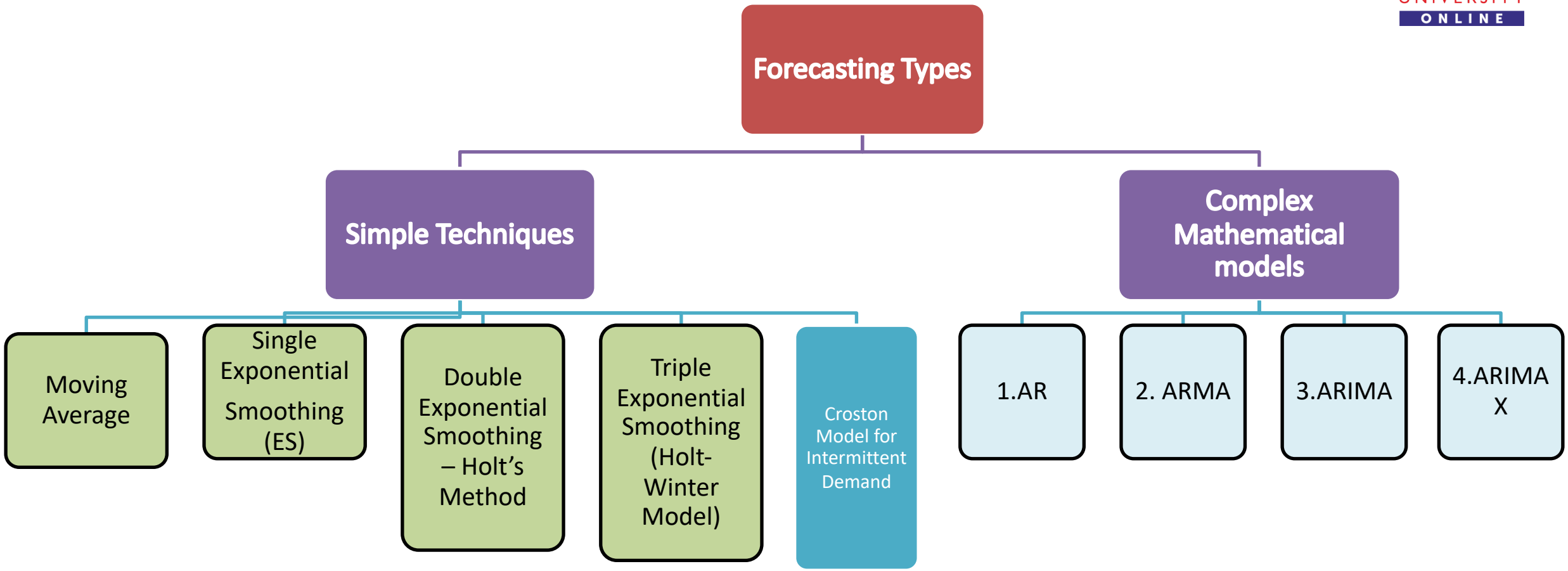
DATA ANALYTICS

Unit 3: Forecasting with Regression, Stationary Signals,ACF,PACF, Correlogram. and ARMA

Jyothi R., Gowri Srinivasa
Department of Computer Science and Engineering

Regression for Forecasting



Regression for Forecasting

- Parker and Segura (1971) claimed regression can predict more accurately than exponential smoothing

- Regression is particularly useful when there is one or more explanatory variable in addition to the dependent variable Y_t

The forecast value at time *t* can be written as

$$F_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt} + \varepsilon_t$$

Here F_t is the forecasted value of Y_t, and X_{1t}, X_{2t}, etc. are the predictor variables measured at time *t*.

Forecasting with Regression – An Example

Forecast the demand of Kesh for months 37 to 48. Calculate the values of MAPE and RMSE.

TABLE 13.1 Data on sales of shampoo, promotion expenses (in 1000 of rupees), and dummy variable for promotion by competition

Month	Sale Quantity	Promotion Expenses	Competition Promotion	Month	Sale Quantity	Promotion Expenses	Competition Promotion
1	3002666	105	1	25	4634047	165	0
2	4401553	145	0	26	3772879	129	1
3	3205279	118	1	27	3187110	120	1
4	4245349	130	0	28	3093683	112	1
5	3001940	98	1	29	4557363	162	0

Table continued in the next slide

Forecasting with Regression – An Example

TABLE 13.1 Data on sales of shampoo, promotion expenses (in 1000 of rupees), and dummy variable for promotion by competition——Continued

Month	Sale Quantity	Promotion Expenses	Competition Promotion	Month	Sale Quantity	Promotion Expenses	Competition Promotion
6	4377766	156	0	30	3816956	140	1
7	2798343	98	1	31	4410887	160	0
8	4303668	144	0	32	3694713	139	0
9	2958185	112	1	33	3822669	141	1
10	3623386	120	0	34	3689286	136	0
11	3279115	125	0	35	3728654	130	1
12	2843766	102	1	36	4732677	168	0
13	4447581	160	0	37	3216483	121	1
14	3675305	130	0	38	3453239	128	0
15	3477156	130	0	39	5431651	170	0
16	3720794	140	0	40	4241851	160	0
17	3834086	167	1	41	3909887	151	1
18	3888913	148	1	42	3216438	120	1
19	3871342	150	1	43	4222005	152	0
20	3679862	129	0	44	3621034	125	0
21	3358242	120	0	45	5162201	170	0
22	3361488	122	0	46	4627177	160	0
23	3670362	135	0	47	4623945	168	0
24	3123966	110	1	48	4599368	166	0

Forecasting with Regression – An Example

$$F_t = \beta_0 + \beta_1 \text{ promotion expenses at time } t + \beta_2 \text{ competition promotion at time } t$$

Model	*R*	*R*-Square	Adjusted *R*-Square	Std. Error of the Estimate	Durbin–Watson
1	0.928	0.862	0.853	207017.359	1.608

Note

- We need a high R^2 value for forecasting applications
- Durbin-Watson Statistic D = 1.608
  Recall: D=2 $\Rightarrow$ autocorrelation; 1.608 $\Rightarrow$ no autocorrelation among the errors
- The presence of autocorrelation may lead to the inclusion of nonsignificant variables in the equation (since the standard error of the regression coefficient is underestimated when autocorrelation errors are present)

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
		B	Std. Error	Beta		
1	(Constant)	808471.843	278944.970		2.898	0.007
	Promotion Expenses	22432.941	1953.674	0.825	11.482	0.000
	Competition Promotion	−212646.036	77012.289	−0.198	−2.761	0.009

Forecasting with Regression – An Example

$$F_t = 808471.843 + 22432.941X_{1t} - 212646.036X_{2t}$$

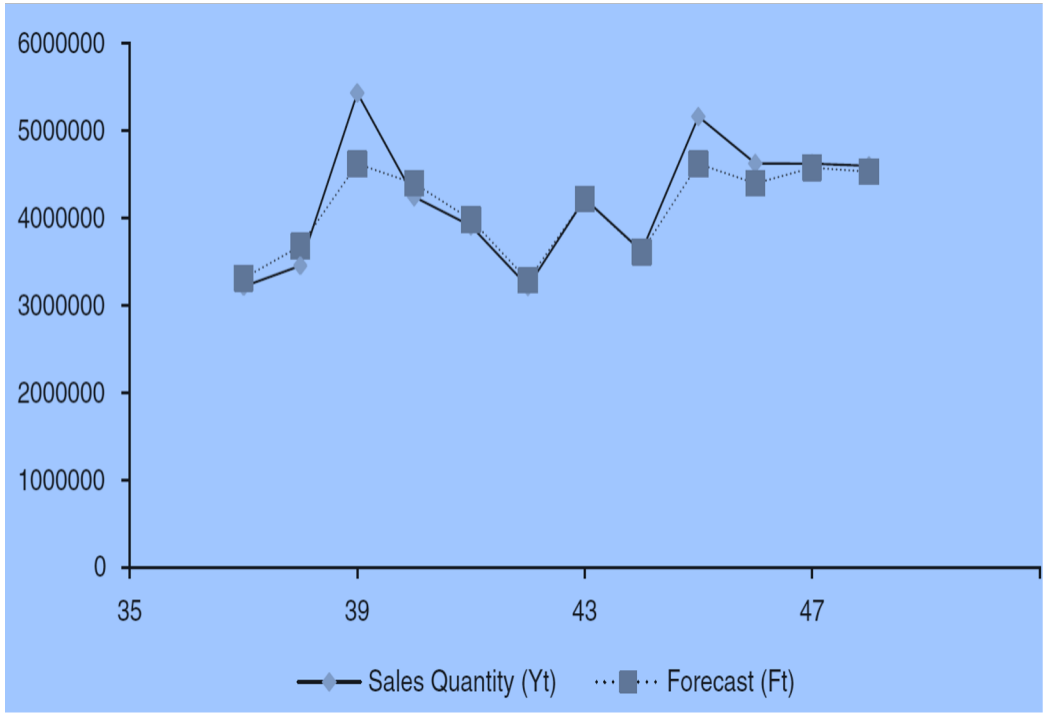
X_{1t} = Promotion expenses at time t

$$X_{2t} = \begin{cases} 1 & \text{Competition is on promotion} \\ 0 & \text{Otherwise} \end{cases}$$

- Sales increases when promotions expenses increase and the sales decrease when the competition is on the promotion.

