CHAPTER 1
INTRODUCTION

1.1 MOTIVATION

Due to the wide use of software in various systems in modern life, software failure, ranging from inconvenience like malfunctions of switching systems to life threats like failures of medical systems or spaceship systems due to unreliable software, has been reported to be the major cause of various systems outages according to the study made in Lyu et al. (1994). The reliability of software is becoming more and more crucial. In order to manage the projects effectively, it is hoped software is made more reliable, can be modeled through the use of Software Reliability Growth Models (SRGMs).

In the past decades, various statistical tools have been applied to solve the problems of software reliability. Software reliability is defined as the probability that a software system operates with no failure occurring for a specified time on specified operating conditions. By proposing statistical models, i.e. software reliability models, reliability can be measured.

By estimating, predicting and managing the reliability, cost and performance of software-based systems, the quality of software products is aided to get improved and the satisfaction of customers are gained. Software reliability models are applied to help estimate the total number of faults latent in a software system and evaluate whether the software meets the reliability requirements, and provide a guiding measurement to determine the optimal release time of a software system before the software is released to customers.

Software reliability models set up an effective, quantitative bridge to understand and measure software quality. Since the first influential software reliability model, the Jelinski-Moranda deeutrophication model (Jelinski and Moranda, 1972) published, many efforts have been made in the area of software reliability modeling. Of the existing stochastic software reliability models, a large subgroup is the Non-Homogeneous Poisson Process (NHPP) software reliability growth model. It has been theoretically justified the appropriateness of using NHPP to characterize the failure process (Miller, 1984). The Goel-Okumoto
model is one of the first such models (Goel and Okumoto, 1979). More on software reliability can be found in Musa et al. (1987), Xie (1991), Lyu (1996), and Pham (2000).

In this research, the NHPP software reliability models are studied. Various NHPP models have been developed based on some specific assumptions. Many models such as the G-O model assume perfect debugging, i.e. each time a failure occurs, the fault that has caused the failure is immediately and completely removed and no new faults are introduced. Another assumption of NHPP reliability models is that each failure occurs independently and randomly from the same distribution during the fault detection process (Musa et al., 1987). The failure distribution can be affected by many factors, such as the running environment, testing strategy and resource allocation. Once these factors are changed during the software-testing phase, the failure process may have changed from one distribution to another. This is the so-called change-point problem (Zhao, 1993).

The desirable characteristic of any product is the quality. The quality of software can be judged by its correctness, adaptability and testability. A software product has a typical development process. Some activities of the process are related to particular phases of development. The other activities of the software process called umbrella activities are applied throughout the software development. One of those activities is assessing quality of the software. As in the case of any other product to obtain quality software the only way is to identify the opportunities of improvement at every stage of software development. The tests and reviews conducted throughout the development come under the quality control.

Software reliability is estimated by the use of an analytical model whose unknown parameters are estimated from the available failure data. A perfect program is said to be reliable while an imperfect one is said to be unreliable. The probability that software faults do not cause a failure during a specified exposure period in a specified environment is called software reliability.

Numerous SRGMs have been proposed, and some appear to be better overall than others. Unfortunately, models that are good overall are not always the
best choice for a particular data set, and it is not possible to know which model to use a priori. Even when an appropriate model is used, the predictions made by a model may still be less accurate than desired. For this reason, a great deal of research has gone into trying to make more effective use of existing models.

The goal of this thesis is to discover methods which will allow us to calculate the reliability and stability of the software using software reliability growth models. To do this we need quantitative ways of measuring the model parameters. It is to measure how well a model fits the observed data; we are concerned with a model’s ability to examine the stability of software. How the parameters should be estimated and what other techniques should be employed to ensure that traditional software reliability growth models give the best results possible. In the process of doing this we will examine exactly what the parameters of these models mean. This study will examine how existing software reliability growth model can be used to know the reliability and stability of the software under consideration using the failures.

