SAMPLE
EVALUATED
PROJECT
REPORTS

A Causal Inference Analysis of Average Treatment Effects using IPTW and TMLE

Project Work

Submitted to

PONDICHERRY UNIVERSITY

in partial fulfilment of the requirements

for the award of the degree of

Master of Science

in

Statistics

BY

AKRITI SRIVASTAVA

Reg No: 21375005

DEPARTMENT OF STATISTICS

PONDICHERRY UNIVERSITY

PONDICHERRY

MAY, 2023

PONDICHERRY UNIVERSITY

R.V. NAGAR, KALAPET,

PUDUCHERRY-605014

Certified that the project work entitled **A Causal Inference Analysis of Average Treatment Effects using IPTW and TMLE** is a bonafide record of work carried out by the following student.

AKRITI SRIVASTAVA , 21375005

of M.Sc. (Statistics) submitted in partial fulfilment of the requirement for the award of degree of Master of Science in Statistics, during the academic year **2021-23**.

(Supervisor) Head of the Department

Submitted for M.Sc. Degree Examination held on _____

Examiners

1.

2.

GSK India Global Services Pvt. Ltd.
Level 1, 2 & 3 Luxor North Tower
Bagmane Capital Business Park
Outer Ring Road, Mahadevpura
KR Puram Hobli
Bangalore - 560037

To Whomsoever it may concern

Date: 18/05/2023

This is to certify that **Akriti Srivastava** pursuing internship with GSK from 23 Jan 2023 and will be completing on 24 Jul 2023

He/She has worked and completed below project with our organization.

Project Title: Causal Inference Analysis of Average Treatment Effects using IPTW and TMLE.

Project Guide: Ramiya Ravindranath.

We acknowledge your commitment and contribution to our organization and wish you all the very best for your future endeavor

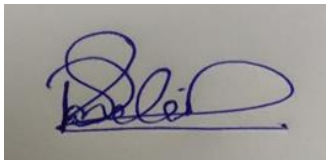
Yours sincerely

Nilesh Kumar

Human Resources Business Lead- Global Capability Centre

CERTIFICATE

This is to certify that **Akriti Srivastava** (**21375005**), a student of M.Sc. Statistics from the Department of Statistics, Pondicherry University has completed her fourth semester project under our guidance at **GlaxoSmithKline (GSK), Bengaluru** from **January to May, 2023**. The project work entitled **"A Causal Inference Analysis of Average Treatment Effects using IPTW and TMLE"** embodies the novel work done by her.



_Signature_____

Dr R Vishnu Vardhan
Associate Professor
Department of Statistics
Pondicherry University
Puducherry – 605014
(Faculty Guide)

Sebin Thomas
Lead Statistician, Statistics
Dev Biostats, India
GSK, Bangalore
Bangalore - 560001
(Guide)

ACKNOWLEDGEMENT

I take this opportunity to acknowledge my deep sense of gratitude to ***Dr. R. Vishnu Vardhan*** **and Dr. V.S. Vaidyanathan**, Coordinators, Internships and Placements, Department of Statistics, Pondicherry University for providing this opportunity to work with the reputed company **GlaxoSmithKline (GSK), Bengaluru**.

I am thankful to my guide, **Dr. R. Vishnu Vardhan,** Associate Professor, Department of Statistics, Pondicherry University, Puducherry, for his guidance and overall support throughout the project without which it would have been impossible to complete this work.

I express my sincere thanks to **Mr. Sebin Thomas**, Lead Statistician, GSK, Bengaluru for providing invaluable guidance throughout the project. His sincerity, dynamism, motivation has inspired me. It was a great privilege to work under his guidance.

I am thankful to **Mrs. Ramiya Ravindranath**, Senior Manager Statistics, GSK, Bengaluru for her generous support.

I express my sincere gratitude to **Dr. Kiruthika,** Professor and Head of Department of Statistics, Pondicherry University, Puducherry for her valuable support and **Dr. Angshuman Sarkar,** Head of the R&D Division, GSK, Bengaluru for this wonderful opportunity to work with them.

I am thankful to **Prof. P. Tirupathi Rao**, Dean, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry for his generous support.

I am very thankful to all the faculty members of the Department **Dr. Navin Chandra, Dr. Sudesh Pundir and Dr. J. Prabhakara Naik** for their generous and moral support.

I wish to place on record the valuable support of my family members without whose support it would have been impossible to complete this project.

I am thankful to my friends for their cooperation and support in bringing out this project as a thriving one.

Akriti Srivastava

CONTENTS

ABSTRACT

To be confident in the leap of faith, we have to accept that our old ways of turning data analysis into business recommendations was a bit naive. We had heard that correlation is not causation, but we effectively ignored that distinction. At best, we admitted that we were uncertain, dug a little more, or compared our results to intuition to build confidence in our conclusions. But this unstructured way of drawing conclusions is subject to biases, especially if we don't exactly know what we are doing. Hence analyzing causal effects is very important.

In this paper we have tried to introduce causal inference and related concepts from basics and took it forward into an application-based study on the two methods, for determining the estimate (ATE), namely IPTW (Inverse Probability Treatment Weighting) and TMLE(Targeted Maximum Likelihood Estimator).

The methodological and implementation approach for both the methods are explored with the help of a simple simulation study and the bias reduction in the TMLE method has been portrayed. Given TMLE's appealing statistical properties, we consider it as a suitable method to be added to the analytical toolbox for estimation of causal effects in large population-based observational studies.

CHAPTER 1 : INTRODUCTION AND OBJECTIVES

1.1 Causal Inference

A causal relationship is a relationship of cause and effect. A line of reasoning uses causal relationships to draw a conclusion. By exploring causal relationships, we can study the difference between fact and opinion.

Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system. The main difference between causal inference and inference of association is that causal inference analyzes the response of an effect variable when a cause of the effect variable is changed.

We study causation because we need to make sense of data, to guide actions and policies, and to learn from our success and failures. We need to estimate the effect of smoking on lung cancer, of education on salaries, of carbon emissions on the climate. Most ambitiously, we also need to understand how and why causes influence their effects, which is not less valuable.

Causal inference and **correlation** are related concepts, but they are fundamentally different. Correlation refers to the strength and direction of the relationship between two variables, while causal inference seeks to establish whether that relationship is causal. A causal relationship is so powerful that it gives enough confidence in making decisions, preventing losses, solving optimal solutions, and so forth.

1.2 RCT vs Observational

A **randomized clinical trial (RCT)** is an experiment where every person in a trial is randomly assigned to either a treatment group or a control group. A clinical trial is performed in a controlled setting (e.g., a clinic or hospital), targets a specific disease, and has an event as a measure of trial outcome (e.g., cure/no cure).

Randomized controlled trials (RCTs) are considered the gold standard for studying the efficacy of an intervention. Randomization highly increases the likelihood that both intervention and control groups have similar characteristics and that any remaining differences will be due to chance, effectively eliminating confounding. Any difference in the outcome between groups can then be attributed to the intervention and the effect estimates may be interpreted as causal.

Under randomization, association does imply causation (of course within the potential outcome framework with assumptions).

However, many research questions cannot be studied in RCTs, as they can be too expensive and time-consuming (especially when studying rare outcomes), tend to include a highly selected population (limiting the generalizability of results) and in some cases randomization is not feasible (for ethical reasons).

Observational studies are a fundamental part of epidemiological research. They are called observational studies because the investigator observes individuals without manipulation or intervention.

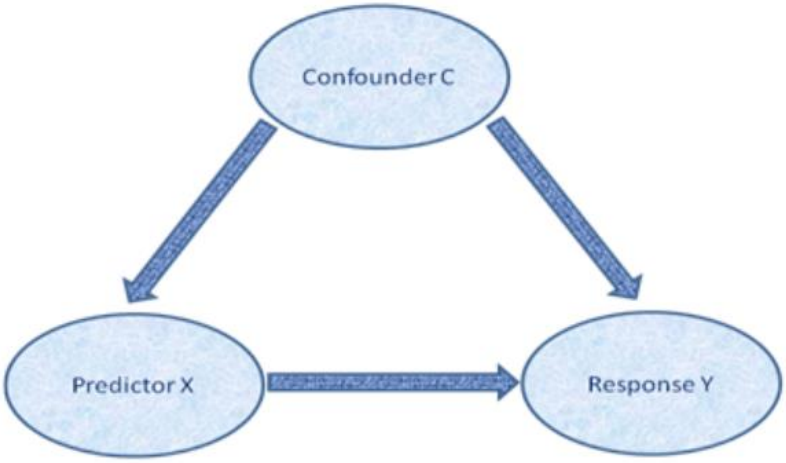
Observational studies suffer less from limitations, as they simply observe unselected patients without intervening. Observational research may be highly suited to assess the impact of the exposure of interest in cases where randomization is impossible, for example, when studying the relationship between body mass index (BMI) and mortality risk.

1.3 Confounding Effect

Because of the lack of randomization, a fair comparison between the exposed and unexposed groups is not as straightforward due to measured and unmeasured differences in characteristics between groups. Certain characteristics that are a common cause of both the observed exposure and the outcome may obscure—or confound—the relationship under study, leading to an over, or underestimation of the true effect.

A confounder is thus a third variable not the exposure, and not the outcome, that biases the measure of association we calculate for the particular exposure/outcome pair.

For example, suppose a researcher collects data on ice cream sales and shark attacks and finds that the two variables are highly correlated. Does this mean that increased ice cream sales cause more shark attacks? That’s unlikely. The more likely cause is the confounding variable **temperature**. When it is warmer outside, more people buy ice-cream and more people go in the ocean.



To control for confounding in observational studies, various statistical methods have been developed that allow researchers to assess causal relationships between an exposure and outcome of interest under strict assumptions. Besides traditional approaches, such as multivariable regression and stratification, other techniques based on so-called propensity scores, have been increasingly used in the literature.

1.4 Simpsons Paradox

Simpson’s paradox is a fascinating phenomenon that illustrates the importance of causality in reasoning. Simpson’s Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations. For instance, two variables may be positively associated in a population, but be independent or even negatively associated in all subpopulations.

Simpson’s paradox reminds researchers that causal inferences, particularly in nonexperimental studies, can be hazardous. Uncontrolled and even unobserved variables that would eliminate or reverse the association observed between two variables might exist.

A common example of Simpson's paradox involves the batting averages of players in professional baseball. It is possible for one player to have a higher batting average than another player each year for a number of years, but to have a lower batting average across all of those years.

In the classical example used by Simpson (1951), a group of sick patients are given the option to try a new drug. Among those who took the drug, a lower percentage recovered than among those who did not. However, when we partition by gender, we see that more men taking the drug recover than do men are not taking the drug, and more women taking the drug recover than do women are not taking the drug! In other words, the drug appears to help men and women, but hurt the general population. It seems nonsensical, or even impossible—which is why, of course, it is considered a paradox.

1.5 Propensity Score

The propensity score (PS) was first defined by Rosenbaum and Rubin in 1983 as 'the conditional probability of assignment to a particular treatment given a vector of observed covariates'. In other words, the propensity score gives the probability (ranging from 0 to 1) of an individual being exposed (i.e., assigned to the intervention or risk factor) given their baseline characteristics.

We use the propensity score (or probability of getting a treatment given a set of covariates) as a balancing score. A balancing score is any function of the set of covariates that captures all the information of the set that is dependent on treatment. Such a balancing score would allow us to model the relation between the confounders and treatment in a relatively simple way. And the minimal expression of a balancing score is the propensity score.

Propensity score analysis typically involves two stages:

Stage 1- Estimate the propensity score, by e.g., a logistic regression or a machine learning method

Stage 2- Given the estimated propensity score, estimate the causal effects through one of these methods: | Stratification | Weighting | Matching| Regression | Mixed procedure of the above.

The aim of the propensity score in observational research is to control for measured confounders by achieving balance in characteristics between exposed and unexposed groups.

The basic idea of propensity score is to focus on the prediction of treatments rather than on outcomes and to replace the confounding variables that play a role in the choice of a given treatment with a function of these covariates.

Two people receiving different treatments have the same propensity value, it means that they would have the same probability to receive treatment 1 or 2, randomly, and thus they may be more confidently compared.

Statistical Definition of PS:

The estimated propensity score, for subject *i, (i = 1,…, N)* is the conditional probability of being assigned to a particular treatment given a vector of observed covariates *xi:*

$$e(x_i) = \Pr(z_i=1/x_i)$$

where,

- z_i= 1, for treatment
- z_i= 0, for control
- x_i, the vector of observed covariates for the i^{th}subject

Since the propensity score is a probability, it ranges in value from 0 to 1.

Idea is that if we combine the individual confounders in a summary measure (propensity score) so that if you just control for propensity score indirectly you will be controlling for confounding. The propensity score–based methods are thus able to summarize all characteristics to a single covariate (the propensity score) and may be viewed as a data reduction technique. These methods are therefore warranted in analyses with either a large number of confounders or a small number of events.

Methods of estimating Propensity Score:

a) **Logistic regression**

- ❖ Logistic regression is the most commonly used method for estimating propensity score
- ❖ The model is used to predict the probability that an event occurs.

$$\ln\frac{e(x_i)}{1-e(x_i)} = \ln\frac{\Pr(z_i=1\mid x_i)}{1-\Pr(z_i=1\mid x_i)} = \alpha+\beta\ x_i$$

where,

$$e(x_i) = \Pr(z_i=1\mid x_i)$$

b) **Classification and regression tree (CART)**

- **CART (Classification And Regression Tree)** is a variation of the decision tree algorithm. It can handle both classification and regression tasks.
- CART was first produced by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone in 1984.
- Not widely used as Logistic regression for estimating propensity scores because it may not be as readily understood
- CART has advantageous properties for estimating propensity scores, including the ability to handle categorical, ordinal, continuous, and missing data.

CHAPTER 2 : METHODOLOGIES

2.1 Inverse probability of treatment weighting (IPTW)

Propensity Score Analysis Methods involves four methods majorly:

- Matching
- Stratification
- Covariate Adjustment
- Inverse probability of treatment weighting (IPTW)

We are focusing on the IPTW method in our study.

Key concepts of IPTW:

- Inverse probability of treatment weighting (IPTW) can be used to adjust for confounding in observational studies. IPTW uses the propensity score to balance baseline patient characteristics in the exposed and unexposed groups by weighting each individual in the analysis by the inverse probability of receiving the actual exposure.
- It is considered a good practice to assess the balance between exposed and unexposed groups for all baseline characteristics both before and after weighting.
- An important methodological consideration is that of the extreme weights. These can be dealt with either weight stabilization or weight truncation or both.
- To adjust for confounding measured over time in the presence of treatment-confounder feedback, IPTW can be applied to appropriately estimate the parameters of marginal structural model. Weights are calculated at each time point as the inverse probability of receiving his/her exposure level, given an individual's previous exposure history, the previous values of time-dependent confounder and baseline confounders.
- In time-to-event analyses, inverse probability of censoring weights can be used to account for informative censoring by upweighting those remaining in the study, who have similar characteristics to those who were censored.

Statistical Methodology of IPTW:

- let Z_i be an indicator variable denoting whether the *i*th subject was treated; furthermore, let e_i denote the propensity score for the *i*th subject. Weights can be defined as

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

- A subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received.

The application of these weights to the study population creates a pseudopopulation in which confounders are equally distributed across exposed and unexposed groups.

Some more facts about IPTW:

- Inverse probability of treatment weighting was first proposed by Rosenbaum (1987)
- IPTW has been used in observational studies to reduce selection bias
- IPTW may be sensitive to whether the propensity score has been accurately estimated
- These weights assure that for each combination of baseline characteristics, the sum of contributions of all experimental and control patients are equal
- Subjects in the comparison group who are more similar to those in the treatment group are given greater weight and those more dissimilar are downweighted. If the propensity scores are properly estimated, then the weighted covariate distributions between treatment groups should be similar and the average treatment effect can be estimated as the difference of weighted means.

Advantages and Limitations:

One of the **advantages** of using propensity score weighting, as opposed to matching, is that you're able to include all patients; none of the patients are excluded because they can't be matched to a patient in the other treatment arm. Including all the patients is especially

important when you have small sample sizes. Propensity score weighting allows you to leverage information from all patients included in your sample. If we want to estimate the average treatment effect assuming that every patient (both treated and comparison group patients) in the population would otherwise be offered the treatment, which is known as the average treatment effect (ATE), we use IPTW. In addition, whereas matching generally compares a single treatment group with a control group, IPTW can be applied in settings with a categorical or continuous exposures. Also, compared with propensity score stratification or adjustment using the propensity score, IPTW has shown to estimate hazard ratios with less bias.

IPTW also has **limitations**. Some simulation studies have demonstrated that depending on the setting, propensity score– based methods such as IPTW perform not better than multivariable regression, and others have cautioned against the use of IPTW in studies with sample sizes of <150 due to underestimation of the variance (i.e., standard error, confidence interval and P-values) of effect estimates. The IPTW is also sensitive to misspecifications of the propensity score model, as omission of interaction effects or misspecification of functional forms of included covariates may cause imbalanced groups, biasing the effect estimate.

