DATA ANALYTICS

Unit 3: Forecasting with Regression, Stationary Signals and ARMA

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

DATA ANALYTICS

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*.

DATA ANALYTICS

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 R2 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

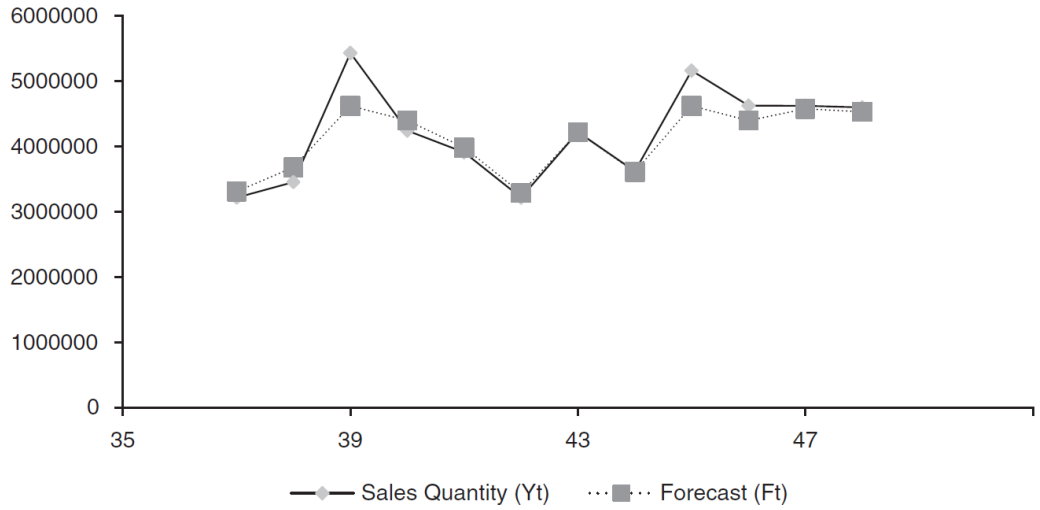
DATA ANALYTICS

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%**

DATA ANALYTICS

Forecasting with Regression – Seasonality

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 $F_{t+1} = F_{d,\,t+1} \times S_{t+1}$.

DATA ANALYTICS

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.