Quantitative approaches, particularly statistical techniques, have successfully been used to support project control and process improvement in manufacturing environments. Such approaches have also been applied within software engineering to enable a better understanding of software development and to improve software product quality. This thesis demonstrates how a sound adoption of Statistical Process Control (SPC) based control charts and Sequential probability Ratio Test (SPRT) can form the basis for identifying the process variation using software defect data, thereby, improving software quality and reliability.

1.2 SIGNIFICANCE OF THE RESEARCH

The three problems which were studied in the thesis will help the developers during development. They can monitor the failure process and analyze whether the process is within the control limits or not.

A Two step approach is proposed and developed to estimate the model parameters and then with the help of these parameters the model is used to test the reliability of the software.
Two approaches were developed in our study to deal with the failure detection problem.

A control chart is developed to monitor the failure process to detect the points of failure. How well the chart performs is studied by applying on the real time data available from different sources of literature.

The use of Sequential probability Ratio Test in deciding, whether the test process is to be accepted, rejected or to continue.

1.3 ORGANIZATION OF THE RESEARCH

This research is organized into six parts:

Chapter 1 is the introduction which includes the problem that is dealt with in this study.

Chapter 2 is the review of relative literature mainly focused on NHPP software reliability models and summarizes the works concerning the different concepts used and the application of statistical control chart used in the field of software engineering.

Chapter 3 studies the goodness of fit of the data sets to the HLSRGM using two step approach. Applies the chart to a real software testing failure dataset, and studies the performance of the chart and its effectiveness;

Chapter 4 contains the use of control chart, applies the chart to a real software testing failure dataset, and studies the performance of the chart and its effectiveness;

Chapter 5 is on the application of SPRT (Sequential Probability Ratio Test) in detecting failures in advance.

Chapter 6 gives the conclusion of the study; further study and limitation are discussed as well.

1.4 SOFTWARE RELIABILITY

The ability of a system or a component to perform its required functions under stated conditions for a specified period of time is referred to as software reliability. It is a probabilistic measure and can be defined as the probability that
software faults do not cause a failure during a specified exposure period in a specified use environment. The probabilistic nature of this measure is due to the uncertainty in the usage of the various software functions and the specified exposure period, that may mean a single run, a number of runs, or time expressed in calendar or execution time units. Software reliability is a useful measure in planning and controlling resources during the development process so that high quality software can be developed.

Usually logic errors in the software are not hard to fix but diagnosing logic bugs is the most challenging. For many reasons, the fault is usually subtle. A computer system consists of two major components: hardware and software. Although extensive research has been carried out on hardware reliability, the growing importance of recent software in complex applications dictates that the major focus has shifted to software reliability. Failures of the software may result in an unintendent system state, may cause property damage or destruction, people are injured or killed, and / or monetary costs are incurred. As software projects become large, the rate of software defects increases geometrically, and locating software faults is extremely difficult and costly.

Software faults are more insidious and much more difficult to handle than physical defects. In theory software can be error free, and unlike hardware, does not degrade or wear out but it does deteriorate. The deterioration here, however, is not a function of time. Rather, it is a function of the results of changes made in the software during maintenance, through correcting latent defects, modifying the code to changing requirements and specifications, environments and applications, or improving software performance. All design faults are present from the time the software is installed in the computer. In principle, these faults could be removed completely, but in reality the goal of perfect software remains elusive (Friedman and Voas, 1995). Computer programs, which vary for fairly critical applications between hundreds and millions of lines of code, can make the wrong decision because the particular inputs that triggered the problem were not tested and corrected during the testing phase. Such inputs may even have been misunderstood or unanticipated by the designer who either correctly programmed the wrong interpretation or failed to identify the problem. These situations and
other such events have made it apparent that we must determine the reliability of the software systems before putting them into operation.

It is also a useful measure for giving the user confidence about software correctness. Planning and controlling the testing resources via the software reliability measure can be done by balancing the additional cost of testing and the corresponding improvement in software reliability. As more and more faults are exposed by the testing and verification process, the additional cost of exposing the remaining faults generally rises very quickly. Thus, there is a point beyond which continuation of testing to further improve the quality of software can be justified only if such improvement is cost effective. An objective measure like software reliability can be used to study such a tradeoff.

