5-2011

Modeling Mortality Rates for Leukemia Between Men and Women in the United States

Blessed Quansah
University of Nevada, Las Vegas

Follow this and additional works at: http://digitalscholarship.unlv.edu/thesesdissertations

Part of the Epidemiology Commons, Mathematics Commons, Oncology Commons, and the Vital and Health Statistics Commons

Repository Citation
http://digitalscholarship.unlv.edu/thesesdissertations/1088

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Theses, Dissertations, Professional Papers, and Capstones by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
MODELING MORTALITY RATES FOR LEUKEMIA BETWEEN MEN AND WOMEN IN THE UNITED STATES

By

Blessed Quansah

Bachelor of Science
Kwame Nkrumah University of Science and Technology, Ghana
2001

A thesis submitted in partial fulfillment of the requirements for the

Master of Science in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
May 2011
THE GRADUATE COLLEGE

We recommend the thesis prepared under our supervision by

**Blessed Quansah**

entitled

**Modeling Mortality Rates for Leukemia Between Men and Women in the United States**

be accepted in partial fulfillment of the requirements for the degree of

**Master of Science in Mathematical Sciences**
Department of Mathematical Sciences

Chih-Hsiang Ho, Committee Chair

Anton Westveld, Committee Member

Amei Amei, Committee Member

Chad Cross, Graduate College Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies and Dean of the Graduate College

**May 2011**
ABSTRACT

Modeling Mortality Rates for Leukemia between Men and Women in the United States

By

Blessed Quansah

Dr. Chih-Hsiang Ho, Examination Committee Chair
Professor of Mathematical Sciences
University of Nevada, Las Vegas

Leukemia related deaths increased dramatically over the last forty years. Leukemia is a malignant disease or cancer of the bone marrow and blood. It is characterized by the uncontrolled accumulation of blood cells. Leukemia is divided into two categories: myelogenous or lymphocytic, each of which can be acute or chronic. The terms, myelogenous or lymphocytic denote the cell type involved.

In this thesis, the proposed modeling techniques are applied to leukemia deaths data from the Surveillance Epidemiology and End Results (SEER). In particular, annual deaths data from 1969 to 2007 are used in the data analysis, which includes three major parts: 1) male and female death rate comparisons using the conditional test (Przyborowski and Wilenski, 1940); 2) development of the empirical recurrence rate (Ho, 2008) and the empirical recurrence rates ratio time series; and 3) the Autoregressive Integrated Moving Average (ARIMA) model: selection, validation, and forecasting for the leukemia death rates and ratio.
ACKNOWLEDGEMENTS

I would never have been able to finish my thesis without the guidance of my committee members, help from friends, and support from my family. I would like to express my deepest gratitude to my advisor, Dr. Chih-Hsiang Ho for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I would like to thank Dr. Westveld who let me experience field and practical issues beyond the textbooks, and patiently corrected my writing. I would also like to thank Dr. Amei and Dr. Cross for guiding my research for the past year and helping me develop my background in statistics.

Special thanks go to Dr Dieudonne Phanord, who as a good friend was always willing to help and give his best suggestions. It would have been a lonely lab without him. I would also like to thank my parents, my two elder brothers and my younger brother; they supported and encouraged me with their best wishes. Finally, I would like to thank my office mate Xundug Sun.
TABLE OF CONTENTS

ABSTRACT..................................................................................................................III
ACKNOWLEDGEMENTS...............................................................................................IV
LIST OF TABLES.........................................................................................................VII
LIST OF FIGURES......................................................................................................IIIX
CHAPTER 1 INTRODUCTION.....................................................................................1

CHAPTER 2 THEORY AND METHOD FOR POINT PROCESSES.........................4
   2.1 Leukemia Data....................................................................................................4
   2.2 Poisson Process................................................................................................4
   2.3 Conditional Test...............................................................................................5
   2.4 Conditional Test for Leukemia.........................................................................7
   2.5 Empirical Recurrence Rates............................................................................8

CHAPTER 3 THEORY AND METHOD FOR ARIMA MODELS.............................10
   3.1 Empirical Recurrence Rates Ratio.................................................................10
   3.2 ARIMA Models..............................................................................................11
      3.2.1 Autoregressive Model of Order p, AR(p).............................................11
      3.2.2 Moving Average Model of Order q, MA(q).........................................12
      3.2.3 Autoregressive Moving Average Model of Order p, q, ARMA(p, q).....12
      3.2.4 Stationary Time Series.......................................................................13
   3.3 Data Transformation.......................................................................................14
      3.3.1 Box-Cox Transformation.....................................................................14
      3.3.2 Differencing.........................................................................................14
      3.3.3 Subtracting the Mean...........................................................................17
LIST OF TABLES

Table 3.1 Behavior of the ACF and PACF for causal invertible ARMA models…………….17
Table 4.1 The numerical values of the actual ERRRs in the prediction set and the predicted
ERRRs and their confidence intervals using the MA(2) based on the training
sample……………………………………………………………………………………………………29
Table 4.2 The numerical values of the predicted ERRRs and their confidence intervals using the
MA(2) based on the full data………………………………………………………………………………30
Table 4.3 Numerical values of the predicted ERRRs with their confidence intervals using
AR(2)………………………………………………………………………………………………35
Table 4.4 Numerical of the actual and the predicted ERRRs with their confidence interval
based on the training sample using ARMA(1, 1)………………………………………40
Table 4.5 Actual and Model predicted values for MA (2), ARMA (1, 1) and AR (2)…41
Table 4.6 The numerical values of the predicted ERRRs with their confidence intervals using
AR(2)………………………………………………………………………………………………45
Table 4.7 The numerical values of the predicted ERRRs with their confidence intervals using
ARMA(1, 1)…………………………………………………………………………………………50
Table 4.8 Predicted value of the three models based on the full data………………………52
LIST OF FIGURES

Figure 2.1 Annual leukemia related deaths data in the United States between 1969 and 2007…4

Figure 2.3 ERR plots for male and female deaths within the study period with time-step h = 1 year.................................................................9

Figure 4.1 ERRR plots with time-step h = 1 year........................................21

Figure 4.2 ERRR plots of the training sample and prediction set with h = 1 year...........22

Figure 4.3 a, Time-plot; b, Sample ACF; c, Sample PACF of the training sample with time-step h = 1 year........................................................................24

Figure 4.4 a, Time-plot; b, Sample ACF; c, Sample PACF of a lag-1 differenced training sample with h = 1 year............................................................25

Figure 4.5 a, Time-plot; b, Sample ACF; c, Sample PACF of the twice-differenced training sample with h = 1 year.........................................................26

Figure 4.6 Diagnostics for the MA(2)fitted and twice-differenced training sample. Residual a; Time-plot; b, Sample ACF; c, Sample PACF...........................................28

Figure 4.7 ERRR plot with prediction intervals..................................................31

Figure 4.8 Comparison of three forecasted ERRRs with the Prediction set.............32

Figure 4.9 The complete data (training sample and prediction set) with three appended to training sample for model validation; Inset: Comparison of three ERRRs with prediction set .................................................................32

Figure 4.10 a, ERRR plots after Box-Cox transformation at λ = 0; b, Sample ACF; c, Sample PACF of the full data with h = 1 year................................................35

Figure 4.11 ERRR plot with prediction intervals using AR(2)...............................36

Figure 4.12 a, ERRR plot after first differencing at lag 1; b, Sample ACF; c, Sample PACF based on the training Sample with h = 1 year........................................37
Figure 4.13 a, ERRR plot after twice-differencing at lag 1; b, Sample ACF; c, Sample PACF based on the training sample with h = 1 year.

Figure 4.14 ERRR plots with prediction intervals using ARMA(1, 1).

Figure 4.15 a, Rescaled Residual-plots; b, Residual ACF; c, Residual PACF with using ARMA(1, 1).

Figure 4.16 Comparison of the models with the actual values based on the training sample.

Figure 4.17 a, ERRR plots after Box-Cox transformation at λ = 0; b, Sample ACF; c, Sample PACF of the full data with h = 1 year.

Figure 4.18 ERRR plot with prediction intervals using AR(2).

Figure 4.19 a, Residual-plot; b, Residual ACF; c, Residual PACF of the full data with h = 1 year.

Figure 4.20 a, ERRR plots after differencing at lag 1; b, Sample ACF; c, Sample PACF of the full data with h = 1 year.

Figure 4.21 a, ERRR plots after twice-differencing at lag 1; b, Sample ACF; c, Sample PACF of the full data with h = 1 year.

Figure 4.22 ERRR plot with prediction intervals Using ARMA(1, 1).

Figure 4.23 a, Residual-plot; b, Residual ACF; c, Residual PACF of the full data with h = 1 year.

Figure 4.24 Comparison of three models based on the full model.
CHAPTER 1

INTRODUCTION

Leukemia is cancer of the human blood cells. It starts in the bone marrow, the soft tissue inside most bones. Bone marrow is where blood cells are made. When you have leukemia, the bone marrow starts to make a lot of abnormal white blood cells, called leukemia cells. The leukemia leukocytes, do not work like the normal white blood cells (leukocytes), instead they grow faster and fail to stop growing than normal leukocytes. Over time, leukemia cells can crowd out the normal white blood cells. The abundance of leukemia leukocytes can lead to serious problems such as anemia, bleeding, and infections. Leukemia cells can also spread to other organs and cause swelling or pain. The four main types of leukemia are as follows:

• Acute lymphoblastic (ALL) is the most common leukemia in children. Adults can also get it.

• Acute myelogenous leukemia (AML) affects both children and adults.

• Chronic lymphocytic leukemia (CLL) is the most common leukemia in adults, who are mostly older than 55 years. Children almost never get it.

