ÖZYEĞİN ÜNİVERSİTESİ

M7
Predictive Analytics

ENİS KAYIŞ

Exploring Data Patterns & Choosing a Forecasting Technique

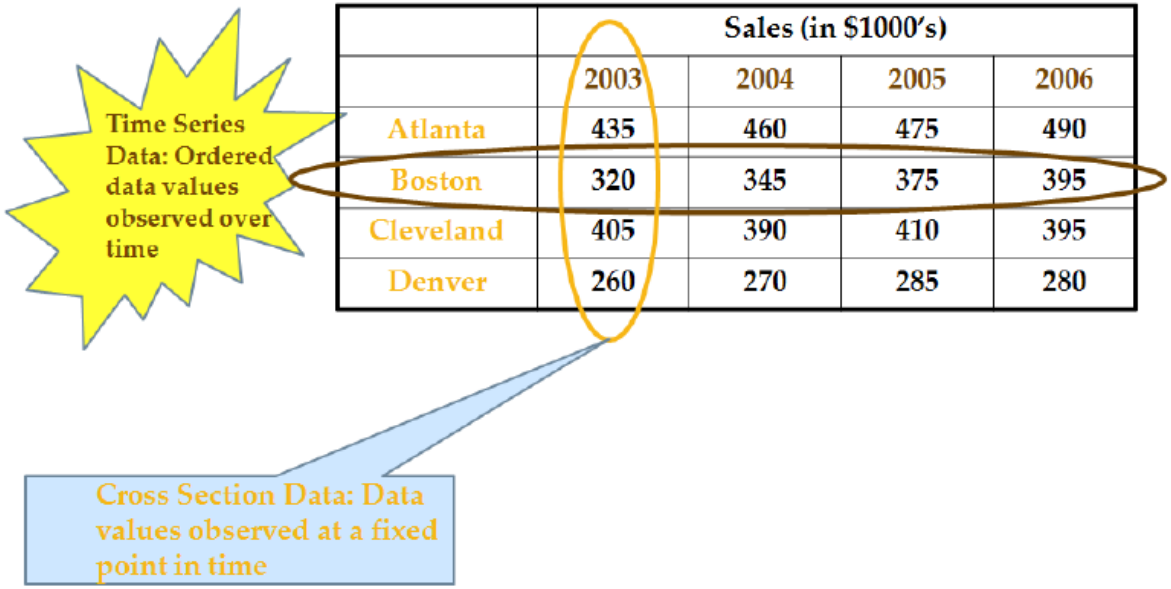
ÖZYEĞİN UNIVERSITY

- Collection of valid and reliable data is the most time consuming and difficult part of forecasting.
  - The difficult task facing most forecasters is how to find relevant data that will help solve their specific decision making problems.
  - GIGO: garbage in and garbage out
  - The following four criteria can be applied to the determination of whether data will be useful:
    1. Data should be reliable and accurate.
    2. Data should be relevant.
    3. Data should be consistent.
    4. Data should be timely.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- There Are Two Types of Data :
  - One are observations collected at a single point in time called cross-sectional data and,
  - The other one are observations collected over successive increments of time called time series data.



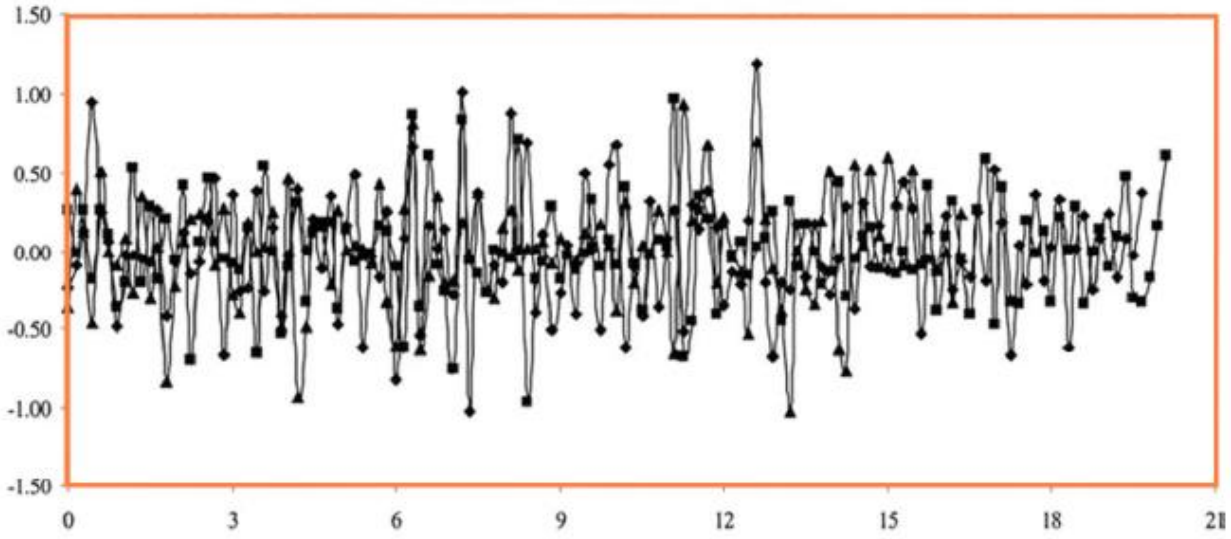
	Sales (in $1000's)			
	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Exploring Time Series Data Patterns :