2.2 Average Treatment Effect (ATE)

Assume that Yi denotes the outcome variable measured on the ith participant.

- For each subject, the effect of treatment is defined to be

  $Y_i(1) - Y_i(0)$

- The average treatment effect (ATE) is defined to be

  $\mathrm{E}[Y_i(1) - Y_i(0)]$

The ATE is the average effect, at the population level, of moving an entire population from untreated to treated.

It is the mean difference in outcomes in a world in which everyone had received the treatment compared to a world in which everyone had not.

Average Treatment Effect For IPTW:

Assume that Yi denotes the outcome variable measured on the ith participant. The estimate of the average treatment effect (ATE) is,

$$\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e_i}$$

where n denotes the number of participants in the full sample.

2.3 Targeted Maximum Likelihood Estimation (TMLE)

- Targeted Learning was proposed by van der Laan & Rubin in 2006 as an automated causal inference method.
- TMLE is used to analyze observational data from a non-controlled experiment in a way that allows effect estimation even in the presence of confounding factors.
- Targeted Maximum Likelihood Estimation (TMLE) is a semiparametric estimation framework to **estimate a statistical quantity of interest**.
- Semiparametric estimation methods like TMLE can rely on machine learning to avoid making unrealistic parametric assumptions about the underlying distribution of the data (e.g., multivariate normality).

It includes the following steps majorly:

We consider Y is binary outcome, (which can also be continuous with a slight change in the algorithm) ; W is confounders (W1,W2,W3,..,Wn). A is Treatment (binary exposure of interest) – "control"-> A=0 ; "treatment"->A=1

Step 1: Estimate the Outcome using

$$Q(A, \mathbf{W}) = \mathrm{E}[Y|A, \mathbf{W}]$$

Step 2: Estimate the Probability of Treatment (propensity score)

$$g(\mathbf{W}) = \Pr(A = 1|\mathbf{W})$$

This step involves defining a function called clever covariate, which is given by

$$H(A, \mathbf{W}) = \frac{\mathrm{I}(A = 1)}{\Pr(A = 1|\mathbf{W})} - \frac{\mathrm{I}(A = 0)}{\Pr(A = 0|\mathbf{W})}$$

Step 3: Estimate the Fluctuation Parameter (ϵ) (it provides information about how much to change, or fluctuate, our initial outcome estimates.) Optimizes the Bias-Variance Tradeoff.

$$logit(\mathrm{E}[Y|A,\mathbf{W}]) = logit(\hat{\mathrm{E}}[Y|A,\mathbf{W}]) + \epsilon H(A,\mathbf{W})$$

Step 4: Update the Initial Estimates of the Expected Outcome

Step 5: Compute the Statistical Estimand of Interest

We can compute the ATE as the mean difference in the updated outcome estimates under treatment and no treatment:

$$\hat{ATE}_{TMLE} = \hat{\Psi}_{TMLE} = \frac{1}{N}\sum_{i=1}^{N}[\hat{E}^*[Y|A=1,\mathbf{W}] - \hat{E}^*[Y|A=0,\mathbf{W}]]$$

Basic Introduction to SuperLearners:

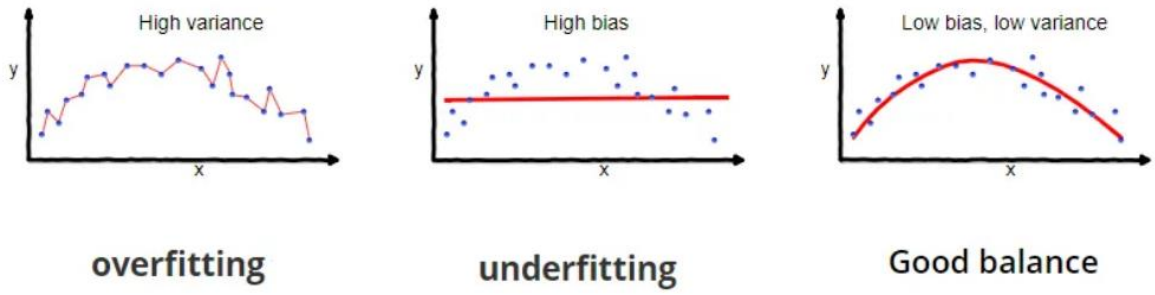
- Superlearning is a technique for prediction that involves combining many individual statistical algorithms (commonly called "machine learning" algorithms) to create a new, single prediction algorithm that is expected to perform at least as well as any of the individual algorithms.
- Superlearning is also called stacking, stacked generalizations, and weighted ensembling by different specializations within the realms of statistics and data science.
- The motivation for this type of ensembling is that a mix of multiple algorithms may be more optimal for a given data set than any single algorithm. For example, superlearners like random forests and LASSO improve predictive performance of the model.

Bias-Variance Tradeoff

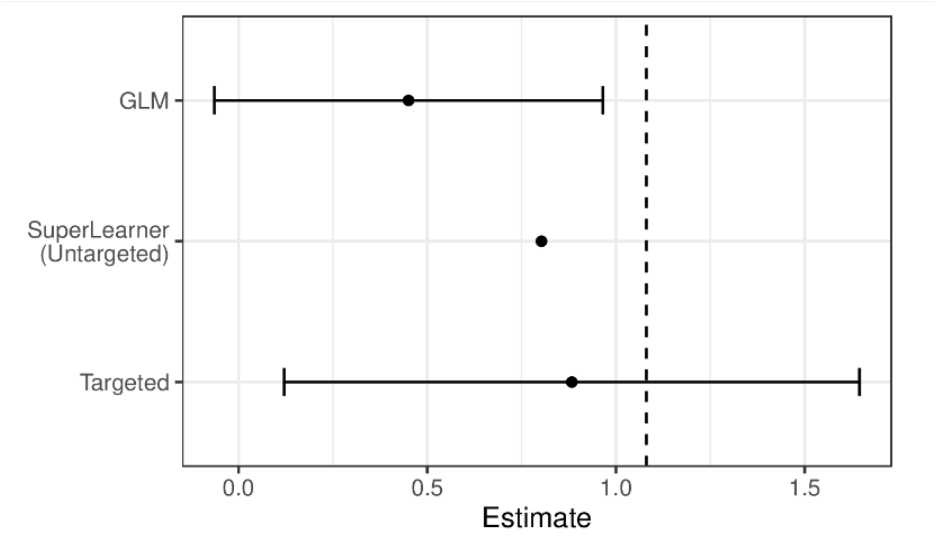
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model.

Whereas model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.

So, we need to find the right/good balance without overfitting and underfitting the data. This tradeoff in complexity is why there is a tradeoff between bias and variance.



Predictive Ability of TMLE is shown by a graphical example:



We see that Super Learner, estimates the true parameter value (indicated by the dashed vertical line) more accurately than GLM. Still, it is still less accurate than TMLE, and valid inference is not possible. TMLE achieves a less biased estimator and valid inference.

TMLE does not balance the covariates or adjust the sample in anyway but it estimates potential outcomes for each individual using an outcome model and adjusts the difference

in the estimated potential outcome means using a function of the propensity score (called the "**clever covariate**").

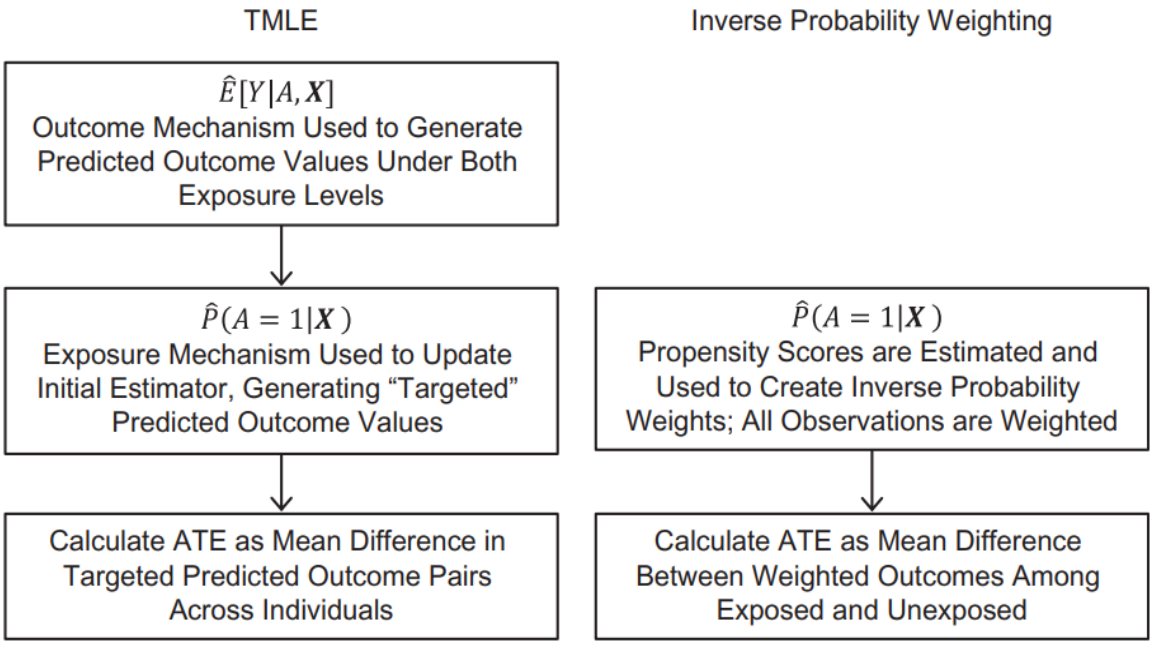
In Targeted Maximum Likelihood Estimation (TMLE), the **fluctuation parameter** is used to adjust for variability in the outcome variable that is not explained by the treatment and covariates. The purpose of the fluctuation parameter is to reduce bias in the estimation of the treatment effect, particularly in settings where there may be unmeasured confounding or other sources of unexplained variability.

The TMLE algorithm involves iteratively estimating the outcome regression model and the treatment probability model, and at each iteration, updating the estimate of the treatment effect based on the current estimates of these models. The fluctuation parameter is added to the outcome regression model at each iteration as a way to capture the unexplained variability in the outcome that is not accounted for by the treatment and covariates.

By including a fluctuation parameter in the outcome regression model, TMLE can account for potential misspecification of the model and improve the efficiency and robustness of the treatment effect estimate. However, it is important to note that the choice of the fluctuation parameter can have an impact on the results and it may be necessary to try different values or functional forms of the parameter to optimize the performance of the TMLE algorithm.

Two frequently used alternatives to estimating the ATE are G-computation and Inverse Probability of Treatment Weighting. In general, neither of them yield valid standard errors unless *a-priori* specified parametric models are used, and this reliance on parametric assumptions can bias the results. There are many simulation studies that show this. TMLE is found to perform better than these.

2.4 TMLE vs IPTW



CHAPTER 3 : ANALYSIS AND INTERPRETATION

3.1 Data Generation

- We plan to estimate the ATE for cancer patients treated with monotherapy (A = 1) versus dual therapy (A = 0) while controlling for confounder (W) and outcome as death(Y).
- We consider W as confounder Age.
- Treatment and Outcome are considered to be binary whose probability is defined using inverse of the logit function.

Controlling for confounding bias is crucial in causal inference study. Distinct methods are currently employed to mitigate the effects of confounding bias. We conduct a simulation study to compare the relative performance results obtained by using IPTW and TMLE method to estimate the average treatment effect. Our simulations are in the context of a binary treatment, a binary outcome and a baseline confounder.

Cancer treatment is independent of the potential mortality outcomes after conditioning on W. Also assume that within strata of W, every patient had a nonzero probability of receiving either of the 2 treatment conditions, i.e., $0 < P(A = 1|W) < 1$. We assume consistency and noninterference, meaning that the counterfactual outcome of one subject was not influenced by the treatment of any other. If we believe these assumptions to hold and the sample size to be sufficient, we may interpret our estimate of the ATE.

W is generated as a Bernoulli variable with probability 0.65. The treatment variable and the potential outcomes were generated as binary indicators using log-linear models.

First, we generated a sample of 5 million patients to estimate the true ATE. Afterwards, we generated a sample of 10,000 patients used to illustrate the implementation of the algorithm and run simulations.

The true ATE implies that the risk of death among cancer patients treated with monotherapy is *calculated-percentage* higher than for those treated with dual therapy considering only one confounder that is age.

At the end of the illustration, we present the results of 1000 Monte Carlo simulations with a sample size of 1000 patients aiming to calculate the mean bias for both IPTW and TMLE method.

3.2 Final R-Code

```
# Load the required packages
library(earth)
library(tidyverse) # for data manipulation
library(SuperLearner) # for ensemble learning
library(WeightIt)# for weighting
library(dplyr)
library(magrittr)

set.seed(7) # for reproducible results
sl_libs <- c('SL.glmnet', 'SL.ranger', 'SL.earth') # a library of machine learning algorithms (penalized regression, random forests, and multivariate adaptive regression splines)

generate_data <- function(n){
  W2 <- rbinom(n, size = 1, prob = 0.65) # binary confounder W2
  A  <- rbinom(n, size = 1, prob = plogis(-5 + 0.05*W2 )) # binary treatment depends on confounders
#plogis(x) represents the inverse logit function
  # counterfactual
  Y.1 <- rbinom(n, size = 1, prob = plogis(-1 + 1 + 0.35*W2))
  Y.0 <- rbinom(n, size = 1, prob = plogis(-1 + 0 + 0.35*W2))
  # observed outcome
  Y <- Y.1*A + Y.0*(1 - A)
  return(tibble(Y, W2, A,Y.1,Y.0))
}
```

```
# True ATE
set.seed(7777)
ObsData <- generate_data(n = 5000000)
True_EY.1 <- mean(ObsData$Y.1)
True_EY.0 <- mean(ObsData$Y.0)
True_ATE <- True_EY.1 - True_EY.0; True_ATE
```
 # 0.238578 TRUE_ATE

```
# Initialize arrays to store iptw ate
iptwate_arr <- numeric(1000)
# Initialize arrays to store tmle ate
tmleate_arr <- numeric(1000)
# Initialize arrays to store biases
bias1_arr <- numeric(1000)
bias2_arr <- numeric(1000)

# Monte Carlo simulation loop
for(i in 1:1000) {
# Data for simulation
ObsData <- generate_data(n = 10000)
Y <- ObsData$Y

# Estimate the treatment weights using IPTW
weights <- weightit(A ~ W2 , data = ObsData, method = "ps", estimand = "ATE")

# Calculate the ATE using the weighted data
weighted <- ObsData %>% mutate(weight = weights$weights)
weighted_summary <- weighted %>% group_by(A) %>% summarize(mean_Y = weighted.mean(Y, weight))
```

```
iptw_ate <- weighted_summary[weighted_summary$A == 1, "mean_Y"] - weighted_summary[weighted_summary$A == 0, "mean_Y"]

iptw_ate=as.numeric(iptw_ate)


#TMLE

W_A <- dplyr::select(ObsData, -Y,-Y.1,-Y.0) # remove the outcome to make a matrix of predictors (A, W1, W2, W3, W4) for SuperLearner


### Step 1: Estimate Q


Q <- SuperLearner(Y = Y, # Y is the outcome vector

                  X = W_A, # W_A is the matrix of W1, W2, W3, W4, and A

                  family=binomial(), # specify we have a binary outcome

                  SL.library = sl_libs) # specify our superlearner library of LASSO, RF, and MARS

Q_A <- as.vector(predict(Q)$pred) # obtain predictions for everyone using the treatment they actually received

W_A1 <- W_A %>% mutate(A = 1)  # data set where everyone received treatment

Q_1 <- as.vector(predict(Q, newdata = W_A1)$pred) # predict on that everyone-exposed data set

W_A0 <- W_A %>% mutate(A = 0) # data set where no one received treatment

Q_0 <- as.vector(predict(Q, newdata = W_A0)$pred)

dat_tmle <- tibble(Y = ObsData$Y, A = ObsData$A, Q_A, Q_0, Q_1)


### Step 2: Estimate g and compute H(A,W)


A <- ObsData$A

W <- dplyr::select(ObsData, -Y,-A) # matrix of predictors that only contains the confounders W1, W2, W3, and W4

g <- SuperLearner(Y = A, # outcome is the A (treatment) vector

                  X = W, # W is a matrix of predictors

                  family=binomial(), # treatment is a binomial outcome
```

```
                SL.library=sl_libs) # using same candidate learners; could use different learners


g_w <- as.vector(predict(g)$pred) # Pr(A=1|W)

H_1 <- 1/g_w

H_0 <- -1/(1-g_w) # Pr(A=0|W) is 1-Pr(A=1|W)

dat_tmle <- # add clever covariate data to dat_tmle

  dat_tmle %>%

  bind_cols(

    H_1 = H_1,

    H_0 = H_0) %>%

  mutate(H_A = case_when(A == 1 ~ H_1, # if A is 1 (treated), assign H_1

                         A == 0 ~ H_0))  # if A is 0 (not treated), assign H_0


### Step 3: Estimate fluctuation parameter

glm_fit <- glm(Y ~ -1 + offset(qlogis(Q_A)) + H_A, data=dat_tmle, family=binomial) # fixed intercept logistic regression

eps <- coef(glm_fit) # save the only coefficient, called epsilon in TMLE lit


### Step 4: Update Q's

H_A <- dat_tmle$H_A # for cleaner code in Q_A_update

Q_A_update <- plogis(qlogis(Q_A) + eps*H_A) # updated expected outcome given treatment actually received


Q_1_update <- plogis(qlogis(Q_1) + eps*H_1) # updated expected outcome for everyone receiving treatment

Q_0_update <- plogis(qlogis(Q_0) + eps*H_0) # updated expected outcome for everyone not receiving treatment


### Step 5: Compute ATE

tmle_ate <- mean(Q_1_update - Q_0_update) # mean diff in updated expected outcome estimates
```

```
### Step 6: compute standard error, CIs and pvals

infl_fn <- (Y - Q_A_update) * H_A + Q_1_update - Q_0_update - tmle_ate # influence function
tmle_se <- sqrt(var(infl_fn)/nrow(ObsData)) # standard error
conf_low <- tmle_ate - 1.96*tmle_se # 95% CI
conf_high <- tmle_ate + 1.96*tmle_se
pval <- 2 * (1 - pnorm(abs(tmle_ate / tmle_se))) # p-value at alpha .05

#Calculating Bias
bias1 <- abs(iptw_ate - True_ATE)
bias2 <- abs(tmle_ate - True_ATE)

# Store iptw ate in arrays
iptwate_arr[i] <- iptw_ate
# Store tmle ate in arrays
tmleate_arr[i] <- tmle_ate
# Store biases in arrays
bias1_arr[i] <- bias1
bias2_arr[i] <- bias2
}

# Calculate mean iptw ate
mean_iptw <- mean(iptwate_arr)
# Calculate mean tmle ate
mean_tmle <- mean(tmleate_arr)
# Calculate mean biases
mean_bias1 <- mean(bias1_arr)
mean_bias2 <- mean(bias2_arr)
```

```
#output
cat("Mean ate for iptw:", mean_iptw, "\n")
cat("Mean ate for tmle:", mean_tmle, "\n")
cat("Mean bias for iptw:", mean_bias1, "\n")
cat("Mean bias for tmle:", mean_bias2, "\n")
```

There are R packages so that we don't have to hand code TMLE ourselves. R packages to implement the TMLE algorithm include tmle, tmle3, ltmle, drtmle, and lmtp.