Method	MAPE	RMSE
Moving Average	734725.84	14.03%
Exponential Smoothing	742339.22	13.94%
Regression	302969	**4.19%**

The plot of actual demand and forecasted demand using regression model



Forecasting with Regression – Seasonality

One can expect seasonal variation in demand for many products and services.
The following steps are used to forecast time-series data with seasonal variations:

STEP 1

Estimate the seasonality index (using techniques such as method of averages or ratio to moving average).

STEP 2

De-seasonalize the data using either additive or multiplicative model. For example, in multiplicative model, the de-seasonalized data $Y_{d,t} = Y_t / S_t$, where $Y_{d,t}$ is the de-seasonalized data and S_t is the seasonality index for period t.

STEP 3

Develop a forecasting model on the de-seasonalized data ($F_{d,t}$).

STEP 4

The forecast for period $t + 1$ is $F_{t+1} = F_{d,t+1} \times S_{t+1}$.

REGRESSION MODEL FOR FORECASTING Example 13.3, page 446

Hiccup Viking (HV) is The Vice President of Viking Cookies that specialized into chocolate chip cookies (Choco-Chip). Viking Cookies believes that demand for cookies is seasonal and is driven by several factors such as school holidays, festivals, etc. The shelf life of Choco-Chip cookies is 6 months and excess inventory and running out of stock can have financial impact.

Hiccup would like to develop a forecasting model that they can use for forecasting the demand. The past monthly demand (quantity of 200 gram packets) for four years (January 2013 to December 2016) along with average price per unit during that month is shown in Table 13.14.

Develop a forecasting model using regression to predict demand between months 37 and 48, given that the data is seasonal.

REGRESSION MODEL FOR FORECASTING Example 13.3

TABLE 13.14 Monthly demand (quantity of 200 gram packets) along with average price per unit

Period	Month	Demand in Units	Average Price	Period	Demand in Units	Average Price
1	January	10500472	37	25	10658309	36
2	February	10123572	34	26	8677622	38
3	March	7372141	36	27	7330354	37
4	April	7764303	38	28	8115471	37
5	May	6904463	40	29	8481936	34
6	June	10068862	34	30	8778999	37
7	July	6436190	40	31	10145039	32
8	August	9898436	34	32	8497839	38
9	September	6803825	39	33	8792138	34
10	October	8333787	36	34	8485358	36
11	November	7541964	39	35	8575904	36
12	December	8540662	37	36	9885156	32
13	January	10229437	37	37	11023467	35
14	February	8453201	38	38	7942451	40
15	March	7997459	35	39	12492798	32
16	April	8557825	35	40	9756258	32
17	May	7818397	36	41	8992741	32
18	June	8944499	37	42	7397807	40
19	July	8904086	36	43	9710611	32
20	August	8463682	39	44	8328379	39
21	September	7723957	37	45	11873063	32
22	October	7731422	39	46	10642507	32
23	November	8441834	35	47	10635075	32

REGRESSION MODEL FOR FORECASTING Example

Solution:

Since the demand is seasonal, the first step in forecasting is to estimate the seasonality index.

We can use first 36 months data to estimate the seasonality index using method of averages explained in Section 13.7.1.

Table 13.15 gives the seasonality index for various months.

For example, the seasonality index for January is 1.2251.

That is, in January the demand will increase by 22.51% from the trend.

REGRESSION MODEL FOR FORECASTING Example

REGRESSION MODEL FOR FORECASTING Example

Step 1 : Estimate the seasonality index

TABLE 13.15 Seasonality index for various months

Month	Demand (2012)	Demand (2013)	Demand (2014)	Average	Seasonality Index
1	10500472	10229437	10658309	10462739	1.2251
2	10123572	8453201	8677622	9084798	1.0637
3	7372141	7997459	7330354	7566651	0.8860
4	7764303	8557825	8115471	8145866	0.9538
5	6904463	7818397	8481936	7734932	0.9057
6	10068862	8944499	8778999	9264120	1.0847
7	6436190	8904086	10145039	8495105	0.9947
8	9898436	8463682	8497839	8953319	1.0483
9	6803825	7723957	8792138	7773307	0.9102
10	8333787	7731422	8485358	8183522	0.9582
11	7541964	8441834	8575904	8186567	0.9585
12	8540662	7485122	9885156	8636980	1.0113
Average of monthly averages				8540659	

REGRESSION MODEL FOR FORECASTING Example

Step 2 : De-seasonalize the data using either additive or multiplicative model.

- De-seasonalized data is calculated by dividing the value of Y_t with the corresponding seasonality index. The de-seasonalized data for periods 1 to 48 is shown in Table 6:
- TABLE 13.16 De-seasonalized demand seasonality index is rounded to 2 decimals

Month	Demand	Seasonality Index	De-seasonalized Demand	Month	Demand	Seasonality Index	De-seasonalized Demand
1	10500472	1.23	8571459.88	25	10658309	1.23	8700301.09
2	10123572	1.06	9517214.68	26	8677622	1.06	8157870.71
3	7372141	0.89	8321110.54	27	7330354	0.89	8273944.56
4	7764303	0.95	8140603.02	28	8115471	0.95	8508790.52
5	6904463	0.91	7623682.26	29	8481936	0.91	9365476.36
6	10068862	1.08	9282556.42	30	8778999	1.08	8093422.43

Refer page No.449 for the instances for 7 to 24 table

REGRESSION MODEL FOR FORECASTING Example

Step 3 :Develop a forecasting model on the de-seasonalized data ($F_{d,t}$).

- Regression output for the de-seasonalized demand and average price using
- Microsoft Excel are shown in Table 13.17:
- TABLE 13.17 Regression output using SPSS for data in Table 13.16 (based on first 36 cases)

Model		Unstandardized Coefficients		*T*	Sig.
		B	Std. Error		
1	(Constant)	20812014.673	717702.417	28.998	0.000
	Average Price	−335945.859	19616.915	−17.125	0.000

REGRESSION MODEL FOR FORECASTING Example

Step 4 :The forecast for period t + 1 is $F_{t+1} = F_{d,\,t+1} * S_{t+1}$

- Regression model for demand forecasting based on first 36 months of de-seasonalized data is given by
- $F_{d,t}$ = 20812014.673 – 335945.859 × Average Price
- The forecasted values are given in Table 13.18.
- TABLE 13.18 Forecasted values for the data in Table 13.16

REGRESSION MODEL FOR FORECASTING Example

- TABLE 13.18 Forecasted values for the data in Table 13.16

Month	Demand	Seasonality Index (S_t)	De-seasonalized Demand	F_{dt}	$F_t = F_{dt} * S_t$	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
37	11023467	1.2251	8998377	9053910	11091497	4628131462	0.006171
38	7942451	1.0637	7466733	7374180	7844001	9692313467	0.012395
39	12492798	0.8860	14100918	10061747	8914269	1.2806×10^{13}	0.286447
40	9756258	0.9538	10229099	10061747	9596642	2.5477×10^{10}	0.01636
41	8992741	0.9057	9929491	10061747	9112521	1.4347×10^{10}	0.01332
42	7397807	1.0847	6820092	7374180	7998831	3.6123×10^{11}	0.081244
43	9710611	0.9947	9762683	10061747	10008080	8.8488×10^{10}	0.030633
44	8328379	1.0483	7944523	7710126	8082657	6.0379×10^{10}	0.029504
45	11873063	0.9102	13045128	10061747	9157730	7.373×10^{12}	0.228697
46	10642507	0.9582	11106956	10061747	9641005	1.003×10^{12}	0.094104
47	10635075	0.9585	11095071	10061747	9644592	9.8106×10^{11}	0.093134
48	10578547	1.0113	10460573	10061747	10175223	1.6267×10^{11}	0.038127

RMSE and MAPE values are 1381119.09 and 0.0775 (7.75%), respectively.

Autoregressive Models

Auto-regression simply means regression of a variable on itself measured at different time periods. One of the fundamental assumptions of AR model is that the time series is assumed to be a stationary process.

