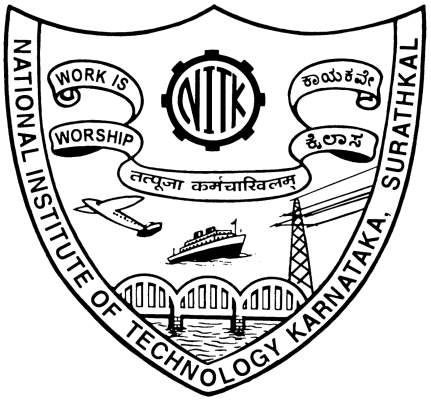
National Institute of Technology Karnataka
Department of Information Technology



Time Series Analysis and Forecasting with R

Submitted To

Mr. Biju R Mohan

Department of Information Technology

Group Members

Tushaar GVS	15IT117
Himadri Pal	15IT119
Pratyush Prakash	15IT130
Suraj Hegde	15IT146

TABLE OF CONTENTS

ABSTRACT

A time series is a sequence of numerical data points in successive order. In investing, a time series tracks the movement of the chosen data points, such as a security's price, over a specified period of time with data points recorded at regular intervals. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity. Time series analysis can be useful to see how a given asset, security or economic variable changes over time. It can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period. Delving a bit deeper, you might be interested to know whether the stock's time series shows any seasonality to determine if it goes through peaks and valleys at regular times each year. Analysis in this area would require taking the observed prices and correlating them to a chosen season. This can include traditional calendar seasons, such as summer and winter, or retail seasons, such as holiday seasons.

Keywords— Time Series Forecasting, Trend, Seasonality, Residuals, Stationarity, Confidence Interval, Hypothesis Testing

1. INTRODUCTION

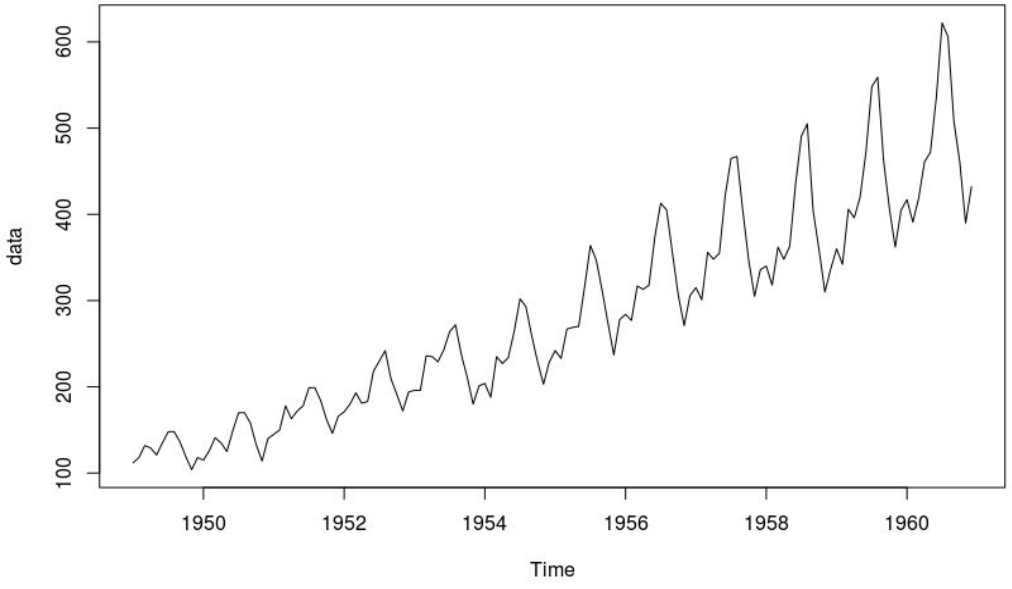
The analysis of experimental data that have been observed at different points in time leads to new and unique problems in statistical modeling and inference. The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods traditionally dependent on the assumption that these adjacent observations are independent and identically distributed. The systematic approach by which one goes about answering the mathematical and statistical questions posed by these time correlations is commonly referred to as time series analysis. The impact of time series analysis on scienti c applications can be partially documented by producing an abbreviated listing of the diverse fields in which important time series problems may arise. For example, many familiar time series occur in the eld of economics, where we are continually exposed to daily stock market quotations or monthly unemployment guresments. An epidemiologist might be interested in the number of in uenza cases observed over some time period. In medicine, blood pressure measurements traced over time could be useful for evaluating drugs used in treating hypertension. Functional magnetic resonance imaging of brain-wave time series patterns might be used to study how the brain reacts to certain stimuli under various experimental conditions.

Many of the most intensive and sophisticated applications of time series methods have been to problems in the physical and environmental sciences. This fact accounts for the basic engineering avor permeating the language of time series analysis. One of the earliest recorded series is the monthly sunspot numbers studied by Schuster (1906). More modern investigations may center on whether a warming is present in global temperature measurements. Social scientists follow population series, such as birth rates or school enroll or whether levels of pollution may in uence daily mortality in Los Angeles. The modeling of speech series is an important problem related to the efficient transmission of voice recordings. Common features in a time series characteristic known as the power spectrum are used to help computers recognize and translate speech. Geophysical time series such as those produced by yearly depositions of various kinds can provide long-range proxies for temperature and rainfall. Seismic recordings can aid in mapping fault lines or in distinguishing between earthquakes and nuclear explosions.

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data, like that encountered in the previous section. In order to provide a statistical setting for describing the character of data that seemingly fluctuate in a random fashion over time, we assume a time series can be de ned as a collection of random variables indexed according to the order they are obtained in time. For example, we may consider a time series as a sequence of random variables, x_1, x_2, x_3, ..., where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, and so on. In general, a collection of random variables, $\{x_t\}$, indexed

by *t* is referred to as a stochastic process. Here, *t* will typically be discrete and vary over the integers t = 0, 1, 2, ..., or some subset of the integers. The observed values of a stochastic process are referred to as a realization of the stochastic process. Because it will be clear from the context of our discussions, we use the term time series whether we are referring generically to the process or to a particular realization and make no notational distinction between the two concepts.

So, in a nutshell, the sequence of random variables {Yt: t = 0, ±1, ±2, ±3, ...} is called a stochastic process and serves as a model for an observed time series. It is known that the complete probabilistic structure of such a process is determined by the set of distributions of all finite collections of the Y's. Fortunately, we will not have to deal explicitly with these multivariate distributions. Much of the information in these joint distributions can be described in terms of means, variances, and covariances. Consequently, we concentrate our efforts on these first and second moments. If the joint distributions of the Y's are multivariate normal distributions, then the first and second moments completely determine all the joint distributions. The models of forecasting and the theory of time series has been explained below using R tool. The data for the count of air passengers has been plotted below, the data has been collected from year 1949 till 1960 (see Fig. 1).




```
> data <- datasets::AirPassengers
> plot (data)
```

Fig. 1. Time Series representing Air Passengers from 1949 to 1960

2. Simple Linear Regression

The linear model and its applications are at least as dominant in the time series context as in classical statistics. The classical statistical method of regression analysis may be readily used to estimate the parameters of common non constant mean trend models. The primary ideas depend on being able to express a response series, say x_t, as a linear combination of inputs, say z_{t1}, z_{t2}, ..., z_{tq}. Estimating the coefficients β_1, β_2, ..., β_q in the linear combinations by least squares provides a method for modeling xt in terms of the inputs. In the time domain applications, for example, we will express x_t as a linear combination of previous values x_{t-1}, x_{t-2}, …, x_{t-p}, of the currently observed series. The outputs x_t may also depend on lagged values of another series, say y_{t-1}, y_{t-2}, …, y_{t-q}, that have influence. It is easy to see that forecasting becomes an option when prediction models can be formulated in this form. Time series smoothing and filtering can be expressed in terms of local regression models. Polynomials and regression splines also provide important techniques for smoothing. Extensions to filters of infinite extent can be handled using regression in the frequency domain. In particular, many regression problems in the frequency domain can be carried out as a function of the periodic components of the input and output series, providing useful scientific intuition into fields like acoustics, oceanographics, engineering, biomedicine, and geophysics. The assumption of linearity, stationarity, and homogeneity of variances over time is critical in the regression context.

