Lecture 4

3.7 Finding the best model

One of the most important task in time series analysis is the model selection. In many cases several models may seem to fit well.

One can select the model according some numerical criteria, i.e. using AC, PACF or AIC criterions.

In time series: we fit the models selected on the properties of ACF and PACF. Such selection is somewhat subjective.

George Box: "all models are wrong but some are useful".

Advise: exercise also your judgement.

Why? because in time series models may be "useful" and "adequate", but there is anything like a "correct" model for a given time series.

Models are just approximations to the dynamic systems generating the data.

General observation: if we fit an AR model with a large number of parameter, we will get a better fit, using more parameter. Drawback: Overfitting

- information will be dissipated over large number of parameters. So, they will be poorly estimated.
- Overfitting, will produce poor forecast
- Overfitted models fit "locally" nearly perfectly , but "globally" perform poor.

Example: we fit 10-th order polynomial to US population data from 1900 to 2000, and extrapolate to 2010. We get good fit for the data 1900-2000, but poor unrealistic prediction for 2010.

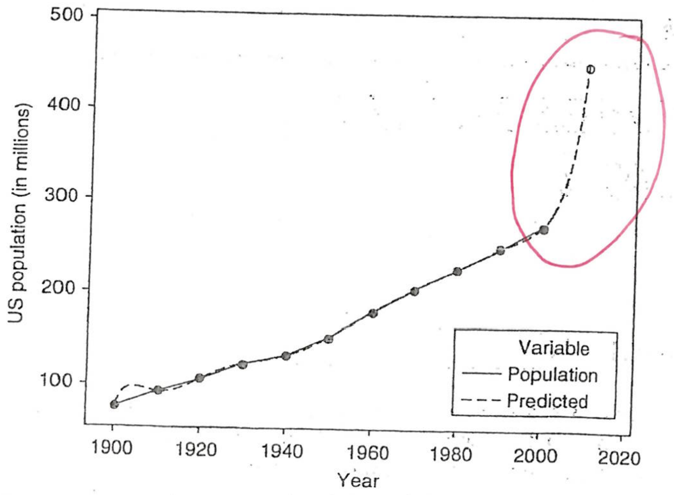


Figure 6.1 The US population according to the census versus time from 1900 to 2000. A perfect fit to the past data, but a poor and unrealistic forecast for the year 2010.

Conclusion: we prefer models with good explanatory power - they are called parsimonious model. They represent data with then minimum number of parameter.

3.8 Case analysis: Internet user data, model selection

What is model selection? By model selection fitting ARMA(p,q,) model, we understand selection of p and q.

Popular criterions for selection for p, q are:

- Akaike's information criterion (AIC)
- Baysian information criterion BIC.
- These criterions include a penalty for over parameterizing. They try to select a model with a minimal number of parameters.

Model selection criterions

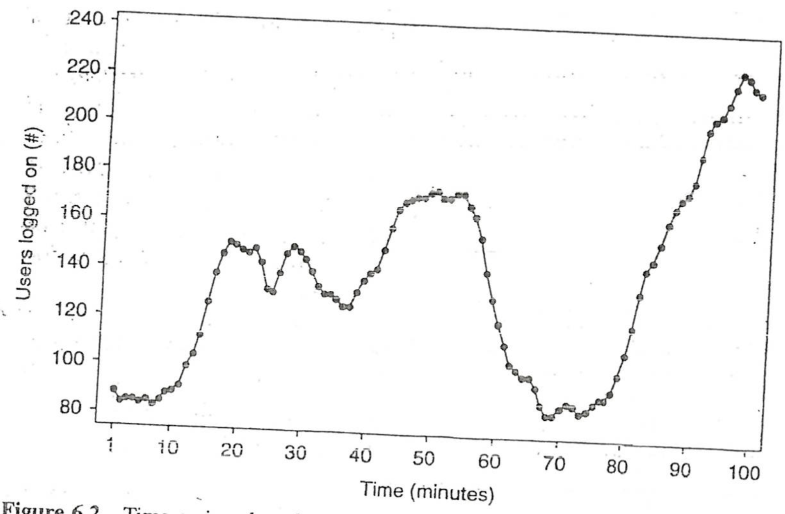
We dicuss below how they are used.

Example: Consider the time series of the number X_t of internet users logged on each minute for a period 100 min.

Figure 6.2. shows the plot of the data

Figure 6.3. shows the sample ACF, which is remains large for a large number of lags.

We conclude: The series appears to be non-stationary. (If a process is stationary we expect it to have a long term constant mean)



No constant mean !

Figure 6.2 Time series plot of the number, z_t, of internet server users over a 100-minute period.

ACF

Note: for non-stationary time series ACF does not decay to 0 fast

ACF of stationary time series

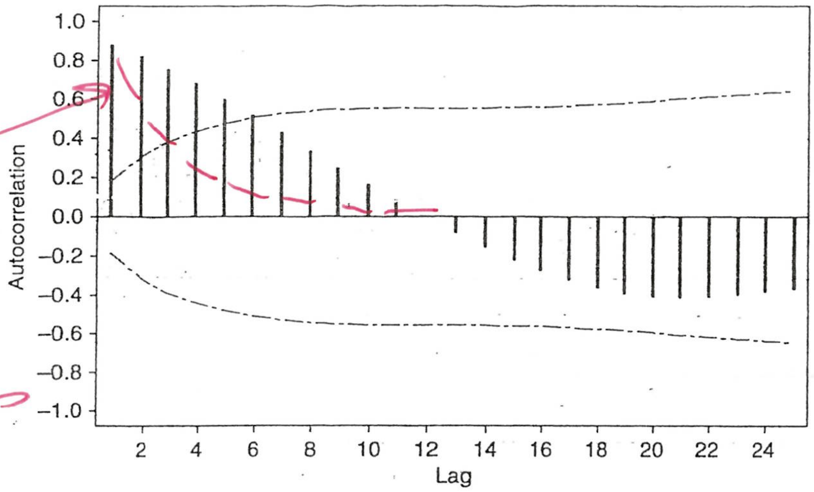


Figure 6.3 The sample autocorrelation for the number, z_t, of internet server users over a 100-minute period.

