Exploring Data Patterns and An Introduction to Forecasting Techniques

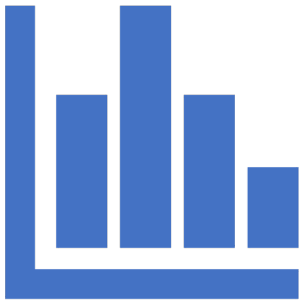
Dr. Khaing S Htun

Applications of Forecasting

- **Sales Forecasting**
- **Forecasting Economic Trends**
- **Forecasting Staffing Needs**
- **Forecasting in education environment**
- **Forecasting in a rural setting**
- **Ministry of Petroleum**
- **Department of Technology**

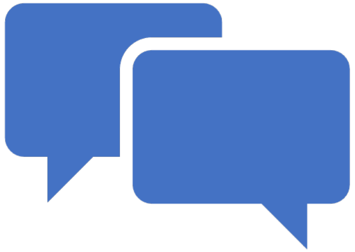
... and in many business and social science-related situations

Forecasting approach



Quantitative

Historical data from
time-series or
correlation information



Qualitative

Opinions from experts,
decision makers, or
customers

Types of data

Cross-sectional data

collected at a single point in time

Company	Exchange	Annual Sales ($ millions)	Earnings per Share ($)
Advanced Comm. Systems	OTC	75.10	0.32
Ag-Chem Equipment Co.	OTC	321.10	0.48
Aztec Manufacturing Co.	NYSE	79.70	1.18
Cal-Maine Foods, Inc.	OTC	314.10	0.38
Chesapeake Utilities	NYSE	174.50	1.13
Dataram Corporation	AMEX	73.10	0.86
Energy South, Inc.	OTC	74.00	1.67
Gencor Industries, Inc.	AMEX	263.30	1.96
Industrial Scientific	OTC	43.50	2.03
Keystone Consolodated	NYSE	365.70	0.86

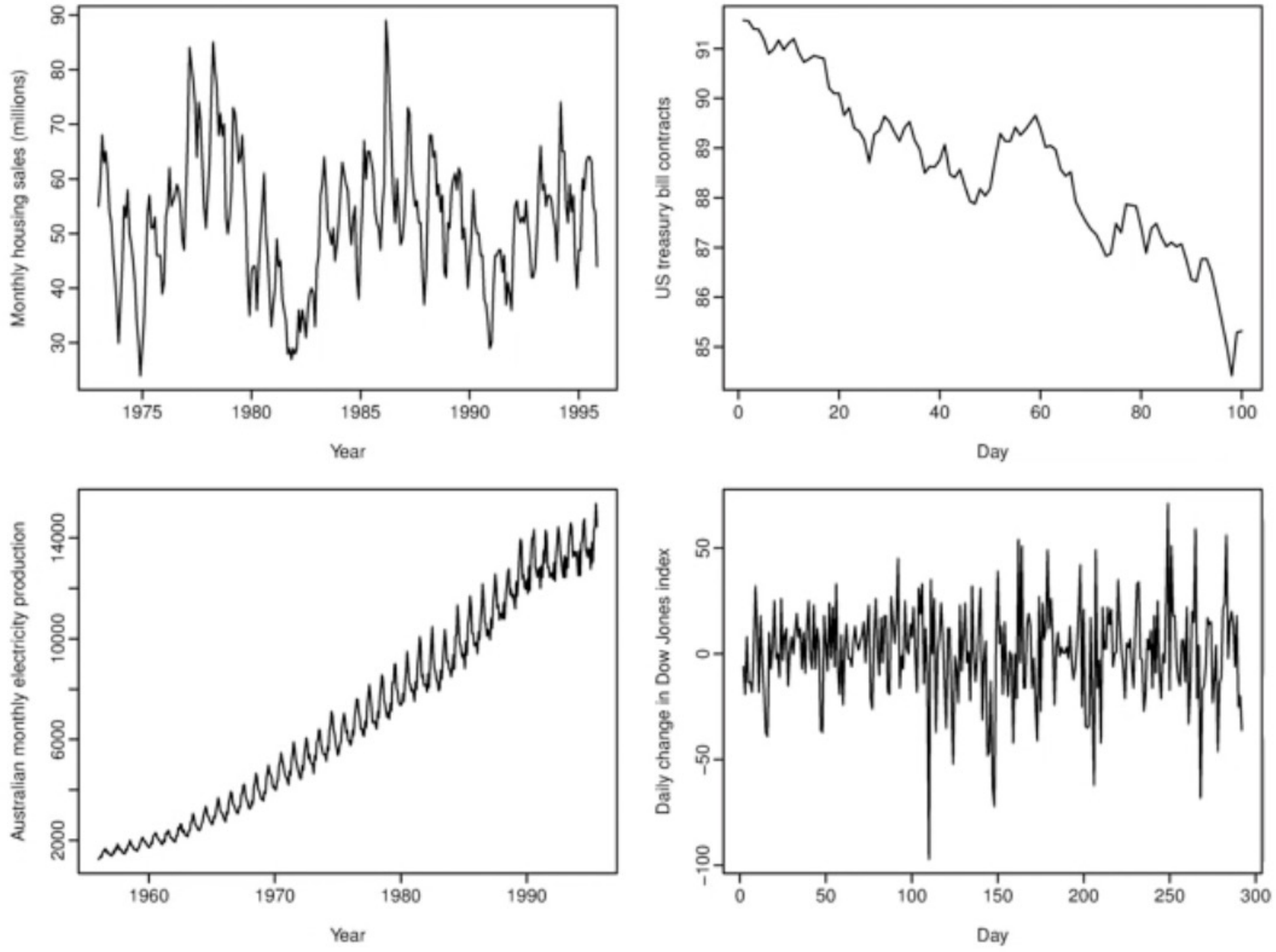
Cost ($)	859	682	471	708	1094	224	320	651	1049
Age (years)	8	5	3	9	11	2	1	8	12

A time series data

collected at regular intervals over time

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

What is Time Series?



