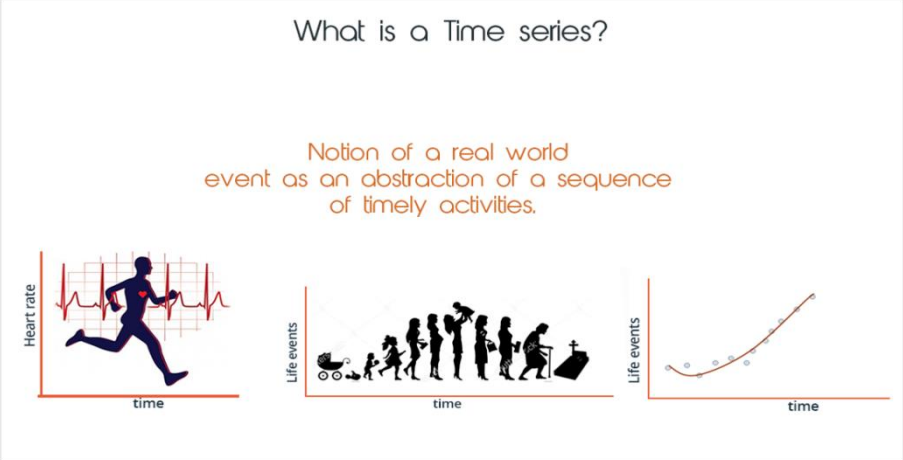
MAT 3103: Computational Statistics and Probability

Chapter 8: Time Series Analysis



Time series is a sequence of data points recorded in time order, often taken at successive equally paced points in time. Data are taken yearly, monthly, weekly, hourly or even by the minute. Stock prices, Sales demand, website traffic, daily temperatures, quarterly sales etc. are some examples.

Why is Time series analysis important?

As the world is becoming more technology-oriented, the collection of data dependent on time is becoming very easy. The proper analysis and forecast of this data can give more valuable and insightful results to enhance various domains including data science and machine learning.

- Time series analysis can be useful to see how a given asset, security, or economic variable changes over time.
- Forecasting methods using time series are used in both fundamental and technical analysis.

Terms and concepts:

Dependence: Dependence refers to the association of two observations with the same variable, at prior time points.

Stationarity: Shows the mean value of the series that remains constant over a time period; if past effects accumulate and the values increase toward infinity, then stationarity is not met.

Differencing: Used to make the series stationary, to De-trend, and to control the auto correlations; however, some time series analyses do not require differencing and over-differenced series can produce inaccurate estimates.

Components of a Time Series

1. **Trend**: is a general direction in which something is developing or changing. A trend can be upward (uptrend) or downward (downtrend). It is not always necessary that the increase or decrease is consistently in the same direction in a given period.
2. **Seasonality**: Predictable pattern that recurs or repeats over regular intervals. Seasonality is often observed within a year or less.
3. **Cycles**: Occur when a series follows an up and down pattern that is not seasonal. Cyclical variations are periodic in nature and repeat themselves like business cycle, which has **four** phases (1) Peak, (2) Recession, (3) Trough/Depression, and (4) Expansion. Seasonality is different from cycles, as seasonal cycles are observed within a calendar year, while cyclical effects can span duration shorter or longer than a calendar year.
4. **Irregular fluctuation**: Variations that occur due to sudden causes and are unpredictable. For example, the rise in prices of food due to war, flood, earthquakes, farmers striking etc.

When not to use a time series analysis?

- When the values are constant, that is, they are not dependent on time [the data is not a time series data and it is pointless as the values never change].
- Values in the form of a function like $\sin(x)$, $\cos(x)$ etc. It is, again, pointless to use time series analysis as you calculate the values using a function.

Time Series and Forecasting

The long-term tendency of the data to move in an upward or downward direction can be measured by the following methods:

1. Graphical Method

The values of a time series are plotted on a graph paper by taking time variable on the *X*-axis and the values variable on the *Y*-axis. After this, a smooth curve is drawn with free hand through the plotted points. The trend line drawn above can be extended to forecast the values.

Advantages

- It is very easy and simple.
- No mathematical calculations are needed.
- It can be used even if the trend is not linear.