DATA ANALYTICS

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

$$E(y_t) = \mu$$

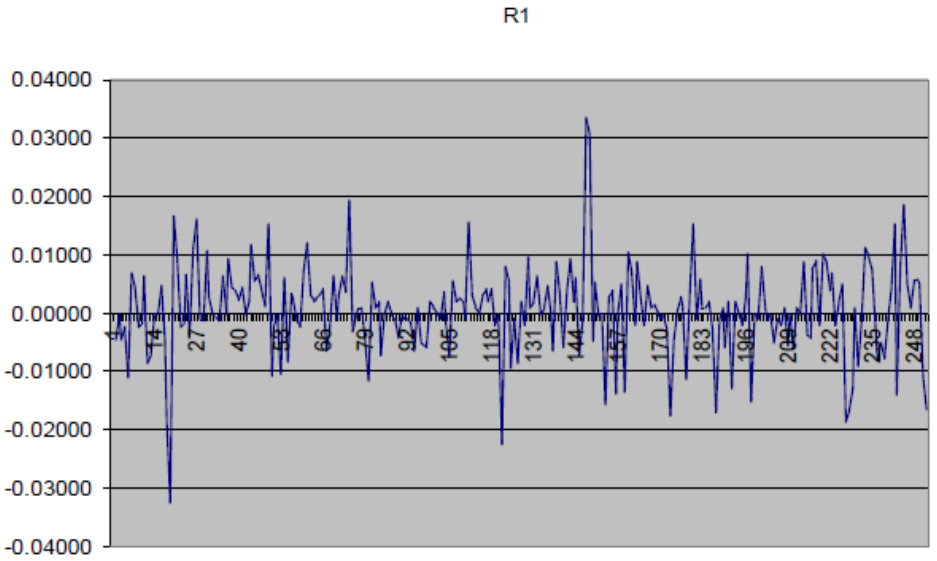
$$E(y_t - \mu)^2 = \sigma^2$$

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

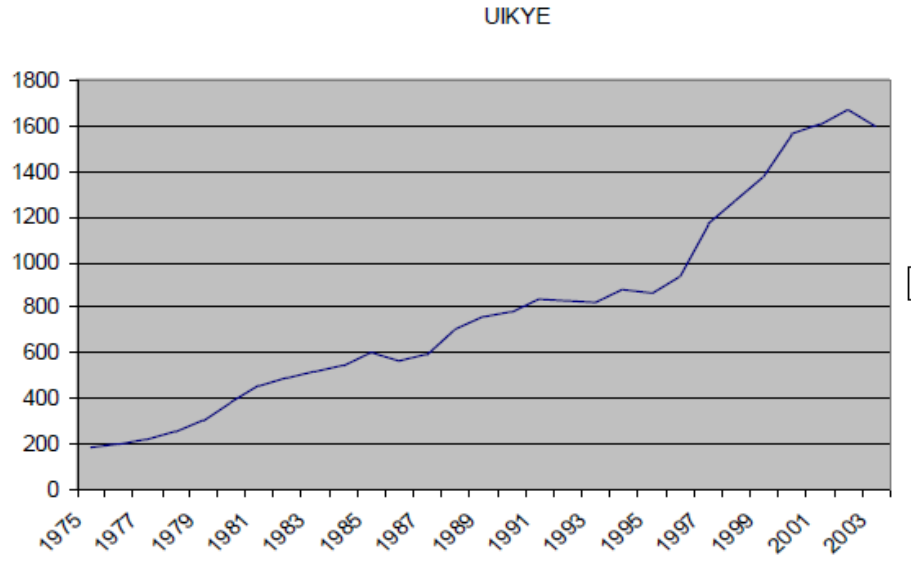
- The latter refers to the covariance between y(t-1) and y(t-2) being the same as y(t-5) and y(t-6).