Differencing

How to deal with non-stationarity?: Although the number of users X_t is non-stationary process, we can expect the first difference $\boxed{w_t = X_t - X_{t-1}}$ in the numbers of users logged in from minute to minute to be stationary

Such hypothesis is confirmed by the data:

- Figure 6.4 shows the plot of w_t. We see that the process of changes looks rather stationary.

Figure 6.5a shows sample ACF, and 6.5b sample PACF of the data w_t.

- ACF shows a damped sin wave. PACF cuts off after lag 3. It suggest we should fit $AR(3)$ model for w_t:

$$w_t = \phi_1 w_{t-1} + \cdots + \phi_3 w_{t-3} + \varepsilon_t.$$

AR(3)

$W_t = X_t - X_{t-1}$

This means that the original series we model by ARIMA(3,1,0) model

$$X_t - X_{t-1} = \phi_1(X_{t-1} - X_{t-2}) + \cdots + \phi_3(X_{t-3} - X_{t-4}) + \varepsilon_t.$$

- Both ACF and PACF are like damped sin waves, which is a pattern of an ARMA(1,1) model. This suggests we could try also ARMA(1,1) model:

$$w_t = \phi_1 w_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}.$$

Plot of $w_t = X_t - X_{t-1}$

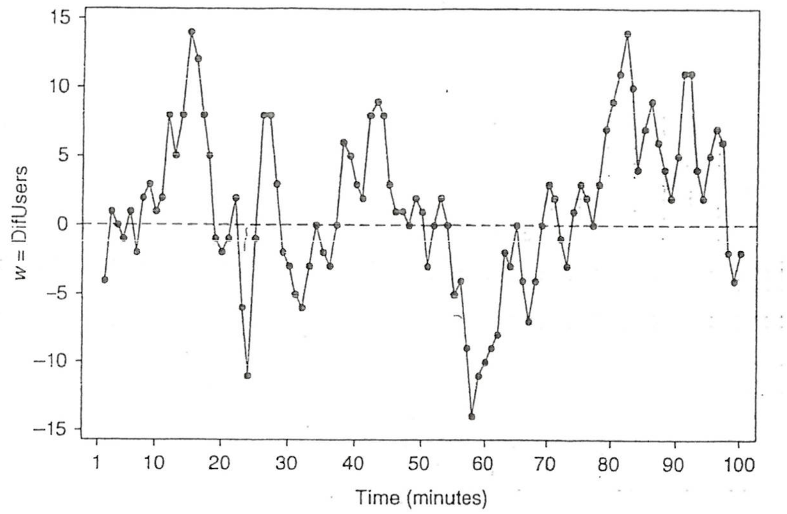
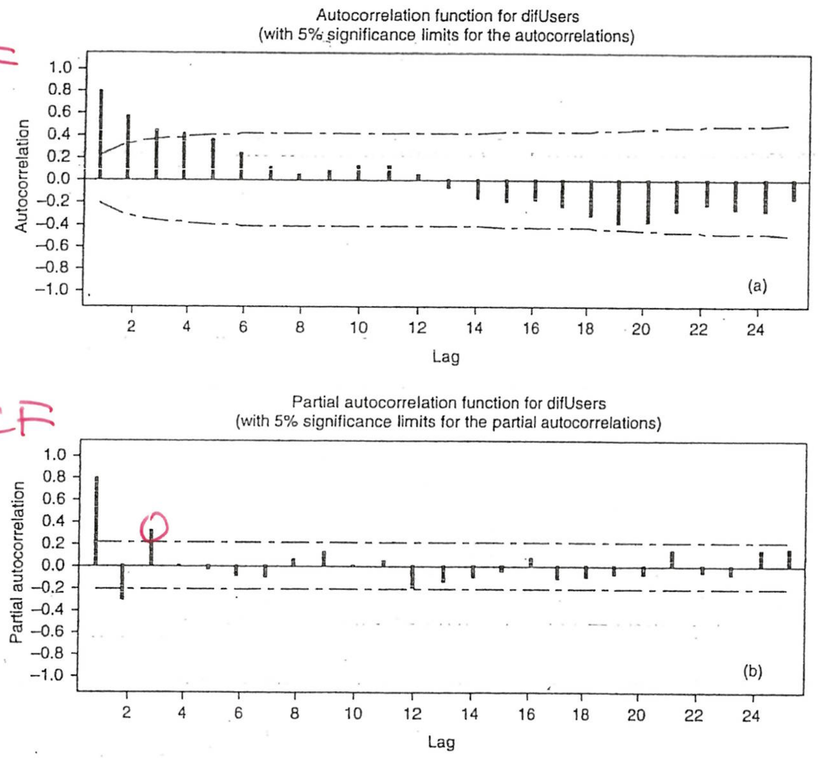


Figure 6.4 Time series plot of the difference $w_t = \nabla z_t$ (changes) of the number of internet server users over a 100-minute period.

Poll 1
Fit MA(q)

ACF

PACF


Autocorrelation function for difUsers
(with 5% significance limits for the autocorrelations)

Partial autocorrelation function for difUsers
(with 5% significance limits for the partial autocorrelations)

Figure 6.5 (a) The sample ACF and (b) the sample PACF of the differences (changes) of the number of internet server users over a 100-minute period.

PACF sugests fitting AR(3) model

? ?

Question: which model to use for w_t: AR(3) or ARMA(1,1)? To check which model suits better, we shall compare the residuals, obtained fitting these two models.

$H_0: \phi_1 = 0$

$H_1: \phi_1 \neq 0$

Poll 2

Poll 3

TABLE 6.1 Estimation Summary for Fitting the ARIMA(3, 1, 0) Model to the Internet Server Data z_t

Model term	Estimate	Standard error	t	p
AR 1 : ϕ_1	1.1632	0.0955	12.18	0.000
AR 2 : ϕ_2	−0.6751	0.1360	−4.96	0.000
AR 3 : ϕ_3	0.3512	0.0956	3.67	0.000

Differencing: 1 regular.

