STATISTICAL ANALYSIS AND FORECASTING OF THE PATIENTS AFFECTED BY CANCER

(A CASE STUDY OF IBADAN CANCER REGISTRY, UCH, IBADAN, OYO STATE)

BY

ODESOLA, OLUWASEGUN ISREAL

MATRIC NO: 2011/2828

A PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF BACHELOR OF SCIENCE DEGREE IN STATISTICS, DEPARTMENT OF STATISTICS

COLLEGE OF NATURAL SCIENCES
FEDERAL UNIVERSITY OF AGRICULTURE, ABEOKUTA, OGUN STATE NIGERIA

AUGUST, 2014

DECLARATION

I ODESOLA OLUWASEGUN ISREAL with matriculation number 2011/2828, hereby declare that this project has been written by me and is a record of my own research work submitted to the Department of statistics, College of Natural Sciences of the Federal University of Agriculture Abeokuta, in partial fulfillment for the award of Bachelor of Science (B.Sc) degree, Statistics.

CERTIFICATION

This is to certify that this project work was carried out by ODESOLA OLUWASEGUN ISREAL with Matriculation number 2011/2828, in the Department of Statistics, College of Natural Sciences, Federal University of Agriculture, Abeokuta, Ogun State Nigeria.

Mr O.S Ariyo Project Supervisor	Date
Dr. S.O.N Agwuegbo Head of Department	Date
External Supervisor	Date

DEDICATION

I dedicate this project to God Almighty, the King of kings, Lord of lords, the Alpha and Omega, the Father that goes about daily doing my business; cleaning up my mess, supplying all my needs and keeping me warm in His wide protective arms. I will forever love You Sir.

AKNOWLEDGEMENT

All praises and honour due to Almighty God, the Lord of University for sparing my life up to this moment, even though I encountered different challenges during the journey but He proofed Himself as I AM that I AM and He scaled me through.

My sincere gratitude to Mr. Ariyo, who took much of his time in guiding me through this work. He treated me like a child, brother and family, I am very grateful Sir.

My unlimited thanks goes to my parent, Prophet (Dr) E.A Odesola and Evangelist (Mrs) Odesola, for bring me into this world and giving me the best legacy, may the Lord allow you to eat the fruit of your labour.

My special thanks also goes to Odesola Oluwabukola, Odesola Oluwaseyi, Odesola Oluwatimileyin, Adeyinka Oluwasola and Adeyinka Oluwawumi for your contributions during my stays in school, may God continue to bless you all.

My sincere gratitude also goes my mummy in then Lord Mrs Adeyinka, thank you for your support morally, financially and in prayers, my God will bless you and your family.

Also I would like to express my sincere appreciations to my colleagues, Adeniran Michaeal, Olasupo mayowa, Olojesiku Titilope, Fafore remilekun, Ogunkunle Ridwan and Oluwmuyiwa Olawale, Okunola Jumoke, Taiwo Adeola and Abdulrauf Yusuf, I Love you all.

My daddy, Prophet (DR) E.A Odesola, this portion is specially created for you. You matter so much to my destiny.

You will be forever be a part of my success story

CHAPTER ONE

1.0 INTRODUCTION

1.1 BACKGROUND OF STUDY

Cancer is the second leading cause of death in the world after cardiovascular disease. Half of men and one third of women in the world will develop cancer during their lifetimes. Today, millions of cancer people extend their life due to early identification and treatment. Cancer is not a new disease and has afflicted people throughout the world (Akulapalis; 2008).

The word cancer came from a Greek words karkinos to describe carcinoma tumors by a physician Hippocrates (460–370 B.C), but he was not the first to discover this disease. Some of the earliest evidence of human bone cancer was found in mummies in ancient Egypt and in ancient manuscripts dates about 1600 B.C. The world's oldest recorded case of breast cancer hails from ancient Egypt in 1500 BC and it was recorded that there was no treatment for the cancer, only palliative treatment. According to inscriptions, surface tumors were surgically removed in a similar manner as they are removed today.

Cancer develops when normal cells in a particular part of the body begin to grow out of control. There are different types of cancers; all types of cancer cells continue to grow, divide and re-divide instead of dying and form new abnormal cells. Some types of cancer cells often travel to other parts of the body through blood circulation or lymph vessels (metastasis), where they begin to grow. For example when a breast cancer cell spread to liver through blood circulation, the cancer is still called as breast cancer, not a liver cancer. Generally cancer cells develop from normal cells due to damage of DNA. Most of the time when ever DNA was damaged, the body is able to repair it, unfortunately in cancer cells, damaged DNA is not repaired. People can also inherit damaged DNA from parents, which accounts for inherited cancers. Many times though, a person's DNA becomes damaged by exposure to something in the environment, like smoking.

Cancer generally forms as a solid tumor. Some cancers like leukemia (blood cancer) do not form tumors. Instead, leukemia cells involve the blood and blood forming organs and circulate through other tissues where they grow. Not all tumors are cancerous, some tumors are benign (non-cancerous). Benign tumors do not grow and are not life threatening. Different types of cancer cells can behave differently. The risk of developing many types of cancers can be reduced by changes in lifestyle by quitting smoking and eating low fat diet. If cancer is identified in early stage it is easy to treat and may have better chances for living many years.

DIAGNOSIS AND CANCER TREATMENT METHOD

Surgery and use of modern technology

Ancient surgeons knew that cancer would usually come back after it was removed by surgery. Many people even today consider that many types of cancers are incurable and may delay to consult a doctor in early stage. After anesthesia was invented in 1846, surgeons Bilroth, Handley and Halsted led cancer operations by removing entire tumor together with lymph nodes. Later Paget a surgeon reported that cancer cells were spread from primary tumor to other places through the blood stream (metastasis). Understanding the mechanism(s) of cancer spreading became a key element in recognizing the limitations of cancer surgery.

In the beginning of 1970s, progress in ultrasound (sonography), computed tomography (CT scans), magnetic resonance imaging (MRI scans) and positron emission tomography (PET scans) have replaced most exploratory operations. Using miniature video cameras and endoscopy, surgeons can remove colon, esophagus and bladder tumors through tubes. Recently, less invasive ways of destroying tumors without removing them are being studied including liquid nitrogen spray to freeze and kill cancer cells (cryosurgery). Lasers also can be used to cut the tumor tissue of cervix, larynx, liver, rectum, skin and other organs.

Chemotherapy

During the last decades of the 20th century, surgeons developed new methods for cancer treatment by combining surgery with chemotherapy and/or radiation. Roentgen discovered X-rays after 50 years of anesthesia was discovered. Later doctors identified that nitrogen mustard can kill rapidly proliferating lymphoma cancer cells. Over the years, use of many chemotherapy drugs has resulted in the successful treatment of many types of cancers. Now new approaches are being studied to reduce the side effects of chemotherapy including use of, (a) new combinations of drugs, (b) liposomal and monoclonal antibody therapy to target specifically cancer cells, (c) chemoprotective agents to reduce chemotherapy side effects, (d) hematopoietic stem cell transplantation and (e) agents that overcome multidrug resistance.

Hormonal therapy

In 1878 Thomas Beatson discovered that the breasts of rabbits stopped producing milk after he removed ovaries. Later scientists identified that dramatic regression of metastatic prostate cancer following removal of the testes. Now new classes of drugs (aromatase inhibitors, LHRH analogs) are being used to treat prostate and breast cancers. How hormones influence growth of cancer has guided progress in developing as well as reducing the risk of breast and prostate cancers.

Radiation therapy

In 1896 Roentgen discovered "X-ray" and after 3 years later radiation was used for cancer diagnosis and in treatment. In the early 20th century, researchers discovered that radiation could cause cancer as well as cure it. Now several radiation therapies are being used, these include: (a) c*onformal proton beam therapy* (proton beam will be used for killing tumor cells instead of X-rays); (b) *stereotactic surgery* and *stereotactic therapy (*gamma knife can be

used to deliver and treat common brain tumor); *(c) intra-operative radiation therapy*(cancer has been removed surgically followed by radiation to the adjacent tissues).

Adjuvant therapy

It is the use of chemotherapy after surgery to destroy the few remaining cancer cells in the body. Adjuvant therapy was used in colon and testis cancers.

Immunotherapy

Use of biological agents that mimic some of the natural signals that body uses to control tumor growth is called immunotherapy. These natural biological agents can now be produced in the laboratory including interferons, interleukins, cytokines, endogenous angioinhibitors and antigens. In 1990s scientists produced therapeutic monoclonal antibodies rituximab and trastuzumab that specifically targeted lymphoma and breast cancer cells. At present scientists are developing vaccines to boost the body's immune response against cancer cells.

Age

Cancer can attack anyone, since occurrence of cancer increases as individual ages. Most of the cases are seen in adult, middle age or older. Sixty percent of all cancer is diagnosed in who are older than 65 years of age

1.2 JUSTIFICATION OF STUDY

The world health organization (WHO) defines health as "a state of total physical, mental and social well-being". Also, talking about foreign reserves, the Gross Domestic products(GDP) growth rate, exchange rate and other indices, which symbolize the strength or weakness of each member of a nations, it all depends on productivity, which is conditioned by state of health of the productive force, the workforce. When diseases such cancer threaten a great percentage of the working population, it poses a greater risk to productivity. Hence, a

study or research of the trend of a common disease like cancer, which is a treat to the physical, mental and social well-being of the community, cannot be over emphasized. Thus to improve and sustain our quality of life, cancer cannot be overlooked.

1.3 AIM AND OBJECTIVE OF STUDY

The aim of this research work is the use of time series analysis to:

1. Analyze incidence of cancer in Nigeria for the period of 10years using time series.
2. To forecast for the future behavior of the series for planning and prevention.
3. To determine the trend of cancer in Nigeria and to declare necessary recommendations for the control.

