2022 Fall
IE 313 Time Series Analysis

6. Model Specification

Yongjae Lee
Department of Industrial Engineering
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Chapter 6. Model Specification

- 6.1 Properties of the Sample Autocorrelation Function
- 6.2 The Partial and Extended Autocorrelation Functions
- 6.3 Specification of Some Simulated Time Series
- 6.4 Nonstationarity
- 6.5 Other Specification Methods
- 6.6 Specification of Some Actual Time Series

How can we choose models?

- So far, we have developed a large class of parametric models for both stationary and nonstationary time series – the ARIMA models

- Now we will begin to study statistical inference for such models
  - How to choose appropriate values for p, d, and q for a given series
  - How to estimate the parameters of a specific ARIMA(p,d,q) model
  - How to check on the appropriateness of the fitted model and improve it if needed

How can we choose models?

- For this part, I will skip most of the details and just summarize some contents from our textbook and reference book
  - Time Series Analysis and Its Applications (with R Examples), 4th edition by Robert H. Shumway & David S. Stoffer, Springer, 2017
    (eBook in pdf format can be downloaded from https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf)

Chapters 6.1 & 6.2

ACF, PACF, and EACF

Autocorrelation function (ACF)

- Recall from Chapter 4.2 that

General MA(q) process

- Consider a general MA(q) process

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad \leftarrow \text{stationary!}$$

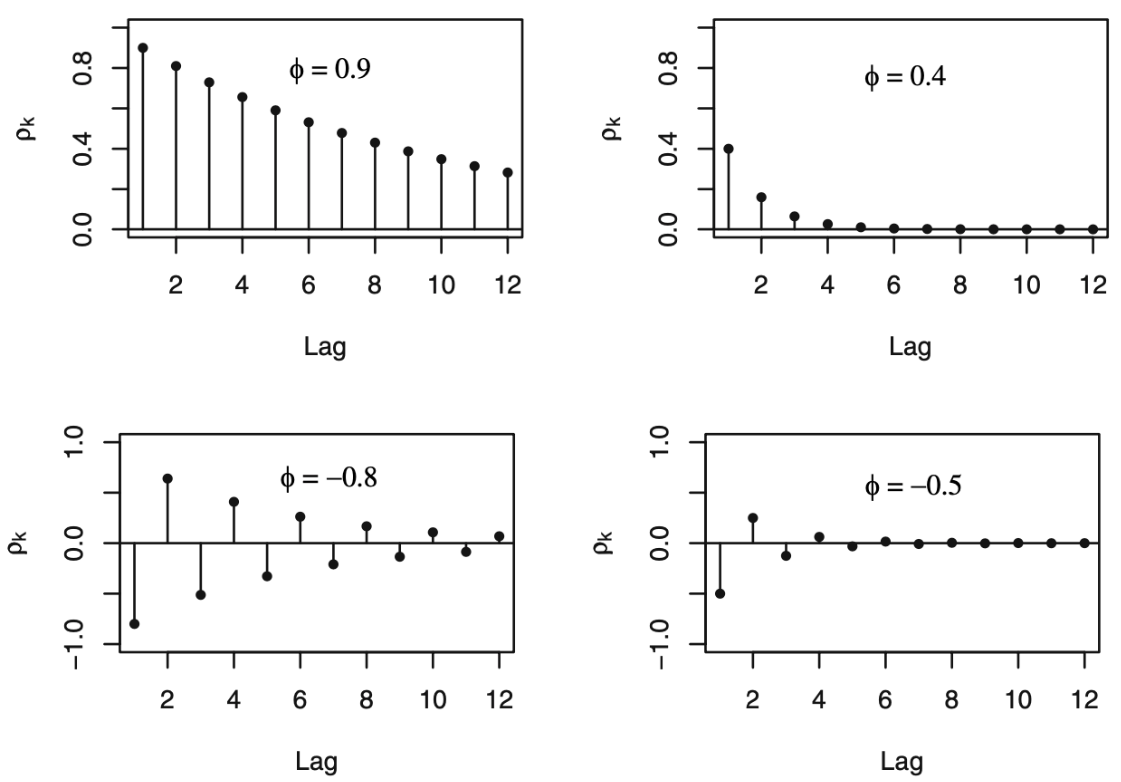
  - Similar calculations show that
    - $\gamma_0 = Var(Y_t) = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_e^2$
    - $\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} & \text{for } 1 \le k \le q \\ 0 & \text{for } k > q \end{cases}$
  - The autocorrelation function $\boxed{\textit{“cuts off”}}$ after lag q (become zero)
  - Its shape can be almost anything for the earlier lags

MGE313 Time Series Analysis 23

Autocorrelation function (ACF)

- Recall from Chapter 4.3 that

Exhibit 4.12 Autocorrelation Functions for Several AR(1) Models



Autocorrelation function (ACF)

- Suppose that we have a stationary time series data and plot its sample autocorrelation function
  - If the sample autocorrelation function becomes *insignificant* after lag q, a **good candidate model** would be MA(q)
    - Because we know that the ACF of MA(q) *cuts off* after lag q
  - Otherwise, we need to do some more tests
    - Because, if it is not an MA process, then we can say nothing from the ACF
      - The ACF of AR(p) *tails off* (its absolute value decareases as the lag increases, but there is no clear cut after some definite point)
      - The ACF of ARMA(p,q) also *tails off*

Partial autocorrelation function (PACF)

- Do we have any similar method for identifying the order of AR processes?

  – Yes, we can use the **partial autocorrelation function (PACF)**

  – First, let's see the **partial correlation** between X and Y given Z

$$\rho_{XY|Z} = corr(X - \hat{X}, Y - \hat{Y})$$

  - $\hat{X}$ is the linear regression of X on Z
    $\hat{Y}$ is the linear regression of Y on Z
  - The idea is that $\rho_{XY|Z}$ measures the correlation between X and Y with the linear effect of Z removed (or partialled out)

Partial autocorrelation function (PACF)

- Do we have any similar method for identifying the order of AR processes?

  – Yes, we can use the **partial autocorrelation function (PACF)**

  – The PACF of a stationary series is defined as

    - $\phi_{11} = corr(Y_{t+1}, Y_t) = \rho_1$
    - $\phi_{kk} = corr(Y_{t+k} - \hat{Y}_{t+k}, Y_t - \hat{Y}_t)$ for $k \geq 2$
      - $\hat{Y}_{t+k}$ is the linear regression of Y_{t+k} on $\{Y_{t+1}, \dots, Y_{t+k-1}\}$
      - $\hat{Y}_t$ is the linear regression of Y_t on $\{Y_{t+1}, \dots, Y_{t+k-1}\}$

  – Hence, PACF ϕ_{kk} is the correlation between Y_{t+k} and Y_t with the linear dependence of $\{Y_{t+1}, \dots, Y_{t+k-1}\}$ on each, removed

Partial autocorrelation function (PACF)

- Do we have any similar method for identifying the order of AR processes?
  - Yes, we can use the **partial autocorrelation function (PACF)**
  - Interesting thing about the PACF is that
    - The PACF of an AR(p) process *cuts off* after lag p
    - The PACF of an MA(q) process *tails off*
      (as well as the PACF of an ARMA(p,q) process)

ACF & PACF

Exhibit 6.3 General Behavior of the ACF and PACF for ARMA Models

	AR(p)	MA(q)	ARMA(p,q), $p>0$, and $q>0$
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

- Suppose that we have a stationary time series data and plot its sample ACF and sample PACF
  - If the sample ACF becomes *insignificant* after lag q, a **good candidate model** would be MA(q)
    - Because we know that the ACF of MA(q) *cuts off* after lag q
  - If the sample PACF becomes *insignificant* after lag p, a **good candidate model** would be AR(p)
    - Because we know that the PACF of AR(p) *cuts off* after lag p

EACF

- Then, how can we determine the orders of ARMA(p,q)?
  - Many methods were proposed by researchers
    - Corner method (Becuin et al., 1980)
    - Extended ACF (EACF) method (Tsay and Tiao, 1984)
    - Smallest canonical correlation (SCAN) method (Tsay and Tiao, 1985)
    - …
  - Among them, let’s briefly see how the EACF method works
    - The EACF method uses the fact that if the AR part of a mixed ARMA model is known, then “*filtering out*” the autoregression from the observed time series results in a pure MA process that enjoys the cutoff property in its ACF
    - Hence, we can try this for various orders of AR part p=1,2,3,…

EACF

Exhibit 6.4 Theoretical Extended ACF (EACF) for an ARMA(1,1) Model

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	0*	0	0	0	0	0	0	0	0	0	0	0	0
2	x	x	0	0	0	0	0	0	0	0	0	0	0	0
3	x	x	x	0	0	0	0	0	0	0	0	0	0	0
4	x	x	x	x	0	0	0	0	0	0	0	0	0	0
5	x	x	x	x	x	0	0	0	0	0	0	0	0	0
6	x	x	x	x	x	x	0	0	0	0	0	0	0	0
7	x	x	x	x	x	x	x	0	0	0	0	0	0	0

- Element at (i,j)
  - X: After filtering out AR(i), ACF at lag j+1 is significantly different from 0
  - O: Otherwise
- Theoretically, an ARMA(p,q) process will show a triangle of O's
  - Upper left vertex of the triangle would be a good candidate