If a time-series data, Y_t, is stationary, then it satisfies the following conditions:
1. The mean values of Y_t at different values of t are constant.
2. The variances of Y_t at different time periods are constant (Homoscedasticity).
3. The covariances of Y_t and Y_{t-k} for different lags depend only on k and not on time t

When the time series data is not stationary (that is, any one of the above conditions are not satisfied), then we have to convert the non-stationary times-series data to stationary data before applying AR models

Another important concept associated with forecasting based on regression-based models is the white noise of residuals. White noise is a process of residuals that are uncorrelated and follow normal distribution with mean 0 and constant standard deviation. In AR models, one of the important assumptions that we make is that the errors follow a white noise.

REGRESSION MODEL FOR FORECASTING – New Terminologies

1. Time lags
2. Correlation over time (*serial correlation*, a.k.a. *autocorrelation*)
3. White noise
4. Stationarity
5. **Autocorrelation Function (ACF)**
6. Partial Autocorrelation Function (PACF)
7. **Correlogram :** A **correlogram** (also called Auto Correlation Function ACF Plot or Autocorrelation plot) is a visual way to show serial correlation in data that changes over time (i.e. time series data).
8.

Time lags and Auto correlation(Revisit to the chapter 10- MLR, 10.15 Auto correlation, page 298)

- Auto-correlation is the correlation between successive error terms in a time-series data.
- Consider a timeseries model as defined below:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- In the regression model [Eq.above)], the values of the response variable Y are measured at different time points t and X_t is the value of the independent variable at time t.

- One of the assumptions of regression model is that, there should be no correlation between error terms, e_t and e_{t-1} (known as auto-correlation of errors of lag 1).

- In general, errors e_t and e_{t-k} may be correlated (known as autocorrelation of lag k).

- If there is an auto-correlation, the standard error estimate of the beta coefficient may be underestimated and that will result in overestimation of the t-statistic value, which, in turn, will result in a low p-value.

- Thus, a variable which has no statistically significant relationship with the response variable may be accepted in the model due to the presence of auto-correlation. The presence of auto-correlation can be established using Durbin–Watson test.

Time lags and Auto correlation

Lags (Lag Operator)

The lag operator (also known as backshift operator) is a function that shifts (offsets) a time series such that the "lagged" values are aligned with the actual time series.

The lags can be shifted any number of units, which simply controls the length of the backshift.

The picture below illustrates the lag operation for lags 1 and 2.

Date	Value	$Value_{t-1}$	$Value_{t-2}$
1/1/2017	200	NA	NA
1/2/2017	220	200	NA
1/3/2017	215	220	200
1/4/2017	230	215	220
1/5/2017	235	230	215
1/6/2017	225	235	230
1/7/2017	220	225	235
1/8/2017	225	220	225
1/9/2017	240	225	220
1/10/2017	245	240	225

Lag and autocorrelation analysis is a good way to detect seasonality. We used the autocorrelation of the lagged values to **detect "abnormal" seasonal patterns**.

DATA ANALYTICS

Time lags and Auto correlation

Date	Value	$Value_{t-1}$	$Value_{t-2}$
1/1/2017	200	NA	NA
1/2/2017	220	200	NA
1/3/2017	215	220	200
1/4/2017	230	215	220
1/5/2017	235	230	215
1/6/2017	225	235	230
1/7/2017	220	225	235
1/8/2017	225	220	225
1/9/2017	240	225	220
1/10/2017	245	240	225

Autocorrelation called serial correlation measures correlation between neighbours

Correlation between pairs of values at a certain lag.
Lag-1 autocorrelation: between y_t and y_{t-1}
Lag-2 autocorrelation: between y_t and y_{t-2}

Time lags and Auto correlation

Date	Value	$Value_{t-1}$	$Value_{t-2}$
1/1/2017	200	NA	NA
1/2/2017	220	200	NA
1/3/2017	215	220	200
1/4/2017	230	215	220
1/5/2017	235	230	215
1/6/2017	225	235	230
1/7/2017	220	225	235
1/8/2017	225	220	225
1/9/2017	240	225	220
1/10/2017	245	240	225

Lags are very useful in time series analysis because of a phenomenon called autocorrelation, which is a tendency for the values within a time series to be correlated with previous copies of itself.

One benefit to autocorrelation is that we can **identify patterns within the time series**, which helps in determining seasonality.

Autoregression is the basis for one of the most widely used forecasting techniques, the autoregressive integrated moving average model or **ARIMA** for short.

Stationarity

- A strictly stationary process is one where the distribution of its values remains the same as time proceeds, implying that the probability lies in a particular interval is the same now as at any point in the past or the future.

- However we tend to use the criteria relating to a 'weakly stationary process' to determine if a series is stationary or not.

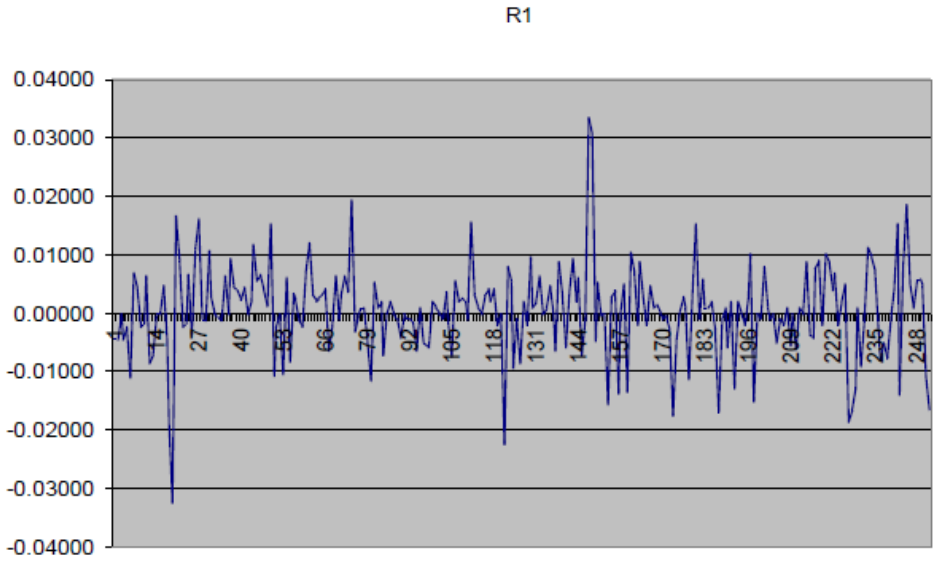
Weakly Stationary Series

- A stationary process or series has the following properties:
  - constant mean
  - constant variance
  - constant auto covariance structure

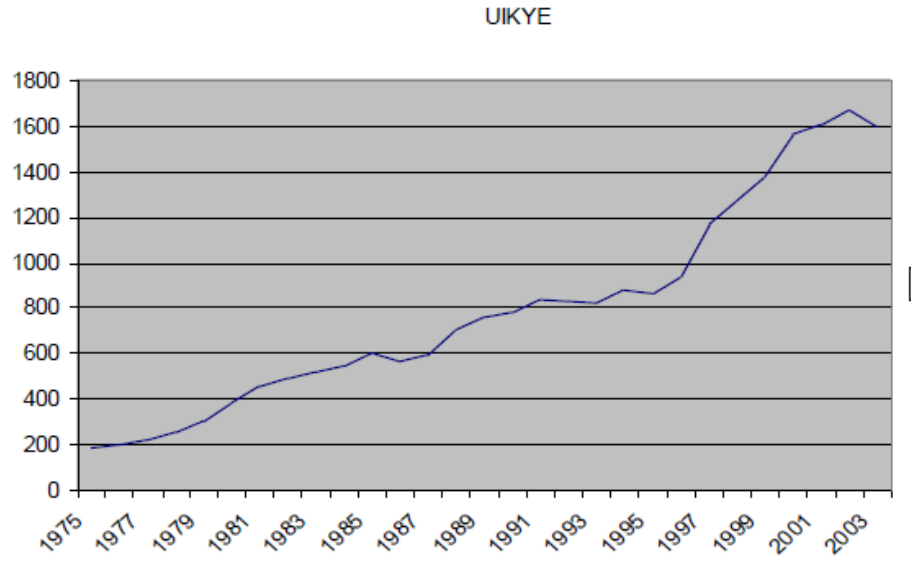
- $E(y_t) = \mu$
- $E(y_t - \mu)^2 = \sigma^2$
- $E(y_{t1} - \mu)(y_{t2} - \mu) = \gamma_{t2-t1}, \forall t_1, t_2$

- The latter refers to the covariance between y(t-1) and y(t-2) being the same as y(t-5) and y(t-6).

