DATA ANALYTICS

Unit 3: Ljung Box and Theil's coefficient

Jyothi R. and Bharathi R

Department of Computer Science and Engineering

Differencing

- In Figure 1 to 9: The Google stock price was non-stationary in panel (a)
- But the daily changes were stationary in panel (b). This shows one way to make a non-stationary time series stationary — compute the differences between consecutive observations. This is known as **differencing**.
- Transformations such as logarithms can help to stabilise the variance of a time series.
- Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating or reducing trend and seasonality.
- By looking at the time plot of the data, the ACF plot is also useful for identifying non-stationary time series.
- For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.
- Also, for non-stationary data, the value of r_1 is often large and positive.

Differencing

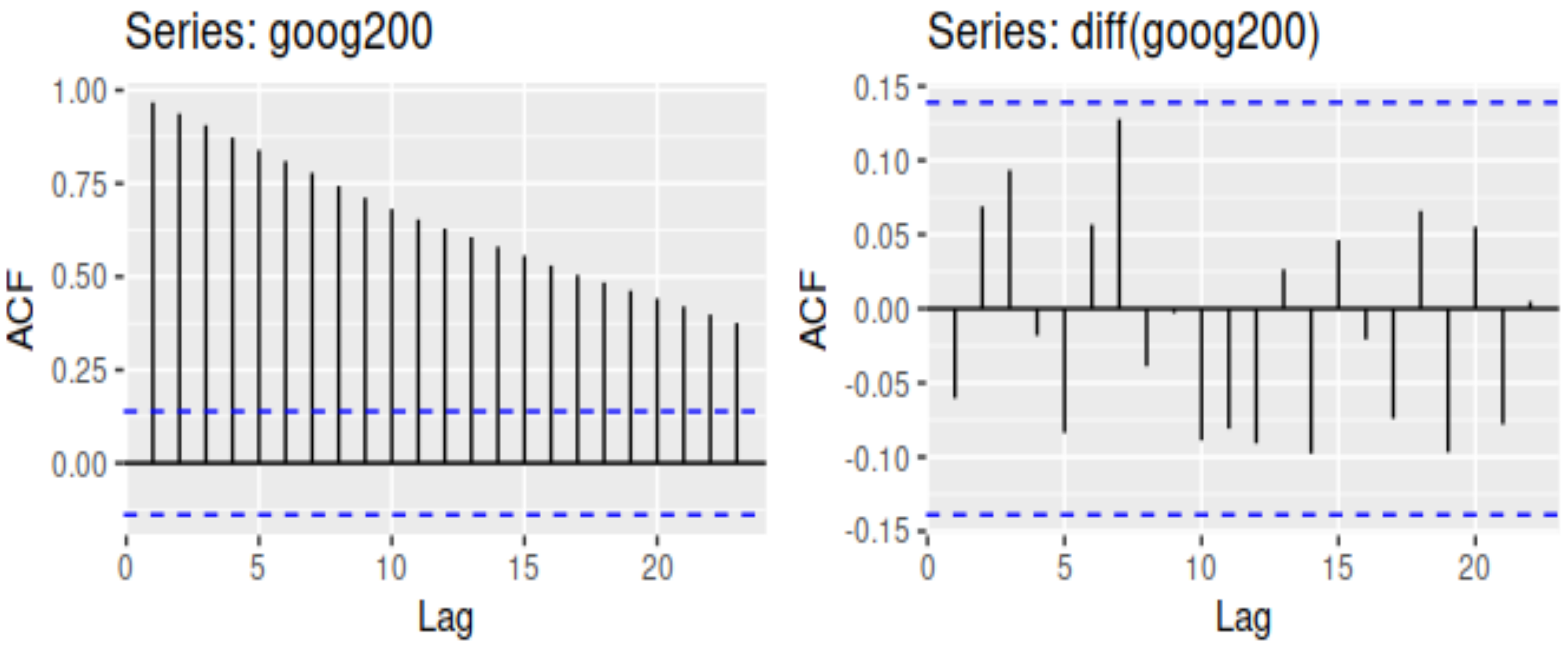


Figure 10:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

Differencing

Figure 8.2:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

- The ACF of the differenced Google stock price looks just like that of a white noise series.
- There are no autocorrelations lying outside the 95% limits, and
- The Ljung -Box Q∗statistic has a *p*-value of 0.355 (for h=10).
- This suggests that the *daily change* in the Google stock price is essentially a random amount which is uncorrelated with that of previous days.

EXAMPLE 13.6, Page No.467

Daily demand for Omelette at Die Another Day (DAD) hospital for the past 115 days is given in the excel sheet Example 13.6.xlsx. Develop an appropriate ARIMA model that DAD hospital can use for forecasting demand for Omelette.

Solution:

The time-series plot of the daily demand for Omelette is shown in Figure 13.15. The corresponding ACF plot is shown in Figure 13.16. From Figure 13.15, it is evident that the mean is not constant for different values of t.