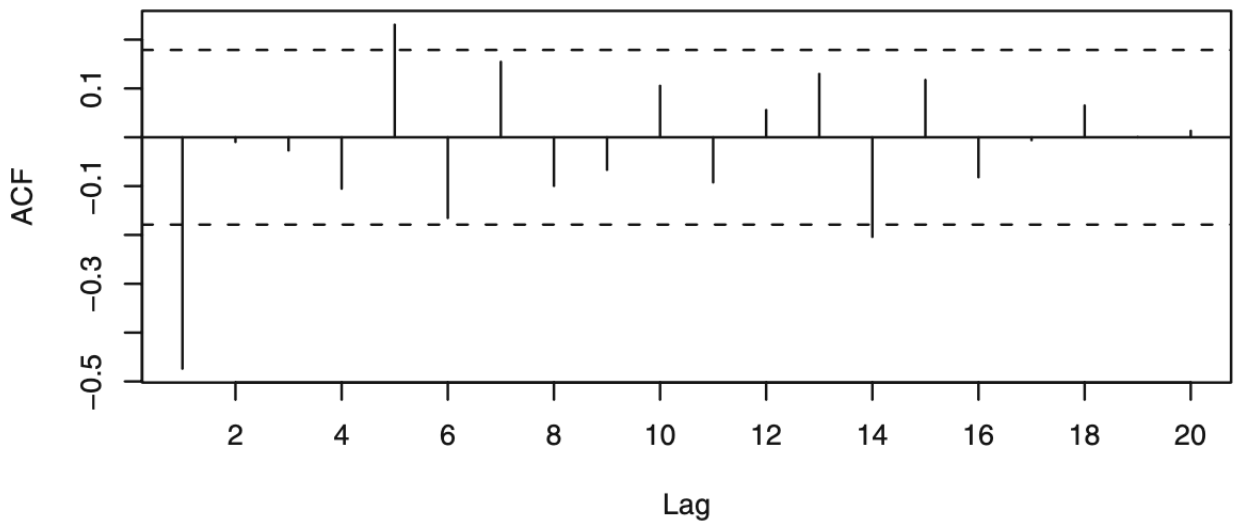
Chapter 6.3

Specification of Some Simulated Time Series

MA(1) with $\theta = 0.9$

Dashed horizontal lines represent two standard errors $\pm 2/\sqrt{n}$

Exhibit 6.5 Sample Autocorrelation of an MA(1) Process with $\theta = 0.9$

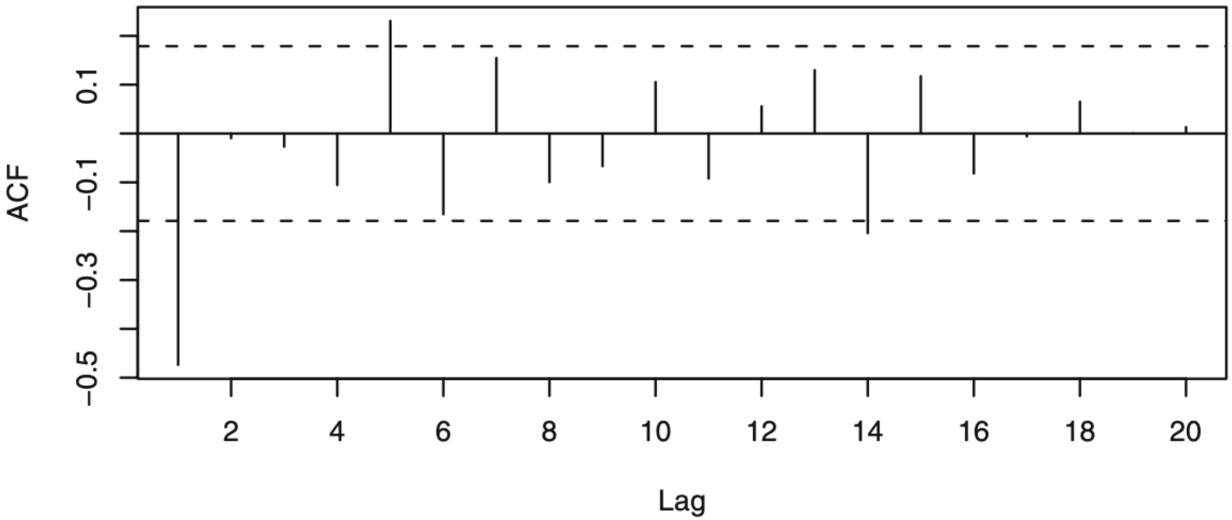


- Even though we know that the underlying process is MA(1), the above plot is slightly different from theoretical results because
  - It is the **sample** ACF based on a **simulation** of the MA(1) process
    (The above is based on 120 points, and more points would make the plot much closer to theoretical values)

MA(1) with $\theta = 0.9$

Dashed horizontal lines represent two standard errors $\pm 2/\sqrt{n}$

Exhibit 6.5 Sample Autocorrelation of an MA(1) Process with θ = 0.9



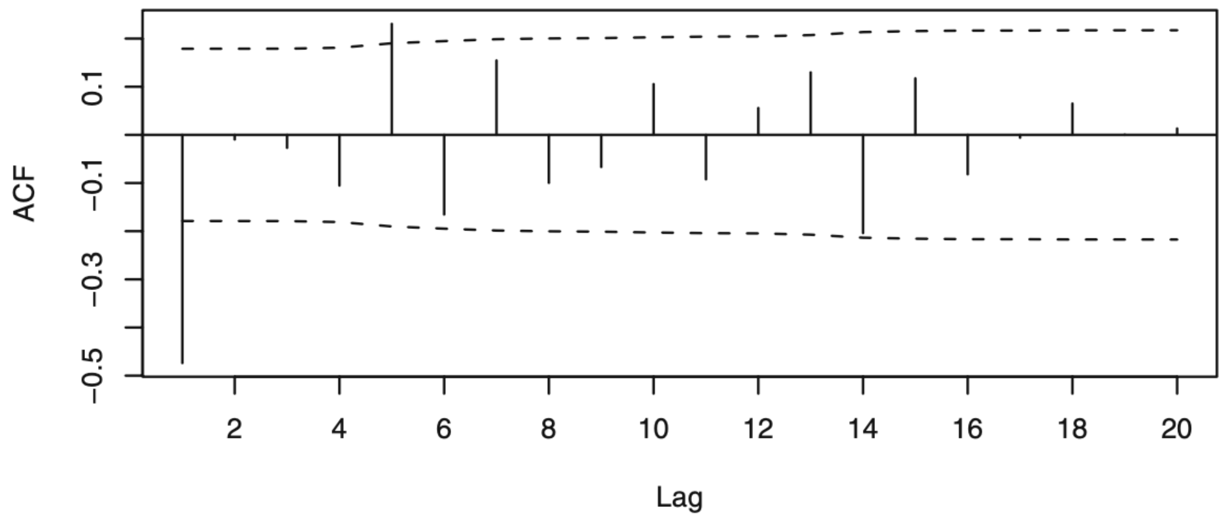
- Very significant at lag 1, slightly significant at lags 5 and 14

- Suggested models?
  - May be MA(1), MA(5), MA(14)?

MA(1) with $\theta = 0.9$

Alternative bound calculated using Equation (6.1.11) on page 112 of the textbook

Exhibit 6.6 Alternative Bounds for the Sample ACF for the MA(1) Process

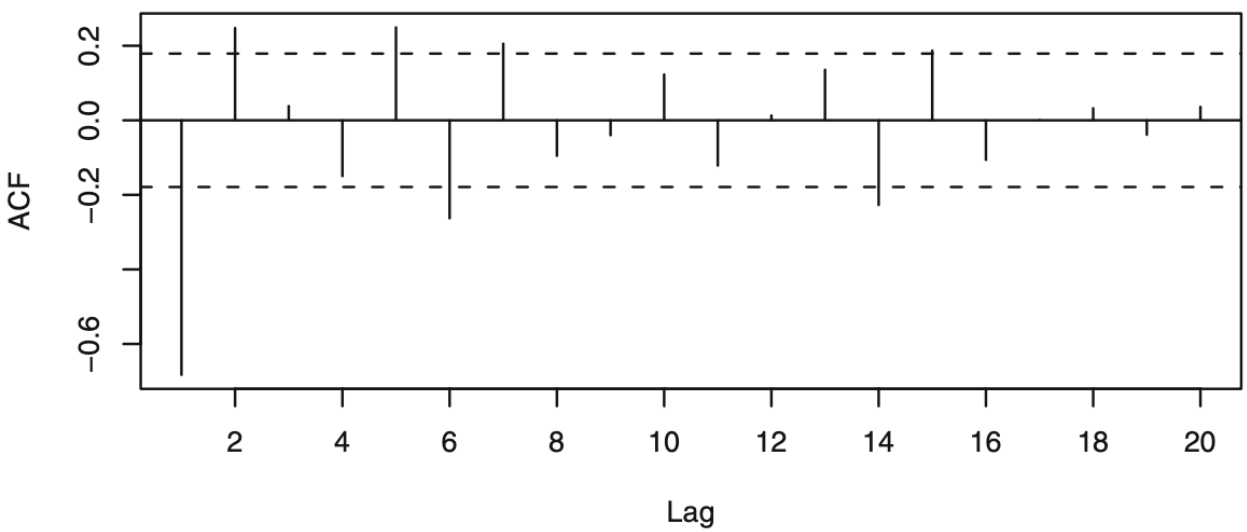


- Very significant at lag 1, slightly significant at lag 5

- Suggested models?
  – More like MA(1), may be MA(5)?

MA(2) with $\theta_1 = 1$ and $\theta_2 = -0.6$

Exhibit 6.8 **Sample ACF for an MA(2) Process with θ_1 = 1 and θ_2 = −0.6**

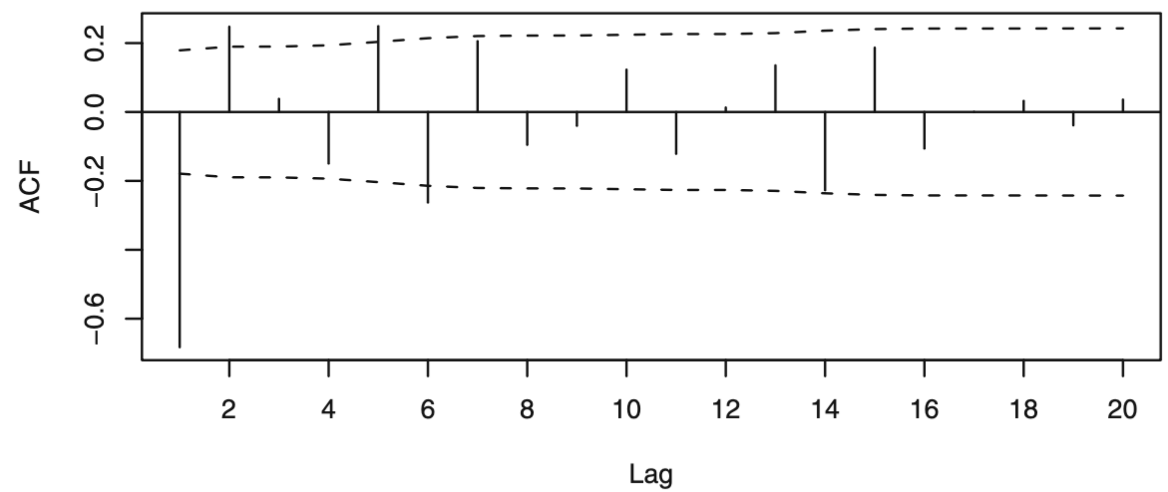


- Very significant at lag 1, slightly significant at lags 2, 5, 6, 7, and 14

- Suggested models?
  – May be MA(1), MA(2), MA(5), MA(6)?