• Chronic myelogenous leukemia (CML) occurs mostly in adults.

Experts do not know the causes of leukemia, but some factors are known to increase the risk of some types of leukemia. One is more likely to develop leukemia if exposed to large amounts of radiation, certain chemicals at work such as benzene, chemotherapy to treat another cancer, Down syndrome or other genetic problems, and cigarette smoke. However few people who have these risk factors develop leukemia. Most people who acquire leukemia do not have any known risk factors (National Institute of Health, 2011).
The following report presents detailed data from 1969 to 2007 on death rates according to a number of social, demographic, and medical characteristics. This data provides information on mortality patterns among residents of the United States by variables such as age, sex, and marital status.

In 2007, a total of 2,423,712 resident deaths were registered in the United States. The five leading causes of death in 2007 were:

1. Heart disease
2. Malignant neoplasm (cancer)
3. Cerebrovascular disease
4. Chronic lower respiratory disease
5. Accidents (unintentional injuries)

With 77.9 being the current Life expectancy a continuing increasing is seen based on data from 2006 and 2007. Life expectancy increased for the total population, including both the black and white populations. Both black and white males and females experienced an increase in life expectancy in 2007 compared with 2006. Rates for the top three leading causes for death: heart disease, cancer, and stroke, continued a decreasing trend. The difference in mortality rates between men and women increased slightly in 2007 from 2006 (National Cancer for Health Statistics, 2010).

In this study, the proposed modeling techniques are applied to the leukemia deaths data from the SEER. First, the data of deaths will be divided into two, based on the gender, as follows: 1) Male deaths, and 2) Female deaths. In particular, annual data from 1969-2007 are used in the data analysis, which includes three major parts: 1) leukemia deaths rates comparisons using the conditional test (Przyborowski and Wilenski, 1940); 2) development of the empirical
recurrence rate (Ho, 2008) and the empirical recurrence rates ratio time series; and 3) the Autoregressive Integrated Moving Average (ARIMA) model selection: validation, and forecasting for the Leukemia death rates and ratio.

Death rate comparisons using the conditional test and the empirical recurrence rate time series will be presented in Chapter 2. The fundamental tools of ARIMA are introduced in chapter 3. Chapter 4 illustrates the ARIMA modeling techniques using the empirical recurrence rates ratio generated from annual leukemia deaths data. Chapter 5 concludes our work.
2.1 Leukemia Data

Statistics for deaths that occurred in the United States during the period 1969 to 2007 are obtained from Surveillance Epidemiology and End Results (SEER) Program (www.seer.cancer.gov). From 2003 to 2007 the median age of death for leukemia was 74 years of age. Approximately 3.0% died under age 20; 3.1% between 20 and 34; 3.3% between 35 and 44; 6.4% between 45 and 54; 12.6% between 55 and 64; 21.6% between 65 and 74; 31.6% between 75 and 84; and 18.4% 85+ years of age. (www.revolutionhealth.com)

In the data set, the year 1969 is the time origin \( t_0 \), and 2007 is the present time 0. There were 709,534 leukemia related deaths during the past 39 years (Appendix Table 1). By using the raw data, we construct a line plot to observe any possible trends (Figure 2.1). It is clear from the line plots that the number of deaths due to leukemia is increasing for male and leveling off for female in the last five years.

![Annual leukemia related deaths data in the United States between 1969 and 2007.](image)

**Figure 2.1** Annual leukemia related deaths data in the United States between 1969 and 2007.
2.2 Poisson Process

To reveal hidden characteristics of the leukemia data, we employ a point process to investigate the data and then conduct a conditional test to support our claim. A point process is a stochastic model that describes the occurrences of events. These occurrences are thought of as points on the time axis. Let \( N(t) \) be the random variable that denotes the number of events in the interval \((0, t]\). The intensity function of the process is defined as \( \lambda(t) = \lim_{\Delta t \to 0} \frac{P(N(t,t+\Delta t]=1)}{\Delta t} \). A counting process \( N(t) \) is called a Poisson process, if and only if it satisfies the three conditions: (1) \( N(0) = 0 \); (2) the random variables \( N(a, b] \) and \( N(c, d] \) are independent, for any \( a < b \leq c < d \); and (3) for any \( a < b \), \( N(a, b] \) has the Poisson distribution with mean \( \int_a^b \lambda(x)dx \). If \( \lambda(t) \) is constant over \( t \), the process is referred to as a homogeneous Poisson process (HPP). For an HPP, \( \lambda \) is treated as the rate of occurrences.

2.3 The Conditional Test.

The problem of hypothesis testing about two Poisson means is will be addressed. The usual conditional test (C-test) and a test based on estimated p-values (E-test) are considered. The exact properties of the tests are evaluated numerically. Numerical studies indicate that the E-test is almost exact because its size seldom exceeds the nominal level, and it is more powerful than the C-test. Power calculations for both tests are outlined below.

Let \( X \) and \( Y \) be respectively independent samples, from \( \text{Poisson}(\lambda_1) \) and \( \text{Poisson}(\lambda_2) \) processes, the joint distribution of \( X \) and \( Y \):

\[
f(x, y) = \frac{\lambda_1^x e^{-\lambda_1}}{x!} \frac{\lambda_2^y e^{-\lambda_2}}{y!} \frac{\lambda_1 x \lambda_2 y}{x! y!} e^{-(\lambda_1 + \lambda_2)}
\]

Note that

\[
X + Y = S \sim \text{Poisson}(\lambda_1 + \lambda_2),
\]
\[ X = 0, 1, 2, 3 \ldots \]
\[ Y = 0, 1, 2, 3 \ldots \]

The well-known method of testing the difference between two Poisson means is the conditional test (Przyborowski and Wilenski, 1940). The conditional distribution of \( X \) given \( X + Y = S \) follows a binomial distribution whose success probability is a function of the ratio \( \frac{\lambda_1}{\lambda_2} = \rho \).

Considering the conditional distribution, \( X \) given \( S = s > 0 \), the probability function:

\[
f(x \mid S = s) = \frac{P(X=x, X + Y = s)}{P(X + Y = s)}
\]

\[
= \frac{e^{-\lambda_1} \frac{\lambda_1^x}{x!} e^{-\lambda_2} \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{s-x}}{e^{-(\lambda_1 + \lambda_2)} \left( \frac{(\lambda_1 + \lambda_2)^s}{s!} \right)}
\]

\[
= \binom{s}{x} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^x \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{s-x}
\]

\[
= \binom{s}{x} \left( \frac{1}{1 + \rho} \right)^x \left( \frac{\rho}{1 + \rho} \right)^{s-x} \sim \text{Binomial}(s, \frac{1}{1 + \rho})
\]

Let \( \frac{1}{1 + \rho} = p \), then to test the equality of two Poisson means is to test the following hypothesis:

\[
H_0: p = \frac{1}{2} \text{Vs} \ H_1: p \neq \frac{1}{2}
\]

Which is equivalent to

\[
H_0: \rho = 1 \text{ Vs } H_1: \rho \neq 1.
\]

It can be generalized as follows for comparison of leukemia deaths:

\[
H_0: p \leq p_0 \text{ Vs } H_1: p > p_0
\]

where \( 0 < p_0 < 1 \). And it is equivalent to

\[
H_0: \rho \geq \rho_0 \text{ Vs } H_1: \rho < \rho_0
\]

where \( \rho_0 = \frac{1 - p_0}{p_0} \).

The conditional test rejects \( H_0 \), when \( X = k \) is observed, whenever

\[
P\text{-value} = P(X \geq k \mid S = s) = \sum_{i=k}^{s} \binom{s}{i} p_0^i (1 - p_0)^{s-i} \leq \alpha,
\]
where $\alpha$ is the level of significance. Of course normal approximation can be implemented for the above binomial test for large number of $s$.

2.4 Conditional test for leukemia deaths.

In this thesis, I will divide the number of leukemia deaths into two main groups: female and male. For each death group, I will assume that the number of deaths follows a homogeneous Poisson process. Let $\lambda_1$ be the death rate of the male group, and $\lambda_2$ that of the female group. For the conditional test,

$$\rho_{12} = \frac{\lambda_2}{\lambda_1} \quad \text{and} \quad p_{12} = \frac{1}{1+\rho_{12}},$$

Then the hypothesis for death rates between any two groups is equal to a reference value:

$$H_0 : \rho_{12} \geq \rho_{12}^0 \quad \text{Vs} \quad H_1 : \rho_{12} < \rho_{12}^0$$

where $\rho_{12}^0$ is a known reference ratio from female and male leukemia death rates and the corresponding Binomial (Conditional) test is

$$H_0 : p_{12} \leq p_{12}^0 \quad \text{Vs} \quad H_1 : p_{12} > p_{12}^0,$$

Where $0 < p_{12}^0 < 1$ and $p_{12}^0 = \frac{1}{1+\rho_{12}^0}$.

Define the average leukemia death rates ratio from the male and female groups as a reference ratio $\rho_{12}^0$, throughout the entire observation period. That is, we wish to test whether the rate ratio of the male leukemia deaths is significantly lower than the female group. In other words, if the death rate ratio ($\rho_{12}$), is significantly higher than that of the reference value $\rho_{12}^0$, male has a higher death rate. Let the reference value, $\rho_{12}^0$ for the female death rate be 1, while $p_{12}^0$ = 0.5. The cumulated number of female death rate from 1969 to 2007 is 314,456 while that of men is 395,078. So the total number is 709,534. Based on the conditional test, $p$-value = $P(X \geq 395078|S = 709534)$

$$= \sum_{k=395078}^{709534} \left( \binom{709534}{k} (0.5)^k (1-0.5)^{709534-k} \right) \approx 0$$
The null hypothesis is rejected, that is, males are more likely to die from leukemia than female. A 95% one-sided confidence interval for $p_{12}$ is $[0.5562236078, 1]$.