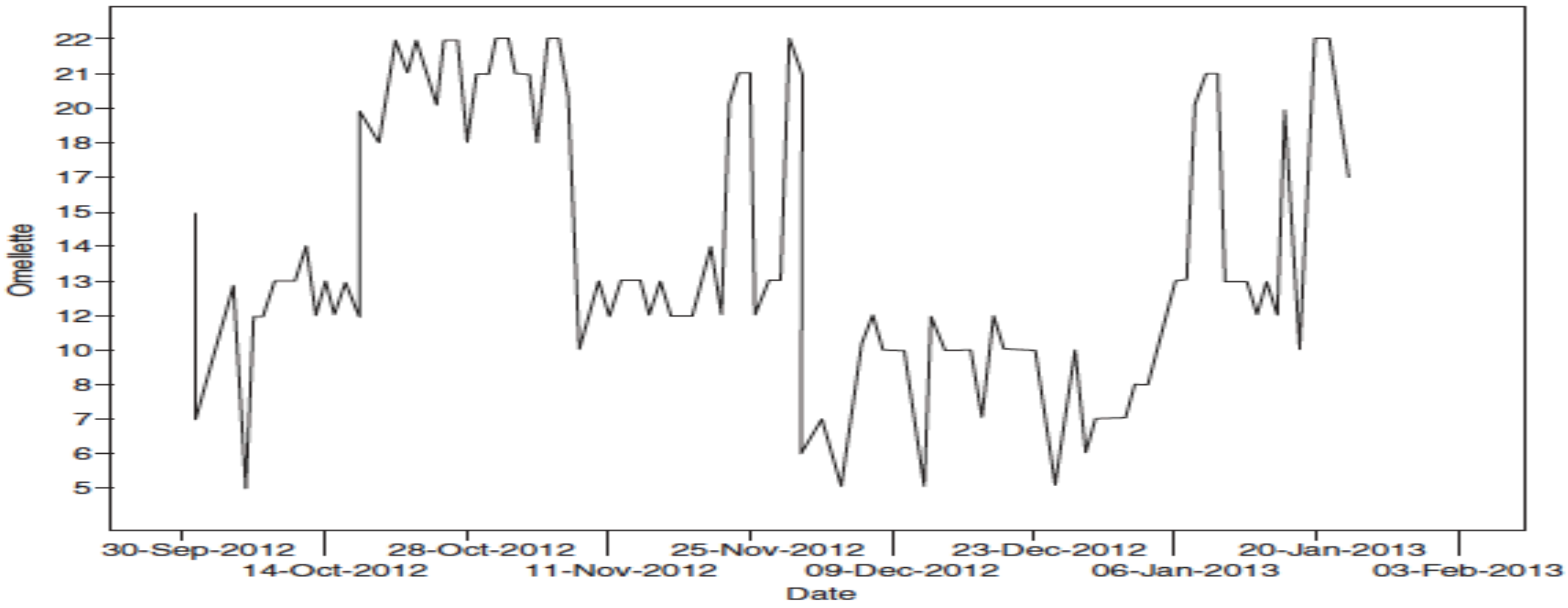


FIGURE 13.15 Time-series plot of demand for Omelette at DAD hospital.

EXAMPLE 13.6, Page No.467

Since the ACF plot shows a very slowly decreasing pattern, we may conclude that the time series is not stationary. We have to convert the process to a stationary process before we can develop a forecasting model. The ACF and PACF plots after differencing (d = 1) are shown in Figures 13.17 and 13.18, respectively.

