Appendix A

An Analysis of Behavioural Bias and Investment Performance among Equity Mutual Fund Investors in Kerala

QUESTIONNAIRE

Please ✓ for each question.

1. Gender: a. Male ☐ b. Female ☐
2. District: ……..
3. Residential Location:
   a. Corporation ☐ b. Municipality ☐ c. Panchayath ☐
4. Age:

a.	Below 20 years
b.	20 – 40 years
c.	40 – 60 years
d.	Above 60 years

5. Education level:

a.	Higher Secondary & Below
b.	Graduate
c.	Post Graduate
d.	Professional
e.	Vocational/Technical

6. Occupation:

a.	Employed
b.	Professional
c.	Businessman
d.	Retired
e.	Others

7. Marital status:
   a. Married ☐ b. Unmarried ☐

8. Annual Income :

a.	Less than Rs. 5,00,000
b.	Rs. 5,00,000 - 10,00,000
c.	Rs. 10,00,000- 15,00,000
d.	More than Rs. 15,00,000

9. Annual mutual fund Investment:

a.	Less than Rs. 25,000
b.	Rs. 25,001 – 50,000
c.	Rs. 50,001 – Rs. 1,00,000
d.	More than Rs. 1,00,000

10. Mode of Investment:

a.	Lumpsum
b.	SIP
c.	SIP & Lumpsum

11. Years of experience in mutual fund investment :

a.	Less than 1 year
b.	1-3 years
c.	3-5 years
d.	Above 5 years

Behavioural Aspects

Read each statement and ✓ the following according to your agreement/disagreement.

SA = Strongly Agree, A = Agree, N = Neutral, D = Disagree, SD = Strongly Disagree

1.	I make investment decisions by monitoring the performance of a few samples.	**SA**	**A**	**N**	**D**	**SD**
2.	I invest in funds that have performed better recently.	**SA**	**A**	**N**	**D**	**SD**
3.	I avoid investing in funds that have performed poorly in the recent past.	**SA**	**A**	**N**	**D**	**SD**
4.	I prefer to buy hot stocks instead of poorly performed stocks.	**SA**	**A**	**N**	**D**	**SD**
5.	I have sufficient knowledge about the Indian mutual fund industry.	**SA**	**A**	**N**	**D**	**SD**
6.	My experience in trading with funds helps me choose funds that outperform the market.	**SA**	**A**	**N**	**D**	**SD**
7.	I have confidence in my ability to pick better funds.	**SA**	**A**	**N**	**D**	**SD**
8.	I never commit mistakes while making investment decisions.	**SA**	**A**	**N**	**D**	**SD**
9.	I believe that I can master the future trend of my investment.	**SA**	**A**	**N**	**D**	**SD**
10.	I think that market trends are often consistent with my perspectives.	**SA**	**A**	**N**	**D**	**SD**
11.	I rely heavily on one piece of information in making investment decision.	**SA**	**A**	**N**	**D**	**SD**
12.	I forecast the changes in net asset value of funds in the future based on the recent net asset values.	**SA**	**A**	**N**	**D**	**SD**
13.	I invest in a fund because I heard good news about it when I decided to make a investment.	**SA**	**A**	**N**	**D**	**SD**
14.	I become more optimistic when the market rises.	**SA**	**A**	**N**	**D**	**SD**
15.	I become more pessimistic when the market	**SA**	**A**	**N**	**D**	**SD**