Stationary and NonStationary Series



Stationary Series



Non-stationary Series

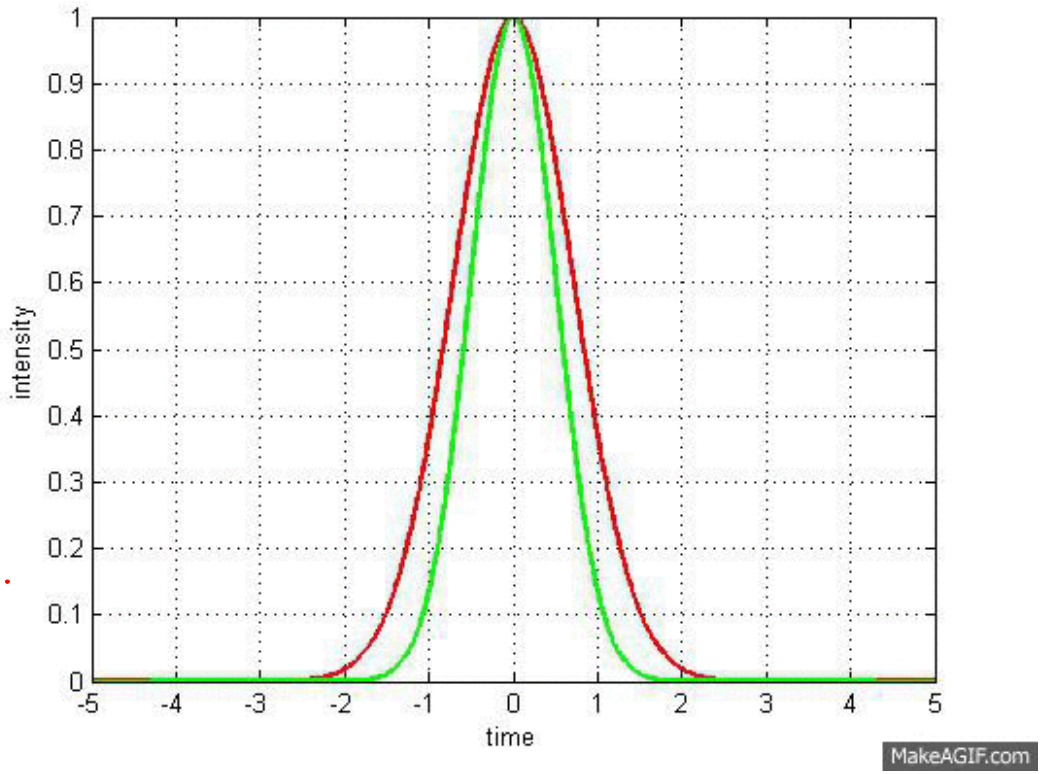
Implications of Nonstationary Data

- If the variables in an OLS regression are not stationary, they tend to produce regressions with high R-squared statistics and low Durbin-Watson statistics, indicating high levels of autocorrelation.
- This is caused by the drift in the variables often being related, but not directly accounted for in the regression, hence the omitted variable effect.
- It is important to determine if our data is stationary before the regression.
- This can be done in a number of ways:
  - plotting the data
  - assessing the autocorrelation function
  - Using a specific test on the significance of the autocorrelation coefficients.
  - Specific tests such as DF, ADF, etc. (to be covered later)

Autocorrelation Function (ACF) at lag *k*

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{covariance at lag k}}{\text{variance}}$$

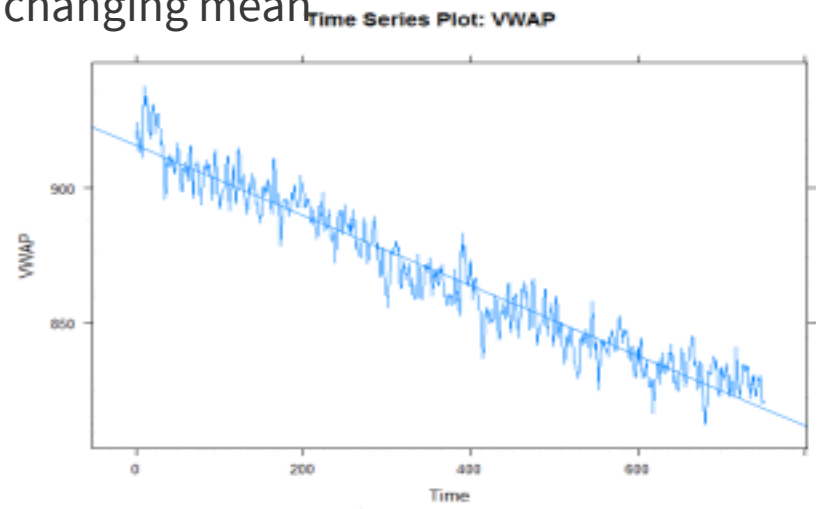
$$\rho_k = \frac{\sum_{t=k+1}^{n}\left(Y_{t-k}-\bar{Y}\right)\left(Y_t-\bar{Y}\right)}{\sum_{t=1}^{n}\left(Y_t-\bar{Y}\right)^2}$$



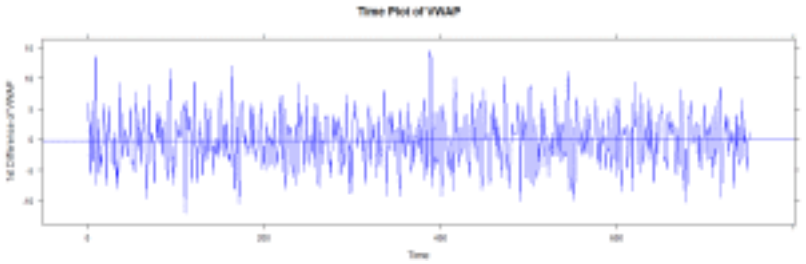
DATA ANALYTICS

The Autocorrelation function is one of the widest used tools in timeseries analysis.
It is used to determine **stationarity** and **seasonality**.

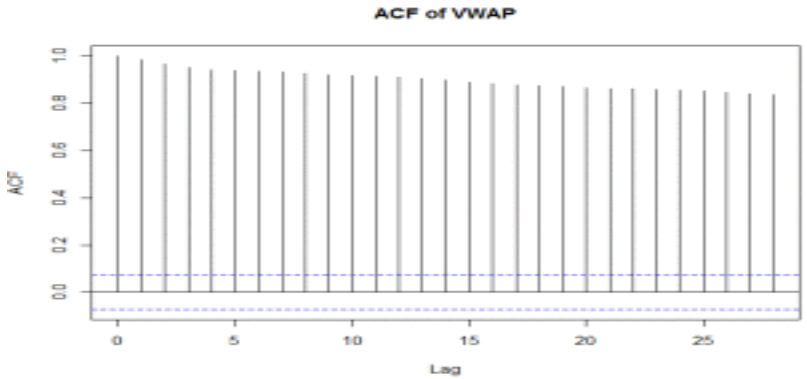
non-stationary series looks like. Note the changing mean



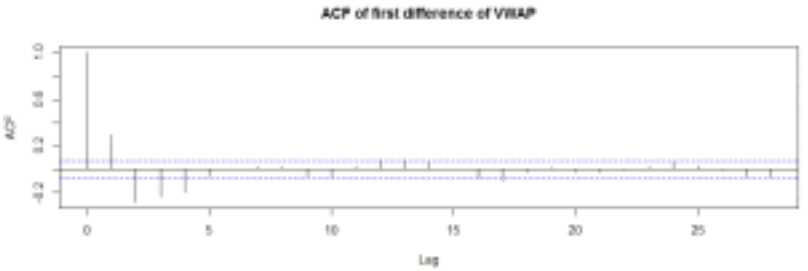
stationary series Note the constant mean (long term)



If a series is non-stationary (moving), its ACF may look a little like this

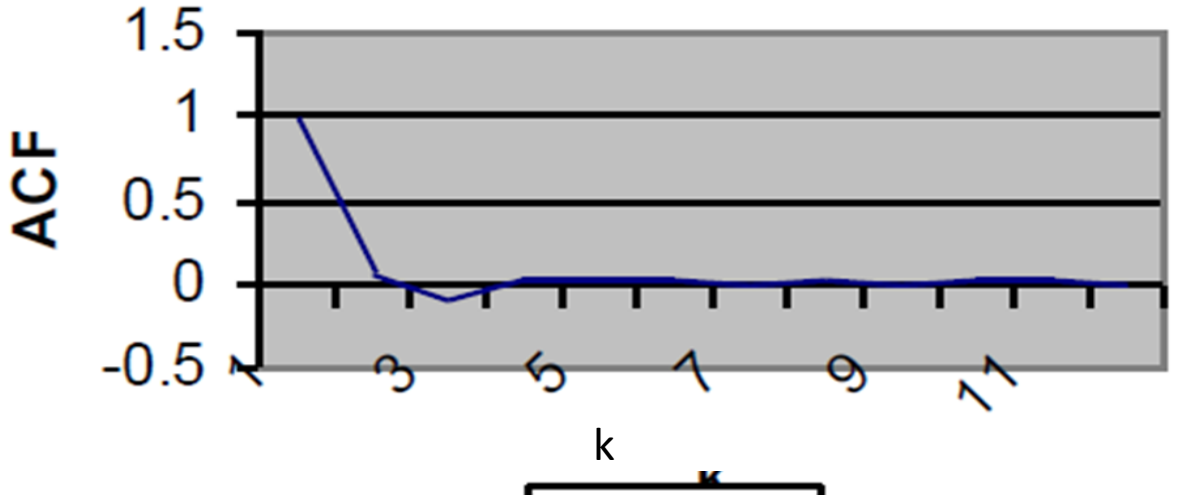


the ACF of a stationary (not going anywhere) series:



Correlogram

- The sample Correlogram is the plot of the ACF against k.
- As the ACF lies between -1 and +1, the Correlogram also lies between these values.