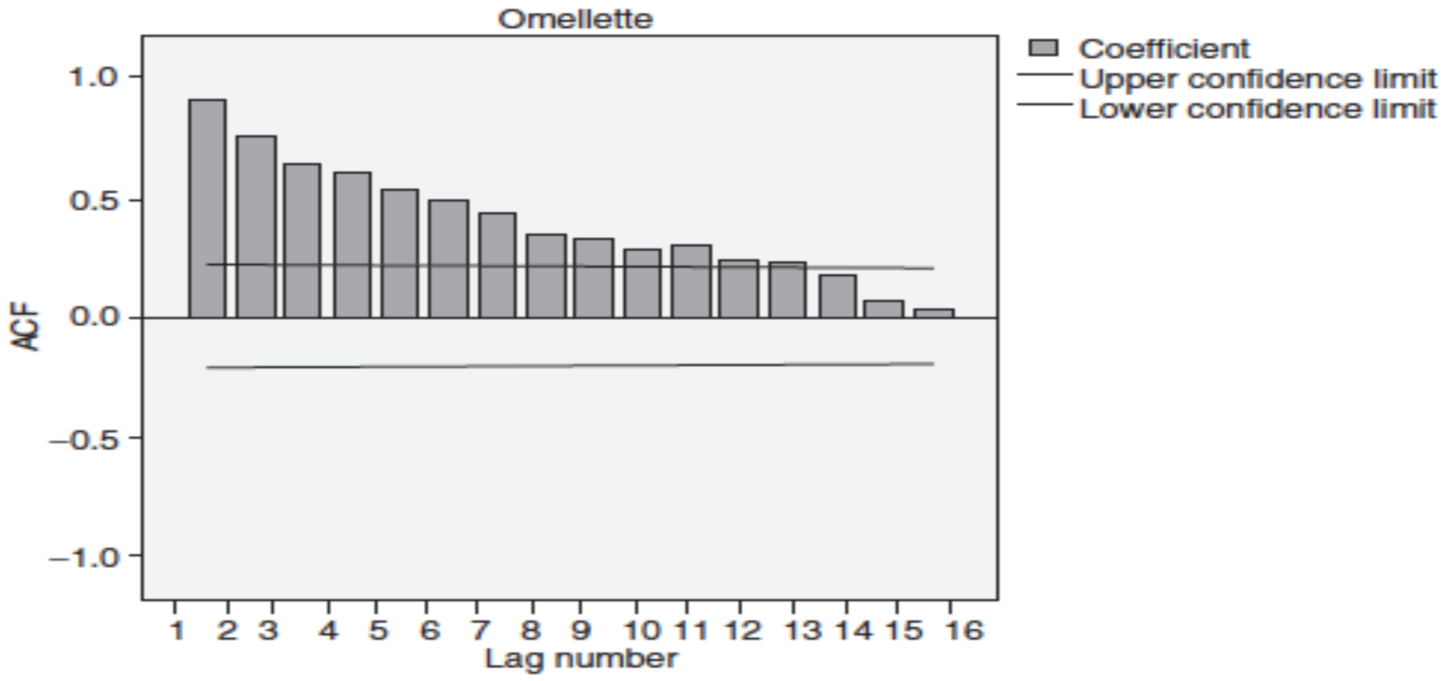


FIGURE 13.16 ACF plot of demand for Omelette at DAD hospital.

EXAMPLE 13.6, Page No.467

Since both ACF and PACF values are cutting off to zero after the first difference, we may conclude that the appropriate model is ARIMA(1,1,1). Note that subsequent correlations once it cuts off to zero is not useful and we will ignore them (for example, in PACF plot in Figure 13.18, the partial auto-correlation value with lag 3 is beyond the critical line).

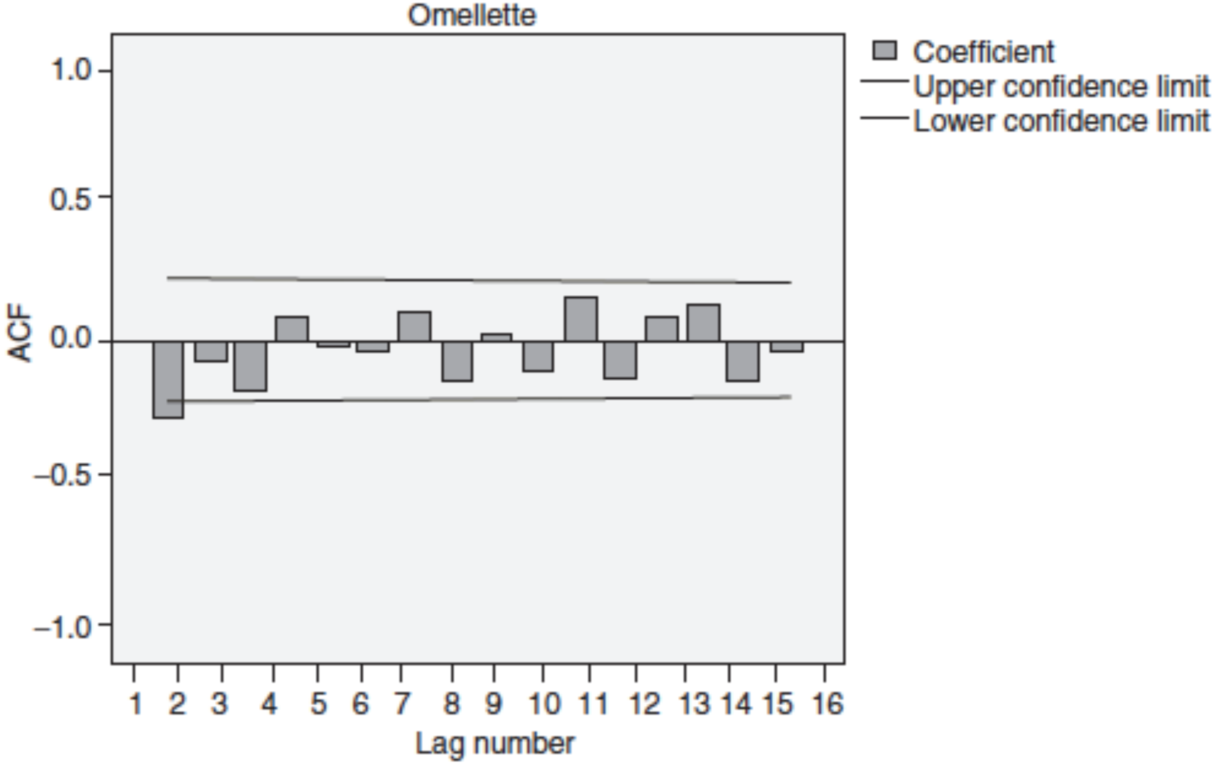


FIGURE 13.17 ACF plot of demand for Omelette after differencing ($d = 1$).