2.5 Empirical Recurrence Rates.

A time series empirical recurrence rates are developed in order to monitor the deaths rates of the individual groups that is male and female.

Let $t_1, \ldots, t_n$ be the time of the $n$-ordered leukemia deaths during an observation period $(t_0, 0)$, where $t_0$ is the time-origin and 0 is the present time. If $h$ is the time-step, then a discrete time series $\{z_t\}$ is generated sequentially at equidistant time intervals $t_0 + h, t_0 + 2h, \ldots, t_0 + \ell h, \ldots, t_0 + Nh = 0 = \text{present time})$. $z_t$ is regarded as the observation at time $t \ (= t_0 + \ell h)$, for the leukemia deaths to be modeled. A key parameter desired by the modelers is the recurrence rate of the targeted leukemia deaths data. Therefore, a time series of the empirical recurrence rates (Ho, 2008) is generated as follows:

$$z_l = \frac{n_l}{lh} = \frac{\text{total number of leukemia deaths in } (t_0, t_0 + lh)}{lh}$$

where $\ell = 1, 2 \ldots N$. Note that $z_t$ evolves over time and is simply the maximum likelihood estimator (MLE) of the mean, if the underlying process observed in $(t_0, t_0 + \ell h)$ is a homogeneous Poisson process. The time-plot of the empirical recurrence rate (ERR-plot), offers the possibility of further insights into the data. ERR plots for male and female leukemia deaths within the study period with time-step $h = 1$ year. If we start at time $T$, the value $z_{T+k}$, $k \geq 1$ needs to be predicted based on the sample observation $(z_1, \ldots, z_T)$ of an ERR time series. In a regression modeling, let $X$ denote the time index, $z$ be the response values, and then use the fitted regression model to obtain $z_{T+k}$ .ERR plots for male and female leukemia deaths within the study period with time-step $h = 1$ year are shown in Figure 2.2. It is clear that the death rate for male and
female are rising approximately at the same rate. To enable us compare the leukemia death rates ratio between men and women we introduce empirical recurrence rates ratio chapter 3.

Figure 2.2 ERR plots for male and female leukemia deaths within the study period with time-step $h = 1$ year.
CHAPTER 3
THEORY AND METHOD FOR ARIMA MODELS

3.1 Empirical Recurrence Rates Ratio

We produce an empirical recurrence rates ratio time series for the leukemia deaths rates ratio as follows: The C-test examines the relationship of two means of homogenous Poisson processes, which have constant expected values. Motivated by the ideas of the C-test and the empirical recurrence rate developed by Ho (2008), the empirical recurrence rates ratio time series for the leukemia deaths rates ratio is produced as follows:

Let \( t_1, t_2, \ldots, t_n \) be the time of the n-ordered leukemia deaths during an observation period \( (t_0, t_0+Nh) \) from the past to the present. Then a discrete time series \( \{d_l\} \) is generated sequentially as \( t_0 + h, t_0 + 2h, \ldots, t_0 + lh, \ldots, t_0 +Nh \) (= the present time). \( h \) represents the time step. Let \( X_{ij} \) be the number of leukemia deaths in \( i^{th} \) group at \( j^{th} \) lag, where \( i = 1, 2 \) and \( j = 1, 2, \ldots, N \); and the Empirical Recurrence Rates Ratio (ERRR) is defined as follows:

\[
d_l = \frac{\sum_{j=1}^{l} X_{1j}}{\sum_{j=1}^{l}(X_{1j}+X_{2j})}, \quad l = 1, 2, \ldots, N.
\]

Both the ERR and EERRR offer the possibility of developing a model, monitoring and predicting leukemia death rate ratios. Moreover, if both of the targeted processes are homogeneous Poisson processes, then the EERRR is the maximum likelihood estimator (MLE) of \( p \), and the MLE of \( p \) can be obtained by the invariance property of the MLE.
3.2 ARIMA Models

Since the 1970s, primarily due to the work of Box and Jenkins (1976), a class of mixed autoregressive (AR) and moving average (MA) models originally proposed by Yule (1927) and Slutsky (1937), have been useful in representing the serial dependent relationship of many time series encountered in practice. Autoregressive integrated moving average (ARIMA) models allow us not only to uncover the hidden patterns in the data, but also to generate forecasts and predict a variable’s future values from its past values.

A branch of the ARIMA model known as the autoregression refers to a special kind of regression analysis aimed at analysis of time series. It rests on autoregressive models – that is, models where the dependent variable is the current value and the independent variable is previous p-values of the time series. The p is called “the order of the autoregression”.

The moving average (MA) model is another form of ARIMA model in which the time series is described as a linear function of its prior errors plus a noise term.

Given a time series of data \( x_t \) the ARMA model is a tool for understanding and perhaps predicting future value in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually referred to as the ARMA (p,q) model where p is the order of the autoregressive part and q is the order of the moving average part.

3.2.1 Autoregressive model of order p, AR(p)

An autoregressive model of order p is of the form

\[
x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t.
\]

Where \( x_t \) is stationary, \( \phi_1, \phi_2, \ldots, \phi_p \) are constants (\( \phi_p \neq 0 \)) and \( w_t \) is a Gaussian white noise series with mean zero and variance \( \sigma_w^2 \). The mean of \( x_t \) is zero. If the mean, \( \mu \), of \( x_t \) is not zero, replace \( x_t \) by \( x_t - \mu \); that is

\[
x_t - \mu = \phi_1 (x_{t-1} - \mu) + \phi_2 (x_{t-2} - \mu) + \cdots + \phi_p (x_{t-p} - \mu) + w_t
\]
The autoregressive operator is defined to be $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$.

### 3.2.2 Moving average model of order q, MA(q)

The moving average model of order q is defined to be

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

Where there are q lags in the moving average and $\theta_1, \theta_2, \ldots, \theta_q$ ($\theta_q \neq 0$) are parameters. The noise $w_t$ is assumed to be Gaussian white noise. The moving average operator is

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

### 3.2.3 Autoregressive Moving average model of order p, q, ARMA(p, q)

A sequence $\{w_t\}$, of uncorrelated random variables, each with zero mean and variance $\sigma^2$, is referred to as white noise. This is indicated by the notation

$\{w_t\} \sim WN(0, \sigma^2)$.

The general ARMA models are a combination of the AR operators and MA operators.

A time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ is ARMA if it is stationary and

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q},$$

where $\phi_p \neq 0, \theta_q \neq 0$. The parameter p and q are called the autoregressive and the moving average orders, respectively.

The following are the problems for ARMA(p, q):

1. **Parameter redundant models**: A model is parameter redundant if it can be reparameterized in terms of a smaller number of parameters than the size of its defining parameter set, so that using classical inference it would not be possible to estimate all
the original parameters. One approach to removing parameter redundancy is to include covariates in a model, that set parameters to be appropriate functions of covariates.

(2) Stationary AR models that depend on the future: To overcome this problem of future-dependent model, we formally introduce the concept of causality. An ARMA (p,q) model is causal if and only if $\varphi(z) \neq 0$ for $|z| \leq 1$.

(3) MA models that are not unique: To address the problem of uniqueness we choose the model that allows an infinite autoregressive representation.

The introduction of correlation as a phenomenon that may be generated through lagged linear relations leads to proposing the autoregressive (AR) and autoregressive moving average (ARMA) models. Adding nonstationary models to the mix leads to the autoregressive integrated moving average (ARIMA) models popularized in the landmark work by Box and Jenkins (1970).

3.2.4 Stationary Time Series

A stationary process is a stochastic process whose joint probability distribution does not change when shifted in time or space. As a result, parameters such as the mean and variance, if they exist, also do not change over time or position. A weak stationary time series, $x_t$, is a finite variance process such that

(i) the mean value function, $u_t$ is constant and does not depend on time $t$, and

(ii) the covariance function, $\gamma(s, t)$ depends on $s$ and $t$ only through their difference $|s - t|$.

Stationarity is used as a tool in time series analysis, where the raw data are often
transformed to become stationary; most data are often seasonal and/or dependent and are therefore nonstationary.

Although the theoretical autocorrelation functions are useful for describing the properties of the data, most of the analysis must be performed using sampled points $x_1, x_2, \ldots x_n$ that are available for estimating the mean, autocovariance, and autocorrelation functions. From the point of view of classical statistics, this poses a problem because we will typically not have iid copies of $x_t$ that are available for estimating the covariance and correlation functions. In the usual situation of only one realization, however, the assumption of stationarity becomes critical.

3.3 Data Transformation

In statistics, data transformation refers to the application of a deterministic mathematical function to each point in a data set that is, each data point $z_i$ is replaced with the transformed value $y_i = f(z_i)$, where $f$ is a function. Transformations are applied so that the data appear to more closely meet the assumptions of a statistical inference procedure that is to be applied or to improve the interpretability or appearance of graphs.

Nearly always, the function that is to be used to transform the data is invertible and generally is continuous. The transformation is usually applied to a collection of comparable measurements. We will introduce three common transformations that are called Box-Cox, differencing and subtracting the mean as follows.

3.3.1 Box-Cox Transformation

In statistics, the power transform is from a family of functions that are applied to create a rank-preserving transformation of data using power functions. This is a useful data processing technique used to stabilize variance, make the data more normal distribution-like, improve the
correlation between variables and other data stabilization procedures. The Box–Cox
transformation, by statisticians George E.P. Box and David Cox, is one particular way of
parameterising a power transform that has advantageous properties.