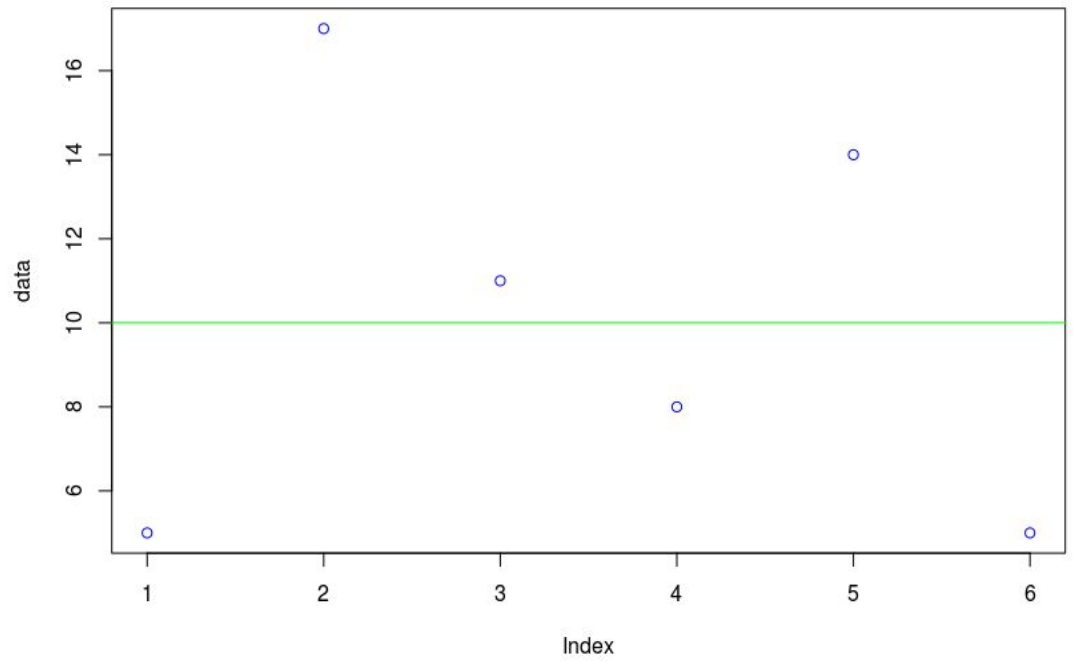
In simple linear regression that is Identically and Independently Distributed (IID), we compare the goodness of the fit of our deployed model with respect to the base model where the assumption of missing independent variable has been made (best prediction in such a case would be the mean). The following figures elucidate and illustrate the steps followed in predicting the tip amount using a simple linear regression model (see Fig. 2 and Fig. 3). Initially we assume that the meal amounts (the independent variable is missing) and the best assumption in this case would be the mean of the existing tip amounts, which may not be the most relevant prediction but this is the best prediction we could make from the given data. This constitutes the base model and the following case considers the model where both the dependent and independent variables are specified, in which case, some of the error caused is eaten up by the regression or trend line. Here the method of least squares estimates is used. The following equations represent a linear regression model, coefficients of which are β_0 and β_1 and are determined by the equations that follow (Eqn. 2 and Eqn. 3)

$$Y = \beta_0 + \beta_1 X \qquad \textbf{Eqn. 1.}$$

$$\beta_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \qquad \textbf{Eqn. 2.}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X} \qquad \textbf{Eqn. 3.}$$

In the models deployed below, the sum of squared errors (SSE), sum of squared error curtailed by regression (SSR) and sum of squares total (SST = SSE + SSR) are used to evaluate the regression models.



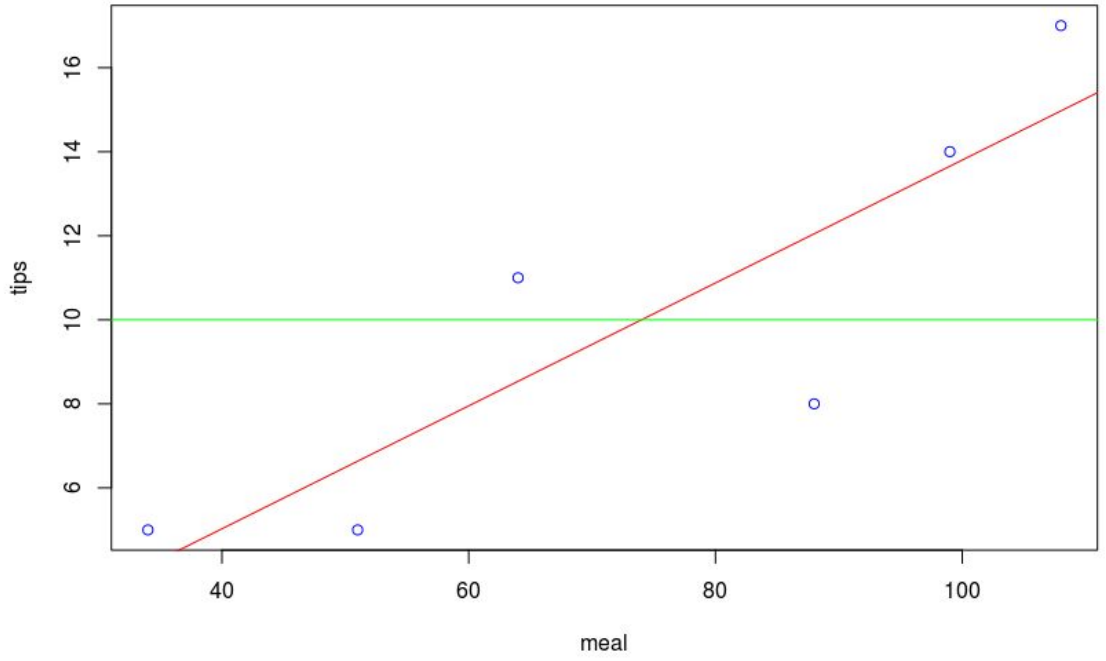

```
> tips <- array(c(5.00, 17.00, 11.00, 8.00, 14.00, 5.00) , dim=c(1, 6, 1))
> mealNumber <- array(c(1, 2, 3, 4, 5, 6), dim=c(1, 6, 1))
> plot(tips, col='blue')
> abline(h=mean(tips), col='green')

> SST <- sum((tips-mean(tips))^2)
> SST
[1] 120
```

Fig. 2. The base model in Regression estimation with no independent variable

It can be seen that the SST (= SSE) obtained in the case of base model is far greater than the value of SSE obtained when the dependent variable was considered and the regression line actually eats up most of the error (SSR = SST - SSE), which indicates that the tip amount's dependency on the meal amount (see Fig. 2 and Fig. 3). The coefficient of determination for the given dataset can be calculated using Eqn. 4, which yields coefficient of determination (r^2) to be 0.7493759, which is considered to be a good r^2 value (anything

above 0.5 is considered good). Thus, around 75% of the error is eaten up by the regression model.

$$r^2 = \frac{SSR}{SST} \qquad \textbf{Eqn. 4.}$$




```
> tips <- array(c(5.00, 17.00, 11.00, 8.00, 14.00, 5.00) , dim=c(1, 6, 1))
> meal <- array(c(34.00, 108.00, 64.00, 88.00, 99.00, 51.00) , dim=c(1, 6, 1))
> myDataFrame <- data.frame(x=meal, y=tips)
> plot(tips ~ meal, data=myDataFrame, col='blue')
> model <- lm(tips ~ meal, data=myDataFrame)
> abline(model, col='red')
> abline(h=mean(tips), col='green')

> coefficients(model)
(Intercept)        meal
 -0.8202568   0.1462197

> SSE <- sum(model$residuals^2)
> SSE
[1] 30.07489
```

Fig. 3. The regression model where the independent variable is considered for Regression

2. Fundamentals Of Statistics

Selecting a finite segment of the time series is called windowing. It is as if the original time series is viewed through a window that only allows a finite segment to be seen. Often this window is a simple box-car, i.e. a function that is zero until the start of the segment, one during the segment, and zero after the segment. Windowing is accomplished by multiplying the original series by this window. This form of discretization also limits the frequency co
ntent of the data to frequencies higher than the fundamental frequency of the segment, $f_0 = \frac{1}{T}$, where T is the length of the sample.

Assuming that the measurement is not saturated the time series will take on various values in the finite segment. One of these will be larger than all others, and one will be smaller. These extreme values for the time series are called the maximum and minimum. Simple programs often use these extremes to establish the scaling of time series plots. This scaling is usually a bad choice, particularly if the data contains a few outliers, or samples that are far from the typical values of the time series. Outliers are often produced by noise in the measurements, for example telemetry dropouts, lightning strikes, etc. When outliers are used to establish the extremes the plot consists mostly of white space with the data producing a nearly straight line and the noise producing the only variation.

2.1. Measures Of Central Tendency

There are several common quantitative measures of the tendency for a variable to cluster around a central value including the mean, median, and mode. The mean of a set of N observations of a discrete variable X_i is defined as in Eqn. 5.

$$\mu = \frac{X_1+X_2+...+X_n}{n} \qquad \textbf{Eqn. 5.}$$

The median of a probability distribution function (pdf) p(x) is the value of X_{med} for which larger and smaller values are equally probable. For discrete values, sort the samples X_i into ascending order and if N is odd find the value of X_i that has equal numbers of points above and below it. If it is even this is not possible so instead take the average of the two central values of the sorted distribution.