Similarly, for IPTW we have packages like WeightIt. WeightIt is a one-stop package to generate balancing weights for point and longitudinal treatments in observational studies.

3.3 Results And Summary

For single computation:

True_ATE	0.238578
Iptw_ate	0.2338803
tmle_ate	0.2366373

After Monte Carlo Simulation (To calculate BIAS):

Mean ATE for iptw: 0.2390988

Mean ATE for tmle: 0.2398114

Mean bias for IPTW	0.04706625
Mean bias for TMLE	0.03726208

- The true ATE implies that the risk of death among cancer patients treated with monotherapy is approximately 23.8% higher than for those treated with dual therapy considering only the age as confounder.

For a sample of 10000,

- The IPTW - ATE is found to be 23.4% approximately
- The TMLE - ATE is found to be 23.7% approximately

After 1000 Monte Carlo Simulation over a sample of 10000,

- The Mean Bias for IPTW method (compared to True ATE) is approximately 4.7%
- The Mean Bias for TMLE method (compared to True ATE) is approximately 3.7%

Under the single computation, we see that the TMLE_ATE is nearer to that of TRUE_ATE when compared with the IPTW_ATE. Hence, we can say that TMLE leads to a more reliable value in this case.

TMLE reduces the bias as compared to IPTW in our study. TMLE has been proven to provide better estimates using superlearners and our attempt to implement the same proves it right. In a way IPTW is a subset of TMLE as we see the methodology. The concept of IPTW is a part of TMLE. The model misspecification is taken care of in the case of TMLE and thus this method has several advantages over other methods.

In summary, we have provided an overview with R code considering a very simple data with one confounder only, for implementing IPTW and TMLE to estimate the ATE for a binary outcome in observational studies. TMLE's appealing statistical properties convinces us to consider it a suitable method for estimation of causal effects in large population-based observational studies.

REFERENCES

1. Miguel Angel Luque-Fernandez, Michael Schomaker, Bernard Rachet and Mireille E. Schnitzer (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial, Statistics in Medicine, John Wiley & Sons Ltd.
2. Nicholas C. Chesnaye , Vianda S. Stel , Giovanni Tripepi , Friedo W. Dekker , Edouard L. Fu, Carmine Zoccali and Kitty J. Jager (2022). An introduction to inverse probability of treatment weighting in observational research, Clinical Kidney, vol. 15, no. 1, 14-20.
3. An Illustrated Guide to TMLE by Katherine Hoffman, Dec 2020, https://www.khstats.com/blog/tmle/tutorial-pt2
4. Megan S. Schuler and Sherri Rose (2017). Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies, Am J Epidemiology, 185(1):65-73.
5. The TMLE framework by Jeremy Coyle, 2021 Targeted Machine Learning with Big Data in the tlverse, https://tlverse.org/tmlcimx2021-workshop/tmle3.html
6. Skylar Kerzner (2022). A Complete Guide to Causal Inference,Towards Data Science.
7. Aleix Ruiz de Villa (2021). Propensity Scores and Inverse Probability Weighting in Causal Inference, Towards Data Science.
8. Judea Pearl, Madelyn Glymour Philosophy, Nicholas P. Jewell (2016). Causal Inference in Statistics- A Primer, John Wiley & Sons.

SARIMA Modelling of WPI and CPI Indices

Project Work

Submitted to

PONDICHERRY UNIVERSITY

in partial fulfilment of the requirements

for the award of the degree of

Master of Science

in

Statistics

BY

MANISHITA SARKAR

Reg No: 20375029

DEPARTMENT OF STATISTICS

PONDICHERRY UNIVERSITY

PONDICHERRY

JUNE 2022

PONDICHERRY UNIVERSITY

R.V. NAGAR, KALAPET,

PUDUCHERRY-605014

Certified that the project work entitled **SARIMA Modelling of WPI and CPI Indices** is a bonafide record of work carried out by the following student.

MANISHITA SARKAR 20375029

of **M.Sc (Statistics)**., Statistics submitted in partial fulfilment of the requirement for the award of degree of Master of Science in Statistics, during the academic year **2020-22**.

(Supervisor) Head of the Department

Submitted for M.Sc, Degree Examination held on _____

Examiners

1.

2.

INDIAN STATISTICAL INSTITUTE

Telephone: +91-(0) 33-25752627
FAX: +91-(0) 33-25778893
Email: samarjit@isical.ac.in

Samarjit Das
Professor
ERU, Indian Statistical Institute
203, B. T. Road
Kolkata-700108
INDIA

TO WHOM IT MAY CONCERN

Subject: Internship Certificate

This is to certify that Ms. Manishita Sarkar, has successfully completed internship for 6 months under my guidance. She was working on "SARIMA Modelling of WPI and CPI Indices". The duration of this project was from 14 January 2022 to 10 June 2022.

I hereby certify that her work is excellent to the best of my knowledge.

Thank you,

S. Das

Prof Samarjit Das
Professor
Economic Research Unit
ISI, Kolkata

CERTIFICATE

This is to certify that **MANISHITA SARKAR** (**20375029**), a student of M.Sc. Statistics from the Department of Statistics, Pondicherry University has completed his/her fourth semester project under our guidance at **Indian Statistical Institute (ISI), Kolkata from January to June 2022**. The project work entitled **"SARIMA Modelling of WPI and CPI Indices"** embodies the novel work done by him/her.

_Signature_____

Dr R Vishnu Vardhan
Assistant Professor
Department of Statistics
Pondicherry University
Puducherry – 605014
(Faculty Guide)

Prof. Samarjit Das
Associate Professor,
Economic Research Unit,
ISI, Kolkata
Kolkata - 700108
(Guide)

ACKNOWLEDGEMENT

I am thankful to my guide, **Dr. R. Vishnu Vardhan,** Associate Professor, Department of Statistics, Pondicherry University, Puducherry, for his/her guidance and overall support throughout the project without which it would have been impossible to complete this work.

I express my sincere thanks to **Prof. Samarjit Das**, **Associate Professor**, ISI, Kolkata for providing invaluable guidance throughout the project. His/her sincerity, dynamism, and motivation has inspired me. It was a great privilege to work under his/her guidance.

I take this opportunity to acknowledge my deep sense of gratitude to **Dr. R. Vishnu Vardhan**, Coordinator, Internships and Placements, Department of Statistics, Pondicherry University for providing this opportunity to work with the reputed company **Indian Statistical Institute (ISI), Kolkata.**

I express my sincere gratitude to **Dr. Kiruthika,** Head of Department of Statistics, Pondicherry University, Puducherry for her valuable support and **Prof. Debasis Sengupta,** Dean of Studies, ISI, Kolkata for this wonderful opportunity to work with them.

I am thankful to **Prof. P.Tirupathi Rao**, Dean, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry for his generous support.

I am very thankful to all the faculty members of the Department **Dr. Navin Chandra, Dr. Sudesh Pundir, Dr. J.Prabhakar Naik and Dr. V.S. Vaidyanathan** for their generous and moral support.

I wish to place on record the valuable support of my family members without whose support it would have been impossible to complete this project.

I am thankful to my friends for their cooperation and support in bringing out this project as a thriving one.

Manishita Sarkar

Abstract

India is a rising economic power at a global level. So, inflation of any price change at an institutional level is an important factor in viewing the overall economic view of a country. Wholesale Price Index (WPI) helps in calculating these price changes of goods in the stages before the retail level. The Wholesale Price Index (WPI) model is dependent on time; hence we perform time-series analysis. In order to understand the model, we need to analyse the price changes over a period of time. For further studies, we need to find an appropriate model to account for these changes.This model can be found by using the auto covariance function and the partial auto covariance function of the model. The model will have a lot of noise and so to get more accuracy a box cox transformation is used to make the model more accurate. This model then needs to be used to forecast the WPI which will help us in calculating the growth rate and hence predict the inflation at an institutional level, for a future time. However, discrepancies due to customer inflation will not be included in this model. So we can take the Consumer Price Index also into consideration and forecast the data to create two model for better understanding of the Inflation in India.

Contents

1. Introduction

A time series is nothing but a sequence of various data points that occurred in a successive order for a given period of time. Time series analysis is one of the important feature for prediction and forecasting analysis which is specific to time based datasets. Its used in

- Analyzing the historical dataset and its patterns.
- Understanding and matching the current situation with patterns derived from the previous stage.
- Understanding the factor or factors influencing certain variable(s) in different periods.

With help of "Time Series" we can prepare numerous time-based analyses and results.

- Forecasting
- Segmentation
- Classification
- Descriptive Analysis
- Intervention Analysis

1.1 Components of Time Series Analysis

The various components of time series analysis are:

- Trend
- Seasonality
- Cyclical
- Irregularity

To further elaborate on this:

- **Trend:** In which there is no fixed interval and any divergence within the given dataset is a continuous timeline. The trend would be Negative or Positive or Null Trend
- **Seasonality:** In which regular or fixed interval shifts within the dataset in a continuous timeline. Would be bell curve or saw tooth
- **Cyclical:**In which there is no fixed interval, uncertainty in movement and its pattern
- **Irregularity:**Unexpected situations/events/scenarios and spikes in a short time span.

Let's discuss the time series' data types and their influence. While discussing TS data-types, there are two major types.

- Stationary
- Non-Stationary
- **Stationary:** A dataset should follow the below thumb rules, without having Trend, Seasonality, Cyclical, and Irregularity component of time series
  - The MEAN value of them should be completely constant in the data during the analysis
  - The VARIANCE should be constant with respect to the time-frame
  - The COVARIANCE measures the relationship between two variables
- **Non-Stationary:** It is just the opposite of Stationary, as the name suggests.

1.2 Objectives

The objective of my project is to collect the WPI of India from April 1980 to August 2021 and then link it to a single base year. We will understand what is linking factor and how it is used. Then model selection of the dataset is done and also, we check the stationarity of the data. Fitting of the model is done using the appropriately selected model. Lastly we forecast the model and check the accuracy of it keeping in mind to forecast the WPI of India for a future month.

1.3 Wholesale Price Index (WPI)

A wholesale price index (WPI) is an index that measures and tracks the changes in the price of goods in the stages before the retail level. This refers to goods that are sold in bulk and traded between entities or businesses (instead of between consumers). It is to be accurate is the price of the representative basket of wholesale goods. Usually expressed as a ratio or percentage, the WPI shows the included good's average price change.It is often seen as one of the indicator of a country's level of inflation.

It also influences stock and fixed price markets. The WPI is published by the Economic Advisor in the Ministry of Commerce and Industry. The Wholesale Price Index focuses on the price of goods traded between corporations, rather than the goods bought by consumers, which is measured by the Consumer Price Index. The purpose of the WPI is to monitor price movements that reflect supply and demand in industry, manufacturing and construction. This helps in analyzing both macroeconomic and microeconomic conditions.
Wholesale price indices (WPIs) are reported monthly in order to show the average price changes of goods. The total costs of the goods being considered in one year are then compared with the total costs of goods in the base year. The total prices for the base year are equal to 100 on the scale. Prices from another year are compared to that total and expressed as a percentage of change.
A WPI typically takes into account commodity prices, but the products included vary from country to country. They are also subject to change, as needed, to better reflect the current economy. Some small countries only compare the prices of 100 to 200 products, while larger countries tend to include thousands of products in their WPIs. Price data used to construct the indexes are usually gathered from business firms by mail, less frequently from trade journals and trade associations, and also from government purchasing agents. Weights are generally based on relative sales volume. Data from censuses of production (manufacturing, mining, agriculture, etc.) are used for weights when they are available.

1.4 Major Components of WPI

1. Primary articles is a major component of WPI, further subdivided into Food Articles and Non-Food Articles.

2. Food Articles include items such as Cereals, Paddy, Wheat, Pulses,

Vegetables, Fruits, Milk, Eggs, Meat and Fish, etc.

3. Non-Food Articles include Oil Seeds, Minerals and Crude Petroleum
4. The next major basket in WPI is Fuel and Power, which tracks price movements in Petrol, Diesel and LPG
5. The biggest basket is Manufactured Goods. It spans across a variety of manufactured products such as Textiles, Apparels, Paper, Chemicals, Plastic, Cement, Metals, and more.
6. Manufactured Goods basket also includes manufactured food products such as Sugar, Tobacco Products, Vegetable and Animal Oils, and Fats.

1.5 Main Uses of WPI

1. to provide estimates of inflation at the wholesale transaction level for the economy as a whole. This helps in timely intervention by the Government to check inflation in particular, in essential commodities, before the price increase spill over to retail prices.
2. WPI is used as deflator for many sectors of the economy including for estimating GDP by Central Statistical Organisation (CSO).
3. WPI is also used for indexation by users in business contracts.
4. Global investors also track WPI as one of the key macro indicators for their investment decisions.

1.6 Consumer Price Index (CPI)

Consumer Price Index or CPI is an index measuring the retail inflation in the economy by collecting the changes of the most common goods and services used by consumers. A Consumer Price Index (CPI) is designed to measure the changes over time in general level of retail prices of selected goods and services that households purchase for the purpose of consumption. Such changes affect the real purchasing power of consumers' income and their welfare. The CPI measures price changes by comparing, through time, the cost of a fixed basket of commodities. The basket is

based on the expenditures of a target population in a certain reference period. Since the basket contains commodities of unchanging or equivalent quantity and quality, the index reflects only pure price. Traditionally, CPI numbers were originally introduced to provide a measure of changes in the living costs of workers, so that their wages could be compensated to the changing level of prices. However, over the years, CPIs have been widely used as a macroeconomic indicator of inflation, and also as a tool by Government and Central Bank for targeting inflation and monitoring price stability. CPI is also used as deflators in the National Accounts. Therefore, CPI is considered as one of the most important economic indicators.

The Reserve Bank of India and other statistical agencies study CPI so as to understand the price change of various commodities and keep a tab on inflation. CPI is also a helpful pointer in understanding the real value of wages, salaries and pensions, the purchasing power of a country's currency; and regulating prices.

In India, there are four consumer price index numbers, which are calculated, and these are as follows:

- CPI for Industrial Workers (IW)
- CPI for Agricultural Labourers (AL)
- CPI for Rural Labourers (RL) and
- CPI for Urban Non-Manual Employees (UNME).

While the Ministry of Statistics and Program Implementation collects CPI (UNME) data and compiles it, the remaining three are collected by the Labour Bureau in the Ministry of Labour.

1.7 Major Components of CPI

- Food and beverages
- Pan , Tobacco and intoxicants
- Clothing and footwear.
- Housing

- Fuel and Light
- Miscellaneous items like transport and communication, Health , Education, Recreation and amusement, personal care and effects etc.

1.8 Wholesale Price Index (WPI) Vs Consumer Price Index (CPI)

WPI reflects the change in average prices for bulk sale of commodities at the first stage of transaction while CPI reflects the average change in prices at retail level paid by the consumer.

The prices used for compilation of WPI are collected at ex-factory level for manufactured products, at ex-mine level for mineral products and mandi level for agricultural products. In contrast, retail prices applicable to consumers and collected from various markets are used to compile CPI.

The reasons for the divergence between the two indices can also be partly attributed to the difference in the weight of food group in the two baskets. CPI Food group has a weight of 39.1 per cent as compared to the combined weight of 24.4 per cent (Food articles and Manufactured Food products) in WPI basket. The CPI basket consists of services like housing, education, medical care, recreation etc. which are not part of WPI basket. A significant proportion of WPI item basket represents manufacturing inputs and intermediate goods like minerals, basic metals, machinery etc. whose prices are influenced by global factors but these are not directly consumed by the households and are not part of the CPI item basket.

Thus even significant price movements in items included in WPI basket need not necessarily translate into movements in CPI in the short run. The rise or fall in prices at wholesale level spill over to the retail level after a lag. Similarly, the movement in prices of non-tradable items included in the CPI basket widens the gap between WPI and CPI movements. The relative price trends of tradable vis a vis non-tradable is an important explanatory factor for divergence in the two indices in the short term.

1.9 Data Source

The data used over here has also been collected from the https://eaindustry.nic.in/ site. It was found in many fragments over different base year as published by the Office of Economic Advisor of India.

The CPI Data has been found from the site https://fred.stlouisfed.org/series/INDCPIALLMINMEI#0 and the base year has been changed to 1970-71.

1.10 Linking Factor

To maintain continuity in the time series data on Wholesale Price Index, there is a need for a linking factor so that the new series, when released, may be compared with the outgoing one. For this purpose, there are several methods available in the literature for linking a new series with an old one. Some of the most common and widely used methods among these are:

- Arithmetic conversion method
- Ratio Method
- Regression Method

There are three commonly used methods for linking new WPI series with the old one:

1. **Arithmetic conversion method:** The relationship between indices in the new series(x) and the old series(y) is assumed to be linear, i.e., y=cx, where c is the conversion factor. Hence, c is calculated using y(bar) and x(bar). Generally x(bar) is 100
2. **Ratio method:** In this method, month wise ratios of new indices and old indices are calculated first and then their average is taken as linking factor.
3. **Regression method:** In this method, the relation is based on y=a+bx , where a and b are regression coefficients.

As we are discussing about linking the dataset its important to know what is the base year.

Base year refers to the base point in time of a time series. A base year is used for comparison in the measure of a business activity or economic index.

Over here we have used the ratio method by using the last one as 100 and converting it and linking the fragmented datasets and made it into a single dataset for further use.

1.11 Base Year and Growth Rates

Many financial ratios are based on growth because analysts want to know how much a particular number changes from one period to the next. The growth rate equation is (Current Year - Base Year) / Base Year. For example, for a WPI to understand Inflation from it we need to find the growth rates of the dataset.

Now, as we have obtained the final dataset, we need to understand it and see how it is. For this we have to plot the data and see it and understand the different components which will help in model selection of the dataset.

First, we can decompose the dataset into the components of trend, seasonal, cyclical and random.

Then we have to find the parameters for our model building by which we can identify the model we need to use for the dataset.

1.12 Parameters of the Model

1. **Autoregressive Component:** AR stands for autoregressive. Autoregressive parameter is denoted by p. When p =0, it means that there is no auto-correlation in the series. When p=1, it means that the series auto-correlation is till one lag.

2. **Integrated Component:** In ARIMA time series analysis, integrated is denoted by d. Integration is the inverse of differencing. When d=0, it means the series is stationary and we do not need to take the difference of it. When

d=1, it means that the series is not stationary and to make it stationary, we need to take the first difference. When d=2, it means that the series has been differenced twice. Usually, more than two time difference is not reliable.

3. **Moving Average Component:** MA stands for moving the average, which is denoted by q. In ARIMA, moving average q=1 means that it is an error term and there is auto-correlation with one lag.

   To find the parameters, we need to plot the ACF and PACF of the time series.

2. Methodologies

2.1 ACF

Let x_t denote the value of a time series at time t. The ACF of the series gives correlations between x_t and x_{t-h} for h = 1, 2, 3, etc. Theoretically, the autocorrelation between x_t and x_{t-h} equals,

$$\frac{covariance(x_t, x_{t-h})}{std.dev(x_t) * std.dev(x_{t-h})} = \frac{covariance(x_t, x_{t-h})}{var(x_t)}$$

The denominator in the second formula occurs because the standard deviation of a stationary series is the same at all times.

The last property of a weakly stationary series says that the theoretical value of autocorrelation of particular lag is the same across the whole series. An interesting property of a stationary series is that theoretically it has the same structure forwards as it does backward.

Many stationary series have recognizable ACF patterns. Most series that we encounter in practice, however, is not stationary. A continual upward trend, for example, is a violation of the requirement that the mean is the same for all t. Distinct seasonal patterns also violate that requirement.

2.2 PACF

In general, a partial correlation is a conditional correlation. It is the correlation between two variables under the assumption that we know and take into account the values of some other set of variables. For instance, consider a regression context in which y is the response variable and x_1, x_2, and x_3 are predictor variables. The partial correlation between y and x_3 is the correlation between the variables determined taking into account how both y and x_3 are related to x_1 and x_2.

In regression, this partial correlation could be found by correlating the residuals from two different regressions:

1. Regression in which we predict y from x_1 and x_2,
2. Regression in which we predict x_1 from x_1 and x_2. Basically, we correlate the "parts" of y and x_3 that are not predicted by x_1 and x_2.

More formally, we can define the partial correlation just described as

$$\frac{\text{Covariance}\,(y, x_3 \mid x_1, x_2)}{\sqrt{\text{Variance}\,(y \mid x_1, x_2)\,\text{Variance}\,(x_3 \mid x_1, x_2)}}$$

$$y = \beta_0 + \beta_1 x^2 \text{ and } y = \beta_0 + \beta_1 x + \beta_2 x^2$$

In the first model, β_1 can be interpreted as the linear dependency between x2 and y. In the second model, β_1 would be interpreted as the linear dependency between x_2 and y WITH the dependency between x and y already accounted for.

For a time series, the partial autocorrelation between x_t and x_{t-h} is defined as the conditional correlation between x_t and x_{t-h}, conditional on x_{t-h+1}, ... , x_{t-1}, the set of observations that come between the time points t and t-h.

- The 1st order partial autocorrelation will be defined to equal the 1st order autocorrelation.
- The 2nd order (lag) partial autocorrelation is

  $$\frac{\text{Covariance}\,(x_t, x_{t-2} \mid x_{t-1})}{\sqrt{\text{Variance}\,(x_t \mid x_{t-1})\,\text{Variance}\,(x_{t-2} \mid x_{t-1})}}$$

  This is the correlation between values two time periods apart conditional on knowledge of the value in between. (By the way, the two variances in the denominator will equal each other in a stationary series.)
- The 3rd order (lag) partial autocorrelation is

  $$\frac{\text{Covariance}\ \ (x_t, x_{t-3} \mid x_{t-1}, x_{t-2})}{\sqrt{\text{Variance}\,(x_t \mid x_{t-1}, x_{t-2})\,\text{Variance}\,(x_{t-3} \mid x_{t-1}, x_{t-2})}}$$

And so on for any lag.

2.3 Augmented Dickey-Fuller Test

We consider the stochastic process of form

$$y_i = \emptyset y_{i-1} + \varepsilon_i$$

where $|\varphi| \leq 1$ and ε_j is white noise. If $|\varphi| = 1$, we have what is called a unit root. In particular, if $\varphi = 1$, we have a random walk (without drift), which is not stationary. In fact, if $|\varphi| = 1$, the process is not stationary, while if $|\varphi| < 1$, the process is stationary. We won't consider the case where $|\varphi| > 1$ further since in this case the process is called explosive and increases over time.

This process is a first-order autoregressive process, $AR(1)$, which we study in more detail in Autoregressive process. We will also see why such processes without a unit root are stationary and why the term "root" is used.

The Dickey-Fuller test is a way to determine whether the above process has a unit root. The approach used is quite straightforward. First calculate the first difference, i.e.

$$y_i - y_{i-1} = \emptyset y_{i-1} + \varepsilon_i - y_{i-1}$$

i.e. $y_i - y_{i-1} = (\emptyset - 1)y_{i-1} + \varepsilon_i$ If we use the delta operator, defined by $\Delta y_i = y_i - y_{i-1}$ and set $\beta = \varnothing - 1$, then the equation becomes the linear regression equation

$$\Delta y_i = \beta y_{i-1} + \varepsilon_i$$

where $\beta \leq 0$ and so the test for φ is transformed into a test that the slope parameter $\beta = 0$. Thus, we have a one-tailed test (since β can't be positive) where $\mathrm{H}_0 : \beta = 0$ (equivalent to $\emptyset = 1$) $\mathrm{H}_1 : \beta < 0$ (equivalent to $\emptyset < 1$) Under the alternative hypothesis, if b is the ordinary least squares (OLS) estimate of β, and so $\emptyset$-bar $= 1 + b$ is the OLS estimate of $\emptyset$, then for large enough n,

$$\sqrt{n}(\emptyset - \widehat{\emptyset}) \sim N\left(0, s, e_{\cdot}\right)$$

Where,s.e. $= \sqrt{1 - \emptyset^2}$

We can use the usual linear regression approach, except that when the null hypothesis holds the t coefficient doesn't follow a normal distribution and so we can't use the usual t-test. Instead, this coefficient follows a tau distribution, and so our test consists of determining whether the tau statistic τ (which is equivalent to

the usual t statistic) is less than τ_{crit} based on a table of critical tau statistics values shown in Dickey-Fuller Table.

If the calculated τ value is less than the critical value in the table of critical values, then we have a significant result; otherwise, we accept the null hypothesis that there is a unit root and the time series is not stationary. There are the following three versions of the Dickey-Fuller test:

Type o	No constant, no trend	$\Delta y_i = \beta_1 y_{i-1} + \varepsilon_i$
Type 1	Constant, no trend	$\Delta y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i$
Type 2	Constant and trend	$\Delta y_i = \beta_0 + \beta_1 y_{i-1} + \beta_2 i + \varepsilon_i$

Each version of the test uses a different set of critical values, as shown in the Dickey-Fuller Table. It is important to select the correct version of the test for the time series being analyzed. Note that the type 2 test assumes there is a constant term (which may be significantly equal to zero).

Stationarity can be checked by performing an Augmented Dickey-Fuller (ADF) test:

✓ p-value > 0.05 : Fail to reject the null hypothesis (HO), the data has a unit root and is non-stationary.

✓ p-value ≤ 0.05 : Reject the null hypothesis (HO), the data does not have a unit root and is stationary.

2.4 AR Model

In a simple linear regression model, the predicted dependent variable is modelled as a linear function of the independent variable plus a random error term.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A **first-order autoregressive process**, denoted **AR(1)**, takes the form

$$y_i = \phi_0 + \phi_1 y_{i-1} + \varepsilon_i$$

Thinking of the subscripts i as representing time, we see that the value of y at time $i + 1$ is a linear function of y at time i plus a fixed constant and a random error

term. Similar to the ordinary linear regression model, we assume that the error terms are independently distributed based on a normal distribution with zero mean and a constant variance σ^2 and that the error terms are independent of the y values. Thus

$$\varepsilon_i \sim N(0, \sigma)$$
$$\text{cov}\,(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j \quad \text{cov}\,(\varepsilon_i, y_j) = 0 \text{ for all } i, j$$

It turns out that such a process is stationary when $|\varphi_1| < 1$, and so we will make this assumption as well. Note that if $|\varphi_1| = 1$ we have a random walk.

Similarly, a **second-order autoregressive process**, denoted $\mathbf{AR}(2)$, takes the form

$$y_i = \phi_0 + \phi_1 y_{i-1} + \phi_2 y_{i-2} + \varepsilon_i$$

and a **p-order autoregressive process**, denoted $\mathbf{AR}(\boldsymbol{p})$, takes the form

$$y_i = \phi_0 + \phi_1 y_{i-1} + \phi_2 y_{i-2} + \cdots + \phi_p y_{i-p} + \varepsilon_i$$

2.5 MA Model

Time series models known as ARIMA models may include *autoregressive* terms and/or *moving average* terms. Previously, we saw that an autoregressive term in a time series model for the variable x_{t-1} is a lagged value of x_t. For instance, a lag 1 autoregressive term is x_{t-1} (multiplied by a coefficient). This lesson defines moving average terms.

A moving average term in a time series model is a past error (multiplied by a coefficient).

Let $w_t \sim N\,(0, {\sigma_w}^2)$, meaning that the w_t are identically, independently distributed(iid), each with a normal distribution having mean 0 and the same variance.

The $1^{\mathbf{st}}$ order moving average model, denoted by MA(1) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1}$$

The $2^{\mathbf{nd}}$ order moving average model, denoted by MA(2) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$$

The $q^{\mathbf{th}}$ order moving average model, denoted by MA(q) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}$$

2.6 ARMA Model

An **autoregressive moving average (ARMA)** process consists of both autoregressive and moving average terms. If the process has terms from both an AR(p) and MA(q) process, then the process is called ARMA (p, q) and can be expressed as

$$y_i = \phi_0 + \phi_1 y_{i-1} + \phi_2 y_{i-2} + \cdots + \phi_p y_{i-p} + \varepsilon_i + \theta_1 \varepsilon_{i-1} + \cdots + \theta_q \varepsilon_{i-q}$$

$$\text{or, } y_i = \phi_0 + \sum_{j=1}^{p} \phi_j y_{i-j} + \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j}$$

We can define an ARMA (p, q) process with zero mean by removing the constant term (i.e. φ_0) and saying that $\mathrm{y}_1, \ldots, \mathrm{y}_n$ has an ARMA(p, q) process with mean μ if the time series $z_1, \ldots, z_n$ has an ARMA(p, q) process with zero mean where $z_i = y_i, -\mu$.

If we include the constant term, then as in the AR(p) case, for a stationary ARMA(p, q) process

$$\mu = \frac{\phi_0}{1 - \sum_{j=1}^{p} \phi_j}$$

An equivalent expression for an ARMA(p, q) process with zero mean is

$$y_i - \sum_{j=1}^{p} \phi_j y_{i-j} = \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j}$$

which can be expressed using the lag (or backshift) operator as follows

$$\phi(L)y_i = \theta(L)\varepsilon_i$$

or even as

$$y_i = \frac{\theta(L)}{\phi(L)}\varepsilon_i$$

2.7 ARIMA

An **autoregressive integrated moving average (ARIMA)** process (aka a **Box-Jenkins** process) adds differencing to an ARMA process.

2.8 SARIMA

The **seasonal ARIMA** model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the model is

$$\text{ARIMA}(p, d, q) \times (P, D, Q)S$$

with p = non-seasonal AR order, d = non-seasonal differencing, q = nonseasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

Without differencing operations, the model could be written more formally as

$$(1)\ \Phi\left(B^S\right)\varphi(B)\left(x_t - \mu\right) = \Theta\left(B^S\right)\theta(B)w_t$$

But over here I have used the dummy variables $\alpha_1, \alpha_2, \ldots . \alpha_{11}$ to indicate seasonal variation with the original ARMA model. Which is

$$y_i = \sum_{s=1}^{11} \alpha_s I_{s,i} + \sum_{j=1}^{p} \emptyset_j y_{i-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j} + \epsilon_i$$

2.9 Identification Of the Model

Calculate ACF and PACF: As we have seen, AR processes have ACF values that converge to zero as the lag increases. MA processes have PACF values that converge to zero as the lag increases. The order of the process may not be obvious using this approach.

AR(p) processes have PACF values that are small (near zero) for lags $> p$. MA(q) processes have ACF values that are small for lags $> q$.

If the ACF and PACF values don't seem to converge to zero, then differencing may be needed.

If all the ACF values are near zero, then the time series is probably random. We can model such processes as $y_i = \varphi_0 + \varepsilon_i$ (white noise process).

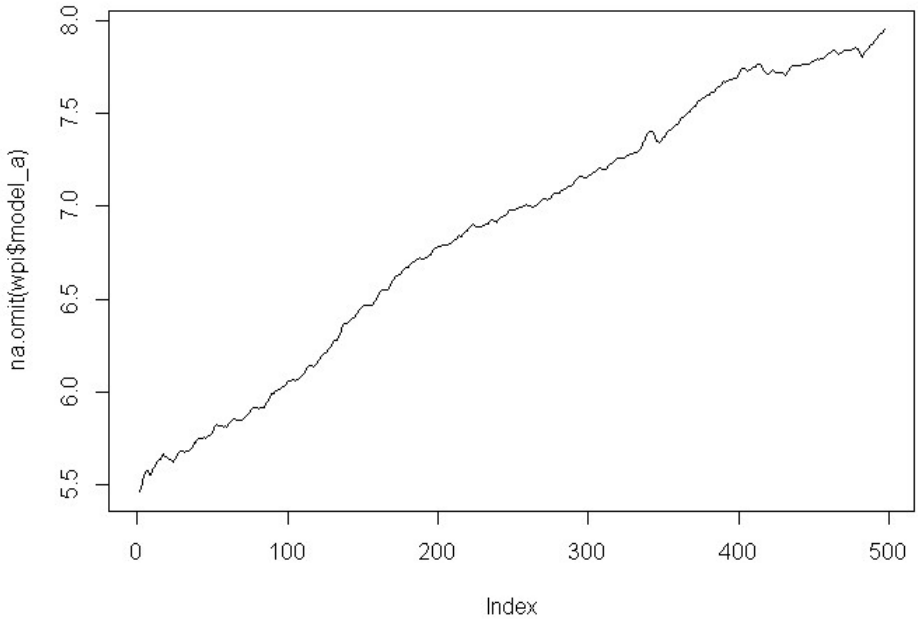
When all the ACF values of first differences are near zero, then the time series is probably a random walk, which can be modelled as $y_i = \varphi_0 + y_{i-1} + \varepsilon_i$.

3. Results And Discussion

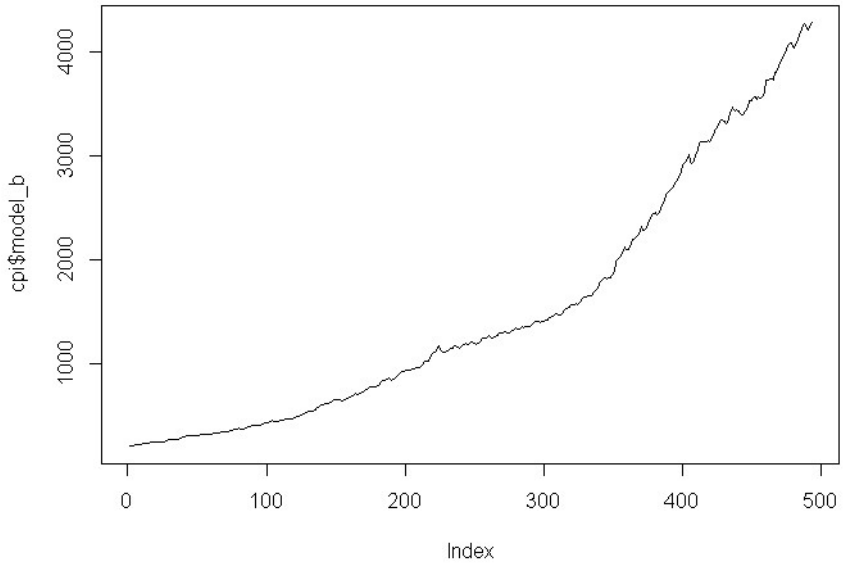
We have understood the meaning of the WPI and CPI. We have found the data to be used. The WPI data used over here has also been collected from the https://eaindustry.nic.in/ site. It was found in many fragments over different base year as published by the Office of economic Advisor of India. We have made this dataset into a continuous dataset by the use of Linking Factor. The CPI Data has been found from https://fred.stlouisfed.org/series/INDCPIALLMINMEI#0 this site and the base year has been changed to 1970-71 from 2015-16.

(i) We have plotted the Datasets by using R and considered it as Model B.

WPI Model B

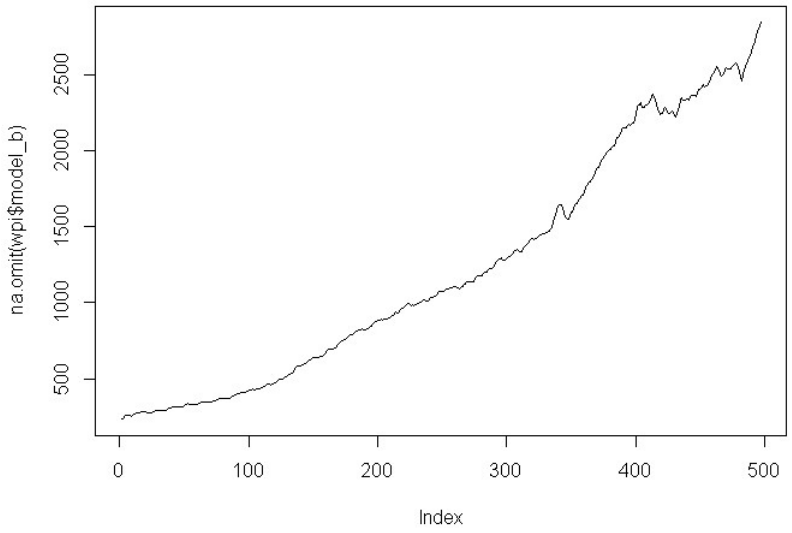


CPI Model B

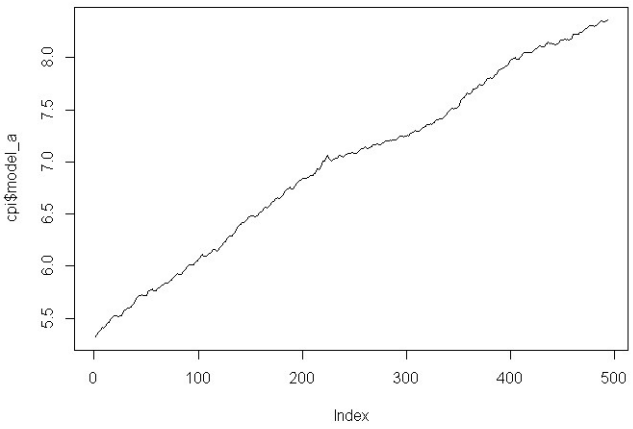


(ii) Now, to make another model A which is log transformation of the original model and note it as Model B. This is done so that the approximation of the model becomes a Normal distribution for better representation of the data and also plot it in R.

WPI Model A

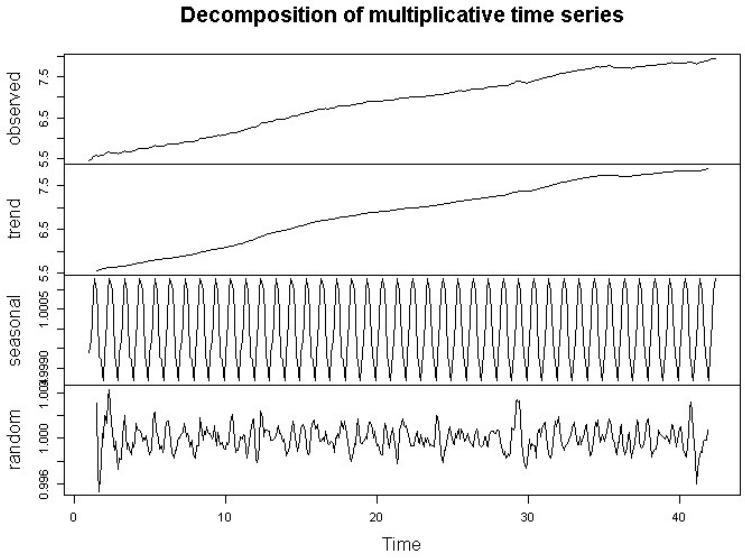


CPI Model A

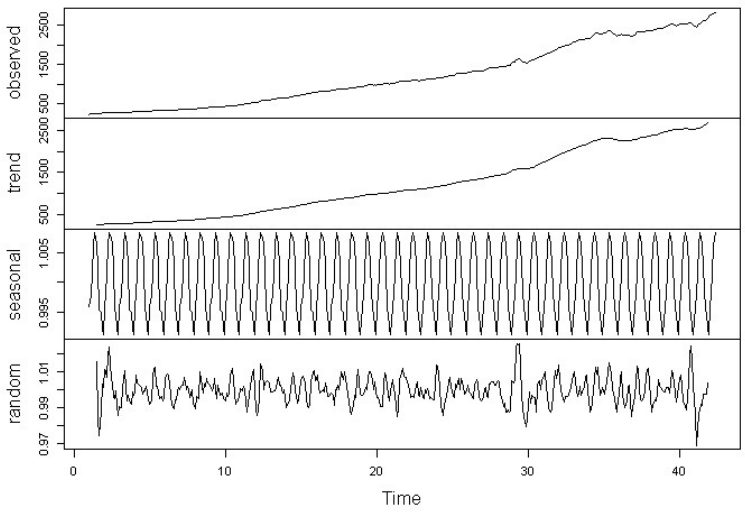


(iii) Decompose the dataset to see the components present in it.

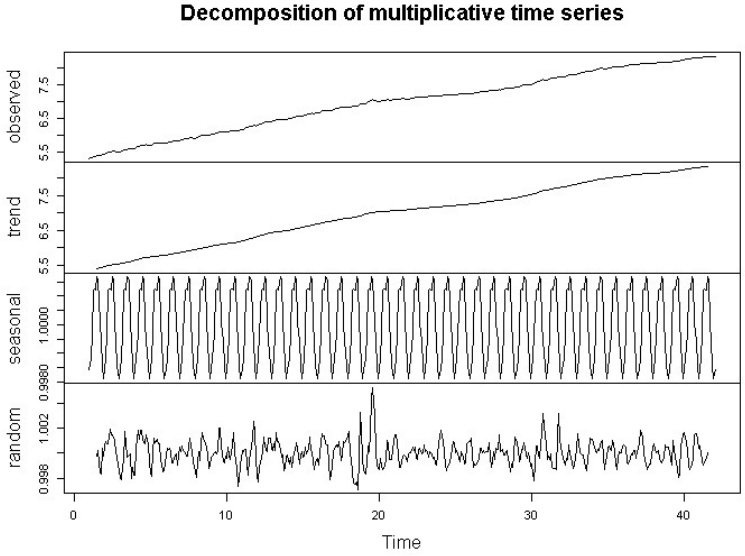
WPI Model A



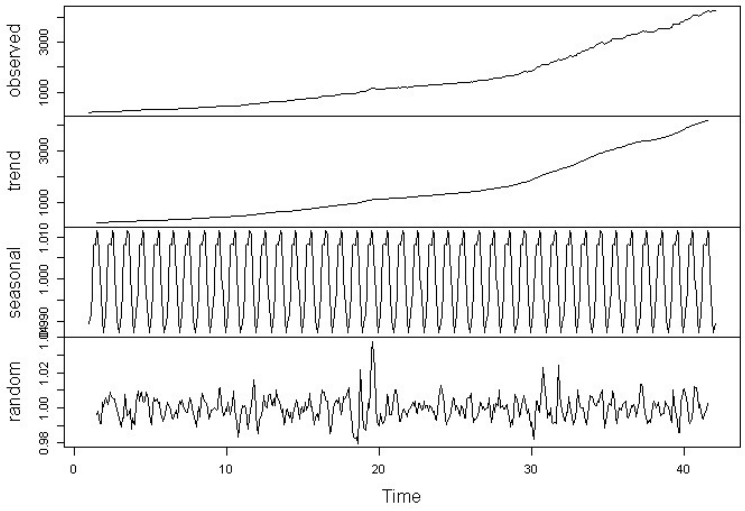
WPI Model B



CPI Model A

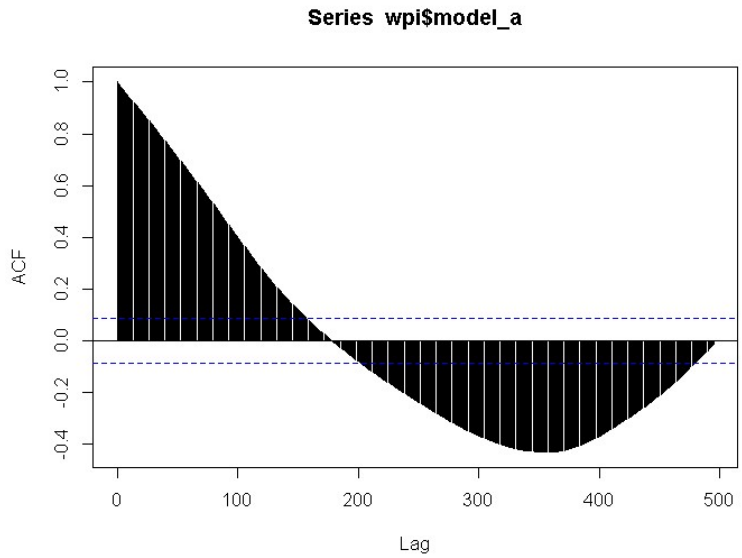


CPI Model B

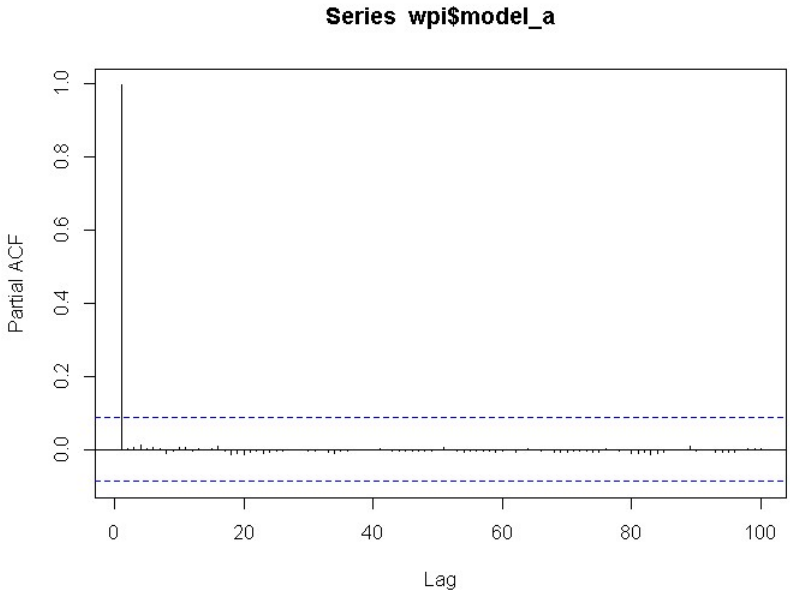


(iv) Find the ACF and PACF of both the models to find the parameters of the models.

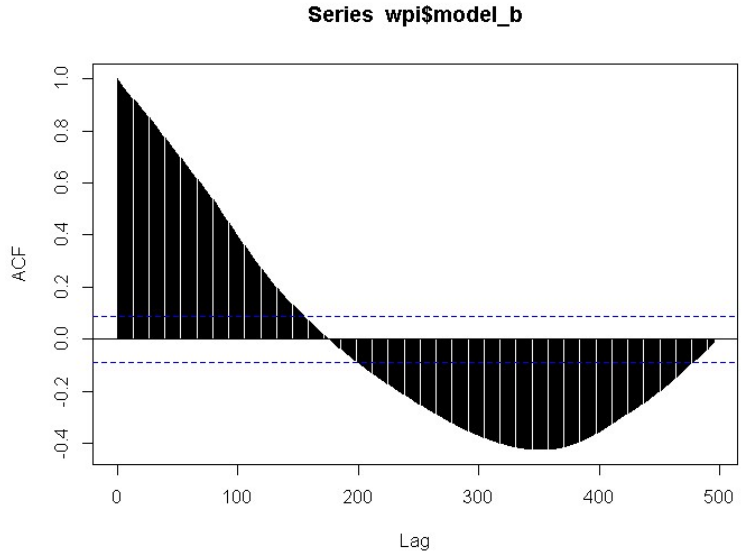
ACF of WPI Model A



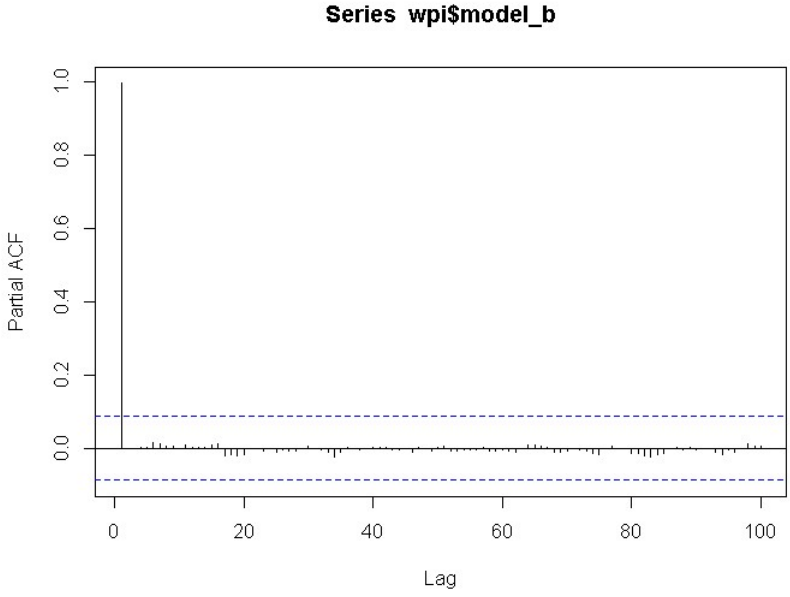
PACF of WPI Model A



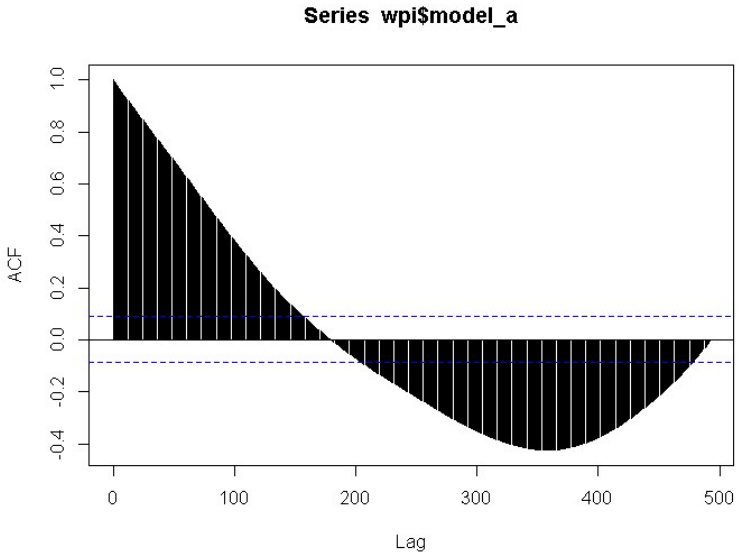
ACF of WPI Model B



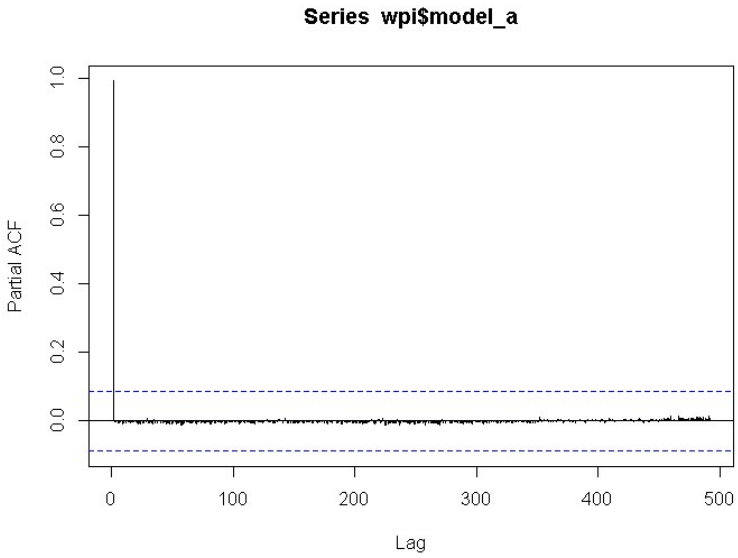
PACF of WPI Model B



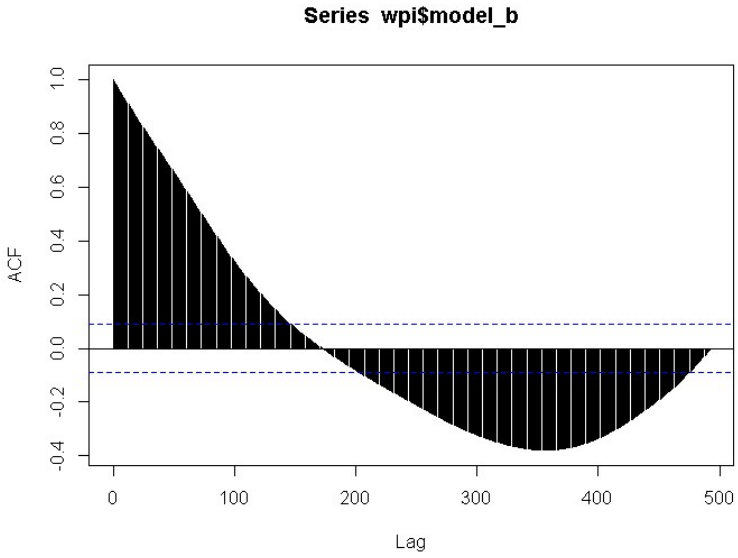
ACF of CPI Model A



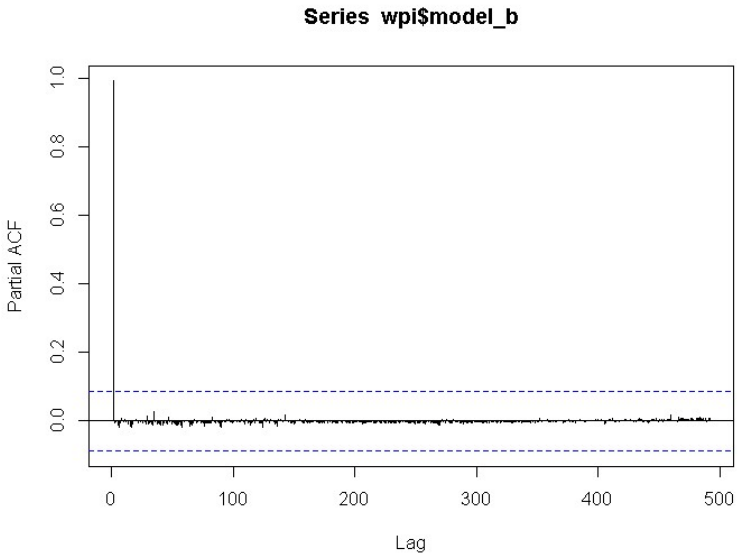
PACF of CPI Model A



ACF of CPI Model B



PACF of CPI Model B



(v) Perform Augmented Dickey-Fuller Test to see the stationarity of the data.We can only use this data for model parameter selection only if stationarity is valid.

Augmented Dickey-Fuller Test for WPI Model A

data: wpi$model_a
Dickey-Fuller = -0.70575, Lag order = 7, p-value = 0.97
alternative hypothesis: stationary

Augmented Dickey-Fuller Test for WPI Model B

data: wpi$model_b
Dickey-Fuller = -1.125, Lag order = 7, p-value = 0.9185
alternative hypothesis: stationary

Augmented Dickey-Fuller Test for CPI Model A

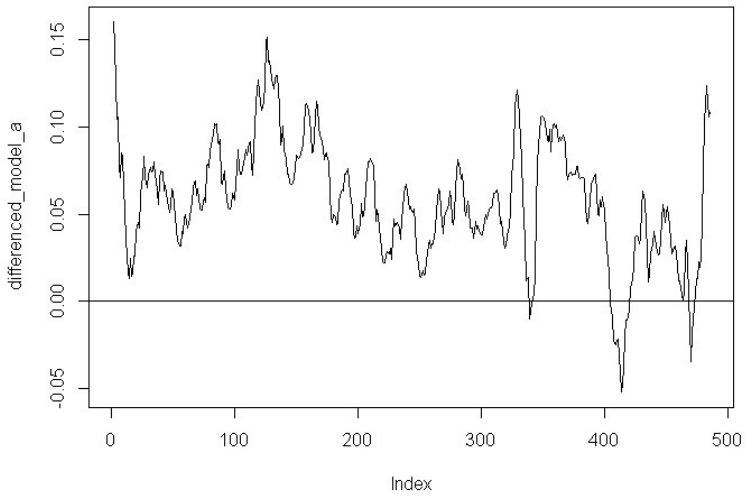
data: cpi$model_a
Dickey-Fuller = -1.2405, Lag order = 7, p-value = 0.8996
alternative hypothesis: stationary

Augmented Dickey-Fuller Test for CPI Model B

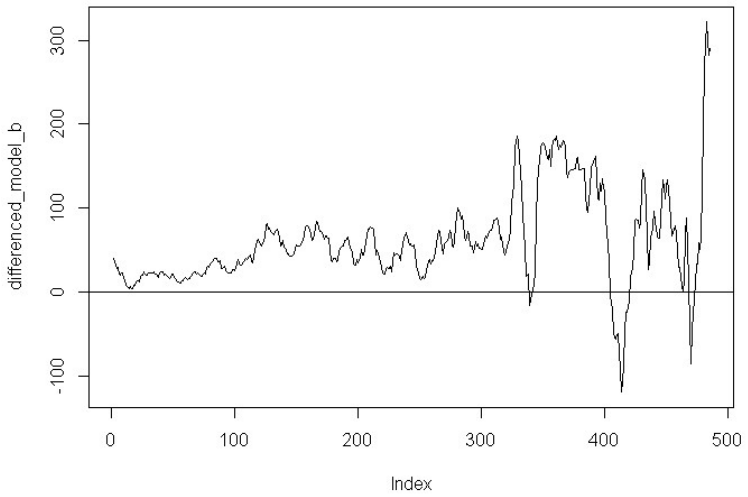
data: cpi$model_b
Dickey-Fuller = -1.8216, Lag order = 7, p-value = 0.99
alternative hypothesis: stationary

(vi) As, the data was not stationary (because the p-value is greater than that of 0.05. So, we can conclude that the datasets is not stationary)we will go for the **First Differencing** to **detrend** the data. Plot This dataset as well.

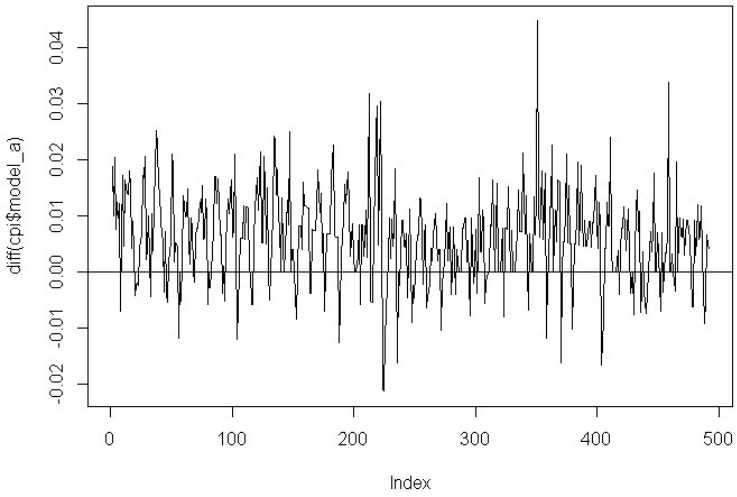
WPI Model A



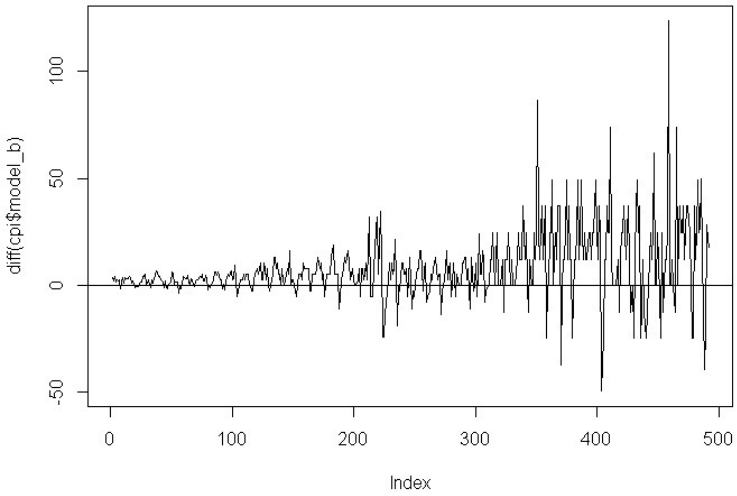
WPI Model B



CPI Model A



CPI Model B



(vii) Again, perform Augmented Dickey-Fuller Test to see the stationarity of the data.

Augmented Dickey-Fuller Test for differenced WPI Model A
data: diff(wpi$model_a)
Dickey-Fuller = -8.9891, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Augmented Dickey-Fuller Test for differenced WPI Model B

data: diff(wpi$model_b)
Dickey-Fuller = -7.3462, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Augmented Dickey-Fuller Test for differenced CPI Model A
data: diff(cpi$model_a)
Dickey-Fuller = -9.9001, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

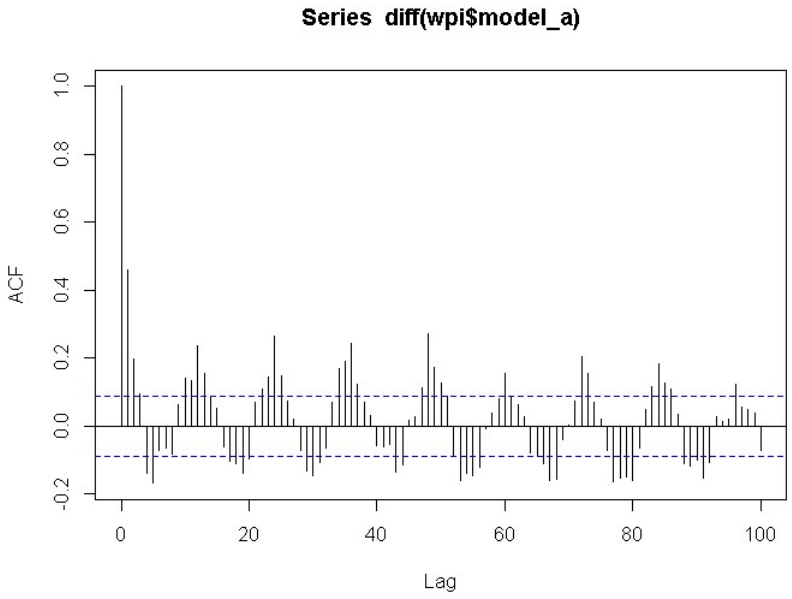
Augmented Dickey-Fuller Test for differenced CPI Model B

data: diff(cpi$model_b)
Dickey-Fuller = -9.5402, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

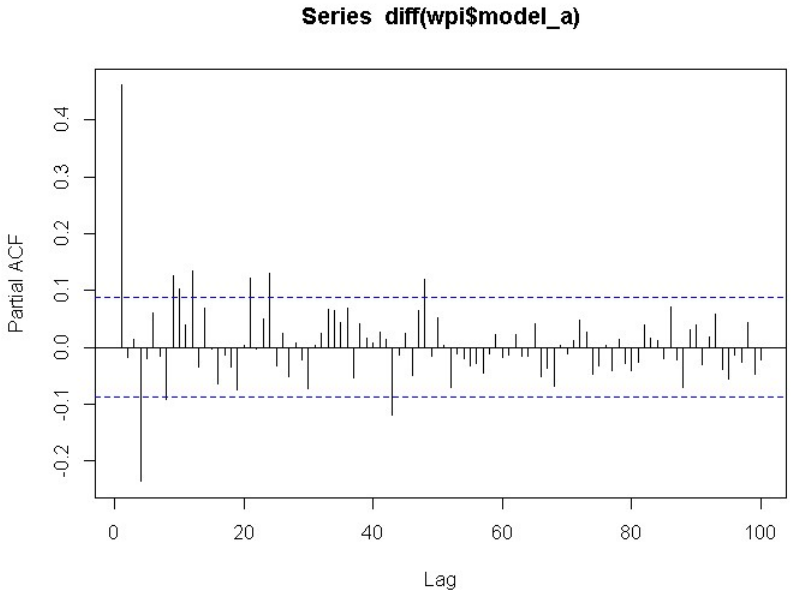
As, the p-value is smaller than 0.05 we reject the Null Hypothesis and can say that the dataset is stationary.

(viii) Then plot the ACF and PACF for the first differenced models.

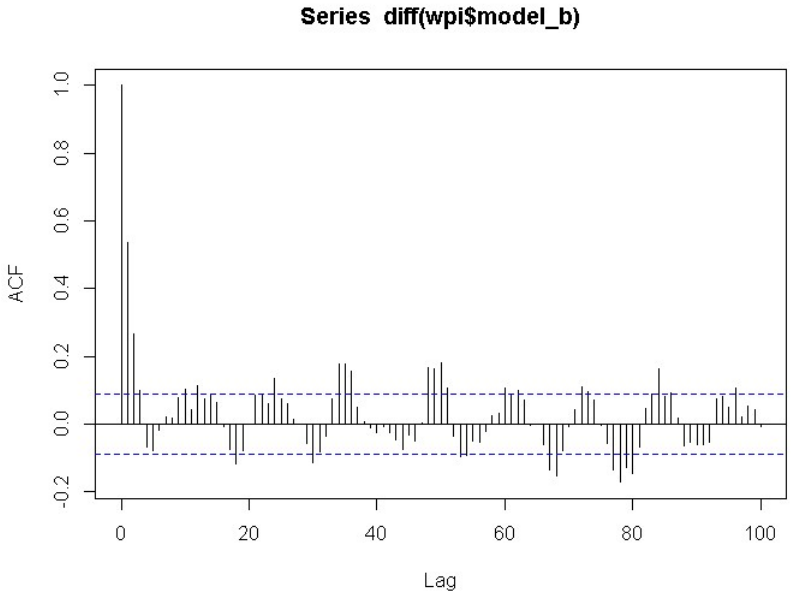
ACF of WPI Model A



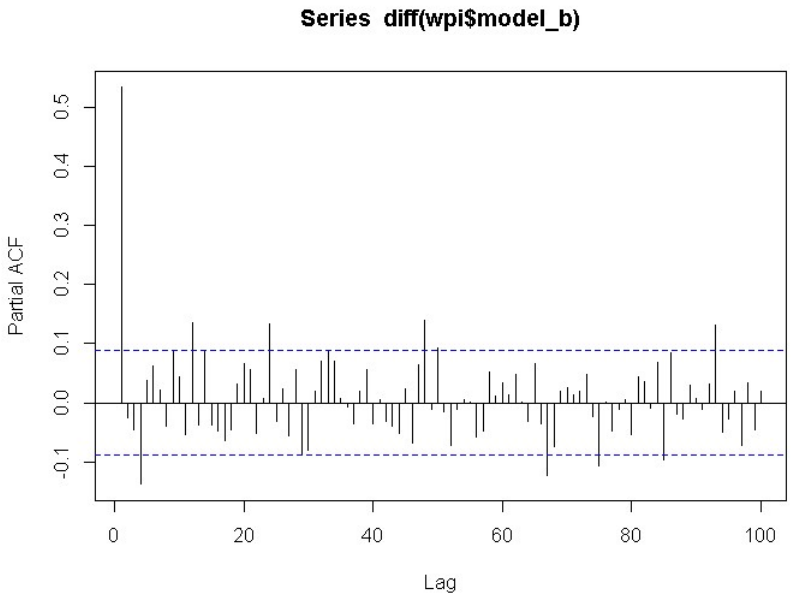
PACF of WPI Model A



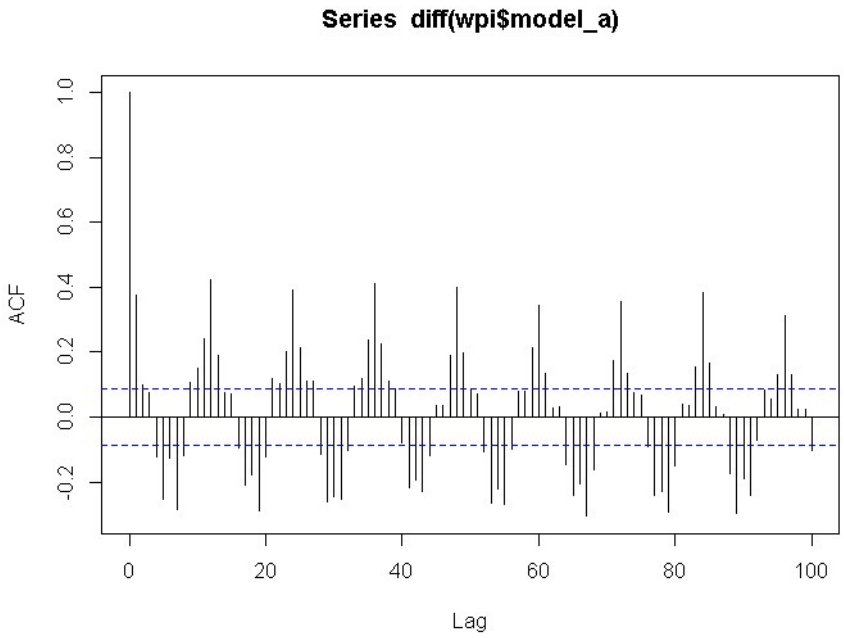
ACF of WPI Model B



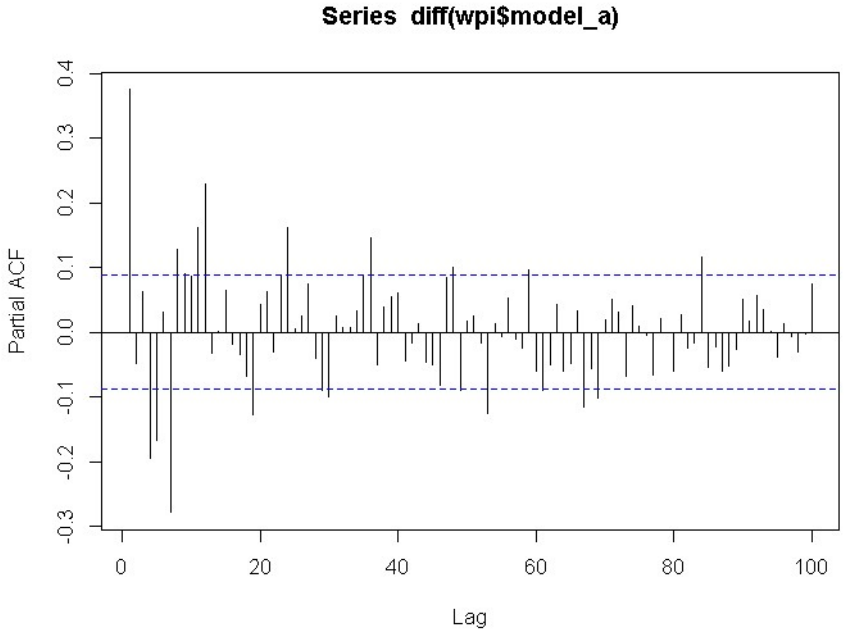
PACF of WPI Model B



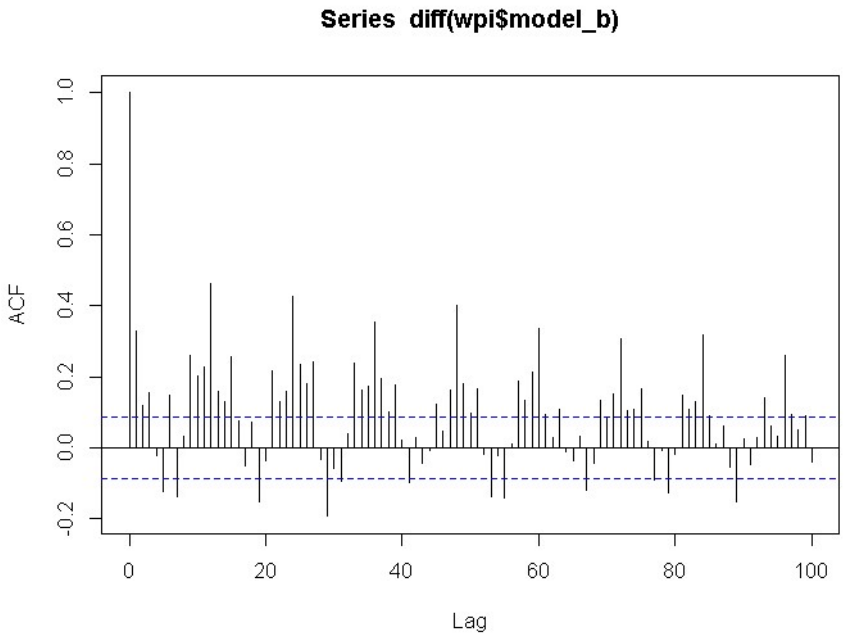
ACF of CPI Model A



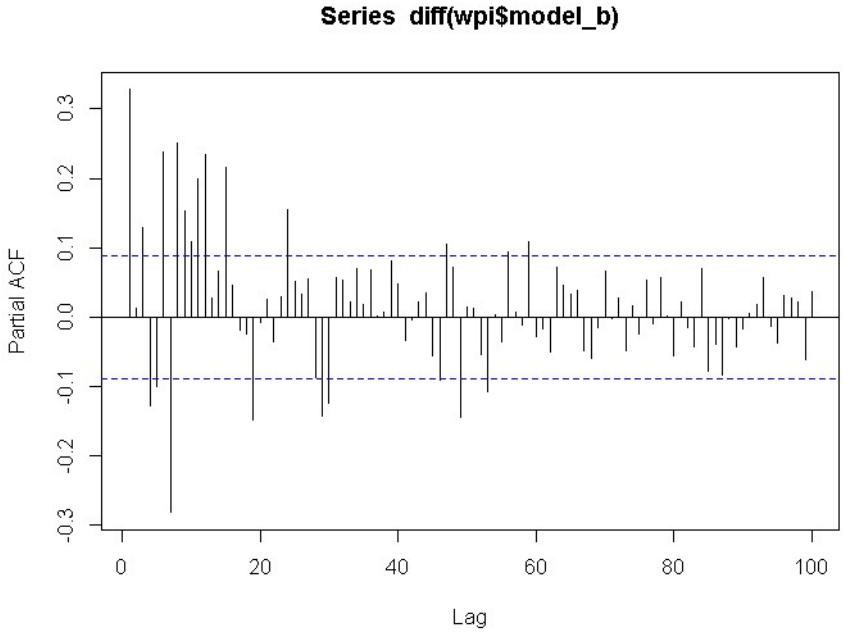
PACF of CPI Model A



ACF of CPI Model B

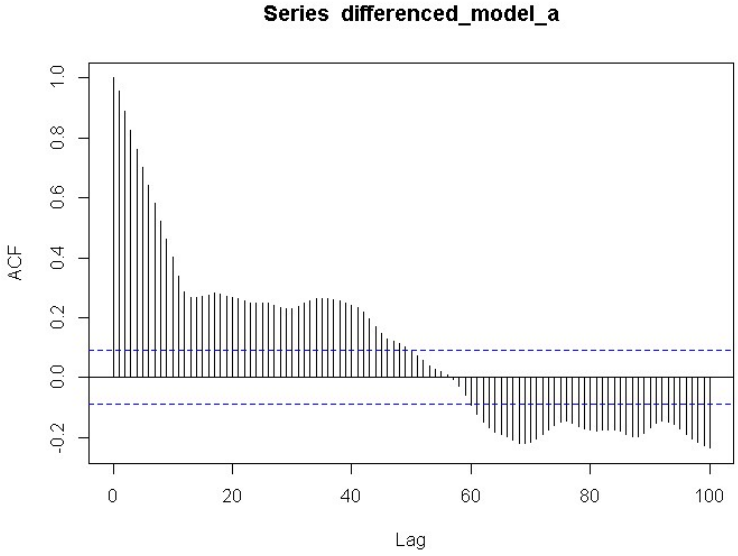


PACF of WPI Model B

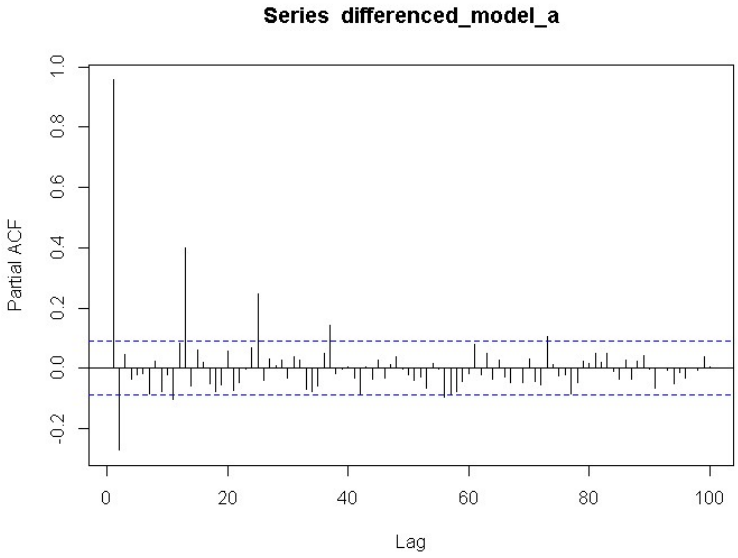


(ix) As we can notice a periodicity over at every twelfth phase, we will try to find ACF and PACF with lag 12 also.

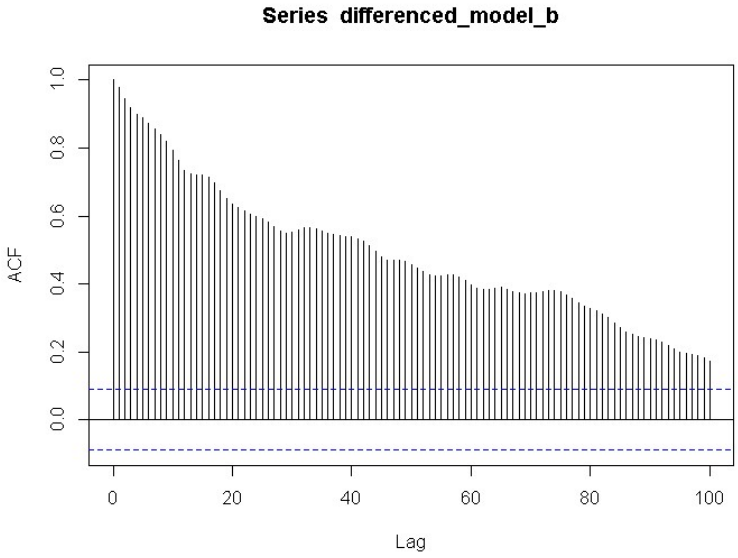
ACF of CPI Model A



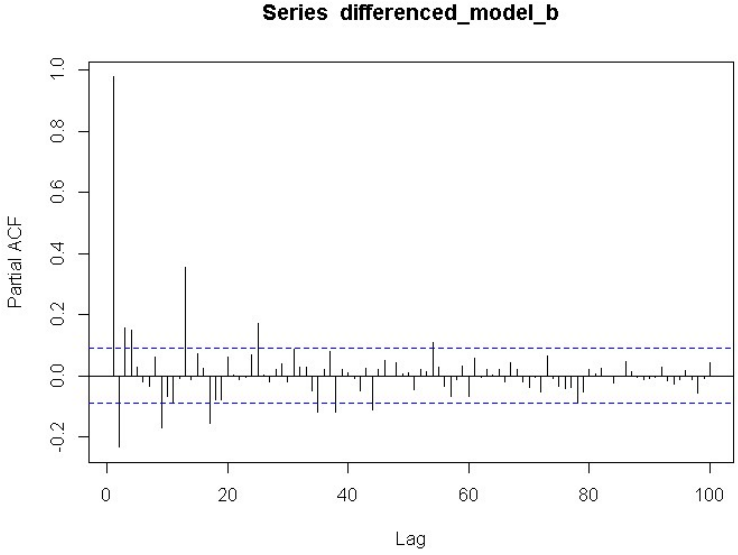
PACF of CPI Model A



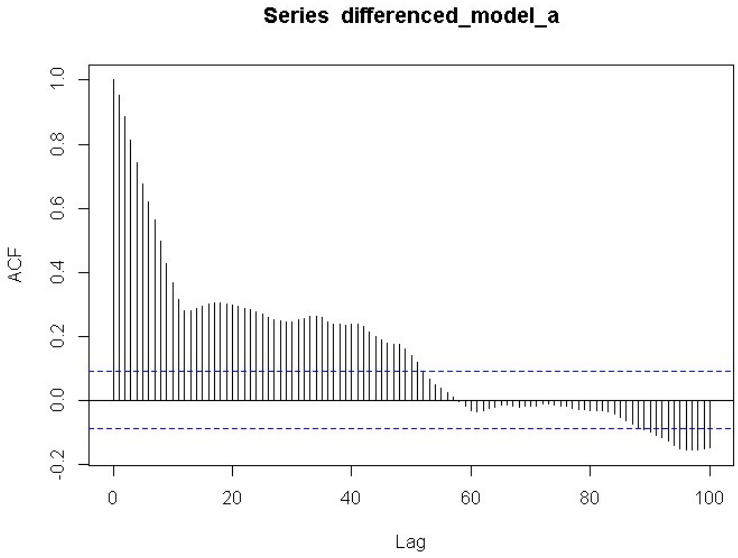
ACF of CPI Model B



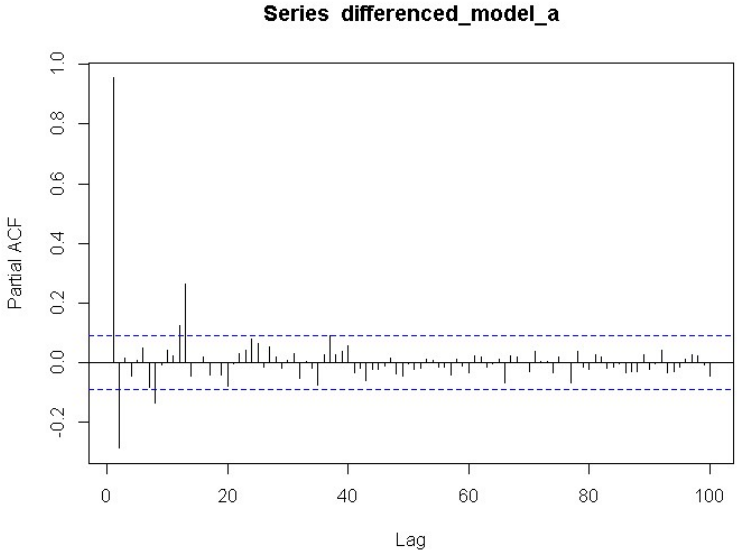
PACF of CPI Model B



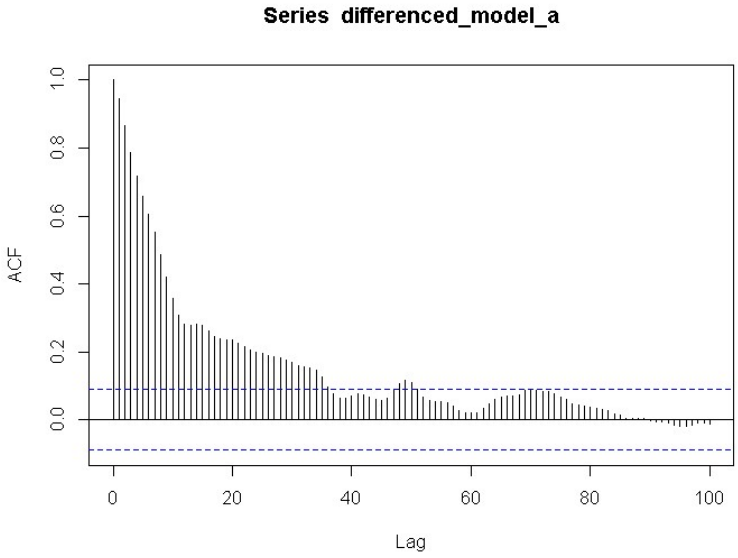
ACF of WPI Model A



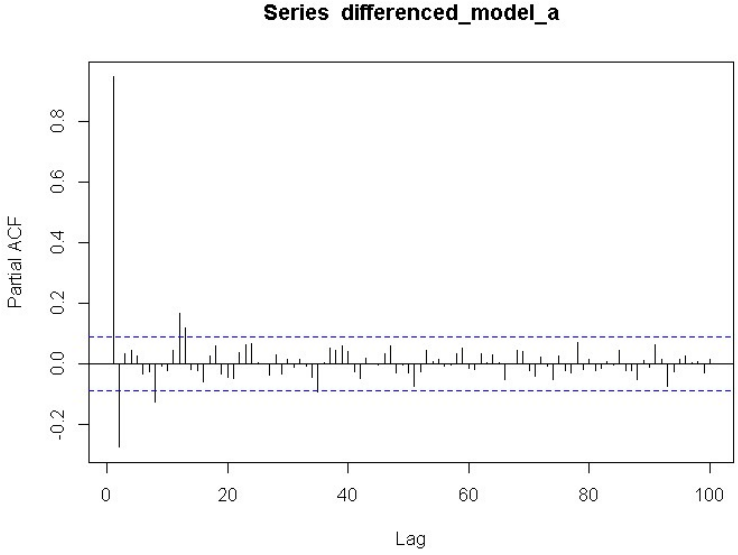
PACF of WPI Model A



ACF of WPI Model B



PACF of WPI Model B



We can see the patterns for the ACF of the seasonal difference models are not clear. We will select the SARIMA model with Indicator Variables i.e.,

$$y_i = \mu + \sum_{s=1}^{11} \alpha_s I_{s,i} + \sum_{j=1}^{p} \emptyset_j y_{i-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j} + \epsilon_i$$

Where a_s and $I_{s,l}$ is indicator variable which is 1 if it is in the s^{th} month where $s = 1$ means April, $s = 2$ means May and so on till $s = 11$ means February and the s=12 is not added as it is a dummy trap as we have an intercept value in the equation.

Now, the fitted model is as follows:

WPI Model A,

Series: differenced_model _ a

Regression with ARIMA(1,0,1) errors

Coefficients:

	ar1	ma1	intercept	a1	a2	a3	a4
	0.3661	0.0874	0.0099	-0.0028	-0.0009	0.0001	-0.0029
s.e.	0.1156	0.1277	0.0010	0.0011	0.0013	0.0014	0.0014
	a5	a6	a7	a8	a9	a10	a11
	-0.0057	-0.0066	-0.0089	-0.0121	-0.0055	-0.0078	-0.0059
s.e.	0.0015	0.0015	0.0015	0.0014	0.0014	0.0013	0.0011

$sigma^2$ = 3.626e-05: log likelihood = 1824.15 AIC=-3618.3 AICc=-3617.3 BIC=-3555.33

WPI Model B,

Series: differenced_model_b

Regression with ARIMA(1,0,1) errors

Coefficients:

	ar1	ma1	intercept	a1	a2	a3	a4
	0.4739	0.0964	10.8898	-4.1633	-1.6571	-1.2837	-4.0045
s.e.	0.0739	0.0827	1.5488	1.4978	1.8973	2.0579	2.1282

	a5	a6	a7	a8	a9	a10	a11
	-5.1151	-7.1320	-8.7492	-14.0753	-7.0717	-9.3385	-5.8039
s.e.	2.1579	2.1659	2.1574	2.1269	2.0552	1.8915	1.4860

sigma^2 = 71.8: log likelihood = -1742.58 AIC=3515.16 AICc=3516.17 BIC=3578.14

sigma^2 estimated as 69.76: log likelihood = -1742.58, aic = 3515.16 AICc=3516.17 BIC=3578.14

CPI Model A,

Series: differenced_model_a

Regression with ARIMA(1,0,1) errors

Coefficients:

	ar1	ma1	intercept	a1	a2	a3	a4
	0.0857	0.2261	0.0085	0.0003	0.0033	0.0085	-0.0021
s.e.	0.1536	0.1510	0.0010	0.0012	0.0014	0.0014	0.0014

	a5	a6	a7	a8	a9	a10	a11
	-0.0026	0.0010	-0.0030	-0.0125	-0.0063	-0.0082	-0.0055
s.e.	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0012

sigma^2 = 3.606e-05: log likelihood = 1825.57 AIC=-3621.15 AICc=-3620.14 BIC=-3558.17

CPI Model B,

Series: differenced_model_b

Regression with ARIMA(1,0,1) errors

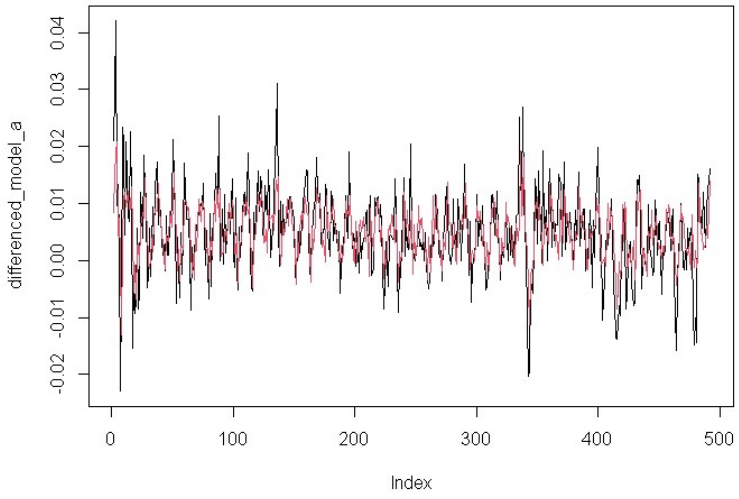
Coefficients:

	ar1	ma1	intercept	a1	a2	a3	a4
	0.4750	-0.1609	12.4910	-1.9447	1.9247	13.0737	-5.7743
s.e.	0.1711	0.1985	2.0936	2.4326	2.7240	2.8509	2.9081
	a5	a6	a7	a8	a9	a10	a11
	-5.5274	1.0892	-5.7822	-18.1250	-7.8897	-13.5235	-8.1628
s.e.	2.9325	2.9391	2.9320	2.9069	2.8484	2.7187	2.4215

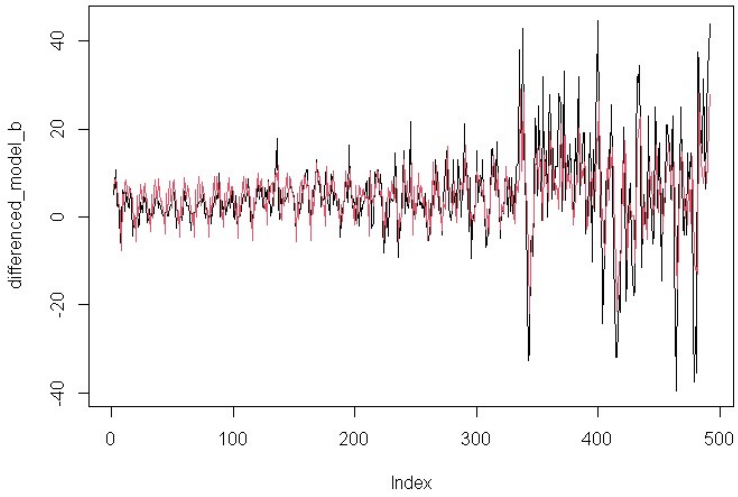
sigma^2 = 164.5: log likelihood = -1946.34 AIC=3922.67 AICc=3923.68 BIC=3985.65

Then we see the fit of it with the original data and we observe that they are almost similar.

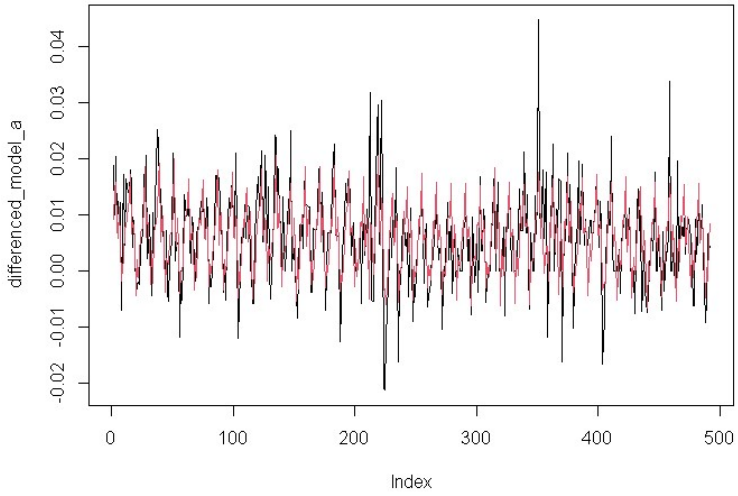
WPI Model A



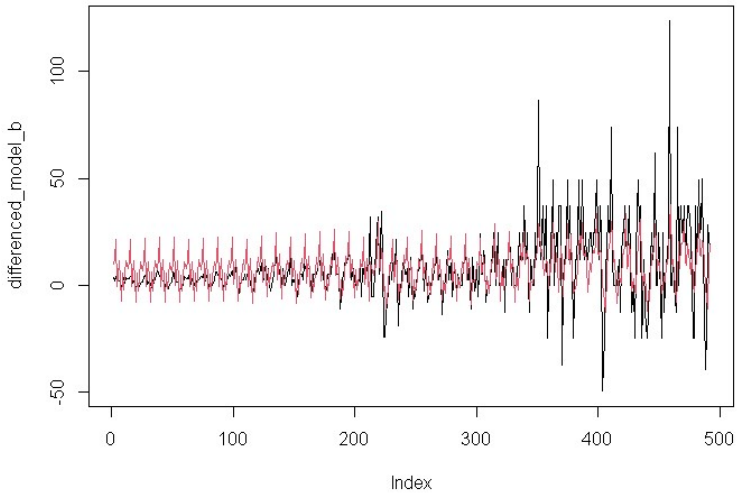
WPI Model B



CPI Model A



CPI Model B



(x) Now we check the forecasting accuracy of the fitted model.

CPI Model B

$pred

Time Series:

Start = 493 End = 503 Frequency = 1

[1] 13.271788 15.710244 26.179586 7.008715 7.102364 13.646096 6.740112 -5.619109 4.608403 -1.029137 4.329852

MAPE ← 1.545625444

predicted values	actual values
4287.575616	4292.099
4303.28586	4331.247
4329.465446	4370.396
4336.474161	4377.513
4343.576525	4388.19
4357.222621	4445.134
4363.962733	4473.605
4358.343624	4462.928
4362.952027	4452.252
4361.92289	4448.693
4366.252742	4484.282

CPI Model A

$pred

Time Series:

Start = 493 End = 503 Frequency = 1

[1] 0.0074161848 0.0116858141 0.0169178643 0.0063754340 0.0058218335 0.0094110263 0.0054308160 -0.0040211919 0.0021173771

[10] 0.0002121588 0.0029799769

MAPE ← 0.98176

predicted values	actual values
3.638282	3.63267
3.649967	3.636613
3.666885	3.640521
3.673261	3.641227
3.679083	3.642285
3.688494	3.647885
3.693924	3.650658
3.696042	3.64962
3.698159	3.64858
3.698371	3.648232
3.701351	3.651693

WPI model A

Time Series:

Start = 493

End = 504

Frequency = 1

[1] 0.0074161848 0.0116858141 0.0169178643 0.0063754340 0.0058218335 0.0094110263 0.0054308160 -0.0040211919 0.0021173771

[10] 0.0002121588 0.0029799769 0.0084600679

MAPE ← -0.664974951

predicted values	actual values
3.448138337	3.443673
3.459824151	3.44628
3.476742015	3.448871
3.483117449	3.453368
3.488939283	3.458135
3.498350309	3.468442
3.503781125	3.477605
3.499759933	3.476394
3.50187731	3.477907
3.502089469	3.482414
3.505069446	3.492751
3.513529514	3.501706

WPI model B

Time Series:

Start = 493

End = 504

Frequency = 1

[1] 23.916093 17.378144 13.465867 8.714283 6.641410 4.168477 2.335259 -3.093199 3.861840 1.572082 5.095805 10.894493

MAPE ← 4.291307478

predicted values	actual values
2782.728383	2777.622
2800.106527	2794.342
2813.572394	2811.063
2822.286677	2840.323
2828.928087	2871.673
2833.096564	2940.643
2835.431823	3003.343
2832.338624	2994.983
2836.200464	3005.433
2837.772546	3036.784
2842.868351	3109.934
2853.762844	3174.724

4. Conclusion

1. MAPE of WPI Model A ← -0.664974951
2. MAPE of WPI Model B ← 4.291307478
3. MAPE of CPI Model A ← -0.98176
4. MAPE of CPI Model B ← 1.545625444

The Mean Absolute Percentage Error is below 5% for all the fitted models forecasted values so we can say it was a good fit and ARIMA(1,1,1) with seasonal Dummies model can be used for forecasting the future values of the CPI and WPI indices. While studying the growth rate we have seen the huge inflation in the economic market of India from the initial years 1970-71. Overall it gives us a good knowledge about the future of Indian market considering that it is turning to be a huge power in the economic field.

References

1. Box,G.E.P., & Jenkins,G.M. 1976. Time series Analysis: Forecasting and Control. San Fransisco: Holden-Day.

2. Modelling the relationship between whole sale price and consumer price indices: cointegration and causality analysis for India AK Tiwari, M Shahbaz - Global Business Review, 2013 - journals.sagepub.com

3. Two-echelon supply chain with selling price dependent demand under wholesale price index and consumer price index D. Nagaraju, A. Ramakrishnarao and S. Narayanan Published Online:November 1, 2012pp 417-439.

4. Exchange rate volatility: A forecasting approach of using the ARCH family along with ARIMA SARIMA and semi-structural-SVAR in Turkey. B Ganbold, I Akram, R Fahrozi Lubis - 2017 - mpra.ub.uni-muenchen.de

5. Singla, C., Sarangi, P. K., Singh, S., & Sahoo, A. K. (2019). Modeling Consumer Price Index: An Empirical Analysis Using Expert Modeler. Journal of Technology Management for Growing Economies, 10(1), 43-50.

6. He, Q., Shen, H., & Tong,Z. 2012. Investigation of Inflation Forecasting. Applied Mathematics & Information Sciences An International Journal6(3): 649-655.

7. Adams, S.O., Awujola,A., & Alumgudu,A.I. 2014. Modeling Nigeria's Consumer Price Index Using Arima Model.International Journal of Development and Economic Sustainability 2(2): 37-47

8. Dongdong, W. 2010. The Consumer Price Index Forecast Based on ARIMA Model.WASE International Conference on Information Engineering.

9. Cryer J.D., & Chan, K.S. 2008. Time Series Analysis with Application in R. New York, Springer.

10. Brocwell,P.J., & Davis,R.A. 2002. Introduction to time series and Forecasting. New York: Springer.

11. The relationship between wholesale price index and consumer price index MF Arby, SP Ghauri - SBP Staff Notes 03/16, 2016 - academia.edu

12. The relationship between wholesale and consumer prices RS Guthrie - Southern Economic Journal, 1981 - JSTOR

13. Singh, S&Sarangi, P.K., 2014. Growth rate of Indian spices exports: Past trend and future prospects. Apeejay - Journal of Management Sciences and Technology 2 (1): 29-34.

14. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679–688

15. W.P. Cleveland, G.C. Tiao, Decomposition of seasonal time series: A model for the census X-11 program, Journal of the American Statistical Association, 71 (1976)

16. Peña, D., Tiao, G. C., & Tsay, R. S. (Eds.). (2001). A course in time series analysis. New York, USA: John Wiley & Sons

17. Theodosiou, M. (2011). Forecasting monthly and quarterly time series using STL decomposition. International Journal of Forecasting, 27(4), 1178–1195.

18. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed). Hoboken, New Jersey: John Wiley & Sons.

19. Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting (3rd ed). New York, USA: Springer.

20. Ord, J. K., Fildes, R., & Kourentzes, N. (2017). Principles of business forecasting (2nd ed.). Wessex Press Publishing Co.