	falls.					
16.	I make investment decisions based on available information.	**SA**	**A**	**N**	**D**	**SD**
17.	I give more importance to current information when I make investment decisions.	**SA**	**A**	**N**	**D**	**SD**
18.	I select the funds of companies which I already know.	**SA**	**A**	**N**	**D**	**SD**
19.	I consider the information from friends and relatives as a reliable reference for my investment decisions.	**SA**	**A**	**N**	**D**	**SD**
20.	I prefer to invest in already known funds.	**SA**	**A**	**N**	**D**	**SD**
21.	I hold the funds when the price decreases, even if it increases the loss.	**SA**	**A**	**N**	**D**	**SD**
22.	I invest in funds that I already own, even if their NAV goes down, to justify my investment decision.	**SA**	**A**	**N**	**D**	**SD**
23.	I believe that I get profit on investment due to my skill.	**SA**	**A**	**N**	**D**	**SD**
24.	The NAV of funds, which I selected by studying myself, increases.	**SA**	**A**	**N**	**D**	**SD**
25.	The NAV of funds, which I selected due to others' recommendations, falls.	**SA**	**A**	**N**	**D**	**SD**
26.	I collect maximum information from experts about funds, to confirm my investment decisions.	**SA**	**A**	**N**	**D**	**SD**
27.	I study the nature of funds and search for information while making investments.	**SA**	**A**	**N**	**D**	**SD**
28.	I seek market news that confirms my investment decision as correct.	**SA**	**A**	**N**	**D**	**SD**
29.	When an investment is not going well, I usually seek information that confirms I made the right decision about it.	**SA**	**A**	**N**	**D**	**SD**
30.	I seek more risk after a prior gain.	**SA**	**A**	**N**	**D**	**SD**
31.	I become more risk averse after a prior loss.	**SA**	**A**	**N**	**D**	**SD**

32.	The pain of financial loss is greater than the pleasure of financial gain.	**SA**	**A**	**N**	**D**	**SD**
33.	I prefer to invest in high-performing funds.	**SA**	**A**	**N**	**D**	**SD**
34.	I tend to hold onto losing funds too long, hoping for a reversal.	**SA**	**A**	**N**	**D**	**SD**
35.	I used to sell winning funds too soon.	**SA**	**A**	**N**	**D**	**SD**
36.	I feel more sorrow about holding onto losing funds too long than about selling winning funds too soon.	**SA**	**A**	**N**	**D**	**SD**
37.	I buy funds in times of bullish trends.	**SA**	**A**	**N**	**D**	**SD**
38.	I sell funds in times of bearish trends.	**SA**	**A**	**N**	**D**	**SD**
39.	I invest in funds in which my friends invest.	**SA**	**A**	**N**	**D**	**SD**
40.	My investment decisions are influenced by the investment behaviour of the majority.	**SA**	**A**	**N**	**D**	**SD**
41.	I would follow the market information to trade.	**SA**	**A**	**N**	**D**	**SD**
42.	I believe I have greater control over my investment.	**SA**	**A**	**N**	**D**	**SD**
43.	I can predict the market in a more logical manner.	**SA**	**A**	**N**	**D**	**SD**
44.	I tend to invest more when I am successful in my previous investment.	**SA**	**A**	**N**	**D**	**SD**
45.	I tend to treat each element of my investment portfolio separately.	**SA**	**A**	**N**	**D**	**SD**
46.	I save a part of my income for investing in the stock market.	**SA**	**A**	**N**	**D**	**SD**
47.	The rate of return on my recent investment meets my expectations.	**SA**	**A**	**N**	**D**	**SD**
48.	My rate of return is equal to or higher than the average rate of return in the market.	**SA**	**A**	**N**	**D**	**SD**
49.	I feel satisfied with my investment decisions over the last year.	**SA**	**A**	**N**	**D**	**SD**

Appendix B

DATABASE FOR THE STUDY

B.1 Database for the First Objective

Table B.1

Average Annual Returns of Equity Mutual Funds in India

Year	Large-cap Funds	Large and Mid-cap Funds	Mid-cap Funds	Small-cap Funds
2011	(21.76)	(23.88)	(23.74)	(27.31)
2012	27.31	34.01	40.52	40.79
2013	5	4.99	3.13	3.07
2014	40.96	52.08	69.73	71.98
2015	1.01	3.55	7.17	8.89
2016	3.30	6.62	3.91	5.82
2017	30.63	38.82	42.40	47.52
2018	(1.91)	(7.33)	(11.37)	(17.27)
2019	11.78	8.54	3.04	(1.51)
2020	14	16.20	24.30	30.66
2021	25.9	37.43	44.6	62.8