Number of observations: Original series 100; after differencing 99.

Residuals: SS = 917.812; MS = 9.561; df = 96.

Modified Box–Pierce (Ljung–Box) Chi–square statistic:

Lag	12	24	36	48
Chi square	7.5	20.2	31.5	46.5
df	9	21	33	45
p-value	0.587	0.509	0.539	0.411

Residuals: $\hat{\varepsilon}_t = w_t - \hat{\phi}_1 w_{t-1} - \hat{\phi}_2 w_{t-2} - \hat{\phi}_3 w_{t-3}$

Note: AR(3) model fits data w_t if serials $\hat{\varepsilon}_t$ are uncorrelated

Poll 4

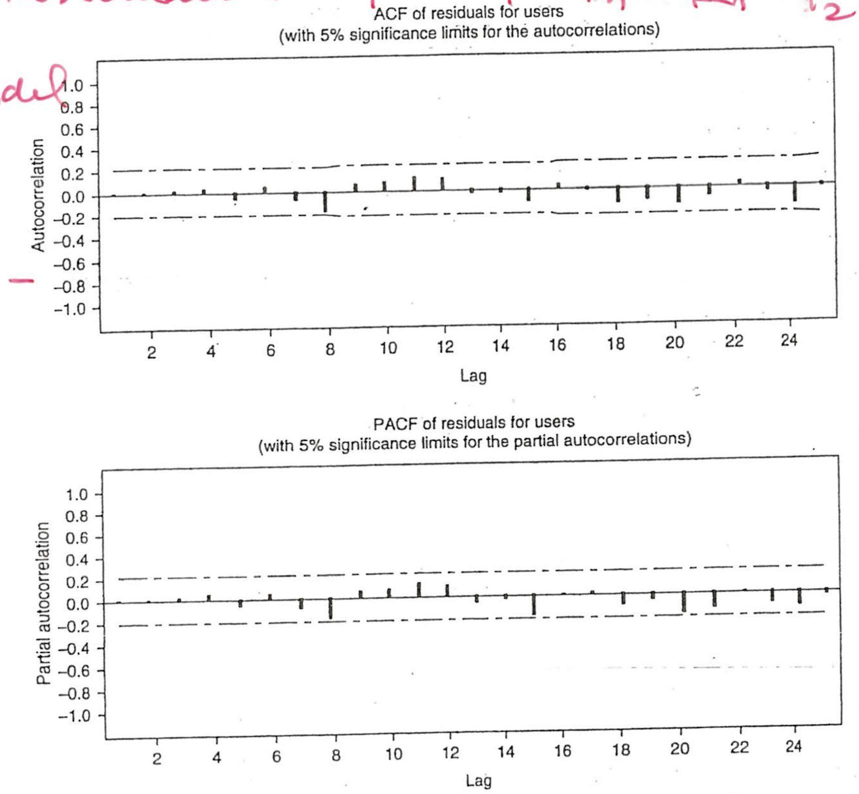
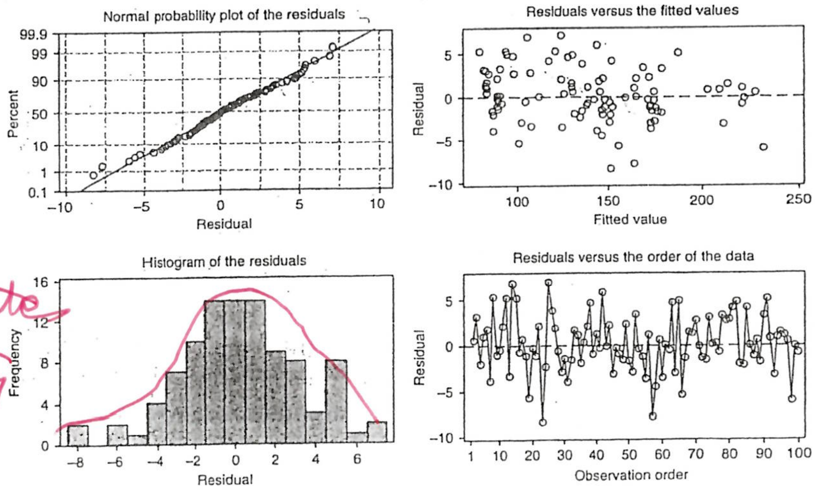


Figure 6.6 Sample ACF and sample PACF for the residuals from the ARIMA(3, 1, 0) model.



approximate probability density f(x)

Figure 6.7 Summary residual check from the ARIMA(3, 1, 0) model.

How to test if the estimate $\hat{\phi}$ is significantly different from 0?

$H_0: \phi = 0$ $\quad$ $H_1: \phi \neq 0$.

Using p-value. Select 5% significance level

If $p > 0.05$, ϕ is not significant (do not reject H_0)

If $p \leq 0.05$, ϕ is significant (reject H_0)

Using SE

If $|\hat{\phi}| > 2SE$ → ϕ significant (reject H_0)

If $|\hat{\phi}| \leq 2SE$ → ϕ not significant (do not reject H_0)

Example 1) $\hat{\phi} = 0.1$ $\quad$ $p = 0.01$

$p < 0.05$ → ϕ is significant

2) $\hat{\phi} = -0.2$ $\quad$ $SE = 0.15$

here $|\hat{\phi}| = 0.2 < 2SE = 0.3$

so ϕ is not significantly different from 0.

