Time Series - Model Building[1],[2]

Given a set of observations, $x_1, \ldots, x_n$, how do we build an appropriate time series model to fit the data?

There are a few basic steps to follow:

[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 9)

I. Transform the data (if necessary)

Plot the data and take a look - if there are no apparent deviations from stationarity and the ACF is decreasing (rapidly), then we'll search for a suitable ARMA model to fit the mean-corrected data (i.e. $x_i - \overline{x}$ where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$).

Otherwise, it is necessary to apply transformations to the data (e.g. log, square root etc.). For example, if the variability in the data increases in time, the Box-Cox class of transformations could be used:

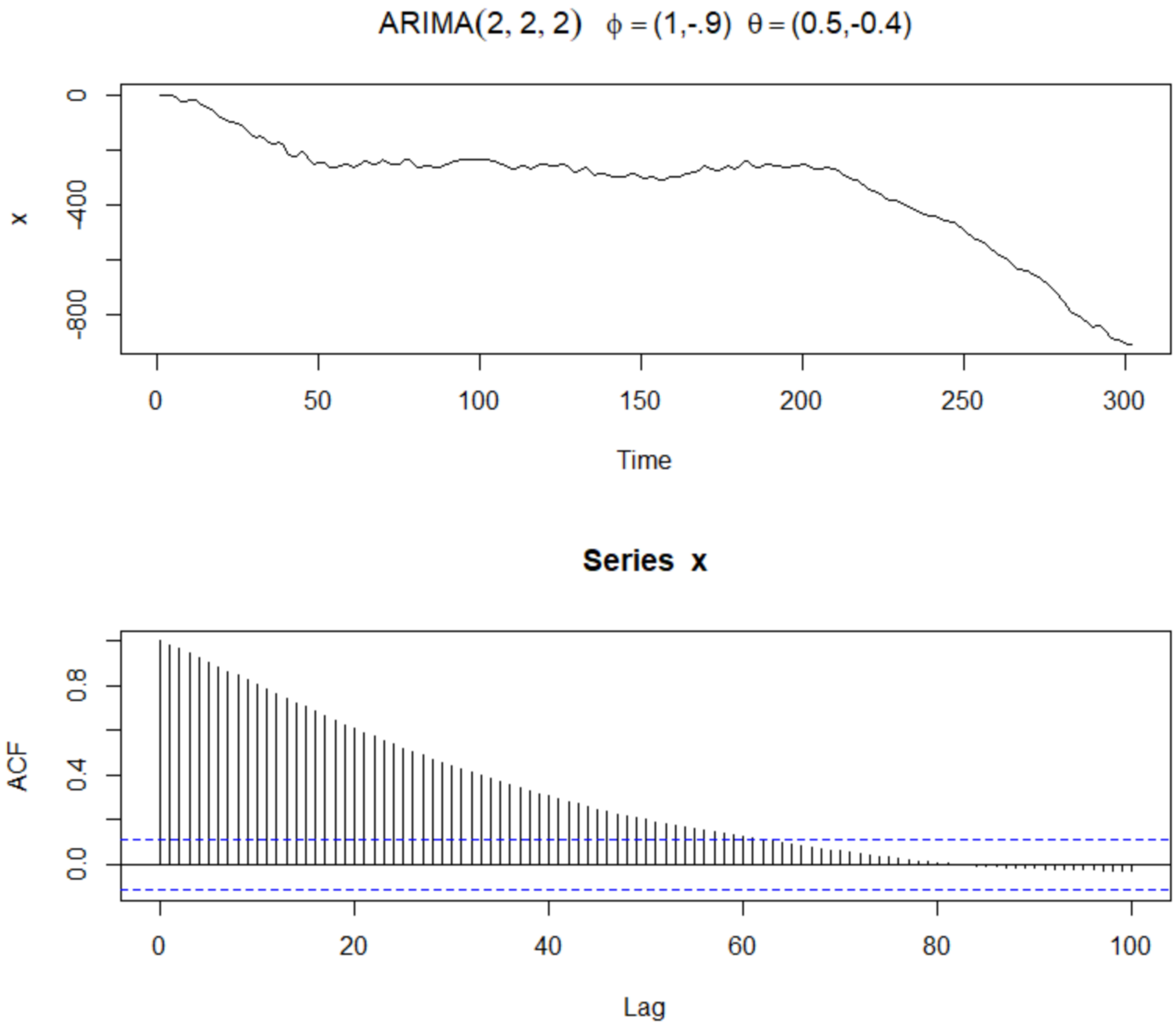
$$y_t = \begin{cases} (x_t^{\lambda} - 1) \cdot \frac{1}{\lambda}, & \lambda \neq 0 \\ \log x_t, & \lambda = 0 \end{cases}$$

(methods for choosing λ are available)

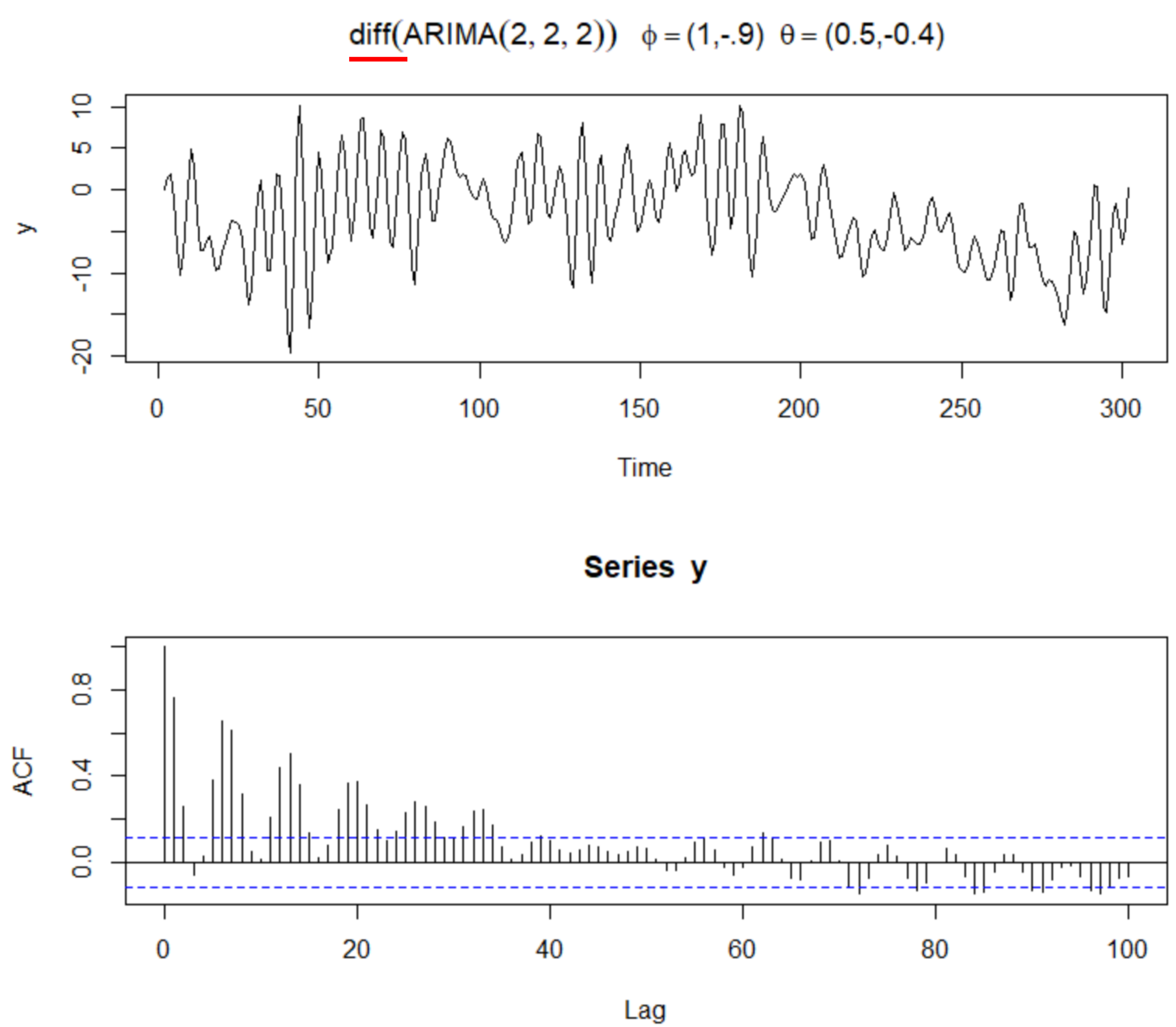
I. Transform the data (if necessary)