Source: Compiled from the Websites of Mutual Fund AMCs

Appendix C

TOOLS USED IN TIME SERIES DATA ANALYSIS

C.1 Augmented Dickey Fuller Test

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Such statistics are useful as descriptors of future behaviour only if the series is stationary. In statistics, a unit root test tests whether a time series variable is non-stationary and possesses a unit root. In this study, ADF tests have been conducted to examine the stationarity properties of the variables. Before understanding ADF Test, one must know the basics of a Dickey Fuller test. Dickey and Fuller (1979) consider three different regression equations that can be used to test the presence of a unit root:

$$\Delta Y_t = \gamma Y_{t-1} + \varepsilon t \quad \text{(C.1)}$$

$$\Delta Y_t = \alpha 0 + \gamma Y_{t-1} + \varepsilon t \quad \text{(C.2)}$$

$$\Delta Y_t = \alpha 0 + \gamma Y_{t-1} + \alpha 2t + \varepsilon t \quad \text{(C.3)}$$

In the above equations, the difference between the three regressions concerns the presence of the deterministic elements a0, a2t. While the first equation represents a pure random walk model, the second equation adds an intercept or drift term into the model and the third equation includes both an intercept and linear time trend. The test is used to identify the value of γ. If γ = 0, it implies that the Yt sequence contains a unit root. The test estimates the value of γ and associated standard error of the equations using OLS method. By analysing the value of t-statistic along with the probability value helps to determine whether to accept or reject the null hypothesis of γ = 0. Dickey Fuller test assumes that the error term εt is uncorrelated. In case when no such assumption regarding εt is taken into consideration, Dickey and Fuller have developed another unit root test which is known as the ADF test. In this test, the lagged difference terms of the variable are

included in the model to make the error term serially independent. This test is conducted by'augmenting'the preceding three equations such as Equation (C.1, C.2 and C.3) by adding the lagged values of the independent variable ΔYt. The ADF test can handle more complex models than the Dickey-Fuller test, and it is also more powerful. The ADF test may be specified as follows:

$$\Delta Yt = \alpha 0 + \alpha 1t + \gamma Yt-1 + X\, k\, i=1\, \beta i Yt-i + \varepsilon t \qquad \text{(C.4)}$$

Where εt represents a pure white noise error term

Δ represents the difference operator

γ and β represents the parameters.

ADF test follows the same asymptotic distribution as the DF statistics, i.e whether γ = 0 so the same critical values can be used. It is important to note that the selection of statistic depends on the deterministic components included in the regression equation. When there is no intercept and trend, τ statistic is used; with only the intercept, τ statistic is used and with both intercept and trend, ττ statistic is used. The statistics labelled τ, τ and ττ are the appropriate statistics to be used in Equations (C.1, C.2 and C.3) respectively. For ADF test, the value of K is determined based on either AIC or SIC.

C.2 Vector Auto Regression (VAR)

VAR method is widely used in the estimation of appropriate lag length of each variable in the system. It is possible to use different lag length for each variable in the equation. Such type of VAR is called as NEAR VAR and can be estimated through Seemingly Unrelated Regression. But for the sake of simplicity the same lag length is used for all equations. Various lag selection criteria are used to select the optimum lag length of the model. These are Likelihood Ratio, Final Prediction Error, Akaike Information Criteria, Schwarz Information Criteria and Hannan-Quinn information criteria. After setting lag length, the next step is to estimate the model through OLS. However, it is difficult to interpret individual coefficients in

estimated VAR models directly. To overcome this problem, advanced techniques like impulse response function and variance decomposition are made use of.
Suppose a multivariate VAR is given as follows:

$$Xt = A0 + A1Xt-1 + A1Xt-2 + ... + ApXt-p + et \quad (C.5)$$

Where, Xt = the (n × 1) vector containing each of the n variables included in the VAR
A0 = an (n × 1) vector of intercept terms.
Ai = an (n × n) matrix of coefficient. et = an (n × 1) vector of error terms.
In the above example, matrix A0 contains n intercept term and each matrix Ai contains n2 coefficients, hence n + pn2 terms need to be estimated. Unquestionably, a VAR will be over parameterized by which many of these coefficient estimates can be properly excluded.