DATA ANALYTICS

Stationary and NonStationary Series



Stationary Series



Non-stationary Series

DATA ANALYTICS

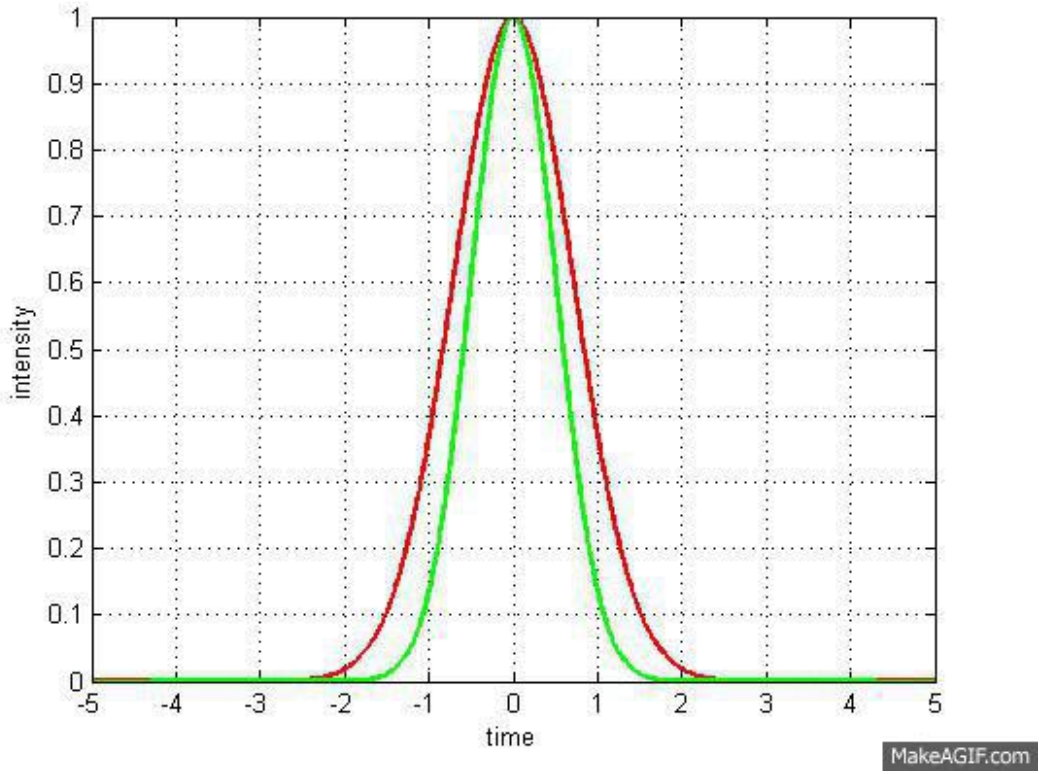
Implications of Nonstationary Data

- 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*

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

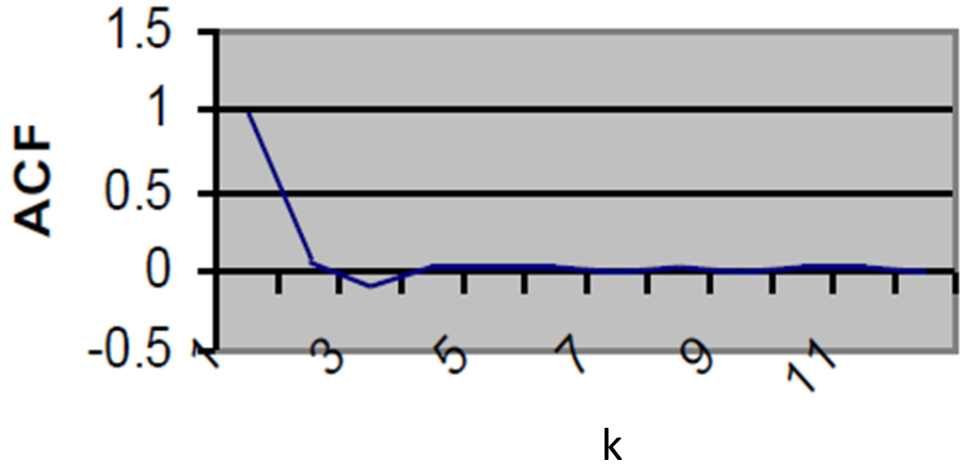
$$R_k = \frac{\sum_{t=k+1}^{n} \left(Y_{t-k} - \bar{Y}\right)\left(Y_t - \bar{Y}\right)}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2}$$



DATA ANALYTICS

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.

DATA ANALYTICS

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.

DATA ANALYTICS

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).

DATA ANALYTICS

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.

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.

DATA ANALYTICS

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($\beta_{t-k} e_k$) is zero.

DATA ANALYTICS

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

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

DATA ANALYTICS

AR Model Identification

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

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.