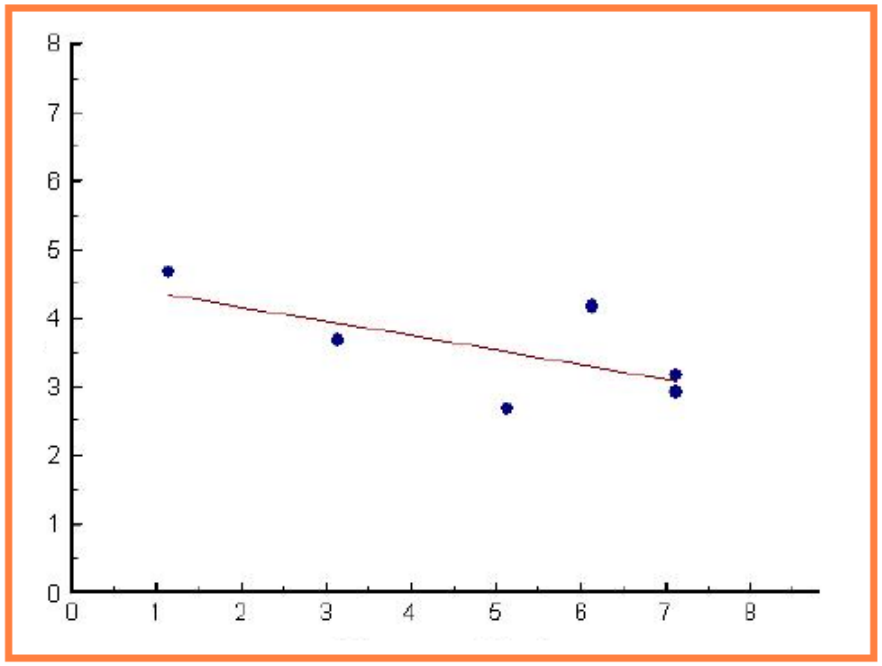
- The important aspects in selecting an appropriate forecasting method for time series data is to consider the following different types of data patterns:
  - 1. Horizontal Pattern
    - The observations fluctuate around a constant level or mean.
    - This type of series is called stationary in its mean.



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

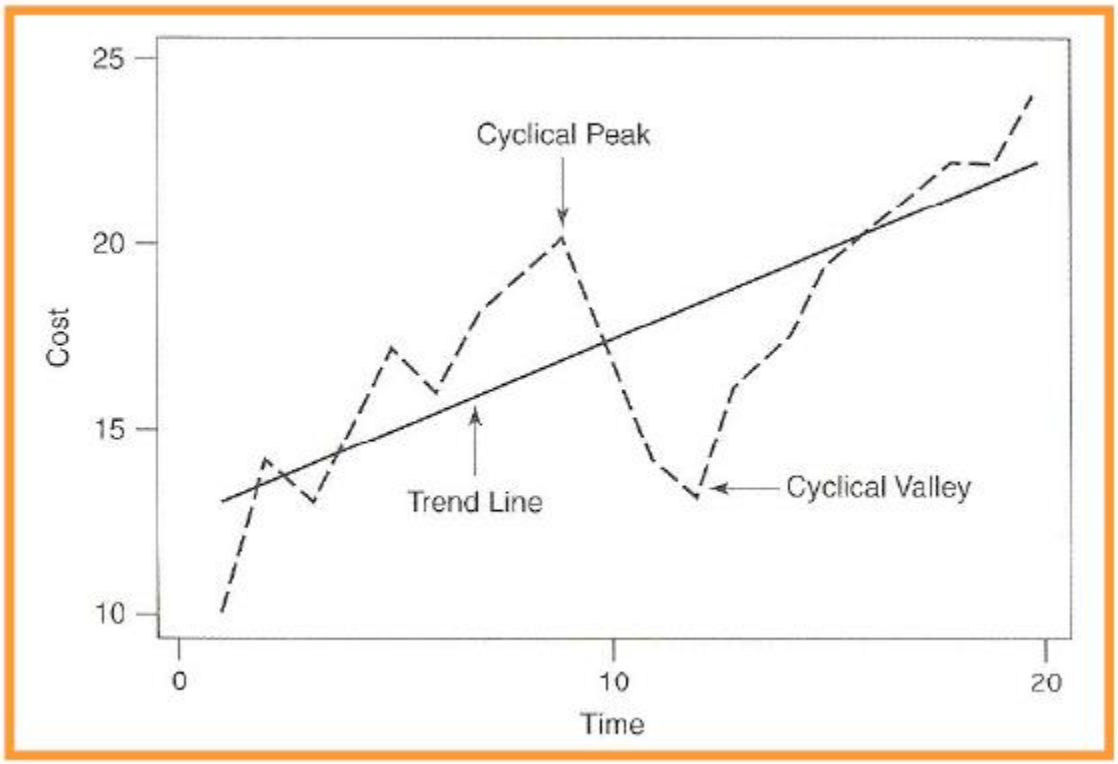
- 2. Trend Pattern
  - The observations grow or decline over an extended period of time.
  - This type of series is called nonstationary.
  - The trend is the long-term component that represents the growth or decline in the time series over an extended period of time.



