TIME SERIES ANALYSIS:

UNIVARIATE AND MULTIVARIATE MODELS

MGS649: Practicum Report By

Mohan Chakradhar Vedurupaka

50200107

INTRODUCTION:

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

The primary difference between time series models and other types of models is that lag values of the target variable are used as predictor variables, whereas traditional models use other variables as predictors, and the concept of a lag value doesn't apply because the observations don't represent a chronological sequence.

To understand time series analysis, first we need to understand following components of a time series

1. Level
2. Trend
3. Seasonality
4. Noise

Level: The level of a time series describes the average value of the series. For example, the below figure 1 shows the total runs scored by a batsman in a year. The red line represents the average runs (which is 1050) of the batsman. You can see that the total runs is fluctuating around the average score or about that level of 1050. Now look at figure 2. It shows the sales over time and clearly there is no steady level in this graph. Instead the level is increasing rapidly. However, it doesn't always have to be increasing. Sometimes, the level can be decreasing as well.

Figure 1

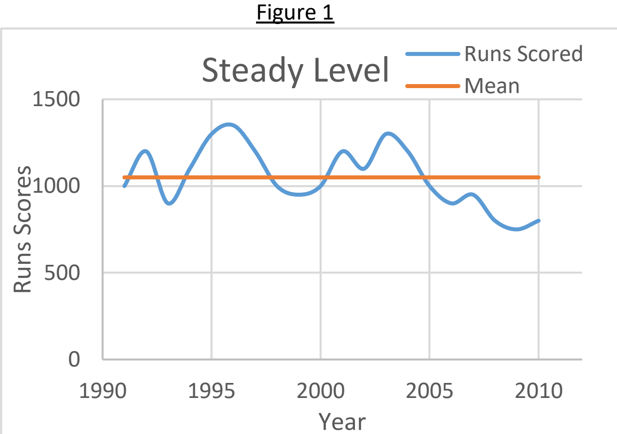
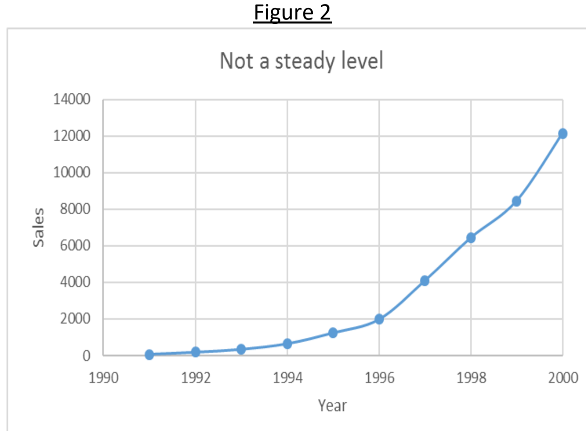


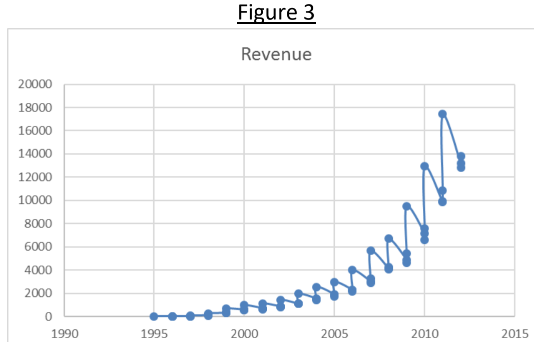
Figure 2



Trend: Trend is simply the change in level from one period to the next in a time series. Again, in the figure 1, there appears to be no significant change in the level over period. Hence, there seems to be no trend. In figure 2, the level is increasing over time which means that there is an increasing trend.

Seasonality: Seasonality means a time series fluctuates consistently at regular intervals. For example, if you look at figure 3 time series plot, you see a peak every fourth quarter. May be due to Christmas 4^{th} quarter is having high number of sales. Hence, regularly 4^{th} quarter is reporting high revenue compared to other three quarters.

Figure 3



Noise: Noise simply refers to random fluctuations in the time series about its typical pattern. If you observe figure 1 again, the data is randomly fluctuating around mean or in other words there is variation around a typical pattern.

STATIONARITY:

A key idea in time series is that of stationarity. Roughly speaking, a time series is stationary if its behavior does not change over time. This means, for example, that the values always tend to vary about the same level and that their variability is constant over time. Stationary series have a rich theory and their behavior is well understood and they therefore play a fundamental role in the study of time series. Obviously, most of the time series that we observe are non-stationary but many of them are related in simple ways to stationary time series.

There is a quite long tradition in time series to focus on only the first two moments of the process rather than on the actual observation distribution. If the process is normally distributed all information is contained in the first two moments and most of the statistical theory of time series estimators is asymptotic and more often than not only dependent on the first two moments of the process.