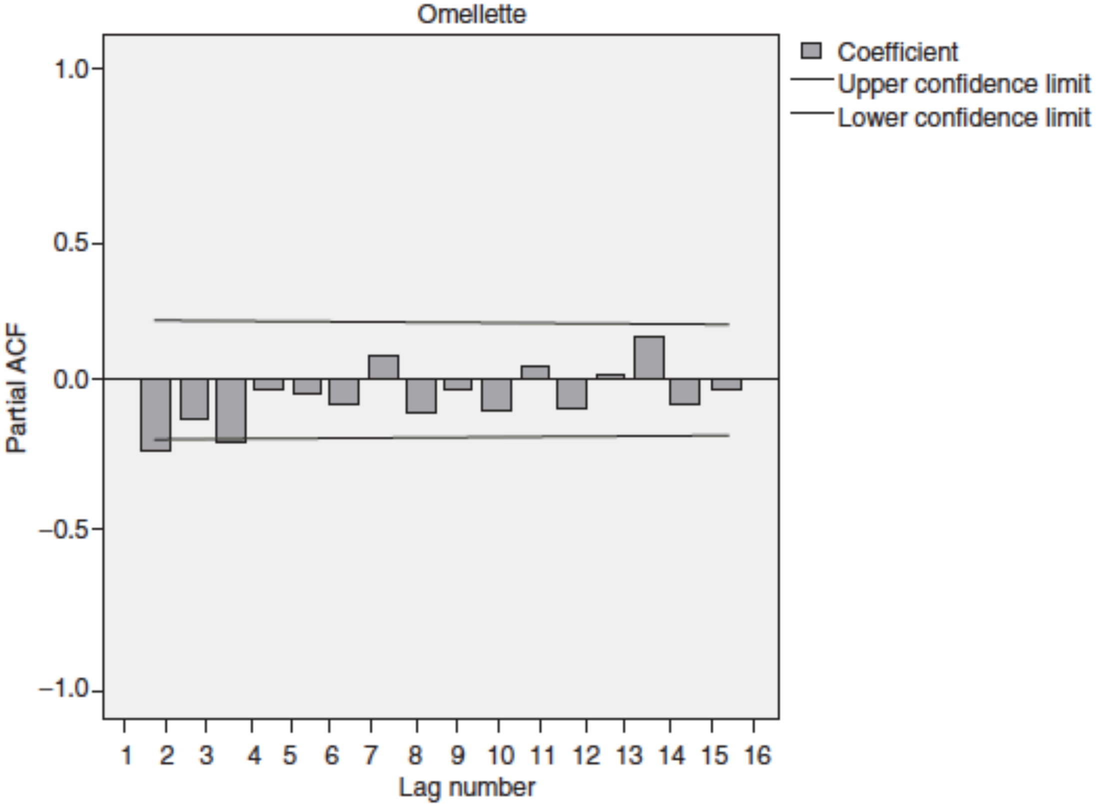


FIGURE 13.18 PACF plot of demand for Omelette after differencing ($d = 1$).

EXAMPLE 13.6, Page No.467

The ARIMA(1, 1, 1) model summary and parameter estimates are shown in Tables 13.29 and 13.30.
AR and MA components in Table 13.30 are statistically significant since the corresponding p-values are less than 0.05.

TABLE 13.29 ARIMA(1, 1, 1) model summary for Omelette demand

Model	Model Fit Statistics			Ljung—Box *Q*(18)		
	R-Squared	RMSE	MAPE	Statistics	*Df*	Sig.
Omellette-Model_1	0.584	3.439	20.830	10.216	16	0.855

TABLE 13.30 ARIMA model parameters

			Estimate	*SE*	*T*	Sig.
Omellette-Model_1	Constant		0.055	0.137	0.402	0.689
	AR	Lag 1	0.439	0.178	2.475	0.015
	Difference		1			
	MA	Lag 1	0.767	0.128	6.004	0.000

EXAMPLE 13.6, Page No.467

The ACF and PACF of residuals are shown in Figure 13.19 which shows white noise of residuals.

Since the residuals follow white noise, we can use ARIMA(1, 1, 1) model for forecasting.