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

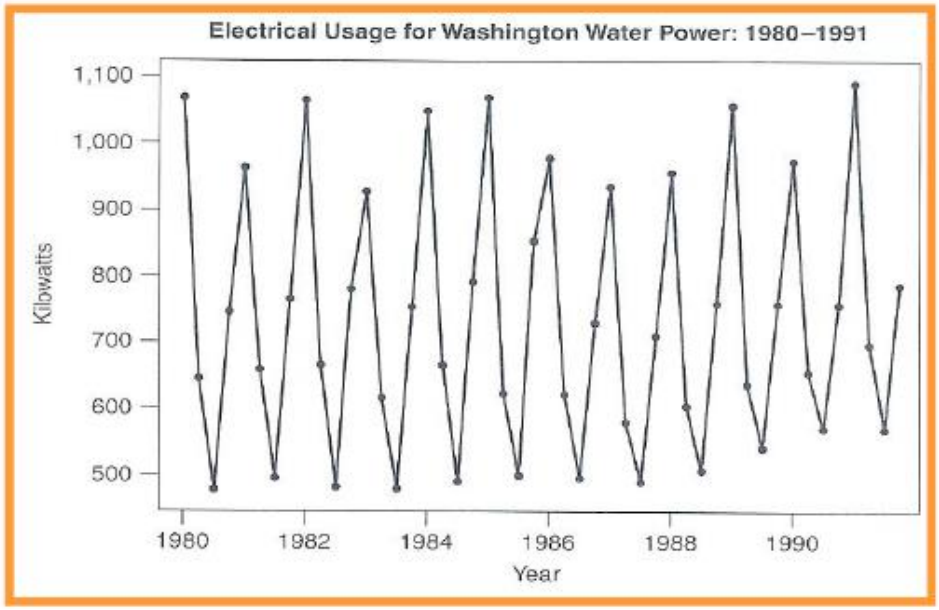
- 3. Cyclical Pattern
  - The observations exhibit rises and falls that are not of a fixed period.
  - The cyclic component is the Wave like fluctuation around the trend.
  - Cyclical fluctuations are often influenced by changes in economic expansions and contractions (business cycle)



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- 4. Seasonal Pattern
  - The observations are influenced by seasonal factors.
  - The seasonal component refers to a pattern of change that repeats it self year after year.
  - In the monthly series the seasonal component measures the variability of the series each month, and in the quarterly series the seasonal component represents the variability in each quarter…etc.



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Autocorrelation Analysis

- The autocorrelation analysis for different time lags of a variable is used to identify time series data patterns including components such as trend and seasonality.
- Autocorrelation is the correlation between a variable lagged one or more periods and itself.
- This is measured using the autocorrelation coefficient at lag k , which is denoted by ρ_k and it's estimated by its sample autocorrelation coefficient r_k at lag k ; k=0,1,2,…

  where;

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2} \text{ for k=0,1,2,…}$$

where Y_t and Y_{t-k} are observations at time period t and t-k respectively.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Autocorrelation analysis can be done by demonstrating the autocorrelation function (ACF).
- The ACF is a graph of the autocorrelations for various lags of a time series.
- Example:
- We have collected data on the number of products sold last year.
- We need to know lag 1 and lag 2 autocorrelation coefficients (r_1 and r_2).

January	123
February	130
March	125
April	138
May	145
June	142
July	141
August	146
September	147
October	157
November	150
December	160

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Solution:

Time t	Month	Original Data Y_t	Y Lagged One Period Y_{t-1}	Y Lagged Two Periods Y_{t-2}
1	January	123		
2	February	130	123	
3	March	125	130	123
4	April	138	125	130
5	May	145	138	125
6	June	142	145	138
7	July	141	142	145
8	August	146	141	142
9	September	147	146	141
10	October	157	147	146
11	November	150	157	147
12	December	160	150	157

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Lag 1

Time, t	Y_t	Y_{t-1}	$(Y_t-\bar{Y})$	$(Y_{t-1}-\bar{Y})$	$(Y_t-\bar{Y})^2$	$(Y_t-\bar{Y})(Y_{t-1}-\bar{Y})$
1	123	—	−19	—	361	—
2	130	123	−12	−19	144	228
3	125	130	−17	−12	289	204
4	138	125	−4	−17	16	68
5	145	138	3	−4	9	−12
6	142	145	0	3	0	0
7	141	142	−1	0	1	0
8	146	141	4	−1	16	−4
9	147	146	5	4	25	20
10	157	147	15	5	225	75
11	150	157	8	15	64	120
12	160	150	18	8	324	144
Total	1,704		0		1,474	843

$$r_1 = \frac{\sum_{t=1+1}^{n} (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} = \frac{843}{1,474} = .572$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Lag 2