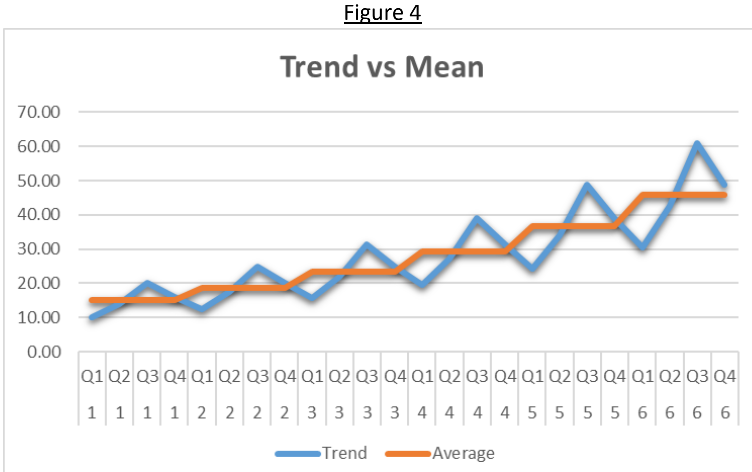
Stationarity is a rather intuitive and is an invariant property which means that statistical characteristics of the time series do not change over time. For example, the yearly rainfall may vary year by year, but the average rainfall in two equal length time intervals will be roughly the same as would the number of times the rainfall exceeds a certain threshold. Of course, over long periods of time this assumption may not be so plausible. For example, the climate change that we are currently experiencing is causing changes in the overall weather patterns. However in many situations, and over shorter intervals the assumption of stationarity is quite a plausible. Indeed often the statistical analysis of a time series is done under the assumption that a time series is stationary. There are two definitions of stationarity, weak stationarity which only concerns the covariance of a process and strict stationarity which is a much stronger condition and supposes the distributions are invariant over time.

CONVERTING NON-STATIONARY SERIES TO STATIONARY SERIES:

Many non-stationary series can be transformed into stationary ones using different transformation methods. One among them is *differencing*. It is relatively simple operation that involves calculating successive changes in the values of a data series. It is mainly used when the mean of a series is changing over time.

Below figure shows an example of such series.

Figure 4



To difference a data series, define a new variable (w_t) which is the change in z_t, that is, $w_t = z_t - z_{t-1}$, t = 2,3,….n

Using the data above, we get the results shown in table 1 when we difference the observations. Also the results plotted are shown in figure 5.

The differencing procedure seems to have been successful i.e. the differenced series in figure 5 appears to have a constant mean. Note that we lost one observation since there is no z_0 available to subtract from z_1. If the first differences do not have a constant mean, we redefine w_t as the first differences of the first differences which is called second differences also shown in table 1. The result of second differences is plotted in figure 6.