Time Series

An orderd squence of values of a variable at equally spaced time intervals.

- yearly, quarterly, monthly, weekly, daily or even hourly

- **Discover the pattern**
- Extrapolate the pattern in the future

Time Series Plot

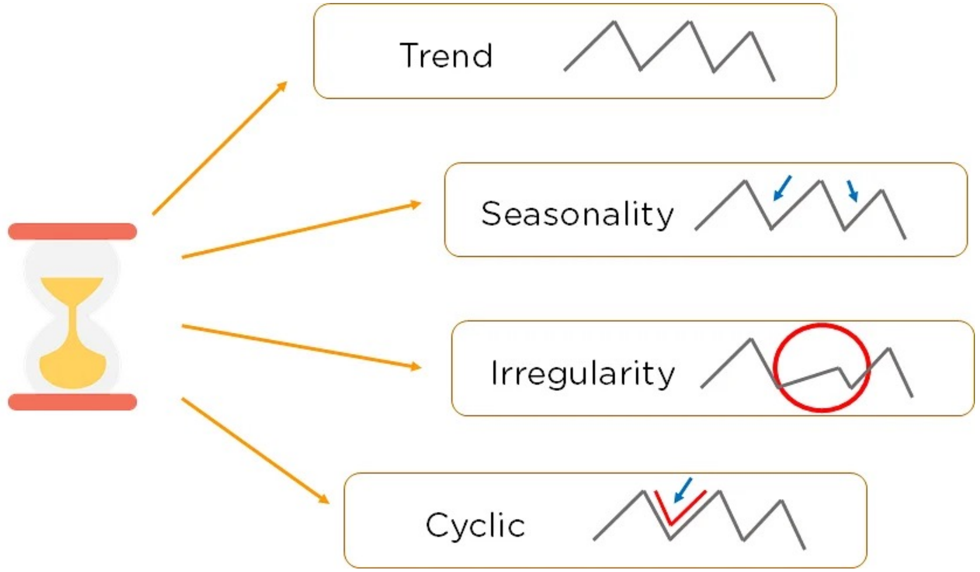
Time Series Plot

relationship between *time and the time series variable*

- Time as independent variable.
- Time series values as dependent variable

Time Series Patterns

- Horizontal pattern
- **Trend** pattern
- **Seasonal** pattern
- **Cycles** pattern
- **Irregular** or Random variations
- *Combination patterns – Trend + Seasonal*



Horizontal pattern

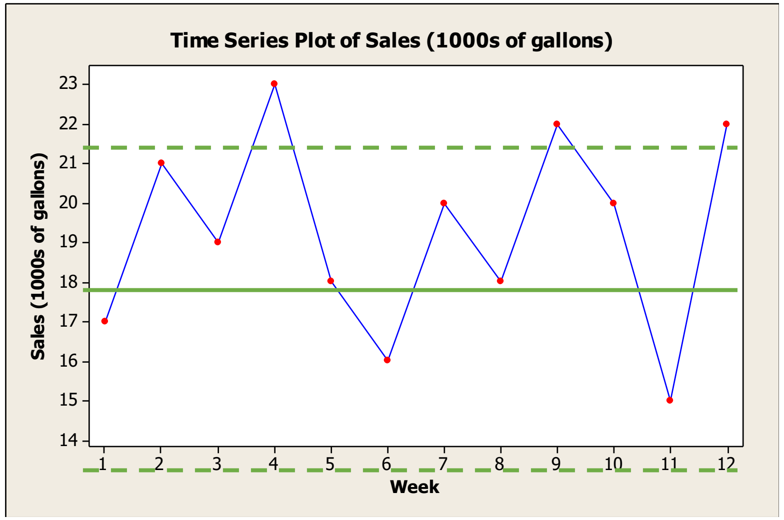
Week	Sales (1000s of gallons)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22

A **horizontal pattern** exists when **the data fluctuate randomly around a constant mean** over time.

Stationary time series

1. Constant mean
2. Constant variability of time series

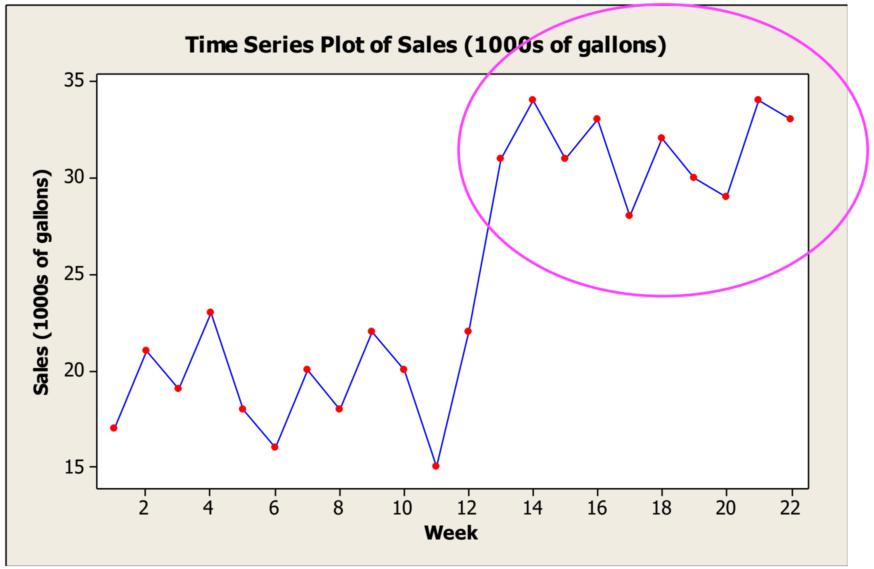
- stationary time series will always exhibit a horizontal pattern
- horizontal pattern is not sufficient evidence to conclude that the time series is stationary