MA(2) with $\theta_1 = 1$ and $\theta_2 = -0.6$

Exhibit 6.9 **Alternative Bounds for the Sample ACF for the MA(2) Process**



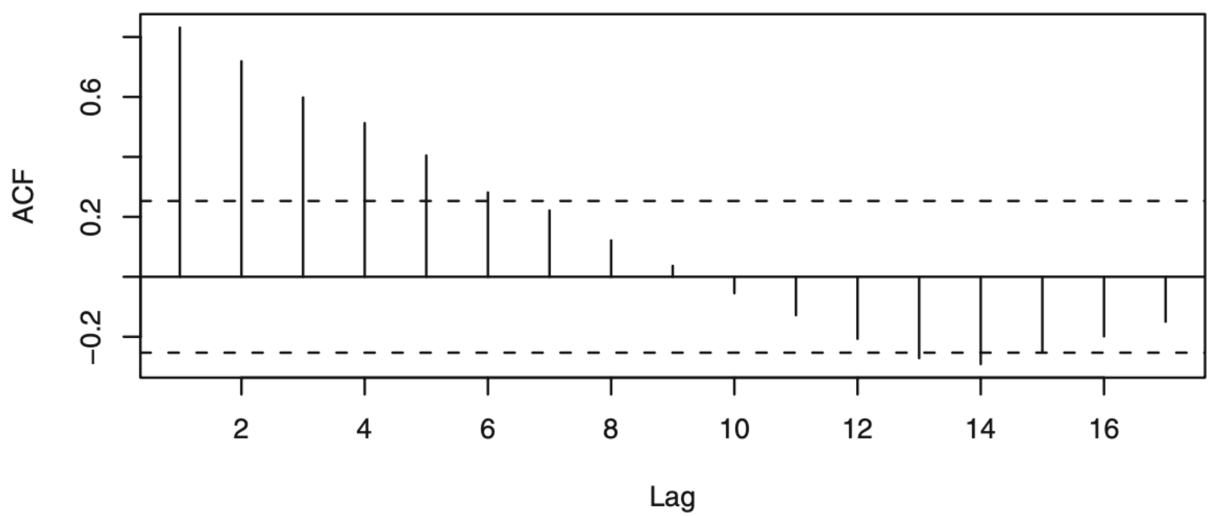
- Very significant at lag 1, slightly significant at lags 2, 5, and 6

- Suggested models?
  - More like MA(1), may be MA(2), MA(5), MA(6)?

(Probably we would need more samples or tests to arrive at the true model MA(2))

AR(1) with $\phi = 0.9$

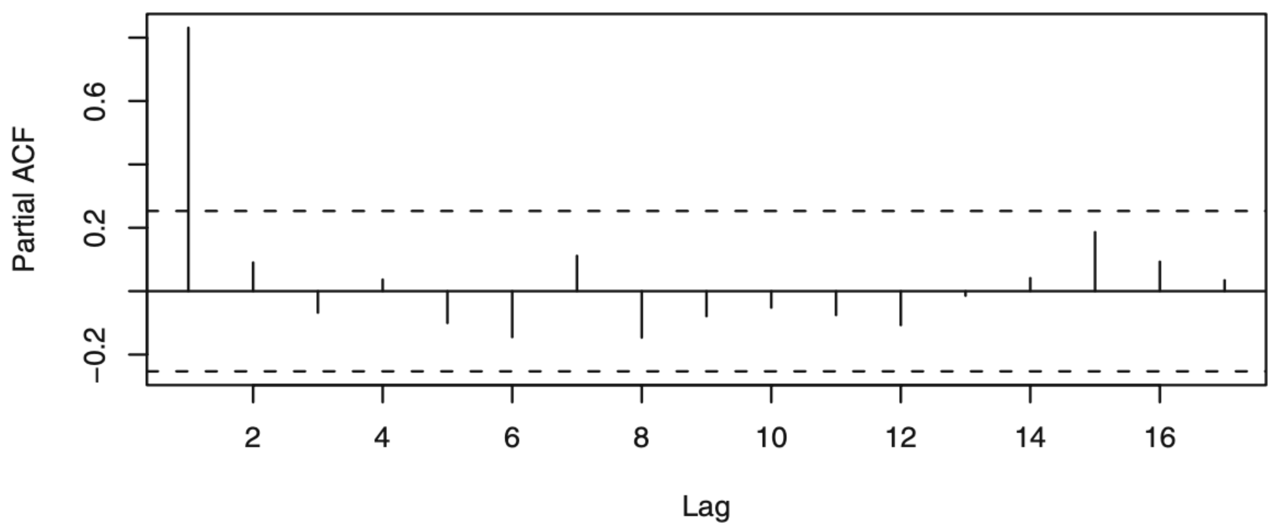
Exhibit 6.10 Sample ACF for an AR(1) Process with $\phi = 0.9$



- ACF seems to tail off
- So we may exclude MA processes

AR(1) with $\phi = 0.9$

Exhibit 6.11 Sample Partial ACF for an AR(1) Process with $\phi = 0.9$

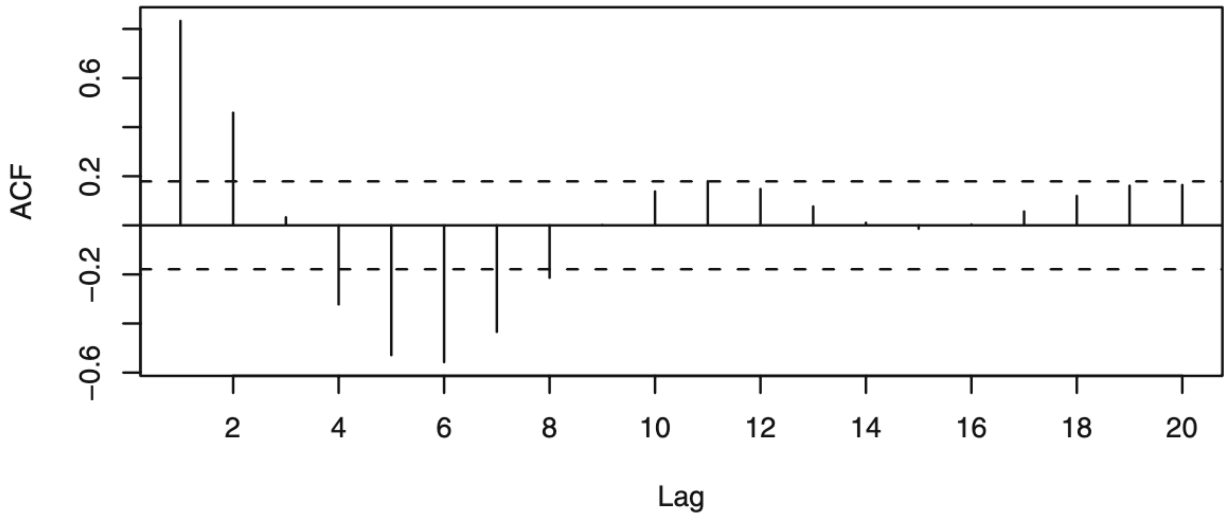


- Very significant at lag 1

- Suggested models?
  - Very much like AR(1)

AR(2) with $\phi_1 = 0.9$ and $\phi_2 = -0.75$

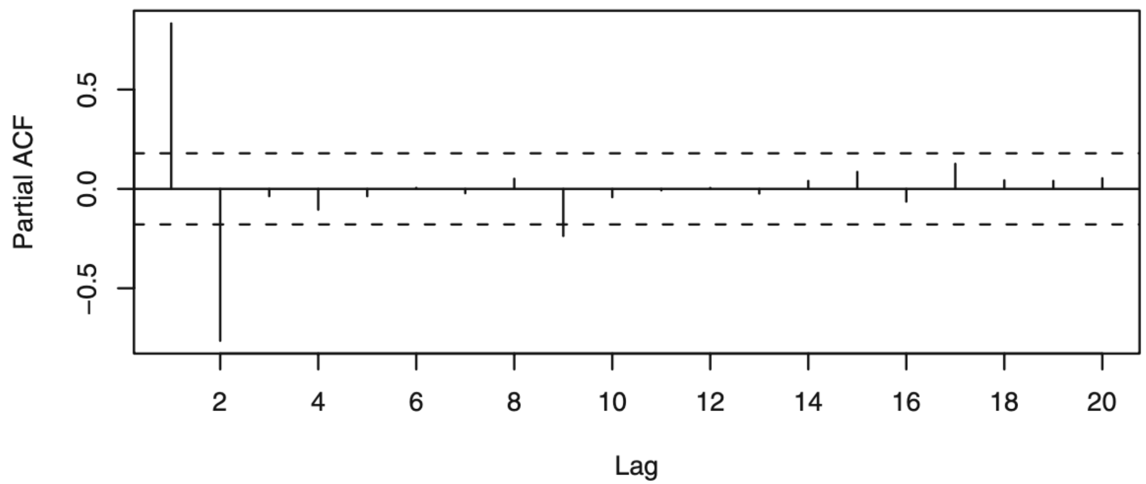
Exhibit 6.12 Sample ACF for an AR(2) Process with $\phi_1 = 1.5$ and $\phi_2 = -0.75$



- ACF seems to tail off
- So we may exclude MA processes

AR(2) with $\phi_1 = 0.9$ and $\phi_2 = -0.75$

Exhibit 6.13 **Sample PACF for an AR(2) Process with $\phi_1 = 1.5$ and $\phi_2 = -0.75$**

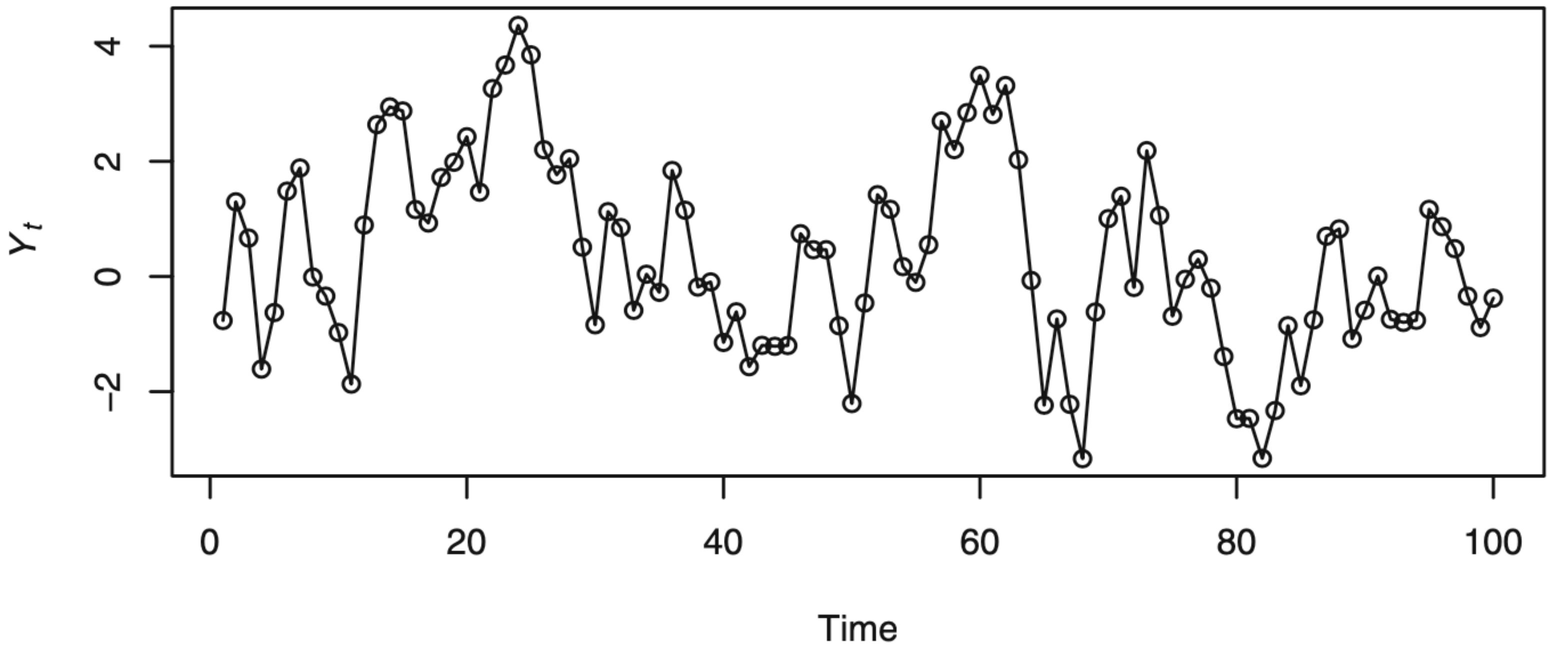


- Very significant at lags 1 and 2, slightly significant at lag 9

- Suggested models?
  - Very much like AR(2), may be AR(9)