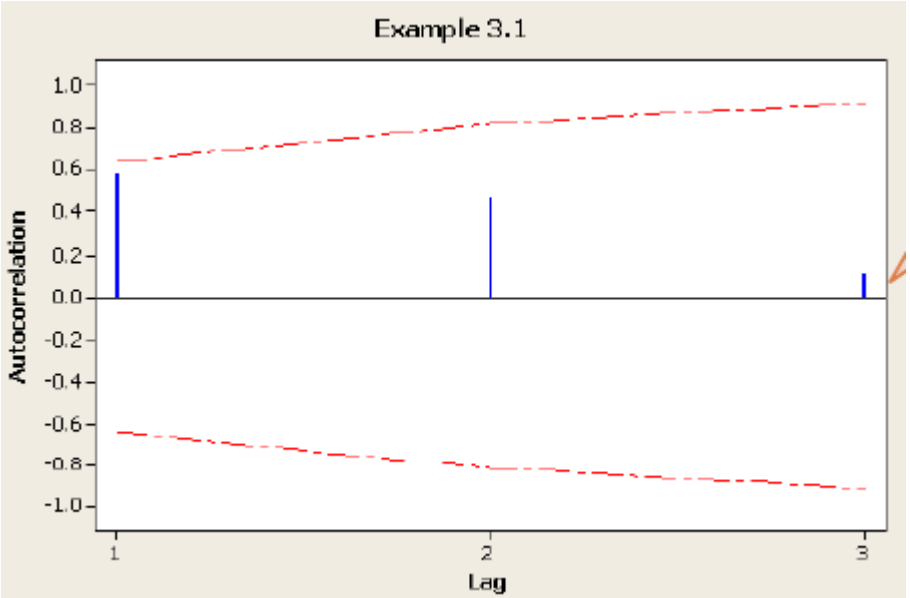
Time, t	Y_t	Y_{t-2}	$(Y_t-\bar{Y})^2$	$(Y_{t-2}-\bar{Y})$	$(Y_t-\bar{Y})$	$(Y_t-\bar{Y})(Y_{t-2}-\bar{Y})$
1	123	-	361	-	-19	-
2	130	-	144	-	-12	-
3	125	123	289	-19	-17	323
4	138	130	16	-12	-4	48
5	145	125	9	-17	3	-51
6	142	138	0	-4	0	0
7	141	145	1	3	-1	-3
8	146	142	16	0	4	0
9	147	141	25	-1	5	-5
10	157	146	225	4	15	60
11	150	147	64	5	8	40
12	160	157	324	15	18	270
Total	1,704		1,474			682

$$r_2 = \frac{\sum_{t=2+1}^{n} (Y_t - \bar{Y})(Y_{t-2} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} = \frac{682}{1,474} = .463$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- The *ACF* is a graph of the autocorrelation for various lags of time series.



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- With the ACF, the data patterns including trend and seasonality can be studied.
- Autocorrelation coefficients for different time lags for a variable can be used to answer the following questions about a time series:
  1. Are the data random?
  2. Do the data have a trend?
  3. Are the data stationary?
  4. Are the data seasonal?
- If a series is random, the autocorrelations between Y_t and Y_{t-k} for any lag k are close to zero. The successive values of a time series are not related to each other.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- If a series has a trend, successive observations are highly correlated and the autocorrelation coefficients are typically significantly different from zero for the first several time lags and then gradually drop toward zero as the number of lags increases.
  - The autocorrelation coefficient for time lag 1 is often very large (close to 1).
  - The autocorrelation coefficient for time lag 2 will also be large. However, it will not he as large as for time lag 1.
- If a series has a seasonal pattern, a significant autocorrelation coefficient will occur at the seasonal time lag or multiples of the seasonal lag.
  - The seasonal lag is 4 for quarterly data and 12 for monthly data.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- How does an analyst determine whether an autocorrelation coefficient is significantly different from zero?
  - Statisticians showed that the sampling distribution of the sample autocorrelation coefficient r_1 is normally distributed with mean zero and approximate standard deviation $1/\sqrt{n}$.
  - Knowing this, we can compare the sample autocorrelation coefficients with this theoretical sampling distribution and determine whether, for given time lags, they come from a population whose mean is zero.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Checking the Significance of the Autocorrelation Coefficients

- We need to determine whether the autocorrelation coefficient ρ_k at lag k ; k=0,1,2,… ; is different from zero for any time series data set.
- The sampling distribution of the sample autocorrelation coefficient r_k at lags k; k=2,3,…; is normally distributed with mean zero and approximate standard deviation $SE(r_k)$ and is given by:

$$SE(r_k) = \sqrt{\frac{1 + 2\sum_{i=1}^{k-1} r_i^2}{n}}; k = 2,3, \ldots$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

1. Testing the randomness of the time series data

- If a 100(1-α)% of the sample autocorrelation coefficient r_k at lags k ; k=1,2,… lie within the interval

$$-t_{\alpha/2,n-1}SE(r_k) \leq \rho_k \leq t_{\alpha/2,n-1}SE(rk)$$

then the time series data is random.