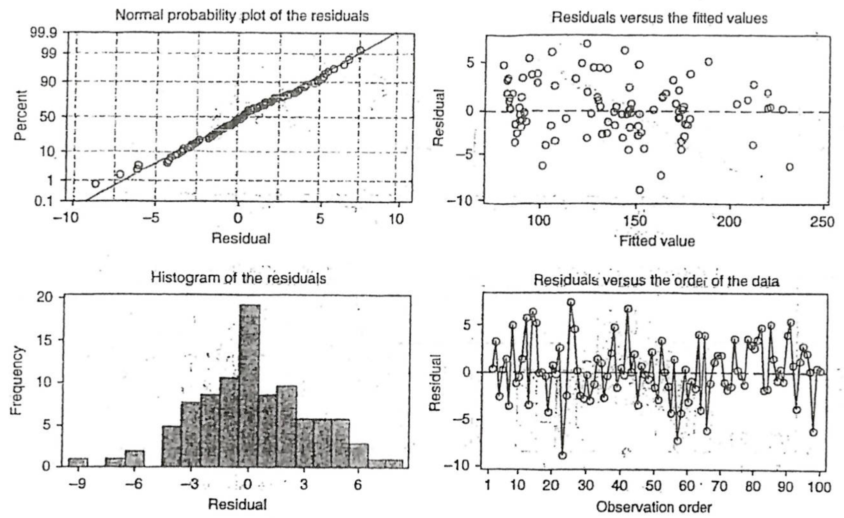


Figure 6.9 Summary residual check from the ARIMA(1, 1, 1) model.

$W_t = \phi_1 W_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}$

TABLE 6.2 Estimation Summary for Fitting the ARIMA(1, 1, 1) Model to the Internet Server Data z_t

Model term	Estimate	Standard error	t	p
AR 1 : ϕ_1	0.6573	0.0868	7.57	0.000
MA 1 : θ_1	−0.5301	0.0974	−5.44	0.000

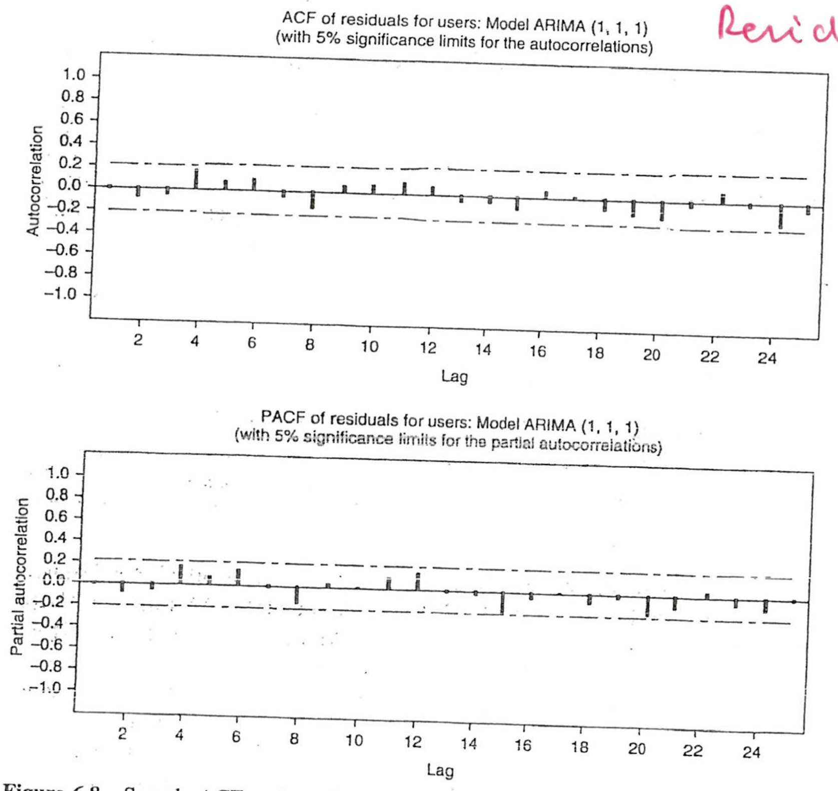
Differencing: 1 regular.

Number of observations: Original series 100; after differencing 99.

Residuals: SS = 961.617; MS = 9.914; df = 97.

Modified Box–Pierce (Ljung–Box) Chi-square statistic:

Lag	12	24	36	48
Chi square	10.3	26.3	38.1	54.2
df	10	22	34	46
p-value	0.417	0.237	0.289	0.190



ACF

Residuals of ARMA(1,1) model

PACF

Figure 6.8 Sample ACF and sample PACF for residuals from the ARIMA(1, 1, 1) model.

Poll 5 AR(3) or ARMA(1,1)?

Var$(\hat{\varepsilon}_t)$ of AR(3)

Comparison:

- both models AR(3) and AR(1,1) fit well, but AR(3) model has smaller residuals variance MS=9.56 than that of ARMA(1,1), MS=9.914. ~ Var$(\hat{\varepsilon}_t)$ of ARMA(1,1)
- ARMA(1,1) has two parameters. They are slightly better estimated than three parameters of AR(3).

Which model to choose? We may go for a model with smaller number of parameter, i.e. ARMA(1,1).

Next we provide formal criterion to answer this question.

3.9 Model selection criteria

A common tool to model selection is a model fitting criterion.

Some software packages provide automatic model selection for stationary ARMA(p,q) models based AIC and BIC criterions.

Note: To apply it for non-stationary data, we may need first to difference the data.

AIC criterion: it is constructed as follows. For different order values (p, q), it fits ARMA(p,q) model. The it computes the sample variance of residuals $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$,

$$\hat{\sigma}_\varepsilon^3 = n^{-1} \sum_{j=1}^{n} \hat{\varepsilon}_j^2$$

and the function

$$AIC = \ln(\hat{\sigma}_\varepsilon^2) + \boxed{\frac{2r}{n}}$$

penalty term

where r denotes the number of estimated parameters including the constant term.

We prefer model with the smallest AIC.

Select p,q for ARMA(p,q) which minimize AIC.

Rule: to fit ARMA(p,q) model, we select p, q which minimizes AIC.

How AIC works? When we add additional parameters we typically reduce the variance $\hat{\sigma}_\varepsilon^2$. Then $\ln(\hat{\sigma}_\varepsilon^2)$ becomes smaller. Since AIC includes penalty $\frac{2r}{n}$ for adding additional parameter, penalty makes AIC larger, and prevents overfitting.