Disadvantages

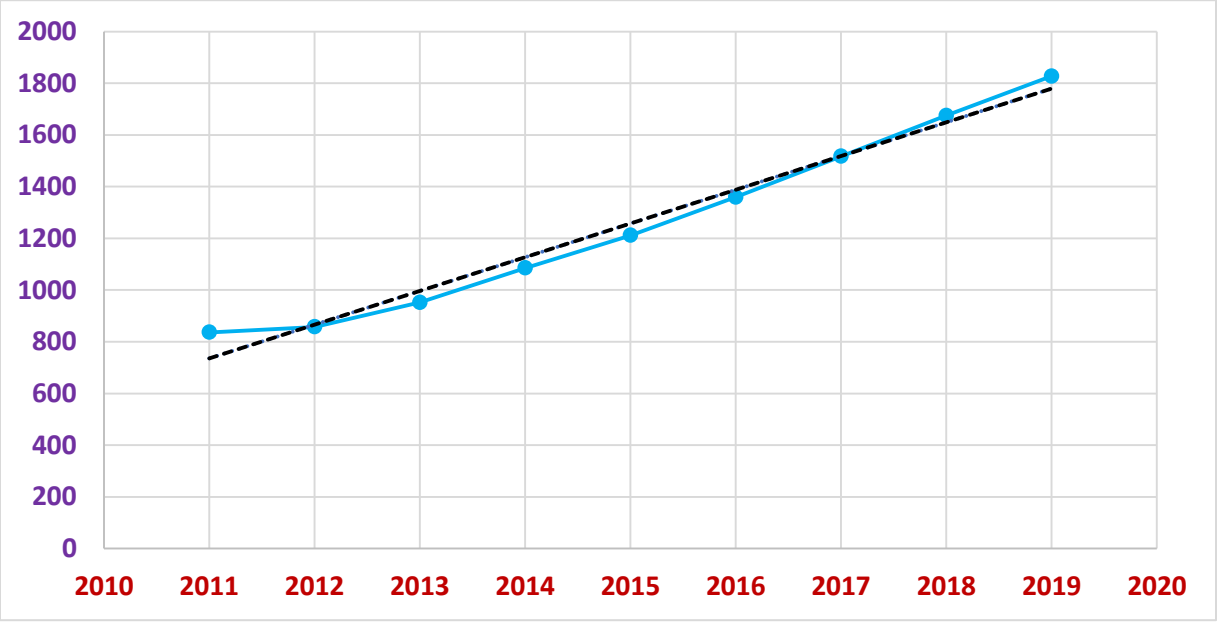
- It is a subjective method.
- Trend values obtained by different statisticians would be different and hence not reliable.

Example 8.1

GDP per capita (in USD) in Bangladesh is listed below:

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
GDP	836	857	952	1085	1211	1359	1517	1625	1827

Fit a trend line applying graphical method to forecast future GDPs.



2. Semi-Average Method

The series is divided into two equal parts and the average of each part is plotted at the mid-point of their time duration.

- In case the series consists of an **even** number of years, the series is divisible into two halves. Find the average of the two parts of the series and place these values in the mid-year of each of the respective durations.
- In case the series consists of **odd** number of years, it is not possible to divide the series into two equal halves. The middle year will be omitted. After dividing the data into two parts, find the arithmetic mean of each part. Thus, we get semi-averages.
- The trend values for other years can be computed by successive addition or subtraction for each year ahead or behind any year.

Advantages

- This method is very simple and easy to understand.
- It does not require many calculations.

Disadvantages

- This method is used only when the trend is linear.
- It is used for calculation of averages and they are affected by extreme values.

Example 8.2

Calculate the trend values using semi-averages methods for the income from the forest department.

Year	2008	2009	2010	2011	2012	2013
Income (in crores)	46.17	51.65	63.81	70.99	84.91	91.64

Solution:

Year	**Income**	**3-year semi total**	**3-year semi average**	**Trend values**
2008	46.17	161.63	53.877	53.877 - 9.545 = 44.332
2009	51.65			44.332 + 9.545 = 53.877
2010	63.81			53.877 + 9.545 = 63.422
2011	70.99	247.54	82.513	63.422 + 9.545 = 72.967
2012	84.91			72.967 + 9.545 = 82.512
2013	91.64			82.512 + 9.545 = 92.057

Difference between the central years = 2012 – 2009 = 3

Difference between the semi-averages = 82.513 – 53.877 = 28.636

Increase in trend value for one year = 28.636 /3 = 9.545

Example 8.3

Calculate the trend values using semi-averages methods for the income from the forest department.