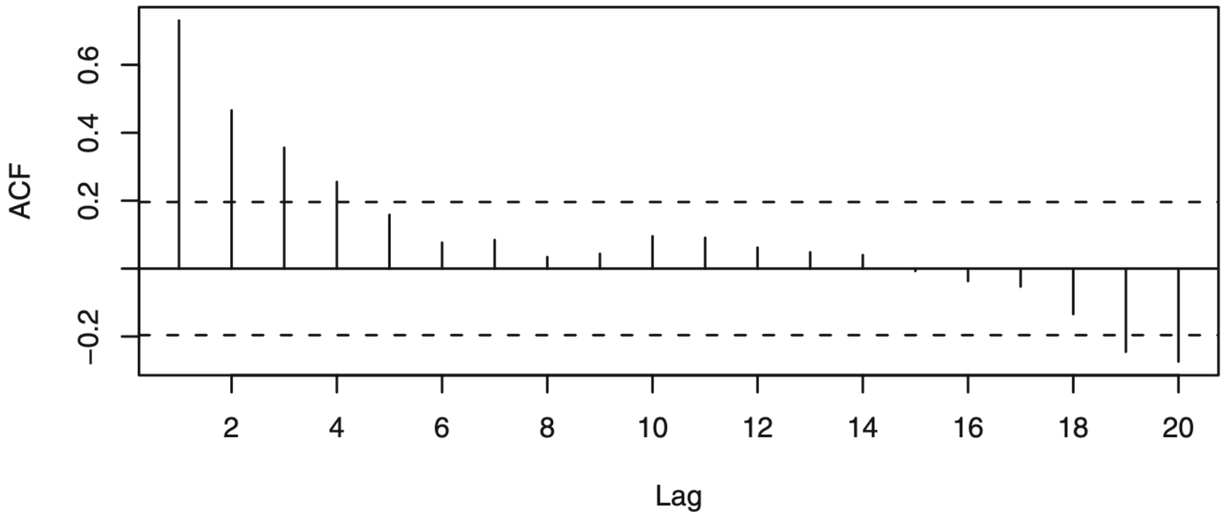
ARMA(1,1) with $\phi = 0.6$ and $\theta = -0.3$

Exhibit 6.14 Simulated ARMA(1,1) Series with $\phi = 0.6$ and $\theta = -0.3$.



ARMA(1,1) with $\phi = 0.6$ and $\theta = -0.3$

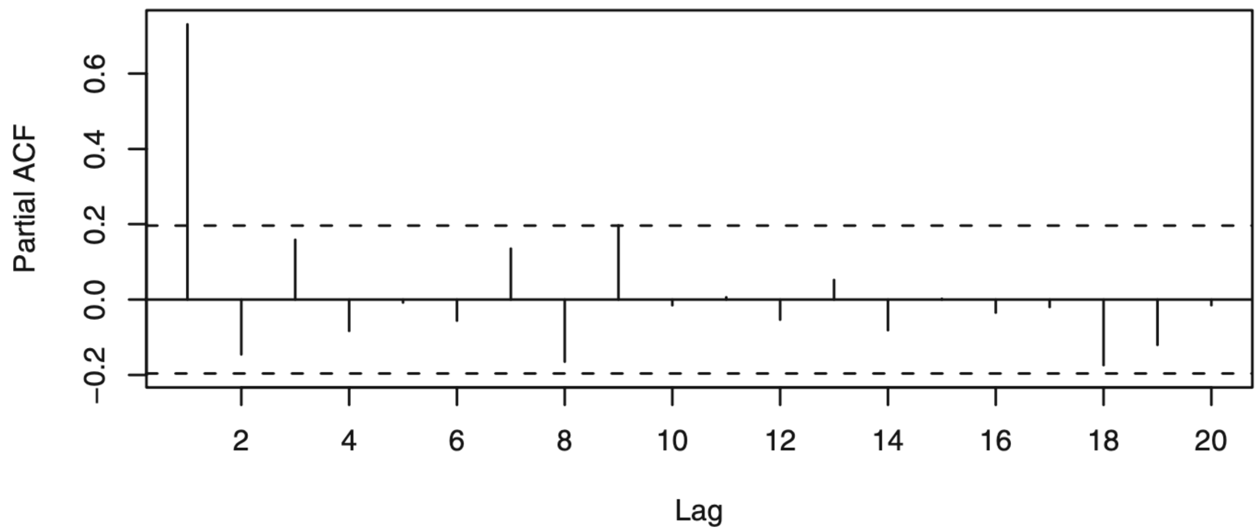
Exhibit 6.15 Sample ACF for Simulated ARMA(1,1) Series



- ACF seems to tail off
- So we may exclude MA processes

ARMA(1,1) with $\phi = 0.6$ and $\theta = -0.3$

Exhibit 6.16 Sample PACF for Simulated ARMA(1,1) Series



- Very significant at lag 1
- Suggested models?
  - Very much like AR(1)..

ARMA(1,1) with $\phi = 0.6$ and $\theta = -0.3$

Exhibit 6.17 Sample EACF for Simulated ARMA(1,1) Series

AR / MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	o	o	o	o	o	o	o	o	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o
4	x	o	x	o	o	o	o	o	o	o	o	o	o	o
5	x	o	o	o	o	o	o	o	o	o	o	o	o	o
6	x	o	o	o	x	o	o	o	o	o	o	o	o	o
7	x	o	o	o	x	o	o	o	o	o	o	o	o	o

- No clear triangle of O’s

- Suggested models?
  - More like ARMA(1,1) or ARMA(2,1)..

Chapter 6.4

Nonstationarity

Nonstationarity

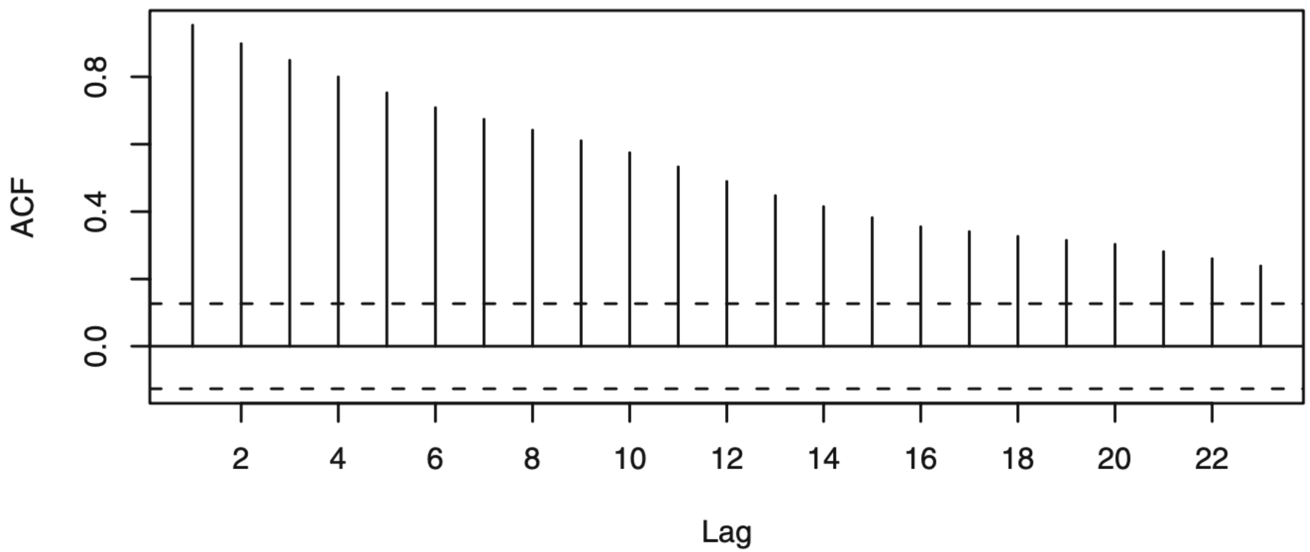
- As we have seen in Chapter 5, many time series exhibit nonstationarity that can be explained by integrated ARMA (ARIMA) models

- The **sample ACF** computed for nonstationarity series will also usually **indicate the nonstationarity**
  - Note that the definition of ACF implicitly *assumes* stationarity
  - Thus it is not at all clear what the sample ACF is estimating for a nonstationary process

- Nevertheless, for nonstationary process, the sample ACF **typically fails to die out rapidly** as the lags increase
  - This is due to the tendency for nonstationary series to drift slowly, either up or down, with apparent “trends”

Checking nonstationarity via sample ACF

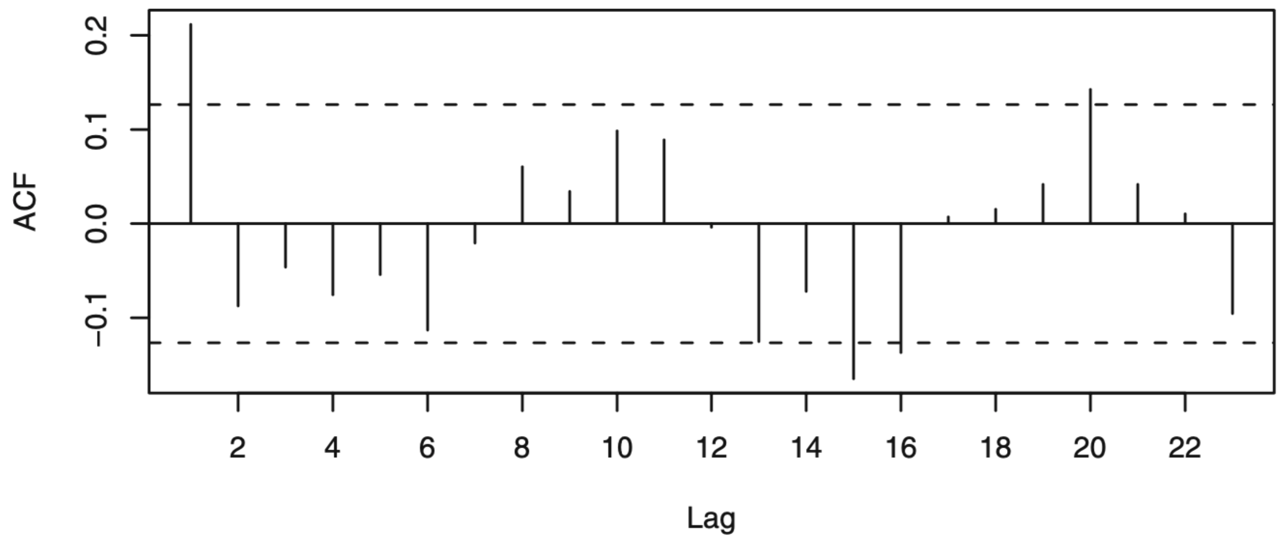
Exhibit 6.18 Sample ACF for the Oil Price Time Series