Note: for AR(p) models, AIC tend to overestimate p. Therefore, often instead of AIC, BIC criterion is used:

$$\boxed{BIC = \ln(\hat{\sigma}_\varepsilon^2) + \frac{r \ln n}{n}.}$$

penalty

It has stronger penalty $\frac{r \ln n}{n}$ for the number of parameters. The prefered model is with the smallest BIC.

Table 6.3. contains the values of AIC and BIC criterions fitted to the internet data.

- We see that AIC would select AR(3) model
- BIC selects ARMA (1,1) model, but difference from AR(3) is very small.

Conclusion:

- We may prefer ARMA(1,1) model since it has less parameters
- often several model fits to the data equally well
- ! unnecessary increase of parameters would increase error in forecasting.

do not overfit!

Poll 6 Use AIC
to select ARMA(p,q) model

$X_t \sim ARIMA(p,1,q)$

means

$w_t = X_t - X_{t-1} \sim ARMA(p,q)$

TABLE 6.3 The AIC, AICC, and BIC Values for ARIMA(p, 1, q), with $p = 0, \ldots, 5$; $q = 0, \ldots, 5$ Models Fitted to the Internet Server Data. The Numbers in Bold Face Indicate the Minimum for Each of the Information Criteria. Note That All of the Above Models are Fitted Without a Constant Term

Model	AIC	AICC	BIC
ARIMA(0, 1, 0)	628.995	628.995	628.995
ARIMA(1, 1, 0)	527.238	527.279	529.833
ARIMA(2, 1, 0)	520.178	520.303	525.368
ARIMA(3, 1, 0)	**509.994**	**510.247**	517.779
ARIMA(4, 1, 0)	511.930	512.355	522.310
ARIMA(5, 1, 0)	513.862	514.507	526.837
ARIMA(0, 1, 1)	547.805	547.847	550.401
ARIMA(1, 1, 1)	512.299	512.424	**517.490**
ARIMA(2, 1, 1)	514.291	514.544	522.077
ARIMA(3, 1, 1)	511.938	512.363	522.318
ARIMA(4, 1, 1)	510.874	511.520	523.850
ARIMA(5, 1, 1)	515.638	516.551	531.209
ARIMA(0, 1, 2)	517.875	518.000	523.065
ARIMA(1, 1, 2)	514.252	514.504	522.037
ARIMA(2, 1, 2)	515.360	515.786	525.741
ARIMA(3, 1, 2)	513.917	514.563	526.893
ARIMA(4, 1, 2)	514.179	515.092	529.750
ARIMA(5, 1, 2)	511.543	512.774	529.709
ARIMA(0, 1, 3)	518.272	518.524	526.057
ARIMA(1, 1, 3)	512.576	513.002	522.957
ARIMA(2, 1, 3)	513.773	514.418	526.749
ARIMA(3, 1, 3)	512.414	513.327	527.985
ARIMA(4, 1, 3)	517.078	518.308	535.243
ARIMA(5, 1, 3)	513.434	515.034	534.195
ARIMA(0, 1, 4)	517.380	517.805	527.760
ARIMA(1, 1, 4)	513.100	513.745	526.076
ARIMA(2, 1, 4)	511.241	512.154	526.812
ARIMA(3, 1, 4)	512.758	513.989	530.924
ARIMA(4, 1, 4)	512.808	514.408	533.569
ARIMA(5, 1, 4)	553.110	555.133	576.466
ARIMA(0, 1, 5)	516.857	517.502	529.833
ARIMA(1, 1, 5)	514.276	515.189	529.847
ARIMA(2, 1, 5)	716.050	717.281	734.216
ARIMA(3, 1, 5)	516.504	518.104	537.265
ARIMA(4, 1, 5)	515.845	517.867	539.201
ARIMA(5, 1, 5)	512.113	514.613	538.064

Example. Figure 2.8 shows

– the plot of monthly simple returns of the CRSP equal weighter index from 1926 to 2003,

– the sample ACF of this series

The dashed lines denote the two standard -error limits for ACF.

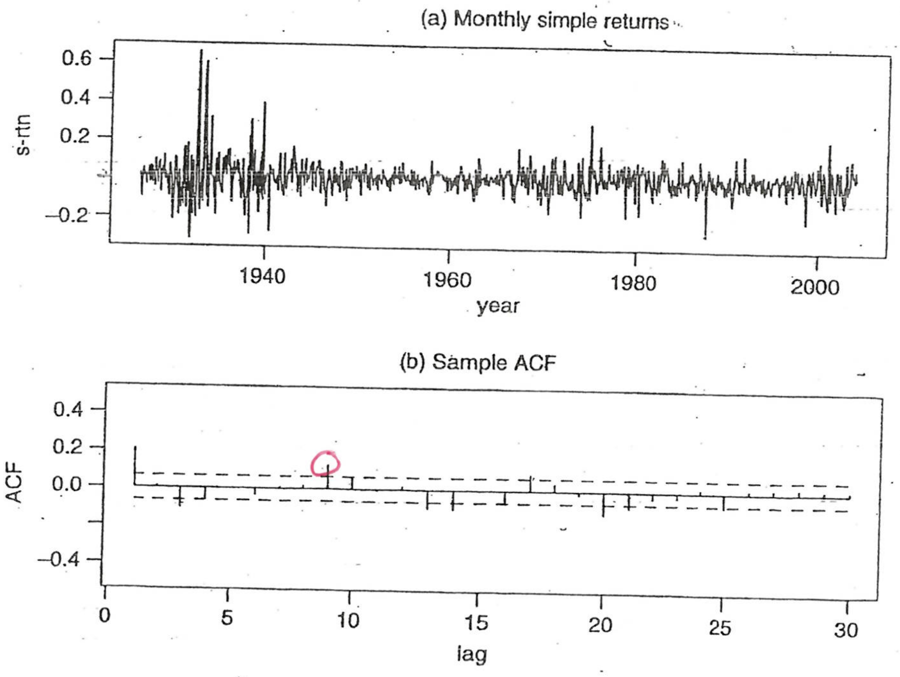
We see that ACF is significant at lags 1, 3 and 9 (and we observe some marginal significance at higher lags).