The mode is defined as the value of X_i corresponding to the maximum of the pdf. More generally it is taken to be the value at the center of the bin containing the largest number of values. For continuous variables the definition depends on the width of bins used in determining the histogram. If the bins are too narrow there will be large fluctuations in the estimated pdf from bin to bin. If the bins are too large the location of the mode will be poorly resolved. t is not necessary to create a histogram to obtain the mode of a distribution. It can be calculated directly from the data by computing the element with the

maximum frequency. Table 1 shows the R code to perform the same on an array [28, 33, 33, 34, 37, 40, 400]. Note that in the above array, the value '400' serves as an outlier, where we get an erroneous mean value and in which case, median and mode serve as unbiased measures of central tendency.

Table 1. R Code for computing the Measures of Central Tendency on an array
data <- c(28, 33, 33, 34, 37, 40, 400)

Function	R Code	Output
Mean	mean(data)	86.42857
Median	median(data)	34
Mode	unique(data)[which.max(tabulate(match(data, unique(data))))]	33

2.2. Measures Of Dispersion

It is obvious from the histogram (if plotted) that, values of this variable are spread around a central value. Three standard measures of this dispersion include the mean absolute deviation, the standard deviation, and the interquartile range. The mean absolute deviation (MAD) is defined by the formula given by Eqn. 6.

$$MAD = mean(abs(X_i - \overline{X}) \quad \textbf{Eqn. 6.}$$

The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. The standard deviation (SD) is given as in Eqn. 7.

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2} \quad \textbf{Eqn. 7.}$$

The upper and lower quartiles are defined in the same way as the median except that the values ¼ and ¾ are used instead of ½. The interquartile range (IQR) is the difference between the upper and lower quartiles (Q_3 and Q_1). The IQR is given by the Eqn. 8.

$$IQR = Q_3 - Q_1 \quad \textbf{Eqn. 8.}$$

For variables with a Gaussian pdf, 68% of all data values will lie within ±1 std of the mean. Similarly, by definition 50% of the data values fall within the interquartile range. Note that the standard deviation is more sensitive to values far from the mean than is the

average absolute deviation. Usually box plots are used in the analysis of the interquartile range. Table 2 summarizes the way of computing the measures of dispersion in R.

Table 2. R Code for computing the Measures of Dispersion on an array
data <- c(28, 33, 33, 34, 37, 40, 400)

Function	R Code	Output
MAD	mad(data)	4.4478
SD	sd(data)	138.3219
IQR	IQR(data)	5.5

2.3. Chebyshev's Theorem

In probability theory, Chebyshev's inequality guarantees that, for a wide class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean. Specifically, no more than $\frac{1}{K^2}$ of the distribution's values can be more than K standard deviations away from the mean. The rule is often called Chebyshev's theorem, about the range of standard deviations around the mean, in statistics. The inequality has great utility because it can be applied to any probability distribution in which the mean and variance are defined. The direct result that follows this theorem is that at least 75% of all the values are within ±2 standard deviations from the mean.

2.4. Measures Of Association

Association is concerned with how each variable is related to the other variable(s). In this case the first measure that we will consider is the covariance between two variables j and k. The population covariance is a measure of the association between pairs of variables in a population. It can range from $-\infty$ to $+\infty$. It depends on the unit of measurement and can be inflated or deflated based on the choice of the units. Negative association can also occur. If one variable tends to be greater than its mean when the other variable is less than its mean, the product of the residuals will be negative, and you will obtain a negative population covariance. Variable j will tend to decrease with increasing values of variable k.

Correlation suggests an alternative measure of association. The population correlation is defined to be equal to the population covariance divided by the product of the population standard deviations. It is very important to note that the population as well as the sample correlation must lie between -1 and 1 and is not affected by the choice of units. For a collection of p variables, the correlation matrix is a $p \times p$ matrix that displays the

correlations between pairs of variables. For instance, the value in the j^{th} row and k^{th} column gives the correlation between variables x_j and x_k. The correlation matrix is symmetric so that the value in the kth row and jth column is also the correlation between variables x_j and x_k. The diagonal elements of the correlation matrix are all identically equal to 1. Table 3 summarizes the way of computing the measures of association in R.

Table 3. R Code for computing the Measures of Association on arrays data_1 <- c(28, 33, 33) and data_2 <- c(34, 37, 40)

Function	R Code	Output
Covariance	cov(data_1, data_2)	7.5
Correlation	cor(data_1, data_2)	0.8660254

2.5. Central Limit Theorem

In probability theory, the central limit theorem establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. The associated mean and standard deviations of the sample with N elements in the sample are given as in Eqn. 9 and Eqn. 10 respectively, where $\overline{X}$ is the sample mean and μ is the population mean; S is the sample standard deviation and σ is the population standard deviation.

$$\overline{X} = \mu \qquad \textbf{Eqn. 9.}$$

$$S = \frac{\sigma}{\sqrt{N}} \qquad \textbf{Eqn. 10.}$$

3. Time Series Forecasting Using AR, MA, ARMA, ARIMA Models

3.1. Time Series Data

Time Series is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals, such as monthly, weekly, quarterly, annually, daily etc. Time series is used in statistics, econometrics, mathematical finance, weather forecasting, earthquake predictions and many other applications. With time series data, one cannot use linear regression as the time series values are not IID. Moreover, regression needs the knowledge of factors that the dependent variable *Y* is dependent on, which in the case of time series is unknown as it is just a series of data of one variable collected over time, i.e. a Univariate Series.

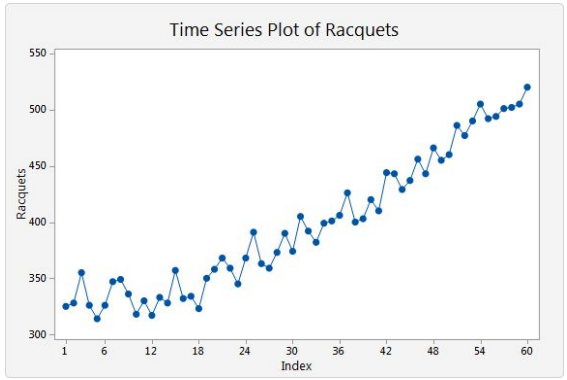
A Univariate Time Series refers to a time series that consists of single observations recorded over regular time intervals. Consider the example of stock prices collected *daily* from 1960 to 2014 for univariate series. Cross-sectional Data is the opposite of Time Series data, i.e. it is the data collected by observing many subjects (variables) at the same point of time or during the same time period (unlike in time series data). Here we are only interested in time series data.

3.2. Patterns Emerging In Time Series Data

Depending on the frequency of the data (hourly, monthly, quarterly, annually etc.) different patterns emerge in the dataset which forms the component to be modelled. Sometimes the time series may just be increasing or decreasing over time with a constant slope or there may be patterns around the increasing slope. Patterns can be widely classified into four types- Trend, Seasonality, Outliers, Abrupt Changes (see Fig. 4).

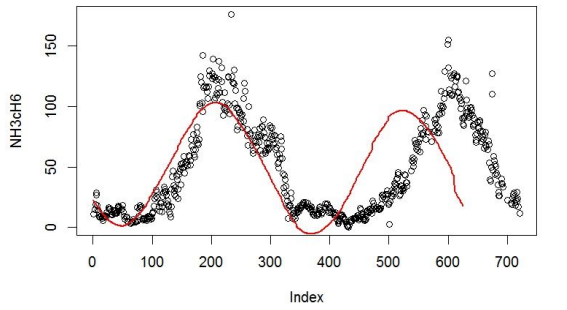
The patterns are sometimes classified based on sloping of the trend; and sometimes based on trend, seasonality, cyclicality and randomness i.e., based on the frequency. Trend is a long term smooth pattern that usually persists more than for one year (upward, downward, both, random). Seasonality is a pattern that appears in regular intervals where the frequency of occurrence is within a year or even shorter. Consider the sales of bridal dresses during the wedding season. Cyclicality is when the repeated pattern that appears in a time series but beyond a frequency of one year. It is a wavelike pattern about a long-term trend and is apparent over a number of years. Cycles are rarely regular and appear in combination with other components. The component of time series that is obtained after the above three patterns is the Random Component. Therefore when we plot the residual

series, the scatter plot should be devoid of any pattern and would be indicating only a random pattern around the mean value.

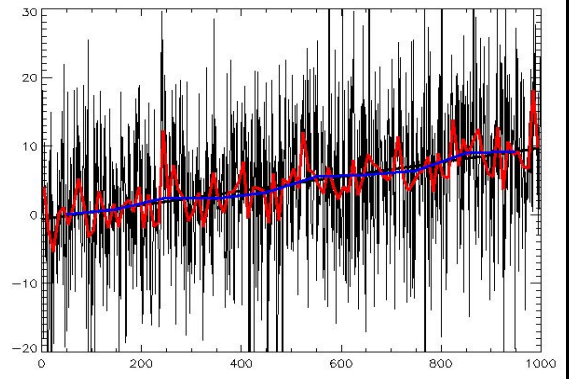


(a) An upward sloping trend

(b) A downward sloping trend



(c) A seasonal $A\cos(\omega t)$ trend



(d) A random trend

Fig. 4. Different Patterns emerging from Time Series Data

3.3. Why Use Univariate Series And Applications Of Univariate Series

The idea behind using univariate time series arises in situations where an appropriate economic theory to the relationship between the series may not be available and hence one considers the only statistical relationship of the given series with the past values. Sometimes, even when the set of explanatory variables may be known it may not be possible to obtain the entire set of such variables required to estimate using a regression model and one would have to use only a single series of the dependent variable to forecast the future values.