- It can be used to determine stationarity, if the ACF falls immediately from 1 to 0, then equals about 0 thereafter, the series is stationary.
- If the ACF declines gradually from 1 to 0 over a prolonged period of time, then it is not stationary.

Statistical Significance of the ACF

- The Q statistic can be used to determine if the sample ACFs are jointly equal to zero.

$$Q = n\sum_{k=1}^{m}\hat{\rho}_k^2$$

- n -> sample size
- m -> lag length
- $\chi^2(m)$ ->degrees of freedom

- If jointly equal to zero we can conclude that the series is stationary.
- It follows the chi-squared distribution, where the null hypothesis is that the sample ACFs jointly equal zero.

Q-statistic Example

- The following information, from a specific variable can be used to determine if a time series is stationary or not.

$$\sum_{k=1}^{4} \hat{\rho}_k^2 = 0.32$$

$$n = 60$$

$$Q = 60 * 0.32 = 19.2$$

$$\chi^2(4) = 9.488$$

$$19.2 > 9.488 \rightarrow reject - H_0$$

- The series is not stationary as the ACFs are jointly significantly different to 0.

Partial ACF

- The Partial Autocorrelation Function (PACF) is similar to the ACF, however it measures correlation between observations that are *k* time periods apart, after controlling for correlations at intermediate lags.

- First order (i.e., k=1), AC and PAC are the same. For second order (k=2),

$$\frac{\text{Covariance}(y_t, y_{t-2} \mid y_{t-1})}{\sqrt{\text{Variance}(y_t \mid y_{t-1})\text{Variance}(y_{t-2} \mid y_{t-1})}}$$

- This can also be used to produce a partial Correlogram, which is used in Box-Jenkins methodology (covered later).

Autoregressive Process

Auto-regression simply means regression of a variable on itself measured at different time periods.
One of the fundamental assumptions of AR model is that the time series is assumed to be a stationary process.

If a time-series data, Y_t, is stationary, then it satisfies the following conditions:
1. The mean values of Y_t at different values of t are constant.
2. The variances of Y_t at different time periods are constant (Homoscedasticity).
3. The covariances of Y_t and Y_{t-k} for different lags depend only on k and not on time t.

When the time series data is not stationary (that is, any one of the above conditions are not satisfied), then we have to convert the non-stationary times-series data to stationary data before applying AR models.

$\varepsilon \sim N(0, \sigma_\varepsilon^2)$

Another important concept associated with forecasting based on regression-based models is the white noise of residuals. White noise is a process of residuals are uncorrelated and follow normal distribution with mean 0 and constant standard deviation. In AR models, one of the important assumptions that we make is that the errors follow a white noise.

Autoregressive Process

$$Y_{t+1} = \beta Y_t + \varepsilon_{t+1} \quad \text{which can be re-written as} \quad Y_{t+1} - \mu = \beta \times (Y_t - \mu) + \varepsilon_{t+1}$$

$$Y_{t+1} - \mu = \beta \times \left[\beta \times (Y_{t-1} - \mu) + \varepsilon_t \right] + \varepsilon_{t+1}$$

$$Y_{t+1} - \mu = \beta^t (Y_0 - \mu) + \beta^{t-1} \varepsilon_1 + \beta^{t-2} \varepsilon_2 + \cdots + \beta \varepsilon_t + \varepsilon_{t+1}$$

$$Y_{t+1} - \mu = \beta^t (Y_0 - \mu) + \sum_{k=1}^{t-1} \beta^{t-k} \times \varepsilon_k + \varepsilon_{t+1}$$

If $|\beta| > 1$, then $[\beta t\ (Y0 - \mu)]$ will result in infinitely large value of $Yt+1$ as the value of t increases and is not very useful for practical applications. The value of $|\beta| = 1$ would imply that the future value of Y depends on the entire past (and will lead to non-stationarity). For practical applications, the value of $|\beta|$ should be less than one.

The second part of the equation can also become infinitely large if the errors do not follow a white noise. When the errors are white noise then the expected value of sum($\beta_{t-k}\varepsilon_k$) is zero.

Estimating β

$$\sum_{t=2}^{n} \varepsilon_t^2 = \sum_{t=2}^{n} [(Y_t - \mu) - \beta \times (Y_{t-1} - \mu)]^2 \tag{13.34}$$

Taking first-derivative of Eq. (13.34) with respect to β and equating that to zero, the estimate of β is given by

$$\hat{\beta} = \frac{\sum_{t=2}^{n} (Y_t - \mu)(Y_{t-1} - \mu)}{\sum_{t=2}^{n} (Y_{t-1} - \mu)^2} \tag{13.35}$$

AR Model Identification

Autocorrelation : $$\rho_k = \frac{\sum_{t=k+1}^{n}\left(Y_{t-k}-\bar{Y}\right)\left(Y_t-\bar{Y}\right)}{\sum_{t=1}^{n}\left(Y_t-\bar{Y}\right)^2}$$

Partial Autocorrelation: Correlation between Y_t and Y_{t-k} when the influence of all intermediate values is removed from both Y_t and Y_{t-k}

Plots of autocorrelation and partial autocorrelation for different values of k are called the ACF and PACF respectively

$H_0: \rho_k = 0$ and $H_A: \rho_k \neq 0$, where ρ_k is the auto-correlation of order k

$H_0: \rho_{pk} = 0$ and $H_A: \rho_{pk} \neq 0$, where ρ_{pk} is the partial auto-correlation of order k

The null hypothesis is rejected when $|\rho_k| > 1.96/\sqrt{n}$ and $|\rho_{pk}| > 1.96/\sqrt{n}$. To select the appropriate p in the auto-regressive model, the following thumb rule may be used. The number of lags is p when

1. The partial auto-correlation, $|\rho_{pk}| > 1.96/\sqrt{n}$ for first p values (first p lags) and cuts off to zero.
2. The auto-correlation function (ACF), ρ_k, decreases exponentially.

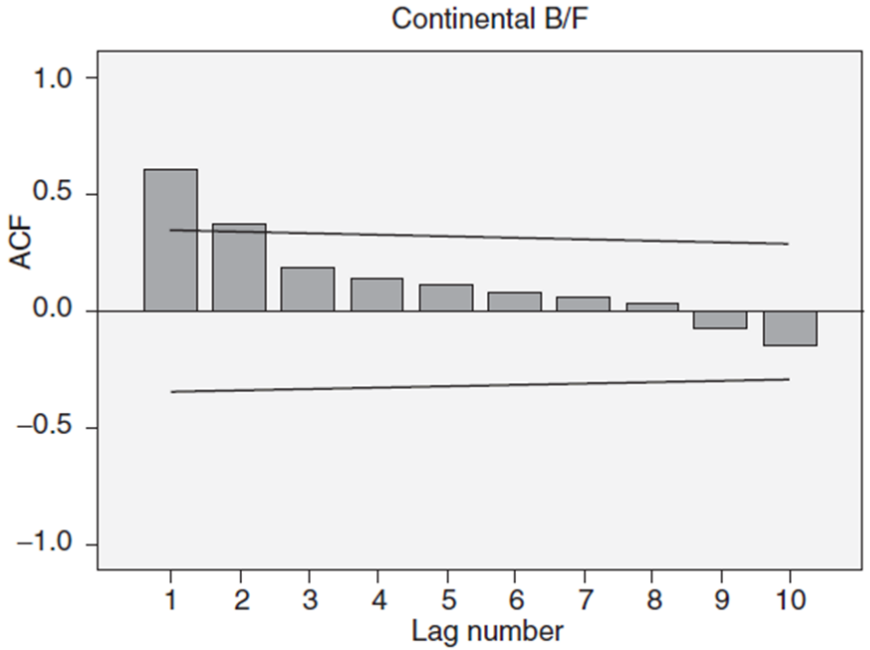
AR Model - Example

Build an auto-regressive model based on the first 30 days of data and forecast the demand for continental breakfast on days 31 to 37. Comment on the accuracy of the forecast.