Year	1951	1961	1971	1981	1991	2001	2011
Income (in crores)	301.2	336.9	412	484.1	558.6	624.1	721.4

Solution:

Year	Income	3-year semi total	3-year semi average	Trend values
1951	301.2	1050.1	350.03	350.03 – 71.17 = 278.86
1961	336.9			278.86 + 71.17 = 350.03
1971	412			350.03 + 71.17 = 421.2
1981	484.1			421.2 + 71.17 = 492.37
1991	558.6	1904.1	634.7	492.37 + 71.17 = 563.54
2001	624.1			563.54 + 71.17 = 634.71
2011	721.4			634.71 + 71.17 = 705.88

Difference between the years = 2001 – 1961 = 40

Difference between the semi-averages = 634.7 – 350.03 = 284.67

Increase in trend value for one year = 284.67 / 4 = 71.17

3. Moving Averages Method

Moving averages is a series of arithmetic means of variate values of a sequence. To find the trend values by the method of 3-yearly moving averages, the following steps have to be considered:

- Add up the values of the first 3 years and place the yearly sum against the median year.
- Leave the first-year value, add up the values of the next three years and place it against its median year.
- This process must be continued till all the values of the data are taken for calculation.
- Each 3-yearly moving total must be divided by 3 to get the 3-year moving averages, which are our required trend values.

Advantages

- It can be easily applied.
- It is useful in case of series with periodic fluctuations.
- It does not show different results when used by different persons
- It can be used to find the figures on either extreme; that is, for the past and future years.

Disadvantages

- In non-periodic data this method is less effective.
- Selection of proper 'period' or 'time interval' for computing moving average is difficult.
- Values for the first few years and as well as for the last few years cannot be found.

Example 8.4

Calculate the 3-years moving averages for the loans issued by co-operative banks for non-farm sector/small scale industries based on the values given below:

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Loan	41.82	40.05	39.12	24.72	26.69	59.66	23.65	28.36	33.31	31.60	36.48

Solution:

Year	**Income**	**3-year semi total**	**3-year semi average**
2004	41.82	---	---
2005	40.05	120.99	40.33
2006	39.12	103.89	34.63
2007	24.72	90.53	30.18
2008	26.69	111.07	37.02
2009	59.66	110	36.67
2010	23.65	111.67	37.22
2011	28.36	85.32	28.44
2012	33.31	93.27	31.09
2013	31.60	101.39	33.80
2014	36.48	---	---

4. Single Exponential Smoothing:

Exponential smoothing weights past observations with exponentially decreasing weights to forecast future values.

This smoothing scheme begins by setting S_2 to y_1, where S_i stands for smoothed observation or Exponentially Weighted Moving Average (EWMA), and y stands for the original observation. The subscripts refer to the time periods, 1, 2, …, n. For the third period, $S_3 = \alpha y_2 + (1-\alpha) S_2$; and so on. There is no S_1; the smoothed series starts with the smoothed version of the second observation.

For any time period t, the smoothed value S_t is found by computing

$$S_t = \alpha y_{t-1} + (1-\alpha) S_{t-1};\ 0 < \alpha \leq 1;\ t \geq 3.$$

This is the basic equation of exponential smoothing and the constant or parameter α is called the smoothing constant.

There is an alternative approach to exponential smoothing that replaces y_{t-1} in the basic equation with y_t, the current observation.

Setting the first EWMA

The first forecast is very important. The initial EWMA plays an important role in computing all the subsequent EWMAs. Setting S_2 to y_1 is one method of initialization. Another way is to set it to the target of the process. Still another possibility would be to average the first four or five observations.

It can also be shown that the smaller the value of α, the more important is the selection of the initial EWMA. The user would be wise to try a few methods, (assuming that the software has them available) before finalizing the settings.

Why is it called "Exponential"?

Expand basic equation

Let us expand the basic equation by first substituting for S_{t-1} in the basic equation to obtain

$$S_t = \alpha y_{t-1} + (1-\alpha)[\alpha y_{t-2} + (1-\alpha) S_{t-2}]$$
$$= \alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + (1-\alpha)^2 S_{t-2}.$$

Let $\alpha = 0.3$. Observe that the weights $\alpha(1-\alpha)^t$ decrease exponentially (geometrically) with time.