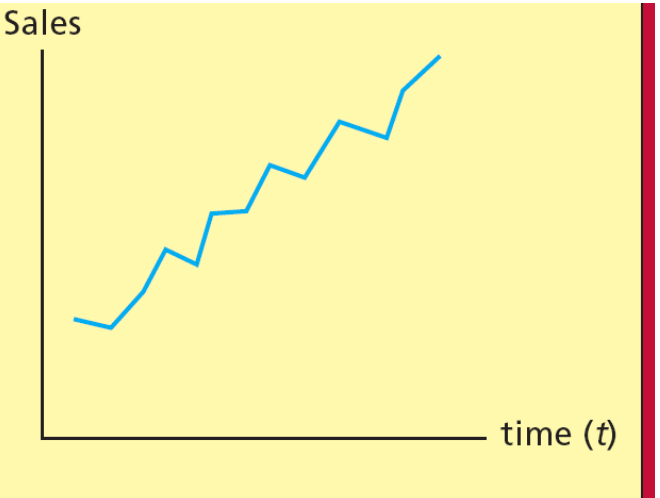
Shifted horizontal pattern

Horizontal pattern shifts to a new level due to changes in business conditions

Week	Sales (1000s of gallons)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22
13	31
14	34
15	31
16	33
17	28
18	32
19	30
20	29
21	34
22	33

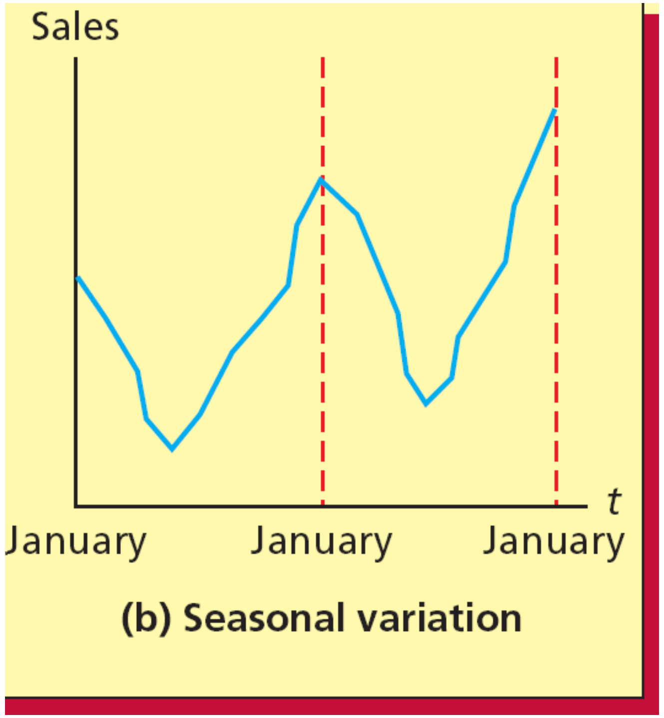


Trend pattern



(a) Trend

- represents a gradual shifting to higher or lower over time
- result of changes in long-term factors
- long-run growth or decline.
  - can be linear or non-linear