The applications of univariate series in forecasting inflation rates, unemployment rates, the net inflow of foreign funds in the near future could be of interest to the

government. Firms may be interested in the demand for their product or the market share of their product. Housing finance companies may want to forecast both the mortgage interest rates and the demand of housing loans. For a gold merchant, forecasting gold or silver prices would be of help.

3.4. Time Series Related Terminology And Processes

3.4.1 White Noise

A series is called white noise if it is purely random in nature. Let $\{ \varepsilon_t \}$ denote such series, then it has zero mean, i.e. $\mu_{\varepsilon_t} = 0$ and constant variance, i.e. $VAR(\varepsilon_t) = \sigma^2$ and is an uncorrelated, i.e. $COR(\varepsilon_t, \varepsilon_s) = 0$ random variable. The scatter plot of such a variable will indicate no pattern at all and hence forecasting the future values of such series is not possible. The best estimation that can be made is the mean value.

3.4.2 Differencing

Most business and economic time series are far from stationary when expressed in their original units of measurement, and even after deflation or seasonal adjustment they will typically still exhibit trends, cycles, random-walking, and other non-stationary behavior. If the series has a stable long-run trend and tends to revert to the trend line following a disturbance, it may be possible to stationarize it by de-trending (e.g., by fitting a trend line and subtracting it out prior to fitting a model, or else by including the time index as an independent variable in a regression or ARIMA model), perhaps in conjunction with logging or deflating. Such a series is said to be *trend-stationary*. However, sometimes even de-trending is not sufficient to make the series stationary, in which case it may be necessary to transform it into a series of period-to-period and/or season-to-season *differences*. If the mean, variance, and autocorrelations of the original series are not constant in time, even after detrending, perhaps the statistics of the *changes* in the series between periods or between seasons *will* be constant. Such a series is said to be *difference-stationary.* (Sometimes it can be hard to tell the difference between a series that is trend-stationary and one that is difference-stationary, and a so-called unit root test may be used to get a more definitive answer.

3.4.3 Stationarity

A series is said to be 'stationary' if the marginal distribution of *Y* at time *t*, $[p(Y_t)]$ is the same as at any other point in time. This implies that the mean, variance and covariance of

the series Y_t are time invariant. However a series is said to be 'weakly stationary' or 'covariance stationary' if the mean is constant, variance is constant but the covariance depends on the lag.

A series that is not stationary can be made stationary by differencing. A series which is stationary after being differences once is said to be integrated of order '1' and is denoted by I(1). In general, a series which is stationary after being differenced 'd' times is said be integrated of order 'd', denoted by I(d). Thus, a series that is stationary without differencing is said to be I(0). The assumption that the series is stationary is a very important assumption due to many reasons, a few of which have been elucidated below.

The results of classical economic theory are derived under the assumption that variables of concern are stationary. Standard techniques are largely invalid when the data is not stationary. Sometimes autocorrelation may result because the time series is not stationary. Moreover, non stationary time series regressions may also result in spurious regressions, i.e. cases where the regression equation shows significant relationship between two variables when actually no such relationship exists.

3.4.4 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. ACF can be computed using Eqn. 11.

$$\rho_k = COR(Y_t, Y_{t-p}) = \frac{COV(Y_t, Y_{t-p})}{\sqrt{VAR(Y_t)}\sqrt{VAR(Y_{t-p})}} = \frac{\gamma_p}{\gamma_0} \qquad \textbf{Eqn. 11.}$$

$$\alpha(1) = COR(Y_{t+1}, Y_t); \qquad \textbf{Eqn. 12.}$$
$$\alpha(k) = COR(Y_{t+k} - P_{t,k}(Y_{t+k}), Y_t - P_{t,k}(Y_t)) \; for \; k \geq 2$$

In time series analysis, the partial autocorrelation function gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags. PACF can numerically be computed using Eqn. 12.

3.4.5 Dickey-Fuller Test of Difference-Stationarity

Assume AR(1) model. The model is non-stationary or a unit root is present if $|\beta| = 1$.

$$Y_t = \beta Y_{t-1} + \varepsilon_t$$
$$Y_t - Y_{t-1} = \beta Y_{t-1} - Y_{t-1} + \varepsilon_t$$
$$\Delta Y_t = (\beta - 1) Y_{t-1} + \varepsilon_t = \gamma^* Y_{t-1} + \varepsilon_t$$

We can estimate the above model and test for the significance of the γ^* coefficient. If the null hypothesis is rejected, $\gamma^* = 0$, then Y_t is not stationary. Difference the variable and repeat the Dickey-Fuller test to see if the differenced variable is stationary. If the null hypothesis is rejected, $\gamma^* > 0$, then Y_t is stationary; use the variable.

3.5. AR, MA, ARMA And ARIMA Prediction Models

3.5.1. Autoregressive Model (AR(p))

An AR model is the one in which Y_t depends only on its own past values $Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-p}$. The order 'p' determines the number of past terms. Eqn. 13 precisely describe this and Eqn. 14 is a common representation of an autoregressive model where it depends on 'p' of its past values, called as AR(p).

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-p}, \varepsilon_t) \quad \textbf{Eqn. 13.}$$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3}, ..., \beta_p Y_{t-p} + \varepsilon_t \quad \textbf{Eqn. 14.}$$

3.5.2. Moving Average Model (MA(q))

A Moving Average model is the one when Y_t depends only on the random error terms which follow a *white noise* process (see Eqn. 15). A common representation of a moving average model where it depends on 'q' of its past errors is called MA(q). Note that the error terms are assumed to be white noise processes, with zero mean and constant variance.

$$Y_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, ..., \varepsilon_{t-q}, \varepsilon_t) \quad \textbf{Eqn. 15.}$$

$$Y_t = \beta_0 + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3}, ..., \phi_q \varepsilon_{t-q} + \varepsilon_t \quad \textbf{Eqn. 16.}$$

3.5.3. Autoregressive Moving Average Model (ARMA(p, q))

There are situations where the time series may be represented as a mix of both MA and AR models referred as ARMA(p, q). The general form of a such a time series model, which depends on 'p' of its past values and 'q' past errors of white noise disturbances, is given in Eqn. 17.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3}, ..., \beta_p Y_{t-p} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3}, ..., \phi_q \varepsilon_{t-q} + \varepsilon_t \quad \textbf{Eqn. 17.}$$

3.5.3. Autoregressive Integrated Moving Average Model (ARIMA(p, d, q))

An autoregressive integrated moving average model is a generalization of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series. ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity.

3.6. Determining Order Of The Models Using ACF And PACF

Table 4 shows the way of determining the order 'p' in an AR(p) model, order 'q' in a MA(q) model and 'p', 'q' in an ARMA(p, q) models. Once the orders are estimated, the models are subjected to goodness of fit tests (AIC, BIC, AIC_C, SBIC etc) with slight variations in the estimations and the estimation with lowest AIC value will be chosen as the deployment model. The whole illustration of R code is included in the last chapter.

Table 4. Determination of the Order of AR, MA and ARMA models

Model	ACF	PACF
AR(p)	Tails off slowly	Cuts off after 'p' lags
MA(q)	Cuts off after 'q' lags	Tails off slowly
ARMA(p, q)	Tails off slowly	Tails off slowly

3.7. Goodness Of Fit

The Akaike information criterion is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection (see Eqn.18 where L is the likelihood).

$$AIC = -2\, ln(L) + 2\, x\, num\_params \quad \textbf{Eqn. 18.}$$

$$BIC = -2\, ln(L) + ln(N)\, x\, num\_params \quad \textbf{Eqn. 19.}$$

Bayesian information criterion or Schwarz criterion is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (see Eqn.19 where L is the likelihood and N is the number of observations).

3.8. Box Jenkins Methodology For ARIMA Model Selection

The Box Jenkins model uses an iterative three-stage modeling approach:

1. *Model identification and model selection*: making sure that the variables are stationary, identifying seasonality in the dependent series (seasonality differencing it if necessary), and using plots of the autocorrelation and partial autocorrelation functions of the dependent time series to decide which (if any) autoregressive or moving average component should be used in the model.
2. *Parameter estimation:* using computation algorithms to arrive at coefficients that best fit the selected ARIMA model. The most common methods use maximum likelihood estimation or nonlinear least-squares estimation.
3. *Model checking:* by testing whether the estimated model conforms to the specifications of a stationary univariate process. In particular, the residuals should be independent of each other and constant in mean and variance over time. (Plotting the mean and variance of residuals over time and performing a Ljung–Box test or plotting autocorrelation and partial autocorrelation of the residuals are helpful to identify misspecification.) If the estimation is inadequate, we have to return to step one and attempt to build a better model.