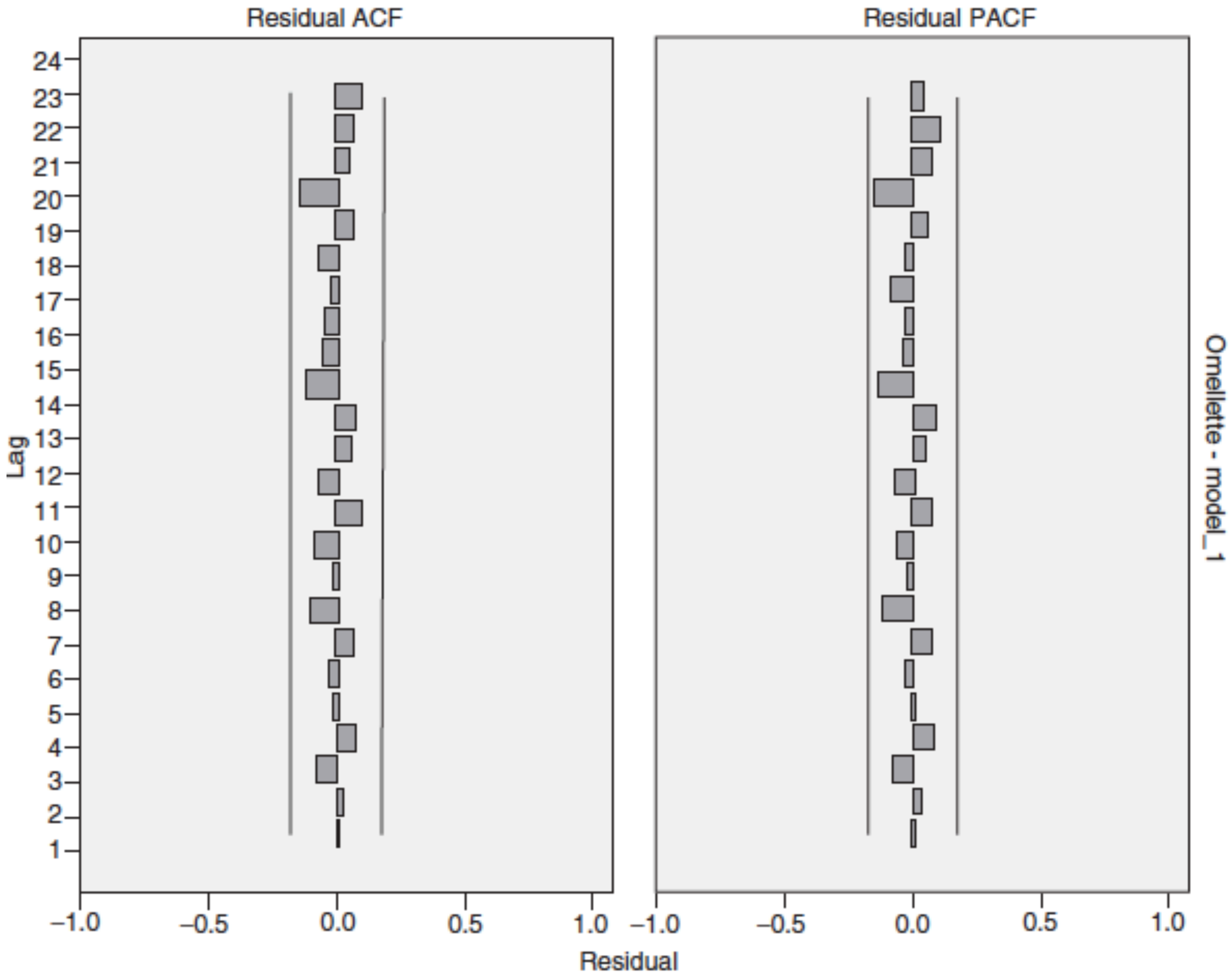


FIGURE 13.19 ACF and PACF of residuals.

Ljung-Box Test for Auto-Correlations

- Ljung−Box is a test of lack of fit of the forecasting model and checks whether the auto-correlations for the errors are different from zero.
- The null and alternative hypotheses are given by

  H_0: The model does not show lack of fit

  H_0: The model exhibits lack of fit

Ljung-Box Test for Auto-Correlations

- The Ljung−Box statistic (*Q*-Statistic) is given by (Ljung and Box, 1978)

$$Q(m) = n(n+2)\sum_{k=1}^{m}\frac{\rho_k^2}{n-k}$$

where *n* is the number of observations in the time series,

  - *k* is the number of lag,
  - ρ_k is the auto-correlation of lag *k*, and
  - *m* is the total number of lags.

Ljung-Box Test for Auto-Correlations

- *Q*-statistic is an approximate chi-square distribution with $m - p - q$ degrees of freedom where p and q are the AR and MA lags.
- The *Q*-statistic for ARIMA(1, 1, 1) is 10.216 (Table 1) and the corresponding *p*-value is 0.855 and thus we fail to reject the null hypothesis.
- Table 1: ARIMA (1, 1, 1) model summary for Omelette demand

Model	Model Fit Statistics			Ljung–Box $Q(18)$		
	R-Squared	RMSE	MAPE	Statistics	*Df*	Sig.
Omellette-Model_1	0.584	3.439	20.830	10.216	16	0.855

- $Q(m)$ measures accumulated auto-correlation up to lag m.