Based on sample ACF, the moving average model is

$$r_t = c_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_3 \varepsilon_{t-3} - \theta_9 \varepsilon_{t-9}.$$

This method provides information on the nonzero MA lags of the model. In this case nonzero coefficients are θ_1, θ_3 and θ_9.

Note: the PACF function we used to determine the order of AR(p) model, does not provide such information.



Plot simple returns

ACF

MA(9)

Figure 2.8: Time plot and sample autocorrelation function of the monthly simple returns of the CRSP equal-weighted index from January 1926 to December 2003.

Estimation

Example. Consider the monthly simple returns of the CRSP equal weighter index and the specified MA(9) model.

The conditional maximum likelihood method produces the fitted model

$$r_t = 0.013 + \varepsilon_t + 0.181\varepsilon_{t-1} - 0.121\varepsilon_{t-3} + 0.122\varepsilon_{t-9}, \quad \hat{\sigma}_\varepsilon = 0.0724,$$

where standard errors for the estimates are 0.003, 0.032, 0.032 and 0.032, respectively. All parameters are significant.

The exact maximum likelihood method produces the fitted model

$$r_t = 0.013 + \varepsilon_t + 0.183\varepsilon_{t-1} - 0.120\varepsilon_{t-3} + 0.123\varepsilon_{t-9}, \quad \hat{\sigma}_\varepsilon = 0.0724,$$

where standard errors for the estimates are 0.003, 0.032, 0.032 and 0.032, respectively. All parameters are significant.

Testing for goodness of fit shows that both models are adequate.

Comparing models, we observe that difference between two estimation methods is negligible.

How to test, that estimated parameter is significant (i.e. different from zero.

We have $\hat{\theta}_1 = 0.181$, SE $= 0.032$

We test

$H_0 : \theta_1 = 0$

against

$H_1 : \theta_1 \neq 0$.

Rule: at 5% significance level,

If $|\hat{\theta}_1| > 2SE \rightarrow$ reject H_0

If $|\hat{\theta}_1| \leq 2SE \rightarrow$ do not reject H_0.

We have

$|\hat{\theta}_1| = 0.181 > 2SE = 2(0.032) = 0.064$

Reject H_0: θ_1 is significantly different from 0.

3.10 Mean, variance and auto-covariance of AR(1) time series

Exercise. Consider a stationary AR(1) process

$$\boxed{X_t = \phi X_{t-1} + \varepsilon_t,}$$

where the process ε_t is white noise process with zero mean and variance $E\varepsilon_t^2 = \sigma_\varepsilon^2$, and $|\phi| < 1$

Prove the following

(i) $EX_t = 0$.
(ii) $Var(X_t) = \frac{\sigma_\varepsilon^2}{1-\phi^2}$.
(iii) Show that autocovariance function is

do not depend on t

$$\gamma_k = \frac{\sigma_\varepsilon^2}{1-\phi^2}\phi^k, \qquad k = 0, 1, 2, \cdots.$$

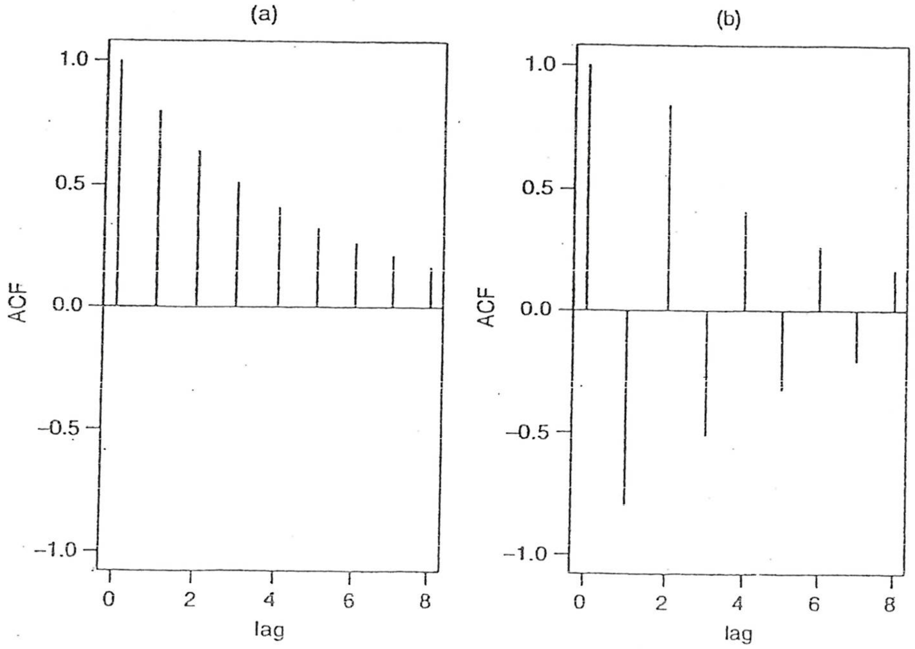
γ_k depends only on the lag k

Show that autocorrelation function

$$\rho_k = \phi^k, \qquad k = 0, 1, 2, \cdots.$$

Examples of ACF and sample ACF of AR(1) process:

They show that ACF ρ_k tends to zero fast, and sample ACF $\hat{\rho}_k$ is close to the true ρ_k, when lag k are small.



$X_t = \phi X_{t-1} + \varepsilon_t$

Figure 2.3. The autocorrelation function of an AR(1) model: (a) for $\phi_1 = 0.8$ and (b) for $\phi_1 = -0.8$.

