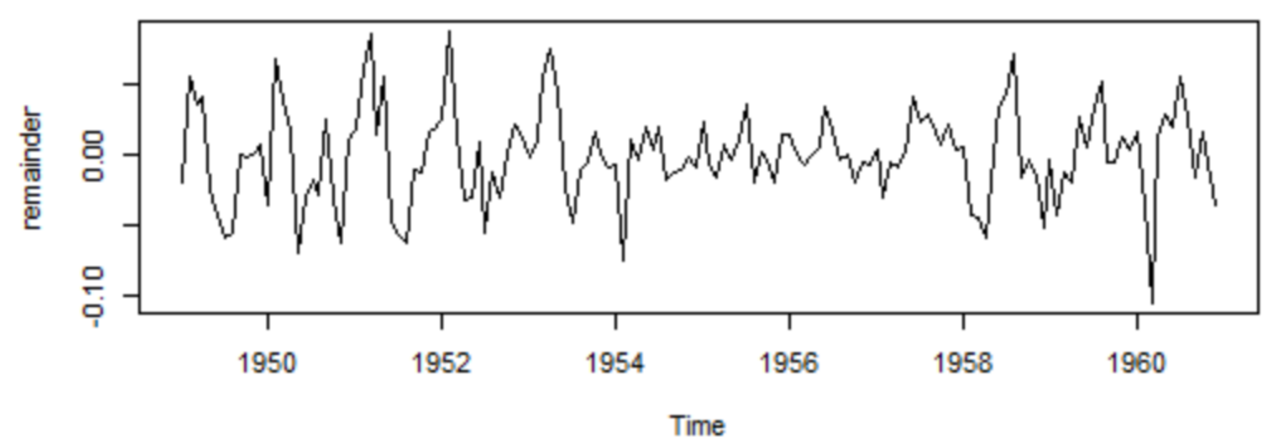
m_t



s_t



Y_t



Dealing with non-stationary time series

The ARIMA(p,d,q) model

It is a process which, after differencing a number of times, reduces to an ARMA(p,q) process.

Def. if d is a non-negative integer, then $\{X_t\}$ is said to be an ARIMA(p,d,q) process if $Y_t = (1-B)^d X_t$ is a causal ARMA(p,q) process.

$$\phi(B)(1-B)^d X_t = \theta(B) w_t, \quad w_t \sim wn(0, \sigma_w^2)$$

Dealing with non-stationary time series

<u>Example</u> $\{X_t\}$ is an ARIMA(1,1,0) process if

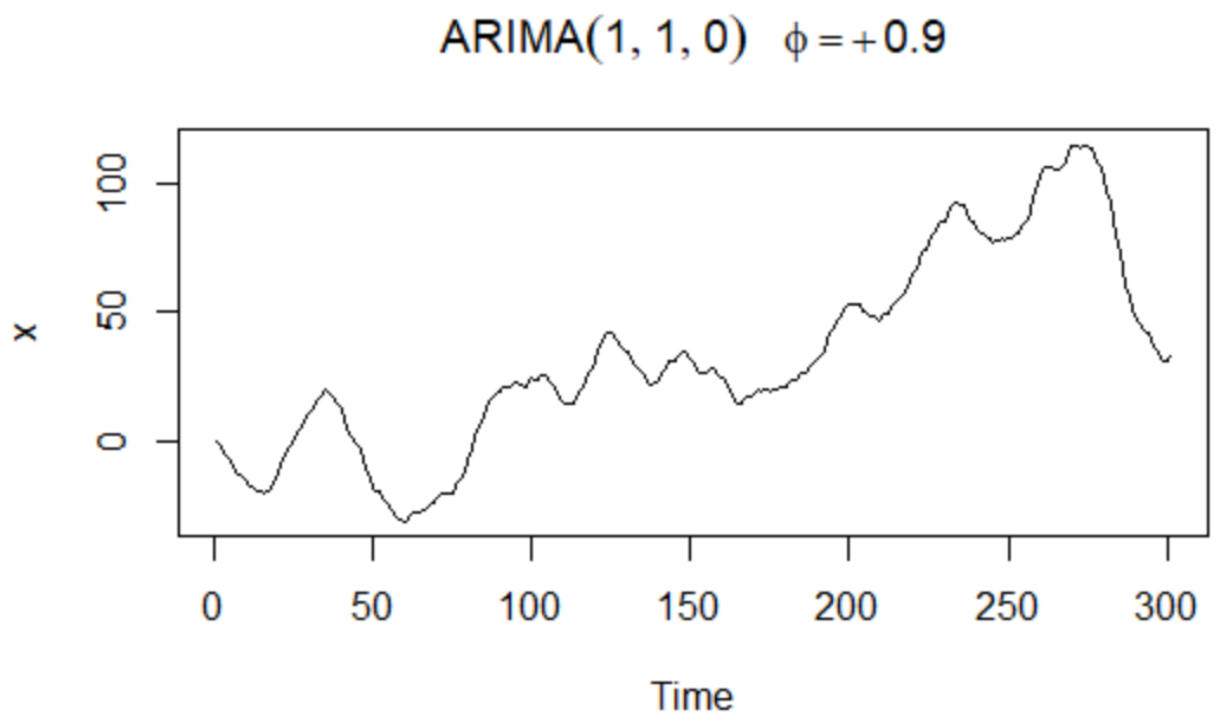
$$(1-\phi B)\underbrace{(1-B)X_t}_{Y_t} = w_t\ , \quad \phi \in (-1,1)$$