Figure 5

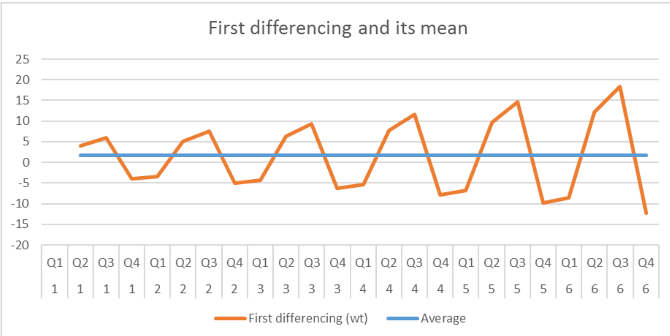


Figure 6

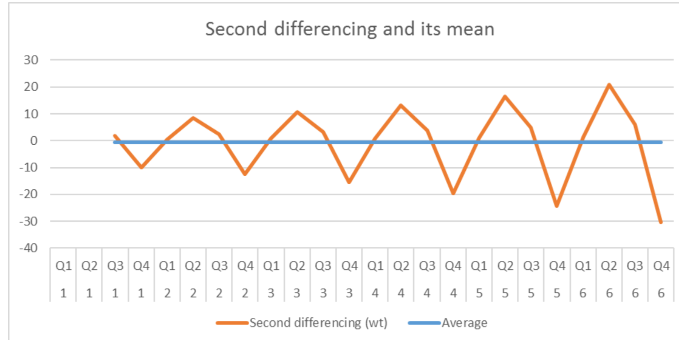


Table 1

Year	Quarter	Trend(z_t)	Average	First differencing (w_t)	Average of 1st difference	Second differencing (w_t)	Average of 2nd difference
1	Q1	10.00	15.00				
1	Q2	14.00	15.00	4.00	1.69		
1	Q3	20.00	15.00	6.00	1.69	2.00	-0.74
1	Q4	16.00	15.00	-4.00	1.69	-10.00	-0.74
2	Q1	12.50	18.75	-3.50	1.69	0.50	-0.74
2	Q2	17.50	18.75	5.00	1.69	8.50	-0.74
2	Q3	25.00	18.75	7.50	1.69	2.50	-0.74
2	Q4	20.00	18.75	-5.00	1.69	-12.50	-0.74
3	Q1	15.63	23.44	-4.38	1.69	0.63	-0.74
3	Q2	21.88	23.44	6.25	1.69	10.63	-0.74
3	Q3	31.25	23.44	9.38	1.69	3.13	-0.74
3	Q4	25.00	23.44	-6.25	1.69	-15.63	-0.74
4	Q1	19.53	29.30	-5.47	1.69	0.78	-0.74
4	Q2	27.34	29.30	7.81	1.69	13.28	-0.74
4	Q3	39.06	29.30	11.72	1.69	3.91	-0.74
4	Q4	31.25	29.30	-7.81	1.69	-19.53	-0.74
5	Q1	24.41	36.62	-6.84	1.69	0.98	-0.74
5	Q2	34.18	36.62	9.77	1.69	16.60	-0.74
5	Q3	48.83	36.62	14.65	1.69	4.88	-0.74
5	Q4	39.06	36.62	-9.77	1.69	-24.41	-0.74
6	Q1	30.52	45.78	-8.54	1.69	1.22	-0.74
6	Q2	42.72	45.78	12.21	1.69	20.75	-0.74
6	Q3	61.04	45.78	18.31	1.69	6.10	-0.74
6	Q4	48.83	45.78	-12.21	1.69	-30.52	-0.74

As shown above, we do differencing to induce a stationary mean by constructing a new series w_t that is different from the original series z_t. We then build an ARIMA model for forecasting on the stationary series w_t. But since we are interested in forecasting original series z_t this can be done using the original definition of how w_t is converted to z_t. A point to note here is a series which has been made stationary by appropriate differencing frequently has a mean of virtually zero (or close to zero).

AUTO-CORRELATION FUNCTION (ACF) & PARTIAL AUTO-CORRELATION FUNCTION (PACF):

The tools acf and pacf measure the statistical relationship between observations in a single data series.

To understand acf, let's look at following simple example.

Let 't' be the time period and z_t be the data at time t. Now if we do a correlation analysis for data at time 't' with data at time 't+1' it gives us how the trend is related with its previous value at lag 1. Following figure 7 shows the scatter plot between trend at 't' vs 't+1' for data shown in table 2. Now suppose we want to see the relationship between observations separated by two time periods. Then we compare data at time 't' with data at time 't+2' for correlation analysis at lag 2. Figure 8 show the scatter plot of the same for data shown in table 2 with lag 2.

Table 2

Year	Quarter	Trend (z_t)	Trend (z_{t+1}) at (t+1)	Trend (z_{t+2}) at (t+2)
1	Q1	10.00	14.00	20.00
1	Q2	14.00	20.00	16.00
1	Q3	20.00	16.00	12.50
1	Q4	16.00	12.50	17.50
2	Q1	12.50	17.50	25.00
2	Q2	17.50	25.00	20.00
2	Q3	25.00	20.00	15.63
2	Q4	20.00	15.63	21.88
3	Q1	15.63	21.88	31.25
3	Q2	21.88	31.25	25.00
3	Q3	31.25	25.00	19.53
3	Q4	25.00	19.53	27.34
4	Q1	19.53	27.34	39.06
4	Q2	27.34	39.06	31.25
4	Q3	39.06	31.25	24.41
4	Q4	31.25	24.41	34.18
5	Q1	24.41	34.18	48.83
5	Q2	34.18	48.83	39.06
5	Q3	48.83	39.06	30.52
5	Q4	39.06	30.52	42.72
6	Q1	30.52	42.72	61.04
6	Q2	42.72	61.04	48.83
6	Q3	61.04	48.83	
6	Q4	48.83		