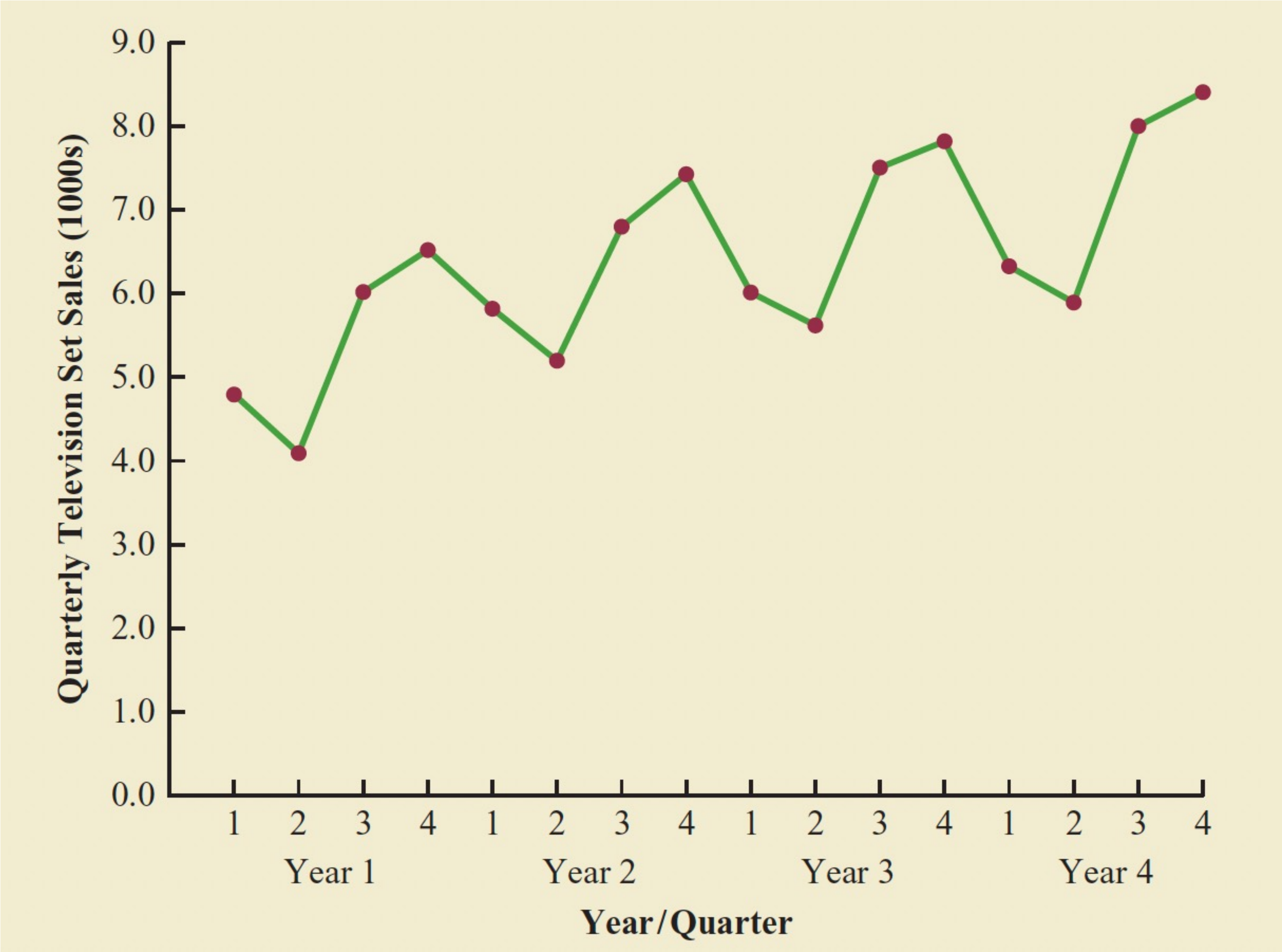


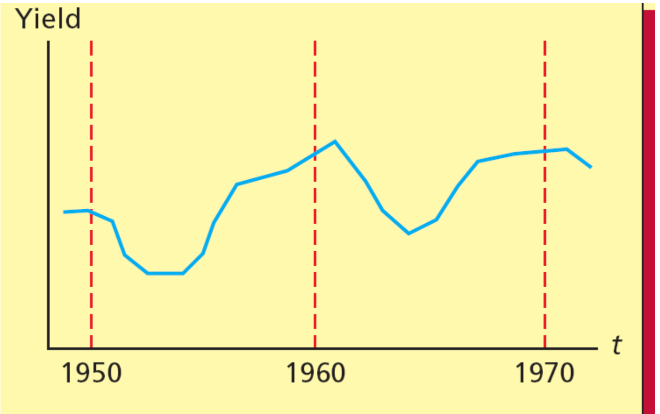
Seasonal pattern

- any repeating pattern, less than one year in duration
- short duration pattern
- regular periodic up and down movements that repeat within the calendar year.

Trend and Seasonal pattern

Combination pattern

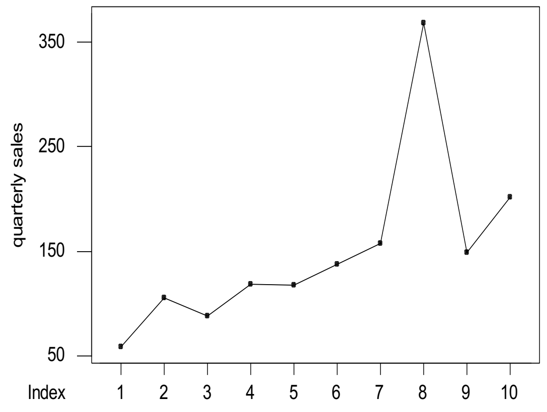




(c) Cycle

Cyclical pattern

- any recurring sequence of points above and below the trend line
- represents multiyear cyclical movements in the economy
- long-run up and down fluctuation around the trend level
- extremely difficult to forecast



Irregular pattern

- random variation
- caused by unforeseeable circumstances
  - nonrecurring factors
  - very short-run movements
  - follow no regular pattern

Time Series Models

Quantitative Methods

Time-Series Models

Time series models look at past patterns of data and attempt to predict the future based upon the underlying patterns contained within those data.

Associative Models

Associative models (often called causal models) assume that the variable being forecasted is related to other variables in the environment. They try to project based upon those associations.

Time Series Models

Stationary models

- Naïve
- Simple Moving Average
- Weighted Moving Average
- Exponential Smoothing (simple and double)