		Value weight
last	y_1	0.2100
	y_2	0.1470
	y_3	0.1029
	y_4	0.0720

What is the "best" value for α?

The speed at which the older responses are dampened (smoothed) is a function of the value of α. When α is close to 1, dampening is quick and when α is close to 0, dampening is slow. This is illustrated in the table below.

----------> towards past observations

α	$(1-\alpha)$	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
0.9	0.1	0.01	0.001	0.0001
0.5	0.5	0.25	0.125	0.0625
0.1	0.9	0.81	0.729	0.6561

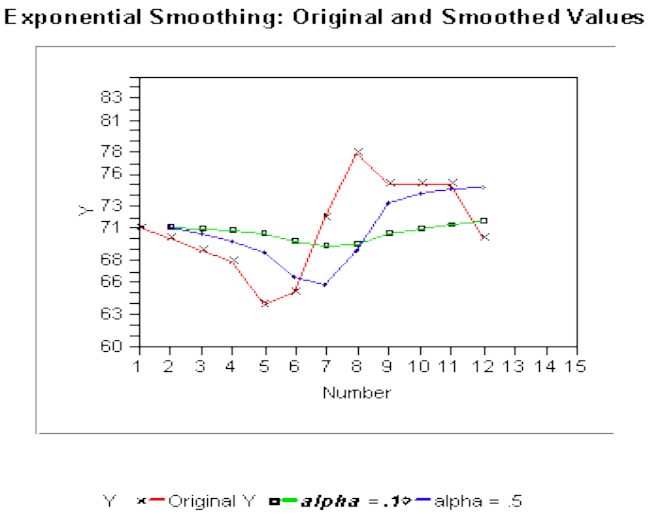
We choose the best value for α so the value which results in the smallest MSE.

Example 8.5: Let us illustrate this principle with an example. Consider the following data set consisting of 12 observations taken over time:

Time	y_t	$S(\alpha = 0.1)$	Error $(y_t - S_t)$	Error squared
1	71			
2	70	71	-1.00	1.00
3	69	70.9	-1.90	3.61
4	68	70.71	-2.71	7.34
5	64	70.44	-6.44	41.47
6	65	69.80	-4.80	23.04
7	72	69.32	2.68	7.18
8	78	69.58	8.42	70.90
9	75	70.43	4.57	20.88
10	75	70.88	4.12	16.97
11	75	71.29	3.71	13.76
12	70	71.67	-1.67	2.79

The sum of the squared errors (SSE) = 208.94. The mean of the squared errors (MSE) is the SSE/11 = 19.0.

The MSE was again calculated for $\alpha = 0.5$ and turned out to be 16.29, so in this case we would prefer an α of 0.5. Can we do better? We could apply the proven trial-and-error method. This is an iterative procedure beginning with a range of α between 0.1 and 0.9. We determine the best initial choice for α and then search between $\alpha - \Delta$ and $\alpha + \Delta$. We could repeat this perhaps one more time to find the best α to 3 decimal places.



Forecasting with Single Exponential Smoothing:

Forecasting formula:

The forecasting formula is the basic equation

$$S_{t+1} = \alpha y_t + (1-\alpha)S_t,\ 0 < \alpha \leq 1,\ t > 0.$$

Example 8.6: The following table displays the forecast:

Period	Data	Single Smoothing Forecast
13	75	71.5
14	75	71.9
15	74	72.2
16	78	72.4
17	86	73.0

Advantages

- It can be used to find average using an entire history of data or output. All other charts tend to treat each data individually.

- User can give weightage to each data point at his/her convenience. This weightage can be changed to compare various averages.
- It displays the data geometrically. Because of that, data doesn't get affected much when outliers occur.

Disadvantages

- It can only be used when continuous data over the time period is available.
- It can be used only when we want to detect a small shift in the process.
- This method can be used to calculate the average. Monitoring variance requires the user to use some other technique.

5. Method of least squares

One way of finding the trend values with the help of mathematical technique is the method of least squares. This method is most widely used in practice and in this method the **sum of squares of deviations of the actual and computed values is least** and hence the line obtained by this method is known as the line of best fit. It helps for forecasting the future values. It plays an important role in finding the trend values of economic and business time series data.