DATA ANALYTICS

AR Model - Example

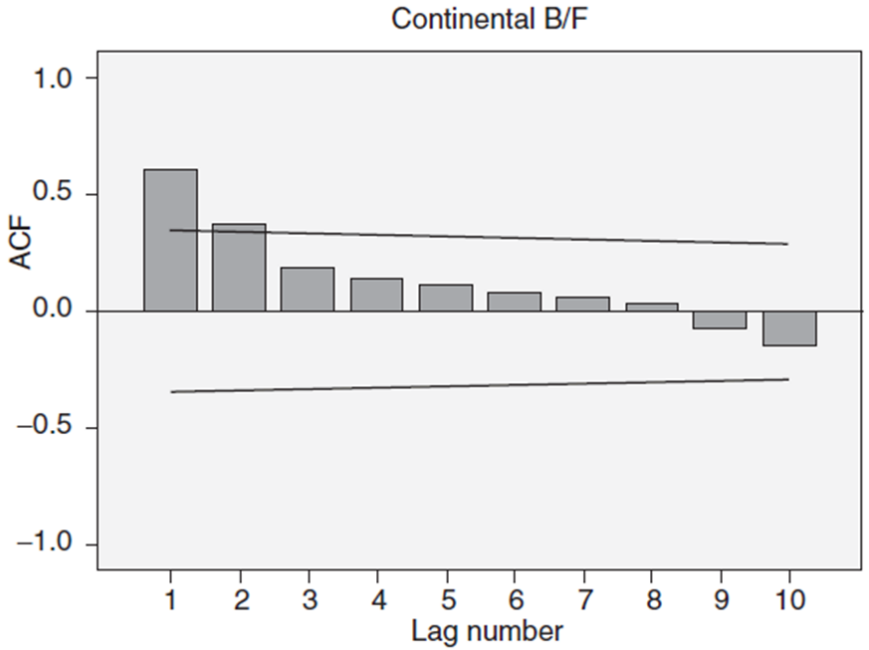
Build an auto-regressive model based on the first 30 days of data and forecast the demand for continental breakfast on days 31 to 37. Comment on the accuracy of the forecast.

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		

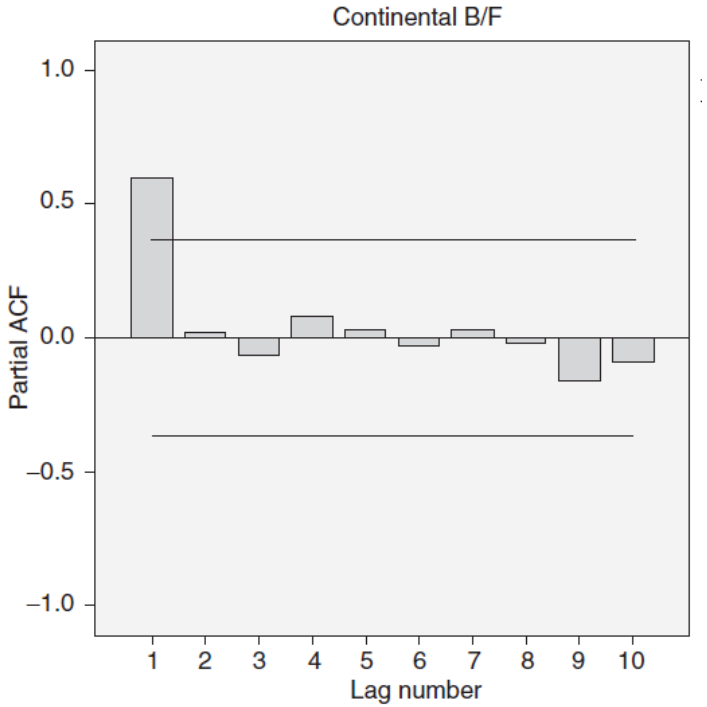
DATA ANALYTICS

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%)

DATA ANALYTICS
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)

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

DATA ANALYTICS

ARMA(p,q): Autoregressive, Moving Average

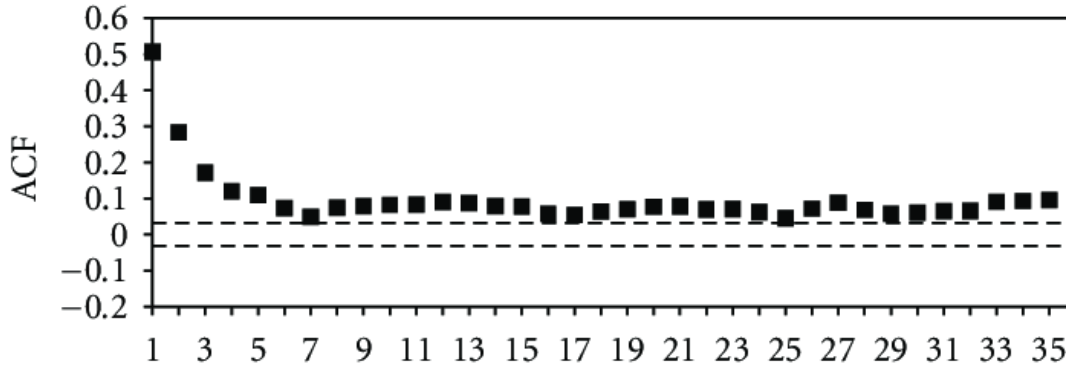
$$Y_{t+1} = \overbrace{\beta_1 Y_t + \beta_2 Y_{t-1} + \ldots + \beta_p Y_{t-p+1}}^{\text{Auto Regressive Part}} + \overbrace{\alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q+1}}^{\text{Moving Average Part}} + \varepsilon_{t+1}$$