Trend Projection/ Time Series Regression models

- Linear trend regression
- Non-linear trend regression

Seasonality

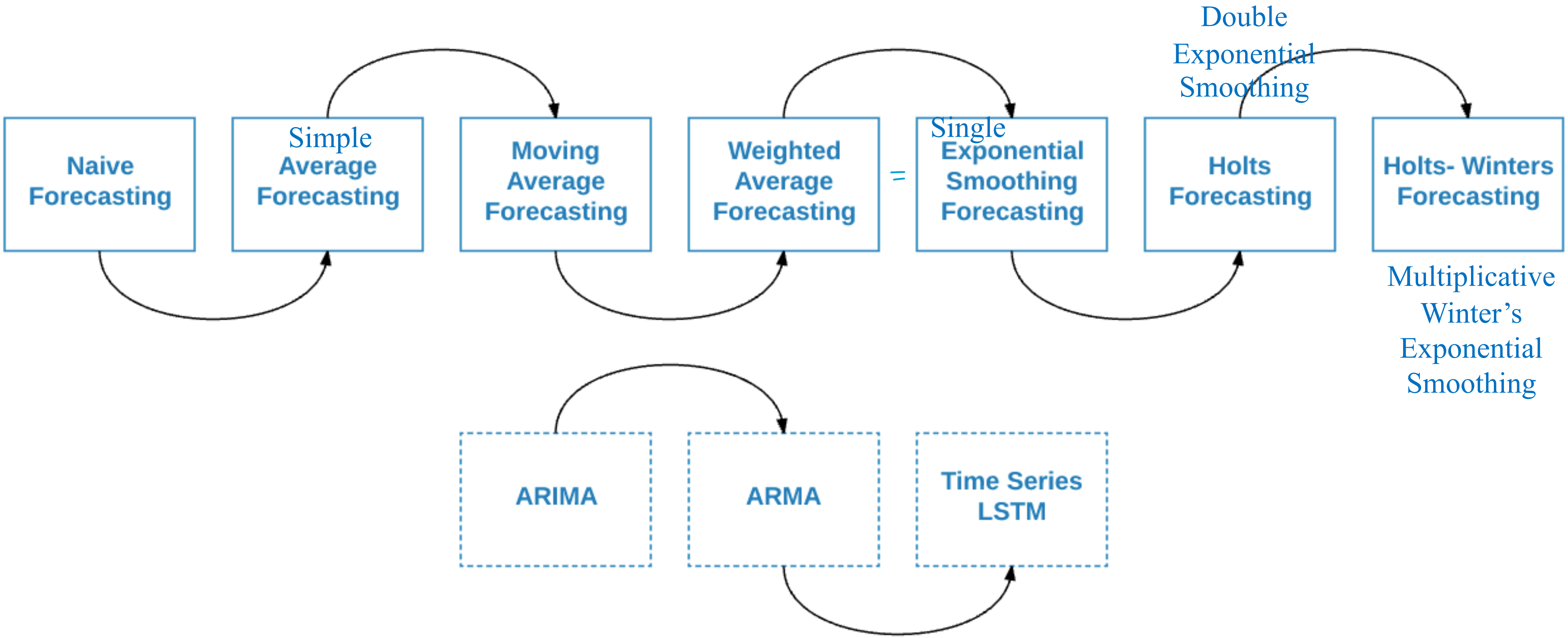
- Without trend
- With trend

Multiplicative decomposition models

- Deseasonalizing the Time Series
- Seasonal index – ratio between actual and average demand

Some of the Time Series Models

Model	Description
Naïve	Uses last period's actual value as a forecast
Simple Average	Uses an average of all past data as a forecast
Moving Average	Uses an average of a specified number of the most recent observations, with each observation receiving the same emphasis (weight)
Weighted Moving Average	Uses an average of a specified number of the most recent observations, with each observation receiving a different emphasis (weight)
Exponential Smoothing	A weighted average procedure with weights declining exponentially as data become older
Trend Projection	Technique that uses the least squares method to fit a straight line to the data
Seasonal Indexes	A mechanism for adjusting the forecast to accommodate any seasonal patterns inherent in the data

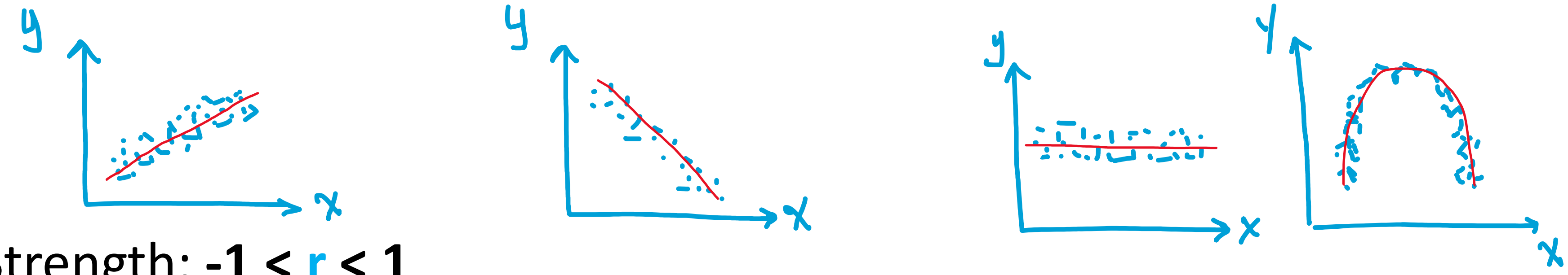


Naive Forecasting
Simple
Average Forecasting
Moving Average Forecasting
Weighted Average Forecasting
=
Single
Exponential Smoothing Forecasting
Double Exponential Smoothing
Holts Forecasting
Holts- Winters Forecasting
Multiplicative Winter's Exponential Smoothing
ARIMA
ARMA
Time Series LSTM

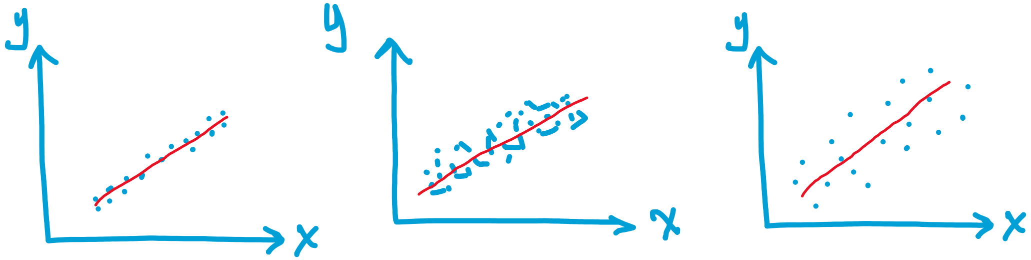
Exploring Data Patterns with Autocorrelation Analysis

Correlation - ***Recall***

- Direction: **Positive, Negative, No correlation**



- Strength: **-1 < r < 1**



Autocorrelation is the correlation between two observations at different points in a time series.