2. Testing for individual ρ_k

- Now using α level of significance, we want to test for k=1,2,…

$$H_0: \rho_k=0$$
$$H_a: \rho_k\neq 0$$

- Using the test statistics

$$t = \frac{r_k}{SE(rk)}$$

Reject H_0 if $|t| \geq t_{\alpha/2,n-1}$ or p-value is less than α.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- 3. Testing a subset of ρ_k; k=1,2,…,m
  - We use one of the common portmanteau tests; the following modified Box-Pierce Q Statistics

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

  - We reject that all the subset of autocorrelations are zero if $Q \geq \chi^2_{\alpha,m}$ (chi-squared distribution) or p-value ≤ α.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Exploring Time Series Data Types

The autocorrelation coefficients are used to know if the time series data are:

- 1. Random data.
  - The time series is random or independent if the auto correlations between Y_t and Y_{t-k} for any lag k are close to zero.
  - This implies that the successive data are not related to each other.
  - We use the option of constructing confidence interval to check that almost all sample autocorrelations should lie within a range specified by zero.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- For example, at 5 % level of significant, the time series data are random if 95 % of the sample autocorrelations will lie within

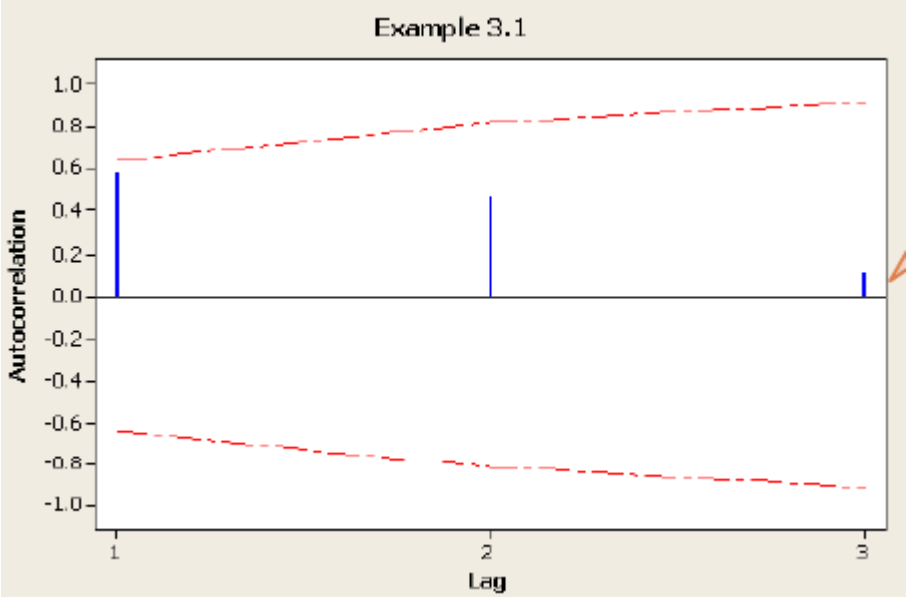
$$-2.2\ SE(r_k) \leq \rho_k \leq +2.2\ SE(r_k) \text{ for all } k = 1,2,3...$$

- Also, it is possible to use the Q Statistic in option of testing a subset of autocorrelations is zero.
- For example, at 5 % level of significant, the time series data are random if Q for a subset of 10 autocorrelations is less than $\chi^2_{0.05,10} = 18.31$.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Method 1: Manual hypothesis testing, or confidence interval construction.
- Method 2: The 95% confidence limits are shown in ACF by the dashed lines in the graphical display.



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- 2. Stationary and nonstationary data.
  - The time series is stationary if the observations fluctuate around a constant level or mean.
    - The sample autocorrelation coefficients decline to zero fairly rapidly, generally after the second or third time lag.
  - The time series is nonstationary or having trend if the successive observations are highly correlated .
    - The autocorrelation coefficients are sufficiently different from zero for the first several time lags and then gradually drop toward zero as the number of lags increases.
  - Nonstationary data to be analyzed the trend should be removed from the data before modeling.
  - One possible technique used to remove the trend is the differencing method.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