Research on software reliability engineering has been conducted during the past few decades and numerous statistical models have been proposed for estimating software reliability (Pham 1999a, 2000). Most of the estimating models for predicting software reliability are based purely on the observation of software product failure, where they required a considerable amount of failure data to obtain an accurate software reliability prediction based on NHPP Models.

1.4.1 Software Reliability: An Attribute of Software Quality

Quality is defined simply as meeting the requirements of the customer and this has been expressed in several ways. Quality control is a series of instructions, reviews and tests used throughout the development of products to ensure that each work product meets the requirements placed upon it. In order to be competent enough in software industry there has to be a reference to quality. The growth of industry depends directly on the quality that can be achieved if it is implemented in all phases of software development. Lots of strategies are adopted for quality like prevention of defects, minimizing the number of bugs so that the quality requirements are met. Certain standards are to be maintained in areas like requirement analysis, coding, integration tests, system testing etc.

Since software is an intellectual creation, it is difficult generally to measure the software quality. However, it can be judged by its correctness, adaptability, testability, obsolescence and intra-operability.
An important quality attribute of a computer system is the degree to which it can be relied upon to perform its intended functions. Evaluation, prediction, and improvement of this attribute have been of concern to designers and users of computers from the early days of their evolutions. Software is essentially an instrument for transforming a discrete set of inputs into a discrete set of outputs. It comprises a set of coded statements whose function may be to evaluate an expression and store the result in a temporary or permanent location, decide which statement to execute next, or to perform input/output operations. Since, to a large extent, software is produced by humans, the finished product is often imperfect. It is imperfect in the sense that a discrepancy exists between what the software can do versus what the user or the computing environment wants it to do. These discrepancies are what we call software faults. Basically, software faults can be attributed to an ignorance of the user requirements, ignorance of the rules of the computing environment, and to poor communication of software requirements between the user and the programmer or poor documentation of the software by the programmer. Even if we know that software contains faults, we generally do not know their exact identity. There are two approaches available for indicating the existence of software faults, namely program proving, and program testing.

In practice neither proving nor testing can guarantee complete confidence in the correctness of a program. Each has its advantages and limitations and should not be viewed as computing tools. They are, in fact, complementary methods for decreasing the likelihood of program failure. Due to the imperfectness of these approaches in assuring a correct program, a metric is needed, which reflects the degree of program correctness and which can be used in planning and controlling additional resources needed for enhancing software quality. One such quantifiable metric of quality that is commonly used in software engineering practice is software reliability. A commonly used approach for measuring software reliability is via an analytical model whose parameters are generally estimated from available data on software failures. Reliability and other relevant measures are then computed from the fitted model.

There are a number of views as to what software reliability is and how it should be quantified. Some people believe that this measure should be binary in nature so that an imperfect program would have zero reliability while a perfect
one would have a reliability value of one. This view parallels that of program proving where by the program is either correct or incorrect. Others, however, feel that software reliability should be defined as the relative frequency of the times that the program works as intended by the user. This view is similar to that taken in testing where a percentage of the successful cases are used as a measure of program quality.

SRGMs can be used to estimate the quantities associated with software reliability, such as the remaining number of faults and the failure rate, etc. NHPP models, as a class of SRGMs, are extensively used. Software reliability \( R(x/t) \) is defined as the probability that no software failure is observed in the time interval \((t, t+x)\), given that the last failure occurred at a time point \( t (t \geq 0, x > 0) \). Based on NHPP, \( R(x/t) \) can be expressed as follows (Zhang et al., 2003).

\[
R(x/t) = e^{-m(t+x)-m(t)}
\]  

(1.4.1.1)

1.4.2 Software Process and Software quality

Process is the transformation of a set of inputs, which can include materials, actions, methods and operations into desired outputs in the form of products, information, services or results. In each area of an organization there will be many processes taking place. Each process may be analyzed by an examination of the inputs and outputs. This will determine the action necessary to improve quality. Clearly, to produce an output which meets the requirements of the customer, it is necessary to define, monitor and control the inputs to the process (Oakland, 2007).