Advantages

- The method of least squares completely eliminates personal bias.
- Trend values for all the given time periods can be obtained
- This method enables us to forecast future values.

Disadvantages

- The calculations for this method are difficult compared to the other methods.
- Addition of new observations requires recalculations.
- It ignores cyclical, seasonal and irregular fluctuations.
- The trend can be estimated only for immediate future and not for distant future.

Computation of Trend using Method of Least squares

Suppose we are given n pairs of observations and it is required to fit a straight line to these data. The general equation of the straight line is: $y = a + bx$, where a and b are constants. Any value

of a and b would give a straight line, and once these values are obtained an estimate of y can be obtained by substituting the observed values of y. In order that the equation $y = a + b\ x$ gives a good representation of the linear relationship between x and y, it is desirable that the estimated values of y_i, say $\hat{y}_i$ on the whole close enough to the observed values y_i, $i = 1, 2, \ldots, n$. According to the principle of least squares, the best fitting equation is obtained by minimizing the sum of squares of differences is minimum. This leads us to two normal equations.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

That is, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad 8.1$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x_i^2 \qquad 8.2$$

Solving these two equations we get the vales for a and b and the fit of the trend equation.

We will apply this least square technique in regression analysis later in the chapter "Correlation and Regression". Example will be illustrated there.

Exercise 8

8.1 Show the principle of Moving average for the given data (Using 3 years).

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Revenue	5.0	8.0	7.0	8.0	8.0	9.0	7.0	9.0	5.0	7.0	5.0	8.0

8.2 Fit trend lines by the methods of i) graphical ii) semi-averages for the given data.

Year	2003	2005	2007	2009	2011	2013
Demand	650	700	800	810	700	900

8.3 Fit a trend line by the method of semi-averages for the given data.

Year	1990	1991	1992	1993	1994	1995	1996
Sales	15	11	20	10	15	25	35

8.4 Measure the trend by the method of semi-averages by using the table given below. Estimate the value for the year 1994-1995.

Years	**Value in Million**
1984 – 85	18.6
1985 – 86	22.6
1986 – 87	38.1
1987 – 88	40.9
1988 – 89	41.4
1989 – 90	40.1
1990 – 91	46.6
1991 – 92	60.7
1992 – 93	57.2
1993 – 94	53.4

8.5 Forecasting the total oil production in millions of tonnes for Saudi Arabia using simple exponential smoothing. For, ($\alpha = 0.3$).

Year	Observation, y_t
2001	440.39
2002	425.19
2003	486.21
2004	500.43
2005	521.28
2006	508.95
2007	488.89
2008	509.87
2009	456.72
2010	473.82
2011	525.95
2012	549.83
2013	542.34

Sample MCQs

1. Semi-averages method is used for measurement of trend when:
(a) Trend is linear (b) Observed data contains yearly values (c) The given time series contains odd number of values (d) None of them

2. A time series consists of:
(a) Short-term variations (b) Long-term variations (c) Irregular variations **(d) All of the above**

3. Moving-averages:
(a) Give the trend in a straight line (b) Measure the seasonal variations (c) **Smooth-out the time series** (d) None of them

4. In time series seasonal variations can occur within a period of:
(a) Four years (b) Three years **(c) One year** (d) Nine years

5. The rise and fall of a time series over periods longer than one year is called:
(a) Secular trend (b) Seasonal variation **(c) Cyclical variation** (d) Irregular variation

6. A time series has:
(a) Two components (b) Three components **(c) Four components** (d) Five components

7. The trend values in freehand curve method are obtained by:
(a) Equation of straight line **(b) Graph** (c) Second degree parabola (d) All of the above

8. Damages due to floods, droughts, strikes fires and political disturbances are:
(a) Trend (b) Seasonal (c) Cyclical **(d) Irregular**

9. In semi averages method, we divide the data into:
(a) Two parts **(b) Two equal parts** (c) Three parts (d) Difficult to tell

10. Find the increase in trend value in a year

Year	2001	2001	2003	2004	2005	2006
June production (in tons)	50	80	90	100	115	130

a) 13.89 b) 12.67 c) 15.76 d) 10.54