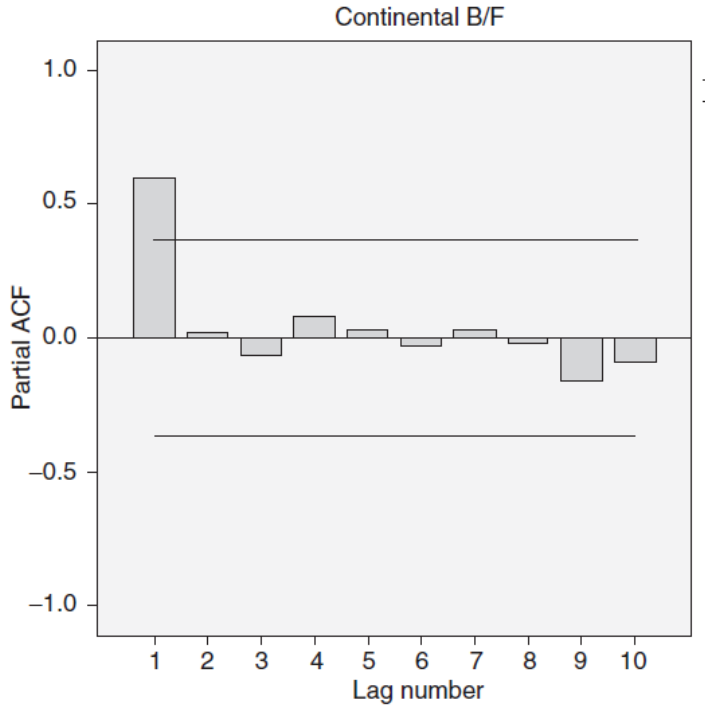
Day	Demand CB	Day	Demand CB
1	25	20	43
2	25	21	41
3	25	22	46
4	35	23	41
5	41	24	40
6	30	25	32
7	40	26	41
8	40	27	41
9	40	28	40
10	40	29	43
11	40	30	46
12	40	31	45
13	44	32	45
14	49	33	46
15	50	34	43
16	45	35	40
17	40	36	41
18	42	37	41
19	40		

Identifying *p*, the order of the AR Model

The first step in AR model building is the identification of the right value of p using ACF and PACF plots. ACF and PACF based on the first 30 observations are given in Figures 13.5 and 13.6, respectively. The horizontal lines in the plot represent the upper and lower 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.



ACF



PACF

DATA ANALYTICS

Results for AR(1)

Model	Model Fit Statistics			
	R-Square	RMSE	MAPE	Normalized BIC
Continental B/F-Model_1	0.373	5.133	10.518	3.498

$$(F_{t+1} - 38.890) = 0.731(Y_t - 38.890)$$

Day	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	45	44.08741	0.832821	0.02028
32	45	43.35641	2.701388	0.036524
33	46	43.35641	6.988568	0.057469
34	43	44.08741	1.182461	0.025289
35	40	41.89441	3.588789	0.04736
36	41	39.70141	1.686336	0.031673
37	41	40.43241	0.322158	0.013844

MAPE 1.5721
RMSE 0.0332 (3.32%)

$$(F_{t+k} - 38.890) = 0.731(F_{t+k-1} - 38.890)$$

Day	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	45	44.0874	0.8328	0.0203
32	45	42.6893	5.3393	0.0513
33	46	41.6673	18.7723	0.0942
34	43	40.9202	4.3256	0.0484
35	40	40.3741	0.1399	0.0094
36	41	39.9749	1.0509	0.0250
37	41	39.6830	1.7344	0.0321

MAPE 2.1446
RMSE 0.04009 (4.009%)

Moving Average (MA) process

- In this simple model, the dependent variable is regressed against lagged values of the past terms or error terms. MA(1) is given by: $Y_{t+1} = \mu + \alpha_1 \varepsilon_t + \varepsilon_{t+1}$

$$Y_{t+1} = \alpha_1 \varepsilon_1 + \varepsilon_{t+1}$$

- MA(q) is given by:

$$Y_{t+1} = \mu + \alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \cdots + \alpha_q \varepsilon_{t-q+1} + \varepsilon_{t+1}$$

- Order q of a MA process:

1. Auto-correlation value, $\left|\rho_p\right| > 1.96/\sqrt{n}$ for first q values (first q lags) and cuts off to zero.
2. The partial auto-correlation function, ρ_{pk}, decreases exponentially.

AR(p) and MA(q)

Steps

1. Before conducting a regression, we need to consider whether the variables are stationary or not.
2. The ACF and Correlogram is one way of determining if a series is stationary, as is the Q- statistic
3. An AR(p) process involves the use of p lags of the dependent variable as explanatory variables
4. A MA(q) process involves the use of q lags of the error term

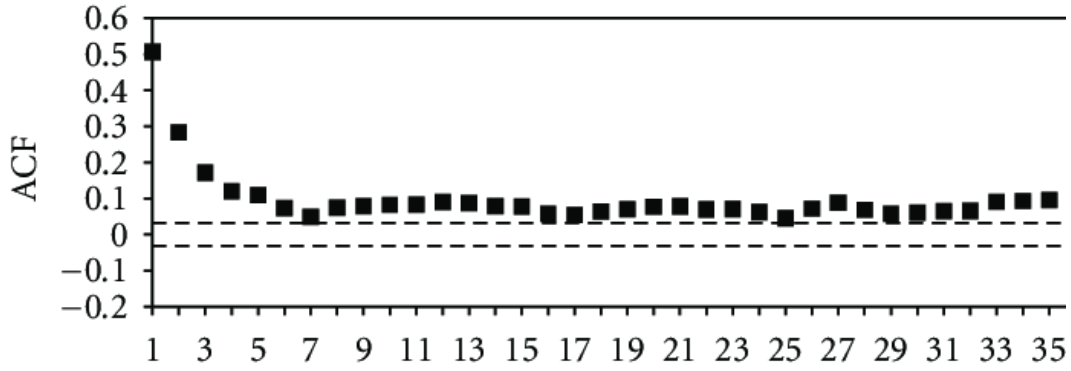
ARMA(p,q): Autoregressive, Moving Average

$$Y_{t+1} = \overbrace{\beta_1 Y_t + \beta_2 Y_{t-1} + \ldots + \beta_p Y_{t-p+1}}^{\text{Auto Regressive Part}} + \overbrace{\alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q+1}}^{\text{Moving Average Part}} + \varepsilon_{t+1}$$

1. Auto-correlation value, $\left|\rho_p\right| > 1.96/\sqrt{n}$ for first q values (first q lags) and cuts off to zero.
2. Partial auto-correlation function, $\left|\rho_{pk}\right| > 1.96/\sqrt{n}$ for first p values and cuts off to zero.