- Sample ACF of the logs of the oil price series (textbook p.88)
  - All values (even lags > 20) are "significantly far from zero"
  - This **linear decay** of the sample ACF is often taken as a symptom of nonstationarity

Checking nonstationarity via sample ACF

Exhibit 6.19 Sample ACF for the Difference of the Log Oil Price Series



- Sample ACF of the **first difference** of the logs of the oil prices
  - Now the pattern emerges much more clearly
    - From this, MA(1) seems to be a good choice
    - i.e., the original series would be a nonstationary IMA(1,1) model

Checking nonstationarity via sample ACF

- If the first difference of a series and its sample ACF do not support a stationary ARMA model, then we take another difference and again compute the sample ACF and PACF to look for characteristics of a stationary ARMA proces

- Usually **one or at most two differences**, perhaps **combinded with a logarithm of other transformation**, will accomplish this reduction to stationarity

Overdifferencing

- Difference of any stationary time series is also stationary (see Exercise 2.7 on page 20)

- However, overdifferencing introduces unnecessary correlations into a series and will complicate the modeling process
  - Consider a random walk process $\{Y_t\}$, which is nonstationary
  - The first difference would make it a white noise series, and thus, stationary
$$\nabla Y_t = Y_t - Y_{t-1} = e_t$$
  - However, if we difference once more, we have
$$\nabla^2 Y_t = e_t - e_{t-1}$$
  - Which is an MA(1) model but with $\theta = 1$

Overdifferencing

- An appropriate modeling of $\{Y_t\}$ would be IMA(1,1) with $\theta = 0$ (or ARI(1,1) with $\phi = 0$ is also plausible)

- But overdifferencing would suggest us an IMA(2,1) model with $\theta = 1$
  - We are having a more complex model for no reason
  - Also, this makes a noninvertible model
    - It creates a serious problems when we attempt to estimate their parameters

Overdifferencing

Exhibit 6.21 Sample ACF of Correctly Differenced Random Walk

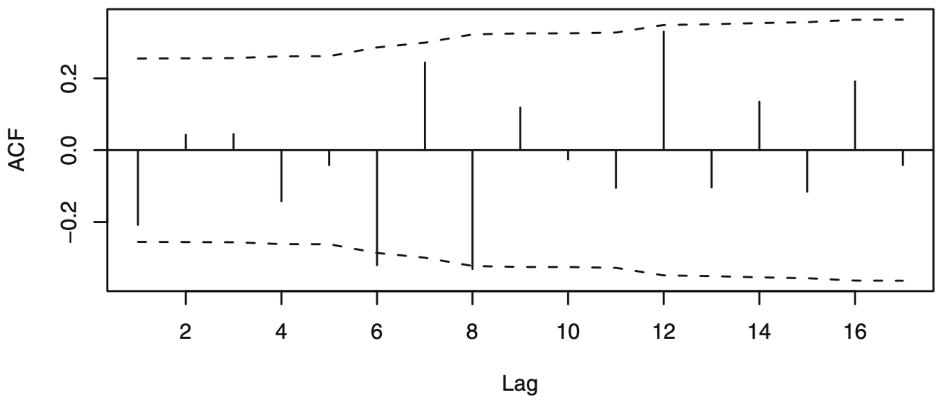
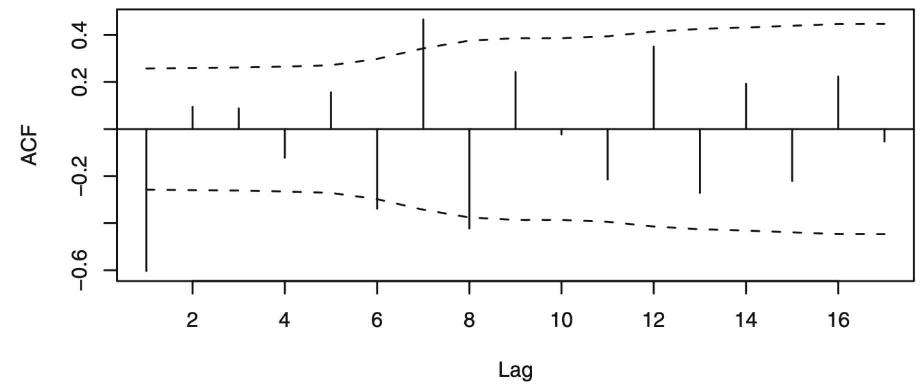


Exhibit 6.20 Sample ACF of Overdifferenced Random Walk



- To avoid overdifferencing,
  - It is recommended to look carefully at each difference in succssion
  - And keep the principle of parsimony always in mind
    - *Models should be simple, but not too simple*

Dickey-Fuller unit-root test

- It is also useful to quantify the evidence of nonstationarity in the data-generating mechanism

- And it can be done via hypothesis testing. Let me give a sketch of idea behind Dickey-Fuller unit-root test

- Consider the model
  - $Y_t = \alpha Y_{t-1} + X_t$, for $t = 1,2, \ldots,$
    - where $\{X_t\}$ is a stationary series
  - $\{Y_t\}$ is nonstationary if $\alpha = 1$
  - $\{Y_t\}$ is stationary if $-1 < \alpha < 1$

Dickey-Fuller unit-root test

$Y_t = \alpha Y_{t-1} + X_t$, for $t = 1,2,\ldots,$
where $\{X_t\}$ is a stationary series

– Suppose that $\{X_t\}$ is an AR(k) process
  - $X_t = \phi_1 X_{t-1} + \cdots + \phi_k X_{t-k} + e_t$

– If we assume that $\alpha = 1$ (i.e., $X_t = Y_t - Y_{t-1}$) and let $a = (\alpha - 1)$,

$$\begin{aligned} Y_t - Y_{t-1} &= (\alpha - 1)Y_{t-1} + X_t \\ &= aY_{t-1} + \phi_1 X_{t-1} + \cdots + \phi_k X_{t-k} + e_t \\ &= aY_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \cdots + \phi_k(Y_{t-k} - Y_{t-k-1}) + e_t \end{aligned}$$

– Hence, if we perform regression of the first difference of the observed time series $(Y_t - Y_{t-1})$ on the past k lags of the first difference of the observed series $(Y_{t-1} - Y_{t-2}, \ldots, Y_{t-k} - Y_{t-k-1})$,
  - If $a = 0$, then $\alpha = 1$, and thus, Y_t is nonstationary but stationary after first differencing
  - Alternative is that Y_t is stationary

Dickey-Fuller unit-root test

- In practice, even after first differencing, the process may not be a finite-order AR process
  – But it may be closely approximated by some AR process with the AR order increasing with the sample size

Chapter 6.5

Other Specification Methods

Akaike's Information Criterion (AIC)

- **Akaike's Information Criterion (AIC)**

$$\text{AIC} = -2\log(\text{maximum likelihood}) + 2k$$

  - Maximum likelihood: how likely our chosen model is given the observed data (will be briefly discussed in Chapter 7)
  - k: number of parameters
    (for ARMA without constant term, $k = p + q$)

  - This criterion says to select the model that *minimizes* AIC
    - -log(maximum likelihood) will have nonnegative value and become close to zero as the model becomes more likely
    - $2k$ can be seen as a "penalty function" to help ensure selection of parsimonious models and to avoid choosing models with too many parameters

Corrected AIC

- When $k/n > 10\%$ (i.e., when we have small sample size n), the corrected AIC might be more appropriate

$$\mathrm{AIC_c} = AIC + \frac{2(k+1)(k+2)}{n-k-2}$$

Bayesian Information Criterion (BIC)

- **Bayesian Information Criterion (BIC)**

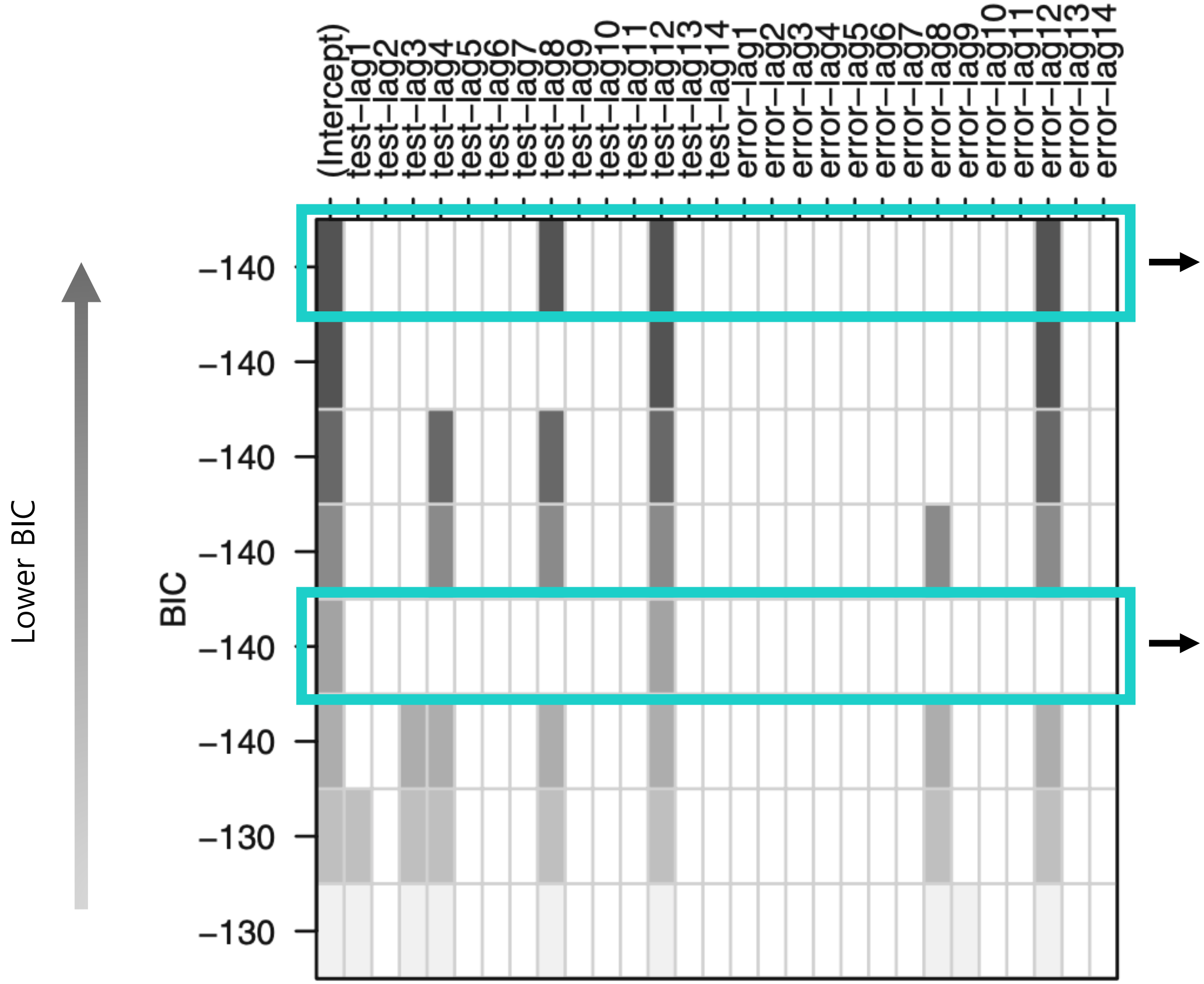
$$\mathrm{BIC} = -2\log(\text{maximum likelihood}) + k\log n$$