The data they used were from a gas furnace. These data are well known as the Box and Jenkins gas furnace data for benchmarking predictive models.

4. Confidence Interval And Hypothesis Testing

4.1. Confidence Interval

When analyzing data, we can't just accept the sample mean or sample proportion as the official mean or proportion. When we estimate the statistics $\overline{X}, \widehat{p}$ (sample mean, sample proportion), we get different answers due to variability. Confidence interval is a range computed using sample statistics to estimate an unknown population parameter with a given confidence level. Confidence level is The proportion of all samples randomly drawn from the population whose confidence intervals contain the estimated population parameter. The center of a confidence interval is the sample statistic, such as a sample mean or sample proportion. This is also known as the *point estimate*. The width of our confidence interval is also known as the *margin of error.* General Form of Confidence Interval is given by Eqn. 20.

$$sample\ statistic \pm margin\ of\ error \qquad \textbf{Eqn. 20.}$$

The margin of error will depend on two factors- the level of confidence and the value of the standard error. The 95% Rule which states that approximately 95% of observations on a normal distribution fall within two standard deviations of the mean. Thus, when constructing a 95% confidence interval we use a multiplier of 2 (see Fig. 5).

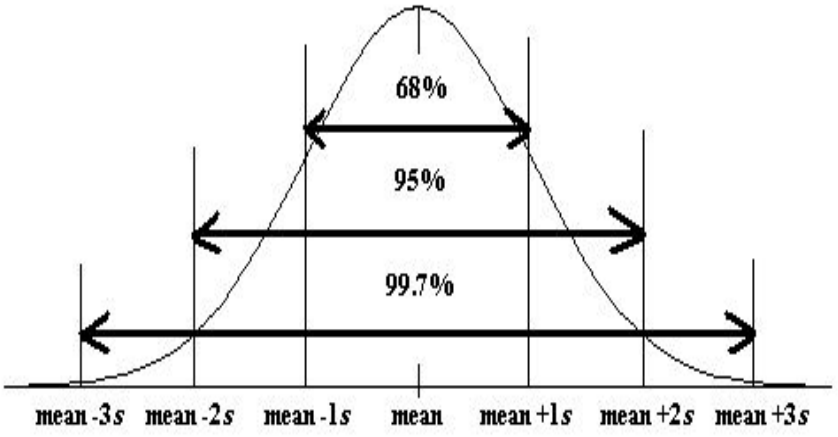


Fig. 5. Probability distribution about the mean

Suppose we want to estimate an actual population mean μ. As we know, we can only obtain, the $\overline{X}$, mean of a sample randomly selected from the population of interest. We can use $\overline{X}$ to find a range of values that we can be really confident contains the population mean μ. The range of values is called a confidence interval.

$$Lower\ value < Population\ mean\ (\mu) < Upper\ value \qquad \textbf{Eqn. 21}$$

If we are interested in estimating a population mean μ, it is very likely that we would use the t-interval for a population mean μ. The formula for the confidence interval is given in Eqn. 22 and Eqn. 23.

$$Sample\ mean \pm (t - multiplier \times standard\ error) \qquad \textbf{Eqn. 22}$$

$$\overline{X} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{N}}\right) \qquad \textbf{Eqn. 23}$$

The 't-multiplier', which we denote as $t_{\alpha/2,n-1}$, depends on the sample size through *N - 1* (degrees of freedom) and the confidence level $(1-\alpha)\times 100$ through $\alpha/2$. The standard error, which is $\frac{s}{\sqrt{N}}$, quantifies how much the sample means $\overline{X}$ vary from sample to sample. That is, the standard error is just another name for the estimated standard deviation of all the possible sample means. The quantity to the right of the ± sign, i.e., [t-multiplier × standard error], is just a more specific form of the margin of error. That is, the margin of error in estimating a population mean μ is calculated by multiplying the t-multiplier by the standard error of the sample mean. Data must be normally distributed. Clearly, the sample mean $\overline{X}$, the sample standard deviation 's', and the sample size 'N' are all readily obtained from the sample data.

4.2. Hypothesis Testing

In reviewing hypothesis tests, we start first with the general idea. Then, we keep returning to the basic procedures of hypothesis testing, each time adding a little more detail. The general idea of hypothesis testing involves three steps:

1. Making an initial assumption.
2. Collecting evidence (data).
3. Based on the available evidence (data), deciding whether to reject or not reject the initial assumption.

Every hypothesis test regardless of the population parameter involved requires the above three steps. Consider the case study of whether normal body temperature is 98.6 degrees F or not? Consider the population of many, many adults. A researcher hypothesized that the average adult body temperature is lower than the often-advertised 98.6 degrees F. That is, the researcher wants an answer to the question: "Is the average adult body temperature 98.6 degrees? Or is it lower?" To answer his research question, the researcher starts by assuming that the average adult body temperature was 98.6 degrees F.

Then, the researcher went out and tried to find evidence that refutes his initial assumption. In doing so, he selects a random sample of 130 adults. The average body temperature of the 130 sampled adults is 98.25 degrees. Then, the researcher uses the data he collected to make a decision about his initial assumption. It is either likely or unlikely that the researcher would collect the evidence he did given his initial assumption that the average adult body temperature is 98.6 degrees: If it is likely, then the researcher does not reject his initial assumption that the average adult body temperature is 98.6 degrees. There is not enough evidence to do otherwise. If it is unlikely, then: either the researcher's initial assumption is correct and he experienced a very unusual event; or the researcher's initial assumption is incorrect.

In statistics, we generally don't make claims that require us to believe that a very unusual event happened. That is, in the practice of statistics, if the evidence (data) we collected is unlikely in light of the initial assumption, then we reject our initial assumption.

4.3. Errors In Hypothesis Testing

There are two types of errors in Hypothesis Testing- one is called a *Type I error*, where the null hypothesis is rejected when it is true and the other is called a *Type II error*, where the null hypothesis is not rejected when it is false. Table 5 summarizes the types of errors in Hypothesis Testing.

Table 5. Types of Errors in Hypothesis Testing

	Truth	
Decision	Null Hypothesis	Alternative Hypothesis
Do not reject null	OK	Type II Error
Reject null	Type I Error	OK

4.3. Making Decisions In Hypothesis Testing

Recall that it is either likely or unlikely that we would observe the evidence we did given our initial assumption. If it is likely, we do not reject the null hypothesis. If it is unlikely, then we reject the null hypothesis in favor of the alternative hypothesis. Effectively, then, making the decision reduces to determining ‘likely’ or ‘unlikely’.

In statistics, there are two ways to determine whether the evidence is likely or unlikely given the initial assumption:

1. Critical value approach.
2. P-value approach.

4.3.1. Critical Value Approach

The critical value approach involves determining likely or unlikely by determining whether or not the observed test statistic is more extreme than would be expected if the null hypothesis were true. That is, it entails comparing the observed test statistic to some cutoff value, called the critical value. If the test statistic is more extreme than the critical value, then the null hypothesis is rejected in favor of the alternative hypothesis. If the test statistic is not as extreme as the critical value, then the null hypothesis is not rejected.

Specifically, the four steps involved in using the critical value approach to conducting any hypothesis test are:

1. Specify the null and alternative hypotheses (H_0, H_A).
2. Using the sample data and assuming the null hypothesis is true, calculate the value of the test statistic. To conduct the hypothesis test for the population mean μ, we use the *t*-statistic $t^* = \frac{\bar{X}-\mu}{s/\sqrt{N}}$ which follows a *t*-distribution with N - 1 degrees of freedom.
3. Determine the critical value by finding the value of the known distribution of the test statistic such that the probability of making a Type I error which is denoted α and is called the *significance level of the test*- is small (typically 0.01, 0.05, or 0.10).
4. Compare the test statistic to the critical value. If the test statistic is more extreme in the direction of the alternative than the critical value, reject the null hypothesis in favor of the alternative hypothesis. If the test statistic is less extreme than the critical value, do not reject the null hypothesis.

4.3.2. Testing

It is necessary for one to first review the critical value approach for conducting each of the following three hypothesis tests about the population mean μ. Table 6 summarizes the same based on the value of μ.

Table 6. Testing Methods in Hypothesis Testing

Type	Null Hypothesis	Alternative Hypothesis
Right-tailed	$H_0 : \mu = 3$	$H_A : \mu > 3$
Left-tailed	$H_0 : \mu = 3$	$H_A : \mu < 3$
Two-tailed	$H_0 : \mu = 3$	$H_A : \mu \neq 3$

5. The Classical Model And Its Assumptions

In statistics, ordinary least squares or linear least squares is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed responses in the given dataset and those predicted by a linear function of a set of explanatory variables.