ARMA(p,q): Parameter selection

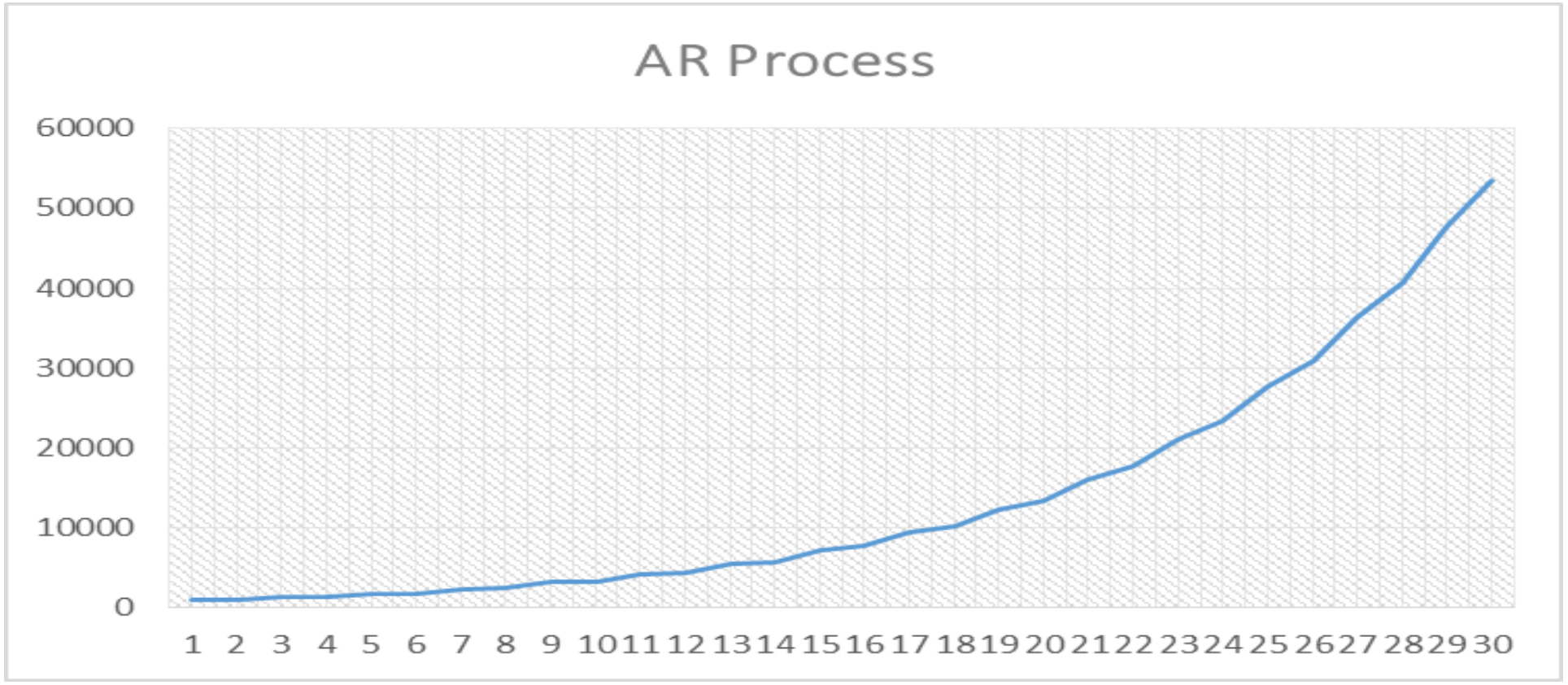
Model	ACF	PACF
AR (p)	Spikes decay towards zero. Coefficients may oscillate.	Spikes decay to zero after lag p
MA (q)	Spikes decay to zero after lag q	Spikes decay towards zero. Coefficients may oscillate.
ARMA (p, q)	Spikes decay (either direct or oscillatory) to zero beginning after lag q	Spikes decay (either direct or oscillatory) to zero beginning after lag p



AR & MAModels

- Autoregressive AR process:
  - Series current values depend on its own previous values
  - AR(p) - Current values depend on its own p-previous values
  - P is the order of AR process
- Moving average MA process:
  - The current deviation from mean depends on previous deviations
  - MA(q) - The current deviation from mean depends on q- previous deviations
  - q is the order of MA process
- Autoregressive Moving average ARMA process

AR Model

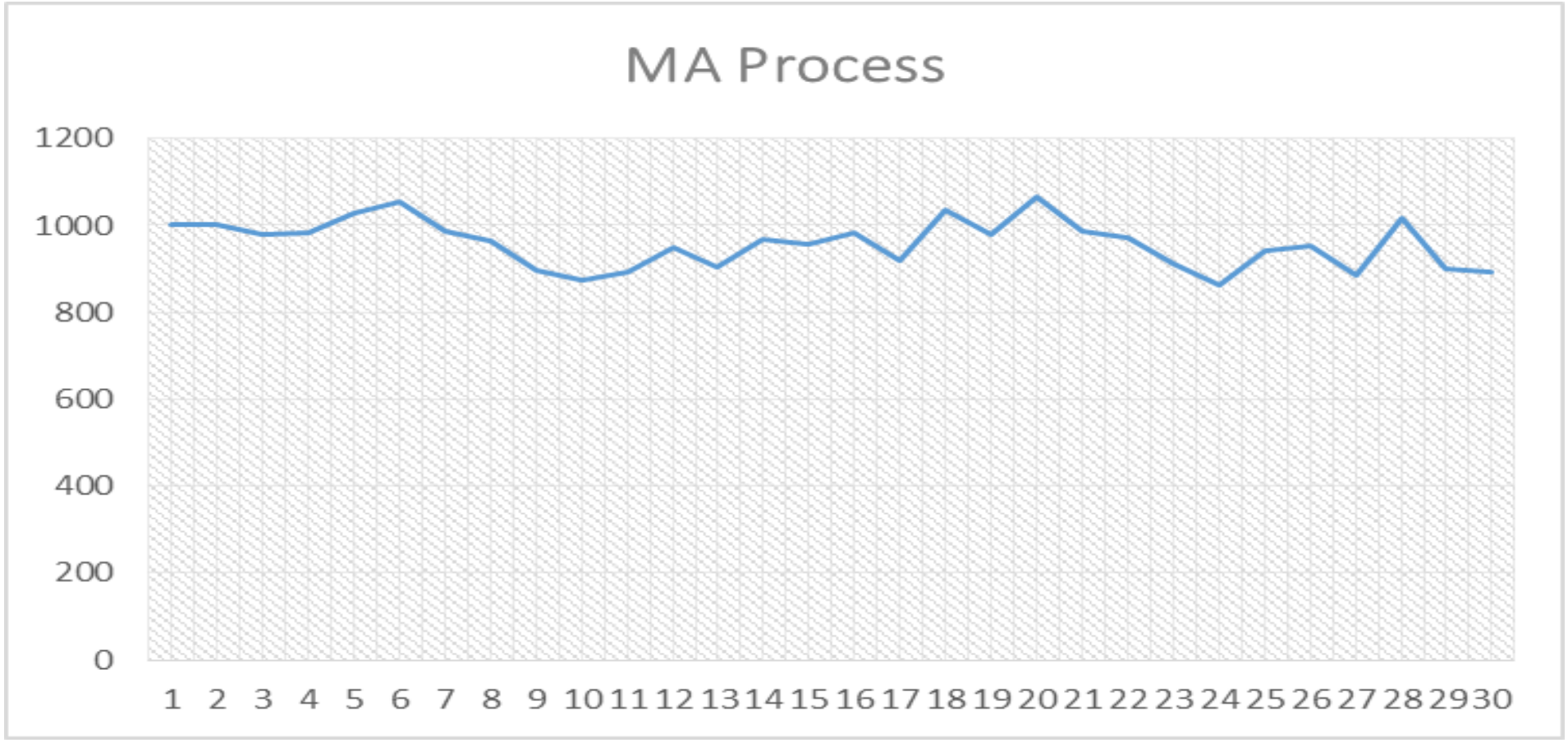


AR(1) $y_t = a1 * y_{t-1}$

AR(2) $y_t = a1 * y_{t-1} + a2 * y_{t-2}$

AR(3) $y_t = a1 * y_{t-1} + a2 * y_{t-2} + a3 * y_{t-2}$

MA Model



MA(1) $\varepsilon_t = b1*\varepsilon_{t-1}$

MA(2) $\varepsilon_t = b1*\varepsilon_{t-1} + b2*\varepsilon_{t-2}$

MA(3) $\varepsilon_t = b1*\varepsilon_{t-1} + b2*\varepsilon_{t-2} + b3*\varepsilon_{t-3}$

ARMA(p, q) – An example

Monthly demand for avionic system spares used in Vimana 007 aircraft is provided.
Build an ARMA model based on the first 30 months of data and forecast the demand
for spares for months 31 to 37. Comment on the accuracy of the forecast.

Month	Demand for Spares	Month	Demand for Spares
1	457	20	516
2	439	21	656
3	404	22	558
4	392	23	647
5	403	24	864
6	371	25	610
7	382	26	677
8	358	27	609
9	594	28	673
10	482	29	400
11	574	30	443
12	704	31	503
13	486	32	688
14	509	33	602
15	537	34	629
16	407	35	823
17	523	36	671
18	363	37	487
19	479		