  – n: (effective) sample size

- This criterion says to select the model that *minimizes* BIC
  – If the true process follows an ARMA(p,q) model, then it is known that the orders specified by minimizing the BIC are consistent
    - That is, they approach the true orders as the sample size increases

Bayesian Information Criterion (BIC)

Exhibit 6.22 Best Subset ARMA Selection Based on BIC



This corresponds to the model that includes

- Intercept θ_0
- Test-lag7 Y_{t-7}
- Test-lag12 Y_{t-12}
- Error-lag12 e_{t-12}

Hence, ARMA(12,12)

This corresponds to the model that includes

- Intercept θ_0
- Test-lag12 Y_{t-12}

Hence, AR(12)

Chapter 6.6

Specification of Some Actual Time Series

Los Angeles annual rainfall series

Exhibit 1.1 Time Series Plot of Los Angeles Annual Rainfall

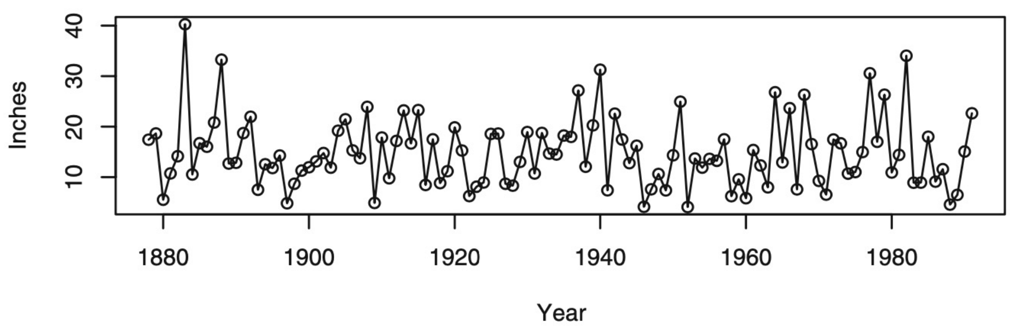
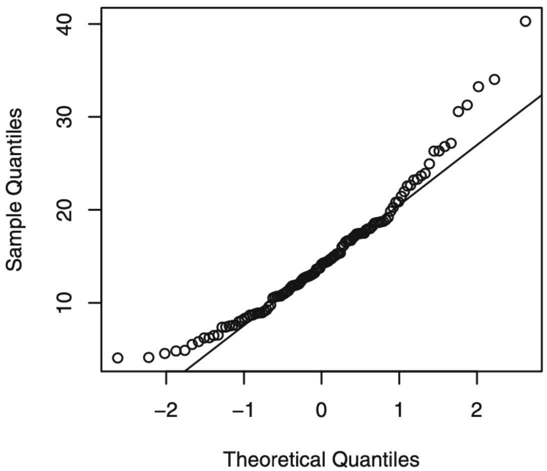


Exhibit 3.17 Quantile-Quantile Plot of Los Angeles Annual Rainfall Series



- In Chapter 3 (page 50), we have seen from Exhibit 3.17 that the rainfall amounts were **not normally distributed**
  - If the Q-Q plot is close to a straight line, then the data can be seen as normally distributed

Los Angeles annual rainfall series

Exhibit 1.1 Time Series Plot of Los Angeles Annual Rainfall

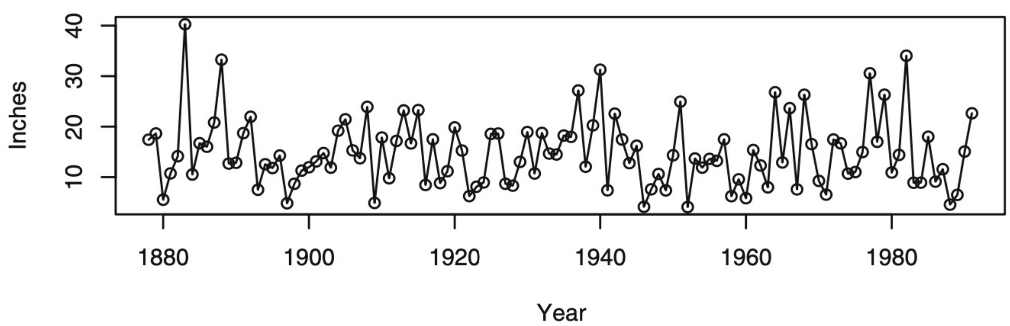
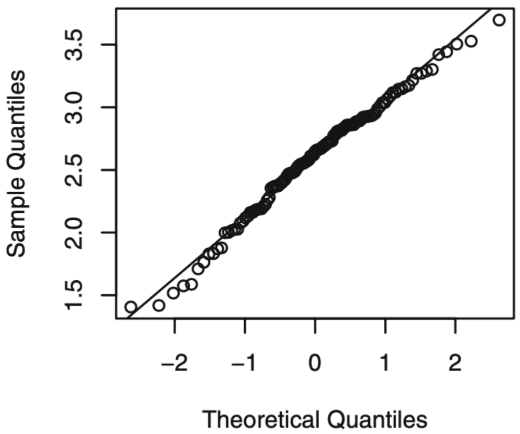


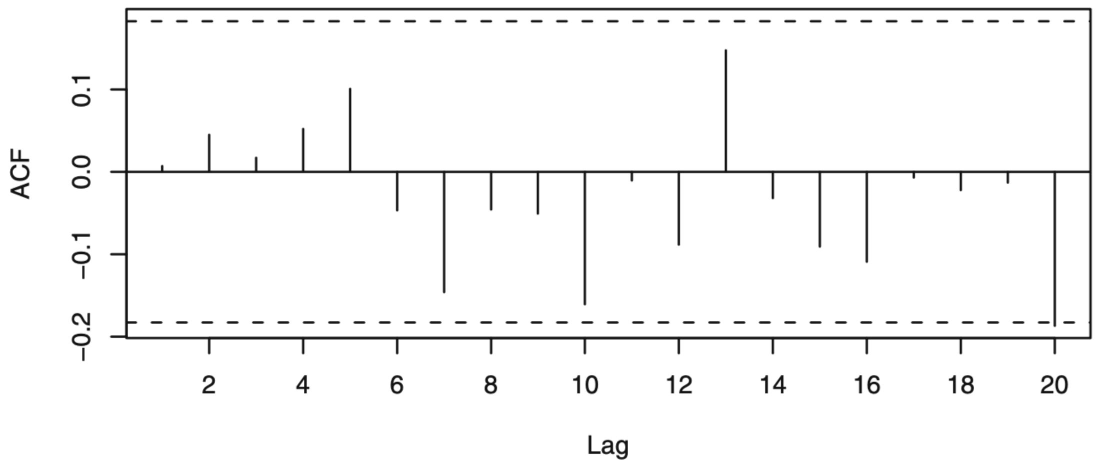
Exhibit 6.23 QQ Normal Plot of the Logarithms of LA Annual Rainfall



- But taking logarithms **improves the normality dramatically**

Los Angeles annual rainfall series

Exhibit 6.24 Sample ACF of the Logarithms of LA Annual Rainfall



- The sample ACF of *logs* of rainfall series shows **no discernable dependence**

- Therefore, we may model the **log** of annual rainfall amount as **independent normal random variables**

Chemical process color property series

Exhibit 1.3 Time Series Plot of Color Property from a Chemical Process

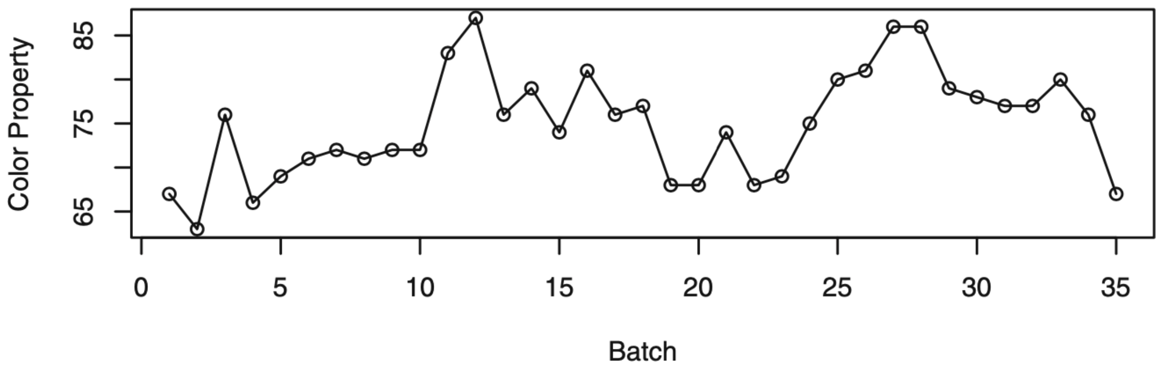
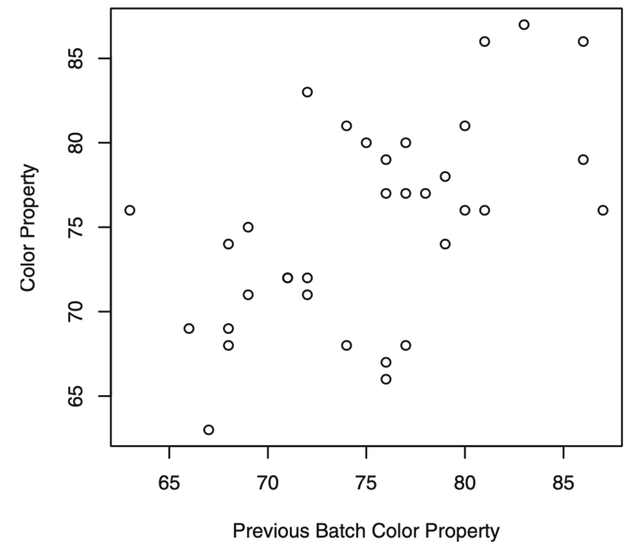