1.4 SIGNIFICANT OF STUDY

The major purpose of any research or study is to understand the path of the existing condition or take appropriate action against degeneration of an improved condition. Thus, this study seeks to create a platform for the improvement of community and nationwide health through the provision of statistical information on cancer.

1.5 DATA COLLECTION

The data for the study is secondary data which is extracted from the Ibadan Cancer Registry IBCR, Department of Pathology, University College Hospital, Ibadan (UCH) for the period of 10 years(2004-2013).

1.6 DATA ANALYSIS AND METHODOLOGY

Data collected will be analyze using Time Series analysis(ARIMA,ACF AND PACF) on the record data of patient diagnosed with cancer in a period of ten years (2004-2013), at University College Hospital(UCH), Ibadan. This research will focus on the trend of cancer and predict the future trend of cancer in Nigeria using (Forecasting method) . The Statistical package that will for the analysis is R. Language.

1.7 DEFINITION OF TERMS

Disease: A **disease** is an abnormal condition that affects the body of an organism. It is often construed as a **medical condition** associated with specific symptoms and signs. It may be caused by factors originally from an external source, such as infectious disease, or it may be caused by internal dysfunctions, such as autoimmune diseases. In humans, "disease" is often used more broadly to refer to any condition that causes pain, dysfunction, distress, social problems, or death to the person afflicted, or similar problems for those in contact with the person. In this broader sense, it sometimes includes injuries, disabilities, disorders, syndromes, infections,isolated symptoms,deviant beh aviors, and atypical variations of structure and function, while in other contexts and for other purposes these may be considered distinguishable categories. Diseases usually affect people not only physically, but also emotionally, as contracting and living with many diseases can alter one's perspective on life, and one's personality.

Time Series: A time series is a set of statistics, usually collected at regular intervals. Time series data occur naturally in many application areas.

• economics - e.g., monthly data for unemployment, hospital admissions, etc.

• finance - e.g., daily exchange rate, a share price, etc.

• environmental - e.g., daily rainfall, air quality readings.

• medicine - e.g., ECG brain wave activity every 2−8 secs. There four objectives of time series analysis, which are:

1. **Descriptive:** This is the exploration of time series data, it involves the use of a time plot rather than descriptive measure of the main properties of the series.
2. **Explanatory:** In order to have a deeper understanding of the mechanism which is generated in given time series, we make use of variation in one time series to explain the variation in another.
3. **Forecasting:** In any business set-up or life situation, forecasting into future is inevitable. Thus for effective purpose, a record of past event is very prominent. This entails the application of time series, and the drawing of trend lines to reduce the magnitude of error. In case where seasonal factors are applicable, the series is also deseasonalised.
4. **Control:** When a time series is generated, for a manufacturing process, the aim of the analysis may be to control the process. This involves numerous tasks ranging from plotting of chart to inspection model.

Time Plot: Time plot is a time series involving variable X_t which is represented picvtorically by constructing a graph of X_t against t. where X_t is a quarterly or yearly representation of an observation and t is the time yearly or quarterly.

CHAPTER TWO

2.0 LTERATURE REVIEW

2.1 LITERATURE REVIEW ON CANCER

Cancer can impose a substantial burden through long-term human suffering for individuals and families, economic impact on active members of society and high costs for health-care systems. A recent report showed that cancer causes the highest economic loss of all the 15 leading causes of death worldwide.(WHO 2011) The total economic impact of premature death and disability from cancer was $895 billion in 2008 not including direct costs of treatment. The top three cancers that had the highest economic impact globally are lung cancer ($188 billion), colon/rectal cancer ($99 billion) and breast cancer ($88 billion). Only 5% of the global resources for cancer are spent in developing countries. In those countries, cancers are usually detected at advanced stages, when many are more difficult to treat, therefore treatment interventions are more costly and less successful. The estimated rise in cancer incidence will have a greater impact on countries that have a low health budget and fragile or absent health systems.(WHO 2008).

Cancer is the second most common cause of death worldwide after cardiovascular diseases. Globally, there were reported 12.7 million new cases of cancer in2008 (6,639,000 in men and 6,038,000 in women) and 7.6 million deaths due to cancer (4,225,000 in men and 3,345,000 in women). Cancer is not a modern disease, but as cancer risk increases steeply with age, it is more common nowadays due to increasing life expectancy. It has been estimated that the incidence of cancer will double between 2000 and 2020 and nearly triple by 2030. Until recently, cancer was considered a disease of westernized, industrialized countries, however, in 2008, 56% of new cases (7.1 millions) and 63% of all cancer deaths (4.8millions) were reported by low- and middle-income countries, this figure is predicted to increase due to a

rise in life expectancy of the populations in low- and middle-income countries and due to their large overall population size.(WHO 2002)

Prevalence of breast cancer could be due to many reasons which vary from increased ability to treat diseases in order to delay their progression to inability to diagnose and treat a disease which leads to disability and death. In essence, the survival rate will determine change in disease prevalence (Crimmins, Hayward and Saito; 1994).

Researchers such as Finlieb (1995), Brown et al (1996), Hann *et al* (1996) Liebson (1997), Shahar *et al* (1997) in Crimmins and Saito (2000) explained that, change in both the prevalence of disease and the processes by which the prevalence changes have come generally can be attributed to increases in lengthening survival after disease diagnosis with varying pattern of change in incidences. In this regard, it was found out that the highest increases in disease prevalence have been heart and cancer related.

There has been considerable research on trends of health which show that changes in the prevalence of disease, for instance, breast cancer is an important indicator of the combined effects of past level of and changes in mortality and disease incidences. In addition, trends in disease presence do not necessarily represent trends in disease of a specified severity. With the constant prevalence of disease overtime, the severity of the disease could change. It is also possible that in more recent years, people are learning of the presence of less severe disease at an earlier stage because of the growing ability to diagnose non-invasively. According to the study carried out by Eileen, Crimmins and Saito (2000), Olopade (2004) Congdon (2004) age is found to be a causal factor for prevalence of diseases. Gender (being female or male) could trigger the prevalence of a disease due to changes in the components of the body. This position was documented by Eileen, Crimmins and Saito (2000),Oncologist (2001) Olopade (2004), Ikpah (2002), Coe (2003), Ferley (2005) and Ogundipe and Obinna (2008) believe that diet and environment which arose as a result of lifestyle or westernization

and lack of awareness, access to health care facilities, no plan for such diseases in the National Health Insurance Scheme (NHIS), lack of empowerment of women, bad economy and other social factors are responsible for prevalence diseases. In addition, inadequate clinical services for life threatening diseases and poor distribution assist in prevalence (Olopade, 2004; Adebamowo 2006; Lambo, 2007; Ogundipe and Obinna, 2008). The issue of limited access and scope of services which does not allow multidisciplinary care, obesity and genetic mutation are also mentioned by Ikpah, (2002), Adebamowo and Ajayi (2000).

2.2 CANCER REGISTRATION IN NIGERIA.

Cancer registration is one the pillars of cancer epidemiology. Cancer Registration in Nigeria was started in 1960 at the Department of Pathology, University College Hospital Ibadan by the late Professor G.M. Edington. The Ibadan cancer registry has since then helped in the establishment of other cancer registries in terms of local training, for other teaching hospitals like Ife, Jos, Calabar and Lagos.

Aim of establishing cancer registries:

>To strengthen cancer statistics collection

>To determine real incidence of cancer

>To determine prevalence of risk factors

>To determine cancer mortality in a population

2.3 RESEARCH FINDINGS ON DEMOGRAPHY

Demographic factors are both determinants and consequences of economic and social development. That is, while demographic variable (whose impact is usually reflected in the size, rate growth, age structure and demographic distribution of a population may influence the tempo social and economic development) the course of demographic variable(fertility, mortality and migration) may also be influence by socio-economic development through health education, economic and research programmers(Feyisetan 1995)

Globally, there were reported 12.7 million new cases of cancer in 2008 (6,639,000 in men and 6,038,000 in women) and 7.6 million deaths due to cancer (4,225,000 in men and 3,345,000 in women) (WHO 2008).

2.4 TIME SERIES RESEARCH AND ANALYSIS

So far the nature, advantages and limitations of different time-series models generally adopted have been critically reviewed. Now comparison with reference to forecast accuracy of different models and their applications in the past are discussed below.

Gerra (1959) presented a series of behaviour relations and identities which were believed to stimulate the basic economic system for the egg industry. He indicated that in using the equations fitted (an econometric model) to forecast values of variables in the egg industry beyond the years for which equations were fitted, better estimates of the annual quantity variable (domestic egg consumption, egg production on farms, average number of layers on farms, and the number sold) were obtained from simultaneous equation approach, while better estimates for some variables like storage movement and price variables were obtained by least square method.

Tobin and Arthur (1964) used a low-pass filter (simple Moving Average) of six months length for broiler chick prices and of twelve months length for hatchery supply flocks. The resulting filtered series revealed cycles of approximately 30 months for both series. A time difference of 12 to 18 57 months existed between the peaks of the two series for the 30 months cycle.

2.5 LITERATURE REVIEW ON PAST PROJECT

In the work, by Ogbebor(2008). It was observed using Least square method there was a decrease in the trend of the number of reported cases of breast, liver, and skin cancer each year. Also there was a decrease in the trend of death resulting from breast and skin cancer while there was an increase in the number of reported death of liver cancer.

Oladipupo (2001), it was also observed that using least square method, the trend was increasing gradually from the first quarter of 1997 to the 4^{th} quarter and also when considering the auto regressive model. It was decreasing from 1^{st} to the 4^{th} quarterly of the same year.