Figure 7

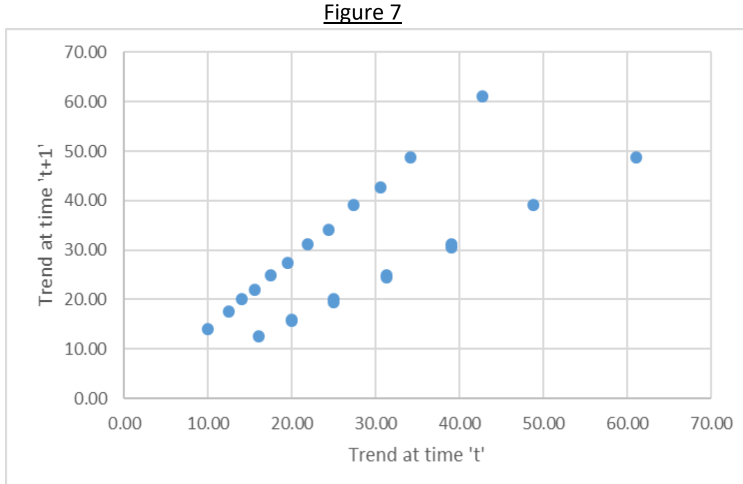
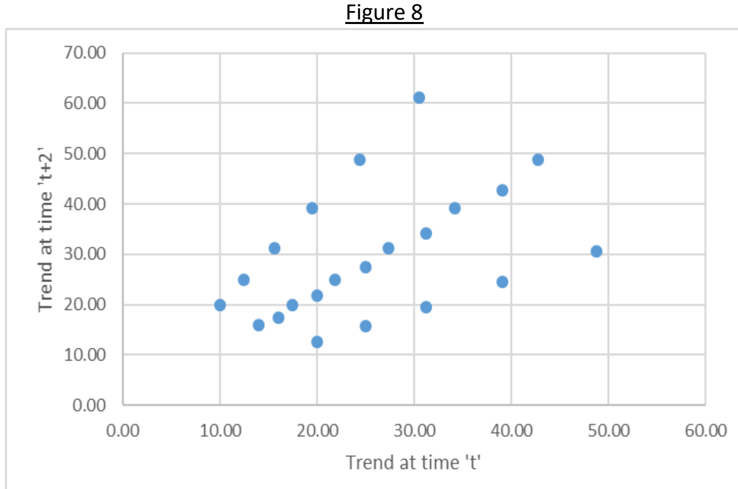
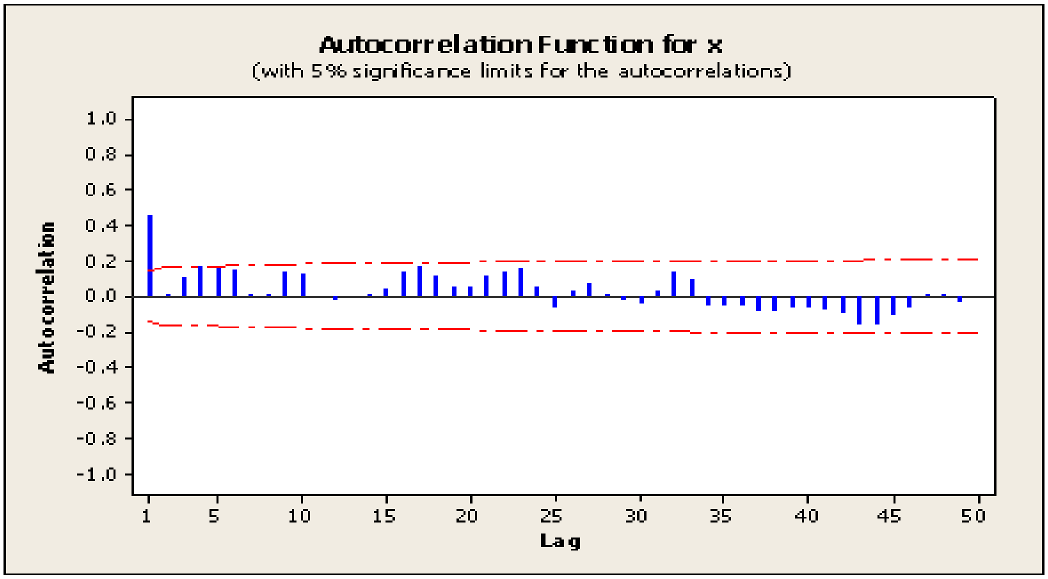


Figure 8



Similarly the idea behind acf is to calculate a correlation coefficient for each set of ordered pairs (z_t, z_{t+k}) where k is the difference in time periods. After calculating acf, we plot them in an acf diagram as shown in figure 9.

Figure 9



The standard formula for calculating autocorrelation coefficients is

$$r_k = \frac{\sum_{t=1}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{n}(z_t - \bar{z})^2}$$

where $\bar{z}$ is the mean of the observations.

The dotted red line shows the 95% confidence interval. If, acf value at lag k crosses the red line, then we say that the correlation is significant at lag k. In the figure 9, at lag 1 r_1 value is around 0.5 and crosses the red line. Hence, at lag 1, the acf is significant. However, the acf for other lag values is not significant which means that z_t is correlated with z_{t+1} and not with observations which are greater than 't+1'.

Partial auto-correlation function (pacf) is broadly similar to an estimated acf. The idea of partial autocorrelation analysis that we want to measure how z_t and z_{t+k} are related, but with the effects of the intervening z's accounted for. For example, we want to show the relationship between the ordered pairs (z_t, z_{t+2}) taking into account the effect of z_{t+1} on z_{t+2}. Next, we want the relationship between the pairs (z_t, z_{t+3}) but with the effects of both z_{t+1} and z_{t+2} on z_{t+3} accounted for and so forth. The pacf coefficient measuring the relationship between z_t and z_{t+k} is designated $\widehat{\emptyset}_{kk}$.