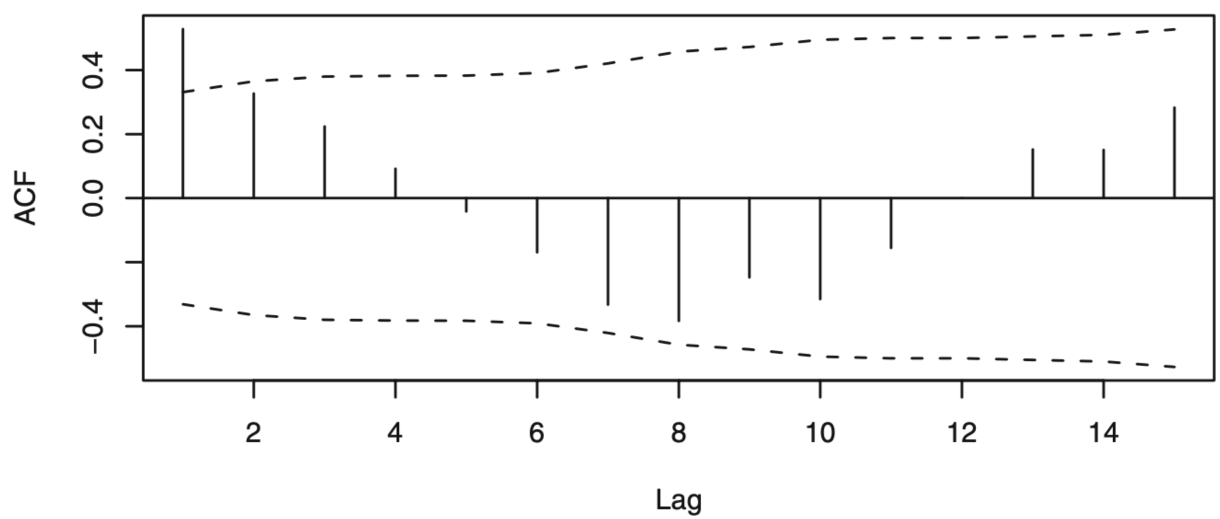
Exhibit 1.4 Scatterplot of Color Value versus Previous Color Value



- Exhibit 1.4 show **some dependence** of successive batches

Chemical process color property series

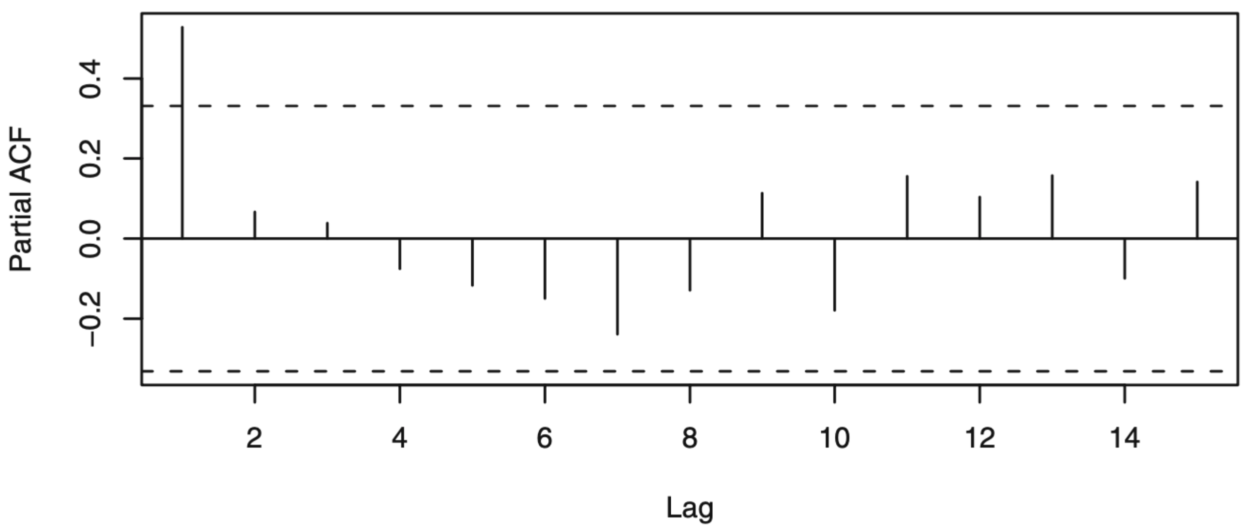
Exhibit 6.25 Sample ACF for the Color Property Series



- The sample ACF might at first glance suggest an **MA(1)** model, as only the lag 1 ACF is significantly different from zero
- However, the **damped sine wave** appearance of the plot encourages us to look further

Chemical process color property series

Exhibit 6.26 Sample Partial ACF for the Color Property Series



- The sample PACF clearly suggest that an **AR(1)** model is worthy of first consideration
- As always, our specified models are tentative and subject to modification during the model diagnostics stage of model building

Abundance of Canadian hare

Exhibit 1.5 Abundance of Canadian Hare

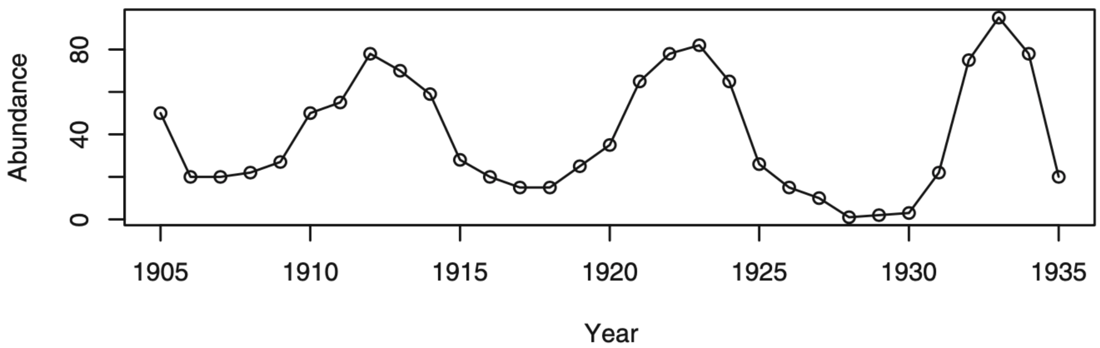
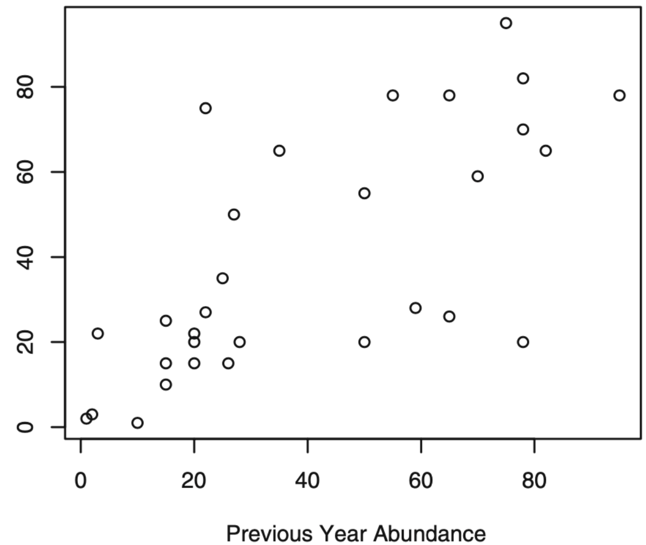


Exhibit 1.6 Hare Abundance versus Previous Year's Hare Abundance



- Exhibit 1.6 shows the **year-to-year dependence**

Abundance of Canadian hare

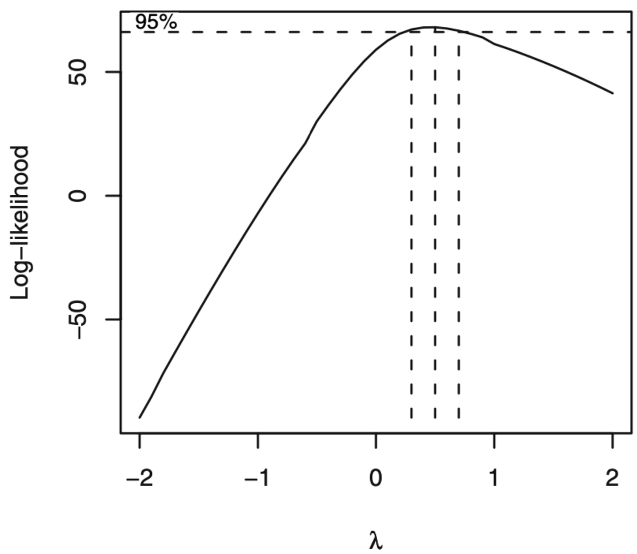
Power transformations

Chapter 5

- A more general **power transformation** was introduced by Box and Cox (1964)
  - For a given value of the parameter λ,

$$g(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log x & \text{for } \lambda = 0 \end{cases}$$

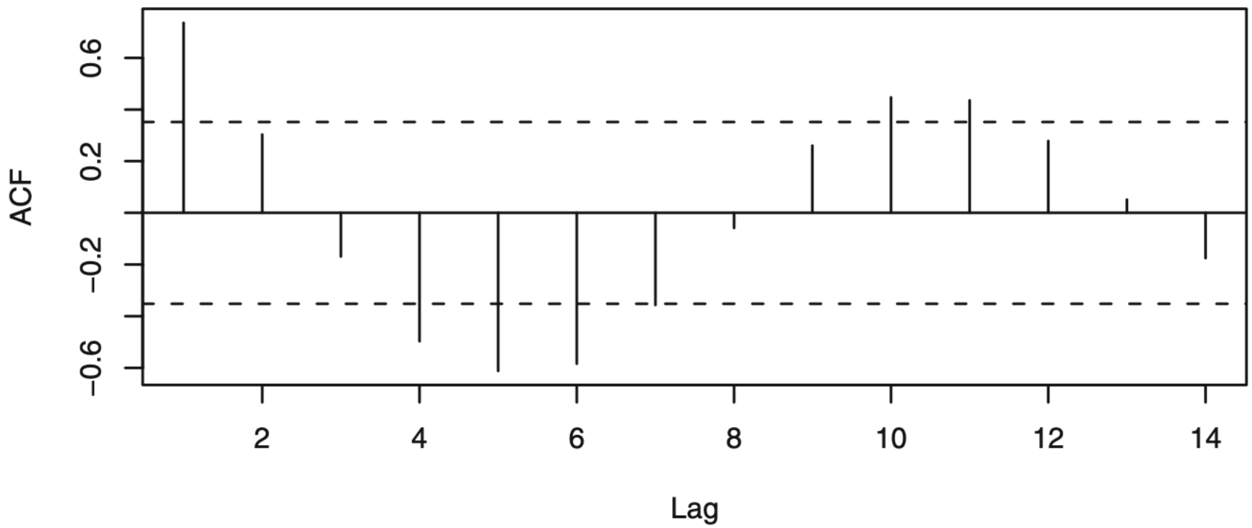
Exhibit 6.27 Box-Cox Power Transformation Results for Hare Abundance



- It has been suggested in the literature that a transformation might be used to produce a good model for these data
  - Exhibit 6.27 displays the log-likelihood as a function of the power parameter λ
  - The maximum occurs at $\lambda = 0.4$, but a **square root** transformation ($\lambda = 0.5$) is well within the confidence interval for λ