Also Ogunjobi (1997) observed that using both least square method and moving average method the trends were increasing from the first quarter of the year 1988 to the fourth quarter of the same year.

According to various project works that were considered, It can be concluded that both the least square methods and the moving average method were mostly adopted for forecasting of the future values. They were found to be the best method. Thus, the least square method which gives a better forecast would be adopted in this project work since both multiplicative model and additive model under moving averages may not give a better forecast. Time Series Analysis is useful in the following

1. To understand the underlying structure of the given data and to isolate as far as possible several broad influences affecting it.
2. To Forecast the future through the pattern of the data

CHAPTER THREE

3.0 METHODOLOGY

3.1 TIME SERIES ANALYSIS

According to Spiegel (1972), time series can be defined as asset of observation taken at specified time usually at equal interval. Generally, it can be define as a sequence of observation generalized sequentially with time. Thus, time series analysis is the research and statistical analysis of variation data.

3.2 PROCESS OF TIME SERIES

- A time series is said to be binary when observations can take two values usually denoted by 0 & 1. This occurs particularly in computer analysis and communication theory.
- Point Processes
- Airplane disaster, record of immigrants within a specified time.

3.3 FORM OF TIME SERIES

DISCRETE TIME SERIES

A time series (Xt) is said to be discrete when observations are taken at discrete times usually equally spaced.

CONTINUOUS TIME SERIES

A time series (Yt) is said to be continuous when observations are taken constantly through time.

3.4 MODELS OF A TIME SERIES

These are: Additive and Multiplicative Model

Additive Model

This model assumes that values of variable (Yt) equal the sum of the four other components.

This means that components are independent of each other such that;

$Yt = St + Tt + \mathrm{Ct} + It$

Multiplicative Model

This assumes that the four are dependent on each other such that;

$Y_t = S_t$ x T_t x C_t x I_t where;

Y_t= Observation data at time t

T_t= Trend variation

S_t= Seasonal Variation

C_t= Cyclical Variation

I_t= Irregular variation

3.5 COMPONENTS OF TIME SERIES

Time series can be classified into four of characteristics movements often called the components. These components are shown below:

- Trend or Secular Movement(Tt): A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend "changing direction" when it might go from an increasing trend to a decreasing trend. It also refers to the general direction in which the graph of time series to going over a long interval of time based on regular data collected over a period of time. There are various types of growth. The most common is the linear trend line. When the trend is determined, the rate of change can be then be ascertained and tentative estimate concerning the future can be made, after the estimation of trend, other can them be estimated.
- Seasonal Variation(St): A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period. It is an identical pattern, which a time series appears to follow during corresponding intervals of successive periods. These intervals may be months, quarterly. They are said to be periodic movement in

business activities, which occur regularly every year. There are many causes of such seasonal variations, which include climate and weather conditions, customs and tradition holiday periods, government policy. e.t.c

- **Cyclical Variation**: A cyclic pattern exists when data exhibit rises and falls that are *not of fixed period*. The duration of these fluctuations is usually of at least 2 years. It is the long oscillation about the trend line or curve. They are sometimes called cycles and may be periodic. A typical example of cyclical variation is the business cycles representing interval prosperity recession, depression and recovery with the pattern repeating itself. It is also extracted by using an approximate moving average of a few months (Say three, four or seven months), which will smooth out any irregular variation.
- Irregular Movement or Random Variation(It): This is the unexpected chance variation which remain when the other influences have been identified and explained. It is assumed that irregular fluctuations are isolated such as bank failures, wars, strikes, drought and earthquakes.

3.6 ESTIMATING THE TREND

The trend line can fitted by estimating using;

- The free hand method
- The method of semi-average
- The moving average method
- The method of Least Squares

Free Hand Method

This method involves the plotting of the series on a graph and fitting a straight line through the plotted plots by mere inspection. This method is of an advantage because it is

easy to use and provides quick description of the general tendency in time series. Its great advantage is that it depends too much on individual judgment. Therefore, accurate prediction cannot be made with this method.

Semi-Average Method

In this method, time series data can be divided into two equal parts and the average is computed for each part. The average of the first is used as the intercept. The ratio of the difference between the two averages over the range of time is used as the slope or gradient. Hence, the variables are coded as follows.

Ty_1= Total of the first half of y observations

Tx_1=Total of the first half of x- observations

Ty_2=Total of the second half of y-observations

Tx_2=Total of the second half of x-observations

So that;

$$a = \frac{Ty_1}{n} - \frac{bTx_1}{n}$$

$$b = \frac{Ty_2 - Ty_1}{Tx_2 - Tx_1}$$

Gives the trend line

$Yt= a + bt$

Method of Moving Average

This is a technique of smoothening out erratic (random) fluctuations by eliminating seasonal, cyclical and irregular movements in time series. It is the easiest and most commonly used. The disadvantage of this method is that the average of other periods are taken into account instead of actual values corresponding to the particular periods and it does not necessarily give us the equation of the line. However, a straight line can be drawn from points obtained by the moving averages.

Method of Least Squares

This is a mathematical method extensively employed for fitting the appropriate trend line which approximates t a straight line trend. A linear relationship of this kind is usually represented by an equation of first degree. That is:

$Y = a + bt$

Where Yt = the estimated trend value at time t

a = the gradient/slope of the trend line

t = time unit

The estimates of parameters of the trend equation are a and b which are estimated below:

Given a regression equation

$Yi = a + bx + ei$

$E(Yi) = a + bx$

$Ei = Yi - E(Yi)$

$\sum(ei)^2 = \sum(Y - a - bX)^2$

Using the method of least square $\sum(ei)^2$ has no be minimized to get a and b.

Minimizing $\sum(ei)^2$ requires setting its partial derivatives with respect to a and b, equating it to zero.

$$\frac{\delta\sum(ei)^2}{\delta a} = -2\sum(Y - a - bX) = 0$$

$\sum Y - na - b\sum X = 0$ ……..(1)

$$\frac{\delta\sum(ei)^2}{\delta b} = -2\sum(Y - a - bX) = 0$$

$\sum XY - a\sum X - b\sum X^2 = 0$……(2)

From eqn(1) & (2)

$\sum Y = na + b\sum X$ ……(3)

$\sum XY = a \sum X + b\sum X^2$ …….(4)

Equation (3) & (4) are called the normal equations, solving them simultaneously.

Multiply equation (3) by $\sum X$ and (4) by n, we have

$\sum X\sum Y = na\sum X + b(\sum X)^2$…….(5)

$n\sum XY = na\sum X + nb\sum X^2$ ……..(6)

Subtract equation (5) from (6), we have

$n\sum XY - \sum X\sum Y = nb\sum X^2 – b(\sum X)^2$

$n\sum XY - \sum X\sum Y = b(n\sum X^2 – (\sum X)^2)$

$$b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 – (\sum X)2}$$

To get a from the equation (3)

$\sum Y = na - b\sum X$

$Na = \sum Y - b\sum X$

Dividing through by n

$$a = \frac{\sum Y}{n} – \frac{b \sum X}{n}$$

$a = Y – bX$

The above mathematical proof gives detail on linear regression $Y = a + b Xi$. Thus, the advantage of this method over other methods is that there are no missing values n the moving average method where data is lost at the beginning and at the end. It also allows for easy forecasting by fitting the required time period on the regression line.

3.7 ESTIMATION OF SEASONAL VARIATION

Seasonal variation is one of the component forces that determine the size variable at any point in time. It is studied in other to eliminate the seasonal movement so that the fluctuations can be more clearly revealed. Once a trend has been established by whatever method, the

seasonal variation can be determined depending on the model we assume either additive model or multiplicative.

There are number of techniques by which seasonal variation can be measured, but the most widely used is the so called moving average method. The steps involved in computing the seasonal index for quarterly data using this method are as follows;

Firstly, we get the 4 – quarterly moving total by adding together period 1 to 4, period 2 to 5, period 3 and so on. Each resulting 4 – quarterly moving total is placed in the middle of the period related to it. For instance, a three year average for period 1 to 3 is centered at period 2 and that of period 2 to 4 at period 3 and so on.

The 2 – quarter moving total are then computed by adding together 2 periods each in pairs. Then, the resulting 2 –quarter central moving total which is the trend.

3.8 DESEASONALISATION OF DATA

A great time series, particularly in meteorology are affected by seasons. Similarly other natural of shorter duration generate periodic effect such as daily rainfall at a give spot. However, if multiplicative model is assumed,

Xt = xt

Corresponding seasonal variation if additive model is

Xt = xt – corresponding seasonal variation

Seasonal Data = Trend + Residual

Original – Seasonal

3.9 FORCASTING METHOD

These can approached in two ways:

- Exponential Smoothing: There exponential smoothening is the most widely used time series forecasting technique. It is easily programmed for computer application, and

offers several computational advantage over moving average time series forecasting model. There are two exponential smoothing models, they are single exponential and double exponential smoothing, the former being applicable in and absence of trend and latter being applicable when the time series is exhibiting some type of growth pattern.

- Autoregressive approach: If observation in a given time series are highly correlated over time, it may be possible to forecast a future of the series using the past observation. A correlation is useful statistical tool for assessing the time-lagged correlations, or the autocorrelation present in time series data. A kth order regressive time series forecasting model is generally denoted by

  $Y_t = b0 + b_1 Y_{t-1} + b_2 Y_{t-2 + \ldots +} b_k Y_{t-k}$

3.9.1 AUTOCOVARIANCE AND AUTOCORRELATION FUNCTION (ACF)

An important guide to the properties of a time series is provided by series of quantities called the sampled correlation coefficients. Auto correlation is a measure of dependence of time series at a certain time on the values at another time. It is a Pearson correlation between all pairs of point in the time series with a given separation in time or lag. It is an important guide to the properties of the time series. It identifies whether a time series is stationary or non stationary. The autocorrelation plot at various lag called a correlogram.