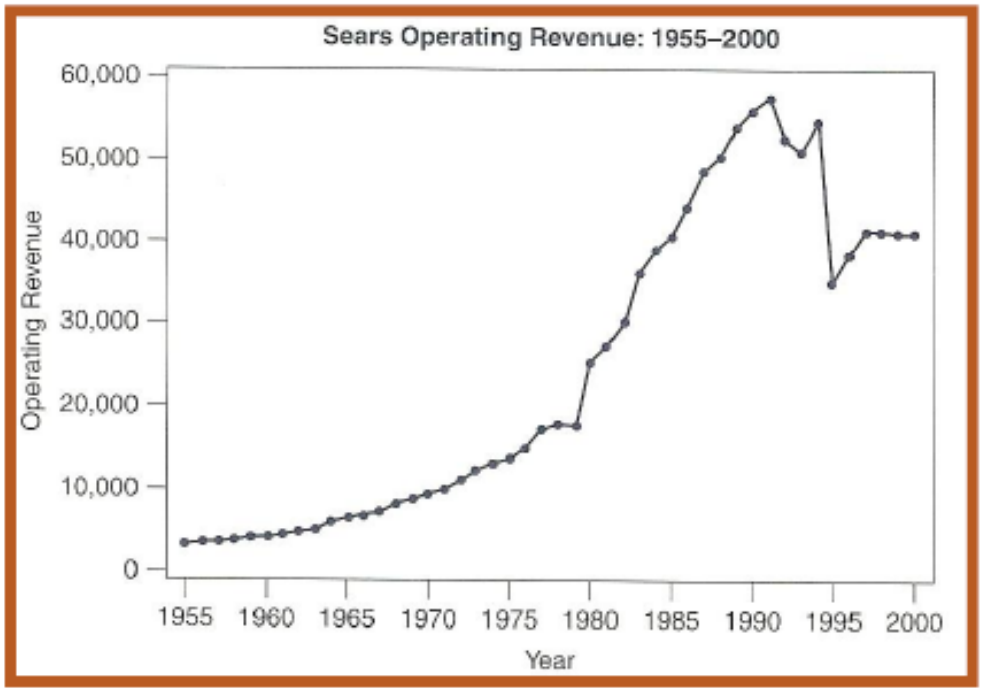
- Difference the data at order 1, $\Delta Y_t = Y_t - Y_{t-1}$ may remove the trend and the time series data becomes stationary.

Time, t	Y_t	Y_{t-1}	$Y_t - Y_{t-1}$
1	123	-	-
2	130	123	7
3	125	130	-5
4	138	125	13
5	145	138	7
6	142	145	-3
7	141	142	-1
8	146	141	5
9	147	146	1
10	157	147	10
11	150	157	-7
12	160	150	10

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

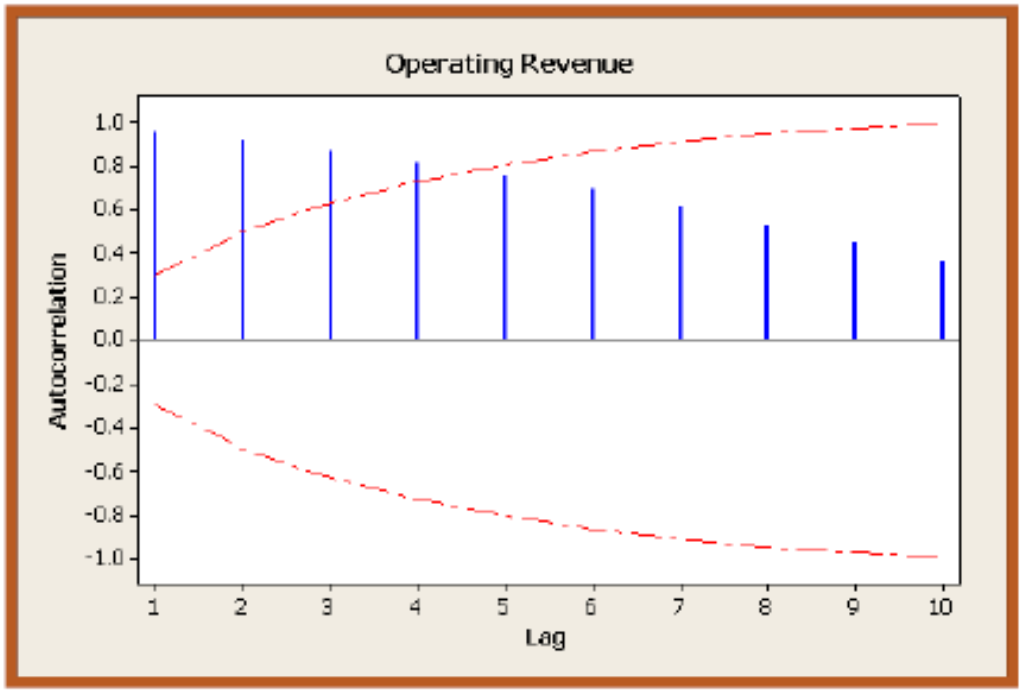
- Example:
- An analyst for Sears company is assigned the task of forecasting operating revenue for 2001. She gathers the data for the years 1955 to 2000.