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- The power of forecasting model is a comparison between Naive forecasting model and the model developed.
- In the Naive forecasting model, the forecasted value for the next period is same as the last period's actual value

$$F_{t+1} = Y_t.$$

Theil's coefficient (*U*-statistic) is given by (Theil, 1965),

$$U = \frac{\sum_{t=1}^{n} (Y_{t+1} - F_{t+1})^2}{\sum_{t=1}^{n} (Y_{t+1} - Y_t)^2}$$

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- Theil's coefficient (*U*-statistic) is given by (Theil, 1965)

$$U = \frac{\sum_{t=1}^{n} (Y_{t+1} - F_{t+1})^2}{\sum_{t=1}^{n} (Y_{t+1} - Y_t)^2}$$

- Theil's coefficient is the ratio of the mean squared error of the forecasting model to the MSE of the Naïve model.
- The value of $U < 1$ indicates that forecasting model is better than the Naive forecasting model.
- $U > 1$ indicates that the forecasting model is not better than Naive model.

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- For the data shown in Table 13.24(page 459) (demand for avionic system spares),
- The *U*-statistic calculations are shown in Table 13.31
- TABLE 13.31:U-statistic calculation

Day	Y_t	ARMA (1,2) Forecast	$(Y_t - F_t)^2$	Naïve Forecast ($F_{t+1} = Y_t$)	$(Y_t - F_t)^2$
31	503	464.8107	1458.423	443	3600
32	688	378.5341	95769.15	503	34225
33	602	444.6372	24763.04	688	7396
34	629	685.8851	3235.909	602	729
35	823	743.5124	6318.281	629	37636
36	671	630.7183	1622.614	823	23104
37	487	649.3491	26357.22	671	33856
		Total	159524.6	Total	140546

The U-statistic value = 159524.6 / 140546 = 1.1350. That is, ARMA(1, 2) model is not better than Naive forecasting.

Practice Quiz

1. Seasonality in time-series data is caused due to

(a) Changes in macro-economic factors such as recession, unemployment, and so on

(b) Festivals and customs in a society

(c) Random events that occur over a period of time

(d) Changes in customer behaviour driven by new products and promotions

Practice Quiz

2. In a simple exponential smoothing method, the low value of smoothing constant a is chosen when

(a) The data has high fluctuations around the trend line

(b) There is seasonality in the data

(c) The data is smooth with low fluctuations

(d) There are variations in the data due to cyclical component

Practice Quiz

3. White noise is

(a) Uncorrelated errors with expected value 0.

(b) Uncorrelated errors that are constant and do not change with time.

(c) Uncorrelated errors that follow normal distribution with mean 0 and constant standard deviation

(d) Errors that follow normal distribution with constant mean and standard deviation

Practice Quiz

4. A stationary process in a time series is a process for which

(a) Mean and variance are constant at different time points

(b) The time series follows normal distribution with zero mean and constant standard deviation

(c) The covariance of the time series depends only on the lag

(d) Mean and standard deviation are constant at different time points and the covariance depends only on the lag between the values and is constant for a given lag

Exercise 1, PageNo. 473

Quarterly demand for certain parts manufactured by Jack and Jill company is shown in Table 13.32.

(a) Calculate the seasonality index for different quarters using the first 3 years of data.

(b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model, $Y_t = T_t * S_t$).

(c) Forecast the demand for 2015 (all four quarters) using moving average, exponential smoothing, and ARMA. Calculate RMSE, MAPE, and Theil's coefficient.

TABLE 13.32 Quarterly demand

Year	Quarter	Value
2012	Q1	75
	Q2	60
	Q3	54
	Q4	59
2013	Q1	86
	Q2	65
	Q3	63
	Q4	80
2014	Q1	90
	Q2	72
	Q3	66
	Q4	85
2015	Q1	100
	Q2	78
	Q3	72
	Q4	93

Exercise 1, PageNo. 473

Quarterly demand for certain parts manufactured by Jack and Jill company is shown in Table 13.32.

(a) Calculate the seasonality index for different quarters using the first 3 years of data.

Solution

Quarter	2012	2013	2014	Average	Seasonality Index
Q1	75	86	90	83.66667	1.174269006
Q2	60	65	72	65.66667	0.921637427
Q3	54	63	66	61	0.856140351
Q4	59	80	85	74.66667	1.047953216
			average	71.25	

TABLE 13.32 Quarterly demand

Year	Quarter	Value
2012	Q1	75
	Q2	60
	Q3	54
	Q4	59
2013	Q1	86
	Q2	65
	Q3	63
	Q4	80
2014	Q1	90
	Q2	72
	Q3	66
	Q4	85
2015	Q1	100
	Q2	78
	Q3	72
	Q4	93

PES UNIVERSITY ONLINE

Exercise 1, PageNo. 473

(b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model, $Y_t = T_t * S_t$).