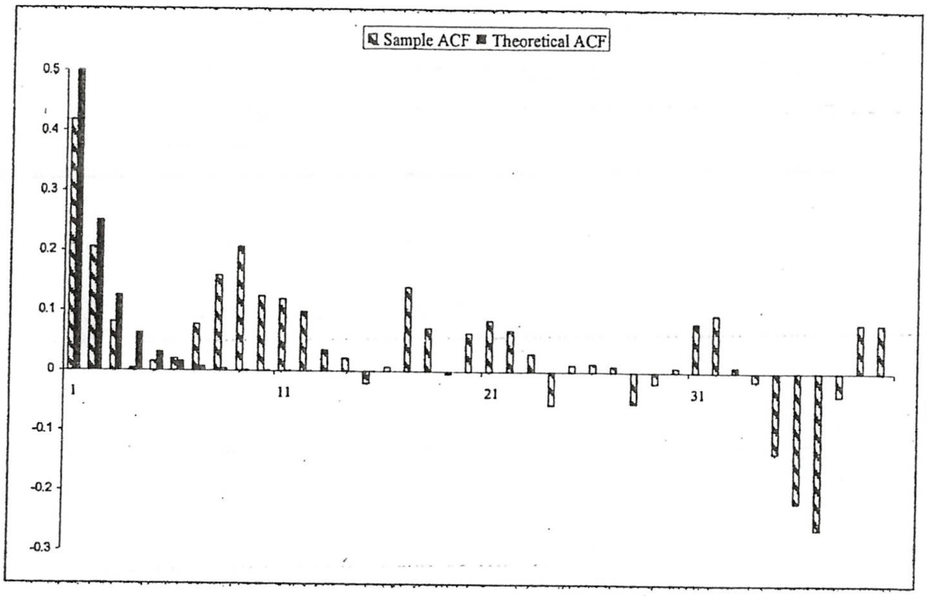
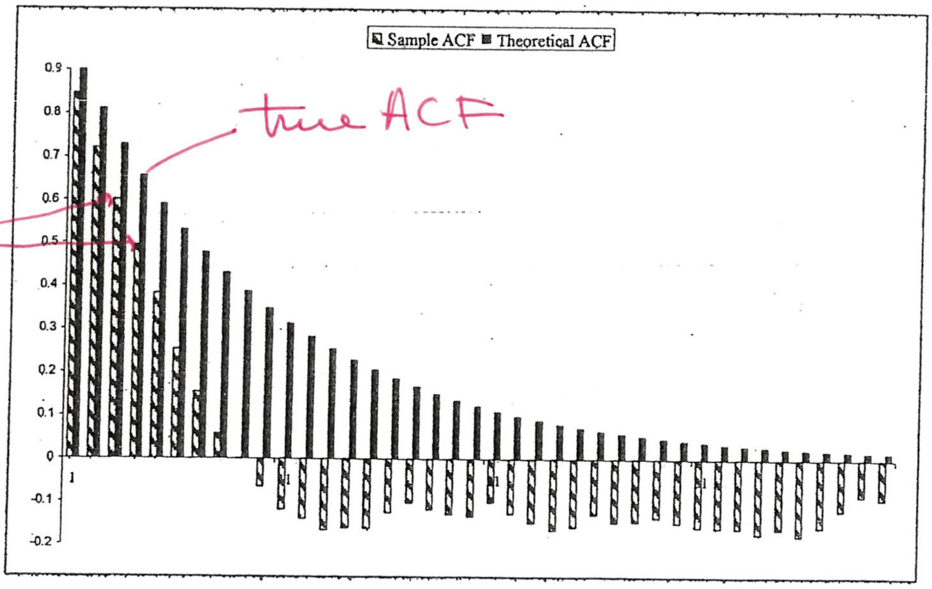


Figure 2.5: Sample and theoretical autocorrelation function (ACF) for the simulated data in Figure 2.1.

$X_t = \phi X_{t-1} + \varepsilon_t, \quad \phi = 0.5$



true ACF

~~Estimated~~

Sample ACF

Figure 2.6: Sample and theoretical autocorrelation function (ACF) for the simulated data in Figure 2.2.

$X_t = \phi X_{t-1} + \varepsilon_t, \quad \phi = 0.9$

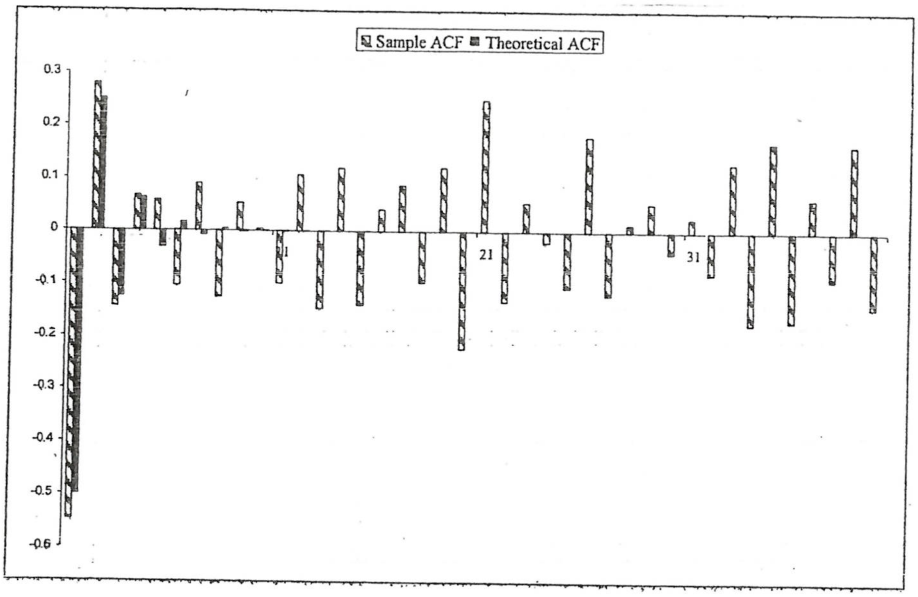


Figure 2.7: Sample and theoretical autocorrelation function (ACF) for the simulated data in Figure 2.3.