If the original observations are \(Y_1, Y_2, Y_3, \ldots, Y_n\), the Box-Cox transformation \(f_\lambda\) converts them to \(f_\lambda(Y_1), f_\lambda(Y_2), \ldots, f_\lambda(Y_n)\), where:

\[
f_\lambda(y) = \begin{cases} 
\frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\
\log(y), & \lambda = 0
\end{cases}
\]

An extended form which could accommodate negative \(y_s\)

\[y(\lambda) = \begin{cases} 
\frac{(y + \lambda_2)^{\lambda_1 - 1}}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\
\log(y + \lambda_2) & \text{if } \lambda_1 = 0.
\end{cases}
\]

Here, \(\lambda = (\lambda_1, \lambda_2)^T\). In practice we could choose \(\lambda_2\) such that \(y + \lambda_2 > 0\) for any \(y\). So, researchers
could only view \(\lambda_1\) as the model parameter. This transformation is useful when the variability of
the data increases or decreases with the level. By suitable choice of \(\lambda\), the variability can be
made nearly constant. For instance, positive data whose standard deviation increases linearly
with level, the variability can be stabilized by choosing \(\lambda = 0\) (Brockwell et al., 2002).

### 3.3.2 Differencing

In the case that the time series data at hand has a trend in it, we should first difference the
data to remove the trend and then consider the autocorrelation function for the differenced data
for signs of seasonality at the seasonal lags. Differencing is an important technique to transform
data, to control autocorrelation, and to achieve stationary time series. The first difference is
denoted as:

\[\nabla X_t = X_t - X_{t-1} = (1 - B)X_t\]
where $B$ is the backshift operator. We may extend the notion further and define the differences of order $d$ as:

$$\nabla^d X_t = (1 - B)^d X_t$$

Usually, single differencing is used to remove linear trends and double differencing is used to remove quadratic trend. We can eliminate seasonality and trend of period $d$ by introducing the lag $d$ difference operator $\nabla_d$:

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t.$$ 

This operator should not be confused with the operator $(1 - B)^d$ (Ho, 2010a). Normally, the correct amount of differencing is the lowest order of differencing that yields a time series which fluctuates around a well-defined mean value and whose autocorrelation function (ACF) plot decays rapidly to zero, either from above or below. Thus, at every stage of differencing, we check the plots of sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) to see where the ACF/PACF “cuts off” the bounds $\pm 1.96 / \sqrt{n}$.

A time plot of the data will typically suggest whether any differencing is needed after the first differencing. However, over differencing may introduce dependence where none exist. In addition to the time plot, the sample ACF can help in indicating whether differencing is needed. The sample ACF will not decay to zero as fast as $h$ increases. Thus a slow decay is an indication that differencing may be needed.

It is desirable to find a sample ACF that decays fairly rapidly. We say that a series is stationary if the sample ACF has very few significant spikes at very small lags and then cuts off drastically or dies down very quickly. If the samples ACF decay slowly, the series still has some trend. If the ACF has periodicity, the series has seasonality. If this occurs we should do some more differencing of the data before continuing. The Behavior of the ACF and PACF for ARMA models are summarized in table 3.1 (Shumway and Stoffer, 2006).
Table 3.1 Behavior of the ACF and PACF for ARMA models.

<table>
<thead>
<tr>
<th></th>
<th>AR(p)</th>
<th>MA(q)</th>
<th>ARMA(p, q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Tails off</td>
<td>Cuts off after lag q</td>
<td>Tails off</td>
</tr>
<tr>
<td>PACF</td>
<td>Cuts off after lag p</td>
<td>Tails off</td>
<td>Tails off</td>
</tr>
</tbody>
</table>

3.3.3 Subtracting the Mean

The term, ARMA model, is used in the program ITSM2000 (Brockwell et al., 2002) to denote a zero-mean ARMA process. Therefore, the sample mean of the data should be small before modeling. Once the apparent deviations from stationary of the data have been removed, the sample mean of the transformed data should be subtracted from each observation. The search for a fitted ARMA model for a mean-corrected data set then follows.

3.4 Model Diagnostics

Model diagnostics is understood as a more or less formal check of properties that certain residuals should have under certain assumptions that the data were generated by the model which is under investigation. In this thesis we will check the residual ACF/PACF of the models that we develop. Also, the models need to pass the test for randomness of the residuals. After the model diagnostics process, further predictions and comparisons can be done.

3.4.1 The Sample ACF/PACF of the Residuals

The residuals autocorrelation function is the basic model checking tool in time series analysis, but it is useless when its distribution is incorrectly approximated because of parameter estimation or because an unnoticed higher serial dependence have not been taken into account.

The sample autocorrelations of an independent and identically distributed (iid)
sequence $y_1, y_2, \ldots, y_n$ are approximately iid with distribution $N(0, \frac{1}{n})$. We can therefore test whether or not the observed residuals are consistent with iid noise by examining the sample correlations of the residuals and rejecting the iid noise hypothesis if more than two or three out of 40 fall outside the bounds $\pm 1.96 \sqrt{n}$ or if one falls far outside the bounds (Brockwell et al, 2002).

### 3.4.2 Tests for Randomness of the Residuals

A popular test, formulated by Ljung and Box (1978), called the Ljung-Box Test, is commonly used to check whether the residuals of a fitted model are observed values of independent and identically distributed random variables in ARIMA modeling. It is referred to as a portmanteau test, since it is based on the autocorrelation plot and tests the overall independence based on a few lags. Then, the definition of Ljung-Box test is as follows:

- $H_0$: The sequence data are iid
- $H_a$: The sequence data are not iid

And use the test statistic as:

$$
\hat{Q}(\hat{r}) = n(n + 2) \sum_{k=1}^{m} (n - k)^{-1} \hat{r}_k^2,
$$

where $\hat{r}_k = \frac{\sum_{i=k+1}^{n} \hat{a}_i \hat{a}_{i-k}}{\sum_{i=1}^{n} \hat{a}_i^2}$, the estimated autocorrelation at lag $k$,

$n = \text{sample size},$

$m = \text{number of lags being tested}$ (As a rule of thumb, the sample ACF and PACF are good estimates of the ACF and PACF of a stationary process for lags up to about a third of the sample size (Brockwell and Davis, 2002) where $\hat{a}_1, \ldots, \hat{a}_n$ are the residuals after a model has been fitted to a series $z_1, \ldots, z_n$. If no model is being fitted, then $\hat{a}_1, \ldots, \hat{a}_n$ are the “mean corrected” series of $z_1, \ldots, z_n$.}
If the sample size \( n \) is large, the distribution of \( \hat{Q}(\hat{r}) \) is roughly \( \chi^2_{m-p-q} \) under the null hypothesis, where \( m - p - q \) is the degree freedom of Chi-square distribution, and \( p + q \) is the number of parameters of the fitted model. The null hypothesis will be rejected, if \( \hat{Q} > \chi^2_{1-\alpha;m-p-q} \) at level \( \alpha \). Thus, the sequence data are not independent, or their autocorrelations are significantly different from zero.

3.4.3 AIC, BIC and AICC Statistics

We develop a small sample criterion (AICC) for the selection of the order of vector autoregressive model. AICC is an approximate unbiased estimator of the Kullback-Lieber information. Furthermore, AICC provides better model order choices than the Akaike information criterion (AIC) in small sample, but it should be used as a rough guide. The final decision is largely based on maximum likelihood estimation. Some other Model selection statistics, such as the BIC statistic, are available in ITSM 2000. The BIC statistic (Schwarz, 1978) is a Bayesian modification of the AIC statistic. The BIC statistics evaluated at the same time as the AICC, and it is used in the same way as the AICC. Each information statistic is defined as follows:

\[
AIC_{p,q} = N \log \hat{\sigma}_e^2 + 2r
\]

\[
AICC_{p,q} = N \log \hat{\sigma}_e^2 + 2rN / (N - r - 1)
\]

\[
BIC_{p,q} = N \log \hat{\sigma}_e^2 + r \log N
\]

Where \( \hat{\sigma}_e^2 \) is the error variance, the error variance in this case is defined as

\[
\hat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]
One may point out from probability theory, that $\hat{\sigma}^2_{\epsilon}$ is a biased estimator for the true variance, $\sigma^2$, and $r = p + q + 1$ is the number of parameters estimated in the model, including a constant term. The second term in all three equations is a consequence for increasing $r$. Hence, if we want to minimize the values of these criteria, we should minimize the number of parameters. Therefore, the best model is the model that adequately describes data and has the fewest parameters.

3.5 Forecasting

This thesis outlines the practical steps which need to be undertaken to use autoregressive integrated moving average (ARIMA) time series models for forecasting death rates of male and female. The emphasis is on forecast performance which suggests more focus on minimizing death rates forecast errors than on maximizing in-sample “goodness of fit.” Practical issues in ARIMA time series forecasting are illustrated. The candidate ARIMA models will be used to predict future values of the time series from the past values. The forecasting function $z_t = f(z_{t-1}, ..., z_1) + a_t$ has the minimum mean square error. The first part of the above equation $f(z_{t-1}, ..., z_1)$ is a function of the past values of the series and it should be determined by the data. The second part $a_t$, called noise part, is a sequence of iid variables.

Predictions will be achieved by forecasting the residuals and then inverting the transformations adopted to arrive at forecasts of the original series. Also, we will observe which model is the best fitting model by comparing the prediction from the training set with the prediction set. Then, I will combine the training sample and the prediction set as a full data set to forecast death rates ratio for the predicted set, based on the same techniques as before.
CHAPTER 4

ANNUAL LEUKEMIA RELATED DEATHS DATA ANALYSIS

4.1 ERRR-plots

Since there are 709,534 Leukemia related deaths in the 39 years of study, which indicates there is approximately 18,194 Leukemia related deaths in every year. We choose \( h = 1 \) year as the time-step and we will try to predict leukemia-related deaths with \( h = 1 \) year. Figure 4.1 shows ERRR plots with time-step \( h=1 \) year which show a continuous decline from lag 1 to lag 33 and then rises a little from lag 34 to lag 39.

![ERRR plots with time-steps h=1 year.](image)

**Figure 4.1** ERRR plots with time-steps \( h=1 \) year.

4.2 Data Splitting.

In some cases, researchers might want to separate several time series contained in one data set into different data sets: training sample and prediction set. Training sample is used to develop a model for prediction. Prediction set is used to evaluate the reasonableness and predictive ability of the selected model (one round of cross validation).

Cross–validation, sometimes called rotation estimation, is a technique for assessing how the results a statistical analysis will generalize to an independent data set. It is mainly used in
settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. Multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. The application in this regard will be detailed in Section 4.3 and 4.4.

4.3 ARIMA Modeling with h = 1year.

We use the ITSM2000 software to model the ERRR data. The data set with time-step h = 1year has 39 lags in total. At first, we use the technique described in Section 4.2 to split the data into two sets: training sample and prediction set. In this case, our training sample is the original data set excluding the last 3 ERRRs, which is the prediction set (Figure 4.2).

Figure 4.2 ERRR plots of the Training Sample and prediction set with h = 1year.

These three ERRR values in the prediction set, representing the number of leukemia-related deaths in three years, will be used to compare to those of the one to three-step predictions produced by a candidate model. Of course, the size of a prediction set is quite flexible as long as the prediction set fits a common goal of model selection. Then, we focus on the training sample set and plot the sample ACF and PACF to observe the data set (Figure 4.3). From the plot of sample ACF, we find that the spikes die slowly and have periodicity. This indicates non-
stationary behavior. As mentioned in Section 2.4, this data has trend and seasonality. Thus, differencing is considered.
4.3.1 Training Sample modeling

Figure 4.3a, Time-plot; b, Sample ACF; c, Sample PACF of the Training Sample with h = 1 year.

Applying the differencing operator $\nabla$ on the training sample, we take a difference at lag 2. Figure 4.4 tells us that the stationarity has almost been achieved. So we do further difference at lag 1
Figure 4.4 a, Time-plot; b, Sample ACF; c, Sample PACF of a lag-1 differenced Training Sample with \( h = 1 \) year.

Then we subtract the sample mean from each observation of the differenced series to generate a stationary zero-mean time series (Figure 4.5)
Figure 4.5  
(a) Time-plot;  
(b) Sample ACF;  
(c) Sample PACF of the twice-differenced training sample with h = 1 year.
We feel that the ACF and the PACF is tailing off. These suggest that an MA (2) should be considered. Indeed, our initial model selection process concludes that the estimated model is:

**ARMA Model:**

\[ X(t) = Z(t) + 0.08686 Z(t-1) - 0.5965 Z(t-2) \]

WN Variance = 0.000001

MA Coefficients

0.086863      -0.596466

Standard Error of MA Coefficients

0.139721       0.139721

(Residual SS)/N = 0.00000107669

AICC = 352068E+03
BIC  = 355605E+03

-2Log(Likelihood) = 358896E+03

Note that \( X_t \) represents a twice-differenced stationary zero-mean time series and the error term \( Z_t \) represents a white noise process.

A set of diagnostic plots (Figure 4.6) is produced by the ITSM2000 package, consisting of the plot of the residuals, its ACF and its PACF for the MA (2) model in which all the spikes lies within the boundary line. The AICC statistic is 352068E+03 and the Ljung-Box test is not significant (p-value = .88320), indicating that the residuals are white noise. The numerical values of the actual ERRRs in the prediction set and the predicted ERRRs by the model MA (2) with their counterparts are shown in Table 4.1.
Figure 4.6 Diagnostics for the MA (2) fitted and twice-differenced Training Sample.

Residual a, Time-plot; b, Sample ACF; c, Sample PACF.
Table 4.1: The numerical values of the actual ERRRs in the prediction set and the predicted ERRRs and their confidence intervals using the MA (2) based on the training sample.

<table>
<thead>
<tr>
<th>Year</th>
<th>Annual ERRR</th>
<th>Prediction ERRR</th>
<th>Lower Bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>0.55623</td>
<td>0.55624</td>
<td>0.55460</td>
<td>0.55789</td>
</tr>
<tr>
<td>2006</td>
<td>0.55649</td>
<td>0.55664</td>
<td>0.55421</td>
<td>0.55907</td>
</tr>
<tr>
<td>2007</td>
<td>0.55681</td>
<td>0.55712</td>
<td>0.55369</td>
<td>0.556057</td>
</tr>
</tbody>
</table>

We list the ratios of (estimated coefficients)/(1.96×standard error) for each coefficient, calculated from the output of an MA (2) model, shown in Section 3.2. The ratios are:

MA Coefficients

0.086868  -0.596464

Standard Error of MA Coefficients

0.139722  0.139722

Note that the ratio at lag1 of MA(2) in absolute value is less than 1, which indicates the corresponding coefficient is nonzero. We keep the corresponding coefficient.
Table 4.2 The numerical values of the predicted ERRRs and their confidence intervals using the MA (2) based on the full data set.

<table>
<thead>
<tr>
<th>Year</th>
<th>Annual ERRR</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prediction</td>
<td>Lower Bound</td>
</tr>
<tr>
<td>2008</td>
<td>0.55730</td>
<td>0.55567</td>
</tr>
<tr>
<td>2009</td>
<td>0.55786</td>
<td>0.555546</td>
</tr>
<tr>
<td>2010</td>
<td>0.55850</td>
<td>0.55511</td>
</tr>
</tbody>
</table>

Table 4.2 shows numerical values of the predicted ERRRs and their confidence intervals using the MA (2) whilst Figure 4.7 depicts the confidence intervals for the predicted values based on the full data set.

![ERRR plot with Prediction intervals.](image)

Figure 4.7 ERRR plot with Prediction intervals.

Comparisons of the results with the prediction set model are defined in Table 4.3. The predicted
values are very similar, indicating that this model is acceptable. Figure 4.8 shows a comparison of three forecasted ERRRs with the prediction set which appears to be moving in the same direction from lag 1 to lag 3 for both the predicted and the actual ERRR values.

Figure 4.8 Comparison of three forecasted ERRRs with the prediction set.

Figure 4.9 depicts the complete Data (training sample and prediction set) with three predicted values appended to the training Sample for model validation; Inset: Comparison of three ERRRs with Prediction set

Figure 4.9 The complete Data (Training Sample and Prediction set) with three appended to training Sample for model validation; Inset: Comparison of three ERRRs with prediction set
4.3.2 Full-Data Forecasting

Finally, we use the full ERRR time series to forecast the probable number of leukemia related deaths in the future. This yields the best-fitted MA (2) model for the mean-corrected and twice-differenced value at lag 1 (same as before). The estimated MLE:

<table>
<thead>
<tr>
<th>ARMA Model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(t) = Z(t) + 0.08003 Z(t-1) - 0.6162 Z(t-2)$</td>
</tr>
<tr>
<td>WN Variance = $0.982271E-06$</td>
</tr>
<tr>
<td>MA Coefficients</td>
</tr>
<tr>
<td>0.080027</td>
</tr>
<tr>
<td>Standard Error of MA Coefficients</td>
</tr>
<tr>
<td>0.209105</td>
</tr>
<tr>
<td>(Residual SS)/N = $0.982271E-06$</td>
</tr>
</tbody>
</table>

The AICC statistic is $-0.388089E+03$, and the Ljung-Box test is significant (p-value = 0.80782).

Then, we check the ratios as follows:

<table>
<thead>
<tr>
<th>MA Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.080027</td>
</tr>
<tr>
<td>Standard Error of MA Coefficients</td>
</tr>
<tr>
<td>0.209105</td>
</tr>
</tbody>
</table>
4.3.2 ARIMA Models

The training samples with 36 lags are shown in Figure 4.2 above. The plots of sample ACF and PACF on the training sample (Figure 4.3) indicate non stationary behavior. No differencing is considered. This is also a suggestion of the AR(2) model. The estimated (MLE) model is:

ARMA Model:

\[
X(t) = 0.7194 X(t-1) + 0.2658 X(t-2) + Z(t)
\]

WN Variance = 0.000004

AR Coefficients

0.719426 0.265804

Standard Error of AR Coefficients

0.366896 0.367090

(Residual SS)/N = 0.0000377443

AICC = -0.337258E+03

BIC = -0.333769E+03

The AICC statistic is -0.337258E+03 The Ljung-Box statistic is 5.3557 and the p-value is 0.99934, which indicates that the residuals are approximately white noise.
Figure 4.10a, ERRR plots after Box-Cox transformation at $\lambda = 0$; b, sample ACF; c, Sample PACF of the full data with $h = 1$ year.

The plots of the training sample (including 36 lags) and its sample ACF and PACF in (Figure
4.10b, c) show nonstationarity and periodicity since some of the spikes extend beyond the required boundaries from lag 0 to lag7 and from lag 17 to lag 26 in the case of the ACF and at lag 0 in the case of the PACF. Therefore, the Box-Cox transformation will be employed to remove the trend and seasonality. Since the plot shows decreasing variability, we consider the Box-Cox transformation to stabilize the variability. After the $\lambda=0$ Box-Cox transformation. The actual and the predicted value based on the Training Sample using AR(2) are shown in Table 4.3.

**Table 4.3** Numerical of the Actual and the Predicted based on the Training Sample using AR(2)

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55623</td>
<td>0.55595</td>
<td>0.55413</td>
<td>0.55778</td>
</tr>
<tr>
<td>0.55649</td>
<td>0.55602</td>
<td>0.55377</td>
<td>0.55828</td>
</tr>
<tr>
<td>0.55681</td>
<td>0.55608</td>
<td>0.55341</td>
<td>0.55878</td>
</tr>
</tbody>
</table>

A plot of the ERRR values and their prediction intervals are shown in Figure 4.11 in which the predicted values seems to be leveling off from lag 36 to lag 39.

![Figure 4.11 ERRR plot with prediction intervals using AR(2)](image_url)
Another model to be considered based on the training sample is ARMA(1,1). Figure 4.12 shows a, an ERRR plot after first differencing at lag 1; b, sample ACF; c, Sample PACF based on the training sample with h = 1year, its ACF and PACF indicates a stationarity behavior.

Figure 4.12a, ERRR plot after first differencing at lag 1; b, Sample ACF; c, Sample PACF based on the training Sample with h = 1year.

Figure 4.13 depicts a, ERRR plot after twice-differencing at lag 1; b, Sample ACF; b, Sample PACF.
PACF. The MLE is as shown below:

ARMA Model:

\[
X(t) = -0.9016 X(t-1) + Z(t) + 0.9999 Z(t-1)
\]

WN Variance = 0.000001

AR Coefficients
-0.901582

Standard Error of AR Coefficients
0.074359

MA Coefficients
0.999883

Standard Error of MA Coefficients
0.002626

(Residual SS)/N = 0.00000103425

AICC = -364283E+03
BIC = -371031E+03

The AICC statistic is -364283E+03 The Ljung-Box statistic is 6.9634 and the p-value = 0.99680, which indicates that the residuals are approximately white noise.
The plots of the training sample and its sample ACF and PACF in (Figure 4.10) show nonstationarity and periodicity. Therefore, the Box-Cox transformation, and differencing will be
employed to remove the trend and seasonality. Since the plot shows decreasing variability, we consider the Box-Cox transformation to stabilize the variability. After the $\lambda=1$ Box-Cox transformation, we see that the trend still exists. We then take the differencing twice at lag 1. Figures 4.12 and 4.13 tell us the series has reached stationarity. ARMA(1, 1) is then considered as a fitting model for the training sample. The Actual and the Predicted ERRRs with their confidence intervals based on the training sample using ARMA(1, 1) are shown in table 4.4. Figure 4.14 depicts ERRR plots with prediction intervals using ARMA(1, 1) with the predicted values rising from lag 36 to lag 39. Figure 4.15 shows a, Rescaled Residual-plots; b, Residual ACF; c, Residual PACF using ARMA(1, 1), this tells us that stationarity has been achieved.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55623</td>
<td>0.55623</td>
<td>0.55495</td>
<td>0.55788</td>
</tr>
<tr>
<td>0.55649</td>
<td>0.55673</td>
<td>0.55291</td>
<td>0.55055</td>
</tr>
<tr>
<td>0.55681</td>
<td>0.55763</td>
<td>0.55098</td>
<td>0.56375</td>
</tr>
</tbody>
</table>

Figure 4.14 depicts the ERRR values, the predicted values and their confidence intervals using ARMA(1, 1) based on the training sample.

**Figure 4.14** ERRR plots with prediction intervals using ARMA(1, 1).
Figure 4.15a, Rescaled Residual-plots; b, Residual ACF; c, Residual PACF with using ARMA(1, 1).

The Actual ERRR values and the three models predictions, that is MA(2), ARMA(1,1), and AR(2) are then plotted and compared to find which of the three predictions is closer to the actual ERRR values. Table 4.5 shows the actual values and the models predicted values and figure 4.16 displays the actual ERRR values and the predicted values by the three models. There is an upward trend from lag 1 to lag 3 with the MA(2) prediction much closer to the actual
4.3.3 More ARIMA Models

We extend the same techniques from the training sample to the full data to confirm our results. The data set with the time-step $h = 1$ years has 39 lags. The training sample with 36 lags and the prediction set with 3 lags are shown in Figure 4.2 above. The plots of sample ACF and PACF on the training sample (Figure 4.3) indicate nonstationary behavior. Thus no differencing is considered. This is also a suggestion of the AR (2) model. The estimated (MLE) model is:

Table 4.5 Actual and Model predicted values for MA (2), ARMA (1, 1) and AR (2)

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>MA(2)</th>
<th>ARMA(1,1)</th>
<th>AR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55623</td>
<td>0.55624</td>
<td>0.55623</td>
<td>0.55595</td>
<td></td>
</tr>
<tr>
<td>0.55649</td>
<td>0.55664</td>
<td>0.55673</td>
<td>0.55602</td>
<td></td>
</tr>
<tr>
<td>0.55681</td>
<td>0.55712</td>
<td>0.55763</td>
<td>0.55608</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.16 Comparison of the models with the actual values based on the training sample.
The AICC statistic is -0.368929E+03. The Ljung-Box statistic is 5.3557 and the p-value is 0.9953, which indicates that the residuals are approximately white noise. Figure 4.17 shows a, ERRR plots after Box-Cox transformation at \( \lambda = 0 \); b, sample ACF; c, sample PACF of the full data with \( h = 1 \text{year} \). This indicates nonstationary in its ACF and PACF as some of the spikes falls outside its boundaries. Figure 4.18 shows ERRR plot with prediction intervals using AR(2), the prediction values is leveling off from lag 36 to lag 39. Table 4.6 displays the numerical values of the predicted ERRRs with their confidence intervals using AR(2). Figure 4.19 is a, residual-plot; b, residual ACF; c, residual PACF of the full data with \( h = 1 \text{year} \). This figure indicates stationarity.
Figure 4.17  a, ERRR plots after Box-Cox transformation at $\lambda = 0$; b, Sample ACF; c, Sample PACF of the full data with h= 1year.
Table 4.6 The numerical values of the Predicted ERRRs with their confidence intervals using AR(2).

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55677</td>
<td>0.55494</td>
<td>0.55861</td>
</tr>
<tr>
<td>0.55683</td>
<td>0.55460</td>
<td>0.55908</td>
</tr>
<tr>
<td>0.55687</td>
<td>0.55422</td>
<td>0.55953</td>
</tr>
</tbody>
</table>

Figure 4.18 ERRR plot with Prediction intervals using AR(2)
Figure 4.19  (a), Residual-plot;  (b), Residual ACF;  (c), Residual PACF of the full data with h = 1 year.
Twice-differencing the full data set at lag 1, the AICC statistic is \(-399986E+03\), the Ljung - Box statistic is 6.9192 and the p-value is 0.99694, which indicates that the residuals are approximately white noise. This is also a suggestion of the ARMA (1, 1) model. Figure 4.20 shows a, ERRR plots after differencing at lag 1; b, sample ACF; c, sample PACF of the full data with \(h = 1\)year, while Figure 4.21 is a, ERRR plots after twice-differencing at lag 1; b, Sample ACF; c, sample PACF of the full data with \(h = 1\)year. Figure 4.22 is an ERRR plot with prediction intervals Using ARMA(1, 1). There is an upward trend from lag 36 to lag 39. Table 7 shows the numerical values of the predicted ERRR with their confidence intervals using ARMA(1,1). Figure 4.23 depicts a, residual-plot; b, residual ACF; c, residual PACF of the full data with \(h = 1\)year. The estimated (MLE) model is:

**ARMA Model**

\[
X(t) = -0.6858 \ X(t-1) \\
+ Z(t) + 0.6121 \ Z(t-1)
\]

WN Variance = \(.985063E-06\)

AR Coefficients
-0.685763

Standard Error of AR Coefficients
0.745346

MA Coefficients
0.612102

Standard Error of MA Coefficients
0.749744

\((\text{Residual SS})/N = .985063E-06\)

AICC = \(-399986E+03\)

BIC = \(-410756E+03\)
Figure 4.20 a, ERRR plots after differencing at lag 1; b, Sample ACF; c, Sample PACF of the full data with $h=1$ year.
Figure 4.21 a, ERRR plots after twice-differencing at lag 1; b, Sample ACF; c, Sample PACF of the full data with h = 1 year.
Table 4.7 The numerical values of the predicted ERRRs with their confidence intervals using ARMA(1, 1).

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55729</td>
<td>0.55566</td>
<td>0.55892</td>
</tr>
<tr>
<td>0.55792</td>
<td>0.55437</td>
<td>0.56146</td>
</tr>
<tr>
<td>0.55870</td>
<td>0.55278</td>
<td>0.56461</td>
</tr>
</tbody>
</table>

Figure 4.22 ERRR plot with prediction intervals Using ARMA(1, 1)
Figure 4.23a, Residual-plot; b, Residual ACF; c, Residual PACF of the full data with h = 1 year.
Figure 4.24 depicts the temporal trends. All the results point to the same directions: male are more likely to die from leukemia than their female counterparts confirming the results of our finding based on the training sample as before. Table 4.8 shows predicted values of the three models based on the full data.

<table>
<thead>
<tr>
<th>Year</th>
<th>MA(2)</th>
<th>ARMA(1,1)</th>
<th>AR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>0.55730</td>
<td>0.55729</td>
<td>0.55677</td>
</tr>
<tr>
<td>2009</td>
<td>0.55786</td>
<td>0.55792</td>
<td>0.55683</td>
</tr>
<tr>
<td>2010</td>
<td>0.55850</td>
<td>0.55870</td>
<td>0.55687</td>
</tr>
</tbody>
</table>

Figure 4.24 Comparison of the three models based on the full model
CHAPTER 5
CONCLUSIONS

Coupled with the conditional test (Przyborowski and Wilenski, 1940), the empirical recurrence rates ratio extended from the empirical recurrence rate (Ho, 2008), which allows us to apply the well-known ARIMA modeling techniques to compare and forecast leukemia related death rates ratio in the United States of America based on the 39 years mortality data. The ERR and ERRR not only smooth and explain deaths rates modeled by a stochastic process, but also operate as a link between a classical time series and a point process.

We split the leukemia ERRR time series into a training sample and a prediction set. The training sample is used to develop the candidate models. For time-step $h = 1$ year, we used the last three ERRRs as a prediction set to make model comparisons by checking the predictive ability of the candidate models developed from the training sample. Before modeling, we must make sure the ARMA process is stationary. After taking twice difference at lag 1, an MA (2) model yields predictions that are the closest to the actual values, therefore we conclude that MA(2) is the best of the three resulting models.

The limitation to this paper is the fact that the data used in the write up has a present value of 2007, instead of a more current value of 2011. In addition we could not use the empirical recursive rates (ERR) values to predict future counts of the leukemia deaths for the male and female.

The application of ARIMA models for long-term leukemia prediction will further facilitate the research in the areas monitoring the occurrence of death rates of other disease, such as pneumonia and influenza, diabetes, accidents and their advert's effects, teen pregnancy, suicide, as well as other disease of interest. Therefore this research will be beneficial to other researchers in this vital field of study.
Table 1A: Leukemia Deaths in the United States. (www.seer.cancer.gov)

<table>
<thead>
<tr>
<th>Years</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1969</td>
<td>8,256</td>
<td>6,193</td>
<td>14,449</td>
</tr>
<tr>
<td>1970</td>
<td>8,128</td>
<td>6,364</td>
<td>14,492</td>
</tr>
<tr>
<td>1971</td>
<td>8,205</td>
<td>6,263</td>
<td>14,468</td>
</tr>
<tr>
<td>1972</td>
<td>8,325</td>
<td>6,292</td>
<td>14,617</td>
</tr>
<tr>
<td>1973</td>
<td>8,262</td>
<td>6,215</td>
<td>14,477</td>
</tr>
<tr>
<td>1974</td>
<td>8,230</td>
<td>6,344</td>
<td>14,574</td>
</tr>
<tr>
<td>1975</td>
<td>8,382</td>
<td>6,372</td>
<td>14,754</td>
</tr>
<tr>
<td>1976</td>
<td>8,556</td>
<td>6,500</td>
<td>15,056</td>
</tr>
<tr>
<td>1977</td>
<td>8,609</td>
<td>6,717</td>
<td>15,326</td>
</tr>
<tr>
<td>1978</td>
<td>8,682</td>
<td>6,708</td>
<td>15,390</td>
</tr>
<tr>
<td>1979</td>
<td>9,019</td>
<td>7,140</td>
<td>16,159</td>
</tr>
<tr>
<td>1980</td>
<td>9,325</td>
<td>7,383</td>
<td>16,708</td>
</tr>
<tr>
<td>1981</td>
<td>9,201</td>
<td>7,241</td>
<td>16,442</td>
</tr>
<tr>
<td>1982</td>
<td>9,376</td>
<td>7,509</td>
<td>16,885</td>
</tr>
<tr>
<td>1983</td>
<td>9,447</td>
<td>7,561</td>
<td>17,008</td>
</tr>
<tr>
<td>1984</td>
<td>9,392</td>
<td>7,849</td>
<td>17,241</td>
</tr>
<tr>
<td>1985</td>
<td>9,563</td>
<td>7,927</td>
<td>17,490</td>
</tr>
<tr>
<td>1986</td>
<td>9,685</td>
<td>7,851</td>
<td>17,536</td>
</tr>
<tr>
<td>1987</td>
<td>9,487</td>
<td>7,953</td>
<td>17,440</td>
</tr>
<tr>
<td>1988</td>
<td>9,831</td>
<td>7,910</td>
<td>17,741</td>
</tr>
<tr>
<td>1989</td>
<td>10,142</td>
<td>8,264</td>
<td>18,406</td>
</tr>
<tr>
<td>1990</td>
<td>10,290</td>
<td>8,435</td>
<td>18,725</td>
</tr>
<tr>
<td>1991</td>
<td>10,286</td>
<td>8,817</td>
<td>19,103</td>
</tr>
<tr>
<td>1992</td>
<td>10,705</td>
<td>8,712</td>
<td>19,417</td>
</tr>
<tr>
<td>1993</td>
<td>10,872</td>
<td>8,834</td>
<td>19,706</td>
</tr>
<tr>
<td>1994</td>
<td>10,948</td>
<td>8,885</td>
<td>19,833</td>
</tr>
<tr>
<td>1995</td>
<td>11,347</td>
<td>8,976</td>
<td>20,323</td>
</tr>
<tr>
<td>1996</td>
<td>11,265</td>
<td>9,229</td>
<td>20,494</td>
</tr>
<tr>
<td>1997</td>
<td>11,379</td>
<td>9,105</td>
<td>20,484</td>
</tr>
<tr>
<td>1998</td>
<td>11,297</td>
<td>9,172</td>
<td>20,469</td>
</tr>
<tr>
<td>1999</td>
<td>11,543</td>
<td>9,528</td>
<td>21,071</td>
</tr>
<tr>
<td>2000</td>
<td>11,803</td>
<td>9,594</td>
<td>21,397</td>
</tr>
<tr>
<td>2001</td>
<td>11,894</td>
<td>9,638</td>
<td>21,532</td>
</tr>
<tr>
<td>2002</td>
<td>12,058</td>
<td>9,523</td>
<td>21,581</td>
</tr>
<tr>
<td>2003</td>
<td>12,104</td>
<td>9,504</td>
<td>21,608</td>
</tr>
<tr>
<td>2004</td>
<td>12,051</td>
<td>9,421</td>
<td>21,472</td>
</tr>
<tr>
<td>2005</td>
<td>12,273</td>
<td>9,443</td>
<td>21,716</td>
</tr>
<tr>
<td>2006</td>
<td>12,426</td>
<td>9,590</td>
<td>22,016</td>
</tr>
<tr>
<td>2007</td>
<td>12,434</td>
<td>9,494</td>
<td>21,928</td>
</tr>
</tbody>
</table>

January 1969-December 2007
## Table 2A: ERRR with Time step h= 1 year

<table>
<thead>
<tr>
<th>Time-step</th>
<th>Total</th>
<th>Male</th>
<th>ERRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1969</td>
<td>14449</td>
<td>8256</td>
<td>0.571389</td>
</tr>
<tr>
<td>1970</td>
<td>14492</td>
<td>8128</td>
<td>0.566117</td>
</tr>
<tr>
<td>1971</td>
<td>14468</td>
<td>8205</td>
<td>0.566449</td>
</tr>
<tr>
<td>1972</td>
<td>14617</td>
<td>8325</td>
<td>0.567228</td>
</tr>
<tr>
<td>1973</td>
<td>14477</td>
<td>8262</td>
<td>0.567921</td>
</tr>
<tr>
<td>1974</td>
<td>14574</td>
<td>8230</td>
<td>0.567383</td>
</tr>
<tr>
<td>1975</td>
<td>14754</td>
<td>8382</td>
<td>0.567489</td>
</tr>
<tr>
<td>1976</td>
<td>15056</td>
<td>8556</td>
<td>0.567911</td>
</tr>
<tr>
<td>1977</td>
<td>15326</td>
<td>8609</td>
<td>0.566911</td>
</tr>
<tr>
<td>1978</td>
<td>15390</td>
<td>8682</td>
<td>0.566621</td>
</tr>
<tr>
<td>1979</td>
<td>16159</td>
<td>9019</td>
<td>0.565784</td>
</tr>
<tr>
<td>1980</td>
<td>16708</td>
<td>9325</td>
<td>0.565075</td>
</tr>
<tr>
<td>1981</td>
<td>16442</td>
<td>9201</td>
<td>0.564618</td>
</tr>
<tr>
<td>1982</td>
<td>16885</td>
<td>9376</td>
<td>0.563881</td>
</tr>
<tr>
<td>1983</td>
<td>17008</td>
<td>9447</td>
<td>0.563259</td>
</tr>
<tr>
<td>1984</td>
<td>17241</td>
<td>9392</td>
<td>0.561972</td>
</tr>
<tr>
<td>1985</td>
<td>17490</td>
<td>9563</td>
<td>0.560971</td>
</tr>
<tr>
<td>1986</td>
<td>17536</td>
<td>9685</td>
<td>0.560433</td>
</tr>
<tr>
<td>1987</td>
<td>17440</td>
<td>9487</td>
<td>0.559478</td>
</tr>
<tr>
<td>1988</td>
<td>17741</td>
<td>9831</td>
<td>0.559478</td>
</tr>
<tr>
<td>1989</td>
<td>18406</td>
<td>10142</td>
<td>0.558735</td>
</tr>
<tr>
<td>1990</td>
<td>18725</td>
<td>10290</td>
<td>0.558250</td>
</tr>
<tr>
<td>1991</td>
<td>19103</td>
<td>10286</td>
<td>0.557240</td>
</tr>
<tr>
<td>1992</td>
<td>19417</td>
<td>10705</td>
<td>0.556948</td>
</tr>
<tr>
<td>1993</td>
<td>19706</td>
<td>10872</td>
<td>0.556998</td>
</tr>
<tr>
<td>1994</td>
<td>19833</td>
<td>10948</td>
<td>0.556484</td>
</tr>
<tr>
<td>1995</td>
<td>20323</td>
<td>11347</td>
<td>0.556567</td>
</tr>
<tr>
<td>1996</td>
<td>20494</td>
<td>11265</td>
<td>0.556269</td>
</tr>
<tr>
<td>1997</td>
<td>20484</td>
<td>11379</td>
<td>0.556237</td>
</tr>
<tr>
<td>1998</td>
<td>20469</td>
<td>11297</td>
<td>0.556065</td>
</tr>
<tr>
<td>1999</td>
<td>21071</td>
<td>11543</td>
<td>0.555741</td>
</tr>
<tr>
<td>2000</td>
<td>21397</td>
<td>11803</td>
<td>0.555583</td>
</tr>
<tr>
<td>2001</td>
<td>21532</td>
<td>11894</td>
<td>0.555464</td>
</tr>
<tr>
<td>2002</td>
<td>21581</td>
<td>12058</td>
<td>0.555581</td>
</tr>
<tr>
<td>2003</td>
<td>21608</td>
<td>12104</td>
<td>0.555741</td>
</tr>
<tr>
<td>2004</td>
<td>21472</td>
<td>12051</td>
<td>0.555924</td>
</tr>
<tr>
<td>2005</td>
<td>21716</td>
<td>12273</td>
<td>0.556225</td>
</tr>
<tr>
<td>2006</td>
<td>22016</td>
<td>12426</td>
<td>0.556487</td>
</tr>
<tr>
<td>2007</td>
<td>21928</td>
<td>12434</td>
<td>0.556813</td>
</tr>
</tbody>
</table>

January 1969- December 2007
Table 3A. ERR with a Time-step $h = 1\text{ year}$.

<table>
<thead>
<tr>
<th>Count</th>
<th>Number of male</th>
<th>ERR (in 1 yr.)</th>
<th>Number of Female</th>
<th>ERR (in 1 yr.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8,256</td>
<td>8256</td>
<td>6,193</td>
<td>6193</td>
</tr>
<tr>
<td>2</td>
<td>8,128</td>
<td>8192</td>
<td>6,364</td>
<td>6278.5</td>
</tr>
<tr>
<td>3</td>
<td>8,205</td>
<td>8196.333333</td>
<td>6,263</td>
<td>6273.333333</td>
</tr>
<tr>
<td>4</td>
<td>8,325</td>
<td>8228.5</td>
<td>6,292</td>
<td>6278</td>
</tr>
<tr>
<td>5</td>
<td>8,262</td>
<td>8235.2</td>
<td>6,215</td>
<td>6265.4</td>
</tr>
<tr>
<td>6</td>
<td>8,230</td>
<td>8234.333333</td>
<td>6,344</td>
<td>6278.5</td>
</tr>
<tr>
<td>7</td>
<td>8,382</td>
<td>8255.428571</td>
<td>6,372</td>
<td>6291.857143</td>
</tr>
<tr>
<td>8</td>
<td>8,556</td>
<td>8293</td>
<td>6,500</td>
<td>6317.875</td>
</tr>
<tr>
<td>9</td>
<td>8,609</td>
<td>8328.111111</td>
<td>6,717</td>
<td>6362.222222</td>
</tr>
<tr>
<td>10</td>
<td>8,682</td>
<td>8363.5</td>
<td>6,708</td>
<td>6396.8</td>
</tr>
<tr>
<td>11</td>
<td>9,019</td>
<td>8423.090909</td>
<td>7,140</td>
<td>6464.363636</td>
</tr>
<tr>
<td>12</td>
<td>9,325</td>
<td>8498.25</td>
<td>7,383</td>
<td>6540.916667</td>
</tr>
<tr>
<td>13</td>
<td>9,201</td>
<td>8552.307692</td>
<td>7,241</td>
<td>6594.769231</td>
</tr>
<tr>
<td>14</td>
<td>9,376</td>
<td>8611.142857</td>
<td>7,509</td>
<td>6660.071429</td>
</tr>
<tr>
<td>15</td>
<td>9,447</td>
<td>8666.866667</td>
<td>7,561</td>
<td>6720.133333</td>
</tr>
<tr>
<td>16</td>
<td>9,392</td>
<td>8712.1875</td>
<td>7,849</td>
<td>6790.6875</td>
</tr>
<tr>
<td>17</td>
<td>9,563</td>
<td>8762.235294</td>
<td>7,927</td>
<td>6857.529412</td>
</tr>
<tr>
<td>18</td>
<td>9,685</td>
<td>8813.5</td>
<td>7,851</td>
<td>6912.722222</td>
</tr>
<tr>
<td>19</td>
<td>9,487</td>
<td>8848.947368</td>
<td>7,953</td>
<td>6967.473684</td>
</tr>
<tr>
<td>20</td>
<td>9,831</td>
<td>8898.05</td>
<td>7,910</td>
<td>7014.6</td>
</tr>
<tr>
<td>21</td>
<td>10,142</td>
<td>8957.285714</td>
<td>8,264</td>
<td>7074.095238</td>
</tr>
<tr>
<td>22</td>
<td>10,290</td>
<td>9017.863636</td>
<td>8,435</td>
<td>7135.954545</td>
</tr>
<tr>
<td>23</td>
<td>10,286</td>
<td>9073</td>
<td>8,817</td>
<td>7209.043478</td>
</tr>
<tr>
<td>24</td>
<td>10,705</td>
<td>9141</td>
<td>8,712</td>
<td>7271.666667</td>
</tr>
<tr>
<td>25</td>
<td>10,872</td>
<td>9210.24</td>
<td>8,834</td>
<td>7334.16</td>
</tr>
<tr>
<td>26</td>
<td>10,948</td>
<td>9277.076923</td>
<td>8,885</td>
<td>7393.807692</td>
</tr>
<tr>
<td>27</td>
<td>11,347</td>
<td>9353.740741</td>
<td>8,976</td>
<td>7452.407407</td>
</tr>
<tr>
<td>28</td>
<td>11,265</td>
<td>9422</td>
<td>9,229</td>
<td>7515.857143</td>
</tr>
<tr>
<td>29</td>
<td>11,379</td>
<td>9489.482759</td>
<td>9,105</td>
<td>7570.655172</td>
</tr>
<tr>
<td>30</td>
<td>11,297</td>
<td>9549.733333</td>
<td>9,172</td>
<td>7624.033333</td>
</tr>
<tr>
<td>31</td>
<td>11,543</td>
<td>9614.032258</td>
<td>9,528</td>
<td>7685.451613</td>
</tr>
<tr>
<td>32</td>
<td>11,803</td>
<td>9682.4375</td>
<td>9,594</td>
<td>7745.09375</td>
</tr>
<tr>
<td>33</td>
<td>11,894</td>
<td>9749.454545</td>
<td>9,638</td>
<td>7802.454545</td>
</tr>
<tr>
<td>34</td>
<td>12,058</td>
<td>9817.352941</td>
<td>9,523</td>
<td>7853.058824</td>
</tr>
<tr>
<td>35</td>
<td>12,104</td>
<td>9882.685714</td>
<td>9,504</td>
<td>7900.228571</td>
</tr>
<tr>
<td>36</td>
<td>12,051</td>
<td>9942.916667</td>
<td>9,421</td>
<td>7942.472222</td>
</tr>
<tr>
<td>37</td>
<td>12,273</td>
<td>10005.89189</td>
<td>9,443</td>
<td>7983.027027</td>
</tr>
<tr>
<td>38</td>
<td>12,426</td>
<td>10069.57895</td>
<td>9,590</td>
<td>8025.315789</td>
</tr>
<tr>
<td>39</td>
<td>12,434</td>
<td>10130.20513</td>
<td>9,494</td>
<td>8062.974359</td>
</tr>
</tbody>
</table>

January 1969-December 2007
### Notation and acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHPP</td>
<td>Non homogeneous Poisson process</td>
</tr>
<tr>
<td>HPP</td>
<td>Homogeneous Poisson process</td>
</tr>
<tr>
<td>M(t)</td>
<td>Mean function of an NHPP</td>
</tr>
<tr>
<td>( \lambda(t) )</td>
<td>Intensity function of an NHPP</td>
</tr>
<tr>
<td>ARIMA</td>
<td>Autoregressive integrated moving Average</td>
</tr>
<tr>
<td>ARMA</td>
<td>Autoregressive moving average</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimator</td>
</tr>
<tr>
<td>ERRR</td>
<td>Empirical recursive rates ratio</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>MA</td>
<td>Moving Average</td>
</tr>
<tr>
<td>( {z_i} )</td>
<td>A discrete time series</td>
</tr>
<tr>
<td>B</td>
<td>Backshift Operator</td>
</tr>
<tr>
<td>Lag</td>
<td>Time separation or time step</td>
</tr>
<tr>
<td>SAFC</td>
<td>Sample autocorrelated function</td>
</tr>
<tr>
<td>SPACF</td>
<td>Sample partial autocorrelated</td>
</tr>
<tr>
<td>ITSM</td>
<td>Time series computing package</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike mode information formation criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>Schwartz model selection information Criterion</td>
</tr>
<tr>
<td>AICC</td>
<td>Estimated corrected version of AIC</td>
</tr>
</tbody>
</table>
REFERENCES


Shumway and Stoffer, 2006. Time series Analysis and its applications with R example
Degree:

Bachelor of Science in Mathematics, 2001
Kwame Nkrumah University of Science and Technology, Kumasi

Thesis Title: Modeling mortality rates for leukemia between men and women in the United States

Thesis Examination Committee:

Chairperson, Chih-Hsiang Ho, Ph.D.
Committee Member, Amei Amei, Ph.D.
Committee Member, Anton Westveld, Ph.D.
Graduate Faculty Representative, Chad Cross, Ph.D. MS, MFT,LCADC