Autocorrelation

- Observations in different time periods are frequently related or correlated. This correlation is measured using the **autocorrelation coefficient**.
- **Autocorrelation** is the correlation between a variable lagged one or more periods and itself.
- Autocorrelation coefficients for different time lags of a variable are used **to identify time series data patterns**.
- The lag *k* autocorrelation coefficients ($\mathbf{r_k}$) between observations Y_t and Y_{t-k}, which are *k* periods apart is

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \quad ; \quad k = 0, 1, 2, \cdots \qquad -1 \le r_k \le 1$$

k, increases, r_k decrease

r_k = autocorrelation coefficient for a lag of k periods
$\bar{Y}$ = mean of the values of the series
Y_t = observation in time period t
Y_{t-k} = observation k time periods earlier or at time period $t-k$

Autocorrelation *(example)*

Harry Vernon has collected data on the number of VCRs sold for Vernon's Music Store in 2007. The data are presented in Table.

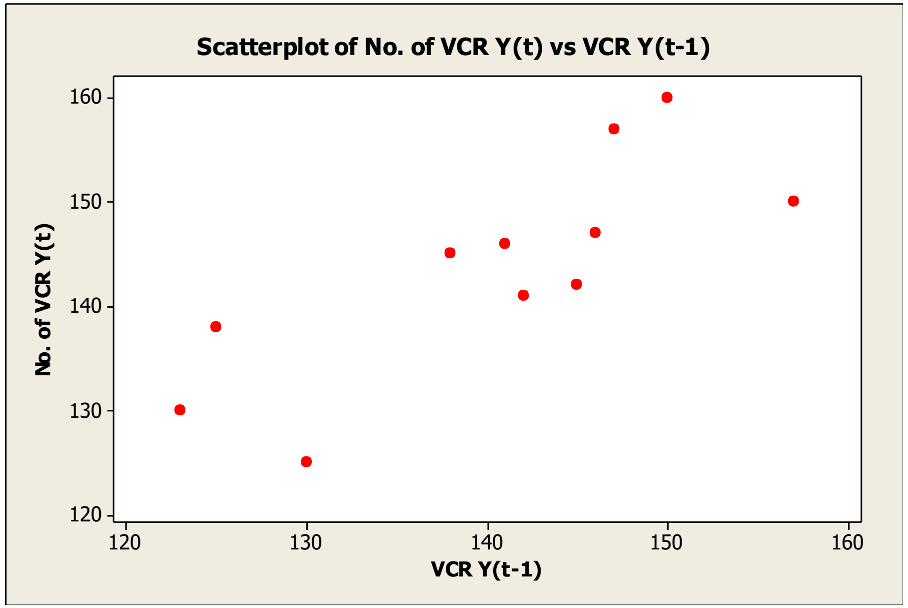
a) Find the value of Y_{t-1} and Y_{t-2} and develop scatter diagram of (Y, Y_{t-1})

b) Compute the lag 1 autocorrelation coefficient (r_1) and the second-order autocorrelation coefficient (r_2)

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

Autocorrelation *(example)*

Time, t	Month	VCRs sold, Yt	Y lagged 1 period, Y(t-1)	Y lagged 2 period, 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



Autocorrelation *(example)*

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

Time, t	Month	VCRs sold, Yt	Y(t-1)	Yt - $\bar{Y}$	Yt-1 - $\bar{Y}$	(Yt - $\bar{Y}$)^2	(Yt - $\bar{Y}$)(Yt-1 - $\bar{Y}$)
1	January	123	-----	–19	-----	261	-----
2	February	130	123	–12	–19	144	228
3	March	125	130	–17	–12	289	204
4	April	138	125	– 4	–17	16	68
5	May	145	138	3	– 4	9	–12
6	June	142	145	0	3	0	0
7	July	141	142	–1	0	1	0
8	August	146	141	4	–1	16	– 4
9	September	147	146	5	4	25	20
10	October	157	147	15	5	225	75
11	November	150	157	8	15	64	120
12	December	160	150	18	8	324	144
		$\bar{Y}$ = 142				**1,474**	**843**

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

$$= \frac{\sum_{t=2}^{12} (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{12} (Y_t - \bar{Y})^2}$$

$$= \frac{843}{1,474} = 0.572$$

Autocorrelation *(example)*

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

Time, t	Month	VCRs sold, Yt	Y(t-2)	Yt - $\bar{Y}$	Yt-2 - $\bar{Y}$	(Yt - $\bar{Y}$)^2	(Yt - $\bar{Y}$)(Yt-2 - $\bar{Y}$)
1	January	123	-----	–19	-----	261	-----
2	February	130	-----	–12	-----	144	-----
3	March	125	123	–17	–19	289	323
4	April	138	130	– 4	–12	16	48
5	May	145	125	3	–17	9	–51
6	June	142	138	0	–4	0	0
7	July	141	145	–1	3	1	–3
8	August	146	142	4	0	16	0
9	September	147	141	5	–1	25	–5
10	October	157	146	15	4	225	60
11	November	150	147	8	5	64	40
12	December	160	157	18	15	324	270
		$\bar{Y}$ = 142				**1,474**	**682**