Example: Step1 – Plot ACF, PACF

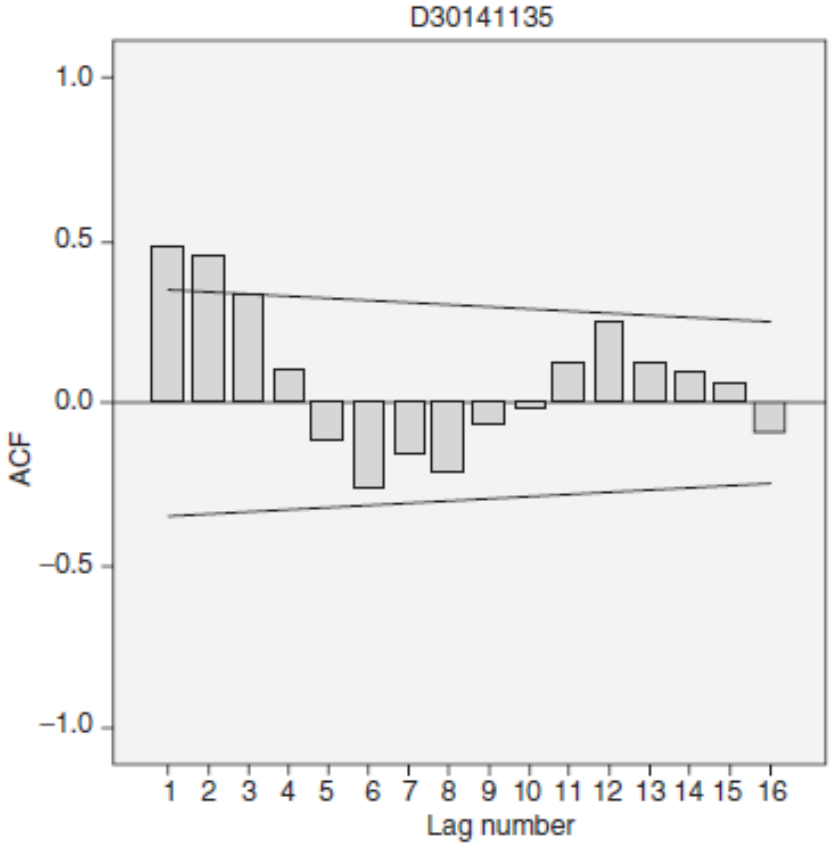


FIGURE 13.9 ACF plot for avionic system spares demand

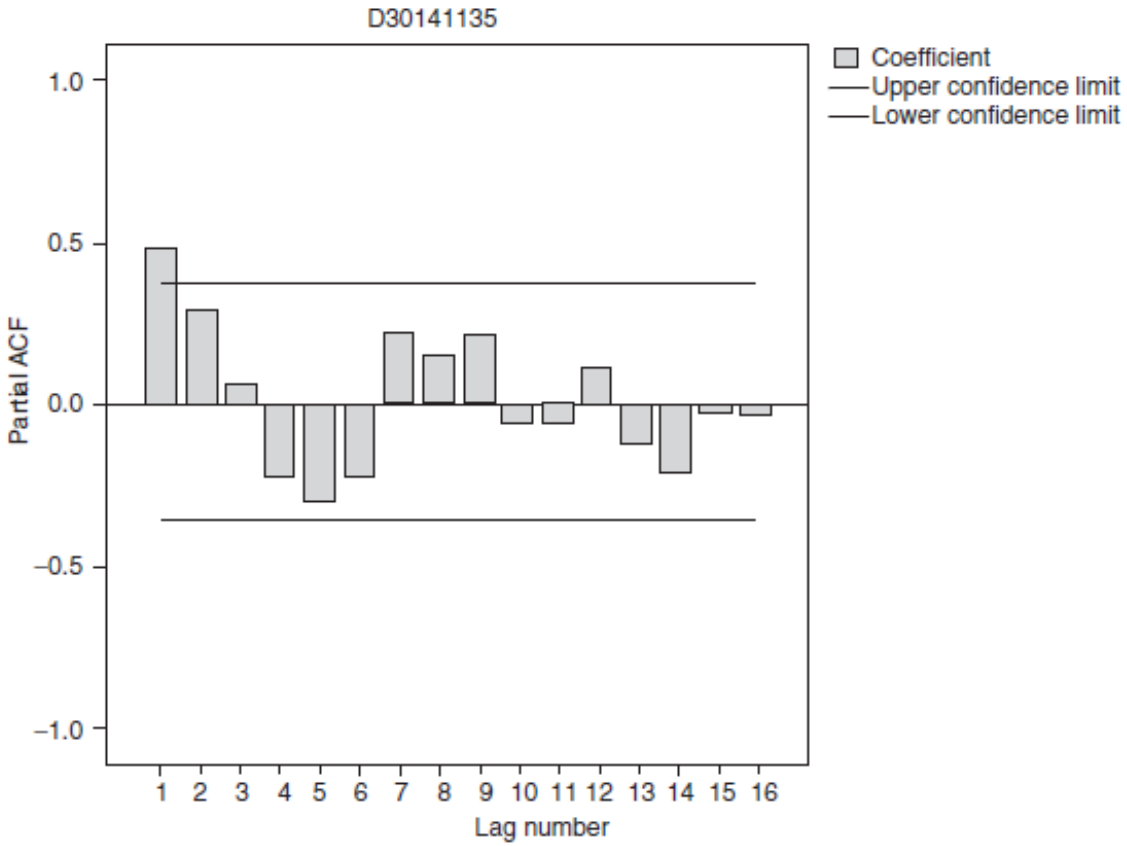


FIGURE 13.10 PACF plot for avionic system spares demand.

Example: Step 2 – Forecast (ARMA(1,2))

Model	Model Fit Statistics		
	Stationary *R*-Squared	RMSE	MAPE
Avionic Spares	0.429	98.824	14.231

TABLE 13.26 l model parameters

			Estimate	*SE*	*T*	Sig.
Avionic Spares	Constant		496.699	57.735	8.603	0.000
	AR	Lag 1	0.706	0.170	4.153	0.000
	MA	Lag 1	0.694	0.173	4.006	0.000
		Lag 2	−0.727	0.170	−4.281	0.000

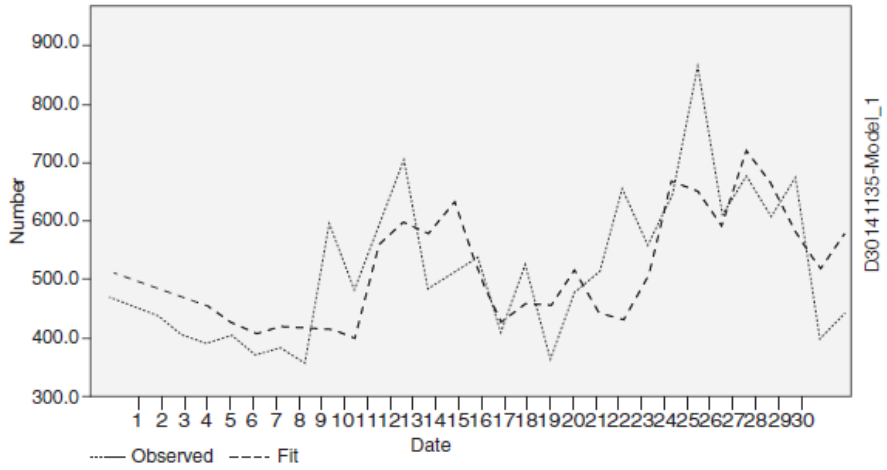


FIGURE 13.11 Observed versus forecasted demand.

All the three components in the ARMA model (AR lag 1 and MA lags 1 and 2) are statistically significant (Table 13.26). The model equation using SPSS is given by

$$Y_{t+1} - 496.669 = 0.706 \times (Y_t - 496.699) - 0.694 \times \varepsilon_t + 0.727 \times \varepsilon_{t-1} \qquad (13.45)$$

Example: Step 3 – Compute MAPE, RMSE

TABLE 13.27 ARMA(1, 2) model forecast

Month	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	503	464.8107	1458.423	0.075923
32	688	378.5341	95769.15	0.449805
33	602	444.6372	24763.04	0.2614
34	629	685.8851	3235.909	0.090437
35	823	743.5124	6318.281	0.096583
36	671	630.7183	1622.614	0.060032
37	487	649.3491	26357.22	0.333366

The RMSE and MAPE for the validation data (months 31 and 37) are 150.961 0.1953 (19.53%), respectively (Table 13.27).

The forecasted values using F_t instead of Y_t when forecasting for more than one period ahead in time are shown in Table 13.28.

TABLE 13.28 ARMA (1, 2) forecast

Month	Y_t	F_t	$(Y_t - F_t)^2$	$\lvert Y_t - F_t \rvert / Y_t$
31	503	464.4239	1488.1147	0.0767
32	688	377.8374	96200.8258	0.4508
33	602	444.5195	24800.1101	0.2616
34	629	687.2082	3388.1980	0.0925
35	823	744.9583	6090.4998	0.0948
36	671	630.5592	1635.4571	0.0603
37	487	648.3959	26048.6313	0.3314

The RMSE and MAPE for the validation data (months 31 and 37) are 151.02 and 0.1954 (19.54%), respectively.

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 (Ch. 13.10-13.13)

Additional reference for the interested reader:
Introduction to Time Series and Forecasting, Second Edition by Peter J. Brockwell, Richard A. Davis Springer 2002.

Image Courtesy

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

Lecture 6: 13.10, 13.12, 13.13 in text - AR, MA and ARMA models
(AṚ https://otexts.com/fpp2/AR.html) + MA (https://otexts.com/fpp2/MA.html) + ARMA Venkat Reddy's slides on ARIMA)

https://www.business-science.io/timeseries-analysis/2017/08/30/tidy-timeseries-analysis-pt-4.html

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu