A STOCHASTIC TIME SERIES ANALYSIS OF SALE OF DOMESTIC CARS IN INDIA:

A STUDY FOR THE PERIOD 1990-91 TO 2019-20

A PROJECT REPORT

A dissertation submitted for the partial fulfillment of M.Sc (INTEGRATED) in Science

Submitted by

Sarupyo Chatterjee (96-STS/170006)
Registration No. 100364 of 2017-2018

Under the supervision of

Prof. Chiranjib Neogi
(Guest Faculty, Department of Statistics, University of Kalyani and Retired Associate Scientist at Economic Research Unit, Indian Statistical Institute)

DEPARTMENT OF STATISTICS
UNIVERSITY OF KALYANI
Kalyani, Nadia, West Bengal-741235,India

2017-2022

DEPARTMENT OF STATISTICS
UNIVERSITY OF KALYANI
Kalyani, Nadia, West Bengal-741235, India

CANDIDATE'S DECLARATION

I, SARUPYO CHATTERJEE, Roll: 96.STS No.170006, Registration No. 100364 of 2017-2018, a student of Integrated M.Sc (Department of Statistics, University of Kalyani), hereby declare that the project Dissertation titled "***A Stochastic Time Series Analysis of Sale of Domestic Cars in INDIA : A Study for the Period 1990-91 to 2019-20***" has been carried out by me under the guidance of Professor Chiranjib Neogi at **Department of Statistics**, **University of Kalyani, Nadia, West Bengal.**

In partial fulfillment of the requirement for the award of the degree of **M.Sc (Integrated Five Years) in Statistics**, this dissertation is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship, or like other similar title or recognition.

Place: Kalyani

Sarupyo Chatterjee

Date: .08.22

DEPARTMENT OF STATISTICS
UNIVERSITY OF KALYANI
Kalyani, Nadia, West Bengal-741235,India

CERTIFICATE

I hereby certify that the Project Dissertation titled "***A Stochastic Time Series Analysis of Sale of Domestic Cars in INDIA : A Study for the Period 1990-91 to 2019-20***" which is submitted by **SARUPYO CHATTERJEE, Roll No. 96-STS/170006, Registration No. 100364 of 2017-2018, Department of Statistics, University of Kalyani,** in partial fulfillment of the requirement for the award of the degree of **Masters in Science (Integrated Five Years)**, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Kalyani — Prof. Chandranath Pal (HOD)

Date: .08.2022 — Prof.Chiranjib Neogi (SUPERVISOR)

DEPARTMENT OF STATISTICS
UNIVERSITY OF KALYANI
Kalyani, Nadia, West Bengal-741235,India

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Professor Chiranjib Neogi for his continuous guidance and mentorship that he provided us during the project. He showed us the path to achieve our targets by explaining all the tasks to be done and explained us the importance of this project, as well as its industrial relevance. He was always ready to clear our doubts and also provide insights regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful. Besides this, I want to thank my group of friends Anjali Sharma, Dishari Chakraborty, Chivukula Vishwanath, and Subhra Bhowmick with whom I have worked on this project. Lastly, I want to thank my family members for their constant support throughout this journey.

Place: Kalyani　　　　Sarupyo Chatterjee

Date: .08.2022

Abstract

It has been observed that in spite of the spread of coronavirus in 2020, there has been a consistent rise in the sale of private vehicles as people are advised to avoid local transportation. In our present study, however, we are not considering the post-covid period. Nowadays, the price hike of Petrol, Diesel, and LPG has become one of the most important issues for the middle-class and lower-middle-class families in INDIA. However, the interesting fact is that even in this high price of Fuel and Power, particularly of Petrol and Diesel the middle-class population is no less interested towards buying cars in the recent years. Basically, the car is becoming a necessary consumer good in today's world.

This paper examines whether there exists a long-run relationship between sale of domestic cars and other independent variables. As we know that there are so many factors that influence the demand of cars. We have taken here some of the variables that may influence the sale of cars in INDIA. Some of the factors we considered are- Price of Fuel and Power (FP), Real Net Per Capita National Income (RNNI), and Price of Manufacturing Products (MP). Since actual all India Prices of these items are rarely available we have taken the price indices of these items. The results of our analysis show that there is a long-term causality between the Price of Fuel and Power (FP), Manufacturing Products (MP),and Real Net Per Capita National Income (RNNI) with the domestic sale of cars.

Contents

List of Tables

List of Figures

List of Symbols

$\in$	Belongs to
$\ni$	Such that
$=$	Equals to
$\neq$	Not equals to
$\equiv$	Equivalent to
$>$	Greater than
$<$	Less than
$\geq$	Greater than equals to
$\leq$	Less than equals to
$\forall$	For all
$\Rightarrow$	Implies
$\sim$	Follows
$\sum_{i=1}^{n} Y_i$	$= Y_1 + Y_2 + ... + Y_n$
Δ	Differencing Operator
L	Lag operator
E	Expectation
Var	Variance
Cov	Covariance
SE	Standard error
$N(0, \sigma^2)$	Normal distribution with $mean = 0, variance = \sigma^2$
H_0	Null hypothesis
H_a	Alternative hypothesis
$\hat{\rho}$	Estimation of ρ

Chapter 1

INTRODUCTION

It has been observed that domestic sale of vehicles has increased noticeably in the past year as people are inclined toward cars due to many reasons. First, in India, most of the middle-class and upper middle-class people have the financial ability to purchase family cars. They prefer their own vehicle rather to avail of public transport for a comfortable journey and saving of time. Also, the safe journey is in consideration of present generation in purchasing own vehicles. However, the interesting fact is that in spite of high fuel prices and prices of other manufacturing products the rise in sales of car is unabated.

In this project, we are interested to find if there exists a long-run relationship between the sales of cars in India and variables like fuel price, price of manufacturing goods, and per capita income of the Indian population.

The primary task of this project is to get a consistent data set over the period of our analysis. So we have tackled the challenge of data mining of real life data set where we try to overcome the problem of missing values and try to make a clean data set as well. We have selected some of the major factors responsible for sale of domestic cars. These are as follows:

Export of Domestic Cars:

Export is one of the vital variables that may be responsible for the variation of sale of cars. Intuitively, the models which are exported in high quantity will have a larger chance of being in demand in the country itself because of brand value. So initially we have considered the variable in our model.

Per Capita Net National Income (PCNNI) :

PCNNI is the most precise measurement that can give the most accurate estimation of a country's average income level, which in turn gives us a rough idea of number of people who can afford luxuries of having private cars.

Price Index of Non Food Articles (NF):

We have taken the wholesale price index for the non food articles, since they indirectly affect the affordability of cars.

Price Index of Fuel and Power (FP):

There are several expenditures which come into play after the commodity has been purchased like- the running charge and maintenance charge- which leads us to consider the price of fuel and power. We thus consider the whole- sale price index of fuel and power to see the relationship with domestic sale of cars. Also, the inflation in rate of fuel and power considerably affects the process of manufacturing of cars as the various parts of a car have to be transported from different units to the warehouse.

Price Index of All Commodities (AC):

The wholesale price index of all commodities is what we consider for our purpose as, commodity prices typically rise when inflation accelerates, therefore directly affecting the rate of sale of cars.Also, we have taken this variable to convert the nominal values of per capita net national income into real term.

Real Per Capita Net National Income (RNNI):

For our purpose we first compute the real per capita net national income (RNNI) by dividing our PCNNI with the wholesale price index of all commodities.

Price Index of Food Articles (FA):

Food articles, being a necessity for survival, directly and majorly affect the expenditure capacity of any individual, and ultimately his standards of living which decide the sale of the car.

Price Index of Manufacturing Products (MP):

A manufactured product is a good that is produced mainly by the application of labour and capital to raw materials and other intermediate inputs. They include steel, chemicals, paper, textiles, machinery, clothing, vehicles, and so on. The increase in the price of these eventually increases the manufacturing cost of cars and affect the sale rate.

DATA SOURCE: We have taken a time series of 30 years starting from 1990-91 to 2019-20. Domestic sales of cars are taken from *Annual Reports of Society of Indian Automobile Manufacturers (SIAM).* Values of the other variables (FP, MP, NNI, AC) are directly taken from the official reports published in the *Statistical Handbook* by the *Reserve Bank of India.*

Chapter 2

OBJECTIVES

In our analysis, we are basically interested in finding the long-run relationship of Domestic Car Sales with other variables that we have chosen in our analysis. The analysis has been done for the period from 1990 − 1991 to 2019 − 2020 using Stochastic Time Series Analysis. For this purpose, we have used influencing variables like price indices fuel and power, manufacturing products, non-food articles, primary articles, and per capita real net national income. This analysis is done entirely based upon secondary data of the number of sales of cars available from the Annual Report by the "Society of Indian Automobile Manufactures" and other financial variables are taken from the Statistical Hand Books of the Reserve bank of India.

Since we have a long time series data over the past three decades, we are going to apply a stochastic time series analysis using Vector Error Correction Model (VECM) for finding dynamic relations among the variables.

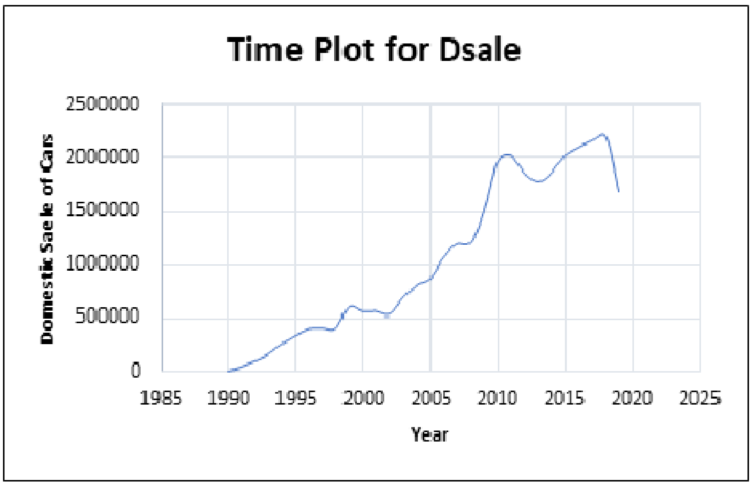

Figure 2.1: Time Trend of Domestic Sales of Car

Over the span of 30 years from 1990 to 2020, there has been a remarkable increase in the sales of Cars, ignoring the slight dip around 2015. We are interested to see the main factors that have contributed to the variation of sales.

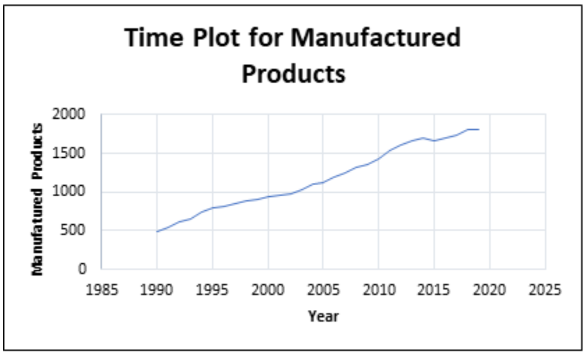

Figure 2.2: Time Trend of Price Index of Manufacturing Products

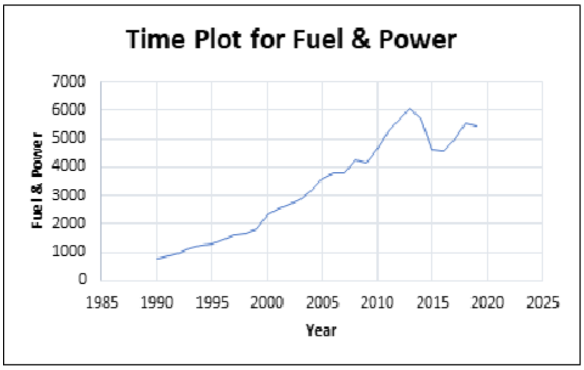

Figure 2.3: Time Trend of Price Index of Fuel and Power

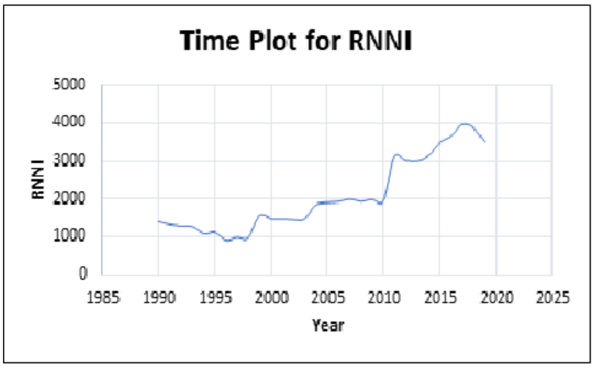

Figure 2.4: Time Trend of RNNI

All of the above time plots show a visibly increasing trend with a few ups and downs over the span of the three decades. Naturally, the visual impression is that these series are all non stationary time series.

The important econometric methods that we have used in this project are Augmented Dickey-Fuller (ADF) analysis, Johansen's Cointegration test, and Vector Error Correction Model (VECM). Before moving to the econometric analysis, we are to follow some data analysis procedures.

For calculation purposes, we have converted all the price indices into a common base year 1972-1973 using the standard splicing method. We have converted each variable in terms of natural logarithmic value. The value of RNNI is calculated as $RNNI$ or $PCRNNI = (PCNNI/AC)100$ [where PCNNI = Per capita Net National Income, AC = Price index of all commodities].

Chapter 3

METHODOLOGY

3.1 Introductory Definitions and Discussions:

- **Stochastic Process:** *A statistical phenomenon defined as: Families of random variables which are functions over time are known as Stochastic(Random) Process. So it is a mapping* $f : T \to Y, \ni \{Y_t : t \in T\}$ *where* Y *is the set of random variables* T *is the time indexing parameter.*

- **Defining Time Series as a Stochastic Process:** *A time series is a variation with time in amplitude and polarity of measured physical quantity.* Based on analytic expression there are two sub categories of time series:
  (a) Deterministic Time Series: It can be expressed explicitly by an analytic expression and it has no random or probabilistic aspects.
  (b) Non-deterministic Time Series: It cannot be an analytic expression and it has some random aspect that prevents its behaviour from being described explicitly.
  Non deterministic time series may be analyzed by assuming that they are the manifestation of stochastic (random) process.

- **Stationary Process:** *The stochastic process* $\{Y_t\}$ *is stationary (or strictly stationary) if , for every collection of time indices* $\{t_j\}_{j=1}^{n}$ *, the joint distribution of* $\{Y_{t_j}\}_{j=1}^{n}$ *is same as the joint distribution of* $\{Y_{t_{j+h}}\}_{j=1}^{n}$ $\forall$ *integers* $h \geq 1$*. It follows that the* Y_i*'s are identically distributed.* In a stationary process we do not observe any trend and systematic changes in variance and it contains no strictly periodic variations.

- **Covariance Stationary Process:** *The stochastic process Y_t is covariance stationary if $\forall$ t and $t-s$,*

  (i) $E(Y_t) = E(Y_{t-s}) = \mu$,

  (ii) $E[(Y_t - \mu)^2] = E[(Y_{t-s} - \mu)^2] = \sigma_Y{}^2$ *[*$Var(Y_t) = Var(Y_{t-s}) = \sigma_Y{}^2$*]*

  (iii) $E[(Y_t - \mu)(Y_{t-s})] = E[(Y_{t-j} - \mu)(Y_{t-j-s} - \mu)] = \gamma_s$ *[*$Cov[(Y_t, Y_{t-s})] = Cov(Y_{t-j}, Y_{t-j-s}) = \gamma_s$*].*

  $*$ **Note:1** A covariance stationary process is referred to as a weakly stationary, second order stationary or wide sense stationary process. A strongly stationary process need not have a finite mean and/or variance.

  $*$ **Note:2** For a stationary process the mean(μ_y) and variance(σ_y^2) do not depend over time, whether for a non-stationary process the mean and variance depend over time (i.e.μ_t and σ_t^2).

  $*$ **Corollary:** Suppose that Z_t is a discrete-time purely random process with mean μ and variance σ_Z^2 and a random walk Y_t is defined as $Y_t = Y_{t-1} + Z_t \ni$ when $t = 0, Y_1 = Z_1$ and $Y_t = \sum_{i=1}^{t} Z_i$ (Z_i's are independent), then Y_t is non- stationary.

- A time series $\{Y_t\}$ can be written as:

$$Y_t = T_t + S_t + I_t + \epsilon_t \quad (3.1.1)$$

  where T_t = value of the trend component in period t,

  S_t = value of the seasonal component in t,

  I_t = the value of the irregular component in t,

  ϵ_t = a pure random disturbance in t.

- **The lag and Difference Operators:** *For a time series Y_t the lag operator is defined as $L(Y_t) = Y_{t-1}$. The lag operator can be defined for linear combinations by $L(aY_{t_1} + bY_{t_2}) = aY_{t_1-1} + bY_{t_2-1}$. In general $L^k Y_t = Y_{t-k}$ and for linear combinations $(aL^k + bL^l)Y_t = aY_{t-k} + bL_{t-l}$*

  The differncing operators can be defined in terms of L by $\Delta Y_t = (1 - L)Y_t = Y_t - Y_{t-1}$

- **Differencing a long memory process:** If $\{Y_t\}$ is a long memory process an old shock to the system (ϵ_{t-h} where h is large) still has an effect on Y_t. When $\{Y_t\}$ is a short memory process the effect of ϵ_{t-h} on Y_t decreases quickly as h increases i.e. the process "forgets" old shocks. In most of the cases differencing a long memory process produces a short memory process. let $Y_t = Y_{t-1} + \epsilon_t$ where ϵ_t is a white noise process.Then,

$$\Delta Y_t = Y_t - Y_{t-1} = \epsilon_t \quad (3.1.2)$$

i.e, the difference between two long memory process becomes a short memory process.

- **Auto Correlation:** *Auto correlation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It measures the relationship between the current value and past value of a variable.*

- **Autocovariance function (ACVF) and Autocorrelation function (ACF):** *Let Y_t be a stationary time series. Then autocovariance function (ACVF) of Y_t at lag s is*

  $$\gamma_y(s) = Cov(Y_{t+s}, Y_t) \quad (3.1.3)$$

  .

  The autocorrelation function (ACF) of Y_t at lag s is

  $$\rho_Y(s) \equiv \frac{\gamma_y(s)}{\gamma_y(0)} = Cor(Y_{t+s}, Y_t) \quad (3.1.4)$$

- The sample autocorrelation function $\hat{\rho_k}$, can be computed for any given time series from a stochastic process. The plot of $\hat{\rho_k}$ vs k is called correlogram.

  For the time series containing a trend the sample auto-correlation function $\hat{\rho_k}$ exhibits slop decay as k increases. For the time series containing a periodic (like seasonality) the sample autocorrelation function $\hat{\rho_k}$, exhibits a behaviour with the same periodicity.Thus $\hat{\rho_k}$, can be useful as an indicator of nonstationarity.

A time series variable may be contained of the four following categories:

- **With no trend:** $Y_t = \epsilon_t$ (3.1.5)
- **With a deterministic trend:** If Y_t contains a deterministic trend then,
  $Y_t = \beta_0 + \beta_1 t + \epsilon_t$ (3.1.6)
- **With a stochastic trend:** If Y_t contains a stochastic trend then,
  $Y_t = \beta_0 + Y_{t-1} + \epsilon_t$ (3.1.7)
- **Contains both deterministic and stochastic trend:** If both deterministic and stochastic trends are presented in Y_t then,
  $Y_t = \rho Z_{t-1} + \beta_0 + \rho Z_{t-1} + \beta_1 t + \epsilon_t$ (3.1.8)

[1] [2]

3.2 Econometric Analysis

Here we are interested to check the non-stationarity of the given variables. Basically there are three ways to check for non-stationarity:

- Using graphical method
- Using ACF and
- Unit root test

The graphical method (time plot) and ACF method are discussed earlier. Now we are to discuss for Unit Root Test.

Unit root test: let us consider the model:

$$y_t = \alpha + \beta y_{t-1} + \epsilon_t \qquad (3.2.1)$$

$*$ **Case:I** If $\beta > 1$, it implies every previous value makes the increment of the current value and ultimately the series will become explosive. So we will not consider this case.
$*$ **Case:II** If $\beta < 1$, it implies that the previous value has a less effect on the current value, and after some lags, it will die out, i.e, there will be no trend and the series will become stationary.
$*$ **Case:III** If $\beta = 1$, it implies that consistently every previous value (lags) is affecting the current values and this is also persistent. This means the series has a trend and it is non-stationary and also the series has a unit root.

From equation (1.1) substracting y_{t-1} from both side we obtain

$$y_t - y_{t-1} = \alpha + (\beta - 1)y_{t-1} + \epsilon_t \qquad (3.2.2)$$

$$\Rightarrow \Delta y_t = \alpha + \gamma y_{t-1} + \epsilon_t \qquad (3.2.3)$$

As, $\gamma = \beta - 1$,
$H_0 : \beta = 1, \gamma = 0$ (implies the series is non-stationary)
vs.
$H_a : \beta < 1, \gamma < 0$ (implies the series is stationary).

This test is known as unit root test and is used to check for stationarity. There are three versions of the unit root test.

Test for a unit root(Dickey-Fuller,1979):

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t \quad (3.2.4)$$

Test for a unit root with an intercept:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \epsilon_t \quad (3.2.5)$$

Test for unit root with constant and deterministic trend :

$$\Delta y_t = \alpha + \phi t + \gamma y_{t-1} + \epsilon_t \quad (3.2.6)$$

To make a good fitness of the model (1.6) more lags are added as independent variables of the dependent variable Δy_t. Now the new model can be written as

$$\Delta y_t = \alpha + \phi t + \gamma y_{t-1} + \sum_{i=1}^{m} \phi_i \Delta y_{t-i} + \epsilon_t \quad (3.2.7)$$

Note: By including lags of the order m the ADF formulation allows for higher-order autoregressive processes. This means that the lag length m has to be determined when applying the test. One possible approach is to test down from high orders and examine the t-values on coefficients. An alternative approach is to examine information criteria such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), or the Hannan–Quinn information (HQI) criterion. The test statistic is defined as: $DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$

Note: Once the test statistic is computed it can be compared to the relevant critical value for the Dickey-Fuller test. As this test is asymmetrical, we are only concerned with negative values of our test statistic DF_τ. If the calculated test statistic is less (more negative) than the critical value, then the null hypothesis of $\gamma = 0$ is rejected and no unit root is present. And if the null hypothesis is accepted then the series is non - stationary thus having a unit root. From this ADF unit root test, it can be determined whether the series are integrated of the same level.

Then in the second step it is needed to find whether the series are cointegrated among themselves or not. Before doing cointegrating the optimal lag has to be calculated. Aikaike information criterion (AIC) 2 lags can be taken. Lower the value of AIC better it will be for the model.

Concept of Cointegration:

It is essentially a time series concept which is concerned with the presence of long-run relationship among time series process Y_t, among two or more variables.

It is necessary that the time series process for individual variables are non - stationary having trend moments. If these trend moments obey same (equilibrium) relationship(s), the variable will have a tendency to drift together over a long period of time and will not move away from one another.

Integrated Process and Order of Integration: A non - stationary process can be transformed to a stationary process by differencing. Suppose Y_t is non stationary and $\Delta^d Y_t$ is stationary, then Y_t is called integrated of order d and denoted as $Y_t \sim$ I(d).

If $Y_t \sim$ I(d) and Z $\sim$ I(e) then, (aY_t + bZ_t) $\sim$ I(d, e). However, in specific case we may have (aY_t + bZ_t) $\sim$ I(0), that is, a linear combination of two given non - stationary process. In fact this happens when the stochastic trends of two variables are related such that they combine to give a stationary process.

Cointegrating Vectors: Consider the regression equation,

$$Y_t = -(b/a)Z_t + \epsilon_t \quad (3.2.8)$$

where Y_t and $Z_t \sim$ I(d). This implies that (aY_t + bZ_t) = ϵ_t. Now, for this regression to be meaningful, ϵ_t must be I(0). Let us take two variables Y_1 and Y_2 which are functionally related as -

$$Y_{1t} = cY_{2t} + \epsilon_{1t} \quad (3.2.9)$$

$$Y_{2t} = Y_{2t-1} + \epsilon_{2t} \quad (3.2.10)$$

where ϵ_1 and ϵ_2 are uncorrelated white noise. c $\neq$ 0. $Y_{2t} \sim$ I(1) as it follows random walk. Taking first difference of the first equation -

$$\Delta Y_{1t} = c\Delta Y_{2t} + \Delta\epsilon_t \sim I(0) \quad (3.2.11)$$

(both are integrated of the same order I(1)), and hence $Y_{1t} \sim$ I(1). Now we can write, $(Y_{1t} - cY_{2t}) = \epsilon_{1t} \sim$ I(0). The variables are cointegrated with cointegrating vectors (1, -c).

In general, if $x_t \sim$ cI(d, b) (where the components of $x_t = (x_{1t}, x_{2t}, ..., x_{nt})'$) and there exists a vector $\beta = (\beta_1, \beta_2, ..., \beta_n)$ such that the linear combination of $\beta_{xt} = \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_n x_{nt}$ is integrated of order (d - b), where $c > 0$. Then the vector β is called cointegrating vector.

Some Important points

- Cointegration refers to a linear combination of non - stationary variables.
- All variables must be integrated of the same order. Of course, this does not imply that all similarly integrated variables are cointegrated.
- A set of k I(1) variables may have 'r' linearly independent cointegrating vectors, $0 \leq r \leq k$.

  If r = 0, the set of variables is said to be non - cointegrated. If r = k, the variables must be all I(0) variables.

  If x has n components, there may be as many as (n - 1) linearly independent cointegrating vectors. Clearly, if x contains only two variables there can be at most one independent cointegrating vector. The number of a cointegrating vector is called the cointegrating rank of x.

Error Correction Models

A principal feature of cointegrated variables is that their time paths are influenced by the extent of any deviation from long run equilibrium. After all, if the system is to return to long run equilibrium, the movements of at least some of the variables must respond to the magnitude of the disequilibrium. Let us consider a simple dynamic model:

$$y_t = \alpha_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \alpha_1 y_{t-1} + \epsilon_t, \epsilon_t \sim IN(0, \sigma^2) \quad (3.2.12)$$

The long run equilibrium between y_t and x_t is given as

$$y_t = \beta_0 + \beta_1 x_t \quad (3.2.13)$$

where $\beta_1 = \frac{\gamma_0+\gamma_1}{1-\alpha_1}$ and $\beta_0 = \frac{\alpha_0}{1-\alpha_1}$

The problems with this specification are -

- likely high level of correlation between current and lagged values of a variable, which will therefore results in problems of multicollinearity.
- Some (if not all) of the variables in a dynamic model of this kind are likely to be non-stationary, since they enter in levels.

- Rectifying this model in terms of first differences removes any information about the long-run relationship from the model and consequently is unlikely to be useful for forecasting purposes.

The dynamic model implied by this discussion is one of error correction. In an error correction model, the short - term dynamics of the variables in the system are influenced by the deviation from equilibrium.

Rearranging and reparameterizing equation (3.2.12) gives:

$$\Delta y_t = \gamma_0 \Delta x_t - (1-\alpha_1)[y_{t-1} - \beta_0 - \beta_1 x_{t-1}] + \epsilon_t \quad (3.2.14)$$

The Error Correction Model (ECM) in equation (3.2.14) has several advantages:

- First, and assuming that x and y are cointegrated, the ECM incorporates both short-run and long-run effects. Thus, if any time the equilibrium holds then $[y_{t-1} - \beta_0 - \beta_1 x_{t-1}] = 0$. During periods of disequilibrium, this term is non-zero and measures the distance, the system is away from equilibrium during time t. Thus, an estimate of $(1-\alpha_1)$ will provide information on the speed of adjustment.

- A second feature of ECM is that all terms in the model are stationary, so standard regression techniques are valid, assuming cointegration and that we have estimates of β_0 and β_1.

- The Simple ECM depicted in equation (3.2.14) can be generalised to capture more complicated dynamic processes by increasing the lag length p and/or q,

  $$A(L)\Delta y_t = B(L)\Delta x_t - (1-\pi)[y_{t-p} - \beta_0 - \beta_1 x_{t-p}] + \epsilon_t \quad (3.2.15)$$

  where A(L) = 1 - $\alpha_1 L - .... - \alpha_p L^p$ and B(L) = $\gamma_0 + \gamma_1 L + ... + \gamma_p L^q$ are the lag operators and $\pi - \alpha_1 + \alpha_2 + ... + \alpha_p$. Further, it is also possible to specify the ECM in multivariate form, explicitly allowing for a set of cointegration vectors.

[3] [4] [5]

Chapter 4

RESULTS and DISCUSSION

4.1 Econometric Analysis:

- **To check stationarity or non-stationarity of the data:**

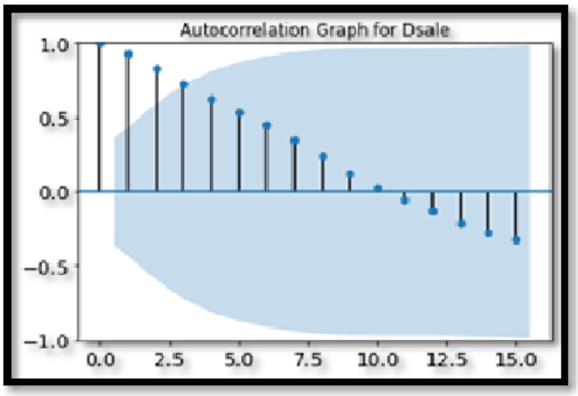

Figure 4.1: Autocorrelation graph for Domestic sale of Cars

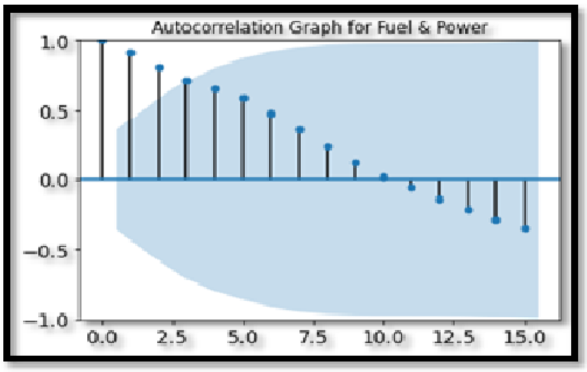

Figure 4.2: Autocorrelation Graph for price index of Fuel and Power

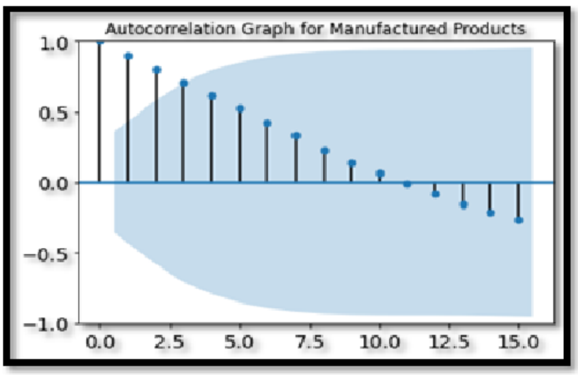

Figure 4.3: Autocorrelation graph for price index of Manufactured products

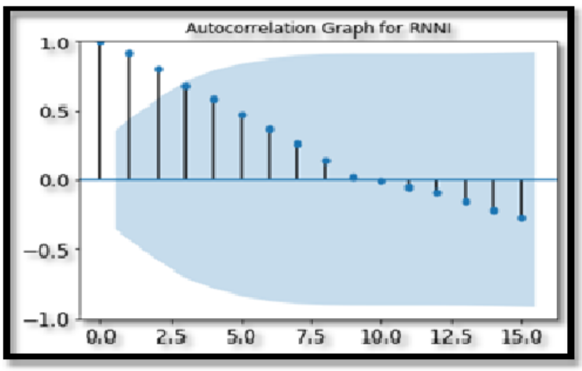

Figure 4.4: Autocorrelation Graph for RNNI

The above ACF plots show that there exists association between the original value with its lagged values as it is evident from the plot that there is a trend present. Hence these plots conform non-stationarity of the variables.

Now, the next step would be to find the lag order and order of differences at which all of the variables become stationary. It is to be kept in mind that all of the variables must have the same lag order and order of differences at which they become stationary or else the analysis will fail.

- **Augmented Dickey-Fuller (ADF) Test:**

  Firstly, the Augmented Dickey-Fuller (ADF) test is used without specifying any lag (i.e., Dickey Fuller test) to check once again the statioarity of the variables. The output below suggests that none of the variables used are stationary as we reject the null hypothesis of stationarity.

Table 4.1: Dickey-Fuller Test

Variable Name	Lag Order	Test Statistic Value	P – Value	Inference	Conclusion
Domestic Sale of Cars (dsale)	3	-1.2373	0.8657	Fail to reject null hypothesis	non-stationary
PCRNNI	3	-3.7772	0.0359	Fail to reject the null hypothesis	non-stationary
Fuel & Power (FP)	3	0.2916	> 0.99	Fail to reject null hypothesis	non-stationary
Manufactured Products (MP)	3	-2.5423	0.3643	Fail to reject null hypothesis	non-stationary

Now, we apply the ADF test but this time lags are specified and we try to see at which lag we can achieve stationarity. We have introduced lag order of 1 and then we have included a difference of 1 and then 2 to achieve acceptance of the null hypothesis of the ADF tests i.e., stationarity of the variables.

Table 4.2: Augmented Dickey-Fuller Test

Variable Name	Lag Order	Test Statistic Value	P–Value	Inference	Conclusion
Domestic Sale of Cars (dsale)	1	-5.203	< 0.01	Reject the null hypothesis	stationary.
PCRNNI	1	-5.0735	< 0.01	Reject the null hypothesis	stationary.
Fuel & Power (FP)	1	-5.3396	< 0.01	Reject the null hypothesis	stationary.
Manufactured Products (MP)	1	-5.0971	< 0.01	Reject the null hypothesis	stationary.

It is seen from the above table that all of the variables are stationary at lag order 1 and order of difference 2 as the null hypothesis is accepted at 1% level of significance. Since the series are integrated of the same differencing order, we can now move on to the next step of the analysis which is Johansen's Co-integration test but only after selecting the optimal lag that needs to be used in the aforementioned test.

- **Optimum Lag Selection:**

  The purpose of choosing optimal lag is to reduce residual correlation. Literature provides various choices such as Akaike's Information Criteria (AIC), Hannah-Quinn (HQ) and Schwarz's Information Criteria (SC) and Final Prediction Error (FPE).

  From the underlying output, using only AIC as our deciding factor, it is seen that for lag order 2, AIC is the minimum most of only $-2.08E+01$ and hence we will be proceeding with a lag order of 2 in our analysis.

Table 4.3: Optimum Lag

	1	2	3	4
AIC(n)	-2.12E+01	-2.08E+01	-2.15E+01	-2.23E+01
HQ(n)	-2.10E+01	-2.03E+01	-2.07E+01	-2.13E+01
SC(n)	-2.03E+01	-1.91E+01	-1.89E+01	-1.90E+01
FPE(n)	6.06E-10	1.01E-09	7.17E-10	5.87E-10

- **Johansen's Co-integration test:**

  Johansen's Co-integration test is a way to determine if three or more time series are co-integrated i.e., whether they have an existing long term relationship. More specifically, it assesses the validity of a co-integrating relationship using a Maximum Likelihood Estimate (MLE) approach. If there exists a relation between two non-stationary series (of same order of Integration) such that the residuals of the regression are stationary, then the variables in question are said to be co-integrated.

  Among the two types of Johansen's test: trace and eigen, we have used trace approach as it is the appropriate one to use here.

  The null hypothesis for the test is that there are no co-integrating relations while the alternate is that there exists atleast one co-integrating relations.

Table 4.4: Johansen's Cointegration Test

		Critical Value		
	Test Statistic Value	10%	5%	1%
$r \leq 3$	5.17	6.50	8.18	11.65
$r \leq 2$	13.30	15.66	17.95	23.52
$r \leq 1$	30.01	28.71	31.52	37.22
$r \leq 0$	51.56	45.23	48.28	55.43

Interpretation of the table: To find the number of co-integrating relations, the method is to find that r where we first fail to reject the null hypothesis. That value of r is the number of existing co-integrating relations, which means at most r number of cointegrating relationships are present in the data. From the above table, using 5% level of significance, at $r = 1$ the null hypothesis was accepted. Hence, we conclude that there exists at most one co-integrating relation.

This test not only examines whether the variables in the model is co-integrated but also determines the number of linearly independent co-integrating vectors and gives the Vector Error Correction Model (VECM).

- **Vector Error Correction Model:**

  VECM is applied when the data is non-stationary. It helps to capture interdependencies of the non-stationary trends. The importance of VECM lies in the interpretation of the results by introducing long term relationship of the variables and then associate the concept of error correction.

Table 4.5: Output of VECM

Full sample size: 30	End sample size: 27	
Number of variables: 4	Number of estimated slope parameters: 40	
AIC: -530.0243	BIC: -474.3034	SSR: 0.7300316

Cointegrating vector (estimated by 2OLS):

	dsale	pcrnni	fp	mp
r1	1	1.382663	-1.444871	-1.772325

Standard Errors:

	dsale -2	pcrnni -2	fp -2	mp -2
	0.05822492	0.04827628	0.08792131	0.72316774

We construct the VECM from the coefficients given in the output as follows-

- **Equation of the fitted VECM:**

$$dsale = 1.38pcrnni - 1.44fp - 1.77mp$$

Here, for 1% increase in the income of the people in India, there is 1.38% increase in domestic sale of cars, whereas for 1% increase in the price of Fuel Power, there is 1.44% decrease in domestic sale of cars. Also, for 1% increase in the price of Manufactured Products, there is 1.77% decrease in domestic sale of cars.

From the above output snippet, we are observing the Standard Error of Estimates of Lag Order 2, as our optimal lag was 2. We can also see that the Standard Error of the coefficients of the VECM is quite low (in between 0 and 1) which means that all the estimates of the coefficients are precise.

We are estimating a VECM model and it was determined by the trace tests that there is one co-integration equation. The independent variables are significant at 5% level of significance.

- **Diagnostic tests for the model:**

  We need to check the adequacy of the fitted model for empirical analysis. To check our model adequacy, we have the following two tests as follows:

- **1. Serial Correlation:**

  To check for Serial Correlation among the residuals, the Portmanteau test which uses autocorrelation or partial autocorrelation in the residuals to criticize the adequacy of the fitted model. Main idea underlying this test is to identify if there is any dependence structure which is yet unexplained by the fitted model. The null hypothesis of this multivariate test is that the autocorrelation function of all series have no significant elements for lags (here a lag order of 1 is specified).

  $*$ **Portmanteau Test (asymptotic test)**

  data: Residuals of VAR object Model1VAR

  $Chi - squared = 14.442$, $df = 4$, $p - value = 0.006009$

  Here, the p-value is < 0.01 and hence our null hypothesis rejected which in turn implies that the ACF of all the series used here has significant elements and thus our model is adequate.

- **2.Normality of residuals:**

  As per standard econometric theory, the error terms of the estimated model must be normally distributed so it becomes mandatory to test the normality property of the residuals.

  Jarque-Bera (JB) test is usually used to check the normality of the residuals (whether residuals are symmetrically distributed or not) in single equation regression. But this may indicate non-normality when applied to a small sample (a sample not as large as $n = 5000$) like the dataset in hand.

  Thus the solution to check the adequacy for a moderately large sample is to plot the histogram of the residuals and check if it is approximately normal or not.

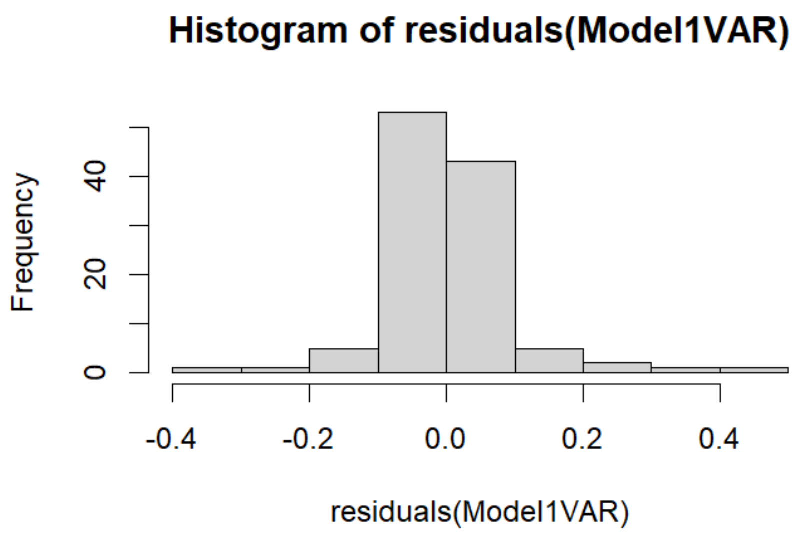

Figure 4.5: Histogram of Residuals

The above plotted histogram of the residuals seems to be approximately normally distributed and thus we can conclude that our fitted model is adequate.

[6] [7]

Chapter 5

CONCLUSION

In our dissertation, we have attempted at understanding how do the chosen variables affect the sale of domestic cars. We have chosen some variables like: Per Capita Real Net National Income, and price index of Non-Food Articles, price index of Fuel and Power, All Commodities combined, Food Articles, Manufactured Products, and also the Primary Articles intuitively. However, after our initial exercise with all these independent variable we consider only three independent variables namely, Price indices of Fuel and Power, Manufacturing products and Per Capita Real Net National Income.

To adjust Per Capita Net National Income with the rate of inflation, we defalte it by corresponding wholesale price index of All Commodities(AC). We name it Real Net National Income (RNNI).

Although we were looking forward to work with a vast dataset, however, we learned that we could proceed with only three variables, namely, RNNI, wholesale price index of Manufactured Products and price index of Fuel and Power. The remaining variables are no longer considerable because of the following two reasons:

- **Order of Integration:** While performing Augmented Dickey Fuller Test it was found that the order of integration was different for the variables not considered for our analysis.
- **Long term relationship:** While performing Co-Integration Test it was found that there was an absence of long-term relationship.

Overall, we observed that there is long run relationship between domestic sale of cars and these three variables: Real Per Capita Net National Income, price index of Fuel and Power and Manufactured Products. This implies that these variables have a long-term relationship towards which they always come back. Since these variables had trend movements, they will have a tendency to move together.

It has also been found from the analysis that RNNI has positive correlation with the increase in sale of domestic cars, whereas the other two explanatory variables MP and FP have negative correlation with our dependent variable. This implies that when there is an increase in the income of the citizens of our country, the sale of cars will increase. However, inflation in the price of Fuel and Power as well as Manufactured products will oversee a decline in the sale of domestic cars.

Bibliography

[1] W. Enders, "Applied econometric time series," *Privredna kretanja i ekonomska politika*, vol. 132, p. 93, 2012.

[2] R. S. Tsay, *Multivariate time series analysis: with R and financial applications.* John Wiley & Sons, 2013.

[3] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting.* Springer, 2002.

[4] W. Enders, *Applied econometric time series.* John Wiley & Sons, 2008.

[5] G. S. Maddala and I.-M. Kim, "Unit roots, cointegration, and structural change," 1998.

[6] W. A. Woodward, H. L. Gray, and A. C. Elliott, *Applied time series analysis with R.* CRC press, 2017.

[7] P. S. Cowpertwait and A. V. Metcalfe, *Introductory time series with R.* Springer Science & Business Media, 2009.