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.

DATA ANALYTICS

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



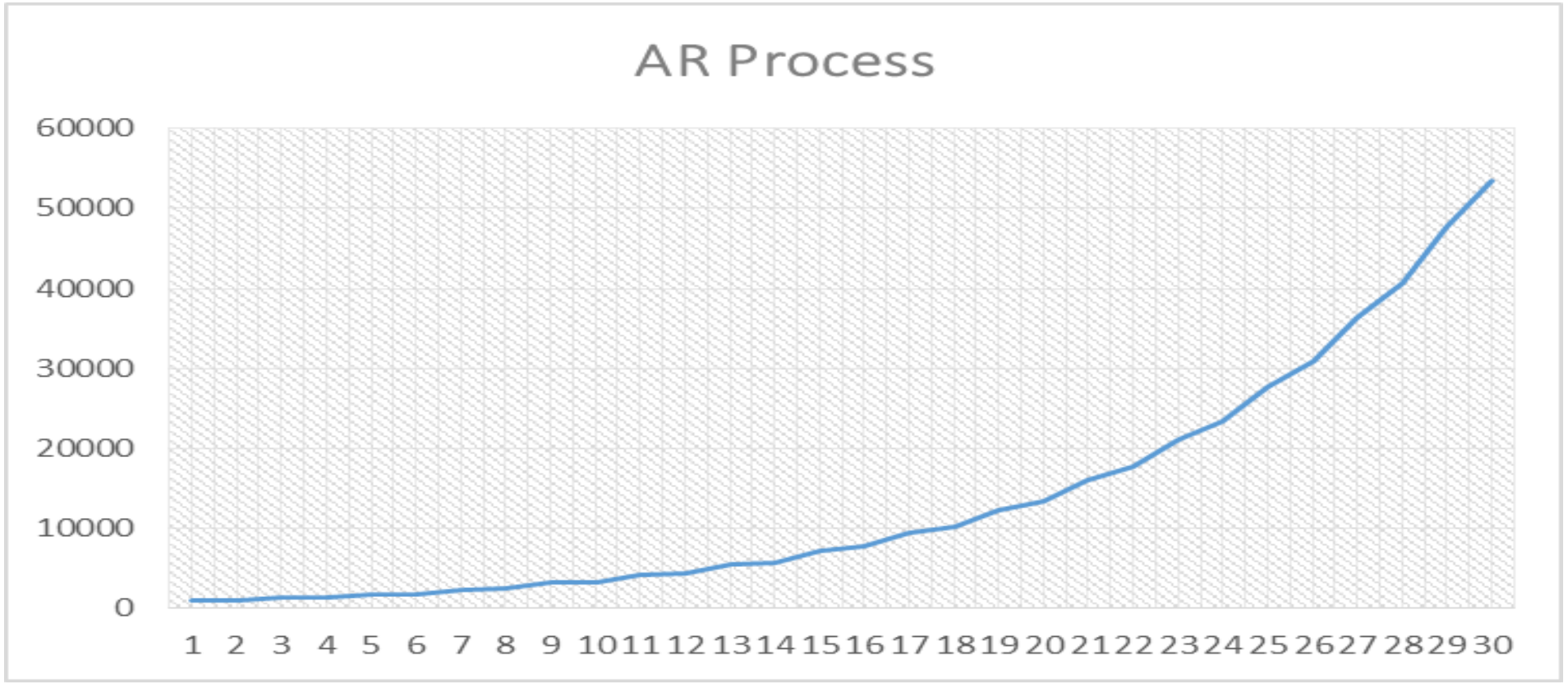
DATA ANALYTICS

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

DATA ANALYTICS

AR Model



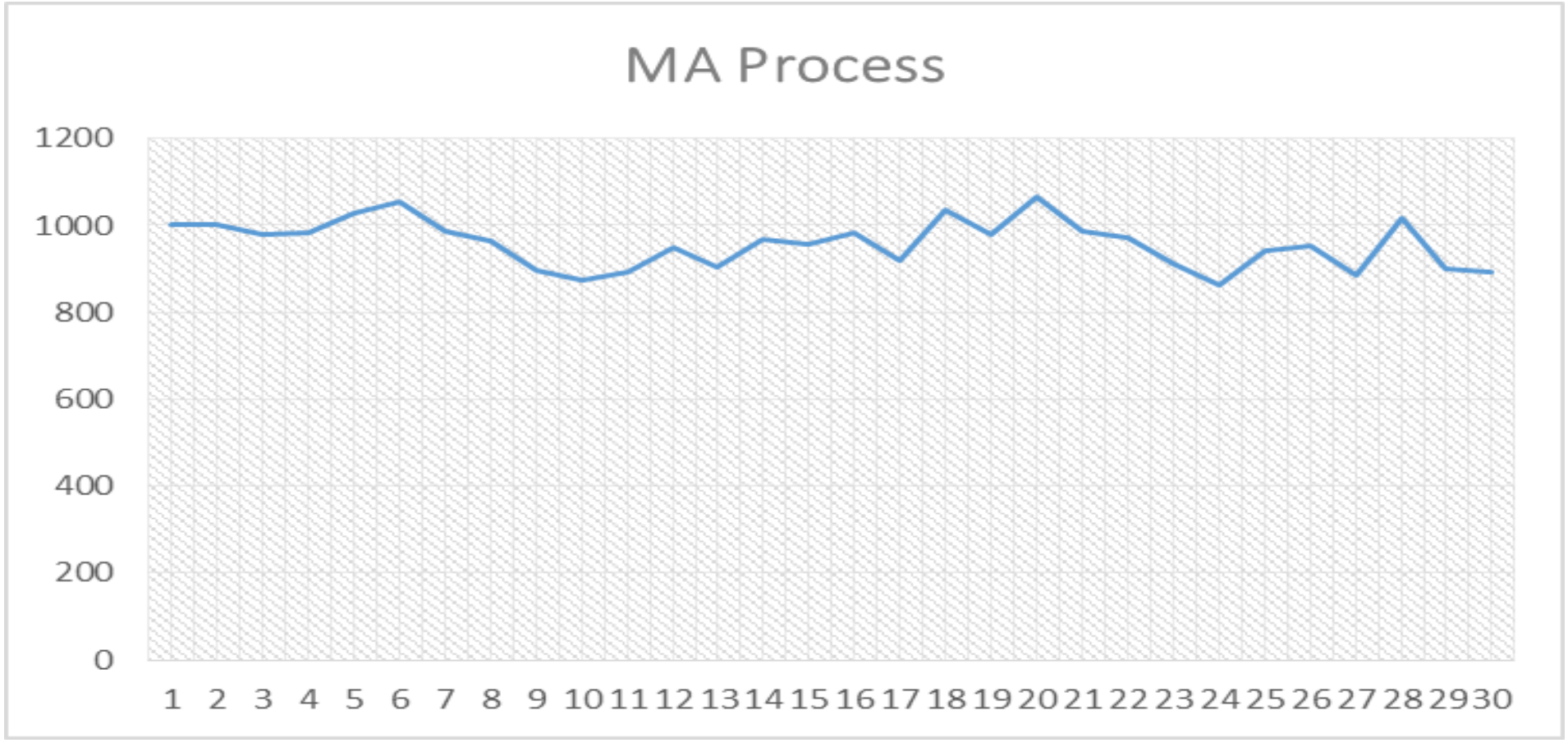
AR(1) $y_t = a1 * y_{t-1}$