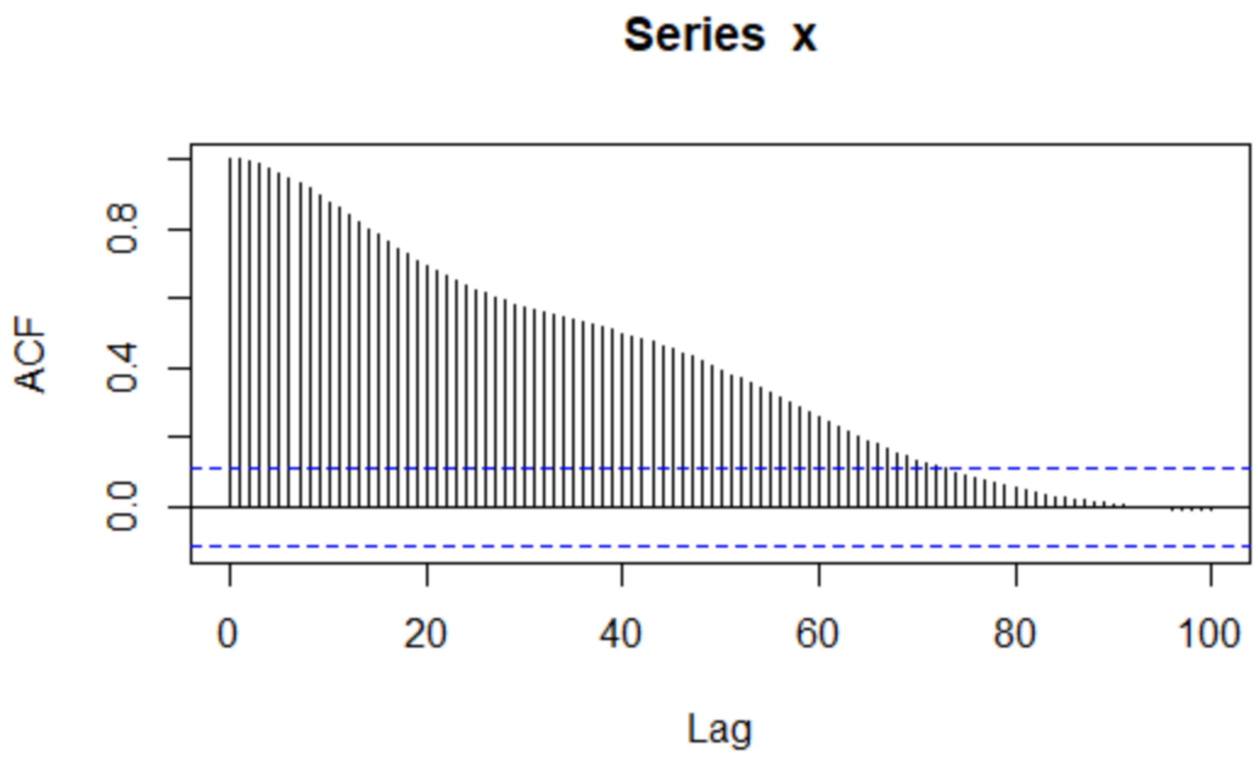
$(1-\phi B)Y_t = w_t$ is causal as $\phi \in (-1,1)$

<u>Obs</u>. Estimation of ϕ, θ and σ_w^2 is based on the observed differences $(1-B)^d X_t$.

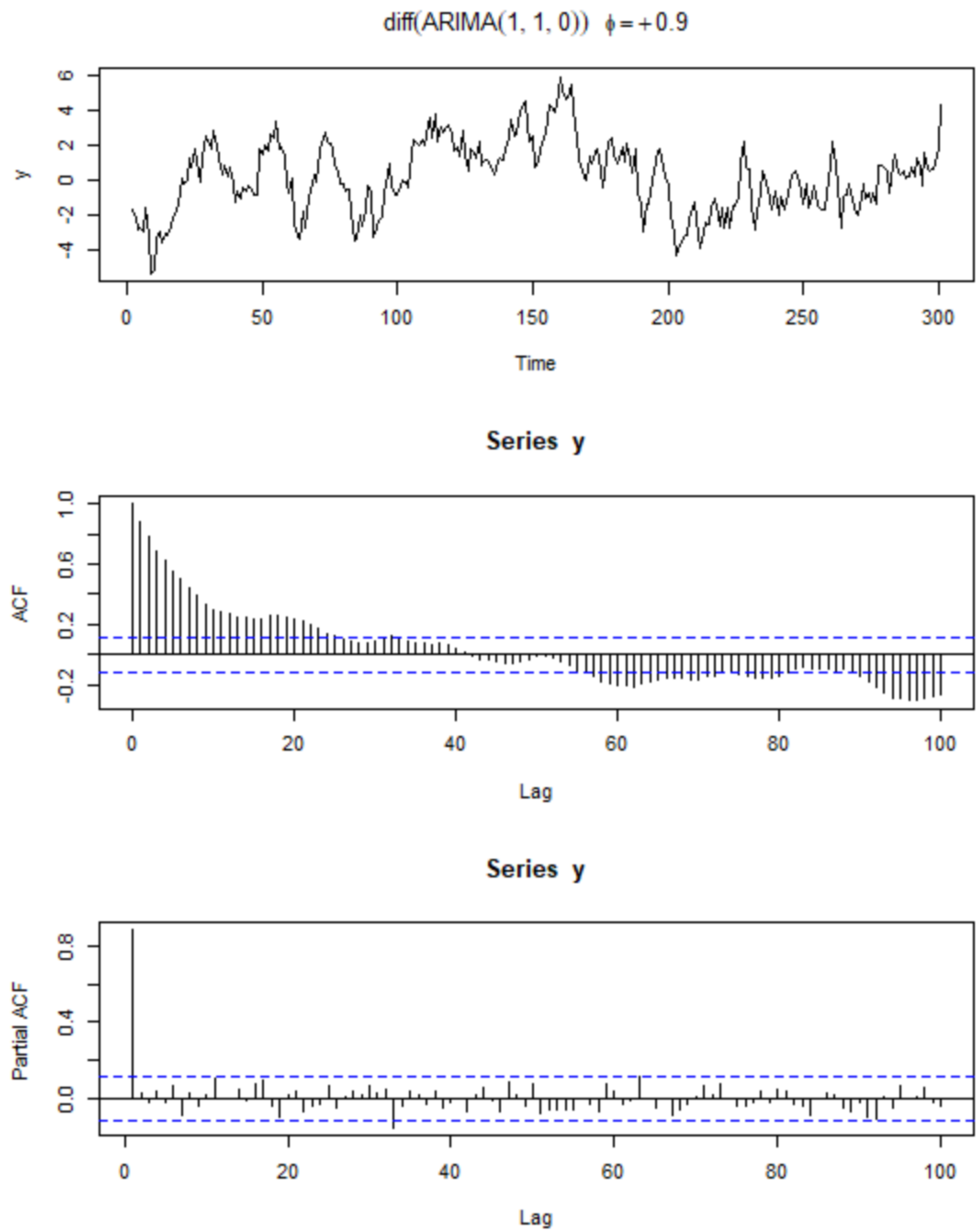
Dealing with non-stationary time series

Obs. A slowly decaying positive ACF is a distinctive feature of the data which suggests the appropriateness of an ARIMA model.