A software process, viewed as a system, has a development life cycle consisting of activities that include requirements definition, design, implementation and test. The requirements for a software process are specified in terms of quality characteristics specified by the customer. Process design is established by partitioning the process into meaningful activities that can be defined in terms of inputs, outputs and constraints. Implementation is accomplished by specifying the procedure for each activity and assigning responsibility for each procedure. In the manufacturing arena, it is not difficult to figure out the relationship between product quality and corresponding production
process. Therefore, we can measure process attributes, work on them, improve according to the results and produce high quality products. On the other hand, software product is difficult to characterize, as it is not concrete, it is difficult to recognize the correlation between a single software process and the quality of the related software product. The software process must be assessed based on its capability to develop software that is consistent with user’s requirement. Actually, there is no specific software measure showing the extent to which customer requirements are met. However, there are processes and products that influence production life cycle by measuring specific characteristics of these processes and products. We can have an idea about the quality of the final product. One difficulty in software production is that, there is not a repetitive production. Each product is distinct and possesses different characteristics. For this reason, it is usually not possible to form a sample of n measurements. In this case, we shall take care of each single measurement, and perform the analysis accordingly.

Beside these general software process characteristics, software measurement is also a very complicated process. Each metric requires different measurement techniques and the reliability of the metric depends on how well the metric is defined, how properly the data collection procedures are performed and how robust the measures are with respect to varying environmental conditions. Excessive variability of target processes cause defects. Defects introduced into a product have multiple effects. They require the effort of skilled personnel to detect, remove, repair and retest to improve reliability. Defects also increase process cost, time and complexity. In addition, defects that escape detection and repair before the product is released to a customer reduce reliability of the product. Activities that decrease the introduction of defects or increase the early detection of defects are prime targets for measuring the effectiveness of the process.

The software process has been identified as a process that is dominated by design risk or cognition because each instance of a software process produces a unique product for which quality is determined by conformance to customer requirements measured in terms of software quality characteristics. Software quality characteristics are those attributes of software that include correctness, reliability, understandability, portability, maintainability, testability, robustness,
usability, cost-to-develop, and time-to develop. Adherence to customer requirements refers to the elimination of variance in quality characteristics, in other words, eliminating differences between customer expectations and the delivered software product. As these requirements vary for each customer and often conflict about a given instance of the software process. Despite the difficulties mentioned, it may be still possible to apply SPC to software process.

1.5 NHPP SRGM

The NHPP group of models provides an analytical framework for describing the software failure phenomenon during testing. They are proved to be quite successful in practical software reliability engineering (Musa et al., 1987). They have been built upon various assumptions. The main issue in the NHPP model is to determine an appropriate mean value function to denote the expected number of failures experienced up to a certain time point. Model parameters can be estimated by using two step approach (i.e one parameter is estimated through Maximum Likelihood Estimate (MLE) and another parameter is estimated through Regression). Various NHPP SRGMs have been proposed upon various assumptions. Many of the SRGMs assume that each time a failure occurs, the fault that caused it can be immediately removed and no new faults are introduced, which is usually called perfect debugging. Imperfect debugging models have proposed a relaxation of the above assumption (Ohba, 1984; Pham, 1993).

If ‘t’ is a continuous random variable with probability density function: 

\[ f(t;\theta_1,\theta_2,\ldots,\theta_k) \]

where \(\theta_1,\theta_2,\ldots,\theta_k\) are k unknown constant parameters which need to be estimated, and cumulative distribution function: \(F(t)\). Let ‘a’ denote the expected number of faults that would be detected given infinite testing time in case of finite failure NHPP models and ‘b’ represents the fault detection rate. In software reliability, the initial number of faults and the fault detection rate are always unknown. The NHPP models are further classified into Finite and Infinite failure models. Then, the mean value function of the finite failure NHPP models can be written as: 

\[ m(t) = aF(t) \]

representing the expected number of software failures by time ‘t’. The failure intensity function \(\lambda(t)\) in case of the finite failure
NHPP models is given by: \( \lambda(t) = aF'(t) \), which is proportional to the residual fault content (Pham, 2006).