For stationary process (X_t), the mean $E(X_t) = \mu$ and variance, $Var(X_t) = E(X_t - \mu)^2 = \sigma^2$, which are constant, and the auto-covariance function between X_t and $X_t + k$ is

$\gamma = cov(X_t\ X_{t+k}) = E(X_t - \mu)(X_{t+k} - \mu)$

The auto-correlation function (ACF) between X_t and X_{t+k}

$$\rho = \frac{cov\ (X_t\ \ X_{t+k})}{\sqrt{Var\ (Xt)\ Var(Xt+k)}} = \frac{\gamma_k}{Y_o}$$

3.9.2 PARTIAL AUTOCORRELATION FUNCTION PACF

Another useful method to examine serial dependencies is to examine partial auto-correlation function (PACF) an extension of auto-correlation, where the dependence on the intermediate elements (those within the lag) is removed. In other words, the partial auto-correlation is similar to autocorrelation, except that when calculating it, autocorrelations with all elements within the lag are partial are partial led out (Box and Jenkins, 1976).

The partial auto-correlation function (PACF) is the quantity Φ_{kk} regarded as function of lag k. the PACF is given as

$$\Phi_{kk} = \frac{\rho_{k+1} - \sum \Phi_{kj}, \rho_{k+1-j}}{1 - \sum \Phi_{kj} \rho_j}$$

Note that $\Phi_{11} = \rho_1$

3.9.3 RANDOM WALKS

Pure random processes are useful in many situation, particularly as building blocks for more complicated processes such as moving average processes (Chatfield, 2004). Purely random processes are sometimes called white noise, particularly by engineers.

Suppose that $(\in_t)$ is a discrete time purely random process with mean μ and variance σ^2. A process (X_t) is said to be a random walk if

$$\mathbf{X_t = X_{t-1} + \in_t}$$

The process is customarily started at zero when t=0, so that and

$$X_t = \sum_{i=1}^{t} e$$

Then we find $E(X_t) = t\mu$ and $var(X_t) = t\ \sigma^2_{\in}$. As the mean and variance change with t, the process is non stationary. However, it is interesting to note that the first difference of a random walk, given by

$\Delta X_t = X_{t-1} - X_{t-1} = \in_t$

Form a purely random process, which is therefore stationary

3.9.4 THE AUTOREGRESSIVE (AR) PROCESS

Whenever autocorrelation exist, it is then possible to predict the value of the time series at the current period given the value of the previous period. Autoregressive states that value of any term in the series depends upon the value of the previous terms.

If X_t is the value of series at time t, and X_{t-1} is the series at time t-1 (the immediate previous value). he first order autoregressive model given as

$$X_t = \Phi X_{t-1} + e_t$$

Where Φ is a measure of closeness between X_t and X_{t-1} and e is the random error term.

Also, the model can be written as

$$X_t = \alpha + \Phi X_{t-1} + e_t$$

The equation indicated that X_t is made up of a constant term and a part or fraction of the term immediately preceding it plus a stochastic error.

AUTO CORRELATION FUCTION OF AN AR(P) PROCESS

Consider a zero mean AR(p) defined by

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + .. \Phi_p X_{t-p} + e_t$$

Multiplying both sides by X_{t-k} and taking expectations, we have

$$\gamma_k = \Phi_1 \gamma_{t-1} + \Phi_2 \gamma_{t-2} + .. \Phi_p \gamma_{t-p} + E(X_t + ke_t)$$

3.9.5 MOVING AVERAGE PROCESSES

Suppose that (e_t) is purely random process with mean zero and variance σ^2_e. then a process (X_t) is said to be a moving average process of order q, denoted as MA(q) process if,

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}$$

Where e_t is a white noise process with variance σ^2_e.

The moving average process given above is always stationary. Therefore, it does not require stationary conditions. The parameters must satisfy a condition similar to those of AR(p) for the process to be invertible.

For invertibility of MA(q) process, all the roots of the characteristics equation

Given $\theta(B)= 1 - \theta_1 B - ..- \theta_q B_q$

3.9.6 AUTOREGRESSIVE MOVING AVERAGE (ARMA) PROCESS

The principle o parsimony is to describe a system by as few parameters as possible. Hence parsimony is achieved by combining AR and MA. The resultant model called an ARMA model.

A time series (X_t) is said to follow an autoregressive moving average model of order p,q ARMA(p,q) if it satisfies.

$X_t - \Phi_1 X_{t-1} - .. \Phi_p X_{t-p} = e_t - \theta_1 e_{t-1} - .. \theta_q B_{t-q}$

Where $\Phi(B)X_t = \theta(B)e_t$

3.9.6 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS (ARIMA)

The general ARIMA model introduced by Box and Jenkins(1976) includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are; autoregressive parameters (P), the number of differencing passes (d), and moving average parameters (q). In the notation introduced by Box and Jenkins, the models are summarized as ARIMA(p,d,q).

Most time series are non stationary. A first step in practical modeling with ARMA models is to transform the observation series into as close possible a stationary series. Transformations is stabilize variance are supplemented in time series analysis with time dependent

transformation called differencing. A time series exhibiting a constant drift in trend may be transformed to a stationary series by taking first differences.

$$Z_1 = X_t - X_{t-1}$$

When a series is difference in this way, the original indifferences series is called integrated. An ARMA model applied to a differenced series is referred to as ARIMA models for the original series. Pole, et al., (1994).

Seasonality can be removed by seasonal differencing. For monthly data exhibiting an annual cycle the twelfth seasonal difference $(1-B_{12})X_{tn}$ removes the seasonality, of course, such differencing removes any linear drift in mean also.

General remarks on Model building

Model building in time series depends on a number o factors, these are the properties o the series as assessed by a usual examination of data, the number of observation available and finally the way the model is to be used.

If a parametric model is required, an ARMA model should be considered. The observed correlogram and partial autocorrelation functions are examined, the appropriate ARMA model identified and the model parameters estimated by least squares.

3.9.7 BOX-JENKINS MODELLING PROCEDURE

The time series model used in Box-Jenkins forecasting are called autoregressive integrated moving average (ARIMA) models.

Box-Jenkins modeling relies heavily on the use of three familiar time series tools: differencing, autocorrelation function(ACF) and the Partial autocorrelation function(PACF). Differencing is used to reduce non-stationary series to stationary ones. The ACF and PACF are then used to identify an appropriate ARIMA model and required number of parameters.

After the model is identified, parameter estimates are obtained; that is, the selected model is fit to the available data. The algorithm used is based on the least squares concept and usually required several iterations before producing the desired estimates.

Box-Jenkins models can only be applied to stationary series which have been made stationary by differencing. the models fall into one of the following three categories.

- Purely autoregressive (AR) models
- Purely moving average (MA) models
- Mixed autoregressive – moving average (ARMA)models

If differencing is required to achieve stationarity, then the series will eventually have to be indifference or integrated before forecasting. The Box-Jenkins iterative cycles o model building consists of the following steps: identification, Estimation, Diagnostic check and Forecasting.

CHAPTER FOUR

DATA PRSENTATION AND ANALYSIS

4.0 RESULTS PRESENTATION AND DISCUSSION OF FINDINGS

This chapter reveals the data interpretation and result. Each data are presented on a table according to each month of collection.

4.1 DATA PRESENTATION

year	Month	No Patient affected by cancer
2004	JAN	31
	FEB	32
	MAR	24
	APR	51
	MAY	42
	JUN	21
	JUL	5
	AUG	20
	SEP	41
	OCT	60
	NOV	29
	DEC	22
2005	JAN	20
	FEB	22
	MAR	45
	APR	24
	MAY	54
	JUN	17
	JUL	49
	AUG	18
	SEP	34
	OCT	32
	NOV	41
	DEC	16
2006	JAN	45
	FEB	60
	MAR	41
	APR	52
	MAY	33
	JUN	18
	JUL	22

	AUG	19
	SEP	21
	OCT	32
	NOV	43
	DEC	48
2007	JAN	48
	FEB	42
	MAR	21
	APR	24
	MAY	20
	JUN	31
	JUL	54
	AUG	29
	SEP	44
	OCT	22
	NOV	40
	DEC	33
2008	JAN	60
	FEB	52
	MAR	22
	APR	45
	MAY	43
	JUN	19
	JUL	32
	AUG	52
	SEP	60
	OCT	28
	NOV	44
	DEC	41
2009	JAN	32
	FEB	41
	MAR	32
	APR	39
	MAY	31
	JUN	61
	JUL	24
	AUG	26
	SEP	10
	OCT	42
	NOV	24
	DEC	32
2010	JAN	52
	FEB	38
	MAR	91

	APR	65
	MAY	44
	JUN	38
	JUL	29
	AUG	90
	SEP	31
	OCT	21
	NOV	24
	DEC	33
2011	JAN	19
	FEB	21
	MAR	22
	APR	18
	MAY	33
	JUN	32
	JUL	43
	AUG	45
	SEP	60
	OCT	48
	NOV	52
	DEC	42
2012	JAN	22
	FEB	39
	MAR	32
	APR	26
	MAY	29
	JUN	32
	JUL	66
	AUG	44
	SEP	37
	OCT	25
	NOV	42
	DEC	22
2013	JAN	11
	FEB	48
	MAR	13
	APR	20
	MAY	21
	JUN	39
	JUL	49
	AUG	59
	SEP	22
	OCT	18
	NOV	31

	DEC	3

Table 4.1

Time Plot

The first step in a time series analysis is to read the data into a choice statistical package for the analysis such as R, S-plus, SPSS and so on. The next step after reading the data into the software is to plot the observations against time referred to as time plot in other to be able to estimate the trend of the data