In constructing an acf, we dealt with only two sets of variables, so autocorrelation analysis is easy to picture graphically using scatter plot. However, in partial autocorrelation analysis we must deal with more than two variables at once. Hence, visualizing pacf on a two dimensional plot is not possible.

The most accurate way of calculating partial autocorrelation coefficients is to estimate a series of least-squares regression coefficients. An estimated regression coefficients is interpreted as a measure of the relationship between the "dependent" variable and the "independent" variable in question, with effects of other variables in the equation taken into account.

First consider the true regression relationship between z_{t+1} and the preceding z_t.

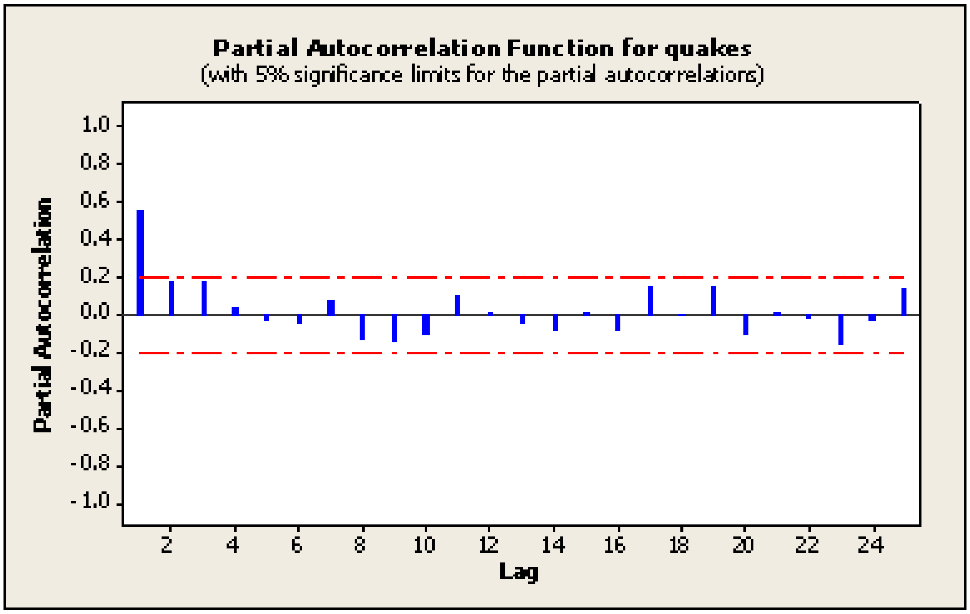
$$z_{t+1} = \emptyset_{11} z_t + u_{t+1}$$

where $\emptyset_{11}$ is the partial autocorrelation coefficient to be estimated for k = 1. u_{t+1} is the error term representing all the things affecting z_{t+1}.

Now to find $\emptyset_{22}$ the regression equation shall be $z_{t+2} = \emptyset_{21} z_{t+1} + \emptyset_{22} z_t + u_{t+2}$ where $\emptyset_{22}$ is the partial autocorrelation coefficient to be estimated for k = 2. Note that $\emptyset_{21}$ is estimated with z_{t+1} included in the equation.

Following figure 10 shows the pacf graph generated on the same data that generated acf graph shown in figure 9.

Figure 10



The acf graph is useful in deciding whether the mean of a series is stationary. If the mean is stationary the estimated acf drops off rapidly to zero. If the mean is not stationary the estimated acf drops off slowly towards zero.

UNIVARIATE FORECASTING MODELS:

1. Autoregressive Integrated Moving Average process (ARIMA).
    a. Autoregressive process (AR) process
    b. Moving Average (MA) process
2. Mixed model ARMA.
3. Three processes and ARIMA(p,d,q) notation

Autoregressive process (AR) process

Processes with past time-lagged z terms are called autoregressive (AR) processes. The longest time lag associated with a z term on the right-hand-side is called the AR order of the process. Below equation is an example of AR(1) model.

$$z_t = C + \phi_1 z_{t-1} + a_t$$

Above equation states that z_t is related to immediately past value of the same variable (z_{t-1}). The coefficient ϕ_1 has a fixed numerical value which tells how z_t is related to z_{t-1}. C is a constant term related to the mean of the process. The variable a_t stands for a random-shock element which represents a probabilistic factor.

Stationary autoregressive processes have theoretical acf's that decay or damp out towards zero. But they have theoretical pacf's that cut off sharply to zero after a few spikes. The lag length of the last pacf spike equals the AR order (p) of the process.

Moving Average (MA) process

Processes with past time-lagged random shocks are called moving average (MA) processes. The longest time lag associated with the 'a' term on the right-hand-side is called the MA order of the process. Below equation is an example of MA(1) model.

$$z_t = C + \theta_1 a_{t-1} + a_t$$

An important assumption here is that the random shocks (a_t) are independent in a process. We cannot observe the random shocks, but we can get estimates of them ($\widehat{a_t}$) at the estimation stage. The $\widehat{a_t}$ are called residuals of a model. We test the shocks for independence by constructing an acf using the residuals as input data. Then we apply t-tests to each estimated residual autocorrelation coefficient and a chi-squared test to all of them as a set. These t-tests and chi-squared test are primary tools at the diagnostic-checking stage. If the residuals are statistically independent, this is important evidence that we cannot improve the model further by adding more AR or MA terms.

Moving Average processes have theoretical acf's that cut off to zero after a certain number of spikes. The lag length of the last acf spike equals the MA order (q) of the process. Their theoretical pacf's decay or die out towards zero.

Mixed model ARMA

Mixed processes have theoretical acf's with both AR and MA characteristics. The acf tails off towards zero after the first q – p lags with either exponential decay or a damped sine wave. The theoretical pacf tails off to zero after the first p – q lags. In practice, p and q are usually not larger than two in a mixed model for non-seasonal data. Below equation is an example of ARMA(1,1) model.

$$z_t = C + \phi_1 z_{t-1} + \theta_1 a_{t-1} + a_t$$

Here the acf and pacf may alternate in sign.

Three processes and ARIMA(p,d,q) notation

Till now we have seen AR(1), MA(1) and ARMA(1,1) models. Now let the AR order of a process be designated p, where p is some non-negative integer. Let q, also a non-negative integer, be the MA order of a process. Let d, another non-negative integer, stand for the number of times a realization must be differenced to achieve a stationary mean. After a differenced series has been modeled, it is integrated d times to return the data to the appropriate overall level. The letter "I" in the acronym ARIMA refers to this integration step, and it corresponds to the number of times (d) the original series has been differenced.

Hence, ARIMA processes are characterized by the values of p, d and q in this manner: ARIMA(p,d,q). For example, following is an example of ARIMA(0,0,2) assuming that the original data is not differenced.

$$z_t = C + \theta_1 a_{t-1} + \theta_2 a_{t-2} + a_t$$

CHARACTERISTICS OF A GOOD ARIMA MODEL

1. A good model is parsimonious, meaning this model fits the available data adequately without using any unnecessary coefficients. For example, if an AR(1) model and an AR(2) model are essentially the same in all other respects, we would select the AR(1) model because it has one less coefficient to estimate.
2. A good AR model is stationary. The conditions that AR coefficients must satisfy for an ARIMA model to be stationary are:

   $$|\phi_1| < 1 \text{ for an AR(1) process}$$

   $$|\phi_2| < 1 \text{ and } \phi_1 + \phi_2 < 1,\ \phi_2 - \phi_1 < 1 \text{ for an AR(2) process}$$

   If p = 0, we have either a pure MA model or a white-noise series. All pure MA models and white noise are stationary, so there are no stationarity conditions to check. Hence, stationarity conditions apply only to AR coefficients.
3. A good MA model is invertible. Invertibility is algebraically similar to stationarity.
4. A good model has high-quality estimated coefficients at the estimation stage. We want to avoid a forecasting model which represents only a chance relationship, so we want each ϕ and θ coefficients to have an absolute t-statistic of about 2.0 or larger. If this condition is met, then these coefficients are statistically different from zero at about the 5% level. In addition, estimated ϕ and θ coefficients should not be too highly correlated with each other. If they are correlated, they tend to be somewhat unstable even if they are statistically significant.
5. A good model has statistically independent residuals.
6. A good model fits the available data sufficiently well at the estimation stage. We use two measures of closeness of fit to check how good the model is. They are root-mean-squared error (RMSE) and the mean absolute percent error (MAPE).

7. A good model has sufficiently small forecast errors. This means, even though a model fit the past data very well, is should also forecast the future data satisfactorily.

MULTIVARIATE ANALYSIS:

The topics we discussed till now are univariate analysis. That is we have forecasted future values based on processes that take past data into consideration. Instead if we take other variables such as time (Year, Month, Weekday) or any other variable instead of single variable (such as past data) to forecast future data, then it is called multivariate analysis.

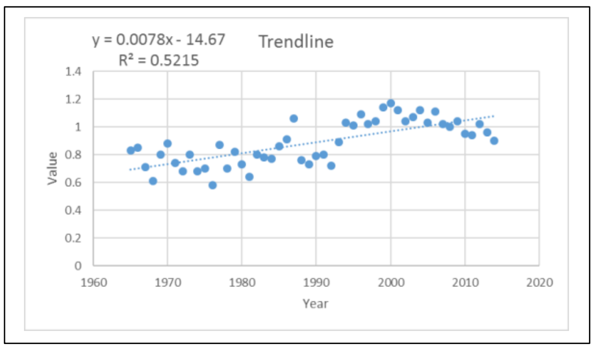
Linear Trend line:

Following example, show the multivariate analysis in which forecasting is based on year.

Following data shows the value for every year from 1965 to 2014.

Year	Value	Year	Value	Year	Value	Year	Value	Year	Value	Year	Value	Year	Value	Year	Value	Year	Value	Year	Value
1965	0.83	1970	0.88	1975	0.7	1980	0.73	1985	0.86	1990	0.79	1995	1.01	2000	1.17	2005	1.03	2010	0.95
1966	0.85	1971	0.74	1976	0.58	1981	0.64	1986	0.91	1991	0.8	1996	1.09	2001	1.12	2006	1.11	2011	0.94
1967	0.71	1972	0.68	1977	0.87	1982	0.8	1987	1.06	1992	0.72	1997	1.02	2002	1.04	2007	1.02	2012	1.02
1968	0.61	1973	0.8	1978	0.7	1983	0.78	1988	0.76	1993	0.89	1998	1.04	2003	1.07	2008	1	2013	0.96
1969	0.8	1974	0.68	1979	0.82	1984	0.77	1989	0.73	1994	1.03	1999	1.14	2004	1.12	2009	1.04	2014	0.9

Following graph shows the trend line for the above data.



The trend line equation is y = 0.0078x – 14.67 and R^2 = 0.5215. This means every year the value is increasing by 0.0078. R^2 value indicates that our trend line explains 52% variance in our data. Using the above equation, we can forecast the value for year 2015 as follows

$$y = 0.0078*2015 – 14.67 = 1.047$$

However, first we need to check if any correlation exists in the data using following method.

Year	Value	Forecast	Error	Sign change
1965	0.83	0.657	0.173	
1966	0.85	0.6648	0.1852	0
1967	0.71	0.6726	0.0374	0
1968	0.61	0.6804	-0.0704	1
1969	0.8	0.6882	0.1118	1
1970	0.88	0.696	0.184	0
1971	0.74	0.7038	0.0362	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2011	0.94	1.0158	-0.0758	0
2012	1.02	1.0236	-0.0036	0
2013	0.96	1.0314	-0.0714	0
2014	0.9	1.0392	-0.1392	0

Forecast all the values from 1965 to 2014 using the equation above. Calculate the error = (Value – Forecast). When the sign of the error gets changed for successive years, put the sign change value as 1. Otherwise 0. Now calculate the autocorrelation as below

Sum of the sign changes = 15

Total possible sign changes = 49

% of sign changes = (15/49)*100 = 30.61%

No of observations 'n' = 50

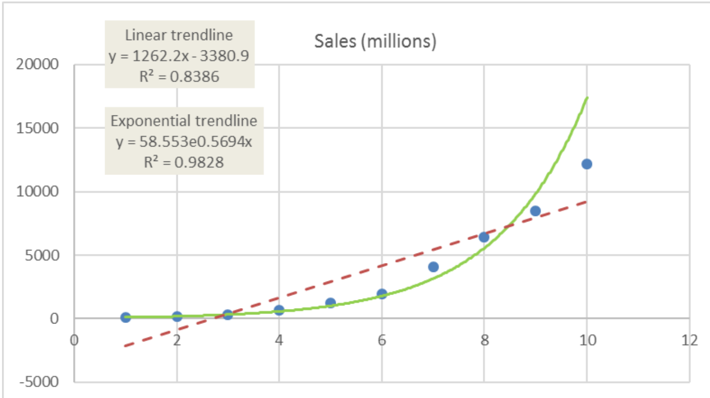
If sign changes <= (n-1)/2 – sqrt(n-1) then autocorrelation exists.

(n-1)/2 – sqrt(n-1) = (50-1)/2 – sqrt(50-1) = 17.5

Since, 15 < 17.5, autocorrelation exists. If autocorrelation is present, then our interpretation of R^2, Trendline coefficients are invalid.

Exponential Trend Curve:

Not always, we can fit a linear trend line to our data. For example, following figure explains that linear trend line is not good fit for that data.



Linear trend line is shown using red dotted lines. The R^2 value for linear trendline is 0.8386. However, if we observe the data, it is exponential. Hence, this data can be best explained by using an exponential trend curve by using the formula:

$$y = ae^{bx}$$

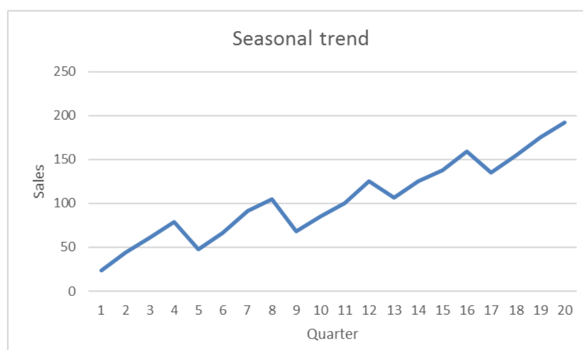
In above exponential trend curve, the R^2 value is 0.9828 which is much higher than linear trend curve. Hence, we have to check which trend line fits well to the given data. Point to note here is that the first time period value should start from 1.

Forecasting Seasonal Data:

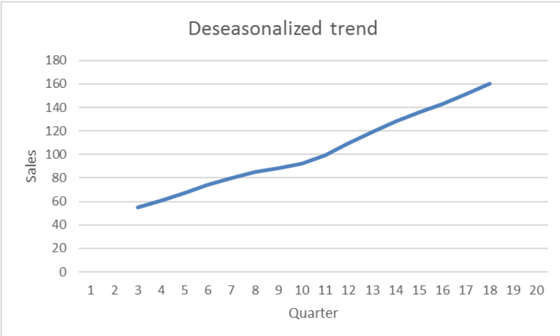
We have seen the seasonal data in figure 3. To make forecasting on seasonal data, we first need to deseasonalize it. Following is a moving average method which helps to deseasonalize the seasonal data.

Quarter#	Year	Quarter	Sales	4 period MA	Centered MA
1	1	1	24		
2	1	2	44	52	
3	1	3	61	58	55.00
4	1	4	79	63.5	60.75
5	2	1	48	71	67.25
6	2	2	66	77.5	74.25
7	2	3	91	82.5	80.00
8	2	4	105	87.25	84.88
9	3	1	68	89.5	88.38
10	3	2	85	94.5	92.00
11	3	3	100	104.25	99.38
12	3	4	125	114.25	109.25
13	4	1	107	123.75	119.00
14	4	2	125	132.25	128.00
15	4	3	138	139.25	135.75
16	4	4	159	146.75	143.00
17	5	1	135	156	151.38
18	5	2	155	164.25	160.13
19	5	3	175		
20	5	4	192		

Column 3 is the quarter number for each year represented by column 2. Column 1 is the cumulative quarter numbers starting from quarter 1 of year 1. Sales column is the sales data for each quarter. Following figure shows the seasonal trend of this data.



Every quarter 4, the sales are high and for every quarter 1, the sale are low for a particular year. To deseasonalize it, first calculate '4 period moving average' by calculating average of sales for every 4 successive time periods as shown in the table. Observe that there are no values for Quarter# 1, 19 and 20 since we don't have data for Quarter# 0, 21 and 22. Now if we observe the '4 period MA' value at quarter# 2, it is the average of sales from quarter# 1 to 4. Hence, the '4 period MA' value belong to time period at quarter# 2.5. Similarly, the '4 period MA' value at quarter# 3 is actually at time period quarter# 3.5. Since we want a value which is at time period equal to actual quarter#, we calculate centered MA by taking the average of 2 successive '4 period MA's. Hence the average of MA value at time period quarter# 2.5 and 3.5 is at time period quarter# 3 which now is equal to actual time period. Now if we plot a graph for new centered MA (deseasonalized) data against quarter#, we get the following graph.



Now this graph looks smooth. However, observe that we have lost the data for quarter# 1, 2, 19 and 20. Now we can do forecasting on this data by calculating the linear trend line. However, to get back the data to appropriate level we multiply the forecasted values with seasonal index.

Following table shows how to calculate a seasonal index.

Quarter#	Year	Quarter	Sales	4 period MA	Centered MA	Actual/CMA
1	1	1	24			
2	1	2	44	52		
3	1	3	61	58	55.00	1.11
4	1	4	79	63.5	60.75	1.30
5	2	1	48	71	67.25	0.71
6	2	2	66	77.5	74.25	0.89
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
16	4	4	159	146.75	143.00	1.11
17	5	1	135	156	151.38	0.89
18	5	2	155	164.25	160.13	0.97
19	5	3	175			
20	5	4	192			

Quarter	Seasonal Index	Normalized
1	0.818547	0.813736785
2	0.93934	0.933819599
3	1.067364	1.06109143
4	1.198394	1.191352187
Average	1.005911	1

Seasonal Index for quarter 1 is the average of all Acual/CMA values for quarter 1. Similarly, Seasonal Index for quarters 2,3 and 4 are the averages of all actual/cma values for quarters 2,3 and 4. Then we normalize the 4 seasonal index's so that the average of all the normalized seasonal indexes comes to 1.

Once, we have seasonal index and deseasonlized trend values, we calculate trendline as shown in linear trendline. Then we forecast future value using this trendline. To get back to original data, we multiply future forecasts with respective normalized seasonal indexes for each quarter.

Forecasting using recent trend:

Sometimes, data from farther time periods give lesser accurate forecasts compared to data from recent past time periods. In such cases, we first calculate seasonal index using all the data. Then we take only recent data for forecasting instead of taking all the data into account.