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

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

DATA ANALYTICS

MA Model



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

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

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

DATA ANALYTICS

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		

DATA ANALYTICS

Example: Step1 – Plot ACF, PACF

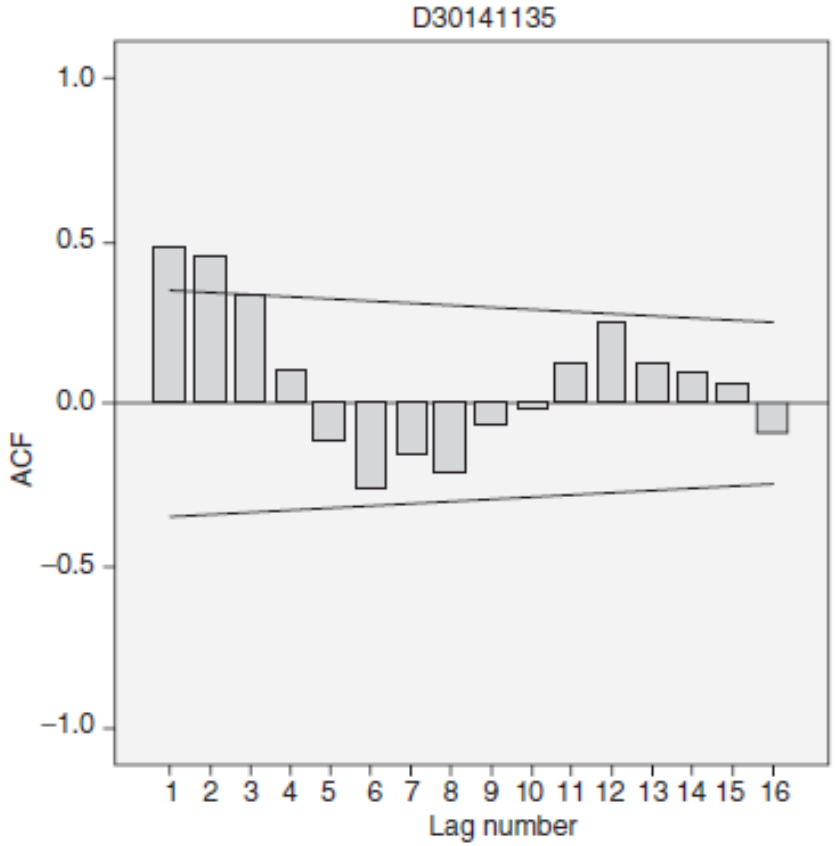


FIGURE 13.9 ACF plot for avionic system spares demand

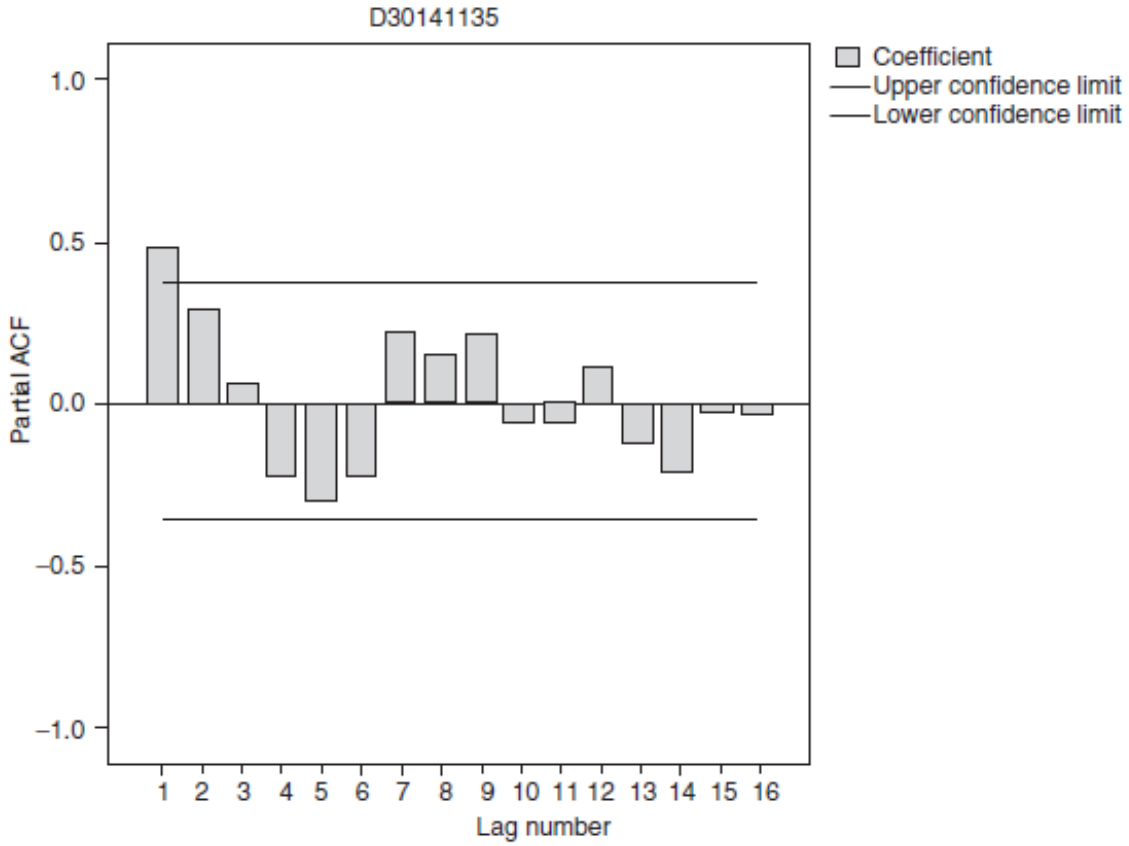


FIGURE 13.10 PACF plot for avionic system spares demand.

DATA ANALYTICS