$X_t = \phi X_{t-1} + \varepsilon_t$, $\phi = -0.7$

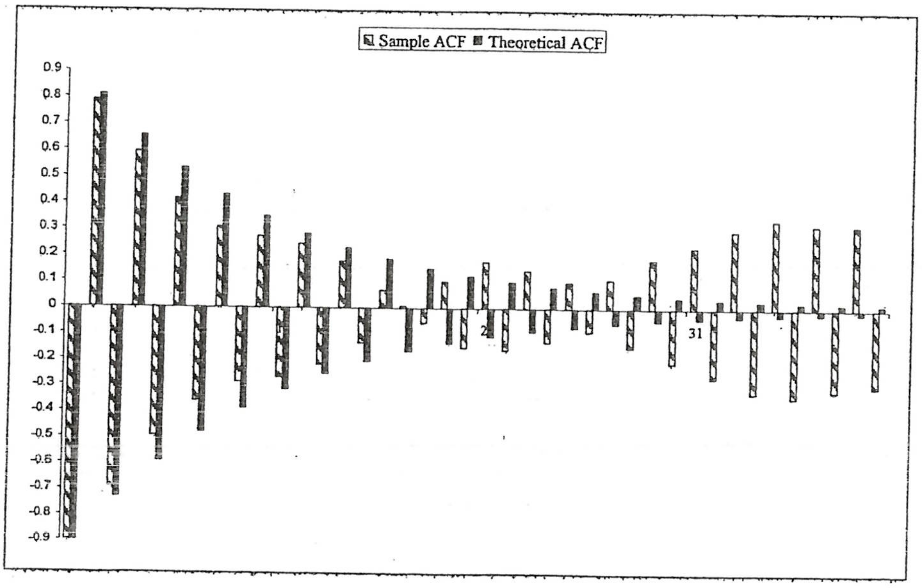


Figure 2.8: Sample and theoretical autocorrelation function (ACF) for the simulated data in Figure 2.4.

$\phi = -0.9$

Solution:

(i) We take expectation of both side of AR(1) equation:

$$\begin{aligned} E[X_t] &= E[\phi X_{t-1} + \varepsilon_t] \\ &= E[\phi X_{t-1}] + E[\varepsilon_t] \\ &= \phi E[X_{t-1}] \end{aligned}$$

$E\varepsilon_t = 0$

since $E[\varepsilon_t] = 0$. Since for $|\phi| < 1$, X_t is a stationary process, then $E[X_t] = E[X_{t-1}] = \mu$ does not depend on time t. Therefore

$$\mu = \phi\mu, \quad or \quad \mu = \frac{0}{1-\phi} = 0.$$

$\rightarrow EX_t = 0$

(ii) We showed that $EX_t = 0$. So, by definition

$$\begin{aligned} \underset{\sigma_Y^2}{\underset{\shortparallel}{Var(X_t)}} &= E(X_t - E[X_t])^2 = EX_t^2 = E(\phi X_{t-1} + \varepsilon_t)^2 \\ &= E(\phi^2 X_{t-1}^2 + 2\phi X_{t-1}\varepsilon_t + \varepsilon_t^2) \\ &= \phi^2 E\underset{\sigma_Y^2}{X_{t-1}^2} + 2\phi E[X_{t-1}\varepsilon_t] + E[\varepsilon_t^2]. \; = \sigma_\varepsilon^2 \end{aligned}$$

Since time series X_t is stationary, its variance remains constant: $Var(Y_t) = EY_t^2 = EY_{t-1}^2 = \sigma_Y^2$. Moreover, future is not correlated with the past, so $E[X_{t-1}\varepsilon_t] = 0$. Thus we obtain

$$\boxed{\sigma_Y^2 = \phi^2\sigma_Y^2 + \sigma_\varepsilon^2,} \quad or \quad \sigma_Y^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}.$$

(iii). Since $EX_t = 0$, then for $k \geq 1$,

$$\begin{aligned} \gamma_k &= Cov(X_t, X_{t-k}) = E[(X_t - EX_t)(X_{t-k} - EX_{t-k})] \\ &= E[X_t X_{t-k}]. \end{aligned}$$

X_t
$\shortparallel$

Since $X_t = \phi X_{t-1} + \varepsilon_t$, then

$$\begin{aligned} \gamma_k &= E[X_t X_{t-k}] = E[(\phi X_{t-1} + \varepsilon_t) X_{t-k}] \\ &= \phi E[X_{t-1} X_{t-k}] + E[\varepsilon_t X_{t-k}] \\ &= \phi E[X_{t-1} X_{t-k}] \end{aligned}$$

because white noise ε_t is uncorrelated with the past and therefore $E[\varepsilon_t Y_{t-k}] = 0$. Because of stationarity,

$$\gamma_k = Cov(X_t, X_{t-k}) = E[X_t X_{t-k}], \quad \gamma_{k-1} = E[X_{t-1} X_{t-k}]$$

and we obtain

$$\boxed{\gamma_k = \phi\gamma_{k-1}, \quad \text{for all} \quad k \geq 0.}$$

$$\gamma_k = \phi\gamma_{k-1}$$
$$= \phi(\phi\gamma_{k-2}) = \phi^2\gamma_{k-2}$$
$$= \phi^2(\phi\gamma_{k-3}) = \phi^3\gamma_{k-3}$$
$$= \phi^4\gamma_{k-4}$$
$$=$$
$$\dots$$
$$= \phi^k\gamma_0 .$$

From here, we deduce that

$$\gamma_k = \phi^2\gamma_{k-2} = \ldots = \phi^k\gamma_0, \quad k \geq 0.$$

By definition $\rho_k = \gamma_k/\gamma_0$. Then

$$\rho_0 = 1$$

$$\frac{\phi^k\gamma_0}{\gamma_0} = \phi^k$$

$$\rho_1 = \phi$$

$$\rho_2 = \phi^2$$

$$\ldots\ldots$$

$$\rho_k = \phi^k.$$

Note that differently from autocoavariance γ_k, autocorrelation ρ_k does not depend on the variance of the white noise ε_t.