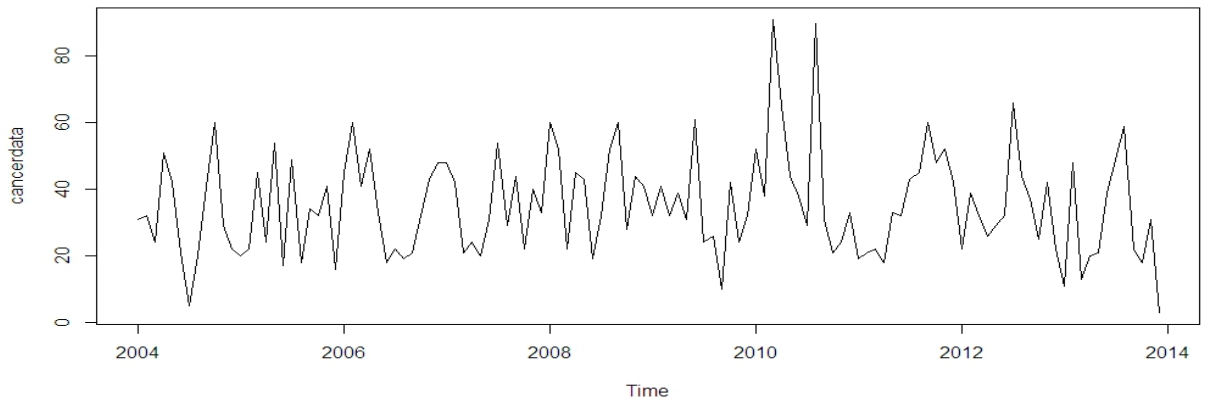

Fig.4.1: Time plot for the monthly record of patients affected by cancer

From the time plot, it appears that the random fluctuations in the time series are roughly constant in size over time, so an additive model is probably appropriate for describing this time series.

We can see from this time series that there seems to be seasonal variation in the number of cancer patient recorded per month: there is a peak every summer, and a trough every winter. Again, it seems that this time series could probably be described using an additive model, as the seasonal fluctuations are roughly constant in size over time and do not seem to depend on

the level of the time series, and the random fluctuations also seem to be roughly constant in size over time.

Furthermore, the time series appears to be stationary in mean and variance, as its level and variance appear to be roughly constant over time. For such a stationary series, the Ljung box and Jenkins ARIMA model approach to times series will be ideal for its analysis and the tools to be used will be: Autocorrelation function (ACF) and partial autocorrelation function (PACF) Therefore, we do not need to difference this series in order to fit an ARIMA model, but can fit an ARIMA model to the original series (the order of differencing required, d, is zero here). The purpose of using both ACF and PACF is for model identification, estimation(like forecasting errors, distribution of the forecasting error), diagnostic check and forecasting.

4.2 DECOMPOSING THE SEASONAL DATA

A seasonal time series consists of a trend component, a seasonal component and an irregular component. Decomposing the time series means separating the time series into these three components: that is, estimating these three components.

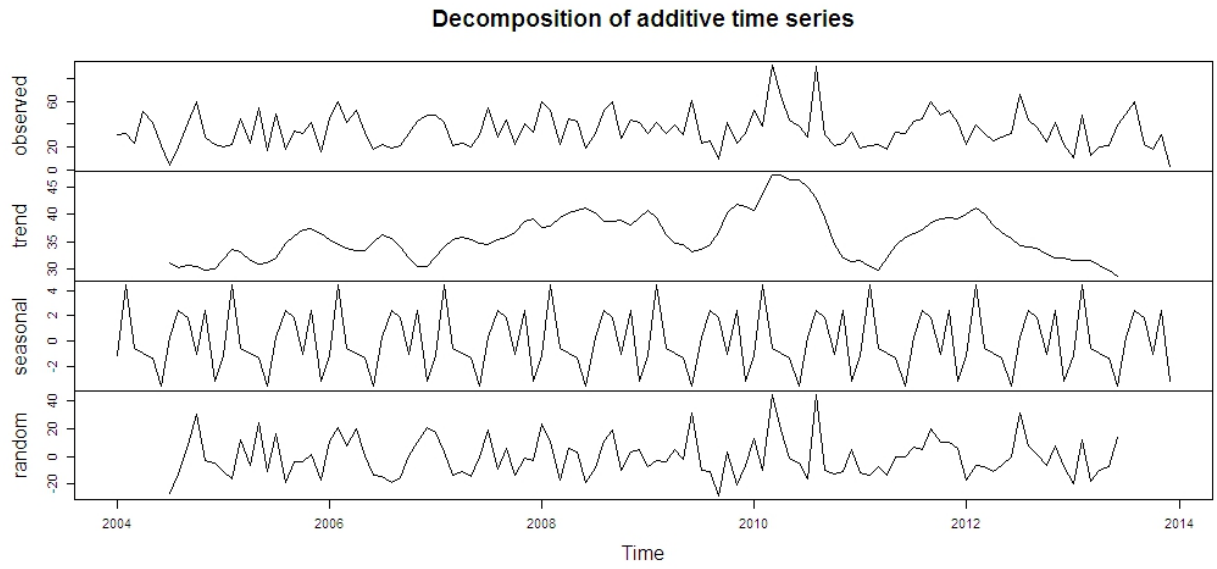

Fig 4.2: decomposition of additive time series into trend, seasonal and irregular component

The plot above shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component (bottom). We see that the estimated trend component shows a small increase from about 35 in 2008 to about 45 in 2010, followed by a steady decrease from then on to about 25 in 2011.

4.3 SEASONAL ADJUSTMENT

Since we have a seasonal time series that is described using an additive model, we can seasonally adjust the time series by estimating the seasonal component, and subtracting the estimated seasonal component from the original time series.

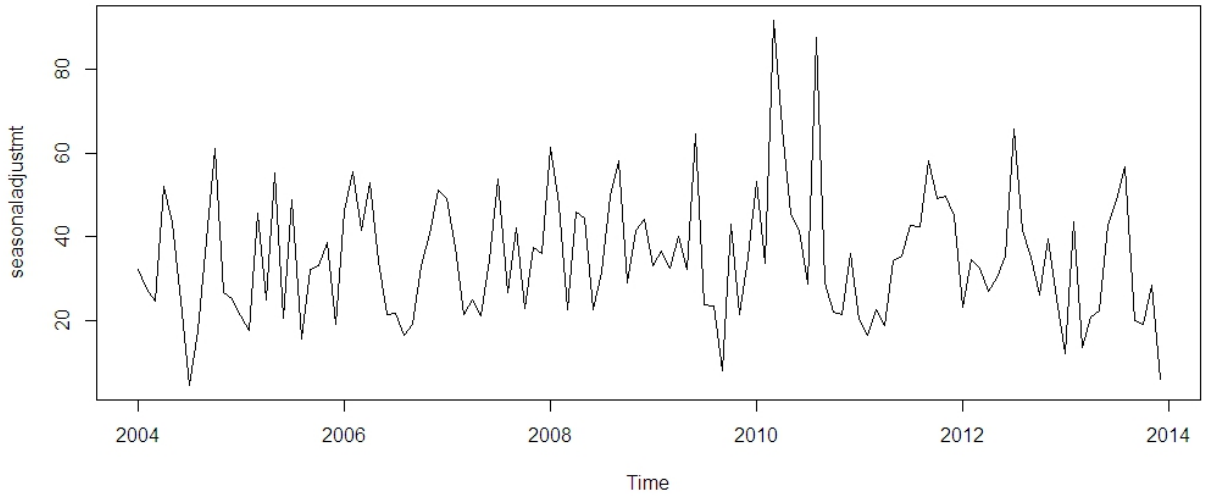

Fig 4.3: Plot of Seasonal adjustment of patient affected by cancer

We can see that the seasonal variation has been removed from the seasonally adjusted time series. The seasonally adjusted time series now just contains the trend component and an irregular component.

4.4 MODEL IDENTIFICATION

The model for the cancer rate in Nigeria can be tentatively achieved by estimating the auto-correlation (ACF) and Partial autocorrelation function (PACF).

The selection of tentative time series model is frequently accomplished by matching estimated sample autocorrelation of the underlying stochastic processes suggests that the series is stationary with ACF and PACF

Below is the estimated ACF

1	2	3	4	5	6	7
0.0000	0.0833	0.1667	0.2500	0.3333	0.4167	.5000
8	9	10	11	12	13	14
0.5833	0.6667	0.7500	0.8333	0.9167	1.0000	1.0833
15	16	17				
1.1667	1.2500	1.3333				
1.000	-0.769	0.386	-0.179	0.058	0.076	0.178
0.245	-0.273	0.216	-0.104	0.024	0.009	0.047
0.096	-0.109	0.078				
1.4167	1.5000	1.5833	1.6667	1.7500	1.8333	1.9167
2.0000	2.0833	2.1667	2.2500	2.3333	2.4167	2.5000
2.5833	2.6667	2.7500				
-0.054	0.045	0.005	-0.108	0.231	-0.316	0.278
-0.129	0.013	0.012	-0.074	0.235	-0.362	0.334
-0.203	0.072	0.023				
2.8333	2.9167	3.0000	3.0833	3.1667	3.2500	3.3333

3.4167 3.5000 3.5833 3.6667 3.7500 3.8333 3.9167

4.0000 4.0833 4.1667

-0.100 0.160 -0.178 0.150 -0.093 0.018 0.034
\

-0.018 -0.010 -0.018 0.054 -0.036 -0.021 0.069

-0.079 0.071 -0.073

Below is the estimated PACF

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833

0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667

1.2500 1.3333 1.4167

-0.769 -0.501 -0.387 -0.440 -0.189 -0.259 0.031
0.040 -0.024 0.044 0.061 0.039 -0.020 0.003
0.037 -0.002 -0.082

1.5000 1.5833 1.6667 1.7500 1.8333 1.9167 2.0000
2.0833 2.1667 2.2500 2.3333 2.4167 2.5000 2.5833
2.6667 2.7500 2.8333

-0.081 0.078 -0.131 0.182 0.058 -0.120 0.045
0.064 0.013 -0.206 0.076 0.008 -0.107 0.035
0.044 0.097 -0.023

2.9167 3.0000 3.0833 3.1667 3.2500 3.3333 3.4167
3.5000 3.5833 3.6667 3.7500 3.8333 3.9167 4.0000
4.0833 4.1667

0.003 0.113 -0.026 -0.026 -0.098 -0.140 0.060
-0.029 0.043 0.061 0.054 -0.065 -0.006 -0.013
-0.098 0.087

Table 4.2 Sample ACF and PACF for monthly record patients affected by cancer in Nigeria.