Let \( N(t) \) be the cumulative number of software failures by time ‘t’. A non-negative integer-valued stochastic process \( N(t) \) is called a counting process, if \( N(t) \) represents the total number of occurrences of an event in the time interval \([0,t]\) and satisfies these two properties:

1. If \( t_1 < t_2 \), then \( N(t_1) \leq N(t_2) \)
2. If \( t_1 < t_2 \), then \( N(t_2) - N(t_1) \) is the number of occurrences of the event in the interval \([t_1, t_2]\).

One of the most important counting processes is the Poisson process. A counting process, \( N(t) \), is said to be a Poisson process with intensity \( \lambda \) if

1. The initial condition is \( N(0) = 0 \)
2. The failure process, \( N(t) \), has independent increments
3. The number of failures in any time interval of length \( s \) has a Poisson distribution with mean \( \lambda s \), that is, \( P\{N(t+s)-N(t)=n\} = \frac{e^{-\lambda s} (\lambda s)^n}{n!} \)

Describing uncertainty about an infinite collection of random variables one for each value of ‘t’ is called a stochastic counting process denoted by \([N(t), t \geq 0]\). The process \( \{N(t), t \geq 0\} \) is assumed to follow a Poisson distribution with characteristic MVF (Mean Value Function) \( m(t) \) representing the expected number of software failures by time ‘t’. Different models can be obtained by using different non decreasing \( m(t) \). The derivative of \( m(t) \) is called the failure intensity function \( \lambda(t) \).

A Poisson process model for describing about the number of software failures in a given time \((0, t)\) is given by the probability equation.

\[
P[N(t) = y] = \frac{e^{-m(t)} [m(t)]^y}{y!}, \quad y = 0, 1, 2, ...
\]
Where, $m(t)$ is a finite valued non negative and non decreasing function of $t$ called the mean value function. Such a probability model for $N(t)$ is said to be an NHPP model.

1.6 STATISTICAL PROCESS CONTROL

Statistical Process Control (SPC) is a methodology that aims to provide process control in statistical terms. Since the great industrial revolution in Japan, SPC has been widely used in manufacturing industries in order to control variability and improve processes (Sutherland et al., 1992). The basic tools used for statistical control are (Paulk and Chrissis, 2002, Montgomery, 1991)

- Check sheets are good means for collecting data efficiently, reliably and easily. They are designed specifically considering the particular needs. In software industry, metric datasheet is in form of a check sheet, which is used to collect measurement data.
- Cause-and-Effect Diagrams are useful tools to visualize, categorize and rank potential causes of a problem, a situation or any outcome.
- Scatter Diagram, data for two variables are collected in pairs $(x_i, y_i)$, and each point $y_i$ is plotted against corresponding $x_i$. This is a useful plot for identifying a potential relationship between two process characteristics. A pattern in the plotted points may suggest that the two factors are associated, perhaps with a cause-effect relationship.
- Run Charts are specialized, time-sequenced form of scatter diagrams that can be used to examine data quickly and informally for trends or other patterns that occur over time.
- Histograms show the frequency distribution of data in a sample.
- Bar charts are like histograms. But they are not only used for depicting the frequencies of occurrences, but also for showing any numerical value of the attribute.
• Pareto chart is another form of bar chart. However, the occurrences are ordered with respect to their frequencies. Pareto charts are good means to visualize the ranking of an attribute among different categories.

• Control Charts are sophisticated statistical data analysis tools, which include upper and lower limits to detect any outliers. They are frequently used in SPC analyses and will be described in detail in the following sections.

1.6.1 Variability in Processes

In 1920s, W. A. Shewhart was working on the idea of quality control and he brought the idea that each process is driven by forces of variation. However, variation results in loss of quality by causing inefficiency and waste. If we can understand the sources of variation, we can take necessary actions to remove inefficiency, and increase quality. If we think of variation in software industry, lines of code produced a day can be considered as a variable parameter. If the same person produces Lines of Code for the same component, the amount of time he spends will be different from one day to the other. This can be explained as the variation in a process attribute.