C.3 Johansen's Co-integration Test

Johansen Co-integration test, named after Søren Johansen, is a procedure for testing cointegration of several, say *k*, I(1) time series. This test permits more than one cointegrating relationship so is more generally applicable than the Engle–Granger test which is based on the Dickey–Fuller (or the augmented) test for unit roots in the residuals from a single (estimated) cointegrating relationship. There are two types of Johansen test, either with trace or with eigenvalue, and the inferences might be a little bit different. The null hypothesis for the trace test is that the number of cointegration vectors is $r = r^* < k$, vs. the alternative that $r = k$. Testing proceeds sequentially for $r^* = 1,2$, etc. and the first non-rejection of the null is taken as an estimate of r. The null hypothesis for the "maximum eigenvalue" test is as for the trace test but the alternative is $r = r^* + 1$ and, again, testing proceeds sequentially for $r^* = 1, 2$ etc., with the first non-rejection used as an estimator for r.

The trace test and maximum eigen value test can be shown in equations

J_{trance}=-T Σ^{N}

i=r+1ln (1- λˆi) (C.6)

J_{tmax}=-T ln (1- λˆr+1) (C.7)

Where T is the sample size

λˆi is the i^{th} largest canonical correlation.

The trace test tests the null hypothesis of r cointegrating vectors against the alternative hypothesis of n cointegrating vectors. The maximum eigen value test, on the other hand, tests the null hypothesis of r cointegrating vectors against the alternative hypothesis of r +1 cointegrating vectors.

C.4 Vector Error Correction Model

If a set of variables are found to have one or more cointegrating vectors, then a suitable estimation technique that can be used to adjust both short run changes in variables and deviations from equilibrium a VECM. Granger (1969) argued that VECM is more appropriate to examine the causality between the series at I (1). VECM is the restricted form of unrestricted VAR and restriction is levied on the presence of the long run relationship between the series. The system of ECM makes use of all series endogenously. This system allows the predicted values to explain itself both by its own lags and lags of forcing variables as well as the lags of the ECT and by residual term. The VECM equation is as follows:

$$\begin{pmatrix} \Delta x_{1t} \\ \Delta y_{1t} \\ \Delta y_{2t} \\ \Delta y_{3t} \\ \dots \\ \Delta y_{nt} \end{pmatrix} = \begin{pmatrix} C_{1t} \\ C_{2t} \\ C_{3t} \\ C_{4t} \\ \dots \\ C_{nt} \end{pmatrix} + \sum_{i=1}^{p} \begin{bmatrix} \beta_{11i} & \beta_{12i} & \beta_{13i} & \beta_{14i} & \dots & \beta_{1ni} \\ \beta_{21i} & \beta_{22i} & \beta_{23i} & \beta_{24i} & \dots & \beta_{2ni} \\ \beta_{31i} & \beta_{32i} & \beta_{33i} & \beta_{34i} & \dots & \beta_{3ni} \\ \beta_{41i} & \beta_{42i} & \beta_{43i} & \beta_{44i} & \dots & \beta_{4ni} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ \beta_{n1i} & \beta_{n2i} & \beta_{n3i} & \beta_{n4i} & \dots & \beta_{nni} \end{bmatrix} \begin{pmatrix} \Delta x_{1_{t-i}} \\ \Delta y_{1_{t-i}} \\ \Delta y_{2_{t-i}} \\ \Delta y_{3_{t-i}} \\ \dots \\ \Delta y_{n_{t-i}} \end{pmatrix} + \begin{pmatrix} \gamma_{1t} \\ \gamma_{2t} \\ \gamma_{3t} \\ \gamma_{4t} \\ \dots \\ \gamma_{nt} \end{pmatrix} ECM_{t-1} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \dots \\ \varepsilon_{nt} \end{pmatrix} \quad \text{(C.8)}$$

Where C's, β's and γ's are the parameters to be estimated

ECM t−1 represents the one period lagged error-term derived from the co-integration vector

ε's are serially independent with mean zero and finite covariance matrix

All variables in the model are treated as endogenous variables. F test is applied to examine the direction of causal relationship between the variables. The coefficients on the ECM represent how fast deviations from the long-run equilibrium become stable.

C.5 Granger Causality Test

Causality refers to the ability of one variable containing useful information to predict and therefore influence the value of another variable based on linear least squares (Diebold 2007). To explain the causality test, the Granger (1969) definition of the proof of causality is that if variable X_t can be predicted with greater accuracy by using past values of the variable Y_t when all other terms or factors remain unchanged, it simply says Y_t that causes X_t. Therefore, the variables Y_t and X_t can affect each other with distributed lags (past period). Causality test reveals which variable is exogenous and which variables are endogenous.

Engle and Granger (1987), find that a causal relationship exists in at least one direction if two individual variables are cointegrated. The VAR model can be constructed in terms of time series at level form, (I(0)). It also can be constructed in terms of the first difference of the variable, (I(1)), with the addition of an ECT to capture the dynamic short-run response. However, if the data are not cointegrated (I(1)), the causality test can be derived from transforming the data into stationarity.

C.6 Variance Decomposition Analysis

Short run variations occurring in a variable are mostly due to its own shocks. However, there are chances of other variables to have an impact on the variable. Forecast Error Variance Decomposition (FEVD) helps to measure the impact of external variables on the selected variable. While Impulse Response Function (IMF) analyses the dynamic behaviour of the target variables due to unanticipated shocks within a VAR model, variance decomposition analysis determines the

relative importance of each innovation on the variables in the system. Variance decompositions analysis can be considered as similar to R2 values associated with the dependent variables in different horizons of shocks. To calculate n-period forecast error Xt+n considering the vector moving average representation of VAR, the following equation is used.

$X_t+n\text{-}EtX_{t+n} = \mu + \Sigma n{-}1\ \theta i{=}0\ i\varepsilon t+n{-}1$ (C.9)

Considering Yt, the first element of the Xt+n matrix in Equation (C.9), the variance

of the n-step-ahead forecast error can be calculated as:

$Y_t + n\text{-}EtX_{t+n} = \theta^{11}(0)\varepsilon_{yt+n} + \theta_{11}(1)\ \varepsilon_{yt+n-1}+....+\ \theta11(n-1)\varepsilon_{yt+1}+\theta12(0)\varepsilon_{zt+n} +\theta12(1)\varepsilon_{zt+n-1} + ... + \theta12(n-1)\ \varepsilon zt+1$ (C.10)

or

$\sigma y(n)^2 = \sigma^2 y\ [\theta11(0)^2 + \theta11(1)^2 +....+\theta11(n-1)^2] + \sigma2z\ [\theta12(0)^2 + \theta12(1)^2 +.... + \theta12(n-1)^2]$ (C.11)

Where $\sigma y(n)^2$ and $\sigma z(n)^2$ denote the n-step-ahead forecast error variance of Y_{t+n} and Z_{t+n}, respectively. While the first part of the Equation (C.10) shows the proportion of variance due to the variables own shock i.e., Y_t, the second part of the Equation (C.11) shows the proportion of variance due to the other variables shock i.e., Z_t.

Theoretically, the first part decreases over time and the second part of the variance increases. However, it is typical for a variable to explain almost all of its forecast error variance at a short horizon and smaller proportions at longer horizons. From this standpoint, variance decomposition analysis is useful to assess how one variable explains a considerable portion of forecast error variance of another variable. That is, when a shock εz explains none of the forecast error variance of the sequence Y_t at all forecast horizons, i.e., $\delta\sigma^2 y/\sigma^2 z \approx 0$, we may say that Y_t evolves indecently of the Z_t shocks i.e, εz. In addition to that, when a shock given to the Z_t sequence i.e.,εz explains the entire forecast error variance of the sequence

Y_t at all forecast horizons, i.e., δσ2 y/σ2z≈ 100%, may say that Y_t sequence is totally endogenous.