Solution

Year	Quarter	Value	S.I.	Deseasonalized Value
2012	Q1	75	1.174269	63.86952191
	Q2	60	0.921637	65.10152284
	Q3	54	0.85614	63.07377049
	Q4	59	1.047953	56.30022321
2013	Q1	86	1.174269	73.23705179
	Q2	65	0.921637	70.52664975
	Q3	63	0.85614	73.58606557
	Q4	80	1.047953	76.33928571
2014	Q1	90	1.174269	76.64342629
	Q2	72	0.921637	78.12182741
	Q3	66	0.85614	77.09016393
	Q4	85	1.047953	81.11049107
2015	Q1	100	1.174269	85.15936255
	Q2	78	0.921637	84.6319797
	Q3	72	0.85614	84.09836066
	Q4	93	1.047953	88.74441964

Exercise 1, PageNo. 473

(b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model, $Y_t = T_t * S_t$).

(c) Forecast the demand for 2015 (all four quarters) us ARMA. Calculate RMSE, MAPE, and Theil's coefficient.

Forecasting model using moving average:

Moving average forecast for the year 2015 for Quarter 1 to Quarter 4 is given by

$$F_{t+1} = \frac{1}{4}\sum_{k=t+1-4}^{t} Y_k, \qquad \text{for } t = Q1, \ldots, Q4$$

The forecasted values using 4 period moving average and the corresponding RMSE and MAPE calculations are given in below Table.

Table Simple moving average forecast, RMSE and MAPE calculations

Solution

	Quarter	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t\rvert / Y_t$
2015	Q1	85.15936255	78.24148	47.85714	0.081234584
	Q2	84.6319797	80.37046	18.16054	0.050353524
	Q3	84.09836066	81.998	4.411518	0.024975057
	Q4	88.74441964	83.75005	24.94374	0.056278143

The RMSE using the moving average forecast is given by 4.8830 and the MAPE value is 0.0532 (or 5.32%).

Exercise 1, PageNo. 473

(b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model, $Y_t = T_t * S_t$).

(c) Forecast the demand for 2015 (all four quarters) using moving average, exponential smoothing, and ARMA. Calculate RMSE, MAPE, and Theil's coefficient.

Forecasting model using Single Exponential Smoothing:

In single ES, the forecast at time (*t* + 1) is given by (Winters, 1960)

$$F_{t+1} = \alpha Y_t + (1-\alpha) F_t$$

Exercise 1, PageNo. 473

(b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model, $Y_t = T_t * S_t$).

(c) Forecast the demand for 2015 (all four quarters) using moving average, exponential smoothing, and ARMA. Calculate RMSE, MAPE, and Theil's coefficient.

Solution :

The first step in ARMA model building is the identification of the right value of *p* and *q* using ACF and PACF plots.

Exercise 1, PageNo. 473

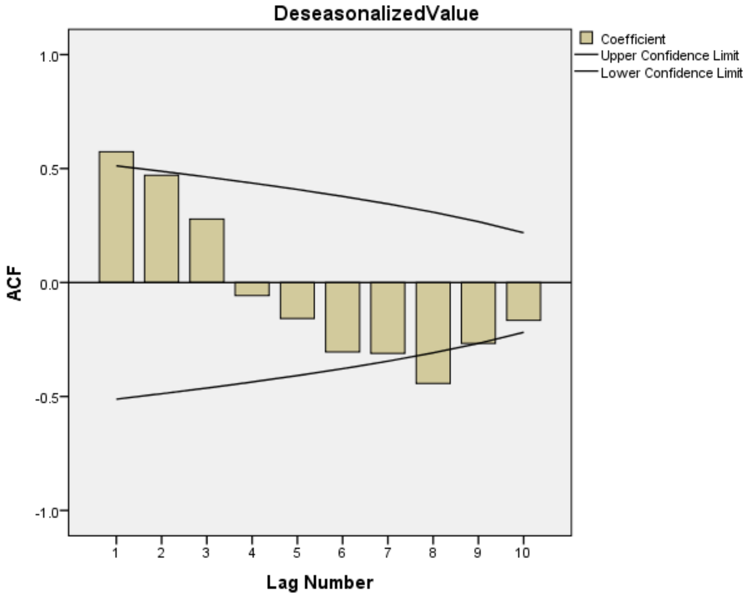
Solution :

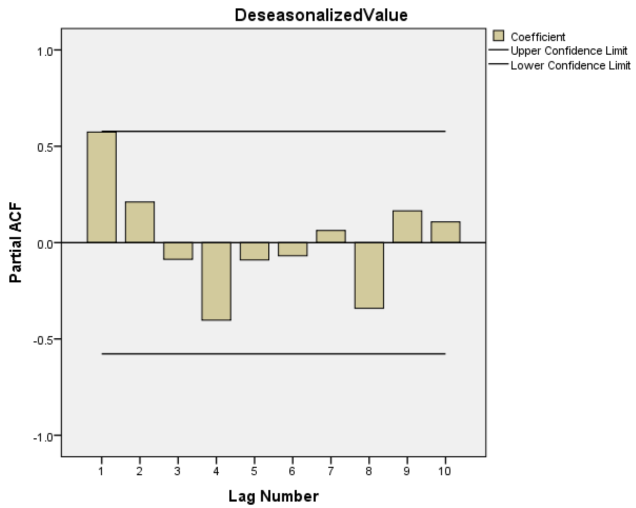
The first step in ARMA model building is the identification of the right value of *p* and *q* using ACF and PACF plots.