According to Shewhart, variation in a process has two types of causes: assignable causes and chance causes. Assignable causes appear in unexpected periods and can be fixed by immediate actions. For instance, if a new tool is being used for coding, the productivity of the coder may be lower during adaptation period. When this is realized, a training program can be implemented to improve productivity. On the other hand, chance causes are the results of the system itself (Deming, 2000) refers them as Common Causes. They are naturally existent within the defined processes and can only be avoided by performing improvement programs. If we think of a software engineering firm that has no reusable code library, we witness that similar code pieces are written separately in each new application and this causes delays due to rework. Such a chance cause can be prevented by creating a reusable code library.

1.6.2 Statistical Control

If the variation in the behavior of a process is predictable in statistical terms, that process is said to be in control. This means that, we can expect (within
certain limits) what the outcome will be the next time we perform the same process. In this way, we can prepare more accurate project plans; do better cost estimations and schedule activities in more reasonable basis. In order to calculate the variance in process behavior, several attributes or variables representing the outcomes of the process shall be defined.

The number of defects found during unit testing, the number of requirements that are changed after requirements analysis phase, amount of CPU utilized to perform a specific application, may all be used to understand the behavior of the processes they represent. The variability in process behavior, then, can be tracked through these measures. The aim of Statistical Process Control is: firstly, to detect assignable causes of variation in the processes and provide process control; secondly, to enable monitoring of the improvement in processes (that are already under statistical control) by demonstrating the chance causes; and Shewhart Control Charts are a good means to achieve Statistical Process Control.

1.6.3 Shewhart Control Charts

During his studies at Bell Labs in 1920s, Shewhart proposed that it is possible to define limits within which the results of routine efforts must lie to be economical. Deviations in the process outcomes resulting in values out of these limits indicate that the process is not performed economically. In order to detect assignable causes, Shewhart utilized statistical quality control charts (Shewhart, 1939). Shewhart control chart model depends on hypothesis testing. First of all, a sample of data is collected for the subject measure (i.e. number of defects in a piece of code). Then, its mean and variance are calculated. The lower and upper control limits are derived and data is analyzed using the statistical evidence on hand. By analyzing the data values with respect to upper and lower control limits together with their location in the zones, assignable causes are detected. Then necessary actions are taken and measurements are repeated. The charts are redrawn with the existing data values, and this process is repeated until no evidence remains for the existence of assignable causes.

Once the process is brought under control, further improvement activities are implemented to minimize the effect of common causes. The measurement can be performed by means of either variables or attributes. A variable as “measure of
a product that can have any value between the limits of the measurement”, while an attribute as “count of things which may or may not be present in the product” (Burr and Owen, 2000). The nature of these two measurement categories necessitates different statistical analyses. A variable normally has normal probabilistic distribution, where as it is likely to be binomial for an attribute. Statistical analysis may be performed by implementing X and R charts for sample data. When variable measures are individual data points, Individuals Charts are mostly utilized.

1.7 SPRT

Sequential analysis was first developed by Abraham Wald in the 1940’s and (Wald, 1945; Wald and Wolfowitz, 1948) introduced the SPRT during that time. The SPRT was initially developed by Wald (1947) for quality control problems during World War II. Neyman and Pearson (1933) result inspired Wald to reformulate it as a sequential analysis problem. The original development of the SPRT is used as a statistical device to decide which of two simple hypotheses is more correct. It has been formulated for use in the computerized testing of human examinees as a termination criterion. The properties of it have been studied intensively by many researchers since Wald (1947). The likelihood based SPRT proposed by Wald is very general in that it can be used for many different probability distributions. Some other people worked on this are Siegmund (1986) and Ghosh and Sen (1991).