C.7 Impulse Response Function

Impulse response function is the reaction of any dynamic system in response to some external change. It is a useful tool in determining the magnitude, direction, and the duration of the variables in the system which are affected by an external variable's shock. Its main purpose is to describe the evolution of a model's variables in reaction to a shock in one or more variables. For estimating impulse response function, VAR model is be transformed into Vector Moving Average (VMA) as it allows to identify the effects of various shocks on variables in the system. In a VAR model which includes two variables, the form of the impulse response function can be written as:

$$\begin{bmatrix} \frac{Y_t}{Z_t} \end{bmatrix} = \begin{bmatrix} \frac{\bar{Y}}{Z} \end{bmatrix} + \sum_{i=0}^{\infty} \frac{A^i}{1 - b_{12}b_{21}} \begin{bmatrix} 1 & -b12 \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{Y_{t-i}} \\ \varepsilon_{Z_{t-i}} \end{bmatrix} \tag{C.12}$$

$$\begin{bmatrix} \frac{Y_t}{Z_t} \end{bmatrix} = \begin{bmatrix} \frac{\bar{Y}}{Z} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \theta_{11}^i & \theta_{12}^i \\ \theta_{21}^i & \theta_{22}^i \end{bmatrix} \begin{bmatrix} \varepsilon_{Y_{t-i}} \\ \varepsilon_{Z_{t-i}} \end{bmatrix} \tag{C.13}$$

and

$$X_t = \mu + \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i} \tag{C.14}$$

Where θi is the impulse response function of disturbances.

Therefore, impulse response function is analysed by reading off the coefficients in the moving average representation of the process. If the innovations εtâĹŠi are contemporaneously uncorrelated, interpretation of the impulse response will be straightforward. For example, the i^{th} innovation of εt is simply a shock to the i^{th} endogenous variable in the system. However, the residuals generated by the VAR models are usually contemporaneously correlated. This is because in a VAR model only lagged endogenous variables are admitted on the right-hand side of each equation

(in addition to a constant term), and hence all the contemporaneous shocks which impact on X_t are forced to feed through the residuals u_{it}. While this may not cause a problem in the estimation of the VAR model, the impulse responses and variance decompositions derived from the initial estimates of the VAR model can be affected because any adjustment made in the order of the variables entered in the system could produce different results. Thus, there is a need to impose some restrictions when estimating the VAR model to identify the impulse response function. In this regard, a common approach is the Cholesky decomposition, which was originally applied by Sims in 1980. The Cholesky decomposition overcomes the problem of contemporaneous relationships among the innovations error terms within the estimated VAR model by identifying structural shocks so that the covariance matrix of the estimated residuals is lower triangular. In fact, the Cholesky decomposition suggests that there is no contemporaneous pass-through from Y_t to the other variable, z_t. More formally, in the VAR, the matrix error structure becomes left triangular. In practice, this means that the Cholesky decomposition attributes all the effect to the variable that comes first to the target variable in the VAR system.

C.8 Auto Regressive Integrated Moving Average (ARIMA)

An Autoregressive Integrated Moving Average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) 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 (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity. The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The equation for the AR model is shown below:

$$Y_t = \beta_1 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} \qquad \text{(C.15)}$$

The respective weights (Φ1, Φ2 …Φp) of the corresponding lagged observations are decided by the correlation between that lagged observation and the current observation. If the correlation is more, the weight corresponding to that lagged observation is high (and vice-versa). This (p) is called the lag order. It represents the number of prior lag observations we include in the model i.e., the number of lags which have a significant correlation with the current observation. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past.

$$Y_t = \beta_2 + \omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \ldots + \omega_q \varepsilon_{t-q} + \varepsilon_t \quad \text{(C.16)}$$

The ε terms represent the errors observed at respective lags and the weights (ω1, ω2 …ωq) are calculated statistically depending on the correlations. (q) represents the size of the moving window i.e., the number of lag observation errors which have a significant impact on the current observation. It's similar to the lag order (p), but it considers errors instead of the observations themselves.

When we combine the AR and MA equations, we get

$$\begin{aligned} Y_t = {} & (\beta_1 + \beta_2) \\ & + (\Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p}) \\ & + (\omega_1 \varepsilon_{t-1} + \ldots + \omega_q \varepsilon_{t-q} + \varepsilon_t) \end{aligned} \quad \text{(C.17)}$$

The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). This is equivalent to performing a transformation of the form:

$$Z_t = Y_{t+1} - Y_t \quad \text{(C.18)}$$

So to revise, the final ARIMA model will take the following form, ARIMA (*p,d,q*).

Where *p* represents Auto Regressive (AR)

d represents order of differencing (I)

q represents moving average (MA)