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 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

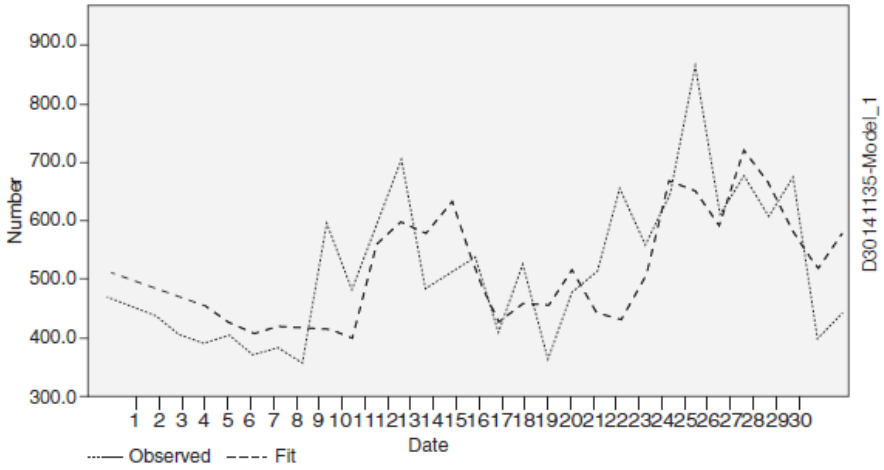


FIGURE 13.11 Observed versus forecasted demand.

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)

THANK YOU

Jyothi R
Assistant Professor, Department of Computer Science
jyothir@pes.edu