Trend and seasonality can be detected by visual inspection of the time series, but they are also characterized by slowly decaying and/or periodic sample ACF. They can be estimated and removed or/and we can do differencing for multiple times – difference the data once and plot ∇x_t; if further differencing is needed, compute $\nabla^2 x_t$ and plot it again and so on.

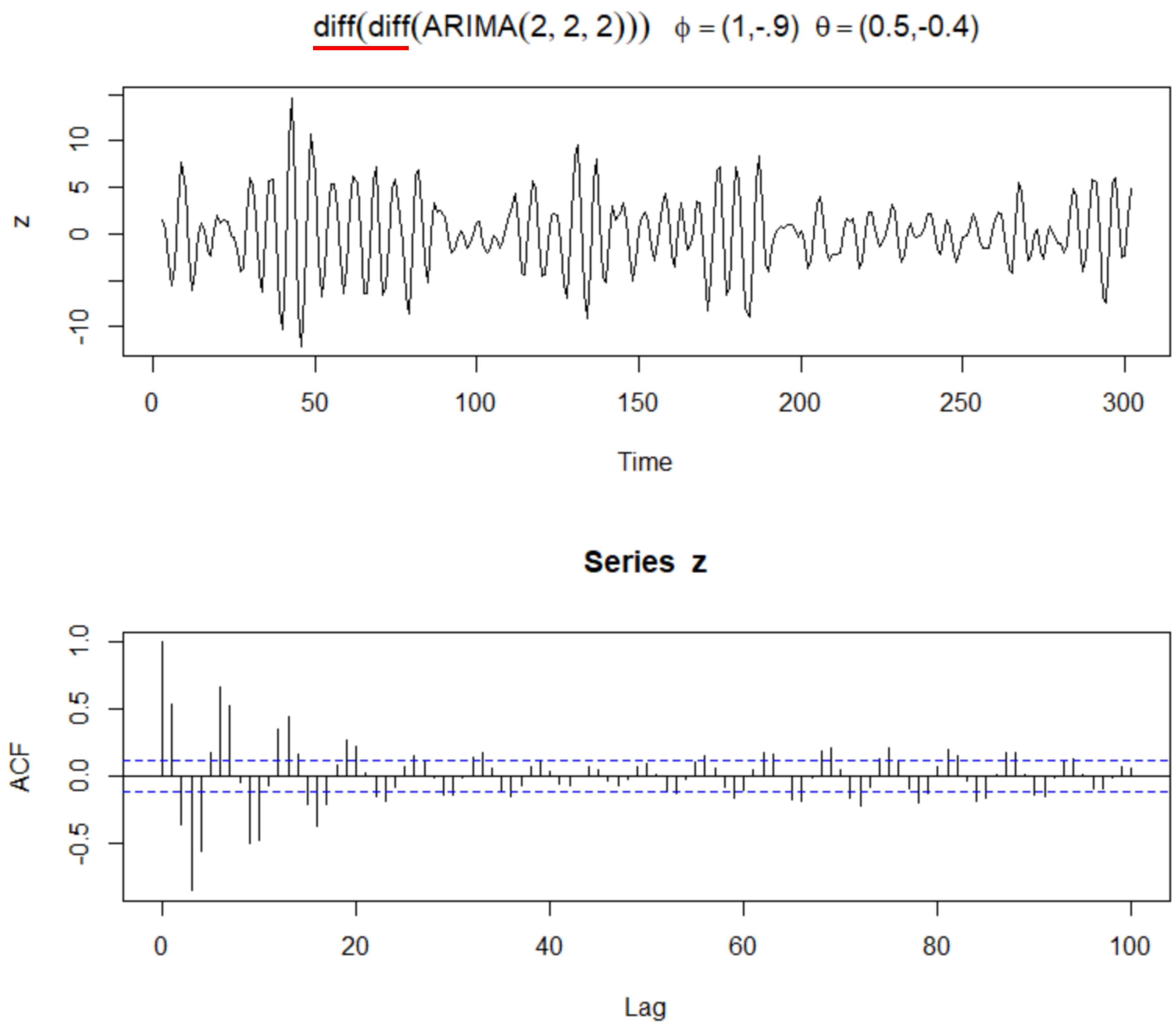
I. Transform the data (if necessary)



I. Transform the data (if necessary)



I. Transform the data (if necessary)



I. Transform the data (if necessary)

Obs. Be careful as overdifferencing may introduce dependencies in the data:

$$x_t = w_t \quad \text{- uncorrelated}$$

but $\nabla x_t = w_t - w_{t-1}$ is MA(1)

Obs. Remember that a slowly decaying sample ACF indicates that further differencing may be needed.

Let d = order of differencing

$$\nabla^d = (1-B)^d \qquad B \text{ - backshift operator}$$

II. Search for the best ARMA(p,q) model

Now $x_1, \ldots, x_n$ denote the mean-corrected transformed values

We search now for the ARMA(p,q) model to fit best the data.

The ACF and the PACF give us information about preliminary values of p and q. Because we deal with estimates, it will not be always clear whether the sample ACF or PACF is decaying or cutting off.

II. Search for the best ARMA(p,q) model

It appears that the higher the values of p and q (that is, a model with more "memory"), the better the model will fit the data. But we must pay attention to overfitting.

Different criteria have been developed (e.g. AIC, BIC, CAT) to prevent overfitting, by penalizing additional parameters

II. Search for the best ARMA(p,q) model

Akaike's AIC criterion

- for an ARMA(p,q) model, we denote by

$\hat{\phi}_{pq}$ = maximum likelihood estimator for $(\phi_1, \dots, \phi_p)'$

$\hat{\theta}_{pq}$ = MLE for $(\theta_1, \dots, \theta_q)'$

$AIC = n \cdot \ell(\hat{\phi}_{pq}, \hat{\theta}_{pq}) + 2(p+q)$, where $\ell(\cdot)$ is the concentrated likelihood (see the previous course)

- for $p, q = 1, 2, \dots$ we choose the values of p and q for which AIC is minimum.

II. Search for the best ARMA(p,q) model

If the ACF and the PACF clearly indicate towards an AR(p) or an MA(q), we fit models of orders 1, 2, 3,... until the minimum value of AIC is found.

If the ACF and PACF are more difficult to interpret (that is, the ACF does not cut off after lag q and the PACF does not cut off after lag p), the search for the best model can be done as following:

II. Search for the best ARMA(p,q) model

- fit ARMA models of orders (1,1), (2,2), ... to the data and select the model of order (p,p) with the smallest value of AIC;

- start from the ARMA(p,p) model and eliminate one or more coefficients and compute the AIC value for each reduced model

$$\left.\begin{array}{l} \text{ARMA}(p, p-1), \ldots, \text{ARMA}(p,1) \\ \text{ARMA}(p-1, p), \ldots, \text{ARMA}(1,1) \end{array}\right\} \text{select the model with smallest AIC}$$

II. Search for the best ARMA(p,q) model

```
x=arima.sim(list(order=c(2,2,2), ar=c(1,-.9),ma=c(0.5,-.4)), n=300) #the same observations as in slide 4
```

```
model1<-arima(x ,order=c(1,2,1))
model1$aic # or AIC(model1)
[1] 1265.129
mean(model1$residuals^2)
[1] 3.780747
```

```
model2<-arima(x ,order=c(2,2,2))
model2$aic
[1] 861.1276
model2$coef
ar1 ar2 ma1 ma2
1.0379151 -0.8980195 0.5141287 -0.4176244
mean(model2$residuals^2)
[1] 0.9699251
```

```
model3<-arima(x ,order=c(3,2,3))
model3$aic
[1] 856.2563
mean(model3$residuals^2)
[1] 0.9405561
```

```
model4<-arima(x ,order=c(4,2,4))
Warning message: In arima(x, order = c(4, 2, 4)) :
possible convergence problem: optim gave code = 1
 model4$aic
[1] 859.7426
mean(model4$residuals^2)
 [1] 0.9388734
```

```
model5<-arima(x ,order=c(5,2,5))
Warning messages: 1: In log(s2) : NaNs produced 2: In
arima(x, order = c(5, 2, 5)) : possible convergence
problem: optim gave code = 1
model5$aic
[1] 863.4055
mean(model5$residuals^2)
 [1] 0.935537
```

III. Diagnostics

Diagnostics - the goodness of fit of an ARMA model to the data is judged by examining the residuals (i.e. the difference between the observed values and the corresponding predicted values obtained from the fitted model $X_t - \hat{X}_t$).

If the model is satisfactory, the standardized residuals

$$e_t = \frac{X_t - \hat{X}_t}{\sqrt{V_{t-1}}}, \text{ where } V_{t-1} = E[(X_t - \hat{X}_t)^2],$$

are i.i.d. with mean zero and variance one.

III. Diagnostics

Proposition. The sample autocorrelations of an i.i.d sequence are for large n approximately iid with distribution $N(0, \frac{1}{n})$.

Hence, a good check on the correlation structure of the residuals e_t is to plot the sample ACF $\hat{\rho}_e(h)$ along with the error bounds $\pm 1.96/\sqrt{n}$.

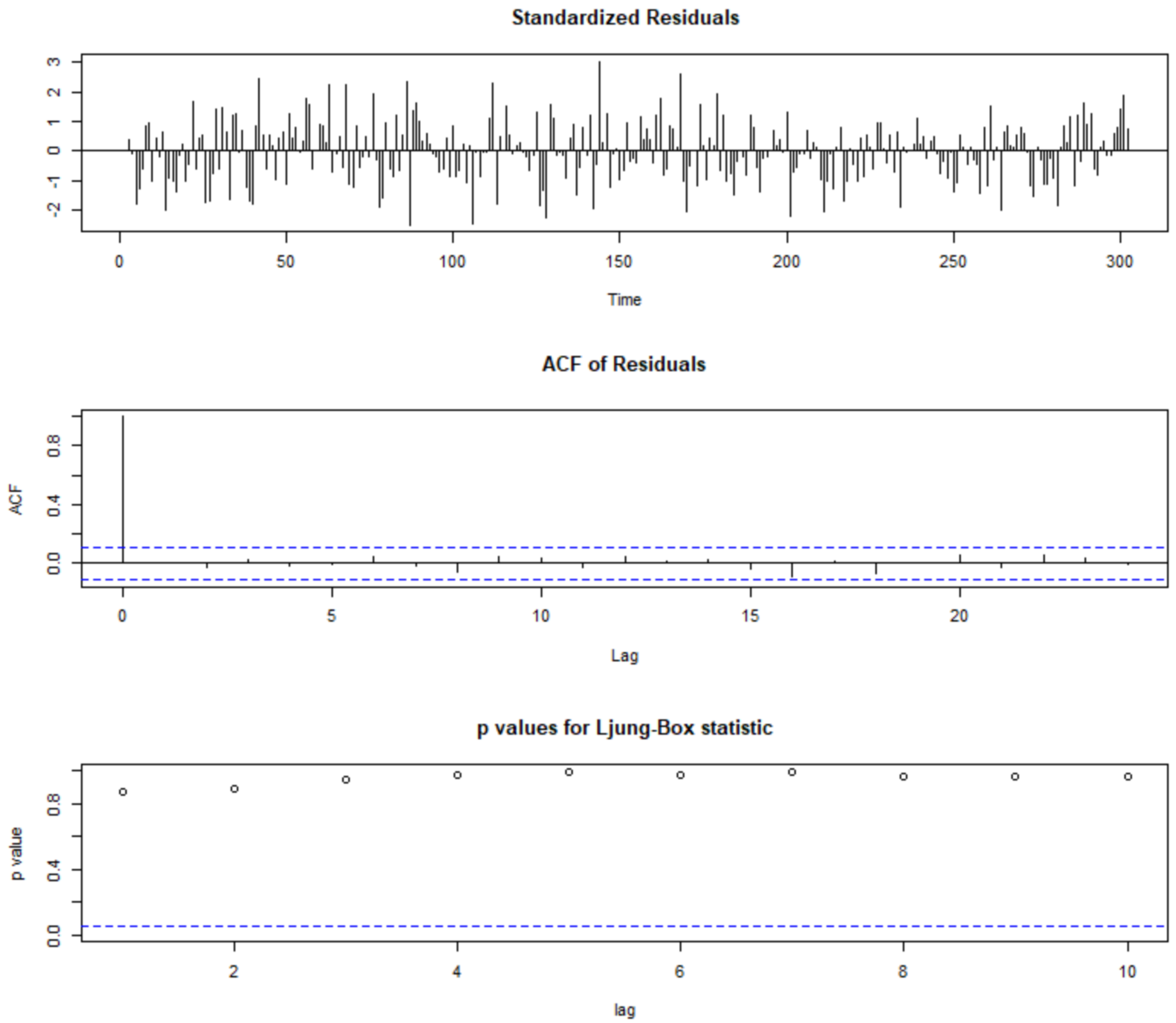
$\left[\text{if } \hat{\rho}_e(h) \sim N(0, \frac{1}{n}) \text{ then } -1.96 < \frac{\hat{\rho}_e(h)}{\frac{1}{\sqrt{n}}} < 1.96 \text{ with probability } 0.95\right]$

In addition to the visual inspection of $\hat{\rho}_e(h)$, the Ljung-Box test can be used to check the magnitudes of $\hat{\rho}_e(h)$ as a group.

III. Diagnostics

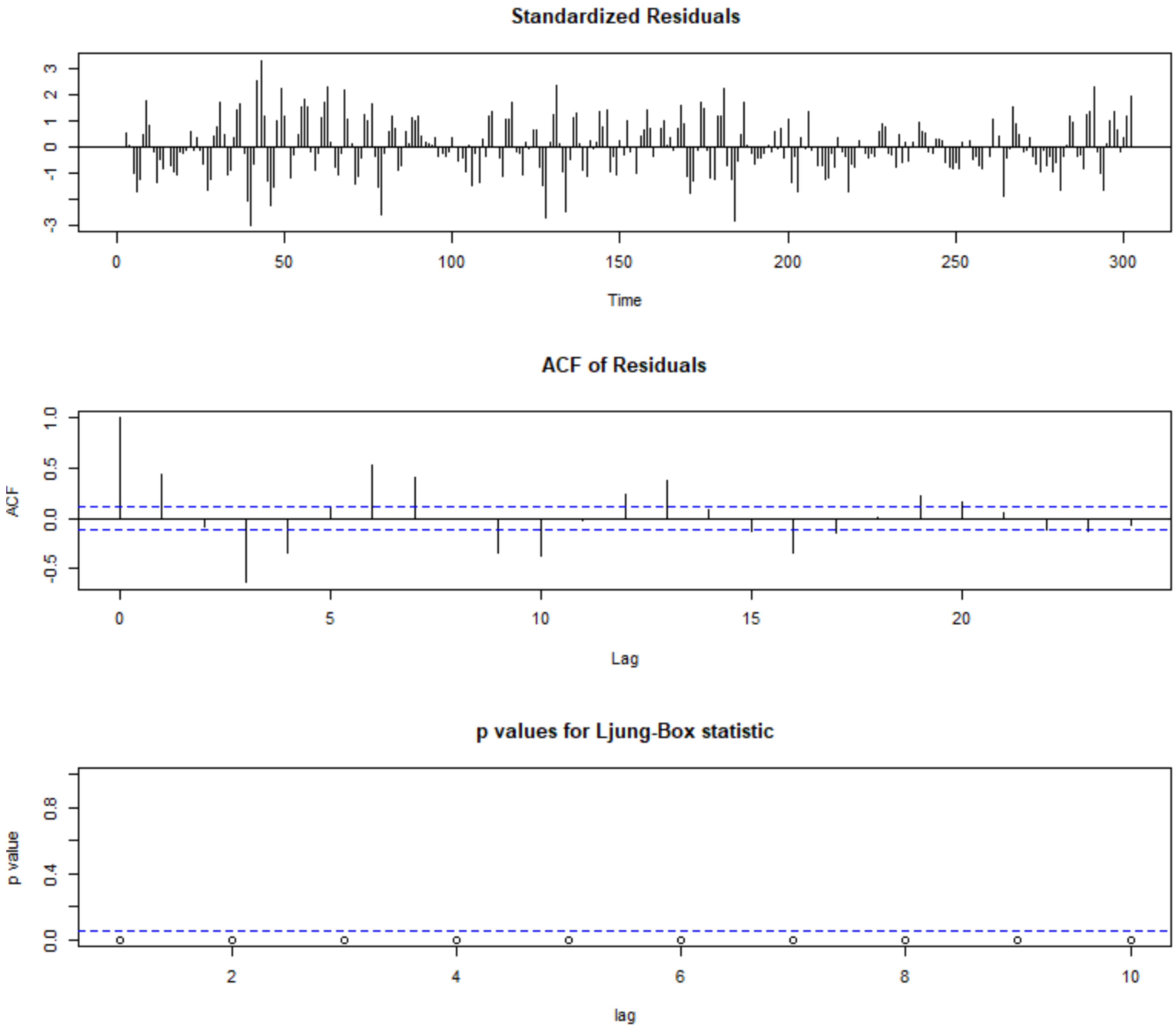
model3 ARIMA(3,2,3)
(good fit)

`tsdiag(model3)`



III. Diagnostics

ARIMA(0,2,3)
(bad fit)

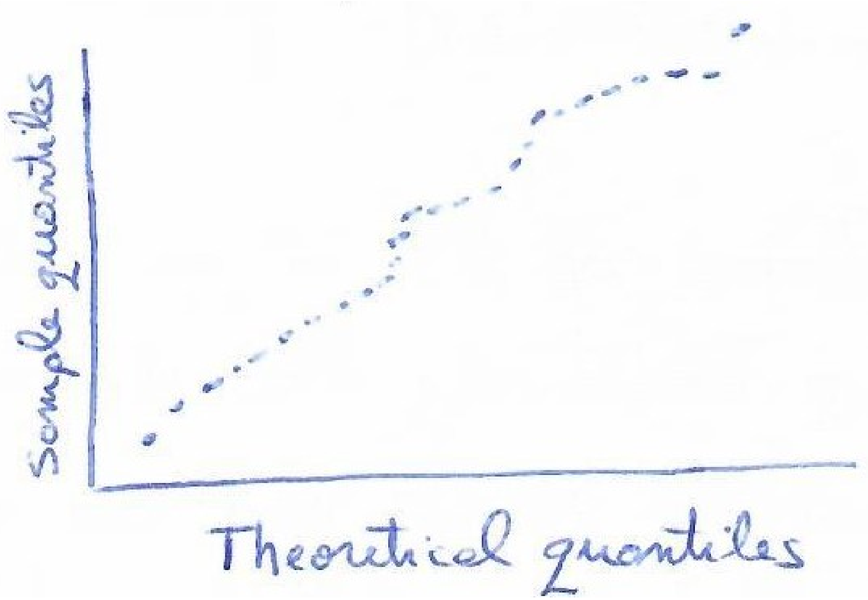


III. Diagnostics

If $\{e_t\}$ is Gaussian, then uncorrelated $\{e_t\}$ means independent $\{e_t\}$. Investigation of normality can be done visually by looking at the histogram of the residuals. In addition to that, a Q-Q plot can be used to identify departures from normality.

Q-Q comes from quantile-quantile

- the Q-Q plot compares the quantiles of the data against the quantiles of the desired distribution (normal in our case)

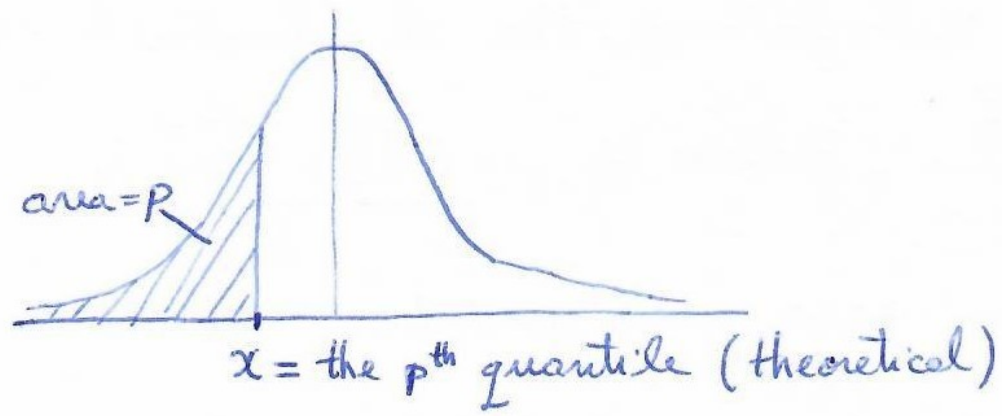


- if the plot is approximately linear, then the data come from the theoretical distribution.

III. Diagnostics

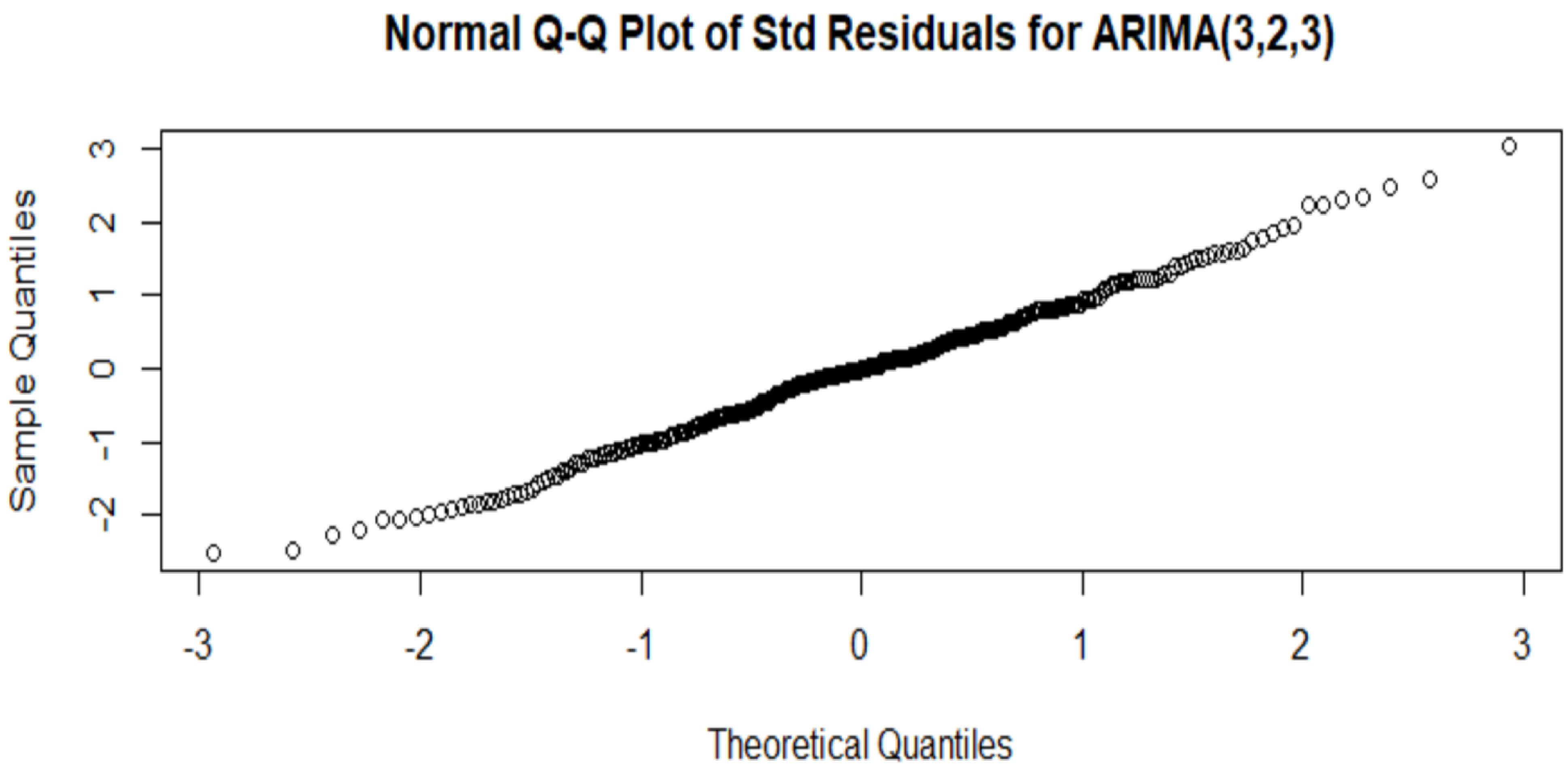
Let $x_{(1)} < x_{(2)} \leq \cdots < x_{(n)}$ the order statistics of the sample $x_1, \ldots, x_n$

$x(\lfloor n \cdot p \rfloor)$ is the sample p^{th} quantile



III. Diagnostics

```
model3<-arima(x ,order=c(3,2,3))
rs <- model3$residuals
stdres <- rs/sqrt(model3$sigma2)  #standardized residuals
qqnorm(stdres, main = "Normal Q-Q Plot of Std Residuals for ARIMA(3,2,3)")
```

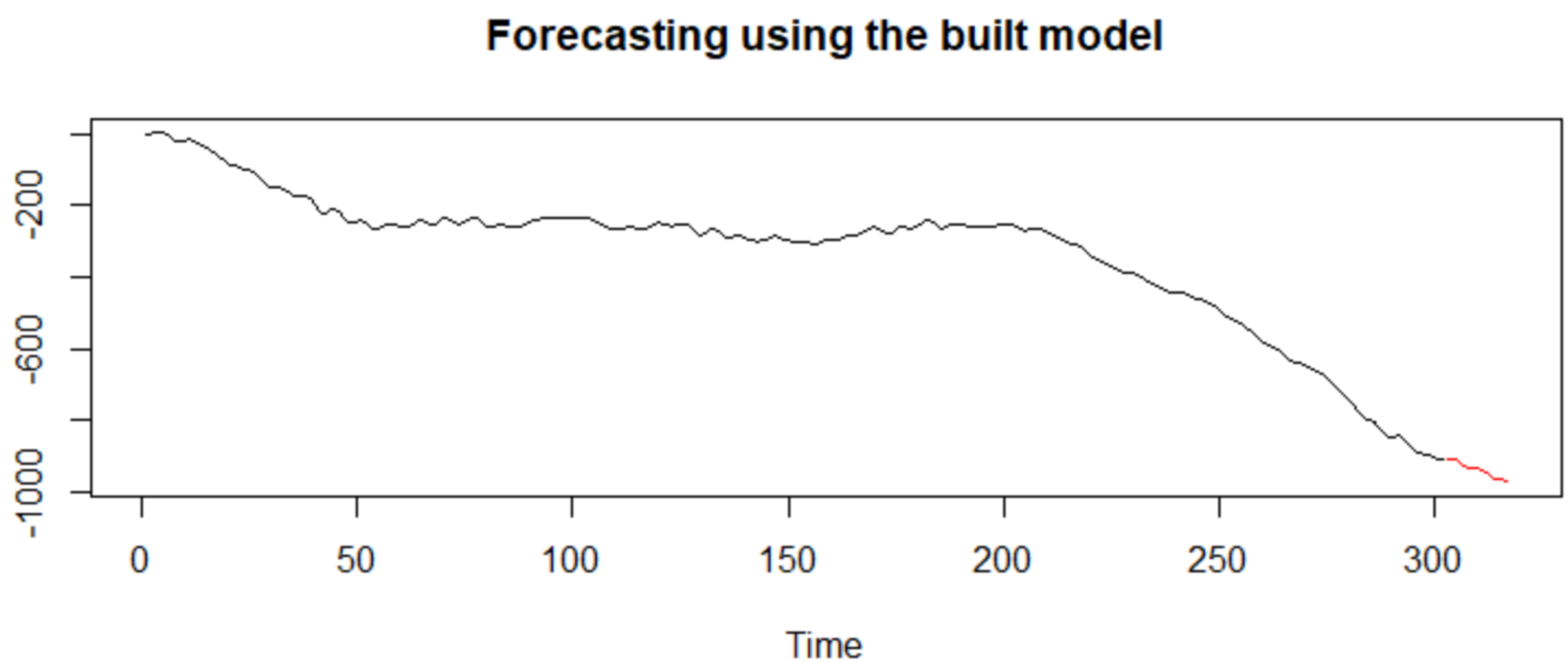


Model building - forecasting

Model building => model3<-arima(x ,order=c(3,2,3))

fore = predict(model3, n.ahead=15)

ts.plot(x, fore$pred, col=1:2)



Seasonal ARIMA models*

Def. X_t is said to be seasonal ARIMA$(p,d,q)\times(P,D,Q)_s$ process with period *s* if the differenced process $Y_t=(1-B)^d(1-B^s)^D X_t$ is a causal ARMA process.

Obs. In applications, D is rarely more than one and P and Q are usually less than 3.

$$\left.\begin{array}{l} X_1, X_{s+1}, X_{2s+1},\dots \\ X_2, X_{s+2}, X_{2s+2},\dots \\ \;\dots \\ X_s, X_{2s}, X_{3s},\dots \end{array}\right\} \text{ are ARMA(P,Q) processes}$$

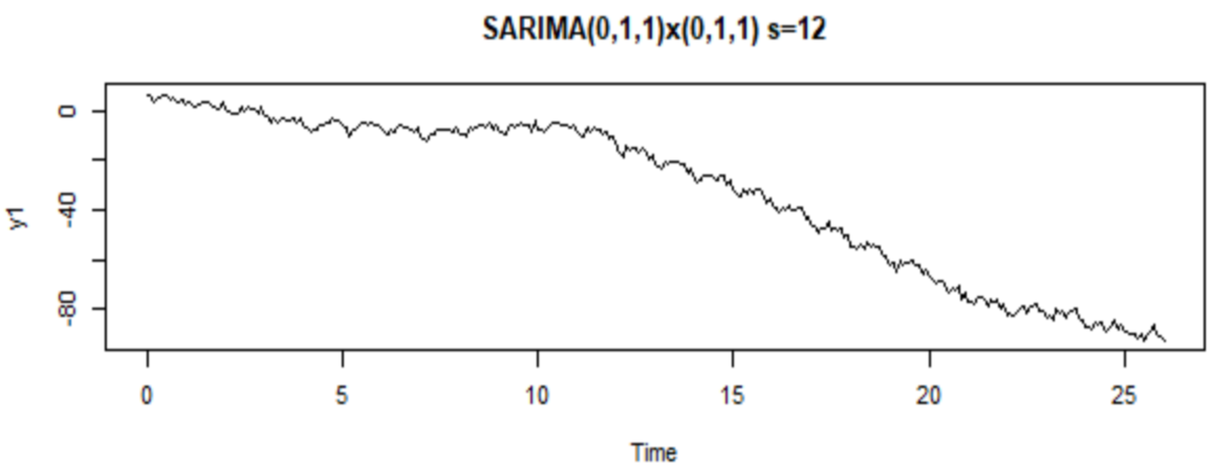
Steps to build a SARIMA model

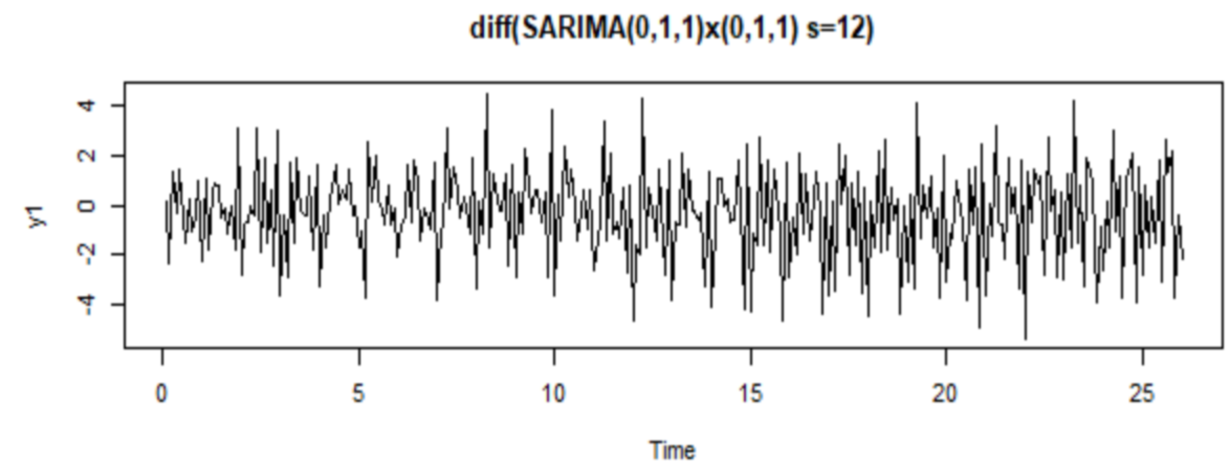
1. Find d, s, (D=1) such that $Y_t=(1-B)^d(1-B^s)^D X_t$ is stationary
2. Examine the ACF and PACF of Y_t
   - Look at the values at $k\cdot s$, k=1,2,3… to choose values for P and Q
   - Look at values at 1,2,…,s-1 to choose values for p and q
3. Use the AIC and the goodness of fit to choose the best SARIMA model

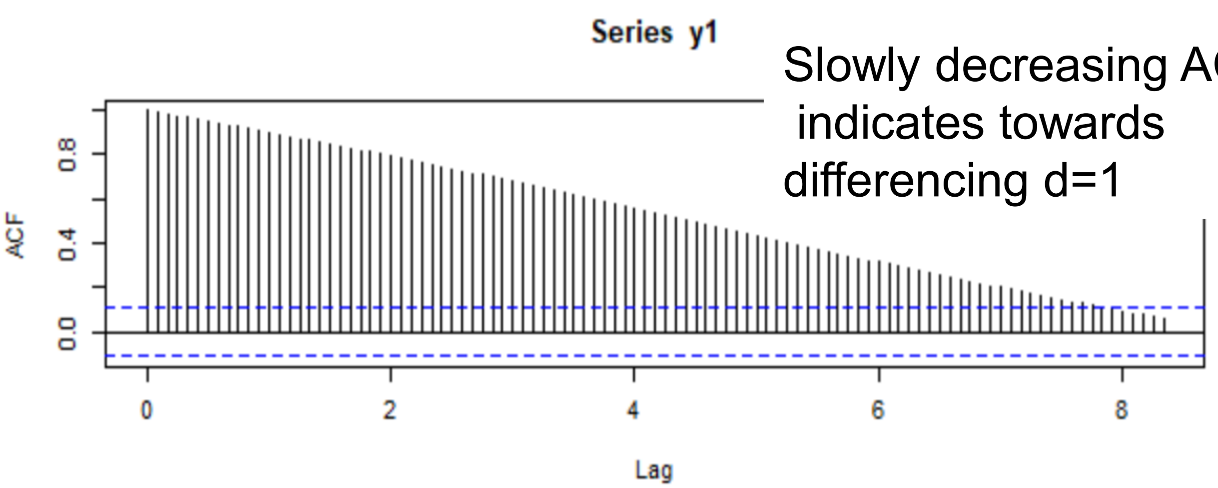
for details see Peter Brockwell, Richard Davis. Time Series: Theory and Methods, Springer-Verlag, 1987 (p. 310 -)

SARIMA models

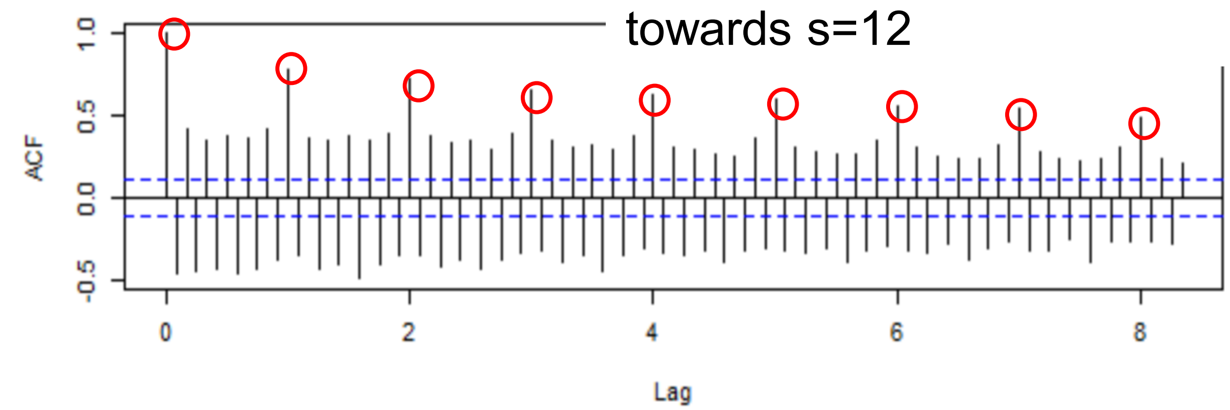
```
x=sarima.sim(d=1, ma=-.4, D=1, sma=-.6, S=12, n=300)          # sarima.sim {astsa}
```







Slowly decreasing ACF indicates towards differencing d=1

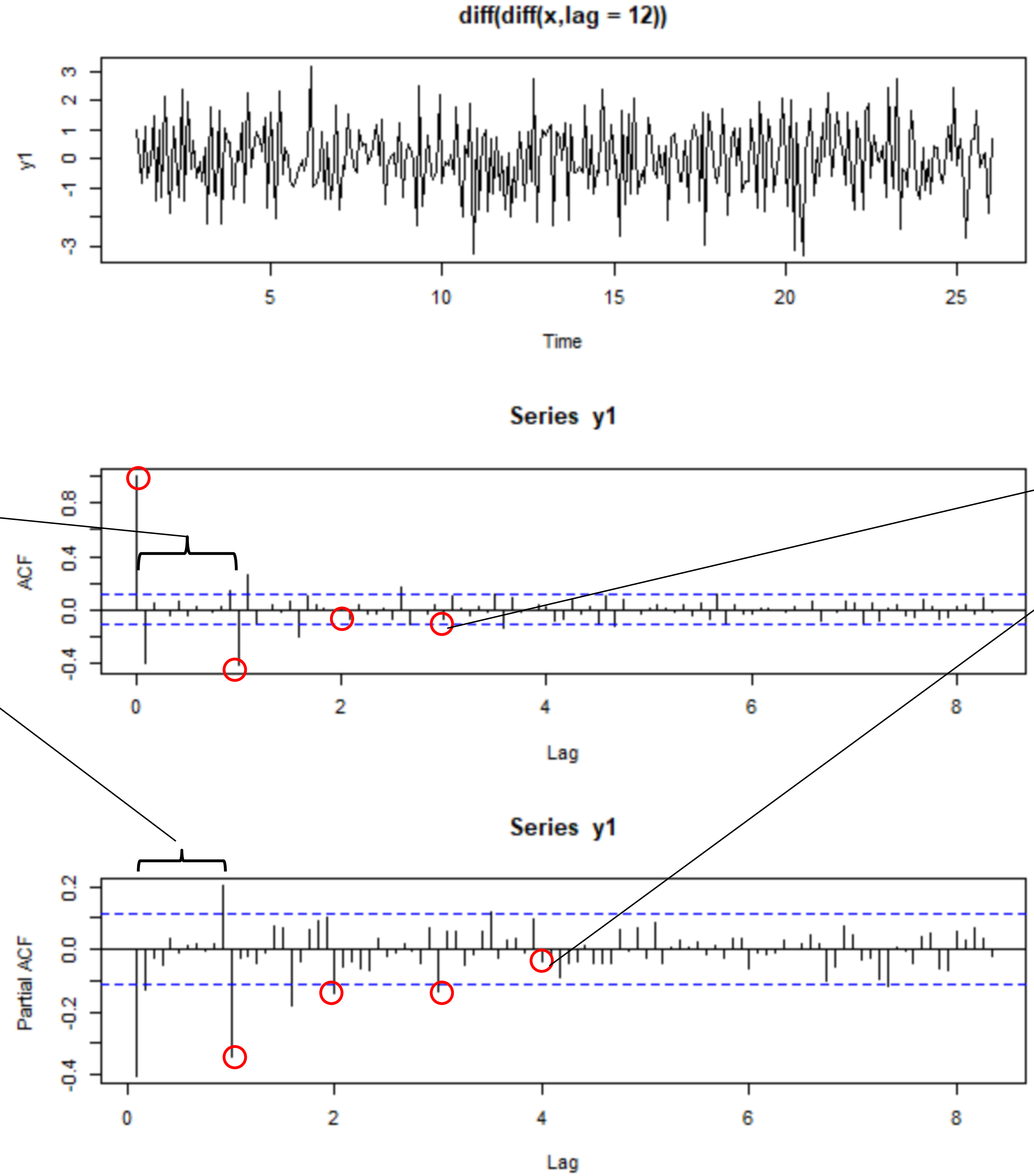


This periodicity indicates towards s=12

SARIMA models

for p,q

p=0,q=1?
p=1,q=0?
p=2,q=0?



for P, Q

P=0,Q=1?

Further use AIC and the diagnostic step to choose the best model to fit the data.

Observations in “transversal” studies – fitting a model

We have $\{X_1, \dots, X_n\}$ random variables, i.i.d like the stochastic model $X : \Omega \to S$.
Statistical data are the observed values $(x_1, \dots, x_n) \in S^n$.
We are interested to fit a parametric model to our sample data.

In parametric statistics, we assume that the model has a known functional form (e.g. Normal, Poisson, Gamma etc.) that depends on an unknown parameter θ. We are looking for estimators of θ, $\hat{\theta}_n(X_1, \dots, X_n)$ (or shorter, just $\hat{\theta}_n$).

Observations in “transversal” studies – fitting a model

Def. $\hat{\theta}_n$ is unbiased for θ iff $E(\hat{\theta}_n) = \theta$.

$\hat{\theta}_n$ is biased (but asymptotically unbiased) for θ iff

$$E(\hat{\theta}_n) \xrightarrow[n \to \infty]{} \theta$$

$$\left[E(\hat{\theta}_n) = \theta + b_n, \text{ where } b_n \xrightarrow[n \to \infty]{} 0; \ b_n \text{ is called bias} \right]$$

Density estimation of the model X

The histogram is actually a density estimator.

Kernel methods are more advanced methods for density estimation.

Let F_n = the sample distribution function

$$F_n(x) = \frac{no(X_i < x)}{n} \qquad \left(\text{recall that } f(x) = F'(x)\right)$$

$$f_n(x) = \frac{F_n\left(x + \frac{b_n}{2}\right) - F_n\left(x - \frac{b_n}{2}\right)}{b_n},$$

where $b_n > 0$ is small enough b_n is called window or bandwidth

Density estimation of the model X

We denote by $I_{[a,b]}(x) = \begin{cases} 1, & x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$

if we take $K = I_{[-\frac{1}{2},\frac{1}{2}]}$ then $f_n(x) = \frac{1}{nb_n} \cdot \sum_{i=1}^{n} K\left(\frac{x - X_i}{b_n}\right)$

Density estimation of the model X

<u>Def</u>. $K : \mathbb{R} \to \mathbb{R}$ is said to be Kernel if:

- $\int_{-\infty}^{\infty} K(x)\,dx = 1$
- it is bounded, positive and symmetrical to zero
  $$K(x) = K(-x)$$
- $\lim_{|x| \to \infty} |x| K(x) = 0$

The estimator f_n associated to a Kernel K is defined as:

$$f_n(x) = \frac{1}{n \cdot b_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b_n}\right) \qquad b_n \text{ - bandwidth}$$

Density estimation of the model X

Examples of Kernels

1. The Parzen – Rosenblatt Kernel $K = I_{[-\frac{1}{2}, \frac{1}{2}]}$

2. The Epanechinikov Kernel $K(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) I_{[-\sqrt{5}, \sqrt{5}]}(x)$

it minimizes the mean integrated square error

$$MISE(f_n(x)) = \int_{-\infty}^{\infty} [f_n(x) - f(x)]^2 \cdot f(x)\, dx$$

3. The Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ – it is the most frequently used.

Parameter estimation - Methods of moments

There are two common methods to estimate the parameters of a density function: Method of Moments (MM) and Maximum Likelihood Estimation (MLE).

X random variable $\sim f(x; \theta_1, \dots, \theta_k)$ the density function

$m_r = E(X^r)$ is the r^{th} moment of X

$\{X_1, X_2, \dots X_n\}$ - random sample of size n

1. we estimate the moments m_r by $\hat{m}_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r$, for $r = \overline{1,k}$

2. we obtain the estimations $\hat{\theta}_1, \dots \hat{\theta}_k$ by solving the equations:

$$\hat{m}_r = m_r(\theta_1, \dots, \theta_k), \quad r = \overline{1,k}$$

Methods of moments - examples

1. $X \sim N(m, \sigma^2)$

We know that $E(X) = m$

$\mathrm{Var}(X) = \sigma^2$

$E(X) = m_1$

$E(X^2) = m_2$

$\mathrm{Var}(X) = E(X^2) - E(X)^2 = m_2 - m_1^2$

$$\left.\begin{array}{l} E(X) = m \\ \mathrm{Var}(X) = \sigma^2 \\ E(X) = m_1 \\ E(X^2) = m_2 \\ \mathrm{Var}(X) = E(X^2) - E(X)^2 = m_2 - m_1^2 \end{array}\right\} \Rightarrow \begin{array}{l} m_1(m, \sigma^2) = m \\ m_2(m, \sigma^2) = \sigma^2 + m^2 \end{array}$$

$$\Rightarrow \hat{m} = \hat{m}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = \hat{m}_2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Methods of moments - examples

2. $X \sim \text{Gamma}(\alpha, \theta)$

$$E(X) = \alpha \cdot \theta$$

$$Var(X) = \alpha \cdot \theta^2$$

$$m_1(\alpha, \theta) = \alpha \cdot \theta$$

$$m_2(\alpha, \theta) = Var(X) + m_1^2(\alpha, \theta) = \alpha \cdot \theta^2 + \alpha^2 \theta^2$$

$$\hat{m}_1 = \hat{\alpha} \cdot \hat{\theta}$$

$$\hat{m}_2 = \hat{\alpha} \cdot \hat{\theta}^2 + \hat{\alpha}^2 \hat{\theta}^2$$

$$\rightarrow \hat{\alpha} = \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2}$$

$$\hat{\theta} = \frac{\hat{m}_2 - \hat{m}_1^2}{\hat{m}_1}$$