We can now plot a correlogram and partial correlogram for lags 1-50 to investigate what ARIMA model to use:

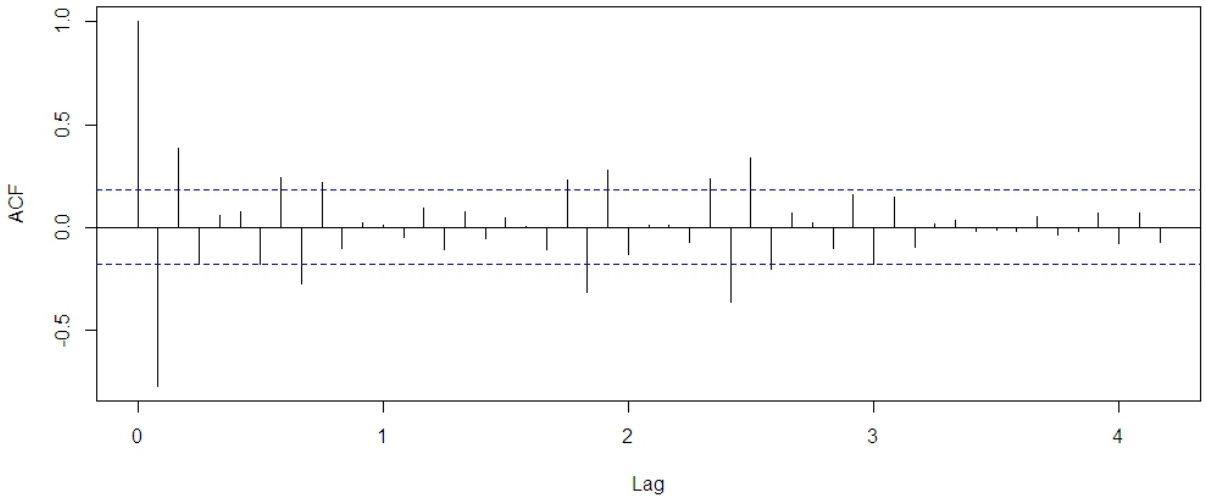

Fig 4.4: ACF for monthly record of rates of patient affected by cancer

We see from the correlogram that the autocorrelations for lags 1 and 2 exceed the significance bounds, and that the autocorrelations tail off to zero after lag 2. The autocorrelations for lags 1, 2 are positive.

The autocorrelation for lags 2.5 and 3 exceed the significance bounds too, but it is likely that this is due to chance, since they just exceed the significance bounds, the autocorrelations for lags 4 do not exceed the significance bounds, and we would expect 1 in 50 lags to exceed the 95% significance bounds by chance alone, which shows that p=1

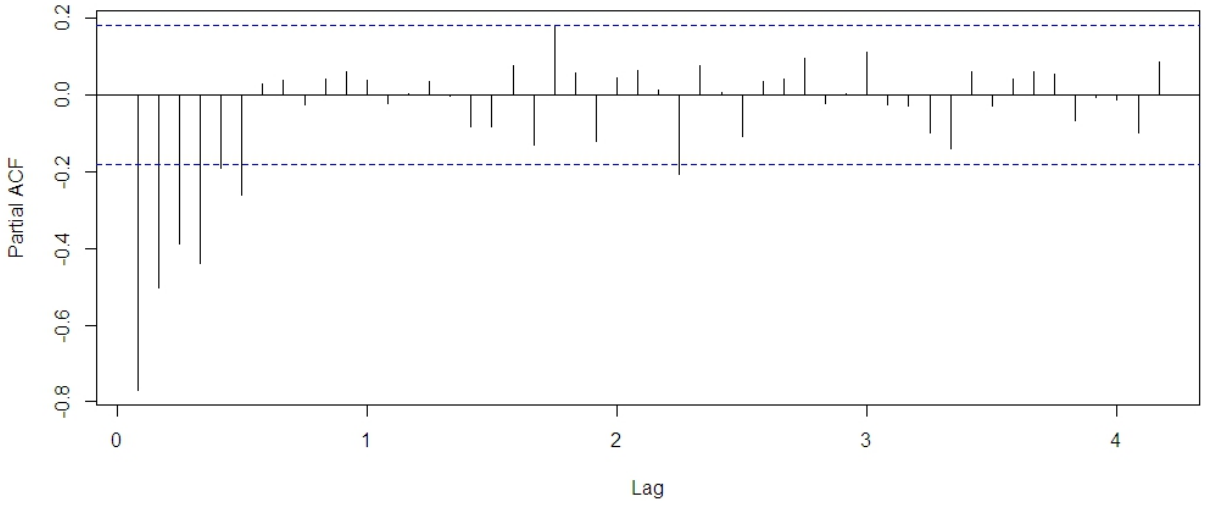

Fig 4.5: PACF for monthly record of rates of patient affected by cancer

From the partial auto-correlogram, we see that the partial autocorrelation at lag 0 is positive and exceeds the significance bounds (0.79).The partial autocorrelations tail off to zero after lag 0.5, which shows that q=0.

Plot of the correlogram greatly assisted in the understanding of the model. From the PACF, the model cut-off after lag1 indicating that the model can either be an ARIMA(1,0,0) or an ARIMA(0,0,1) is the same as AR(1) while ARIMA(0,0,1) is the same as MA(1) model.

The R-language package used the Akaike Information Criteria (AIC) to provide best fit for an autoregressive model to a set of data. The values of the AIC generally listed for autoregressive for order 0, which is the white noise model. The model with smallest value of the AIC is judge to be the most appropriate. From PACF, the model has a cut-off after lag 1 which is along the exponential decay seen in the ACF, we conclude that the model is autoregressive of order one, AR(1). AIC=999.26 AICc=999.47

4.5 PARAMETER ESTIMATION

By fitting an ARIMA(1,0,0) MODEL for rate of patients affected by cancer in Nigeria, the value of the parameter was gotten to be 1. The corresponding fitted autoregressive model is

$$X_t = (0.1722)X_{t-1} + \boldsymbol{e_t}$$

And the AIC for the mode is ***999.26***

4.6 FORCASTING USING ARIMA MODEL

Since we have selected the best candidate ARIMA(p,d,q) model for our time series data, and that we have estimate the parameters of that ARIMA model, then we can use it as a predictive model for making forecasts for future values of patient that will be affected by cancer.

Below table shows the forecast of the number of a patient that will be affected by cancer for the next 10 months.

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2014	29.44592	10.00076	48.89109	-0.2928874	59.18473
Feb 2014	34.13328	14.38504	53.88152	3.9309589	64.33561
Mar 2014	34.96408	15.20640	54.72177	4.7473150	65.18085
Apr 2014	35.11134	15.35336	54.86932	4.8941152	65.32856
May 2014	35.13744	15.37945	54.89543	4.9202007	65.35468
Jun 2014	35.14206	15.38407	54.90005	4.9248263	65.35930
Jul 2014	35.14288	15.38489	54.90087	4.9256462	65.36012
Aug 2014	35.14303	15.38504	54.90102	4.9257915	65.36027
Sep 2014	35.14306	15.38507	54.90104	4.9258173	65.36029
Oct 2014	35.14306	15.38507	54.90105	4.9258218	65.36030

Table 4.3: forecast of the number patient affected by cancer in the next 10 months

We can plot the observed record of patent affected by cancer for the past 120months, as well as the records that would be predicted for these patients for the next 10 months using our ARIMA (1,0,0).

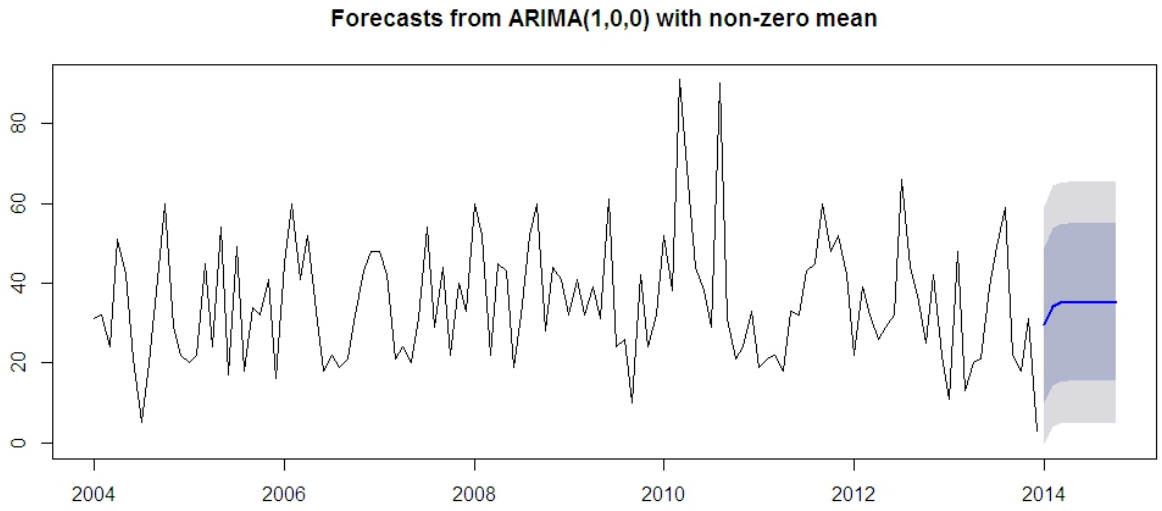

fig 4.6: forecast for record of patients affected by cancer for the next five years

4.7 DIAGNOTSIC CHECKING (FORECAST ERRORS)

The main aims of diagnostics are: to validate the model, or failing that, to point the way to a better model choice. To investigate whether the forecast errors of an ARIMA model are normally distributed with mean zero and constant variance, and whether there are correlations between successive forecast errors, we can make a correlogram of the forecast errors for our ARIMA(1,0,0) model for the patient affected by cancer, and perform the Ljung-Box test for lags 1-50

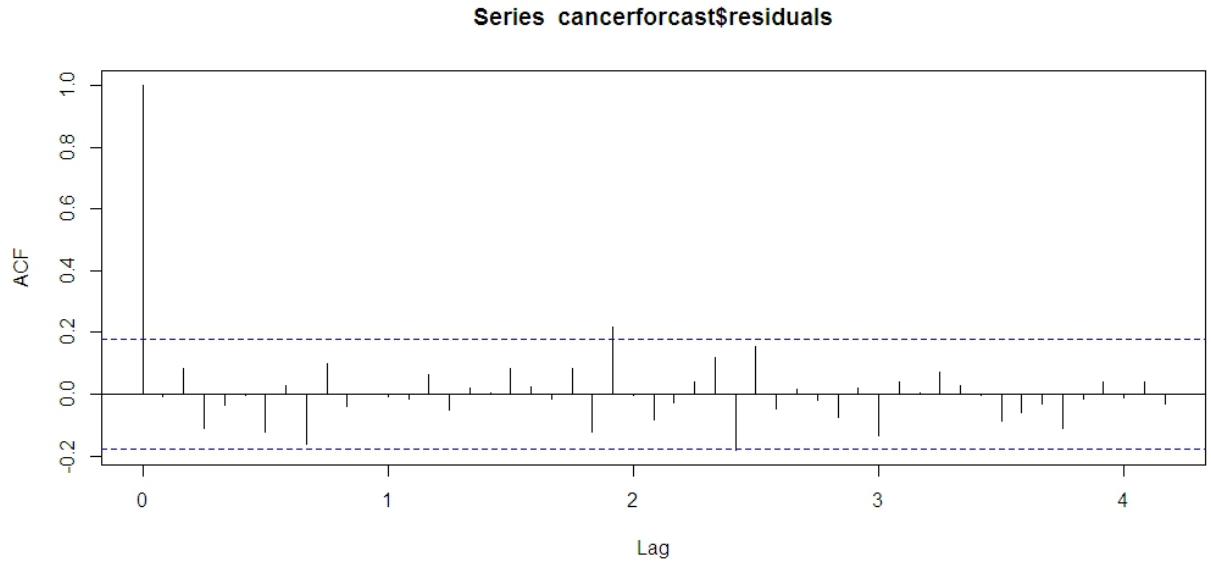

Fig 4.7: ACF of the cancer forecast residual

The correlogram shows that the sample autocorrelation at lag 2 exceeds the significance bounds. However, this is probably due to chance, since we would expect one out of 50 sample autocorrelations to exceed the 95% significance bounds. Furthermore, the p-value for the Ljung-Box test is *0.6199*, indicating that there is little evidence for non-zero autocorrelations in the forecast errors for lags 1-50.

We can construct histogram, to check the distribution of the errors. To check whether the forecast errors are normally distributed with mean zero and constant variance, we make a time plot of the forecast errors, and a histogram:

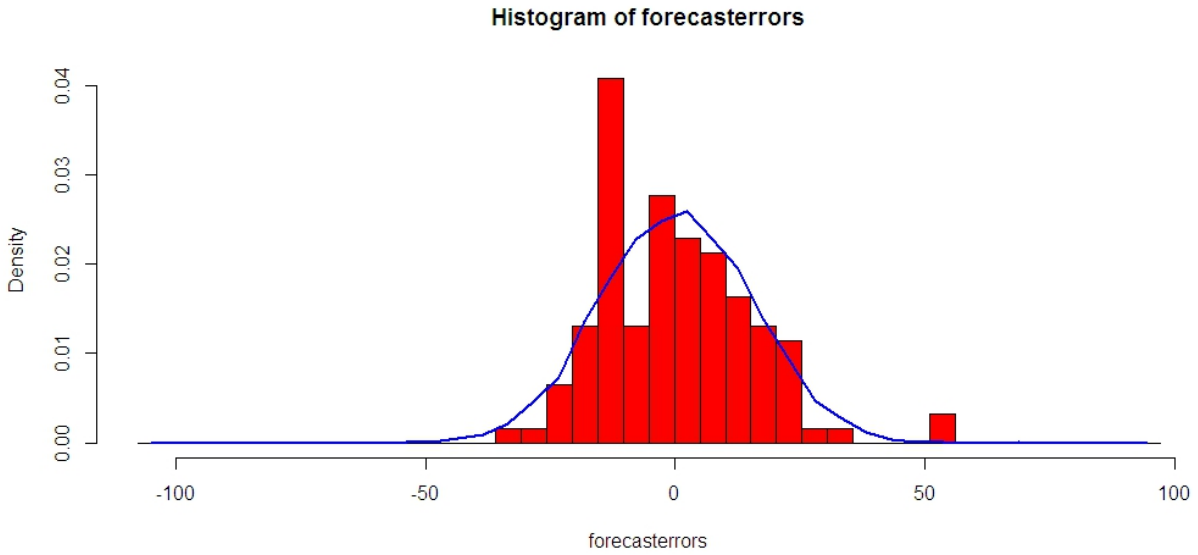

Fig 4.8: Time plot and histogram of the forecast error distribution

The time plot of the forecast errors shows that the variance of the forecast errors seems to be roughly constant over time (though perhaps there is slightly higher variance for the second half of the time series). The histogram of the time series shows that the forecast errors are roughly normally distributed and the mean seems to be close to zero. Therefore, it is plausible that the forecast errors are normally distributed with mean zero and constant variance.

Since successive forecast errors do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the ARIMA(1,0,0) does seem to provide an adequate predictive model for the record of patient that will be affected by cancer.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 SUMMARY

Time series analysis was carried out on the record of patient affected by cancer in Nigeria using the data generated from Ibadan Cancer Registry, Department of Pathology, University College of Hospital(UCH).

The analysis took the Ljung box and Jenkins model approach.

5.2 CONCLUSION

This study demonstrates the use of classical time series analysis in modeling rate at which patient are affected by cancer in Nigeria. The study revealed that there will be an increase in the trend of cancer in Nigeria as from January 2014 to October 2014. The implication of this is that more cases of cancer will be reported for the next 10months which may increase death rate.

5.3 RECOMMENDATION

From the research findings, are recommend that Nigeria authority should focus more on the well being of their people and give more support to the cancer registery in Nigeria

Government should try and establish more cancer registration centre both in the urban area and rural area in other to help in collecting adequate information on the people affected by cancer

Government should help ministry of health in establishing different kind of health program to enlight people on the causes and the prevention of cancer, also government need to make appropriate provision on the needed facilities to help in preventing cancer in Nigeria.

REFERENCES

Adebamowo, C. A. and Ajayi (2000*).*Breast Cancer in Nigeria. *West African Journal of Medicine* 19: 179-171

Adebamowo C.A., T. O. Ogundiran, A. A, Adenipekun, R. A. Oyesegun, O. B. Campbell, E. E. Akang, C. N. Rotimi, and O. I. Olopade (2002) "Waist-Hip Ratio and Breast Cancer Risk in Urbanized Nigerian Women." *Breast Cancer Research* .

Althuis, M.D.Dozier,J. M; Anderson, W.F; Devesa,S.S and Brinton,L.A (2005) *Global Trends in Breast Cancer and Mortality; 1973-1997*. *Int. Journal of Epidemiology,* Accessed on on 14th August 2007

Bakkita, O. B. (1998). "Population Education Reaching Out to Rural Women in Nigeria for Improved Quality of Life." W*omen Education and Development. Vol 1. no.1*

Cancer Statistics Worldwide (2005*). London Cancer Research Report*. (No.104) U.K. Author

Carver, C. S. (2005). *Enhancing Adaptation during Treatment and the Roleof Individual Differences.* (No. 104, 2602-2607) UK Cancer Record. U.K. Author.

Cockburn, C.J., Sabrina, P and Sally Redner (1999*) Perception of screening mammography among women aged 40 – 49 years*. University of New Castle. South Wales Hunter, School of Population HealthScience.

Crimmins, E.M;Hayward,M.D and Saito,Y (1994). "Changing Mortality and Morbidity Rates and the Health Status and Life Expectancy of the Older Population." *Demography* 31(1):159-175.

Crimmins, E.M and Saito,Y (2000). "Change in the Prevalence of Diseases among Older Americans 1984-1994." *Demographic Research* 3(9)

Gallucci BB. Selected concepts of cancer as a disease. From the Greeks to 1900. Oncol Nurs Forum. 1985;12:67–71. [PubMed]

Ikpah, O.F Kuopio, T., Collan, Y.(2002). "Proliferation in African Breast Cancer Biology and Prognostication in Nigeria. *Modern Pathology*.

Kalluri R. Basement membranes: structure, assembly and role in tumour angiogenesis. Nat Rev Cancer. 2003;3:422–433. [PubMed]

Kardinal C, Yarbro JA. Conceptual history of cancer. Semin Oncol. 1979;6:396–408. [PubMed]

Lyons AS, Petrucelli RJ. Medicine: An Illustrated History. New York: Harry N. Abrams Publishers; 1978.

Ogundipe, S. and Obinna, C. (2008). "Why Cancer is on the Rise in Nigeria." Retrieved on 10thMay,2007from*http://search.yahoo.com/search?p=Ogundipe+and+Obinna+%2B 2008&ei=UTF8&fr=yfpt501&xargs=0&pstart=1&b=1&xa=dt0JTLeZxSgKnoJZWu YYcQ--,1247153072*

Olopade, F. (2004). "Why Take it if you don't Have Anything? Breast Cancer Risk Perceptions and Prevention Choices At A Public Hospital." *Canada Pubmed Online Journal of the National Library of Medicine and the National Institute of Health.*

Othman, N., et al., *The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens.* Biochim Biophys Acta, 2001. **1471**(2): p. C1-10.

WHO. *Cancer Prevention Programme*. 2010 Available from: http://www.who.int/cancer/en/. IARC, *World Cancer Report*. 2008, IARC: Lyon.

WHO. *National cancer control programmes: policies and managerial guidelines (2nd edition),* 2002, Geneva: WHO.

APPENDIX

BELOW IS RESULT OUTPUT OF THE PROJECT

```
> cancer<-read.table("newdat.csv",header=TRUE,sep=",")
> cancerdata<-ts(cancer$month, frequency=12,start=c(2004,1))
> cancerdata
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2004  31  32  24  51  42  21   5  20  41  60  29  22
2005  20  22  45  24  54  17  49  18  34  32  41  16
2006  45  60  41  52  33  18  22  19  21  32  43  48
2007  48  42  21  24  20  31  54  29  44  22  40  33
2008  60  52  22  45  43  19  32  52  60  28  44  41
2009  32  41  32  39  31  61  24  26  10  42  24  32
2010  52  38  91  65  44  38  29  90  31  21  24  33
2011  19  21  22  18  33  32  43  45  60  48  52  42
2012  22  39  32  26  29  32  66  44  37  25  42  22
2013  11  48  13  20  21  39  49  59  22  18  31   3
> plot.ts(cancerdata)
> decomcancer<-decompose(cancerdata)
> plot(decomcancer)
> decomcancer$seasonal
          Jan       Feb       Mar       Apr       May       Jun       Jul       Aug       Sep       Oct
2004 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2005 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2006 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2007 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2008 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2009 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2010 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2011 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2012 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
2013 -1.1689815  4.4467593 -0.5347222 -0.9189815 -1.2893519 -3.5439815  0.3402778
2.4699074  1.8912037 -1.0254630
          Nov       Dec
2004  2.4375000 -3.1041667
2005  2.4375000 -3.1041667
2006  2.4375000 -3.1041667
2007  2.4375000 -3.1041667
2008  2.4375000 -3.1041667
2009  2.4375000 -3.1041667
2010  2.4375000 -3.1041667
```

```
2011  2.4375000 -3.1041667
2012  2.4375000 -3.1041667
2013  2.4375000 -3.1041667
> seasonaladjustmt<-cancerdata - decomcancer$seasonal
> plot(seasonaladjustmt)
> cancerdiffstate<-diff(cancerdata,differences=1)
> plot.ts(cancerdiffstate)
> cancerdiffstate2<-diff(cancerdata,differences=2)
> plot.ts(cancerdiffstate2)
> cancerdiffstate3<-diff(cancerdata,differences=3)
> plot.ts(cancerdiffstate3)
> cancerdiffstate4<-diff(cancerdata,differences=4)
> plot.ts(cancerdiffstate4)
> acf(cancerdiffstate3,lag.max=50)
> acf(cancerdiffstate3,lag.max=50, plot=FALSE)

Autocorrelations of series ‘cancerdiffstate3’, by lag

0.0000 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167
1.0000 1.0833 1.1667 1.2500 1.3333
 1.000 -0.769  0.386 -0.179  0.058  0.076 -0.178  0.245 -0.273  0.216 -0.104  0.024  0.009 -
0.047  0.096 -0.109  0.078
1.4167 1.5000 1.5833 1.6667 1.7500 1.8333 1.9167 2.0000 2.0833 2.1667 2.2500 2.3333
2.4167 2.5000 2.5833 2.6667 2.7500
-0.054  0.045  0.005 -0.108  0.231 -0.316  0.278 -0.129  0.013  0.012 -0.074  0.235 -0.362
0.334 -0.203  0.072  0.023
2.8333 2.9167 3.0000 3.0833 3.1667 3.2500 3.3333 3.4167 3.5000 3.5833 3.6667 3.7500
3.8333 3.9167 4.0000 4.0833 4.1667
-0.100  0.160 -0.178  0.150 -0.093  0.018  0.034 -0.018 -0.010 -0.018  0.054 -0.036 -0.021
0.069 -0.079  0.071 -0.073



> pacf(cancerdiffstate3,lag.max=50)
> pacf(cancerdiffstate3,lag.max=50, plot=FALSE)

Partial autocorrelations of series ‘cancerdiffstate3’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000
1.0833 1.1667 1.2500 1.3333 1.4167
-0.769 -0.501 -0.387 -0.440 -0.189 -0.259  0.031  0.040 -0.024  0.044  0.061  0.039 -0.020
0.003  0.037 -0.002 -0.082
1.5000 1.5833 1.6667 1.7500 1.8333 1.9167 2.0000 2.0833 2.1667 2.2500 2.3333 2.4167
2.5000 2.5833 2.6667 2.7500 2.8333
-0.081  0.078 -0.131  0.182  0.058 -0.120  0.045  0.064  0.013 -0.206  0.076  0.008 -0.107
0.035  0.044  0.097 -0.023
2.9167 3.0000 3.0833 3.1667 3.2500 3.3333 3.4167 3.5000 3.5833 3.6667 3.7500 3.8333
3.9167 4.0000 4.0833 4.1667
 0.003  0.113 -0.026 -0.026 -0.098 -0.140  0.060 -0.029  0.043  0.061  0.054 -0.065 -0.006 -
0.013 -0.098  0.087
```

```
> library(forecast)
Loading required package: zoo

Attaching package: ‘zoo’

The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric

Loading required package: timeDate
This is forecast 5.6

> auto.arima(cancerdata)
Series: cancerdata
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1  intercept
      0.1772    35.1431
s.e.  0.0912     1.6810

sigma^2 estimated as 230.2:  log likelihood=-496.63
AIC=999.26   AICc=999.47   BIC=1007.63
> library("forecast")
> cancerarima<-arima(cancerdata,order=c(1,0,0))
> cancerarima
Series: cancerdata
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1  intercept
      0.1772    35.1431
s.e.  0.0912     1.6810

sigma^2 estimated as 230.2:  log likelihood=-496.63
AIC=999.26   AICc=999.47   BIC=1007.63
> cancerforcast<-forecast.Arima(cancerarima,h=10)
> cancerforcast
         Point Forecast    Lo 80    Hi 80      Lo 95    Hi 95
Jan 2014       29.44592 10.00076 48.89109 -0.2928874 59.18473
Feb 2014       34.13328 14.38504 53.88152  3.9309589 64.33561
Mar 2014       34.96408 15.20640 54.72177  4.7473150 65.18085
Apr 2014       35.11134 15.35336 54.86932  4.8941152 65.32856
May 2014       35.13744 15.37945 54.89543  4.9202007 65.35468
Jun 2014       35.14206 15.38407 54.90005  4.9248263 65.35930
Jul 2014       35.14288 15.38489 54.90087  4.9256462 65.36012
Aug 2014       35.14303 15.38504 54.90102  4.9257915 65.36027
Sep 2014       35.14306 15.38507 54.90104  4.9258173 65.36029
Oct 2014       35.14306 15.38507 54.90105  4.9258218 65.36030
> plot.forecast(cancerforcast)
```

```
> acf(cancerforcast$residuals,lag.max=50)
> Box.test(cancerforcast$residuals,lag=50, type="Ljung-Box")

        Box-Ljung test

data:  cancerforcast$residuals
X-squared = 46.368, df = 50, p-value = 0.6199

> plot.ts(cancerforcast$residuals)
> plotforecasterrors(cancerforcasts$residuals)
> plotForecastErrors <- function(forecasterrors)
+   {
+     # make a histogram of the forecast errors:
+     mybinsize <- IQR(forecasterrors)/4
+     mysd   <- sd(forecasterrors)
+     mymin  <- min(forecasterrors) - mysd*5
+     mymax  <- max(forecasterrors) + mysd*3
+     # generate normally distributed data with mean 0 and standard deviation mysd
+     mynorm <- rnorm(10000, mean=0, sd=mysd)
+     mymin2 <- min(mynorm)
+     mymax2 <- max(mynorm)
+     if (mymin2 < mymin) { mymin <- mymin2 }
+     if (mymax2 > mymax) { mymax <- mymax2 }
+     # make a red histogram of the forecast errors, with the normally distributed data
overlaid:
+     mybins <- seq(mymin, mymax, mybinsize)
+     hist(forecasterrors, col="red", freq=FALSE, breaks=mybins)
+     # freq=FALSE ensures the area under the histogram = 1
+     # generate normally distributed data with mean 0 and standard deviation mysd
+     myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
+     # plot the normal curve as a blue line on top of the histogram of forecast errors:
+     points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
+   }
> plotForecastErrors(cancerforcast$residuals)
```