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

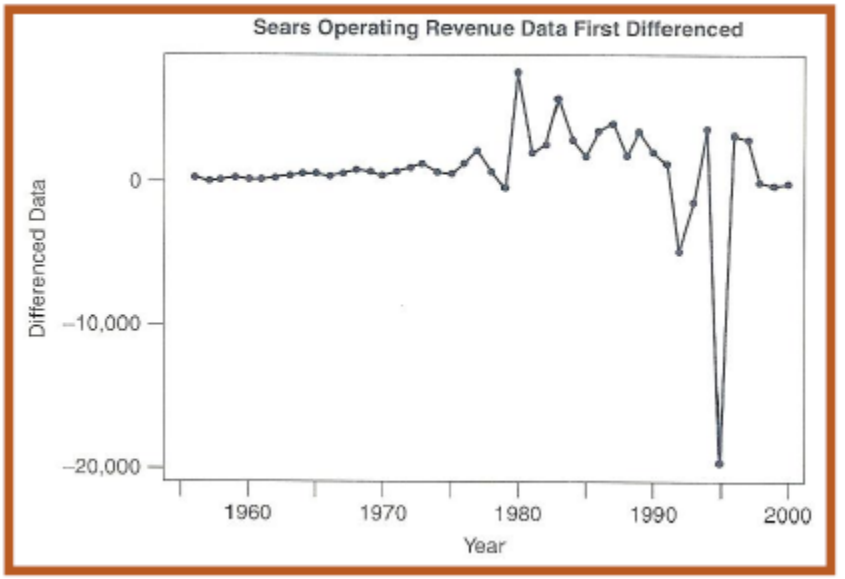
- Time lags are significantly different from zero (0.96, 0.92 and 0.87) and then the values gradually drop to zero.

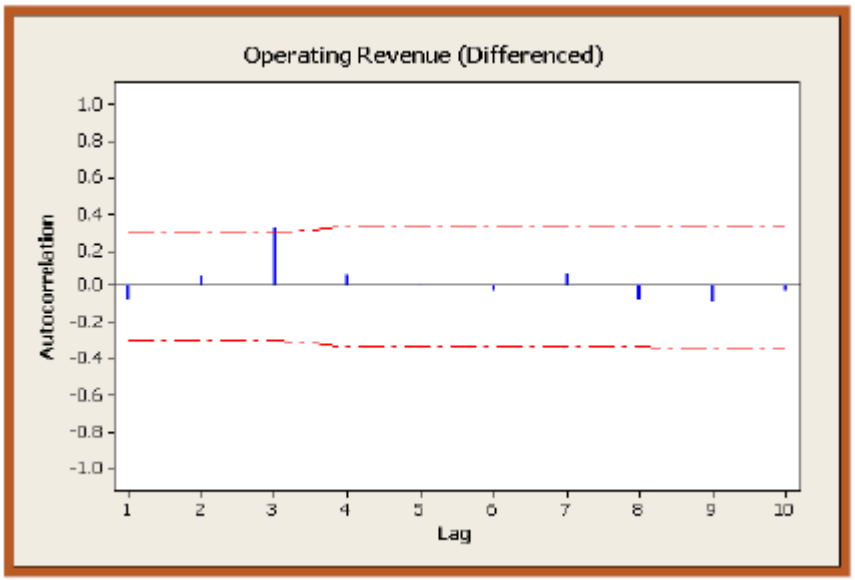


Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- The date series were differenced to remove the trend and to create a stationary series.
- The differenced series shows no evidence of a trend.
- The autocorrelation coefficient at time lag 3 (0.32) is significantly different from zero.
- The autocorrelations at lags other than lag 3 are small.





Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

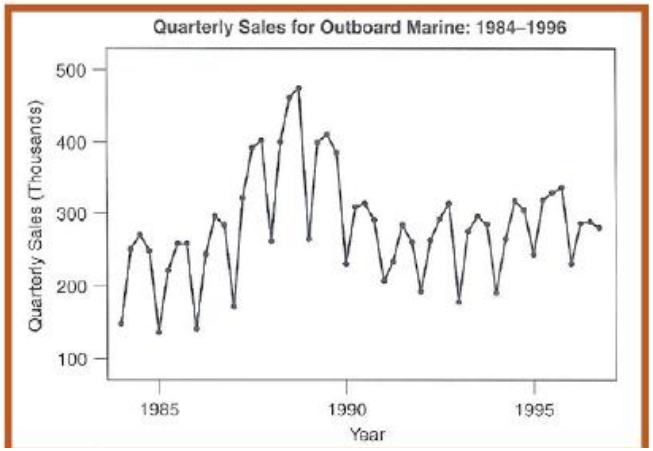
4. Seasonal data.

- The time series is seasonal if significant autocorrelation coefficient will occur at a seasonal time lag or multiple of the seasonal lag.
- For example, for the quarterly seasonal data, a significant autocorrelation coefficient will appear at lag 4, for the monthly seasonal data, a significant autocorrelation coefficient will appear at lag 12, and so forth.
- The time series data to be analyzed the seasonal component should be removed from the data before modeling.
- Different techniques will be studied in future helps in removing the seasonal components.