Dealing with non-stationary time series



Dealing with non-stationary time series

1). Random Walk with Drift is not stationary

$$X_t = \delta + X_{t-1} + W_t, \quad W_t \sim wn(0, \sigma_w^2)$$

Differencing will lead to a stationary process

$$\nabla X_t = X_t - X_{t-1} = \delta + W_t$$

Dealing with non-stationary time series

2) $X_t = m_t + Y_t$

$m_t = \beta_0 + \beta_1 t$

Y_t is stationary

$\nabla X_t = \beta_1 + Y_t - Y_{t-1} = \beta_1 + \nabla Y_t$ is stationary

Examples of (non) stationary time series*

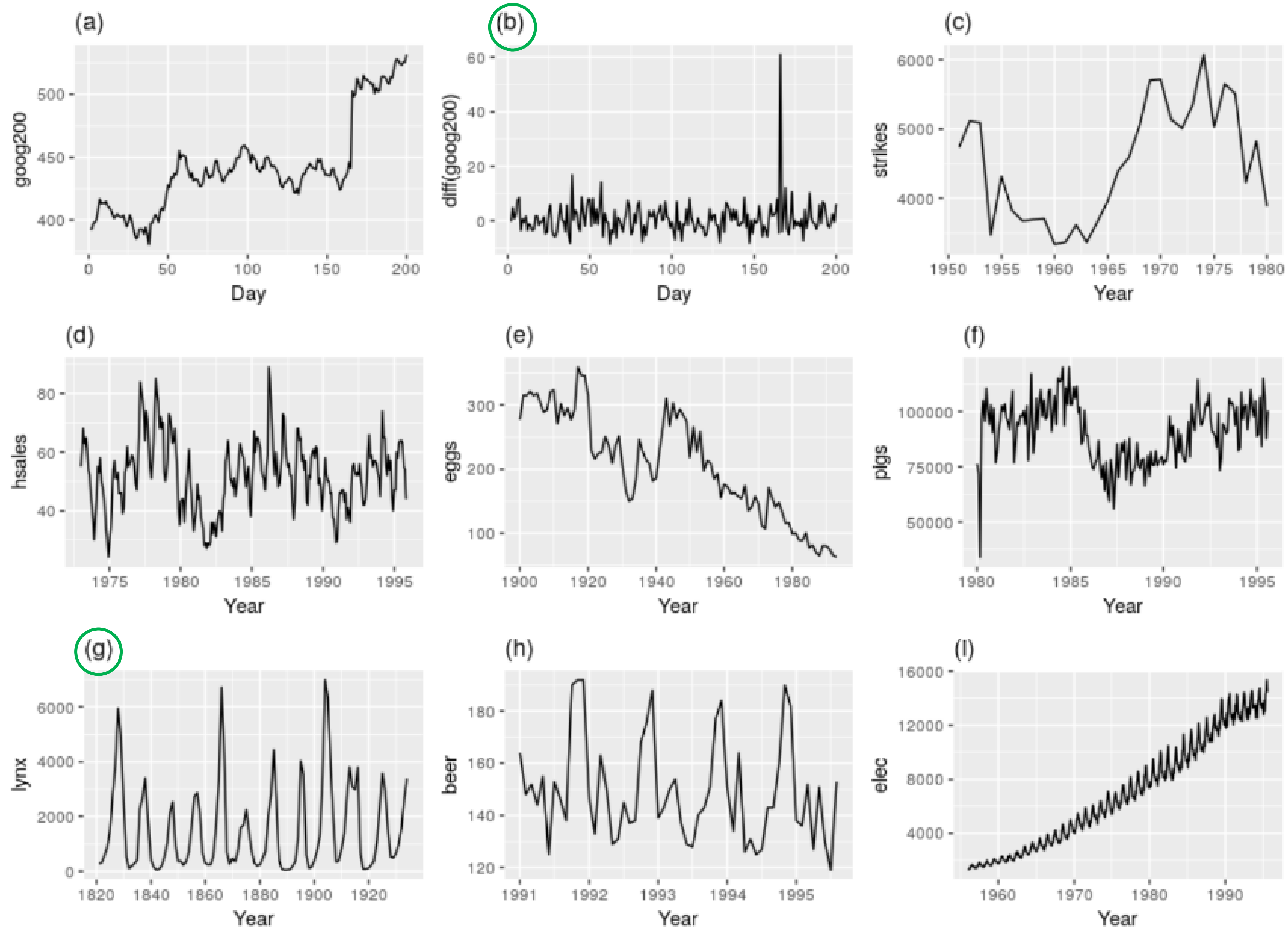


Figure 1: Nine examples of time series data; (a) Google stock price for 200 consecutive days; (b) Daily change in the Google stock price for 200 consecutive days; (c) Annual number of strikes in the US; (d) Monthly sales of new one-family houses sold in the US; (e) Annual price of a dozen eggs in the US (constant dollars); (f) Monthly total of pigs slaughtered in Victoria, Australia; (g) Annual total of lynx trapped in the McKenzie River district of north-west Canada; (h) Monthly Australian beer production; (i) Monthly Australian electricity production. [Hyndman & Athanasopoulos, 2018]

*taken from https://towardsdatascience.com/detecting-stationarity-in-time-series-data-d29e0a21e638