The OLS estimator is consistent when the regressors are exogenous, and optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances. Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator. OLS is used in fields as diverse as economics (econometrics), political science, psychology and electrical engineering (control theory and signal processing).

5.1. The Assumptions In The Classical Model

The Ordinary Least Square Estimates (OLS) i.e. classical model assumes the following seven assumptions, which can be stated as follows-

1. Regression line is *linear* and has an *additive* error term.
2. Error term has *zero* population mean.
3. Explanatory variables are *uncorrelated* with the error terms.
4. No correlation between the error terms (*no autocorrelation*).
5. Explanatory variables are not perfect linear functions of one another (*no multicollinearity*).
6. Error variables have a constant variance (*no heteroskedasticity*).
7. Error variables are *normally distributed*.

5.1.1. Linear Regression and Additive Error Term

Consider Eqn. 24, where the Ordinary Linear Estimates (*OLS*) is *linear* and the error term is *additive* and all the explanatory variables are correctly specified.

$$Y_t = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_{t-1} Y_{t-1} + \varepsilon_t \qquad \textbf{Eqn. 24}$$

5.1.2. Error Term Has Zero Mean

The error term must form *white noise* (zero mean and constant variance). This accounts for the unexplained additional value from the one predicted using OLS and the actual value. Such an error term is said to be *unbiased* (biased otherwise).

5.1.3. Explanatory Variables Are Not Correlated With The Error Terms

If the explanatory variables are correlated with the error term, then some of the variation in *Y* that occurred due to the error term is attributed to *X*. Thus, if error term and the explanatory variables are positively correlated, then the expected value is actually greater than the actual one.

5.1.4. Error Terms Are Not Autocorrelated

If the error terms are serially correlated, that is to say that if the error term at time *t* depends on the error terms at times *t-1*, *t-2* and so on, then it becomes difficult for the OLS estimator to get accurate values.

5.1.5. No Perfect Multicollinearity Between Explanatory Variables

Two variables are perfectly collinear if either both are the same or one is a multiple of the other or if both variables only differ by a constant.

5.1.6. Error Terms Must Be Homoscedastic

If the value of variance of the error terms changes over different intervals of time, then the error terms are heteroskedastic. The OLS estimator cannot be applied here.

5.1.7. Error Term Is Normally Distributed

This assumption is usually violated and is mainly used in Hypothesis Testing. According to Gauss Markov theorem, this is not a necessary assumption.

5.2. The Gauss Markov Theorem

Gauss Markov Theorem states that the OLS estimator is the Best Linear Unbiased Estimator (*BLUE*). The theorem states that only the first *six* assumptions of the classical model are sufficient to apply the OLS estimator which is the best linear estimator with error term centered around zero (zero mean). If the assumption that the error term is normally distributed is also satisfied, then the following four conditions also satisfy-

1. The estimated coefficients are also *normally distributed*.

2. The estimated coefficients are actually centered around the actual values (*unbiased*).
3. No other unbiased estimator has *lower variance* for the estimated coefficient than OLS.
4. As the sample size increases, the variance gets closer to the actual variance (*consistent*).

Thus, by Gauss Markov theorem we can say that if the first six assumptions of the classical model are followed, then OLS is BLUE.

6. HETEROSKEDASTICITY

In statistics, a collection of random variables is heteroscedastic if there are sub-populations that have different variabilities from others. Here "variability" could be quantified by the variance or any other measure of statistical dispersion. Thus heteroscedasticity is the absence of homoscedasticity. A violation of the Classical Assumption V states that “the error terms must have a constant variance”.

6.1. PURE HETEROSKEDASTICITY

Pure heteroskedasticity occurs when Classical Assumption V which assumes constant variance of the error term, is violated. Classical assumption V assumes that-

$$VAR(\varepsilon_i) = \sigma^2 \quad \text{(a constant)}$$

With heteroskedasticity, the error term variance is not constant, so that-

$$VAR(\varepsilon_i) = \sigma_i^2 \quad (i=1, 2, \ldots, n)$$

Heteroskedasticity often occurs in data sets in which there is a large disparity between the largest and smallest observed values (it's reasonable to believe that large values will produce large variances and small value will produce small variances). The simplest case is that of discrete heteroskedasticity, where the observations of the error term can be grouped into just two different distributions, “wide” and “narrow”. The case is illustrated in Fig. 6.

Heteroskedasticity takes on many more complex forms, however, than the discrete heteroskedasticity case. Perhaps the most frequently specified model of pure heteroskedasticity relates the variance of the error term to an exogenous variable Zi as follows-

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

with $VAR(\varepsilon_i) = \sigma^2 Z_i^2$

where Z, the “proportionality factor,” may or may not be in the equation.

6.2. Impure Heteroskedasticity

A non-constant variance caused by incorrect specification such as an omitted variable. If the omitted variable is heteroskedastic, then its effect must be absorbed by the error term, which will in turn be heteroskedastic. The correct remedy is to find the omitted variable and include it in the regression.

6.3. Consequences Of Heteroskedasticity

1. Pure heteroskedasticity does not cause bias in the coefficient estimates.
2. We can no longer be certain that the OLS coefficient estimates are efficient (of minimum variance compared to all other unbiased linear estimators).
3. The OLS formula for calculating the standard errors of the coefficient estimates is no longer valid, and typically produces standard errors that are too low.

Note- these are the same as for autocorrelation.

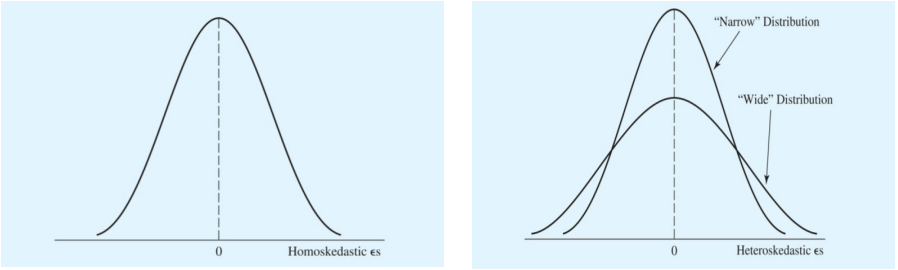


Fig. 6. Constant variance (left) vs. Varying variance (right)

6.4. Testing For Heteroskedasticity

1. Econometricians do not all use the same test for heteroskedasticity because heteroskedasticity takes a number of different forms, and its precise manifestation in a given equation is almost never known.
2. Before using any test for heteroskedasticity, however, ask the following:
    a. Are there any obvious specification errors?
        i. Fix those before testing.
    b. Is the subject of the research likely to be afflicted with heteroskedasticity?
        i. Not only are cross-sectional studies the most frequent source of heteroskedasticity.
    c. Does a graph of the residuals show any evidence of heteroskedasticity?
        i. Specifically, plot the residuals against a potential Z.

ii. In such cases, the graph alone can often show that heteroskedasticity is or is not likely. Fig. 8 below shows an example.

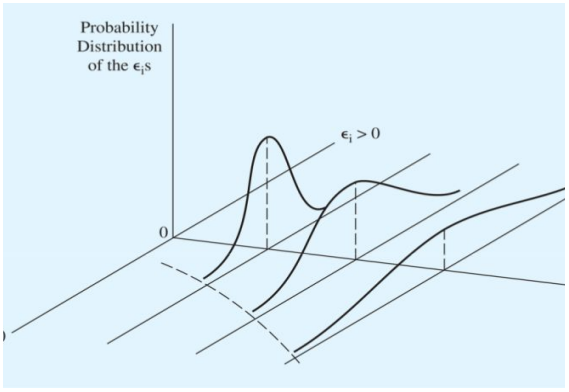


Fig. 7. Heteroskedasticity

6.5. Breusch Pagan Test

Consider the form- $VAR(\varepsilon_i) = \sigma^2 Z_i^2$

There are 3 steps to this test:

1. Obtain the residuals from the estimated equation:

$$e_i = Y_i - \hat{Y}_i$$

2. Use these residuals to form the dependent variable in a second regression.

$$\ln(e_i^2) = \alpha_0 + \alpha_1 \ln Z_i + u_i$$

where Z_i = your best estimate as to the possible proportionality factor (Z)

3. Test the significance of Z with a t-test. If Z is significantly different from zero, it shows that Z helps to explain some of the variation in e, and so there is evidence of heteroskedasticity.