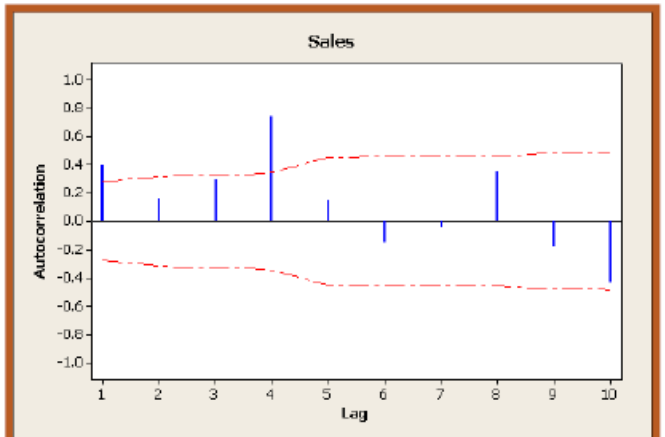
Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Example:
- An analyst for Outboard Marine Corporation always felt that sales were seasonal. He gathers the data for the quarterly sales of Outboard Marine Corporation from 1984 to 1996.



  - By observing the time series plot, he noticed a seasonal pattern
  - He computes the ACF.
  - He notes that the autocorrelation coefficients at time lags 1 and 4 are significantly different from zero
  - He concludes that Outboard Marine sales are seasonal on a quarterly basis



Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Measuring Forecasting Error

- Suppose Y_t be the actual value of a time series at time t and $\hat{Y}_t$ be the forecast value of a time series at time t where t = 1,2,3,…n.
- Then the difference between the actual value and its forecast value is called the residual or forecast error and usually denoted by e_t such that

$$e_t = Y_t - \hat{Y}_t$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Types of Forecast Accuracy Measures

- 1. Mean Absolute Deviation (MAD)
  - It is useful when the analyst wants to measure forecast error in the same units as the original series.

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |Yi - \hat{Y}_i|$$

- 2. Mean Squared Error (MSE)
  - It is useful because it penalizes large forecasting error and therefore the method with moderate errors is more preferable than the method of small errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- 3. Mean Absolute Percentage Error (MAPE)
  - It is useful when the size or magnitude of the forecast variable is important in evaluating the accuracy of forecast and useful when the actual values of a time series are large.
  - MAPE provides an indication of how large the forecast errors are in comparison to the actual values of the series
  - Also it can be used to compare the accuracy of the same or different techniques on two entirely different series.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i}$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- 4. Mean Percentage Error (MPE)
  - It is useful when the analyst wants to determine whether a forecasting method is biased (consistently forecasting low or high).
  - Therefore;
    - If MPE is very close to zero then the forecasting method is unbiased.
    - If MPE is large negative percentage then the forecasting method is consistently overestimating.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)}{Y_i}$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- In general, the above four measures of forecast accuracy are usually used as follows:
  - To compare the accuracy of two or more different techniques.
  - To measure the usefulness and the reliability of a particular technique.
  - To help search for an optimal technique.

Determining the Accuracy of a Forecasting Technique

- To evaluate the adequacy of the forecasting technique, we should check the following:
  - Randomness of the residuals → Use the autocorrelation function for the residuals.
  - Normality of the residuals → Use the histogram or the normal probability plot for the residuals.
  - Significance of parameter estimates → Use the t test for all parameter estimates.
  - Simplicity and understandability of the technique for decision makers.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

- Example:
- Data: The daily number of customers requiring repair work, Y, and a forecast of these data, Y_t, for Gary's Chevron Station.
- The forecasting technique used the number of customers serviced in the previous period as the forecast for the current period.

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Time t	Customers Y_t	Forecast $\hat{Y}_t$	Error e_t	$\lvert e_t \rvert$	e_t^2	e_t / Y_t	$\lvert e_t \rvert / Y_t$
1	58	—	—	—	—	—	—
2	54	58	−4	4	16	−.074	.074
3	60	54	6	6	36	.100	.100
4	55	60	−5	5	25	−.091	.091
5	62	55	7	7	49	.113	.113
6	62	62	0	0	0	.000	.000
7	65	62	3	3	9	.046	.046
8	63	65	−2	2	4	−.032	.032
9	70	63	7	7	49	.100	.100
			—	—	—	—	—
		Totals	12	34	188	.162	.556

$$MAD = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t| = \frac{34}{8} = 4.3$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{Y_t} = \frac{.556}{8} = .0695\ (6.95\%)$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 = \frac{188}{8} = 23.5$$

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(Y_t - \hat{Y}_t)}{Y_t} = \frac{.162}{8} = .0203\ (2.03\%)$$

Exploring Data Patterns & Choosing a Forecasting Technique

ÖZYEĞİN UNIVERSITY

Remarks on Empirical Evaluation of Forecasting Methods

- Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler methods.
- Various accuracy measures (MAD, MSE, MAPE, and MPE) produce consistent results when used to evaluate different forecasting methods.
- Combining the three smoothing methods on the average does well in comparison with other methods.
- The performance of the various forecasting methods depends on the length of the horizon and the kind of the data analyzed.