1.8 MODEL DESCRIPTION: HLSRGM

Software reliability growth models are commonly designed to be used with data collected in terms of the testing time between failures, and it is on such models that this study will focus. These models are usually stated in terms of two equations; the mean value function and the failure intensity function. It is worth noting that the second function is the derivate of the first. The mean value and failure intensity functions are usually denoted as $m(t)$ and $\lambda(t)$, respectively. In this case ‘t’ refers to the total time which has elapsed since the start of testing.
Regardless of how these models were originally formulated, we will refer to the parameters of these models as ‘a’ and ‘b’. Standard practice is to determine the values of these parameters by fitting the model in question to the available data. Once the model has been fitted to the data, it can then be used to obtain estimates of current stability of the software and make predictions about the program's future reliability. The following paragraphs describe the model and will focus on the different functions used in estimating the reliability and stability of the process.

One simple class of finite failure NHPP model is the HLSRGM, assuming that the failure intensity is proportional to the number of faults remaining in the software describing an exponential failure curve. It has two parameters. Where, ‘a’ is the expected total number of faults in the code and ‘b’ is the shape parameter defined as, the rate at which the failure rate decreases. The cumulative distribution function of the model is: 

$$F(t) = \frac{1 - e^{-bt}}{1 + e^{-at}}.$$ 

The expected number of faults at time ‘t’ is denoted by 

$$m(t) = \frac{a(1 - e^{-bt})}{1 + e^{-at}}, a > 0, b > 0, t \geq 0.$$ 

The corresponding failure intensity function is given by 

$$\lambda(t) = \frac{2abe^{-bt}}{(1 + e^{-at})^2}.$$ 

Software reliability is defined as the probability of failure-free software operation for specified period of time ‘t’ in a specified environment.

### 1.9 LEAST SQUARES: REGRESSION

The least squares method is used to compute numerical values of parameters and to fit a function to a set of data. It is one of the oldest techniques of modern statistics as it was first published in 1805 by the French mathematician Legendre. It exists with several variations: its simpler version is called Ordinary Least Squares, more sophisticated versions called Weighted Least Squares, Alternating Least Squares and Partial Least Squares. The oldest use of Ordinary Least Squares was Linear regression, which corresponds to the problem of finding a line that best fits a set of data. In the standard formulation, a set of N pairs of observations
\{Y_i, X_i\} is used to find a function giving the value of the dependent variable ‘Y’ from the values of an independent variable ‘X’. With one variable and a linear function, the prediction is given by the following equation: \( Y_i = a + bX_i \).

Linear regression analysis is to fit lines to data. We can consider a number of approaches. One possibility would be to find a line that would minimize the sum of the error terms. This approach, however, is flawed. The problem with this method is the cancellation of positive errors with negative errors. Our task is to find a line determined by ‘a’ and ‘b’ such that the sum of the squared errors ‘\( e_i \)’ is as small as possible. This, in fact, is called the method of least-squares, sometimes called ordinary least-squares (OLS) so as to distinguish it from other specialized least-squares methods. We can represent this task of minimizing the sum of squared errors mathematically by first noting that the error term \( e_i \) as \( e_i = Y_i - (a + bX_i) \). Where, \( Y_i \), is the observed value and \( (a + bX_i) \) is the predicted value. The OLS method finds a value for ‘a’ and ‘b’ that minimizes the sum of the squared errors. Doing this, we have \( \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \).

At this point, our task now becomes a calculus problem. We have an equation that we want to minimize with respect to two parameters ‘a’ and ‘b’. To carry out this task, we use differential calculus and take the partial derivative of the above Equation with respect to ‘a’ and set it equal to zero, then do the same with respect to ‘b’. We then end up with two equations (the two partial derivatives) with two unknowns. The last step is to solve this system of equations simultaneously for ‘a’ and ‘b’. Leaving out the details, what we end up with is the following formulas for b and a (Gujarati, 2003).

\[
b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\]

\[
a = \bar{Y} - b \bar{X}
\]
Where, in both Equations, the bar above the variable stands for the mean of that variable. Thus, given values for $X_i$ and $Y_i$, we can use these data to calculate ‘a’ and ‘b’.

Statistical approaches such as the linear regression models are used to analyze software failure data. In (Mostafa et al., 2006), it has been demonstrated that for most of the software failure data sets, the linear regression models outperform other popular models from literature in terms of the predictive accuracy of the Mean Time Between Failure (MTBF) estimates. These estimates were measured using the Mean Square Error (MSE) and Mean Absolute Value Difference (MAVD).

Engineers, scientists, and statisticians studied the likelihood of failures of a software system in a mathematically and rigorous procedure. Thus, it is based on the idea of using failure times and time between failures in the past to model the failure behavior of a software system. At the same time, we illustrate software reliability modeling and some other related statistical concepts. The NHPP is a good model for repairable systems because it can model systems that are deteriorating or improving. It is shown that the exponential process can be described as NHPP with intensity function $\lambda(t)$. The shape parameter shows how the system improves or deteriorates over time. The point estimation for the parameters are used for goodness of fit tests, model validation, estimation of the intensity function and estimation of the mean time between failures for exponential process through regression.

Controlling faults in software requires that one can predict problems early enough to take preventive action. In the past 35 years more than 100 software reliability models have been developed to solve reliability models (Lyu, 2007). Most of these models as the models of software reliability growth depend on a certain a priori assumptions about the nature of software faults and the stochastic behavior of software process (Cai et al., 1991). As a result, different models have different predictive performance at different testing phases across various projects. A single universal model that can provide highly accurate predictions under all circumstances without any assumptions is most desirable (Karunanithi, 1992).
Regression models are the most popular methods for building models and are used to calibrate almost all of the models (Alj ahdali and Buragga, 2007). Regression models are one of the proposed models to predict the number of faults in the software as shown by Aljahdali. Least square estimation is the best technique used to estimate linear models as observed in the literature.

1.10 FAILURE DATA ANALYSIS

In software reliability modeling, depending on the format in which test data are available, there are two common types of failure data: time-domain (i.e. ungrouped) data and interval-domain (i.e. grouped) data (Pham, 2006). These data are usually used by practitioners when analyzing, assessing and predicting reliability applications.

Type I Data---inter-failure time data, i.e. the time between two failures. In this type of data, failure data are collected chronologically and the occurrence times are recorded as \(0 \leq t_1 \leq t_2 \leq \cdots \leq t_n\), which are the realizations of the random variables \(0 \leq T_1 \leq T_2 \leq \cdots \leq T_n\). The inter-failure time is given by \(x_i = t_i - t_{i-1}\). In some literature, this type of data is also called as ungrouped data;

Type II Data---cumulative count of failures per interval, i.e. in this type of data, a period of failure observing time is divided into a set of sub-intervals, for instance \((0, t_1), (t_1, t_2) \cdots (t_{i-1}, t_i)\), and the total number of failures in the above each sub-interval is recorded as \(n_1, n_2, \cdots n_i\). In some literature, this type of data is also called as grouped data.

In this research as Type I data is used, it is well known that for a NHPP the joint distribution of the failure times of \(n\) failures that have occurred during the period of \((0, t_n)\) has the same distribution as the order statistics of a random sample with size \(n\), therefore the log likelihood function takes on the following form (Pham, 2006):

\[
L(\Theta, t_1, t_2, \cdots t_n) = e^{-m(t_n)} \prod_{i=1}^{n} \lambda(t_i)
\]

\[
\ln L(\Theta, t_1, t_2, \cdots t_n) = \sum_{i=1}^{n} \ln \left[ \lambda(t_i) \right] - m(t_n)
\]
More discussions regarding the MLE issue can be found in Knafl (1992), Hossain and Dahiya (1993), Zhao and Xie (1996) and Knafl and Morgan (1996).

The estimation of model’s parameters is an important issue too. The maximum likelihood estimation technique is most commonly used in the estimation of parameters of the NHPP models. With respect to the model’s parameters, the partial derivatives of the parameters respectively are taken. Let the equations equal to zero, solve the equations and obtain the values of the parameters.