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

$$= \frac{\sum_{t=3}^{12}(Y_t - \bar{Y})(Y_{t-2} - \bar{Y})}{\sum_{t=1}^{12}(Y_t - \bar{Y})^2}$$

$$= \frac{682}{1,474} = 0.463$$

$r_1 > r_2$

k, increases, r_k decrease

Autocorrelation *(example)* *extra* Y_{t-3}

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

Time, t	Month	VCRs sold, Yt	Y(t-3)	Yt - $\bar{Y}$	Yt-3 - $\bar{Y}$	(Yt - $\bar{Y}$)^2	(Yt - $\bar{Y}$)(Yt-3 - $\bar{Y}$)
1	January	123	-----	–19	-----	261	-----
2	February	130	-----	–12	-----	144	-----
3	March	125	-----	–17	-----	289	-----
4	April	138	123	– 4	–19	16	76
5	May	145	130	3	–12	9	-36
6	June	142	125	0	–17	0	0
7	July	141	138	–1	–4	1	4
8	August	146	145	4	3	16	12
9	September	147	142	5	0	25	0
10	October	157	141	15	–1	225	-15
11	November	150	146	8	4	64	32
12	December	160	147	18	5	324	90
						1,474	**163**

$\bar{Y} = 142$

$$r_3 = \frac{163}{1474} = 0.112$$

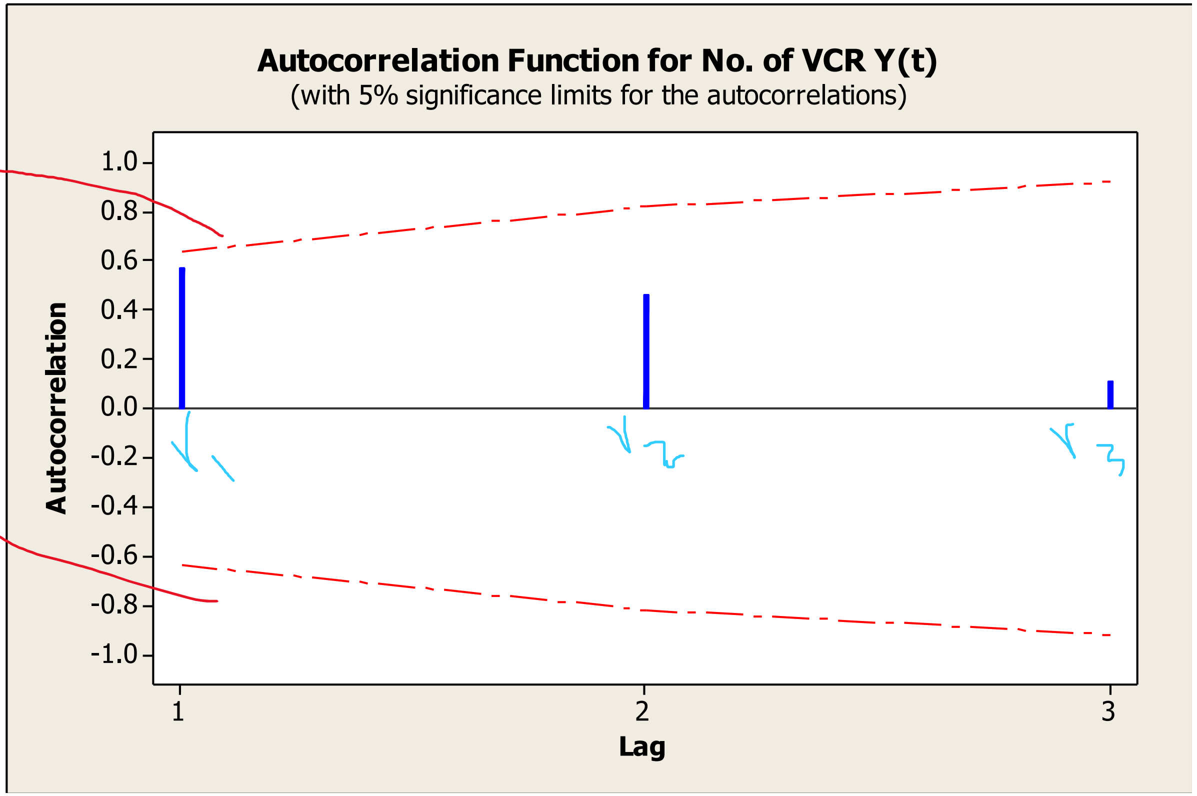
$r_2 > r_3$

k, increases, r_k decrease

Autocorrelation *(example)*

Correlogram or Autocorrelation Function for the VCR Data

95% confidence



A **correlogram** - visual way to show serial correlation in data that changes over time

Autocorrelation Function: No. of VCR Y(t)

Lag	ACF	T	LBQ
1	0.571913	1.98	5.00
2	0.462687	1.25	8.59
3	0.110583	0.27	8.82

Characteristic's Correlogram or Autocorrelation function

- The ***Correlogram*** *or autocorrelation function* is a **graph of the autocorrelations** for various lags of a time series.
- The horizontal scale on the bottom of the graph shows each time lag of interest, 1, 2, 3, and so on.
- The vertical scale on the left shows the possible range of an autocorrelation coefficient, - 1 to + 1.
- The middle of graph is zero.
- The dotted lines are 95% confidence limits of the autocorrelation function.
- The T is the value of test statistics for testing for autocorrelation at the various lags.
- The LBQ is the Ljung-Box Q statistic for testing whether data are correlated at any time lag or there is autocorrelation at any time lag.

Autocorrelation function (ACF)

ACF answers the following questions-

- Are the data random?
- Do the data have a trend (are they nonstationary)?
- Are the data stationary?
- Are the data seasonal?