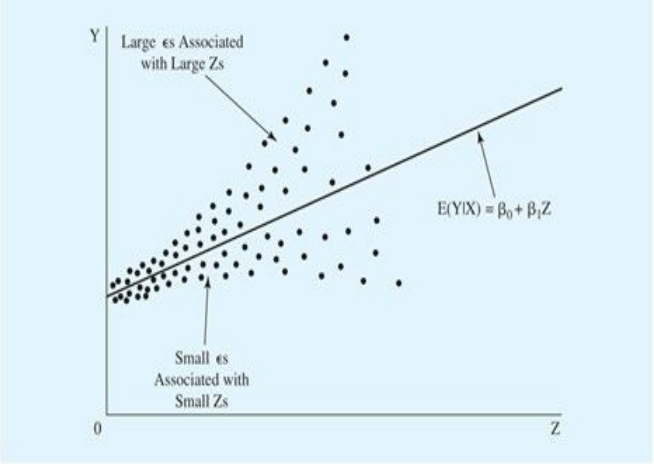


Fig. 8. Heteroskedasticity test

6.6. White Test

The white test also has 3 basic steps-

1. Obtain the residuals from the estimated regression equation. (Same as Park Test)
2. Use these residuals (squared) as the dependent variable in a second equation that includes as explanatory variables each X from the original equation, the square of each X, and the product of each X times every other X—for example, in the case of three explanatory variables:

$$(e_i)^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{1i}^2 + \alpha_5 X_{2i}^2 + \alpha_6 X_{3i}^2 + \alpha_7 X_{1i} X_{2i} + \alpha_8 X_{1i} X_{3i} + \alpha_9 X_{2i} X_{3i} + u_i$$

3. Test the overall significance of Equation 10.9 with the chi-square test
    a. The appropriate test statistic here is NR^2, or the sample size (N) times the coefficient of determination (the unadjusted R^2) of the equation
    b. This test statistic has a chi-square distribution with degrees of freedom equal to the number of slope coefficients in the equation
    c. If NR^2 is larger than the critical chi-square value found in Statistical Table, then we reject the null hypothesis and conclude that it's likely that we have heteroskedasticity
    d. If NR^2 is less than the critical chi-square value, then we cannot reject the null hypothesis of homoscedasticity.

6.7. Remedies For Heteroskedasticity

1. Heteroskedasticity-corrected standard errors

a. The logic behind heteroskedasticity-corrected standard errors is power.
b. The heteroskedasticity-corrected SEs are biased but generally more accurate than uncorrected standard errors for large samples in the face of heteroskedasticity.
c. Typically heteroskedasticity-corrected SEs are larger than OLS SEs, thus producing lower *t-scores*.

2. Redefining the variables
    a. Redefining the variables to avoid heteroskedasticity.
    b. In some cases, the only redefinition that's needed to rid an equation of heteroskedasticity is to switch from a linear functional form to a double-log functional form: double-log form has inherently less variation than the linear form, so it's less likely to encounter heteroskedasticity.

7. Serial Correlation

7.1. Pure Serial Correlation

Pure serial correlation occurs when Classical Assumption IV which assumes uncorrelated observations of the error term, is violated (in a correctly specified equation!)

1. The most commonly assumed kind of serial correlation is **first-order serial correlation**, in which the current value of the error term is a function of the previous value of the error term:

   $$\boldsymbol{\varepsilon_t = \rho\varepsilon_{t-1} + u_t}$$

   where: **ε** = the error term of the equation in question
   ρ = the first-order autocorrelation coefficient
   u = a classical (not serially correlated) error term
2. The magnitude of ρ indicates the strength of the serial correlation:
   a. If ρ is zero, there is no serial correlation
   b. As ρ approaches one in absolute value, the previous observation of the error term becomes more important in determining the current value of ε_t and a high degree of serial correlation exists
   c. For ρ to exceed one is unreasonable, since the error term effectively would "explode"
3. As a result of this, we can state that:

   $$\boldsymbol{-1 < \rho < +1}$$

7.2. Impure Serial Correlation

It is a type of serial correlation that is caused by a specification error such as:

1. an omitted variable and/or
2. an incorrect functional form

7.3. Examples Of Serial Correlation

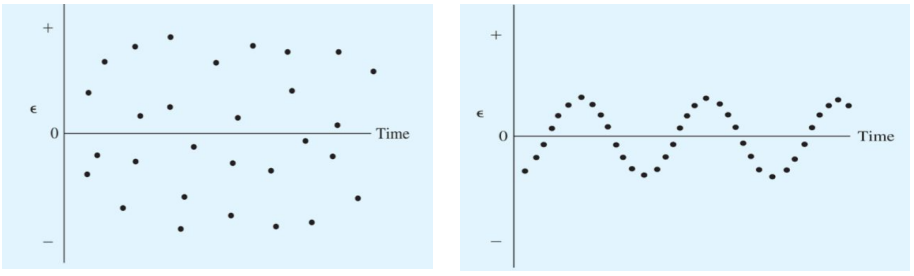


Fig. 9. No correlation (left) vs. Positive correlation (right)

7.4. Consequences Of Serial Correlation

1. Pure heteroskedasticity does not cause bias in the coefficient estimates.
2. We can no longer be certain that the OLS coefficient estimates are efficient (of minimum variance compared to all other unbiased linear estimators).
3. The OLS formula for calculating the standard errors of the coefficient estimates is no longer valid, and typically produces standard errors that are too low.

Note- these are the same as for autocorrelation.

7.5. Durbin Watson D-Test

The Durbin Watson Test is a measure of autocorrelation (also called serial correlation) in residuals from regression analysis. Autocorrelation is the similarity of a time series over successive time intervals. It can lead to underestimates of the standard error and can cause you to think predictors are significant when they are not. The Durbin Watson test looks for a specific type of serial correlation, the AR(1) process.

The Hypotheses for the Durbin Watson test are:
H_0 = no first order autocorrelation.
H_1 = first order correlation exists.
(For a first order correlation, the lag is one time unit).
Assumptions are:

1. That the errors are normally distributed with a mean of 0.
2. The errors are stationary.

The test statistic is calculated with the following formula:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Where $\boldsymbol{\varepsilon_t}$ are residuals from an ordinary least squares regression.
The Durbin Watson test reports a test statistic, with a value from 0 to 4, where:

1. 2 is no autocorrelation.
2. 0 to <2 is positive autocorrelation (common in time series data).
3. >2 to 4 is negative autocorrelation (less common in time series data).

7.6. Remedies For Serial Correlation

The place to start in correcting a serial correlation problem is to look carefully at the specification of the equation for possible errors that might be causing Impure serial correlation:

1. Is the functional form correct?
2. Are you sure that there are no omitted variables?
3. Only after the specification of the equation has been reviewed carefully should the possibility of an adjustment for pure serial correlation be considered.

There are two main remedies for pure serial correlation:

1. Generalized Least Squares
2. Newey-West standard errors

7.7. Generalized Least Squares

Start with an equation that has first-order serial correlation:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \epsilon_t$$

1. Which, if $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ (due to pure serial correlation), also equals:
$$Y_t = \beta_0 + \beta_1 X_{1t} + \rho\epsilon_{t-1} + u_t$$
2. Multiply Equation 1 by ρ and then lag the new equation by one period, obtaining:
$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{1t-1} + \rho\epsilon_{t-1}$$
3. Next, subtract Equation 3 from Equation 2 , obtaining:
$$Y_t - \rho Y_{t-1} = \beta_0(1 - \rho) + \beta_1(X_{1t} - \rho X_{1t-1}) + u_t$$
4. Finally, rewrite the above equation as
$$Y_t^* = \beta_0^* + \beta_1 X_{1t}^* + u_t$$

where,

$$Y_t^* = Y_t - \rho Y_{t-1}$$
$$X_{1t}^* = X_{1t} - \rho X_{1t-1}$$
$$\beta_0^* = \beta_0 - \rho\beta_0$$

Notice that:

1. The error term is not serially correlated
    i. As a result, OLS estimation of Equation 9.19 will be minimum variance
    ii. This is true if we know ρ or if we accurately estimate ρ

2. The slope coefficient β_1 is the same as the slope coefficient of the original serially correlated equation,Equation 2. Thus coefficients estimated with GLS have the same meaning as those estimated with OLS.

3. The dependent variable has changed compared to that in Equation 2. This mean that the GLS is not directly comparable to the OLS.

7.7. Newey-West Standard Errors

1. The logic behind Newey-west standard errors is power.
2. The Newey-west SEs are biased but generally more accurate than uncorrected standard errors for large samples in the face of heteroskedasticity
3. Typically Newey-west SEs are larger than OLS SEs, thus producing lower *t-scores.*

8. ARCH/ GARCH Model

We look at volatility clustering, and some aspects of modeling it with a univariate GARCH(1,1) model.

8.1. Volatility Clustering

Volatility clustering — the phenomenon of there being periods of relative calm and periods of high volatility — is a seemingly universal attribute of market data. There is no universally accepted explanation of it. GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) models volatility clustering. It does not explain it. Fig. 10 is an example of a garch model of volatility.

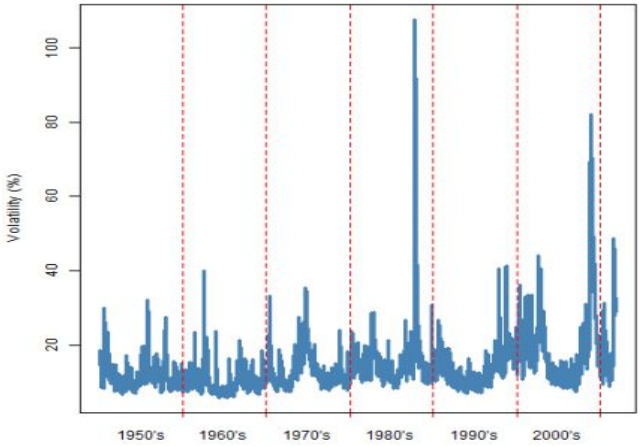


Fig. 10. S&P 500 volatility until late 2011 as estimated by a garch(1,1) model.

Clearly the volatility moves around through time. Fig. 10 is a model of volatility, not the true volatility. But if we had a picture of the true volatility, it would look remarkably like Fig 10.

8.2. Data Demands

The natural frequency of data to feed a garch estimator is daily data. You can use weekly or monthly data, but that smooths some of the garch-iness out of the data. We can use garch with intraday data, but this gets complicated. There is seasonality of volatility throughout the day. The seasonality highly depends on the particular market where the trading happens, and possibly on the specific asset. Fig. 10 does not show true volatility because we never observe volatility. Volatility ever only indirectly exposes itself to us. So we are trying to estimate something that we never see. Fig. 11 is a sketch of a prototypical garch model.

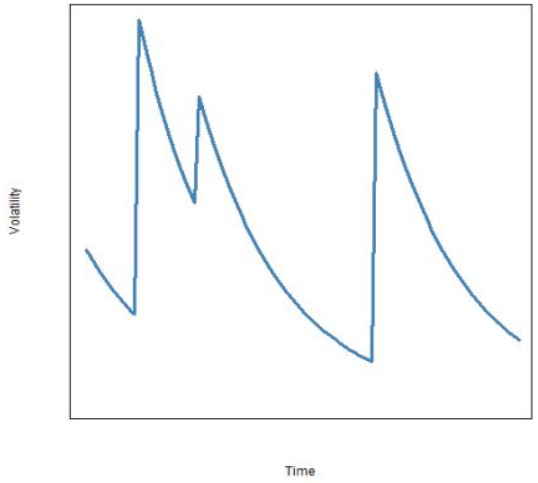


Fig. 11. Sketch of a "noiseless" garch process.

The garch view is that volatility spikes upwards and then decays away until there is another spike. It is hard to see that behavior in Fig. 10 because time is so compressed, it is more visible in Fig 12.

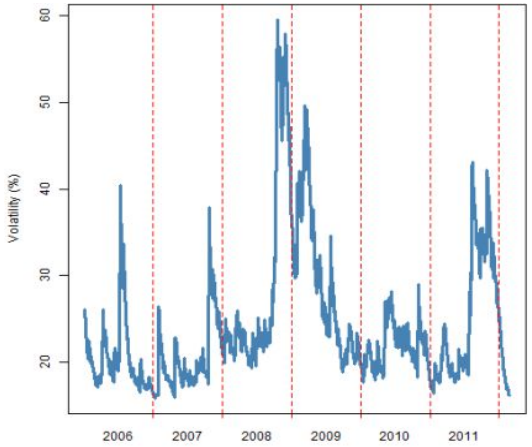


Fig. 12. Volatility of MMM as estimated by a garch(1,1) model.

Of course in the real data there are shocks of all sizes, not just big shocks. Note that volatility from announcements (as opposed to shocks) goes the other way around — volatility builds up as the announcement time approaches, and then goes away when the results of the announcement are known. The estimation of a garch model is mostly about estimating how fast the decay is. The decay that it sees is very noisy, so it wants to see a lot of data. Lots of data as in it would like tens of thousands of daily observations.

So there is a balancing act 2000 daily observations tends to be not unreasonable. If you have less than about 1000 daily observations, then the estimation is unlikely to give

you much real information about the parameters. It is probably better to just pick a "reasonable" model. That is going to be one with about the right persistence (see below), with the alpha1 parameter somewhere between 0 and 0.1, and the beta1 parameter between 0.9 and 1.

8.3. Estimation

We are staying with a GARCH(1,1) model; not because it is the best — it certainly is not. We are staying with it because it is the most commonly available, the most commonly used, and sometimes good enough. Garch models are almost always estimated via maximum likelihood. That turns out to be a very difficult optimization problem. That nastiness is just another aspect of us trying to ask a lot of the data. Assuming that you have enough data that it matters, even the best implementations of garch bear watching in terms of the optimization of the likelihood.

We know that returns do not have a normal distribution, that they have long tails. It is perfectly reasonable to hypothesize that the long tails are due entirely to garch effects, in which case using a normal distribution in the garch model would be the right thing to do. However, using the likelihood of a longer tailed distribution turns out to give a better fit (almost always). The *t* distribution seems to do quite well.

8.4. Autocorrelation

If the volatility clustering is properly explained by the model, then there will be no autocorrelation in the squared standardized residuals. It is common to do a Ljung-Box test to test for this autocorrelation. Below is output for such tests (actually Box-Pierce in this case) on a fit assuming a normal distribution on returns for MMM-

Table 7. Q-Statistics on Standardized Squared Residuals

Lag	Statistic	p-Value
Lag10	2.973	0.9821
Lag15	5.333	0.9889
Lag20	6.532	0.9980

The p-values are suspiciously close to 1. The tests are saying that we have overfit 1547 observations with 4 parameters. That is 1547 really noisy observations. A better

explanation is that the test is not robust to this extreme data, even though the test is very robust. It is probably counter-productive to test the squared residuals. An informative test is on the ranks of the squared standardized residuals.

8.5. USEFULNESS

Garch models are useful because of two things-

1. We can predict with garch models
2. We can simulate with garch models

8.6. PREDICTION

The farther ahead you predict, the closer to perfect your model has to be. Garch models are not especially close to perfect. If you are predicting with a time horizon of a month or more, then If you are predicting a few days ahead, then garch should be quite useful. The persistence of the model is a key driver of the predictions — it determines how fast the predictions go to the unconditional volatility. If there really is a lot of persistence in the volatility and your model accurately captures the persistence, then you will get good predictions far ahead.

There are two different things that might be predicted-

1. The volatility at each time point of the prediction period.
2. The average volatility from the start of the period to each time point in the period (often called the term structure).

For example, the volatility that goes into an option price is the average volatility until expiry, not the volatility on the expiry date. So there are two things you need to know when predicting-

1. Which prediction do you want?
2. Which prediction are you getting?

8.7. SIMULATION

A garch simulation needs-

1. A garch model (including the parameter values).
2. Volatility state for the model.
3. A distribution of standardized (variance 1) innovation values.

Simulation is dependent on the estimated parameters, but not as seriously as with prediction. Model errors compound as we simulate farther into the future, but they compound with a vengeance when we predict far into the future.

8.7. Implementation Using Tseries in R

This package includes a publicly available garch function in R. It is restricted to the normal distribution.

```
> gfit.ts <- garch(sp5.ret[,1])
> coef(gfit.ts)
         a0         a1         b1
6.727470e-06 5.588495e-02 9.153086e-01
> # plot in-sample volatility estimates
> plot(sqrt(252) * gfit.ts$fitted.values[, 1], type="l")
```

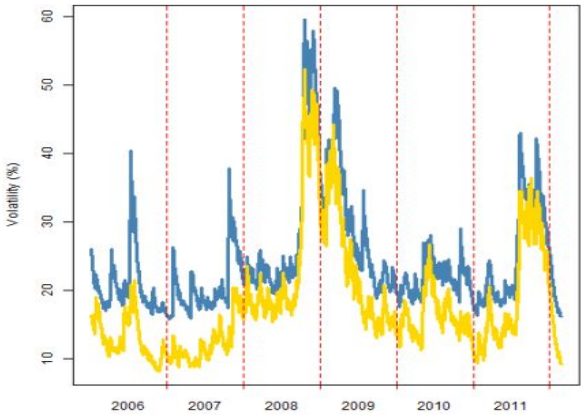


Fig. 13. Volatility of MMM as estimated by a garch(1,1) model (blue) and by the beta-t EGARCH model (gold).

9. CONCLUSIONS

In the duration of the time series course we learnt how the importance and characteristics of time series data. We learnt how to sample the data and remove seasonality and make the dataset stationary. Various models like AR(1), ARMA and ARIMA were applied to the data to predict the future values. These models were determined by Box Jenkins Method. The validity of these models were done using the concept of Hypothesis testing. The classical model (OLS) and its assumptions were studied. Heteroskedasticity and Serial Correlation removal techniques was also implemented.

REFERENCES

[1] Jonathan D. Cryer, Kung-Sik Chan, “Time Series Analysis With Applications in R”, Second Edition.

[2] Robert H. Shumway, David S. Stoffer, “Time Series Analysis and Its Applications With R Examples”, Third Edition.