Abundance of Canadian hare

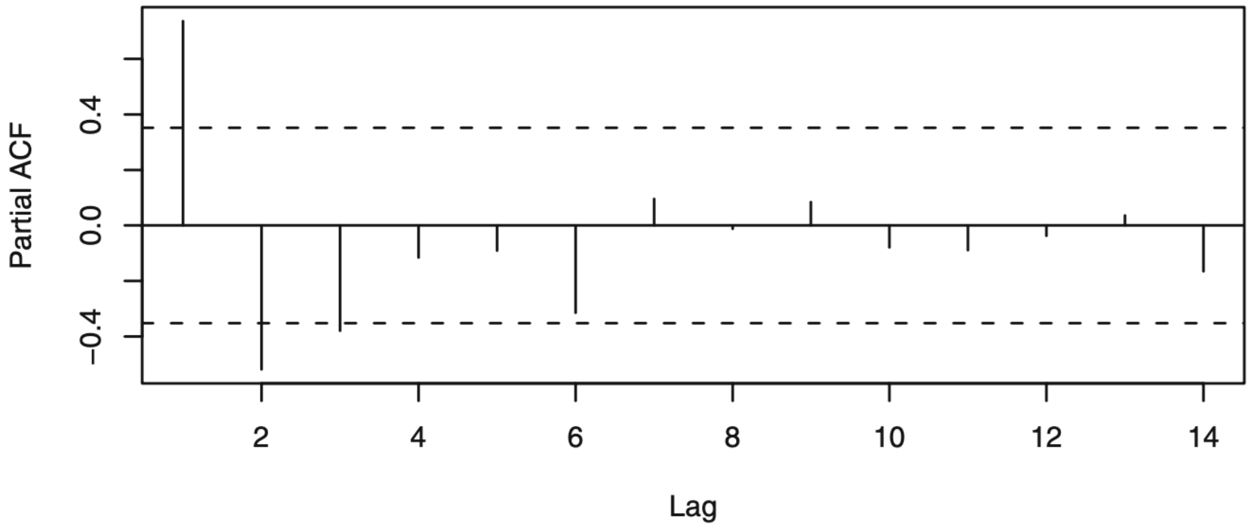
Exhibit 6.28 Sample ACF for Square Root of Hare Abundance



- The sample ACF of the *square root transformed* series has a fairly strong lag 1 ACF, but there is a strong indication of **dampened oscilatory behavior**

Abundance of Canadian hare

Exhibit 6.29 Sample Partial ACF for Square Root of Hare Abundance



- The sample PACF of square root transformed series gives strong evidence to support **AR(2)** or possibly an **AR(3)**

Oil price series

Exhibit 5.1 Monthly Price of Oil: January 1986–January 2006

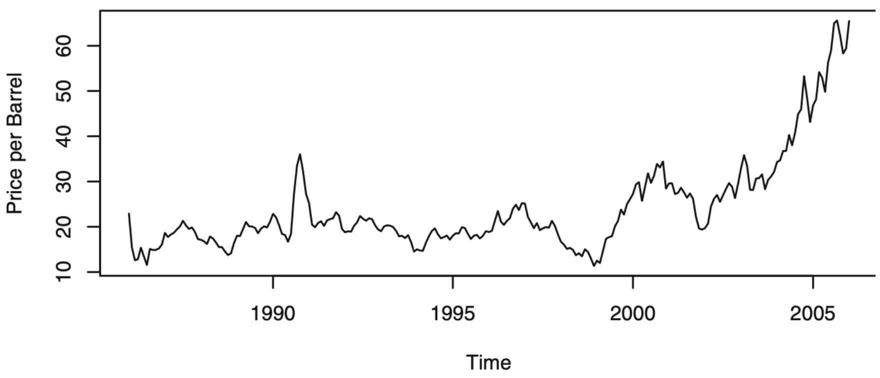
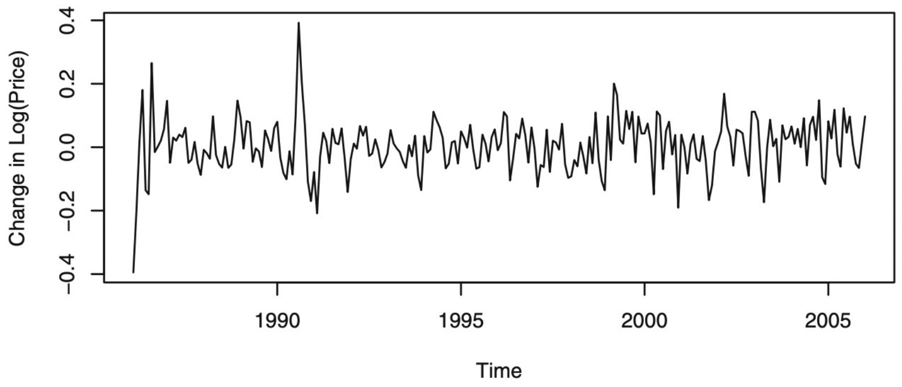


Exhibit 5.4 The Difference Series of the Logs of the Oil Price Time



- In Chapter 5, we have argued graphically that the **difference of the logarithms** could be considered **stationary**
  - (Augmented Dickey-Fuller test result also supports this)

Oil price series

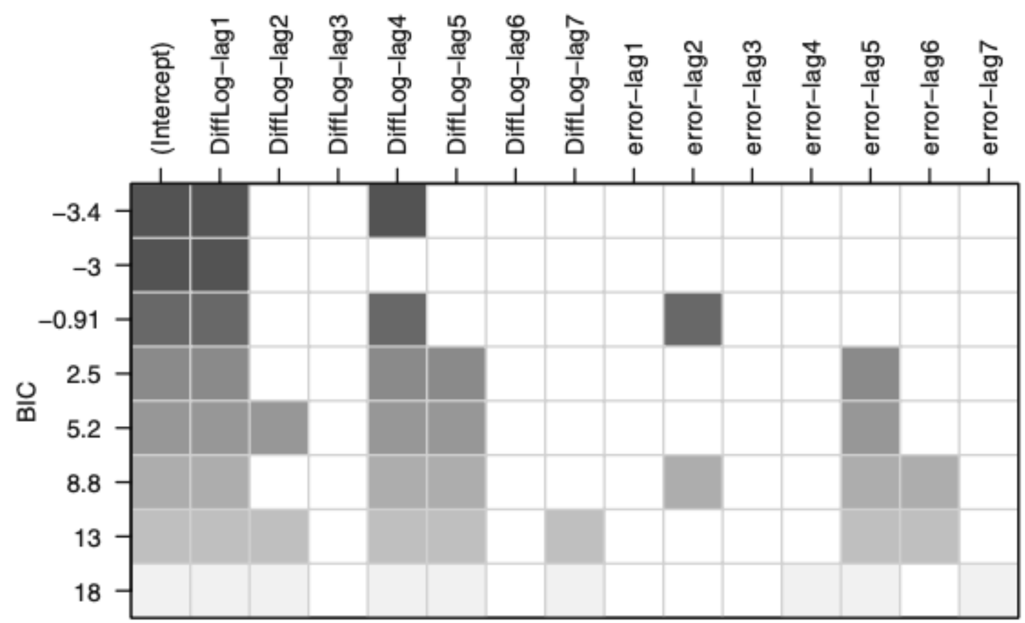
Exhibit 6.30 Extended ACF for Difference of Logarithms of Oil Price Series

AR / MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	o	o	o	o	o	o	o	o	o	o	o	o
1	x	x	o	o	o	o	o	o	o	o	x	o	o	o
2	o	x	o	o	o	o	o	o	o	o	o	o	o	o
3	o	x	o	o	o	o	o	o	o	o	o	o	o	o
4	o	x	x	o	o	o	o	o	o	o	o	o	o	o
5	o	x	o	x	o	o	o	o	o	o	o	o	o	o
6	o	x	o	x	o	o	o	o	o	o	o	o	o	o
7	x	x	o	x	o	o	o	o	o	o	o	o	o	o

- EACF table suggests that an **ARMA model with p = 0 and q = 1 (i.e., MA(1))** would be appropriate for the first difference of logs of oil price series

Oil price series

Exhibit 6.31 Best Subset ARMA Model for Difference of Log(Oil)



- But, best subsets ARMA approach based on BIC suggests
  – A model with θ_0, Y_{t-1}, and Y_{t-4} (i.e., ARIMA(4,1,0) on log)
  – Or a model with θ_0 and Y_{t-1} (i.e., ARIMA(1,1,0) on log)

Oil price series

Exhibit 6.32 Sample ACF of Difference of Logged Oil Prices

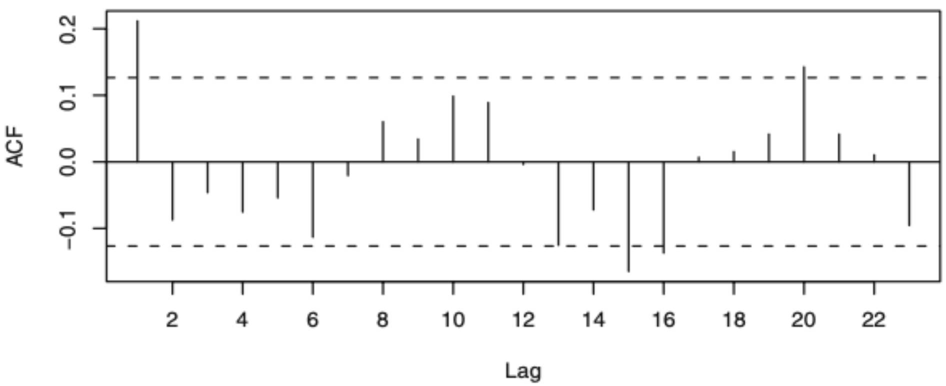
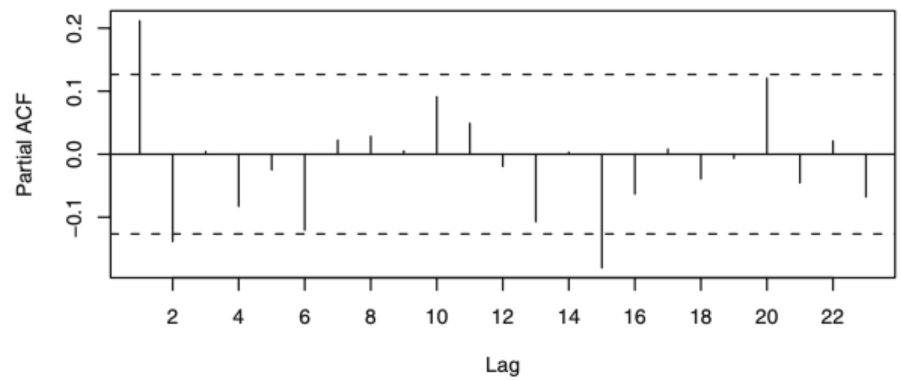


Exhibit 6.33 Sample PACF of Difference of Logged Oil Prices



- Sample ACF of difference of logged oil prices suggests **MA(1)**
- Sample PACF of difference of logged oil prices suggests **AR(2)**

- Different tests give different suggestions
  - Need to look at them further when we estimate parameters and perform diagnostic tests in Chapters 7 and 8
  - (actually, it will be later shown that outliers should be dealt with)