Random data

- Autocorrelation between Y_t and Y_{t-k} for any lag are close to zero
- The successive values of a time series are not related to each other
- All the sample autocorrelation coefficient should lie within a range as:

$$\boxed{0 \pm (t \times SE(r_k))}$$

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

where
$SE(r_k)$ = the standard error of the autocorrelation at time lag
r_i = the autocorrelation at time lag *i*
k = the time lag
n = the number of observations in the time series

Random data

Hypothesis testing

$H_0: \rho_i = 0$

$H_1: \rho_i \neq 0$

Test statistic is

$$t_i = \frac{r_i - \rho_i}{SE(r_i)}$$

H_0 is rejected at α and d.f.=n-1,

if $\boldsymbol{t_i} \leq -\mathbf{t}_{\alpha/2}$ or $\mathbf{t} \geq +\mathbf{t}_{\alpha/2}$

The other common test is the modified Box-Pierce Q statistic which is given

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

H_0 is rejected at α and d.f.=m-a,

if $\boldsymbol{Q \geq \chi^2_\alpha}$

where
m = the number of time lags to be tested
a = the number of the estimated parameters in model

Trend data (nonstationary)

- If the series has a trend, Y_t and Y_{t-k} are highly correlated
- The autocorrelation coefficient are significantly different from zero and will decline toward zero slowly
- A series that contains a trend is said to be **non-stationary**.

Stationary data

- Constant mean (no trend)
- Constant variance
- Constant autocorrelation structure
- No periodic component (no seasonality)
- A series that varies about a fixed level (no growth or decline) over time
- The autocorrelation coefficient for a stationary series decline to zero rapidly, generally after the second- or third-time lag

Stationary data

- To analyze nonstationary series, the trend is removed before additional modeling occurs.
- **Data differencing method** - to remove the trend from a nonstationary series.

Seasonal data

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.

Choosing a Forecasting Technique

- Why is a forecast needed?
- Who will use the forecast?
- What are the characteristics of the available data?
- What time period is to be forecast?
- What are the minimum data requirements?
- How much accuracy is desired?
- What will the forecast cost?

Choosing a Forecasting Technique

- Define the nature of the forecasting problem.
- Explain the nature of the data under investigation.
- Describe the capabilities and limitations of potentially useful forecasting techniques.
- Develop some predetermined criteria on which the selection decision can be made.

Choosing a Forecasting Technique

Pattern of Data:

- ST - stationary;
- T - trend;
- S - seasonal;
- C - cyclical

Time Horizon:

- S - short term (less than 3 months);
- I - intermediate term;
- L - long term

Type of Model:

- TS - time series;
- C - casual

Seasonal:

- S - length of seasonality

Variable:

- V - number of variables

Method	Pattern of Data	Time Horizon	Type of Model	Minimal Data Requirements	
				Non-seasonal	Seasonal
Naïve	ST, T, S	S	TS	1	
Simple averages	ST	S	TS	30	
Moving averages	ST	S	TS	4-20	
Exponential smoothing	ST	S	TS	2	
Linear exponential smoothing	T	S	TS	3	
Quadratic exponential smoothing	T	S	TS	4	
Seasonal exponential smoothing	S	S	TS		2 × s
Adaptive filtering	S	S	TS		5 × s
Simple regression	T	I	C	10	
Multiple regression	C, S	I	C	10 × V	
Classical decomposition	S	S	TS		5 × s
Exponential trend models	T	I , L	TS	10	
S-curve fitting	T	I , L	TS	10	
Gompertz models	T	I , L	TS	10	
Growth cirves	T	I , L	TS	10	
Census X-12	S	S	TS		6 × s
Box-Jenkins	ST, T, C, S	S	TS	24	3 × s
Leading indicators	C	S	C	24	
Econometric models	C	S	C	30	
Time series multiple regression	T, S	I, L	C		6 × s

Evaluation of forecasting model

- BAIS – the arithmetic mean of the errors
- MAD – mean absolute deviation
- MSE – mean square error
  - Standard error - $\sqrt{MSE}$
- MAPE – mean absolute percent w=error

Forecast Accuracy

Forecast error = Actual Value - Forecast

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

A residual/error is the difference between an actual observed value and its forecast value.

where
e_t = the forecast error in time period t
Y_t= the actual value in time period t
$\hat{Y}_t$= the forecast value in time period t

Measuring Forecasting Error and Forecast Error Comparison

1. The mean absolute deviation:

MAE $$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|Y_t - \hat{Y}_t\right|$$

2. The mean square deviation:

MSE $$MSD = \frac{1}{n}\sum_{i=1}^{n}\left(Y_t - \hat{Y}_t\right)^2$$

3. The mean absolute percentage error:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_t - \hat{Y}_t\right|}{Y_t}$$

4. The mean percentage error:

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_t - \hat{Y}_t\right)}{Y_t}$$

Determining the Adequacy of a Forecasting Technique

- Are the autocorrelation coefficients of the residuals indicative of a random series?
- Are the residuals approximately normally distributed?
- Do all parameter estimates have significant *t* ratios?
- Does the technique simple to use, and can planners and policy makers understand it?

Time series forcasting process

Look at the data (Scatter Plot) → Forecast using one or more techniques → Evaluate the technique and pick the best one.

Observations from the scatter Plot	Techniques to try	Ways to evaluate
Data is reasonably **stationary** (no trend or seasonality)	**Heuristics - Averaging methods** • Naive • Moving Averages • Simple Exponential Smoothing	• MAD • MAPE • Standard Error • BIAS
Data shows a consistent **trend**	**Regression** • Linear • Non-linear Regressions (not covered in this course)	• MAD • MAPE • Standard Error • BIAS • R-Squared
Data shows both a **trend** and a **seasonal pattern**	**Classical decomposition** • Find Seasonal Index • Use regression analyses to find the trend component	• MAD • MAPE • Standard Error • BIAS • R-Squared