ACF and PACF based on the first 12 observations are given in Figures below.

The horizontal lines in the plot represent the critical values for ρ_k and ρ_{pk}.

The correlation values (vertical bars) beyond the critical values will result in rejection of the null hypothesis.





Exercise 1, PageNo. 473

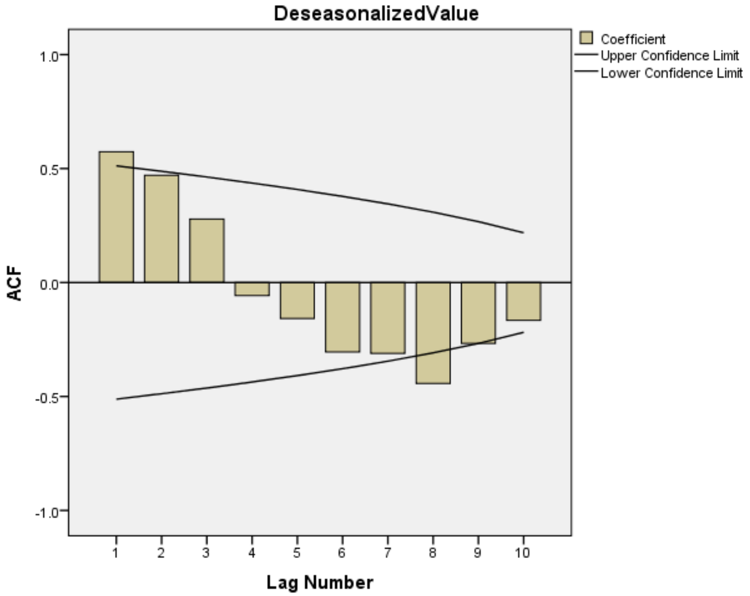
Solution :

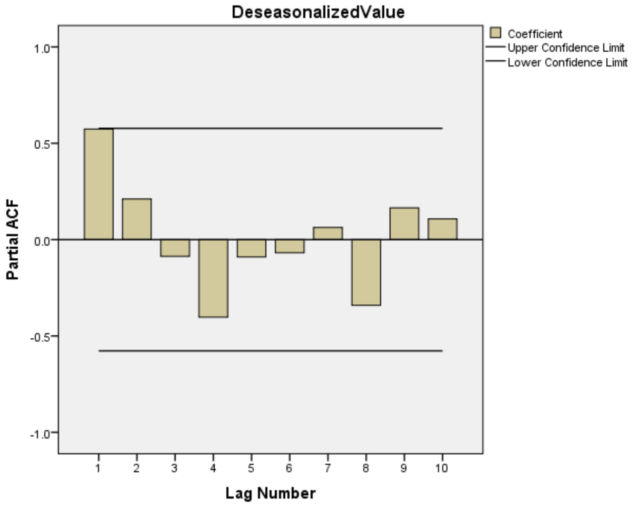
The first step in ARMA model building is the identification of the right value of *p* and *q* using ACF and PACF plots. ACF and PACF based on the first 12 observations are given in Figures below.

The horizontal lines in the plot represent the critical values for ρ_k and ρ_{pk}.

The correlation values (vertical bars) beyond the critical values will result in rejection of the null hypothesis.

In PACF plot, the PACF values cut-off to zero after lag 1 and in ACF plot, the values of auto-correlations cuts off to zero after 2 lags. Thus, we can conclude that the value of *p* and *q* in this case is *p* =1 and *q*=2.





Exercise 1, PageNo. 473

Solution :

The values of $R2$, RMSE, MAPE, and regression parameter estimates of ARMA(1,2) process, using SPSS are shown below.

ARMA(1,2) Model Statistics

Model	Model Fit Statistics			
	R-Square	RMSE	MAPE	Normalized BIC
Manufactured_Parts-Model_1	0.482	5.430	5.145	3.591

Exercise 2, PageNo. 474

Solution :
The values of $R2$, RMSE, MAPE, and regression parameter estimates of ARMA(1,2) process, using SPSS are shown below.
ARMA(1,2) Model Statistics

Model	Model Fit Statistics			
	R-Square	RMSE	MAPE	Normalized BIC
Manufactured_Parts-Model_1	0.482	5.430	5.145	3.591

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 Ch. 13.14.5 and 13.15

Image Courtesy

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu