ESTIMATION OF HIGHWAY PROJECT DURATION AT THE PLANNING STAGE AND ANALYSIS OF RISK FACTORS LEADING TO TIME OVERRUN

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science
In
Transportation Engineering

Submitted By

SIDRA KALEEM

(2011 – NUST – MS PhD – TN – 14)

NATIONAL INSTITUTE OF TRANSPORTATION (NIT)
SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING (SCEE)
NATIONAL UNIVERSITY OF SCIENCES & TECHNOLOGY (NUST)
SECTOR H-12, ISLAMABAD, PAKISTAN

(2014)

This is to certify that the
thesis entitled

ESTIMATION OF HIGHWAY PROJECT DURATION AT THE PLANNING STAGE AND ANALYSIS OF RISK FACTORS LEADING TO TIME OVERRUN

Submitted by

SIDRA KALEEM

Has been accepted towards the partial fulfillment
of the
requirements for
Master of Science in Transportation Engineering

____

Associate Professor (Dr. Muhammad Irfan, PhD)
Department of Transportation Engineering
National Institute of Transportation
School of Civil and Environmental Engineering
National University of Science& Technology, Islamabad

ESTIMATION OF HIGHWAY PROJECT DURATION AT THE PLANNING STAGE AND ANALYSIS OF RISK FACTORS LEADING TO TIME OVERRUN

by

SIDRA KALEEM

(2011 – NUST – MS PhD – TN – 14)

A Thesis

of

Master of Science

Submitted to

Department of Transportation Engineering

National Institute of Transportation

School of Civil and Environmental Engineering

National University of Science and Technology

Islamabad

In partial fulfillment of the requirements for the degree of

Master of Science Transportation Engineering

2014

DISCLAIMER

The contents of this research work purely reflect the views and opinion of the author and their results and findings on the basis of data modeling. This thesis does not constitute any standard, regulation, bench mark, principle, reference or criterion.

Every effort has been made to make this research thesis complete and accurate. However, there may be mistakes in the content. Therefore, this work should be used only as source of information and guidelines but in no case as an ultimate source of information.

Broader publication of research findings can be limited by the use and publication of confidential project data. This could be of concern to organizations that have furnished the information. The data compiled for this research was derived from publicly sourced documents, as most of the project information is also available on NHA's website. There are no ethical considerations required in publishing the data and its subsequent analysis in this report

The author shall have neither liability nor any responsibility to any person or entity with respect to any loss or damage caused, or alleged to have been caused, directly or indirectly, by the information contained in the thesis.

To My Parents

Muhammad Kaleem & Munawar Kaleem

ACKNOWLEDGEMENT

"and Allah is sufficient as a friend, and Allah is sufficient as a helper"

(4:45)

Firstly I would want to thank Almighty Allah for He bestowed me with the opportunity to complete my Masters from one of the best universities of Pakistan. The faculty at NIT is top-notch, I could not have asked for a better program. I am so honored to be a part of this finest institution.

I am indebted to Mansoor Ahmed Malik, who kicked off my MS journey by opening to me the doors of NIT that were beyond my accessibility; words cannot express my gratitude for his numerous time to time favors over me. His belief in me and his support was vital for me to realize my potential.

I am really grateful and fortunate to have Dr. Muhammad Irfan as my supervisor. He remained there with me no matter what hiccups I encountered or how slowly I went. I would especially like to thank Dr. Irfan for guiding me through the entire research and providing prompt and helpful comments and feedback. He brings a rare combination of depth, pragmatism and vision to his work that has inspired me persistently. I have been lucky that he sanctioned me much freedom and encouraged me to pursue and explore ideas that I presumed interesting.

The research would not have been possible without the access to the NHA data that Assistant Director Mr. Ans Farrukh and Assistant Director Mr. Baber Khan provided me. My gratitude and appreciation for their unwavering support is indeed overwhelming. I would also like to extend heartfelt praise to my colleagues and friends Mr. Murtaza Mehmood, Mr. Atif Hameed, Mr. Arsalan Ahmed and Ms. Ramla Qureshi, who read drafts without complaining, and furnished feedback on my writing that was invaluable in shaping of this report and my research paper.

My stay at NUST for my MS was a beautiful journey and I have a network of close friends who traversed this road with me. I am truly thankful to all my friends and colleagues who patiently listened to my ideas and provided valuable insights. If NUST allowed it, I would list everyone who deserves recognition. My good friend Sadia Bashir has been a constant boosting source through times of self-doubt and a sounding board for a number of ideas. My "GIS-Genius" and partner-in-crime roommate Misbah Sehrish made my stay at NUST hostels an absolutely exhilarating experience. I will really miss those fun times we cherished together. I am also indebted to Madiha Saeed for her hospitality, friendship, support and love. I have no words to thank her for her generosity and untiring friendship.

Finally, I would like to thank Mrs. Munawar Kaleem, my mother, and Mr. Muhammad Kaleem, my father, for the sacrifices they made over the years to give me the opportunity to follow my passions and get me where I stand today. They encouraged me to go pursue my dreams at the other side of the country when I had never stepped out of Karachi alone even for a day before. I deem my accomplishments to this point majorly the product of their parenting.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Estimation of time duration of highway projects at the planning stage serves as a vital input for construction planning, scheduling and contract administration. However time overrun, resulting from various factors is the most cardinal issue which eventually leads to cost overrun and hence induces turbulence in the initial cost and time estimates. In this research, highway project duration is estimated on the basis of variables such as planned cost and project type which are known at the planning phase. Data comprises of project types such as pavement construction, improvement, rehabilitation and bridge construction projects of National Highway Authority, Pakistan. A mathematical relationship between highway project duration, planned cost and project type is demonstrated in this research by using various model specifications. Furthermore using multivariate regression analysis correlation of the time overrun with potential risk factors is investigated encompassing attributes such as project type, cost and geographical location. Probability plots are also generated by survivor function in log logistic analysis which provide the likelihood of the project duration being equal or greater than some specified duration. The research identifies a number of significant risk variables and their severity that contributes to the extensive delays and consequently exceeds the planned time estimates. Late funds release from to the funding agencies, land acquisition problems and cash flow problems within NHA were the chief time overrun factors prevailing in most of project across the country. The models developed can assist the project administrators in determining improved estimates of project duration and enhancing the expected delay estimation in completion time of planned projects.

Chapter 1

INTRODUCTION

1.1 Background and Problem Statement

Realistic and precise planning helps to derive maximum benefits out of investments. The key goal of every project is to achieve work completion on time and within the specified budget. Transformation of paper drawings into concrete form while ensuring quality and safety is indeed a daunting task. As time is the essence of every project, development of reliable duration estimates can help agencies to deliver optimum project schedules and thus avoid issues pertaining to time overruns.

Early and reliable time estimates are essential inputs for decision making in the initial stages of construction projects. Construction project duration is a very important factor for the client, consultant and the contractor, yet delay is a typical phenomenon which is bound to occur as construction projects are seldom on schedule, often delays are among the most critical construction disputes.

Government of Pakistan has invested a heavy amount for the infrastructure development. Many development partners like Asian Development Bank and World Bank have also supported the infrastructure projects in the country by committing huge sums of money in this regard. Despite the enormous importance of infrastructure and big sum of dollars committed to it, success rate of most of the highway projects in Pakistan is relatively unsatisfactory; the commonality among the projects is time overrun and cost overrun. These two reasons can be attributed by various risk factors arising at every stage of the project development. It's known worldwide that economic sustainability depends largely upon the expansion of the existing facilities along with their modernization to meet the growing needs. For the rapid socio-economic uplift government of Pakistan desires to execute maximum public development projects in short span of time. 75% to 80% of Pakistan's total commercial traffic is carried by National Highway Network and Motorway system. NHA's official website reports that two-third of the total road network is in relatively poor condition. This scenario indicates a number of upcoming rehabilitation, construction and improvement projects in the pipeline.

Early studies have proposed that only the completion of the project marks its true duration though the final duration is adversely affected by some aberrations. These aberrations could include extreme weather conditions, financial delays, skilled labor scarcity, political situations, force majeure and other project related changes taking place at various phases of the project's life cycle. Nevertheless information about probability of occurrence of a certain problem can help the project planners in various aspects. They can act pro-actively and prepare a contingency plan keeping the historical risk factors under consideration. Planning being one of the major chunks of work sets the milestones for design and execution phases of the projects. The cardinal objective of every project is to achieve work completion on time and within the specified budget while ensuring no compromise on quality and safety. Several studies have been carried out regarding the estimation of project duration and evaluation of risk factors in highway projects. This study aims to add to the body of knowledge by estimating the project duration and identifying the potential risk factors which affect the highway projects and ultimately result into time overrun. The developed models can also act as empirical tools for the contractors who may find it useful for making appropriate project plans for equipment mobilization, material utilization and resource optimization.

1.2 Research Objectives

The aim of the present study is to propose statistical models for the estimation of project durations. The study also seeks to identify the distribution of the time delays and propose model for estimation of project delays. In the course of such estimations and investigations, "Risk Factors" in NHA projects in the four provinces of Pakistan are also to be identified. Reasons and the responsibilities for time delays are to be pinpointed by collecting, reviewing, processing and analyzing historic data. The analysis is undertaken to ascertain if the findings can lead to more accurate construction duration estimates of highway projects.

Comparison of the extent and causes of the time delay problem of NHA projects in four provinces of Pakistan is also to be performed. Moreover, the study also tends to provide recommendations for addressing the present high risk situation so that the prevailing conditions can be rectified.

1.3 Scope of the Thesis

To address the research objective, highway project data comprised of National Highway Authority Pakistan projects. The data collected was from four different provinces of Pakistan: Sindh, Punjab, Balochistan and Khyber Pukhtunkhwa. Total of 120 projects over financial years 2001-2012 were selected for estimation of highway project duration and for the analysis of potential risk factors. Projects costs were rebased to 2012 project prices. The data was related to four different project types: (1) Pavement construction; (2) Pavement rehabilitation; (3) Pavement improvement; (4) Bridge construction.

1.4 Overview of Study Approach

In order to achieve the stated objectives this research follows a sequence of activities. This present study, structured on the findings of past research, seeks to estimate project duration by first describing the time duration data using more traditional functional forms and modeling techniques. The research goes further to provide new insight into the potential risk factors affecting time overrun through examining past incidents.

This dissertation first describes project duration in terms of explanatory variable using traditional linear form. Separate linear models are also formed for different project types. Cognizance of past studies is taken into account to further investigate the project duration. Weibull analysis considered as a robust technique is used to yield survival curves and hazard functions. Survival analysis is used to model the time taken for an event to take place. Time taken to project completion is sought to be modeled in this research. Using historic data correlation between highway project delays and different types of projects is also calculated. This framework can be utilized to develop optimal highway duration estimates. The framework is developed using a case study of NHA projects. NHA not only needs to address delayed project delivery issues but also must scrutinize the types of delayed projects that hinder efficient programming. In an effort to address these issues, an understanding of the characteristics of projects correlating with the problems causing delay, can permit NHA to increase the accuracy of project delivery. The research concludes with a summary of the findings, its contributions and recommendations to cater future projects.

1.5 Organization of the Thesis

The thesis is organized in six chapters. Chapter 1 provides a background and the extent of the duration spillage problem and the need to develop the time and delay estimates. Chapter 2 covers literature review, it provides an introduction to the project duration estimation, along with risks associated with project delivery. Chapter 3 covers Research Methodology used in the research while chapter 4 covers data collection collation and descriptive statistics. Chapter 5 covers data modeling and analysis of the results. Lastly in chapter 6 research summary, conclusions and recommendations for future projects are presented.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Several studies on the nature of relationship between project duration and project cost annotate shifts in their underlying philosophy. This chapter is a review of the researches and studies already carried in the past related to project duration estimation and project related risk factors. In the first phase, studies related to time and cost relationship are included and in the second place, studies related to delay causing factors are discussed. The change in project duration occurs as a result of many related factors all of which are associated with some form of risk. The project management could be sometimes inefficient and take an extended time and so often sets up opposing stances between the project participants which can therefore eventually compromise the measures of success of a project in terms of time, budget and technical performance. Similarly, the main barriers to achieving project success are the changes in the project environment. Problem multiplies with the size of the project as uncertainty in project outcomes increase. Huge scale construction projects are exposed to uncertain environments because of such factors as planning, design and construction complexity. In addition, the presence of numerous interest groups (such as the project owner, consultants and contractors) as well as resources (such as materials, equipment, project funding, climatic, economic and political environment and statutory regulations) all add to project uncertainty. Other factors contributing to uncertainty include the complexity of the project, the speed of its construction, the location of the project, and its degree of unfamiliarity.

Analysis of the reasons for project time overrun of construction projects is a necessary step for the improvement of any given time estimating system and can be used to pinpoint areas where the greatest improvement can be obtained. As part of this process, this chapter identifies previous literature on the subject of estimation of project duration and risk factors leading to time delays.

2.2 Time and Cost Relationships

Projects generally surrounds large, expensive, unique or high-risk undertakings that have to be completed by a certain date, within a certain amount of money, and deliver some expected or anticipated level of performance. These three criteria of success have become widely used. It captures the major task of the project manager, and their essential trade-offs. Project duration is essentially needed for proper project planning and contract administration. A number of studies have blazed the trail from the modeling techniques perspective, for examining the issue of cost overrun and time delays. The problem of time overrun in construction project was studied (Knight and Fayek 2002; Shaheen et al. 2007). Time and cost deviations were also investigated by Zheng and Ng (2005). Prior knowledge of the project expected duration can be useful in bid evaluation and life cycle cost analysis (Irfan et al. 2010). Considerable reliance on engineering judgment is made while a project is planned (Hendrickson et al. 1987). Salapatas and Sawle (1986) define success to have been achieved only when three groups perceive success: the client (based on performance, budget and reputation), the contractor (based on profitability, reputation, client and public satisfaction) and the customer/public (based on environment, reliability and cost). Potter (1987) has found from experience that success and failure can in fact be very close and Sykes (1982) supports this by pointing out that many large projects have been saved from disaster only because of fortuitous circumstances. Schedule is most important in early stages of the project, but during the project it cost becomes most important and after the project only technical performance is remembered.

Early studies assume a linear relationship between project duration and project cost (Fulkerson 1961) but subsequent studies showed flexibility by using variety of nonlinear mathematical functions that include discrete formulations (Skutella 1998; Zheng et al. 2004) convex (Foldes and Soumis 1993), concave (Falk and Horowitz 1972), hybrid of convex and concave (Moder et al., 1995) or quadratic (Deckro et al. 1995). Hierarchical rule based activity duration models were estimated by Hendrickson et al (1987). Chan (2001) carried out a study in Malaysia to estimate average project duration using a time-cost formula expressed as Duration= K x $Cost^B$, where K represents the characteristic of duration performance and B is the indicative constant of sensitivity of time performance to cost level. The possibility of having piecewise discontinuous activity time cost function has also been explained in recent past studies (Moussourakis and Haksever 2004; Yang 2005). Weibull functional form has been used for the analysis to describe the relationship between project cost and duration (Nassar et al. 2005), and contract type and project duration (Anastasopoulos 2007). Several other studies have not only sought out a relationship between cost and duration but have also proceeded using linear and integer programming techniques to investigate the trade-offs between project duration and cost (Chassiakos and Sakellaropoulous 2005). Optimization algorithm was used to develop a time-cost profile considering various mathematical forms (Yang 2007).

Irfan et al. (2011) sought project duration models as a function of the project cost, type, and contract type. Log-linear logistic models and log-linear functional form is used to develop survivor and hazard models. Analyzing the data by linear regression mathematical models subsequently determined that linear forms could be used under certain conditions, like while accounting for the unique character of the empirical project data, and the restriction of Least- Square Estimation (LSE) techniques to incorporate certain project assumptions (Hosmer and Lemeshow 1999). When ordinary least square (OLS) techniques are used certain variables that are not represented by traditional explanatory variables could cause irreducible random noise (Hendrickson et al. 1987). Concept of earned value project management has also been applied by the researchers to predict the project duration (Vandevoorde and Vanhoucke 2006; Lipke et al. 2009). Table 2.1 summarizes the past studies over time-cost relationship.

2.3 Causes of Time Overruns

Ahmed et al (2003) regards delay as universal phenomenon which is usually accompanied by cost overrun. Dias and Ioannou (1995) concluded two types of risk; 1) Pure risk and 2) Speculative risk. Former risk exists when there is the possibility of financial loss but no possibility of financial gain (e.g. physical damages) later involves the possibility of both gains and losses (i.e. financial and production risk). All construction projects by their nature are economically risky undertakings. Risk is termed as an uncertain condition or event which if occurs, causes significant positive or negative effects (Project Management Institute, 2008b). Uncertain situations are characterized by the risk where actual outcome of an event or activity is deviated from the planned value

(Raftery 1994). Kwak and Stoddard (2004) termed identification of risks as the most crucial activity. Risk response measures need to be adapted to prevent the identified risk from materializing (Ropponen and Lyytinen 1997). Project duration and cost is dynamically affected by many variables at the execution stage (del Caño and de la Cruz 2002). Pakkala (2002) emphasized that better practices should be provided to ensure quicker project completion time and cost effective solutions to the owner, since he is most vulnerable to the design and construction risks.

Ibbs and Allen (1995) quantified the project changes impacts on engineering and construction project performance and concluded change as an event, which results in modification of original scope, execution time and cost of work. The problem remains the same that the future is not always predictable. Factor analysis technique was used to identify variable affecting construction time and cost overrun in Indonesia by grouping time and cost overrun variable into factors and then their relationship was determined, the study identified main causes of time delay as inadequate planning, design change and poor labor productivity (Kaming et al. 1997). According to Kaming et al. (1997) the results were specific to Indonesia but they reflected construction management problem in the developing countries. Chan and Kumaraswamy (1997) determined the significant factors causing time delays in Hong Kong and evaluated their relative importance. Their research stated poor supervision, poor site management, poor decision making, unexpected ground conditions and client initiated variations as the major causes of delay. Lo et al. (2006) found the distribution of construction delays in Hon Kong. Studies carried out in Ghana indicated time and cost overruns are related to poor contractor management, material procurement, material and cost escalation, poor technical performance and payment difficulties from agencies (Frimpong et al. 2003). Kaliba et al. (2003) identified that the duration of road construction projects in Zambia is influenced by economic problem, contract modification, material procurement, delayed payment, change in specification and drawings, construction mistakes and poor supervision and coordination on site. Similar studies have been conducted worldwide however no such problem occurs at the planning phase and the root causes for delay are unknown when a project is planned Hence adverse effects of potential risks can be minimized.

A study was conducted by Ellis and Thomas [2002] to investigate the root causes of delays in highway construction. In their study, it was found that 31% to 55% of all highway projects experience an average time delay of 44% in excess of their original contract periods. It was observed that time delays occur more frequently for contracts in urban areas. The focus of that study was to identify the root causes of delays (not only for the apparent causes). A root cause is distinguished from an apparent cause by determining if the cause violated a fundamental principle and if the cause is known or developed in sufficient detail to allow corrective action to be taken. For example, an apparent cause may be plan errors; however, an in-depth investigation may ultimately determine that the root cause was a violation of the "time-cost" principle leading to easily recognizable mistakes. Generally, apparent causes are relatively many while root causes are relatively few in number. According to the authors, the main root causes of delays include business practices, procedures, utilities, unforeseen site conditions, contractor and State Highway Agencies management of scheduling and planning, maintenance of traffic work zones, and design errors and omissions. One of the major causes of business failures is related to the client. Client-generated risk factors can be stated as a client's financial ability to meet the cost of the work, its claims record, changing needs, and the construction sophistication. In turn, these risks can put a strain on the contractor's cash flow and can increase the actual cost of a project during construction.

This study seeks to predict duration for various project types and identify the major risk factors for time overrun dominant in the highway projects of Pakistan. Development of reliable duration and delay estimates can help agencies to deliver optimum project schedules and thus avoid issues pertaining to time overruns that result in cost escalation.

2.4 Discussion of the Literature Review

Inaccurate planning can cause disruption of fiscal planning by both overestimated and underestimated project planning schedules. Projects that are not let at the expected time usually incur either additional expenses causing a deficit in allotted funds or inhibit the programming of additional projects, possibly causing available resources to squander. A wide spectrum of possible methods, including Linear, Log Linear, weibull etc, have been presented in the past research to estimate the duration of projects. Therefore, the literature review first enabled to identify the key technique for the present study. Irfan et al. (2011), for example, provided some detail information on time and cost relationship and demonstrated duration models using statistical and probabilistic modeling. Table 2.1 summarizes the major elements of the time-cost relationship literature review and indicates how the findings of the review relate to the framework of the present study. A number of delayed factors were also discussed in the literature review that; Impairs the efficient use of allotted funds; shifts current and future project programming; disturbs the letting schedule for construction bidding. Agencies are unable to accommodate projects that spill over into current schedules due to programming shifts therefore users incur increased costs in reference to traffic, route change, and increased travel time. Defective highways also place user safety at risk. The information in the present chapter provides an indication of issues faced by transportation agencies in managing time delays.

Table 2.1 Time-Cost Modeling past Findings Summary

Study	Year of Study	Modeling Technique
Fulkerson	1961	Linear Relationship
Falk and Horowitz	1972	Concave
Hendrickson et al.	1987	Hierarchical Rule-Based Activity Duration Models
Hendrickson et al.	1987	Ordinary Least Square (OLS)
Foldes and Soumis	1993	Convex
Moder et al.	1995	Hybrid Of Concave And Convex
Deckro et al.	1995	Quadratic
Skutella	1998	Discrete Formulations
Hosmer and lemeshow	1999	Least-Square Estimation (LSE)
Chan	2001	Duration = K X Cost ^{B}
Zheng et al.	2004	Discrete Formulations
Moussourakis and Haksever	2004	Piecewise Discontinuous Activity Time–Cost Functions

Yang	2005	Piecewise Discontinuous Activity Time–Cost Functions
Nassar et al.	2005	Weibull Functional Forms
Chassiakos and Sakellaropoulos	2005	Linear and Integer Programming Techniques
Vandevoorde and Vanhoucke	2006	Earned Value Project Management
Anastasopoulos	2007	Weibull functional forms
Yang	2007	optimization algoritham
Lipke et al.	2009	Earned Value Project Management
Irfan et al.	2011	Log-Linear Logistic Models & Log-Linear Functional Forms
A. Czarnigowska et al.	2013	Synthesis of past studies of time-cost relationship

2.5 Chapter Summary

The literature review chapter covers the time-cost relationship and causes of cost overruns. Further, from literature, the problems of time delay was identified and gave an insight about the past findings of project duration estimation using various mathematical techniques. Some new and external points of view and definitions of the key concepts were also acquired. Previous studies identified some factors that influence time delays and developed tools that help address such problems. Similarly previous literature also provided a framework to estimate duration of projects keeping cost as a primary predictor variable. Various modeling techniques for time-cost used by researchers are also tabulated which provide basis for the appropriate selection of model technique for this research study.

Chapter 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter explains the overall framework, methods and underlying assumptions for estimating the project duration and analyzing the problem of time delays in NHA projects. The methodology includes preliminary descriptive statistics that examines the general temporal and spatial trends in the data. The methodologies include definitions of dependent variables (time overrun, project duration) and potential influential factor (independent variables; geographical location, risk factors), and selection of model categories and appropriate mathematical forms. The methodology was designed to yield statistical models with a view to predict project duration and time overrun, but more importantly, to identify significant factors that influence project duration and time delays.

Minitab and PASW-17 software are used for the analysis of collected data. The risk collected data is checked by performing a major test for further analysis; this include normality test. Shapiro-Wilk normality test was performed to find out the parametric or non-parametric nature of data. A 5% level of significance is considered to represent statistically significant relationships in the data. Ranking of the risks is performed using Relative Importance Index (RII) method. Multivariate regression analysis is performed to estimate the delay duration due to risk factors.

When the problem under investigation is of more recent historical origin, then data and facts can be available but may not necessarily be collected in the form needed in order to describe and understand the problem (Bennett, 1991). There are two ways to look at the historical approach to a research problem: Data is collected to describe the field at a particular point in time (referred to as cross-sectional study), or the development of the problem is described over a period of time (longitudinal historical study).

Stone (1978) on the other hand points out survey as a strategy that allows a researcher to collect data directly from sources in a systematic fashion but it includes some disadvantages as people refuse to respond to survey probes, because of suspicion or other resistance. Moreover, most surveys have limited capacity to generate data to analyze casual connections among variables and survey could be an extremely expensive research strategy because of administrative and other costs. In addition, surveys may have low response rates. This study seeks to model the historical data, to yield reliable estimates to cater the duration and time overrun problem.

3.2 Framework of Research

The procedures to conduct a research in societal sciences comprises of modeling, analysis, experiments, case studies etc. The selection of the procedure for specific study depends upon the requirements of that study, extent of research and category of research function i.e. how, why, what, focal point of research and control over variables (Yin J, 2006). While choosing an appropriate method for a study, it is mandatory to think the associations between the collection of data and its analysis, also the major questions to be addressed, and the consequences. The objectives of the research have been established in the first chapter. The procedures that can be followed for achieving objectives of the study are elaborated here.

The main source of data was the historic project records of National Highway Authority of Pakistan. Basic Statistics was performed on the data to check the trend of its distribution. Risk factors were identified. Parametric and non-parametric tests were conducted for their spatial distribution. Statistical and probabilistic modeling was carried out for the project duration estimation. In the end multivariate regression analysis was performed to cater delay duration because of different risk factors. Figure 3.1 presents a summary of the study framework adopted in this research.

3.3 Descriptive Statistics

Two software's are used for the analysis of collected data, these are Minitab and PASW-17. Level of significance followed is $\alpha = 0.05$. Identification of project duration and analysis of risk factors that influence the planned time estimates is the paramount objective of this research, and it is expected that a descriptive statistical analysis would throw more light on this issue. For project duration and risk factors leading to time delays, descriptive statistical analysis in terms of their frequency and amounts was carried out. Descriptive graphs (histograms) show any variations in such attributes by geographical location, type of project or project cost. On the other hand, descriptive figures such as pie charts easily and readily show the relative significance of various categories of time overruns.

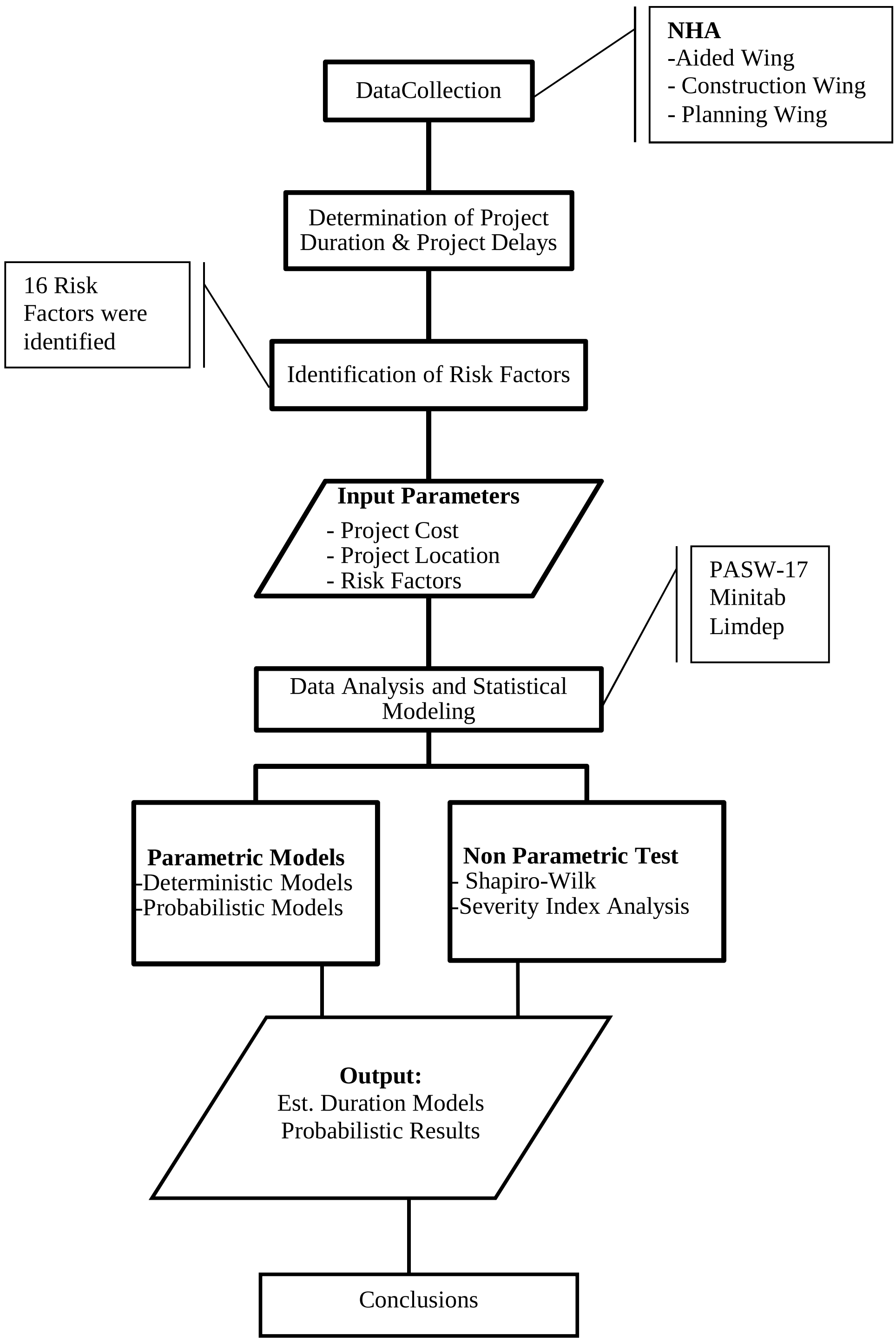


Figure 3.1 Framework of Research

3.4 Description of the Modeling Proces

The next step after examining statistical characteristics was to develop models to confirm the magnitude and direction of the influence of potential factors and to predict the project duration and time delay of any future projects.

3.4.1 Independent Variables

The first step was to identify the independent variables to be used in the statistical models. These were as follows:

Type of Projects: Four types of projects (pavement construction, pavement improvement, pavement rehabilitation and bridge construction) were selected for project duration estimation.

Project Cost: This explanatory variable depicts the size of the project and is measured in terms of total cost. Jahren and Ashe (1990) indicated size of the project a very significant predictor for time delay. Large projects are usually assumed to have greater duration. Involvement of huge number of contractors and subcontractors in large projects often leads to lapses in communication between them, thus makes them prone to longer delays.

Risk Factors: A number of time overrun factors were identified; a common time overrun variable was recorded for common time overrun factors across the projects. These variables, their frequencies and symbols are represented in Table 4.3. Sixteen risk variables were identified and were considered in multivariate delay analysis.

Geographical Location: The province at which a project is located may be a significant variable because of variations in law and order situation and administrative practices/culture. Project locations by provinces were represented using binary variables.

Extent of geographic area is identified as an important factor for competitive bidding in building projects (Dew and Skitmore 1992). Construction cost and time duration are usually observed specific to the geographic areas. The data collected was from four different provinces of Pakistan: Sindh, Punjab, Balochistan and Khyber Pukhtunkhwa. Provincial data was split in the group of two depending upon the strong relationship between the risk factors and other attributes like project cost that could lead to greater time duration and time delays. Sindh and Punjab were placed in group 1 while

Balochistan and Khyber Pukhtunkhwa were stationed in group 2. Two binary variables were created for each of these groups. If the information pertained to the location in the group the variable takes the value of 1 otherwise 0 is inserted.

3.4.2 Response Variables

Project duration is the initial length (days) of the project, computed as the difference between the estimated last day of work and the notice to proceed date. The project duration can be of the following types:

Planned Duration: This explanatory variable is measured as the length in calendar days allocated on the project at the preconstruction phase.

Actual Duration: It is the length in calendar days of the project estimated as a difference between the last day of work and stipulated start date of the project. An initial expectation for this variable would be that longer projects would result in longer delays. Rowland [1981] indeed found that the size of the project is a significant variable. Size can be understood as the total cost or duration of the project. It is intuitive that a high-cost project will likely involve long project duration. Conversely, a project of long duration is likely to have a high cost. So, these two potential influential variables are obviously related to each other. In the present study, the actual project length was used only for time duration models because this is the most relevant variable for these models.

Delay Duration: It is taken as the difference between the planned duration and actual duration of the project.

3.4.3 Investigation and Selection of Mathematical Forms

This terminal stage of the research process included the examination into statistical models which can analyze the correlation between relationships pertinent to project duration and time overrun in construction projects of highways.

In this research, the term model that can be used in many ways (Emory, 1980), is attributed to the dynamic framework, that assists to portray major concepts and propositions of the research. The model developed at the start of the research can be extremely conceptual or theoretical; it is then tested through the process of reasoning, data gathering and analysis. On the contrary, a model can be generated at the end of research, (Bennett, 1991).

Decision making process is valued by the results of analysis of various models which makes significant contribution to this activity. Model development also trades off the complexity to provide an acceptable picture of the system under consideration. The main idea of this study was not only to yield a model but it was also a major concern to produce a model that can be practically used by the stake holders with sufficient credibility and acceptance. Moreover, a robust statistical technique was required.

The method of analysis techniques and quantification are interdependent, available information is one factor that restricts them. Techniques like network-based solutions involve great quantity of information which is not always readily available. Simulation and sensitivity analysis use Monte Carlo analysis in order to determine the significance of individual causes of project related risk and several other factors.

Data mining comes as a handy and powerful research tool for pre-processing structured information from the data available and applying statistical analysis to crack patterns and relationships recondite in project databases. It is often claimed that high value out of repositories of information is compressed by data mining. Multivariate statistical analysis comes into picture where it is required to infer from multiple sets of performance evaluation on a number of individuals of objects. It is a scientific inference and widely used in analytical work. The success of such inferences has been confirmed by the history of science. It is capable of handling inferred reality and can also reduce the number of the variables. Sole reliance on classical analysis methods has been abandoned by the researchers in sciences. Each project in the construction industry is unique and different techniques are required to overcome major complexities. Therefore construction industry has been seen as an apt example.

Two functions can be derived with multivariate statistical techniques; these correspond to the characteristics between inferential and descriptive statistics. Usually no assumptions are required for descriptive statistics; however multivariate significance tests are based on normality and homogeneity assumptions.

In order to determine relationships between projects, project risks and project time overrun Multivariate regression technique is the most powerful tool to administer multiple project variables in the development of a model. In past Multivariate regression has been the most widely used method of modelling construction costs. Conditional expectation of a random variable provided other random variables estimates Multivariate regression. It also involves contributing model solutions which can be utilized as a foundation for decision making.

A linear relationship was observed between dependent and independent variables of the project over preliminary data analysis. This property gave an indication to use multivariate regression analysis as an ideal tool for examination of project variables in the available data. Several other reasons were also considered for taking up linear statistical models: (i) These models are not complex ad can be easily used by the site personnel and site management &; (ii) Scatter diagram showed linear trends by visual data inspection in project duration and cost data.

The multivariate statistical analysis method is formulated for distribution of inferential and descriptive techniques that can evaluate sets of variables. Data can be summarized and nature and strength of the relationships can be qualified among the variables by Regression analysis. Regression analysis can predict new values of dependent variables based on observed values. Regression Analysis is used to evaluate the degree of strength of relationship between the response and the explanatory variable.

There are several techniques used for regression, this research has adopted the method of least squares. The following decisive factors were taken into consideration while developing the multivariate regression: (1) Problem formulation (2) Adequate, high quality project data (3) Selection of appropriate project variables.

Problem formulation: It was difficult to study the relationship of several projects and time overrun variables because a number of factors combinations needed to be analyzed and required knowledge is complex in nature. It is impractical and daunting task to include too many parameters in the model. It was therefore necessary to identify and pin point the key factors and amplify the factor list that were to be included in the model taking into consideration the complexities in construction. The pivot point of this part thesis was a decision-oriented model in the highway projects, A model that defines the correlation between project time overrun and the number of variables associated with the projects.

Adequate, high quality data: Two type of data was present in the project i-e qualitative and quantitative. Quantitative data characterized the quantity or amount of a component. i-e, the dependent variable of the proposed model (Project Duration) is a quantitative variable. On the contrary, qualitative (or categorical) data, can be project delivery method, geographical location, or the reasons for the time overrun.

Selection of appropriate variables: The basic purpose for this process was identifying any correlating project variables to the project duration and delay. Any closely related variable can help to produce more accurate and realistic models for time overruns. They can also pin point the areas of concern to the estimators so they can pay extra attention while planning in the presence of those particular variable which can extend from project location, project cost etc.

It is rather difficult to handle different forms of variable relationships in regression analysis. It is not always suitable to assume a linear relationship between the dependent and independent variables. This fact can be made obvious by plotting scatter plots of those variables from the sample. In this thesis it is assumed that the relationship between the variables is linear i-e relation between independent variables and Y is linear.

In order to investigate the historical data multivariate regression analysis was applied and time overrun models were developed in terms of a regression coefficient for each explanatory variable. The multiple regression took the following form of the in terms of model:

$$Y = a_o + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n + e \qquad (3\text{-}1)$$

Here Y is a dependent variable (i.e. % or project duration), X_n is independent variables, a_0 is a constant indicating Y intersect, b_n are partial regression coefficients, and e is the error term.

The main idea of the model was to drive the value of predictable quantity $Y = f(X)$ in form of a set of quantities X = (X1, X2, X3,….Xn.), and relationship (f) between X and Y. The function f was initially established on the primary assumptions developed from the analysts experience in construction of highways regarding the data being analyzed. The efficiency of the multiple regression models was determined by examination of residuals and the value of the multiple correlation coefficients.

Regression coefficients are determined by applying multivariate regression analysis to a set of data, one coefficient is obtained for each explanatory variable. The estimated change in the response variable associated with unit change in the corresponding variables is obtained by these coefficients, where other explanatory variables remain constant. The multivariate regression model fit can be determined in number of ways, i-e, examination of residuals or by calculation of the multiple correlation coefficients.

3.4.4 Regression analysis primary assumptions

The primary assumptions that needed to be meet in the a model using least squares estimation to yield reliable estimates in the regression models, are stated as follows:

1. Linearity Assumption: Relationship between dependent and independent variables is linear
2. Normality of Error: Error values (ε) are normally distributed for any given value of X
3. Homoscedasticity: Probability distribution of the errors has constant variance
4. Independence of Errors: Error values are statistically independent

If any above conditions are not satisfied then there could be a risk that any inferences can be misleading about confidence intervals and tests for significance. In a model it is rather difficult to identify random disturbances as true regression line location cannot be known, so it is therefore devised to check the hypothesized relationship which is the difference between observed value Y an estimated value regression line or simply residuals.

3.4.5 Correlation analysis

To analyze the performance of the models and relationship between the variables this technique is used. Correlation analysis was performed in order to to identify any correlation of project variables with project time overrun. The actual and predicted values are examined on the basis of the correlation coefficient (R). R range from 0 to 1. A higher the correlation value is 1, which shows that actual and predicted values are more correlated.

There are several other criteria that could be used to develop a ranking order in terms of goodness of fit of the regression model, the widely used one is R^2 and adjusted R^2 statistics. R^2 is adopted for this research. R^2 allows direct comparison of the most suitable model identified (Neter et al., 1990). R^2 is usually applied to multiple regression analysis and it is called as coefficient of multiple determination. It is a statistical indicator. The accuracy of a regression model to the accuracy of a trivial benchmark model is compared by this tool.

3.4.6 t-test

It is one of the hypothesis test which is usually appropriate when there is a need to compare the means of two groups or the difference between averages of two populations is investigated. A t-test produces a t-value which is then transformed into probability or p-value.

3.4.7 p-value

Statistical models require the reporting of appropriate F-tests and t-tests in multivariate regression analysis. For each of these test, a p-value is reported. The p-value provides a measure of evidence for the results of tests, for accepting or rejecting the hypothesis.

The extent of evidence against the null hypothesis is often measured by the p-value. Smaller p-value refers to greater evidence for the null hypothesis to be rejected and vice versa. It is often combined with significance level to make decisions on the test hypothesis. For cases where the p-value is less than some set point or threshold (usually 0.05, sometimes even greater e.g. 0.1, or smaller e.g. 0.01) then the null hypothesis (H_0) is rejected.

p-values lesser to 0.01 are regarded highly significant and lesser to 0.05 significant. A greater p-value could be that deviation can be random. If the p-value is less than to a specified targeted value i-e 0.05, 0.1 or.01 then the null hypothesis is rejected. If a p-value is associated with a data set, It is then the measure of the probability that the data set could have from some population as a random sample defined by the statistical model. P-vale is the extent of evidence against the null hypothesis (H_0). Less p-value refers to more evidence against (H_0). P-value can be combined with the significance level to make a decision on a hypothesis.

Under null hypothesis the distribution of p-value is uniform, and therefore does not depend on any statistical test of particular form. The p-value in a statistical test is probability of examining a test statistic to an extent that the value is actually examined, with an assumption of null hypothesis to be true. The value of p is then specified in accordance to the distribution. It can therefore be termed as model-distributional hypothesis instead of null hypothesis. This concludes that if the null is true, then p-value is the probability versus the null in that case. The p-value is identified by observed value.

3.5 Risk Factors Analysis

3.5.1 Risk

Risk is related to some unpredictable events that have the tendency to occur in future. The exact outcome and likelihood is uncertain but it can potentially affect objectives and interests. Risk can also be defined as the chance of an event that will have an impact upon major objectives and is measured in terms of likelihood and consequences (Standards Association of Australia, 1999).

3.5.2 Identification of Risk

Risk Identification is a process of determining different ways and forms in which an uncertain event can take place (Standards Association of Australia 1999). American National Standard (2004) regards risk identification as a process of determining risks that might affect the project and documenting them.

3.5.3 Quantitative and Qualitative Risk Analysis

Qualitative analysis is a process that involves qualitative descriptive scales i-e high, medium and low for the analysis of opportunities and risks whereas quantitative analysis of risks and opportunities involves numerical estimates. Quantitative is normally conducted on risks and opportunities that are identified critical from qualitative analysis.

3.5.4 Test for Normality

Normality test is performed to check the nature of the data is either parametric or non-parametric. The normality tests are very sensitive to the sample size of the variable concerned. Two software's are used for the analysis of collected data, these are MS excel and PASW-17. Level of significance followed is $\alpha = 0.05$. Kolmogorov-Smirnov test is widely used method for data containing more than two thousands values, it is also known as K-S Lilliefors.

Shapiro-Wilk test is performed for the data sets of about two thousands elements or less than two thousands elements. The Significance value should be non-significant, to count as sufficiently normal, it should be greater than 0.05. Therefore for the present study Shapiro-Wilk test is used to test the normality of the data because of the limitation of size of the sample. The significant value of the data was 0.00, which showed that data is not normal, so the data was treated by non-parametric techniques.

3.5.5 Severity Index

One of the non-parametric techniques is relative index ranking, used for the analysis of compiled data. This technique is widely used by construction management researchers to analyze response data by structured questionnaire concerning ordinal measurement of attitudes. Relative index ranking has one form as Severity index analysis that utilized weighted percentage scores to compare the comparative significance of the criteria under study (Elhag and Boussabaine, 1999; Al-Hammad, 2000; Ballal, 2000).

Severity Indices of the risks aid to make the priority choices, the risk with highest severity index is ranked at the top and that with least severity index is ranked at the bottom. Relative important indices of the risks are calculated from five point likert scale; therefore risks are ranked on the basis of frequency of their occurrence. These ranks of the risk factors determined the relative importance of the different risks as per the frequency of occurrence of risks from the historic temporal and spatial data of NHA.

In this research first frequencies of the occurrence of risks are calculated and then these are used to calculate severity indices using formula 3-1 as under

$$\text{Severity Index (I)} = [\sum a_i .\ x_i\ ] \ /\ [\ 5\sum x_i\ ] * 100\% \qquad (3\text{-}2)$$

Where,
x_i = variable expressing frequency of the occurrences

For i

i = 1, 2, 3, 4, 5 as illustrated below

x_5 = frequency of 'very high extend' occurrence; and corresponds to $a_5 = 5$

x_4 = frequency of 'high' occurrence; and corresponds to $a_4 = 4$

x_3 = frequency of 'moderate' occurrence ; and corresponds to $a_3 = 3$

x_2 = frequency of 'low' occurrence; and corresponds to $a_2 = 2$

x_1 = frequency of 'very low occurrence; and corresponds to $a_1 = 1$

3.6 Probabilistic Modeling

Econometric techniques using hazard models are termed suitable for describing the distributions of the period for which a phenomenon lasts.

Probabilities that change over time are generally suited for hazard function analysis. Probability plots generated by survivor function in log logistic analysis provide the likelihood of the project duration being equal or greater than some specified duration

3.6.1 Weibull Analysis

Weibull analysis is a statistical approach which stochastically evaluates a problem. It is a common method for reliability engineering and failure analysis and is used in wide range of engineering applications. It has been mainly used to predict the lifetime of certain event. It entails fitting a Weibull distribution to the collected data.

With parameters $\lambda>0$ and $P>0$, Weibull distribution has the density function:

$$f(t)= \lambda P(\lambda t)P-1 \exp[-(\lambda t)P] \quad (3\text{-}3)$$

Log-linear functional form was used in this research for survival and hazard models. The parameters are $\lambda > 0$ and $P > 0$. Though Weibull exhibits a flexible means of calculating duration dependence but it has a limitation of keeping the hazard to be monotonic (Washington et al., 2003). While log logistic on the other hand caters non monotonic hazard function.

This study has therefore utilized log logistic distribution, describing that the hazard increases in duration till some extent and then it starts to decrease. Limdep statistical software package (Greene, 2007) is being used to produce log-linear logistic modeling process. Probability plots are provided by the survival function of log-logistic and Weibull duration model, $S(t)$ = Prob [$T \geq t$] = exp [$-(\lambda t)^{P}$] which indicates the duration of the project to be equal or greater than the specified duration.. The percentile of the survival distribution is given by the Equation 3-4, provided α is the probability that the project will survive up to time‘t’ or greater:

$$t = [((1.0-\alpha) / \alpha)^{1/P}] / \lambda \quad (3\text{-}4)$$

3.7 Other considerations

Linear relationship is being assumed by Multivariate linear regression between variables. However, this assumption can effectively never be verified in practice. Multivariate regression analysis is not largely affected by small deviations from assumptions of linearity therefore no explicitly allowing for nonlinear components or transformation of variables was considered.

It is not advisable to make predictions at the levels of unobserved variables i-e the values which are not comparable to the observed data. This may result in misleading predictions. Therefore Multivariate regression should not be used for predictions outside explanatory variables range, for example beyond highway projects or the geographic locations of Pakistan.

3.8 Chapter Summary

The present chapter explained the overall framework, methods, and underlying assumptions for analyzing the problem of time overruns, estimation of time duration for NHA projects. The methodology included preliminary descriptive statistics, and statistical and probabilistic modeling. This chapter provides a description of the methodologies used in the statistical and probabilistic analysis. Independent variables i-e project type, project cost, geographical location were also discussed in detail. An in depth overview risk associated with projects is also lime lighted in this chapter. Appropriate functional forms are also being discussed.

This chapter describes the process of investigating and assessing the correlations between project risks, project types, geographical location, time overruns and project duration on highway projects procured within a public highway agency. Aspects of the methodology adopted included: reviewing literature on project risk and project cost overrun, determining and establishing a source of historic project data, recognizing project risk factors, determining highway project types and undertaking statistical and probabilistic modeling to establish correlations between project duration, time overrun elements and project attributes. These research methods were applied to research data consisting NHA highway construction projects. The description of the highway project data collected for the research, the analysis techniques using the methodology and the statistics are described in the following Chapter 4. An overview of the research methodology is presented in figure 3.1.

Chapter 4

DATA COLLECTION, COLLATION AND DESCRIPTIVE STATISTICS

4.1 Introduction

This chapter describes the data collection and the development of the dataset used for the estimation of project duration and analysis of risk factors leading to time delays in NHA projects. To address the research objective, highway project data was collected from National Highway Authority Pakistan. Total of 120 projects, over financial years 2001-2012 were selected for estimation of highway project duration and for the analysis of potential risk factors. Projects costs were rebased to 2012 project prices. The data was related to four different project types: pavement construction, pavement rehabilitation, pavement improvement and bridge construction. This section describes the detail on the data selection, its measurement and how we desire to elucidate it in framework of modeling results.

This chapter explains trends in the dependent variables used in the various models based on the data obtained from 120 projects. It describes time delays in terms of in terms of risk factors associated with projects. A detailed description of the time duration and time overrun trends classified by categories gives an explicit overview of the data. The chapter also provides a detailed description of the proportion of delay factors in various locations of the country. All PKR amounts are in year 2012 PKRs, and inflation factor was considered because the duration of most projects was large.

4.2 Data Collection

NHA staff members were extremely helpful in providing the required information. The information extracted from their monthly reports and presentations included: (1) Contract Name (2) Project type (bridge, maintenance, etc.) (3) Project location (Province) (4) Planned amount (5) Final amount (6) Project start date (7) Planed end date (8) Completion date (9) Project Size.

Figure 4.1 indicates the geographical breakup of NHA road network length across various provinces in Pakistan. The present study takes into account data from four provinces: Sindh, Punjab, Balochistan and Khyber Pukhtunkhwa.

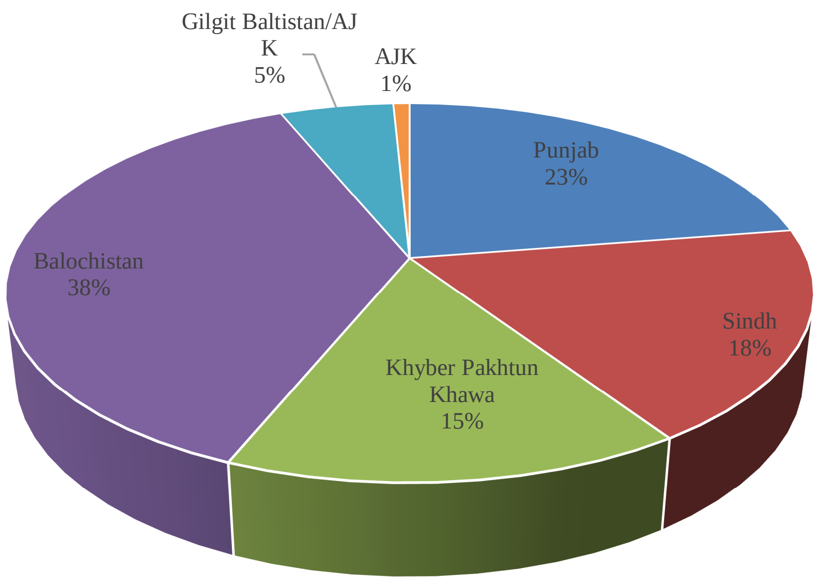


Figure 4.1 Province- wise break up of National Highway

4.3 Database Development

It was ascertained that the presented data was true and factual representations of the National highway authority historic data. A sample of developed data is given in Table 4.1 below. This sample data referred to four types of highway projects completed in the four provinces of Pakistan.

Various columns in table 4.1 contain important project data. This data is described in the following sections.

Table 4.1 Sample of data development

Project Number	Location	Project Type	Project Cost (m)	Commencement Date	Planned Completion	Actual Completion	Delay Duration (Days)	Reason for Delay
S2	Sindh	Bridge	606.80	Feb-09	Nov-09	May-10	181	R1
P25	Punjab	Improvement	653.16	Dec-09	Dec-10	Jun-11	182	R1, R15
B15	Balochistan	Rehabilitation	650.00	Oct-06	Apr-08	Mar-10	699	R8
K1	KPK	Construction	932.00	Aug-10	Dec-10	Dec-12	731	R1

4.3.1 Project Type

The third column of Table 4.1 is comprised of the 'Project Type' and describes the type of highway project constructed for the particular project. This description also indicates whether the project was a road or bridge project. The types of highway projects used in this research are: (1) Pavement construction (2) Pavement rehabilitation (3) Pavement improvement (4) Bridge construction.

4.3.2 Project cost

The project cost shown in the fourth column of Table 4.1 was the final completion cost the project in millions of PKR. This was the project cost at the 'completion' stage of the project. This cost was derived from the overall expenditures spent till the completion of the project and represent all the major activities and acquisition costs of the project, including: (1) Developing the design and conducting investigations.(2) Detailing the design (3) Land acquisition (4) Altering public utility plant (5) Project Execution (5) Project management and handover.

4.3.3 Indexing of project costs to 2012 prices

Analysis including temporally spread data should be carried out with great caution because unexpected deviations in explanatory factors such as advanced construction technology or petroleum prices variations may jeopardize the predictability of future durations that are forecast on the basis of past costs as potential independent variable. An appropriate index, CPI has been used.

Consumer Price Index (CPI)**:** The principle measure of price variation at retail level is Consumer Price Index (CPI) and generally represents inflation rate in the country. Inflation is underlying cause of upward movement of State Bank of Pakistan policy rate and prices of major inputs to construction industry i.e. cement, steel and oil. Figure 4.2 and Figure 4.3 describe the trend of CPI and inflation rate trends in Pakistan over a span of more than a decade.

In order to remove the effect of inflation on the individual project expenditure over the full analysis period, all the reported project expenditures were indexed up to 2012 equivalent PKR prices. This process involved the application of price indices to the project costs for the years 2001 through to 2012.

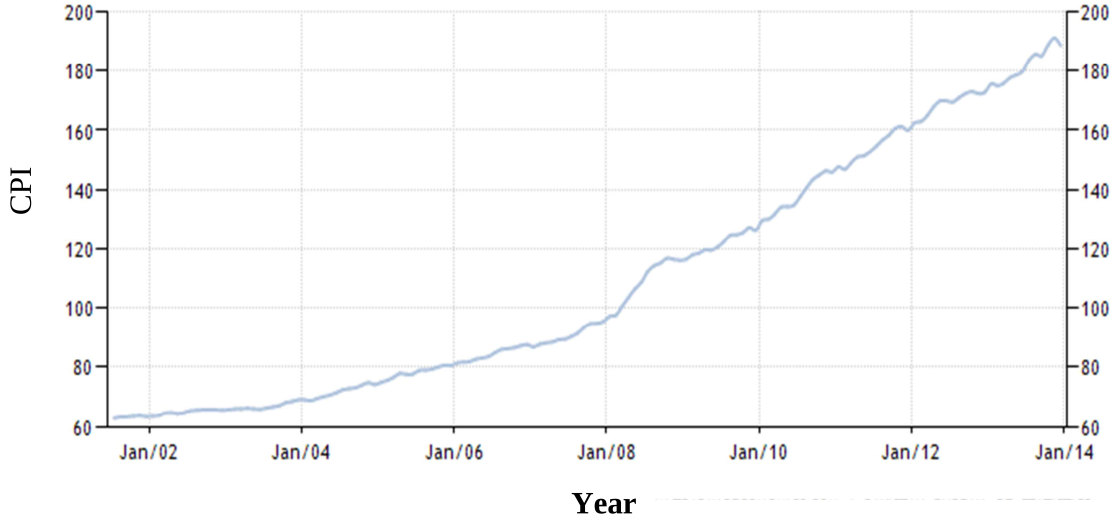


Figure 4.2 Consumer Price Index- CPI

Figure 4.3 Pakistan Inflation Rate

Table 4.2 details the CPI indices adopted over the analysis period. Column 3 lists the factor worth used to factor up the historical project cost information. These factors were applied to projects programmed and actual costs for the corresponding financial years in which the projects were constructed

$$\text{Factor worth}_{x} = \frac{CPI\ 2012}{CPI\ x}$$

4.4 Determination of Delay Factors from Historic Project Data

This step in the research required the determination of project delay factors from historic data. The analysis was focused on the client's exposure to project time overrun. A client focus demands a number of considerations identified in the literature to be taken into consideration when reporting the time overrun factors. Graphical representation of frequent occurring risks is presented in Figure 4.4.

Table 4.2 CPI Applied to the Projects Costs

Year	CPI	CPI Rate	Factor Worth
2001	103.15	3.15	2.8
2002	106.54	3.29	2.7
2003	109.64	2.91	2.6
2004	117.80	7.44	2.4
2005	128.47	9.06	2.2
2006	138.64	7.92	2.1
2007	149.18	7.60	1.9
2008	179.45	20.29	1.6
2009	203.95	13.65	1.4
2010	232.26	13.88	1.2
2011	259.94	11.92	1.1
2012	285.16	9.70	1.0

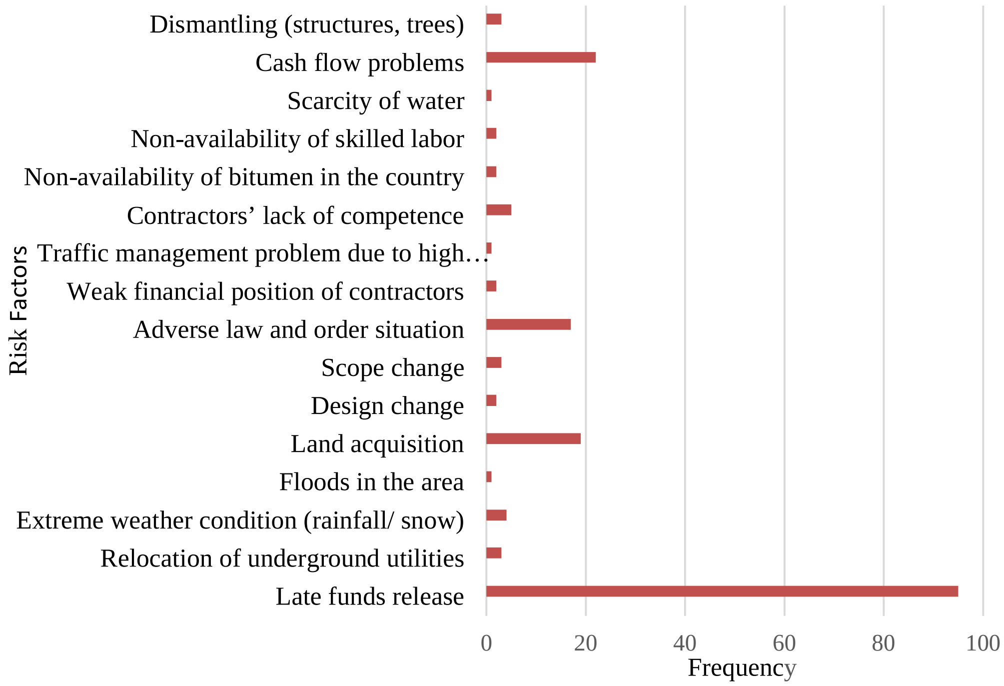


Figure 4.4 Frequent occurring risks

The available highway data comprised of individual descriptions of all the reasons for individual projects stated by NHA as having caused the client's programmed duration for the project to be exceeded. It was therefore considered that the reasons stated were true and factual representations that were documented of the historic project data. These reasons were then recorded in an Excel spreadsheet for further analysis. Where common time overrun factors occurred across projects, single delay factors were recorded to cover incidences. All unique reasons were recorded individually. The research identified 15 factors from the highway data analyzed. The final list of time overrun variables, their symbols and the number of times they occurred across projects are shown in Table 4.3.

Table 4.3 Project Time Overrun Risk Factors

Causes of Delays	Code	Total Indices
Late funds release	R1	95
Relocation of underground utilities	R2	3
Extreme weather condition (rainfall/ snow)	R3	4
Floods in the area	R4	1
Land acquisition	R5	19
Design change	R6	2
Scope change	R7	3
Adverse law and order situation	R8	17
Weak financial position of contractors	R9	2
Traffic management problem due to high traffic	R10	1
Contractors' lack of competence	R11	5
Non-availability of bitumen in the country	R12	2
Non-availability of skilled labor	R13	2
Scarcity of water	R14	1
Cash flow problems	R15	22
Dismantling (structures, trees)	R16	3

4.5 General Description of Data

4.5.1 Distribution of Projects by Provinces

Table 4.4 presents the distribution of the studied projects among provinces. There are sufficient projects in each province for the description statistics and to justify regression analysis. The province with most projects was Balochistan, probably because the largest highway network length in this province.

Table 4.4 Distribution of Projects by Provinces

Province	Number of projects	Highway Length (Km)
Sindh	32	2,204
Punjab	32	2,731
Balochistan	34	4,565
Khyber Pukhtunkhwa	22	1,878

4.5.2 Distribution of Projects by Project Type

In Table 4.5, the distribution of projects across the various project types, each with a good number of observations is presented.

Table 4.5 Distribution of Projects by Project Type

Project Type	Number Of Projects
Pavement Construction	45
Pavement Rehabilitation	21
Pavement Improvement	20
Bridge Construction	34

4.6 Time Overruns

The analysis of the time delay included only 120 projects of the database. Time delay is defined as the difference between the estimated final date and the actual one. It is worth noting that most of these delayed days were actually due to late funds release by the funding agency. Figure 4.5 presents the distribution of time delays. It can be noticed that the distribution is not symmetric. There were few projects indeed that were completed before the estimated final date and for such projects, the time delay was negative.

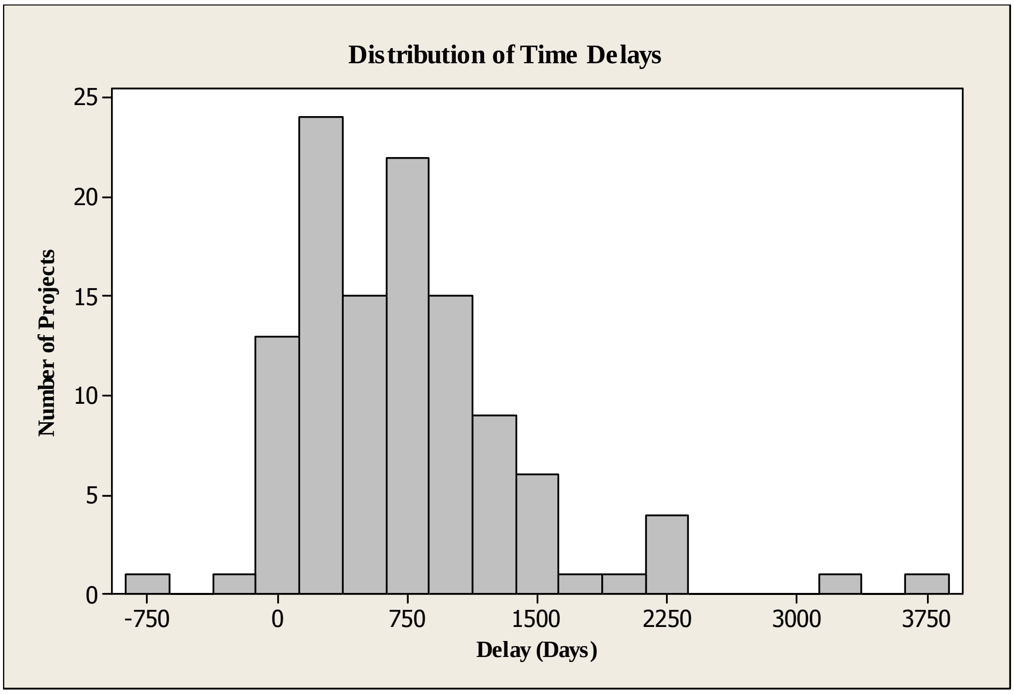


Figure 4.5 Distribution of Project Delays (days) across Projects

4.7 Average Project Duration

Table 4.6 presents the average contract duration for a given project characterized by province and project type.

Table 4.6 Average Project Duration

Project Type	Project Cost (m)	Province			
		Sindh	KPK	Punjab	Balochistan
Bridge	0-1000	927	1340	1137	-
	1000-5000	1178	-	1118	-
	5000-10,000	-	-	-	-
	10,000-15,000	-	-	-	-
Construction	0-1000	1249	959	1119	-
	1000-5000	2071	2161	2333	1510
	5000-10,000	-	3378	2312	1864
	10,000-15,000	-	-	1461	1614
Rehabilitation	0-1000	951	-	1370	-
	1000-5000	1173	943	1278	349
	5000-10,000	-	-	-	-
	10,000-15,000	1112	-	-	-
Improvement	0-1000	1127	1553	1177	472
	1000-5000	1372	-	-	1719
	5000-10,000	-	-	-	-
	10,000-15,000	-	-	-	-

4.8 Average Project Delays

Table 4.7 presents the average contract delays for a given project characterized by province and project type.

Table 4.7 Average Project Delays

Project Type	Project Cost (m)	Province			
		Sindh	KPK	Punjab	Balochistan
Bridge	0-1000	608	685	589	-
	1000-5000	589	-	480	-
	5000-10,000	-	-	-	-
	10,000-15,000	-	-	-	-
Construction	0-1000	703	464	700	-
	1000-5000	1347	1212	1593	844
	5000-10,000	-	2282	1216	867
	10,000-15,000	-	-	365	-
Rehabilitation	0-1000	433	641	497	699
	1000-5000	531	-	548	-
	5000-10,000	-	-	-	-
	10,000-15,000	198	-	-	-
Improvement	0-1000	762	1035	812	107
	1000-5000	961	-	-	548
	5000-10,000	-	-	-	-
	10,000-15,000	-	-	-	-

4.9 Descriptive Statics of the Data

Table 4.8 presents the descriptive statistics of the data of project duration and project cost for different highway projects. It can be observed that mean of project duration lies in range of 1000 to 1600 whereas standard deviation is in between 300 to 850.

Table 4.8 Descriptive Statics of the Data by Project Type

Statistics	Construction	Improvement	Rehabilitation	Bridge
(a) Project Duration				
Mean	1585.46	1285.53	1047.85	1094.14
Std. dev	838.98	558.75	299.16	539.95
Minimum	545	333	395	250
Maximum	4474	26996.31	1500	2500
Observation Count	45	21	20	34
(b) Project Indexed cost (in millions PKR)				
Mean	1667.65	1299.98	1159.8	1031.76
Std. dev	1357.27	986.46	310.85	640.15
Minimum	300	287.29	664.5	148
Maximum	5617.73	27299.77	1616.57	3221.68
Observation Count	45	21	20	34

4.10 Delay Factors Distribution in Balochistan

Figure 4.6 represents the distribution of risk factors across the province of Baluchistan. The majority of the projects had R1 prevailing in 44% projects. These risk factors were taken into account while proposing models for project delay calculation.

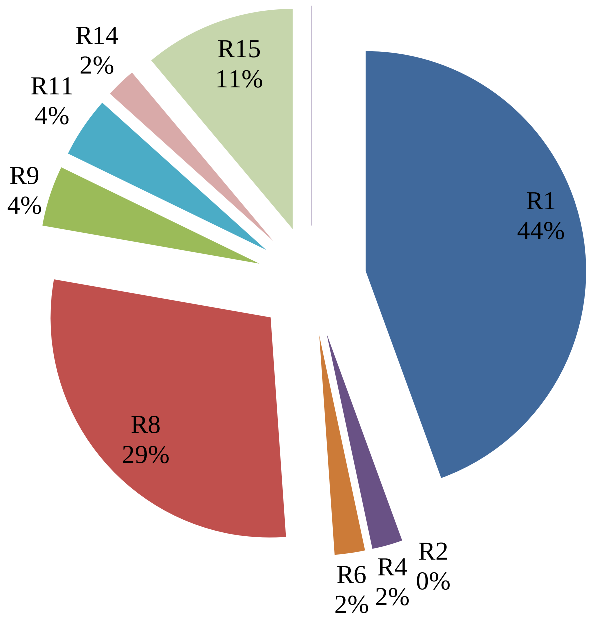


Figure 4.6 Distribution of Risk Factors in Balochistan

4.11 Delay Factors Distribution in Sindh

Figure 4.7 represents the distribution of risk factors across the province of Sindh. The majority of the projects had R1 prevailing in 65% projects. These risk factors were taken into account while proposing models for project delay calculation.

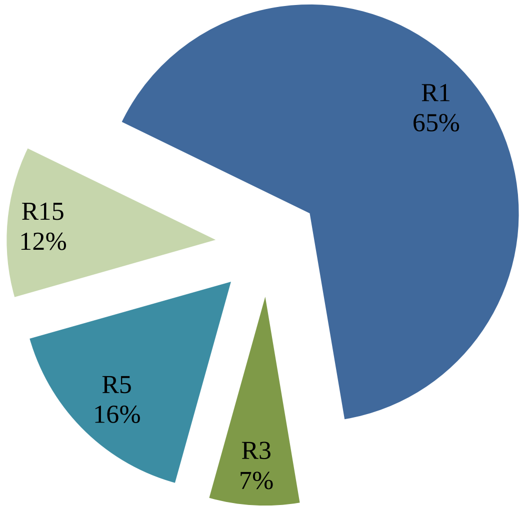


Figure 4.7 Distribution of Risk Factors in Sindh

4.12 Delay Factors Distribution in KPK

Figure 4.8 represents the distribution of risk factors across the province of KPK. The majority of the projects had R1 prevailing in 54% projects. These risk factors were taken into account while proposing models for project delay calculation.

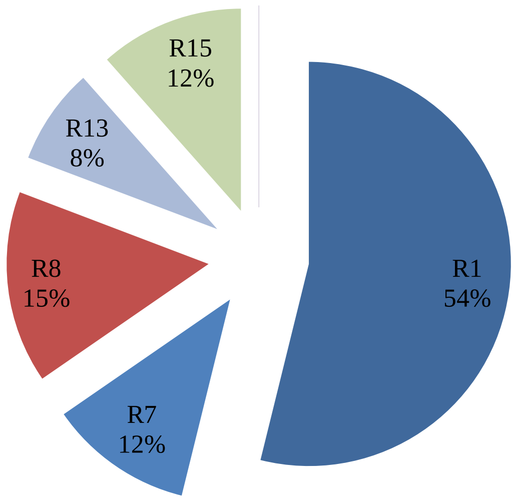


Figure 4.8 Distribution of Risk Factors in KPK

4.13 Delay Factors Distribution in Punjab

Figure 4.9 represents the distribution of risk factors across the province of Punjab. The majority of the projects had R1 prevailing in 49% projects. These risk factors were taken into account while proposing models for project delay calculation.

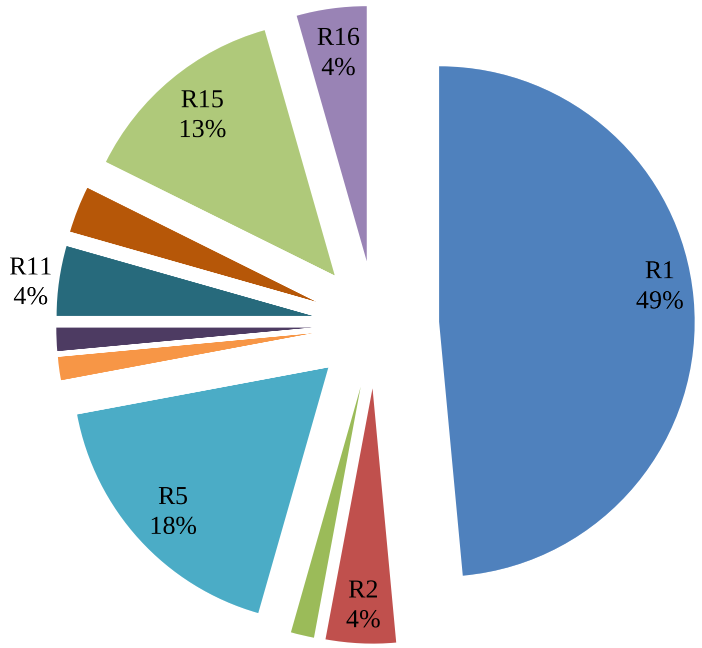


Figure 4.9 Distribution of Risk Factors in Punjab

4.14 Delays due to Risk Factors:

Figure 4.10 represents the distribution of time overrun rates across projects due to individual or combination of risk factors. It has been observed that R1 factor is associated with most number of delays (days). The combination of various factors like R1, R15 and R1, R8 are also causing significant delay.

.

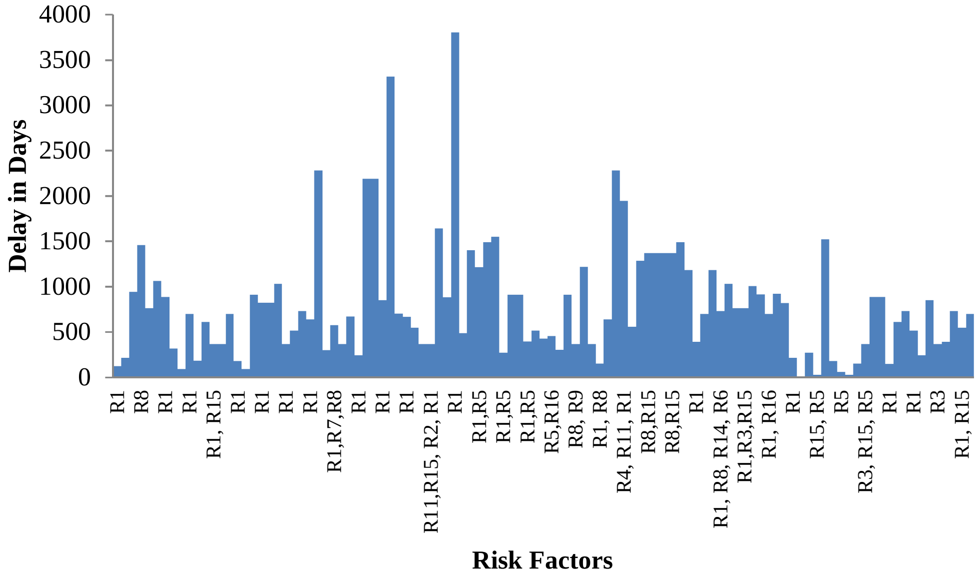


Figure 4.10 Distribution of Delays (days) due to Different Risk Factors

4.15 Chapter Summary

The chapter first provided an insight about the data collection and data development process along with indexing of the project cost. Consumer price index is calculated along with factor worth. The factor worth and CPI is determined for year 2012. Database is developed on the basis of project type and project cost. The chapter further provided the general descriptive of the data to help and identify the trends data follows. The frequency of delay factors is represented by gantt chart. Time overruns, project durations and average project delays are also presented herein. Project duration and average cost delays for all four provinces are tabulated for four different types of projects namely bridge, construction, rehabilitation and improvement. Delay factors along with their average delay statistics are also depicted. Delay factors for each province is calculated and pie chart is also presented separately. It has been observed that R1 is most prevailing factor in all province as indicated by pie chart.

Chapter 5

STATISTICAL ANALYSIS AND DATA MODELLING

5.1 Introduction

5.1.1 Statistical modeling

The aim of this step was to analyze of historical project data based on statistical theories and concepts that identified direct correlations between particular highway construction project types, project cost, geographical location and project duration.

5.1.2 Basic Assumptions

Geographic project type: The reason for including the geographic project type in the proposed models was because there appeared to exist a strong relationship between the remoteness of a project from established workforces and from proven materials and component manufactures that could lead to increases in project costs above those estimated. Drew and Skitmore (1992) identified the density of population and the extent of geographic area as important factors for competitive bidding in building projects. It was therefore postulated that the rural geographic type of highway projects had a higher potential to overrun budgeted costs.

Geographic data and model coding: An analysis was carried out on the project data to split projects down into the geographic area in which the project was constructed. Provincial data was split in the group of two depending upon the strong relationship between the risk factors and other attributes like project cost that could lead to greater time duration and time delays. Sindh and Punjab were placed in group 1 while Balochistan and Khyber Pukhtunkhwa were stationed in group 2. Two binary variables were created for each of these groups. If the information pertained to the location in the group 1 the variable takes the value of 1 otherwise 0 is inserted. A sample of this data coding for the geographic area is shown in Table 5.1.

Table 5.1 Geographic coding sample

Project Number	Location	Dummy Variable
S2	Sindh	1
P25	Punjab	1
B15	Balochistan	0
K1	KPK	0

Indexed highway project programmed cost continuous variable: Generally there is a correlation between the cost of a project and the size of the project. In this research it was adopted that, if projects costs are indexed to a common year, then the project cost can be used as a surrogate for project size. The reason for including the indexed highway programmed cost in the proposed model was because it was thought that there was a strong relationship between the size of a project and project duration. For highway projects, the greatest risk lays below ground level due to the relatively greater physical footprint of the project, and the larger the footprint, then the larger the risk cost should be.

Outlying data values: As a preliminary step in the analysis process, the 145 project cases identified in were analyzed for random disturbance. For the purpose of specifically identifying any project outliers, a linear regression analysis was carried out using the dependent variable as 'project cost in PKR' and the predictor variable as Project Duration (Days)'. Outlying data (exceeding three standard deviations of the mean) of project duration and project delays was expunged using statistical software (SPSS).

Data from the provinces showing similar trends was grouped together. One of the hypothesis thus in this research is everything remains same, difference in geographical groupings may result in different project duration. For example the projects in group 1 are subjected to similar situations different than the projects in group 2. Table 5.2 presents a summary of model variable description.

Table 5.2 Model Variable description

Variable	Description
Project Duration	Highway project duration in days
X_1= Project indexed cost	Final Cost in millions of PKR, rebased to 2012 price
X_2= Geographical Location	X_2=1 indicates that variable pertains to Sindh or Punjab X_2=0 indicates that variable pertains to Balochistan or

Considering the convenience for use linear normal models were formed. The general form of the duration model is presented in Equation 5-1:

$$\text{Project Duration} = \beta_1 \text{ x Cost} + \beta_2 \text{ x Geographical Location} + \text{intercept} \quad (5\text{-}1)$$

The regression results produced an R value of 0.56 indicating reasonably good correlation. It can be seen that marginal increase in project duration with unit increase in project cost seems to be linear, and gets high if the projects are placed in geographical locations in group 1. Table 5.3 represents the model estimation results for the duration model for all project types.

Table 5.3 Parameter Estimates of All Project Types

Variable	Coefficients (t-statistics in parentheses)
Project Duration Model	
Intercept	433.60 (3.343)
Project indexed cost	0.37 (8.369)
Geographical Location	156.17 (1.30)
Number of observations	92
R^2	0.56
Adjusted R^2	0.55

Separate models were also developed for different project types to study the impact of explanatory variables over the project duration, results are reported in Table 5.4

Table 5.4 Parameter Estimates of Individual Project Types (t-statistics in parentheses)

Variable	Construction	Bridge	Rehabilitation	Improvement
Project Duration Model				
Intercept	871.6 (4.23)	663.57 (2.42)	254.80 (1.01)	732.91 (2.79)
Project indexed cost	0.40 (5.20)	0.49 (3.94)	0.620 (3.50)	0.325 (2.64)
Geographical Location	97.09 (2.46)	-89.13 (-2.31)	86.42 (2.574)	246.58 (2.03)
Number of observations	45	34	20	21
R^2	0.4	0.34	0.43	0.28
Adjusted R^2	0.38	0.3	0.36	0.2

5.2 Bridge Duration Model Plots:

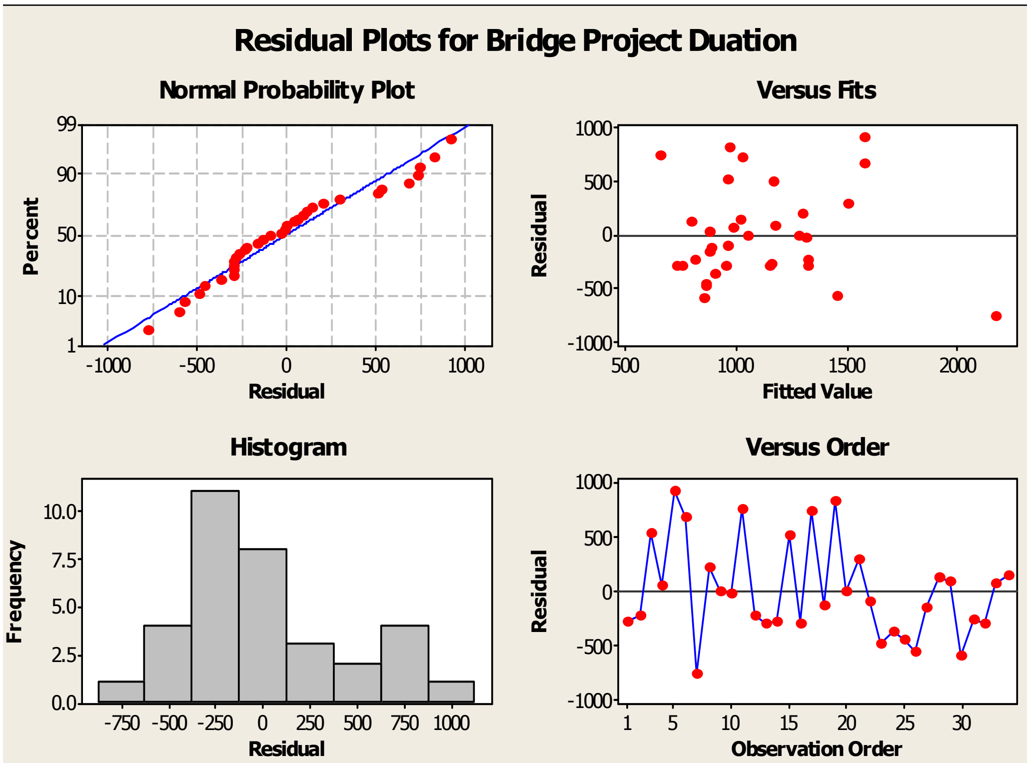


Figure 5.1 Probability plot of residuals (Bridge)

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Following assumptions were satisfied:

Normality Assumption: The check of normality assumption was made by plotting a histogram of residuals as shown in Figure 5.1. It shows that assumptions are almost satisfied and this moderate departure from the normality does not imply a serious violation of the assumption. The plot look like a simple normal distribution.

Another way to check normality is plotting normal probability plot of residuals. As shown in the Figure 5.1, the plot resembles almost straight line which means that the errors are normally distributed and assumption is satisfied.

Residual vs. Fitted values: Residual vs. Fitted values is a scatter plot of residuals on the y axis and fitted values (responses) on the x axis. Plotting of the residuals versus the fitted values or responses must produce a distribution of points that is scattered randomly about 0, regardless of size of fitted values. Commonly the residual values increase as the size of the fitted values increase. Due to this reason the residuals cloud become "funnel shaped" with the larger end toward larger fitted values; From the Figure 5.1, it is clear that the plotted points lie in an approximately horizontal band across the plot. The residuals are structure less. The plot does not reveal any obvious pattern like outward-opening funnel or megaphone. Non constant variance sometime arises when data follow non normal skewed distribution, because in skewed distribution the variance tends to be function of mean

Residual vs. Observation Order: Plotting the residual in time order of data collection (observation order) is helpful in detecting correlation between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation and implies that the independence assumption on the errors has been violated. This is potentially serious problem, and one that is difficult to correct, so it is important to prevent the problem if possible when data are collected. Figure 5.1 shows that the residuals are not correlated and they are independent. The plot is structure less and does not show any kind of definite patterns.

5.3 Construction Duration Model Plots:

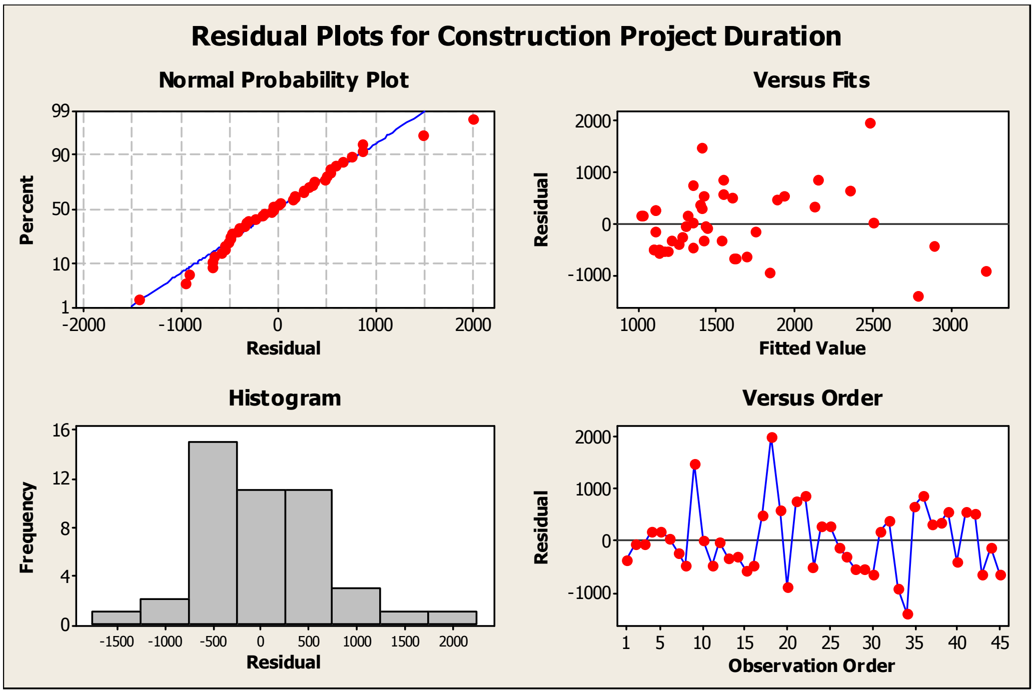


Figure 5.2 Probability plot of residuals (Construction)

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Following assumptions were satisfied:

Normality Assumption: The check of normality assumption was made by plotting a histogram of residuals as shown in Figure 5.2. It shows that assumptions are almost satisfied and this moderate departure from the normality does not imply a serious violation of the assumption. The plot look like a simple normal distribution.

Another way to check normality is plotting normal probability plot of residuals. As shown in the Figure 5.2, the plot resembles almost straight line which means that the errors are normally distributed and assumption is satisfied.

Residual vs. Fitted values: Residual vs. Fitted values is a scatter plot of residuals on the y axis and fitted values (responses) on the x axis. Plotting of the residuals versus the fitted values or responses must produce a distribution of points that is scattered randomly about 0, regardless of size of fitted values. Commonly the residual values increase as the size of the fitted values increase. Due to this reason the residuals cloud become "funnel shaped" with the larger end toward larger fitted values; From the Figure 5.2, it is clear that the plotted points lie in an approximately horizontal band across the plot. The residuals are structure less. The plot does not reveal any obvious pattern like outward-opening funnel or megaphone. Non constant variance sometime arises when data follow non normal skewed distribution, because in skewed distribution the variance tends to be function of mean.

***Residual vs. Observation Order**:* Plotting the residual in time order of data collection (observation order) is helpful in detecting correlation between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation and implies that the independence assumption on the errors has been violated. This is potentially serious problem, and one that is difficult to correct, so it is important to prevent the problem if possible when data are collected. Figure 5.2 shows that the residuals are not correlated and they are independent. The plot is structure less and does not show any kind of definite patterns.

5.4 Delay Models

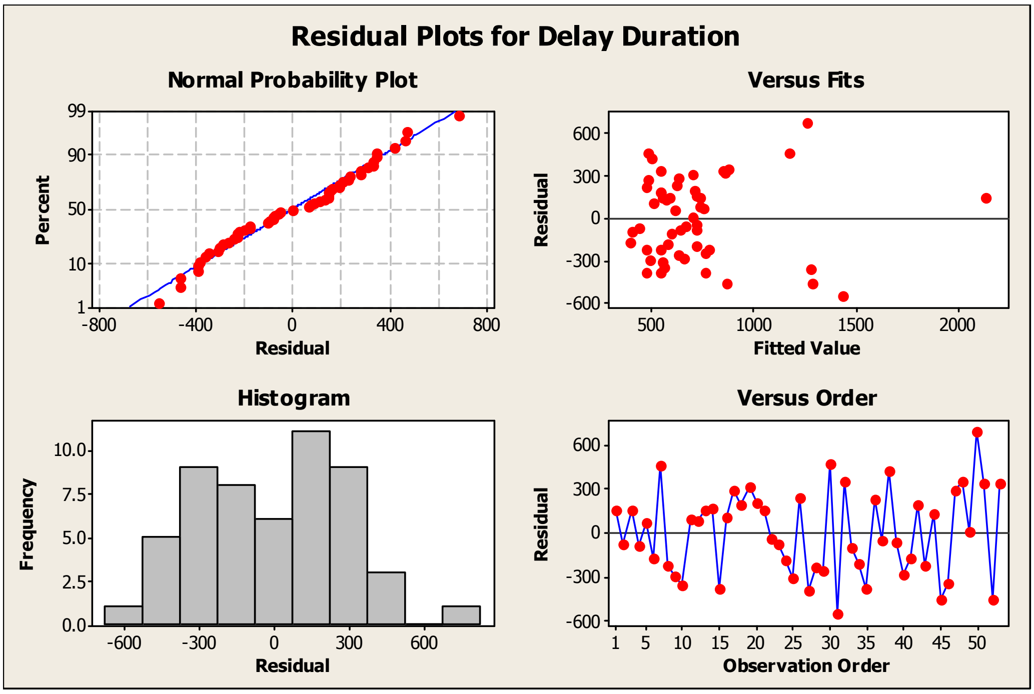


Figure 5.3 Probability plot of residuals (Delay)

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Following assumptions were satisfied:

Normality Assumption: The check of normality assumption was made by plotting a histogram of residuals as shown in Figure 5.3. It shows that assumptions are almost satisfied and this moderate departure from the normality does not imply a serious violation of the assumption. The plot look like a simple normal distribution.

Another way to check normality is plotting normal probability plot of residuals. As shown in the Figure 5.3, the plot resembles almost straight line which means that the errors are normally distributed and assumption is satisfied.

Residual vs. Fitted values*:* Residual vs. Fitted values is a scatter plot of residuals on the y axis and fitted values (responses) on the x axis. Plotting of the residuals versus the fitted values or responses must produce a distribution of points that is scattered randomly about 0, regardless of size of fitted values. Commonly the residual values increase as the size of the fitted values increase. Due to this reason the residuals cloud become "funnel shaped" with the larger end toward larger fitted values; From the Figure 5.3, it is clear that the plotted points lie in an approximately horizontal band across the plot. The residuals are structure less. The plot does not reveal any obvious pattern like outward-opening funnel or megaphone. Non constant variance sometime arises when data follow non normal skewed distribution, because in skewed distribution the variance tends to be function of mean.

Residual vs. Observation Order*:*Plotting the residual in time order of data collection (observation order) is helpful in detecting correlation between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation and implies that the independence assumption on the errors has been violated. This is potentially serious problem, and one that is difficult to correct, so it is important to prevent the problem if possible when data are collected. Figure 5.3 shows that the residuals are not correlated and they are independent. The plot is structure less and does not show any kind of definite patterns.

5.5 Duration Model

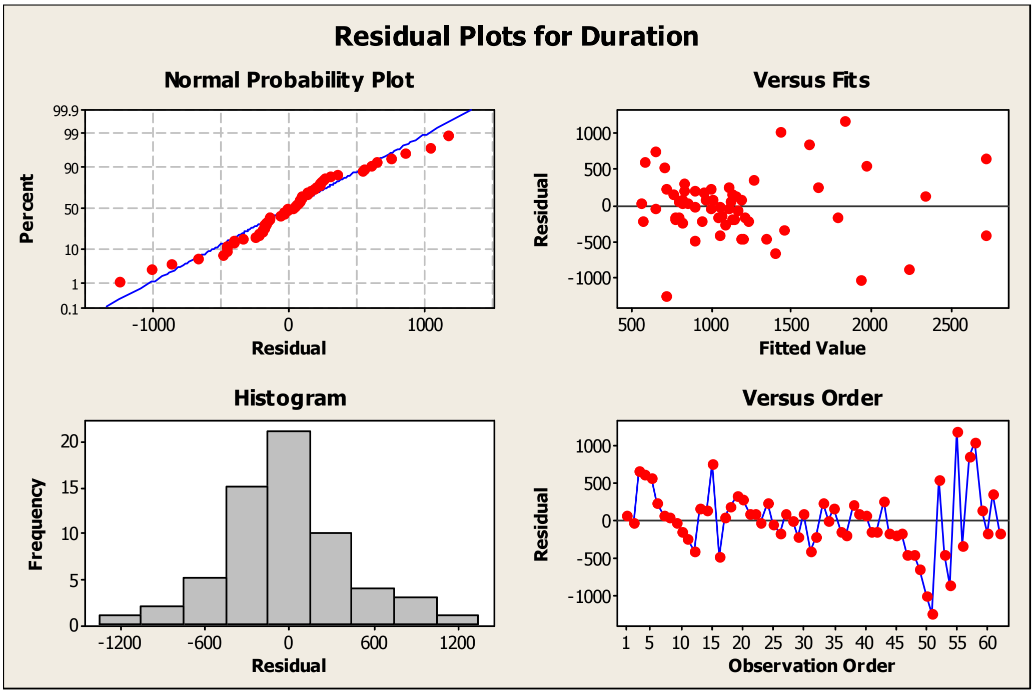


Figure 5.4 Probability plot of residuals (Duration-All Projects)

Normality Assumption: The check of normality assumption was made by plotting a histogram of residuals as shown in Figure 5.4. It shows that assumptions are almost satisfied and this moderate departure from the normality does not imply a serious violation of the assumption. The plot look like a simple normal distribution.

Another way to check normality is plotting normal probability plot of residuals. As shown in the Figure 5.4, the plot resembles almost straight line which means that the errors are normally distributed and assumption is satisfied.

Residual vs. Fitted values*:* Residual vs. Fitted values is a scatter plot of residuals on the *y* axis and fitted values (responses) on the *x* axis. Plotting of the residuals versus the fitted values or responses must produce a distribution of points that is scattered randomly about 0, regardless of size of fitted values. Commonly the residual values increase as the size of the fitted values increase. Due to this reason the residuals cloud become "funnel shaped" with the larger end toward larger fitted values; From the Figure 5.4, it is clear that the plotted points lie in an approximately horizontal band across the plot. The residuals are structure less. The plot does not reveal any obvious pattern like outward-opening funnel or megaphone. Non constant variance sometime arises when data follow non normal skewed distribution, because in skewed distribution the variance tends to be function of mean.

Residual vs. Observation Order*:* Plotting the residual in time order of data collection (observation order) is helpful in detecting correlation between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation and implies that the independence assumption on the errors has been violated. This is potentially serious problem, and one that is difficult to correct, so it is important to prevent the problem if possible when data are collected. Figure 5.4 shows that the residuals are not correlated and they are independent. The plot is structure less and does not show any kind of definite patterns.

5.6 Improvement Model

***Normality Assumption*:** The check of normality assumption was made by plotting a histogram of residuals as shown in Figure 5.5. It shows that assumptions are almost satisfied and this moderate departure from the normality does not imply a serious violation of the assumption. The plot looks like a simple normal distribution.

Another way to check normality is plotting normal probability plot of residuals. As shown in the Figure 5.5, the plot resembles almost straight line which means that the errors are normally distributed and assumption is satisfied

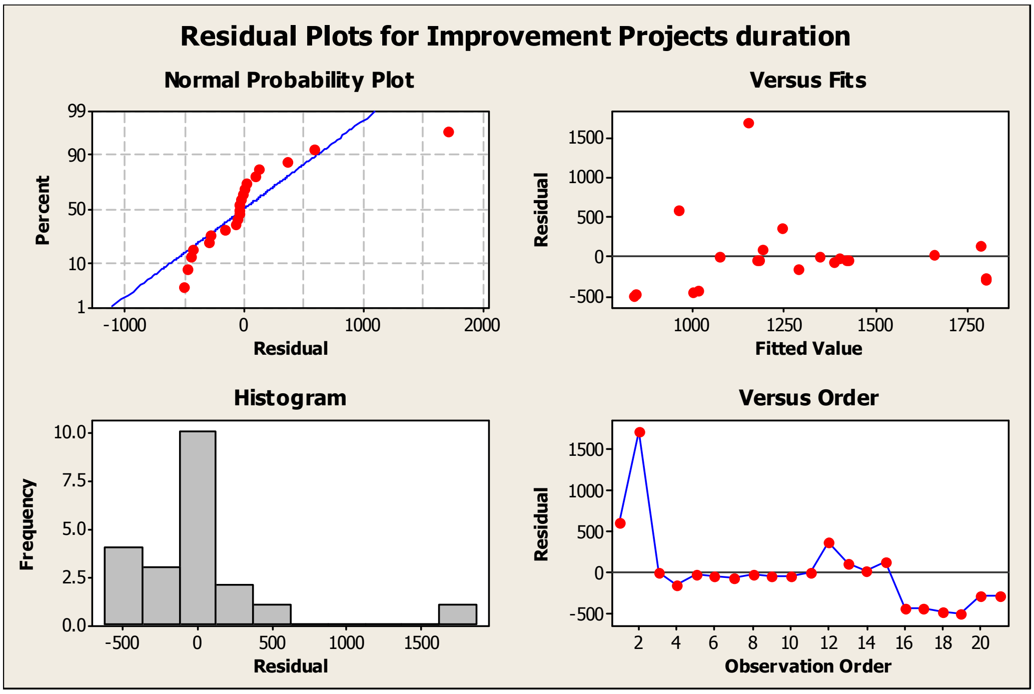


Figure 5.5 Probability plot of residuals (Improvement)

.***Residual vs. Fitted Values****:* Residual vs. Fitted values is a scatter plot of residuals on the *y* axis and fitted values (responses) on the *x* axis. Plotting of the residuals versus the fitted values or responses must produce a distribution of points that is scattered randomly about 0, regardless of size of fitted values. Commonly the residual values increase as the size of the fitted values increase. Due to this reason the residuals cloud become "funnel shaped" with the larger end toward larger fitted values; From the Figure 5.5, it is clear that the plotted points lie in an approximately horizontal band across the plot. The residuals are structure less. The plot does not reveal any obvious pattern like outward-opening funnel or megaphone. Non constant variance sometime arises when data follow non normal skewed distribution, because in skewed distribution the variance tends to be function of mean.

Residual vs. Observation Order*:* Plotting the residual in time order of data collection (observation order) is helpful in detecting correlation between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation and implies that the independence assumption on the errors has been violated. This is potentially serious problem, and one that is difficult to correct, so it is important to prevent the problem if possible when data are collected. Figure 5.5 shows that the residuals are not correlated and they are independent. The plot is structure less and does not show any kind of definite patterns.

5.7 Rehabilitation Models

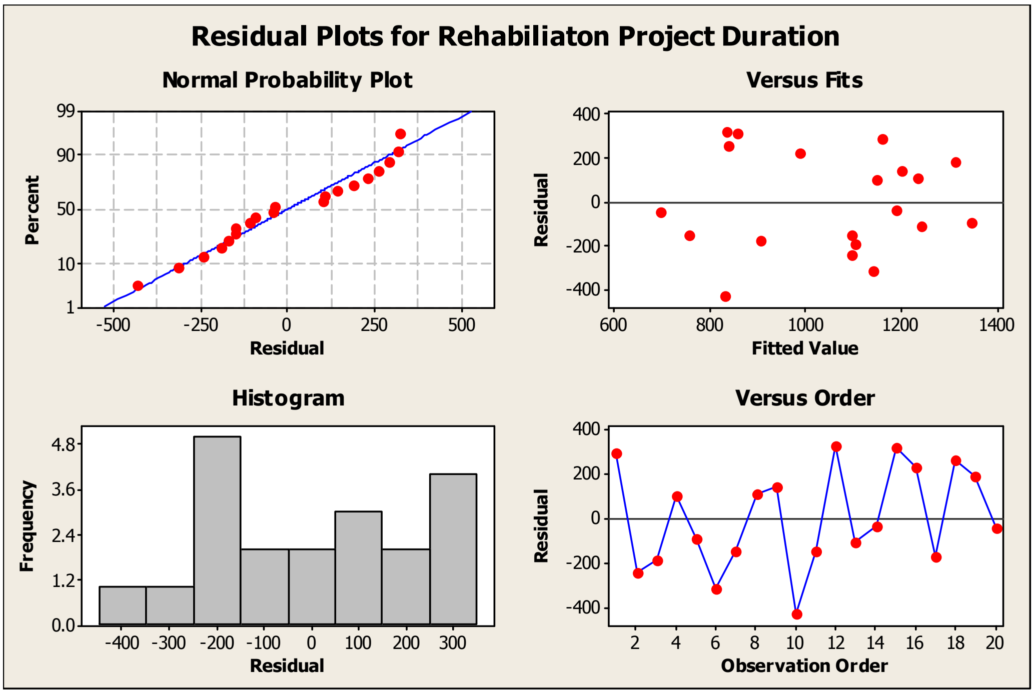


Figure 5.6 Probability plot of residuals (Rehabilitation)

Normality Assumption: The check of normality assumption was made by plotting a histogram of residuals as shown in Figure 5.6. It shows that assumptions are almost satisfied and this moderate departure from the normality does not imply a serious violation of the assumption. The plot look like a simple normal distribution.

Another way to check normality is plotting normal probability plot of residuals. As shown in the Figure 5.6, the plot resembles almost straight line which means that the errors are normally distributed and assumption is satisfied.

Residual vs. Fitted values*:* Residual vs. Fitted values is a scatter plot of residuals on the *y* axis and fitted values (responses) on the *x* axis. Plotting of the residuals versus the fitted values or responses must produce a distribution of points that is scattered randomly about 0, regardless of size of fitted values. Commonly the residual values increase as the size of the fitted values increase. Due to this reason the residuals cloud become "funnel shaped" with the larger end toward larger fitted values; From the Figure 5.6, it is clear that the plotted points lie in an approximately horizontal band across the plot. The residuals are structure less. The plot does not reveal any obvious pattern like outward-opening funnel or megaphone. Non constant variance sometime arises when data follow non normal skewed distribution, because in skewed distribution the variance tends to be function of mean.

Residual vs. Observation Order*:* Plotting the residual in time order of data collection (observation order) is helpful in detecting correlation between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation and implies that the independence assumption on the errors has been violated. This is potentially serious problem, and one that is difficult to correct, so it is important to prevent the problem if possible when data are collected. Figure 5.6 shows that the residuals are not correlated and they are independent. The plot is structure less and does not show any kind of definite patterns.

5.8 Normality Test for Risk Factors across the Country

Shapiro Wilk normality test is conducted as shown in table 5.5. This test is performed to check the normality of data as per the requirements of sample size which is less than 2000. This test was performed to determine the nature of data that is either parametric or non-parametric. Significance value found from the test is 0.00 which shows that the data is not normally distributed, as for sufficiently normal data significance value should be greater than 0.05. Therefore, for current data non-parametric techniques are used for further analysis as data is not normally distributed. Shapiro Wilk normality test conducted on the risk data is shown in Table 5.5. Figure 5.7 and Figure 5.8 graphically represents the behavior of the data.

Table 5.5 Test of Normality Shapiro-Wilk Test

	Shapiro-Wilk		
	statistic	df	Sig.
Indices of Risk Factors	0.479	16	0.00

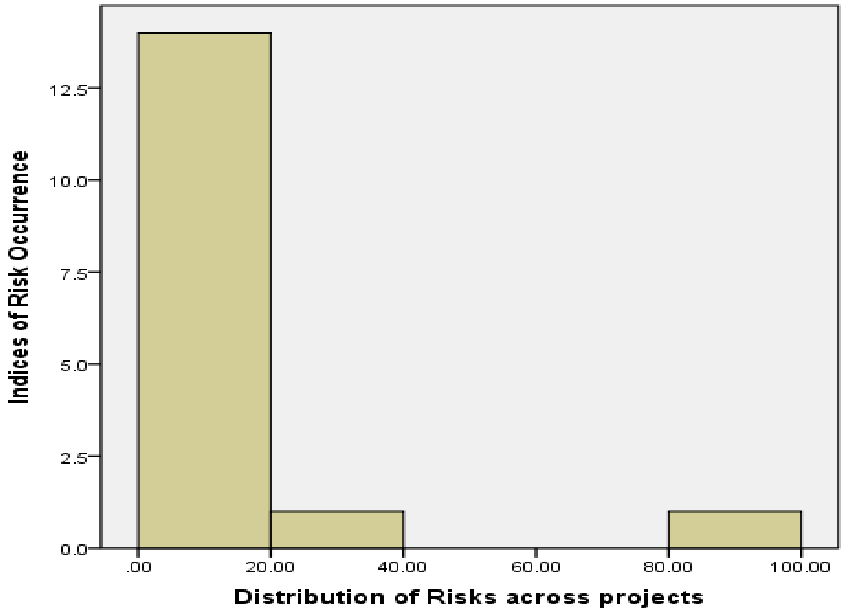


Figure 5.7 Distribution of Risks across projects

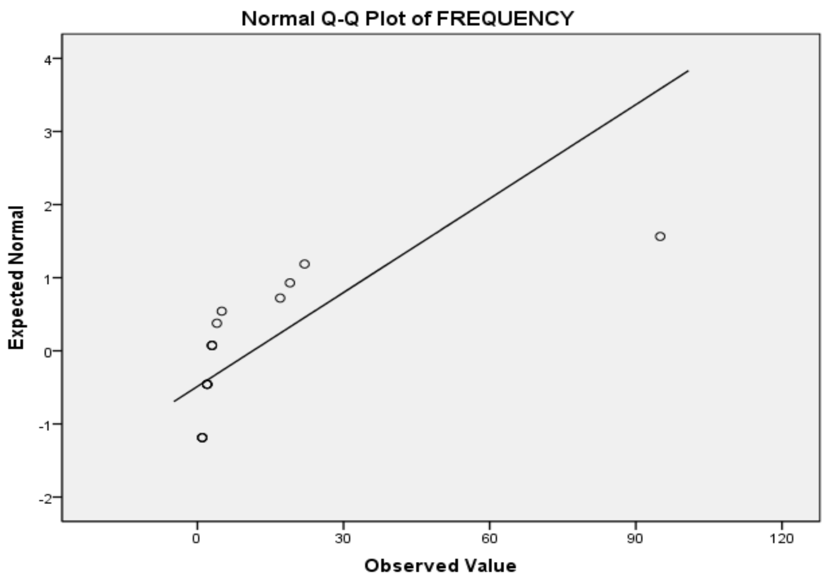


Figure 5.8 Figure Showing Risk Data Behaving in a Nonlinear Fashion

5.9 Severity Index Analysis

Severity index analysis technique is used for the analysis of non-parametric data. Non-parametric techniques are adopted for condition to have meaningful results to analyze data when the current data collected is ordinal i-e distance between any two ratings is unknown and parametric statistics such as mean, standard deviation etc will not produce (Siegel, 1956; Siegel and Castellan, 1988; Johnson and Bhattacharyya, 1996).

Relative index ranking and Frequency analysis are non-parametric techniques used for the analysis of data. Relative index ranking technique is extensively used by construction management researchers to analyze response data of structured questionnaire concerning ordinal measurement of attitudes. Severity index analysis is a form of this Relative index ranking that uses weighted percentage scores to compare the comparative significance of the risks under study (Elhag and Boussabaine, 1999; Al-Hammad, 2000; Ballal, 2000). For the present study; First Frequency analysis was performed to determine the frequency of in the historic data which were then used to calculate severity. Table 5.6 shows the severity index analysis of the risks factors.

The statistics in Table 5.6 show that the top three risks ranked in the NHA projects of Pakistan are late funds release from to the funding agencies, land acquisition problems and cash flow problems within the highway agency.

Statistical models explaining correlation between several relationships relating to time overrun were also identified in this research. Multivariate regression analysis was used to manage the relationship between project cost, project risks and project delay.

The time overrun, adopted as a dependent variable was denoted as the difference between the programmed duration and actual duration. The correlation was identified among the following project variables:

1. Highway completed indexed cost;
2. Highway project type (construction, rehabilitation, improvement, bridge);
3. Highway geographic location (Sindh, Punjab, Balochistan, Khyber Pakhtunkhwa);
4. Risk factors (R1-R16)

Table 5.6 Severity Indices of Risks occurring across the country

Risks Factors	Code	Severity Indices of Risks (%)	Severity Indices of Risks	Ranking of Risks based on Severity Indices
Late funds release	R1	52.198	0.522	16
Relocation of underground utilities	R2	0.330	0.003	9
Extreme weather condition (rainfall/ snow)	R3	0.440	0.004	11
Floods in the area	R4	0.110	0.001	2
Land acquisition	R5	2.088	0.021	14
Design change	R6	0.220	0.002	6
Scope change	R7	0.330	0.003	9
Adverse law and order situation	R8	1.868	0.019	13
Weak financial position of contractors	R9	0.220	0.002	6
Traffic management problem due to high traffic	R10	0.110	0.001	2
Contractors' lack of competence	R11	0.549	0.005	12
Non-availability of bitumen in the country	R12	0.220	0.002	6
Non-availability of skilled labor	R13	0.220	0.002	6
Scarcity of water	R14	0.110	0.001	2
Cash flow problems	R15	4.835	0.048	15
Dismantling (structures, trees)	R16	0.330	0.003	9

Taking into account the convenience for future use linear normal models were formed. Forward, backward and stepwise multivariate regression was used in analysis for finding correlation between the variables. Null hypothesis was assumed that there is no correlation between the project cost, project type, project delays and the project risks. Identification of best models was allowed by the use of coefficients of multiple determinations (R^2 and adjusted R^2 statics).

Multivariate regression taking into account all the explanatory variables provided an R value < 0.02, indicating that the data has not fitted the model very well. Stepwise regression showed risk variable R1 exhibiting reasonably strong correlation then rest of the risk factors responsible for delay.

An interesting finding was the correlation between the delay and project variables like project indexed cost, risk factor R1 and geographical location. The model parameter results are represented in Table 5.7.

Variable representing delay was found to be statistically significant to reasonable extent with the explanatory variables, irrespective of the project type. These results thus go in the favor to reject the null hypothesis, and support the relationship between delays, project indexed cost and risk factors. One of the interesting finding was the impact of geographical location on the delay duration which shows an inverse correlation. Historical data showed risk factor R1 prevailing in most of the projects and influencing the planned duration which is also found to be inversely related with each other and statistically significant (at 80% level of confidence).

Table 5.7 Model Estimation Results

Variable	Coefficients (t-statistics in parentheses)
Project Delay Model	
Intercept	545.41 (3.36)
Project indexed cost	0.29 (6.86)
R1 (Risk)	-176.16 (-2.26)
Geographical Location	-85.40 (-2.96)
Number of observations	90
R^2	0.53
Adjusted R^2	0.54

System weighted R-square related to goodness of fit of the developed model indicates that statistically significant variables that were used to explain the delay showed 50% of variation in the data. It is therefore concluded that the attributes considered in the present study are not the comprehensive representatives of the delay factors and there is an evident possibility of various other common variables which can further explain the variation in time overrun.

To evaluate the accuracy of predicted models in forecasting duration and delays, mean absolute percent error (MAPE) was calculated as shown in Equation 5-2:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}|PE_i| \qquad (5-2)$$

Where $PE_i = (X_i - F_i) / X_i$ is the error for the observations, the percentage error can be obtained by multiplying the above equation with 100. Here X_i is the actual duration and F_i is the predicted duration. The MAPE value closer to zero has better accuracy. The calculated MAPE values in this research suggest the extent of data over estimation or underestimation ranges from 20% to 40% which should be taken into account by the planning agencies while predicting the project durations and delays. For example MAPE value for rehabilitation duration model suggests that duration predicted is 21% over estimated or it is 21% underestimated.

5.10 Probabilistic Modeling

The rate at which the project durations are ending at time'ʻt' are presented in hazard functions plot in Figure 5.9. Level of hazard is presented at the vertical axis and abscissa of the plot presents the duration of project in days. The parameter P is found to be greater than 1, and hence it indicates all the project types to be non-monotonic. This exhibits the property of hazard function as shown in equation (5-3).

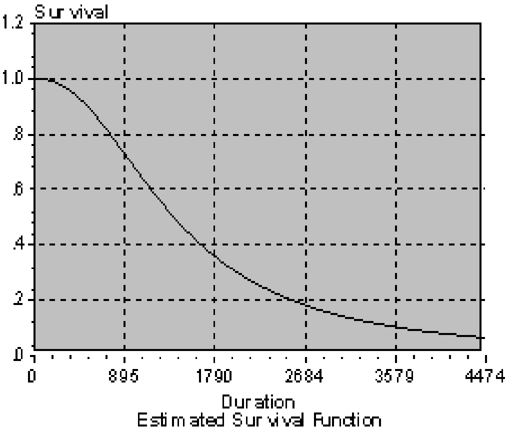
$$h(t) = [(\lambda P)\ (\lambda t)^{P-1}] / [1+(\lambda t)^{P}] \qquad (5\text{-}3)$$

Increasing in duration from zero to an inflation point where duration, $t^{*} = (P-1)^{1/P} / \lambda$, is calculated in Table 5.8. For different project types, project duration functions of the above form and respective survival distributions at percentiles 0.25, 0.50, 0.75 and 0.95 are presented in Table 5.8.

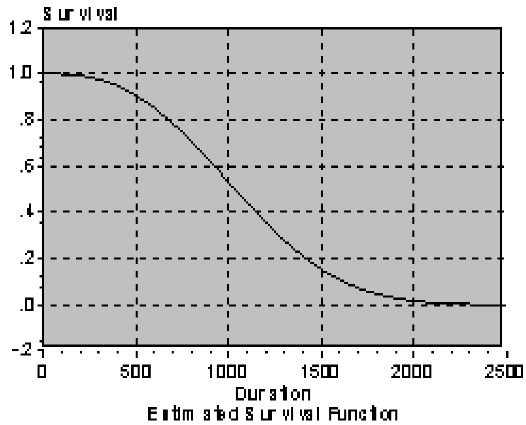
The rate at which the project durations are ending at time'ʻt' are presented in hazard functions plot in Figure 5.10. Level of hazard is presented at the vertical axis and abscissa of the plot presents the duration of project in days. The parameter P is found to be greater than 1, and hence it indicates all the project types to be non-monotonic.

Table 5.8 Model Estimation Results

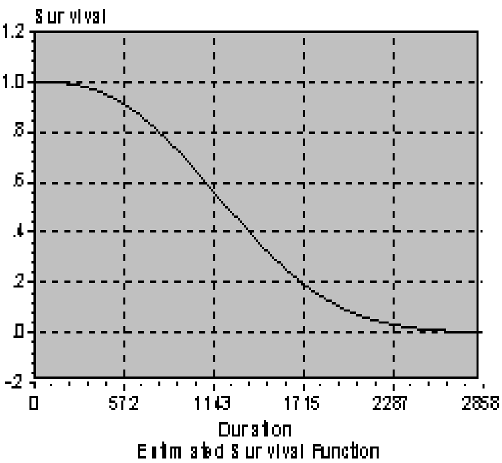
Project Type	Model Parameter (Log-Logistics)		Probability of Surviving to Time				Inflection Duration Point (days)
	λ	P	α =0.25	α =0.50	α =0.75	α =0.95	t^{*}
Pavement Construction	0.00072	2.25	2251	1382	849	374	1535
Pavement Improvement	0.00083	4.24	1554	1200	926	600	1590
Pavement Rehabilitation	0.00098	6.39	1209	1018	857	642	1328
Bridge Construction	0.00103	2.22	1585	966	589	257	1062
Delay Models	0.00167	1.87	1078	600	334	125	558



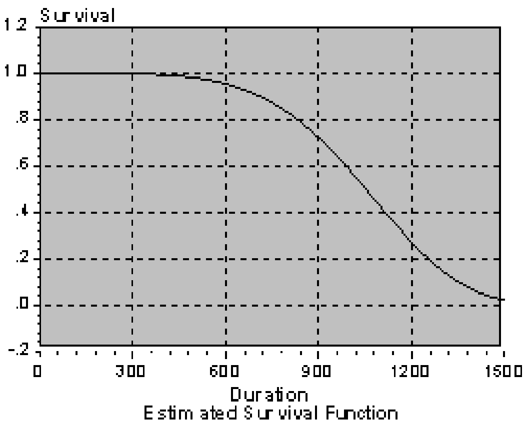
(a) Road Construction Projects



(b) Bridge Construction Projects

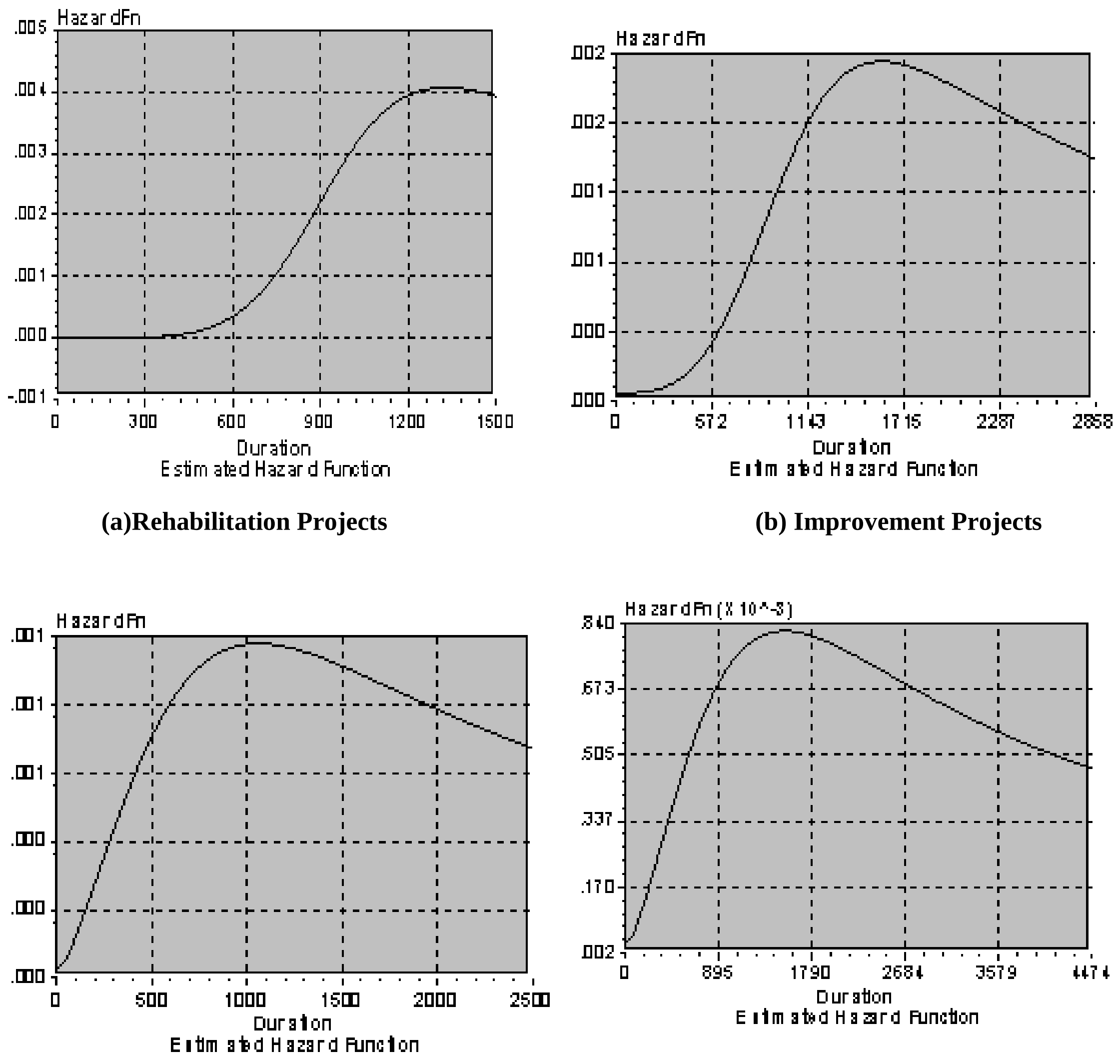


c) Improvement Projects



(d) Rehabilitation Projects

Figure 5.9 Survival plots of different projects



(a)Rehabilitation Projects **(b) Improvement Projects**

(c) Bridge Projects **(d) Construction Projects**

Figure 5.10 Log- Logistic hazard model Plots of different projects

5.10 Chapter Summary

This chapter explains the results of the statistical and probabilistic analysis of the data. It also identifies the parametric or non-parametric nature of the risks available and applies multivariate regression analysis to yield delay models to help the project estimators come up with a reliable estimate encompassing various contingencies.

Chapter 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary

Duration model represents the data pertaining to four different provinces of Pakistan exhibits some similar characteristics and is therefore grouped into two pairs. Sindh and Punjab were sited in group-1 while Balochistan and Khyber Pakhtunkhwa were placed in group-2. Each paired group was prone to the similar risk factors which affect their durations. A finding from the literature review suggested that extent of geographic location is an important factor in building projects (Drew and Skitmore 1992), in this research weak correlation between geographic location of the project and project duration was encountered. The difference could be due to the changed nature and complexities of highway projects as compared to the building projects.

Late funds release from to the funding agencies (R1), land acquisition problems (R5) and cash flow problems were the chief time overrun factors prevailing in most of project across the country. The remedial measure to overcome R1 was investigated that the funds releases on the projects should be balanced for the smooth completion of work. A number of projects are taken up simultaneously by the National Highway Authority with limited resources in hand which ultimately results into exceeding planned time estimates, it is therefore suggested that ongoing projects should accorded the priority and fresh projects should be initiated depending upon the available funds.

The appropriateness of probabilistic model specifications is also investigated in this research by using Weibull analysis which produced survival curves and hazard functions for the project models.

Accuracy estimation of the predicted models was carried out by using MAPE which identified the upper and lower extent of deviation of the true duration and delay values from the predicted models.

This is very important to keep in view that model suggested in this research are based on limited data so there is likely a possibility of many unobserved factors that may influence the duration and time overrun. Unforeseen events could not be considered while modeling the data. There are many other factors that are closely bonded to the project's performance and its timely completion, one of them being the competence of the management team. Construction management practices are in the state of continuous evolution and their consideration in the econometric models can produce more conclusive results.

This work has been published in the second T&DI Congress 2014 which is a peer-reviewed conference proceeding of ASCE.

The research is confined to the study of National Highway Authority projects in Pakistan. Data should be compared with caution with other situations.

6.2 Conclusions

The focus of the present study was to develop models for estimation of expected duration of highway projects on the basis of project cost, project type and geographical location. The project types included were pavement construction, pavement maintenance, pavement rehabilitation and bridge construction, based in four provinces of Pakistan. The data used in this study spans over the years 2001-2012. This research also investigated the risk factors leading to time overrun. Consequences of various explanatory variables were studied on the duration and time overruns of the projects and statistically significant models were presented to equip the project planners beforehand for preparation of duration estimate at the planning stage. The developed models explaining time overrun as a function of variables available at the planning stage can be used as a tool for identifying projects with high time overruns and promptly remedial measures can be put in place to mitigate the risk.

Findings in this research can assist the highway agencies in forecasting project duration during planning stage and significantly improving the process of delay mitigation which can ultimately result into more competent highway project programs. Also, the projected models can help the contractors to prepare project duration estimates for making appropriate plans for labor, equipment and resource utilization which are strongly influenced by project duration.

REFERENCES

A. Czarnigowska, A. Sobotka, (2013). “Time-cost relationship for predicting construction duration”, Archives of Civil and Mechanical Engineering, http://dx.doi.org/10.10.1016/j.acme.2013.05.004

Ahmed, S., Azhar, S., Kappagantula, P., and Gollapudi, D. (2003). “Delays in construction: A brief study of the Florida construction industry.” Proc., 39th Annual Conf. of the Associated Schools of Construction, Clemson Univ., Clemson, SC.

Anastasopoulos, P.Ch., (2007). “Performance-based contracting for roadway maintenance operations in Indiana”. M.S. Thesis, Purdue University, W. Lafayette, Ind.

Arditi, D., Akan, G.T., Gurdamer, S., (1985). “Reasons for delays in public projects in Turkey. J. Constr. Manage.” Econ. 3 (2), 171–181.

Bennett, R. (1991). "How is management research carried out in the Management Research Handbook", eds. N. C. Smith and P. Danity, 85-103. London: Routledge.

Chan, D. W. M., and Kumaraswamy, M. M. (1997). “A comparative study of causes of time overruns in Hong Kong construction projects.” Int. J. Proj. Manage., 15(1), 55–63.

Chan, A.P.C., (2001). “Time–cost relationship of public sector projects in Malaysia.” Int. J. Proj. Manage. 19, 223–229.

Chassiakos, A.P., Sakellaropoulos, S.P., (2005). “Time–cost optimization of construction projects with generalized activity constraints.” J. Constr.Manage. 131 (10), 1115–1124.

Deckro, R.F., Hebert, J.E., Verdini, W.A., Grimsrud, P.H., Venkateshwar,S., (1995). “Nonlinear time/cost trade-off models in project management.” Comput. Ind. Eng. 28 (2), 219–229.

del Caño, A., and Pilar de la Cruz, M. (2002). “Integrated methodology for project risk management.” J. Constr. Eng. Manage., 128(6), 473–484.

Drew, D. S., and Skitmore, M. (1992). “Competitiveness in bidding: A consultant’s perspective.” Constr. Manage. Econom., 10 (3), 227–247.

Elandt-Johnson, R., Johnson, N. (1999). “Survival Models and Data Analysis.” Wiley, New York.

Emory, C. W. (1980). "Business Research Methods", 2nd Edition. Illinois: Irwin & Co.

Falk, J., Horowitz, J.L., (1972). "Critical path problems with concave cost time curve." Manage. Sci. 19, 446–455.

Foldes, S., Soumis, F., (1993). "Pert and crashing revisited: mathematical generalizations." Eur. J. Oper. Res. 64, 286–294.

Frimpong, Y., Oluwoyeb, J., and Crawford, L. (2003). "Causes of delay and cost overruns in construction of groundwater projects in a developing countries; Ghana as a case study." Int. J. Proj. Manage., 21(5), 321–326.

Fulkerson, D.R., (1961). "A network flow computation for project cost curves." Manage. Sci. 7, 167–178.

Greene, W., (2007). Limdep Version 9.0. Econometric Software, Inc., Plainview, NY.

Hendrickson, C., Martinelli, D., Rehak, D., (1987). "Hierarchical rule-based activity duration estimation." J. Constr. Manage. 113 (2), 288–301.

Hosmer, D.W., Lemeshow, S., (1999). "Applied Survival Analysis: Regression Modeling of Time to Event Data." Wiley, New York.

Ibbs, C. W., and Allen, W. E. (1995). "Quantitative impacts of project changes", Univ. of California, Berkeley, Calif

Irfan, M., Khurshid M.B, Anastasopoulos. P ,Labi. S, Moavenzadeh. F, (2011). "Planning-stage estimation of highway project duration on the basis of anticipated project cost, project type, and contract type." International Journal of Project Management 29(1),78–92.

Jahren, C., and Ashe, A. (1990). "Predictors of cost-overrun rates." J. Constr. Eng. Manage., 116(3), 548–551.

Kaliba, C., Muya, M., Mumba, K., (2009). "Cost escalation and schedule delays in road construction projects in Zambia." Int. J. Proj. Manage.27 (5), 522–531.

Kaming, P., Olomolaiye, P., Holt, G., and Harris, F. (1997). "Factors influencing construction time and cost overruns on high-rise projects in Indonesia." Constr. Manage. Econom., 15(1), 83–94.

Knight, K., and Fayek, A. M. (2002). "Use of fuzzy logic for predicting design cost overruns on building projects." J. Constr. Eng. Manage., 128(6), 503–512.

Kraiem, Z.M., (1987). "Concurrent delays in construction projects." J.Constr. Manage. 113 (4), 591–601.

Kwak, Y.H., Stoddard, J., (2004). “Project risk management: lessons learned from software development environment. Technovation 24 (11), 915–920.

Lipke, W., Zwikael, O., Henderson, K., Anbari, F., (2009). “Prediction of project outcome: the application of statistical methods to earned value management and earned schedule performance indexes.” Int. J. Proj. Manage. 27, 400–407.

Lo, T. Y., Fung, I. W. H., and Tung, K. C. F. (2006). "Construction delays in Hong Kong civil engineering projects.” J. Constr. Eng. Manage., 132(6), 636–649

Moder, J.J., Phillips, C.R., Davis, E.W., (1995). “Project Management with CPM, PERT, and Precedence Diagramming”, third ed. Blitz, Middleton, WI.

Moussourakis, J., Haksever, C., (2004). “Flexible model for time–cost tradeoff problem.” J. Constr. Manage. 130 (3), 307–314.

Nassar, K.M., Gunnarsson, Hordur G., Hegab, M.Y., (2005). “Using Weibull analysis for evaluation of cost and schedule performance.” J. Constr. Manage. 131 (12), 1257–1262.

Pakkala, P. (2002). “Innovative project delivery methods for infrastructure—An international perspective”, Finnish Road Enterprise,Helsinki, Finland.

Project Management Institute, (2008b). “The Standard for Portfolio Management - Second Edition”, 2nd ed. Project Management Institute, Newtown Square, PA.

Raftery, J. (1994). “Risk analysis in project management”, E & FN Spoon, London.

Ropponen, J., Lyytinen, K., (1997). “Can software risk management improve system development: an exploratory study.” European Journal of Information Systems 6 (1), 41–50.

Shaheen, A. A., Fayek, A. R., and AbouRizk, S. M. (2007). “Fuzzy numbers in cost range estimating.” J. Constr. Eng. Manage., 133(4), 325–334
.
Skutella, M., (1998). “Approximation algorithms for the discrete time–cost trade-off problem.” Math. Oper. Res. 23 (4), 909–929.

Vandevoorde, S., Vanhoucke, M., (2006). “A comparison of different project duration forecasting methods using earned value metrics.” Int. J. Proj. Manage. 24, 289–302.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., (2003). “Statistical and Econometric Methods for Transportation Data Analysis”. Chapman & Hall/CRC.

Yang, I.T., (2005). “Chance-constrained time–cost tradeoff analysis considering funding variability. “J. Constr. Manage. 131 (9), 1002–1012.

Yang, I.T., (2007a). “Using elitist particle swarm optimization to facilitate bicriterion time–cost trade-off analysis.” J. Constr. Manage. 133 (7), 499–505.

Zheng, D.X.M., Ng, S.T., Kumaraswamy, M.M., (2004). “Applying a genetic algorithm-based multiobjective approach for time–cost optimization.” J. Constr. Manage. 130 (2), 168–176.

Zheng, D. X. M., and Ng, T. S. (2005). “Stochastic time-cost optimization model incorporating fuzzy sets theory and non replaceable front.” J. Constr. Eng. Manage., 131(2), 176–186.

APPENDICES

APPENDIX: I Probabilistic Output

CONSTRUCTION PROJECTS

```
--> RESET
--> sample;1-45$
--> read;nvar=3;nobs=45;file=C:\Const.txt$
--> reject;x1=0$
--> reject;x2=0$
--> create;ltime=LOG(x1)$
--> create;Cost=x2$
--> create;Location=x3$
--> survival;lhs=ltime;rhs=one,Cost,Location;model=weibull;plot$
```

```
+-----------------------------------------------------------------------
+
| Log-linear survival regression model: WEIBULL
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   7.237845140    , S.D.=   .5201822489
|
| Model size: Observations =      45, Parameters =   3, Deg.Fr.=     42
|
| Residuals:  Sum of squares= .8396464131D+01, Std.Dev.=         .44712
|
| Fit:        R-squared=  .294767, Adjusted R-squared =          .26118
|
| Model test: F[  2,     42] =    8.78,    Prob value =          .00065
|
| Diagnostic: Log-L =    -26.0781, Restricted(b=0) Log-L =     -33.9357
|
|             LogAmemiyaPrCrt.=   -1.545, Akaike Info. Crt.=      1.292
|
+-----------------------------------------------------------------------
+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
 Constant  6.935400095      .10335120       67.105   .0000
 COST      .1703844118E-03  .55729511E-04    3.057   .0022  1163.5915
 LOCATION  .1723127635E-02  .45863631E-03    3.757   .0002  60.464000

Normal exit from iterations. Exit status=0.

              +----------------------------------------------+
              | Loglinear survival model: WEIBULL            |
              | Maximum Likelihood Estimates                 |
              | Dependent variable                 LTIME     |
              | Weighting variable                   ONE     |
              | Number of observations                45     |
              | Iterations completed                  10     |
              | Log likelihood function        -29.45983     |
              +----------------------------------------------+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
```

```
          RHS of hazard model
 Constant  7.176099064      .97212585E-01   73.819   .0000
 COST      .1602129322E-03  .40159793E-04    3.989   .0001  1163.5915
 LOCATION  .1568101497E-02  .63921038E-03    2.453   .0142  60.464000
          Ancillary parameters for survival
 Sigma     .4196200483      .68692439E-01    6.109   .0000
```

Matrix: Las
[4,4]

```
    +--------------------------------------------------------------+
    | Parameters of underlying density at data means:              |
    | Parameter   Estimate   Std. Error     Confidence Interval    |
    | ---------------------------------------------------------    |
    | Lambda        .00058       .00004       .0005 to       .0007 |
    | P            2.38311       .39012      1.6185 to      3.1477 |
    | Median    1485.54798    108.33778   1273.2059 to   1697.8900 |
    | Percentiles  of  survival  distribution:                     |
    | Survival      .25       .50       .75       .95              |
    | Time       1987.03   1485.55   1027.14    498.19             |
    +--------------------------------------------------------------+
```

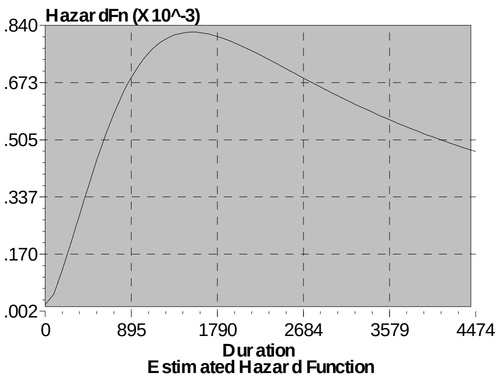
--> survival;lhs=ltime;rhs=one,Cost,Location;model=logistic;plot$

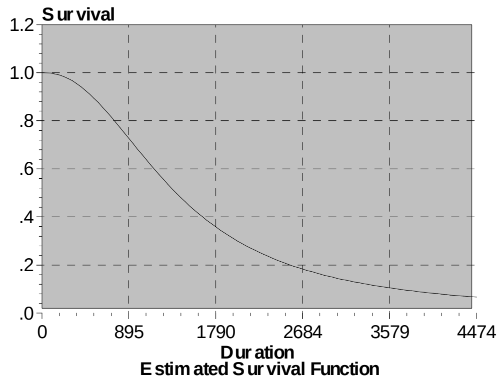
```
+-----------------------------------------------------------------------
+
| Log-linear survival regression model: LOGISTIC
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   7.237845140    , S.D.=   .5201822489
|
| Model size: Observations =      45, Parameters =   3, Deg.Fr.=     42
|
| Residuals:  Sum of squares= .8396464131D+01, Std.Dev.=         .44712
|
| Fit:        R-squared=  .294767, Adjusted R-squared =          .26118
|
| Model test: F[  2,     42] =    8.78,    Prob value =          .00065
|
| Diagnostic: Log-L =    -26.0781, Restricted(b=0) Log-L =     -33.9357
|
|             LogAmemiyaPrCrt.=   -1.545, Akaike Info. Crt.=      1.292
|
+-----------------------------------------------------------------------
+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
 Constant  6.935400095      .10335120       67.105   .0000
 COST      .1703844118E-03  .55729511E-04    3.057   .0022  1163.5915
 LOCATION  .1723127635E-02  .45863631E-03    3.757   .0002  60.464000

Maximum iterations reached. Exit iterations with status=1.
Abnormal exit from iterations. If current results are shown
check convergence values shown below. This may not be a
solution value (especially if initial iterations stopped).
Gradient value: Tolerance= .1000D-05, current value= .1098D+02
Function chg. : Tolerance= .0000D+00, current value= .4113D-02
```

```
Parameters chg: Tolerance= .0000D+00, current value= .4549D+01
Smallest abs. parameter change from start value = .2667D-05
Note:  At least one parameter did not leave start value.

              +---------------------------------------------+
              | Loglinear survival model: LOGISTIC          |
              | Maximum Likelihood Estimates                |
              | Dependent variable                 LTIME    |
              | Weighting variable                   ONE    |
              | Number of observations                45    |
              | Iterations completed                  51    |
              | Log likelihood function        -35.72669    |
              +---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.935417486      .24284334       28.559   .0000
 COST      .1677176932E-03  .13108939E-03    1.279   .2008   1163.5915
 LOCATION  .1667315962E-02  .16875435E-02     .988   .3231   60.464000
          Ancillary parameters for survival
 Sigma     .4437799907      .93801964E-01    4.731   .0000
```

Matrix: Las
[4,4]

```
    +-----------------------------------------------------------------+
    | Parameters of underlying density at data means:                 |
    | Parameter    Estimate    Std. Error     Confidence Interval     |
    | ------------------------------------------------------------    |
    | Lambda         .00072       .00012       .0005 to       .0010   |
    | P             2.25337       .47630      1.3198 to      3.1869   |
    | Median     1382.13539    221.92375    947.1648 to   1817.1059   |
    | Percentiles  of  survival  distribution:                        |
    | Survival       .25       .50       .75       .95                |
    | Time       2250.54   1382.14    848.82    374.17                |
    +-----------------------------------------------------------------+
```





BRIDGE PROJECTS

```
--> RESET
--> sample;1-34$
--> read;nvar=3;nobs=34;file=C:\Bridge.txt$
--> reject;x1=0$
--> reject;x2=0$
--> create;ltime=LOG(x1)$
--> create;Cost=x2$
--> create;Location=x3$
--> survival;lhs=ltime;rhs=one,Cost,Location;model=weibull;plot$
```

```
+-----------------------------------------------------------------------
+
| Log-linear survival regression model: WEIBULL
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.866882654    , S.D.=   .5451738399
|
| Model size: Observations =      34, Parameters =   3, Deg.Fr.=     31
|
| Residuals:  Sum of squares= .6345472952D+01, Std.Dev.=         .45243
|
| Fit:        R-squared=  .353036, Adjusted R-squared =          .31130
|
| Model test: F[  2,     31] =    8.46,    Prob value =          .00117
|
| Diagnostic: Log-L =    -19.7074, Restricted(b=0) Log-L =     -27.1103
|
|             LogAmemiyaPrCrt.=   -1.502, Akaike Info. Crt.=      1.336
|
+-----------------------------------------------------------------------
+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
 Constant  6.280726871      .16755327       37.485   .0000
 COST      .6247800266E-03  .17670342E-03    3.536   .0004  665.10209
 LOCATION  .1657537716E-02  .44043675E-03    3.763   .0002  102.93176

Normal exit from iterations. Exit status=0.

              +---------------------------------------------+
              | Loglinear survival model: WEIBULL           |
              | Maximum Likelihood Estimates                |
              | Dependent variable                LTIME     |
              | Weighting variable                  ONE     |
              | Number of observations               34     |
              | Iterations completed                 10     |
              | Log likelihood function       -19.80417     |
              +---------------------------------------------+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
```

```
+---------+-------------+---------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.668777537      .12523099       53.252   .0000
 COST      .4549058656E-03  .15894714E-03    2.862   .0042   665.10209
 LOCATION  .1038004868E-02  .73845767E-03    1.406   .1598   102.93176
          Ancillary parameters for survival
 Sigma     .3748668823      .64943954E-01    5.772   .0000
```

Matrix: Las
[4,4]

```
    +-----------------------------------------------------------+
    | Parameters of underlying density at data means:           |
    | Parameter    Estimate   Std. Error     Confidence Interval  |
    | ------------------------------------------------------------ |
    | Lambda         .00084       .00008       .0007 to      .0010 |
    | P             2.66761       .46215      1.7618 to     3.5734 |
    | Median     1033.58280     93.48953    850.3433 to  1216.8223 |
    | Percentiles  of  survival  distribution:                    |
    | Survival       .25        .50        .75        .95          |
    | Time        1340.27    1033.58     743.33     389.46          |
    +-----------------------------------------------------------+
```

--> survival;lhs=ltime;rhs=one,Cost,Location;model=logistic;plot$

```
+-----------------------------------------------------------------------
+
| Log-linear survival regression model: LOGISTIC
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.866882654    , S.D.=   .5451738399
|
| Model size: Observations =      34, Parameters =   3, Deg.Fr.=     31
|
| Residuals:  Sum of squares= .6345472952D+01, Std.Dev.=         .45243
|
| Fit:        R-squared=  .353036, Adjusted R-squared =          .31130
|
| Model test: F[  2,     31] =    8.46,     Prob value =          .00117
|
| Diagnostic: Log-L =    -19.7074, Restricted(b=0) Log-L =     -27.1103
|
|             LogAmemiyaPrCrt.=   -1.502, Akaike Info. Crt.=      1.336
|
+-----------------------------------------------------------------------
+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient   | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
 Constant  6.280726871      .16755327       37.485   .0000
 COST      .6247800266E-03  .17670342E-03    3.536   .0004   665.10209
 LOCATION  .1657537716E-02  .44043675E-03    3.763   .0002   102.93176
```

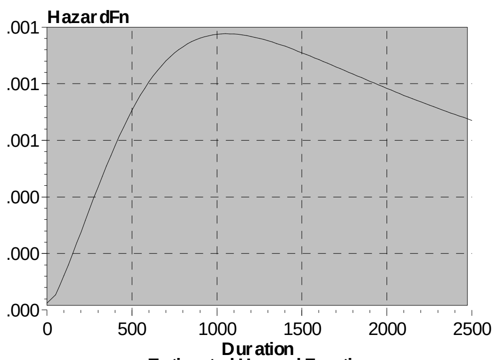
Maximum iterations reached. Exit iterations with status=1.
Abnormal exit from iterations. If current results are shown
check convergence values shown below. This may not be a
solution value (especially if initial iterations stopped).

```
Gradient value: Tolerance= .1000D-05, current value= .1978D+02
Function chg. : Tolerance= .0000D+00, current value= .2054D-02
Parameters chg: Tolerance= .0000D+00, current value= .1864D+02
Smallest abs. parameter change from start value = .8734D-05
Note:  At least one parameter did not leave start value.
```

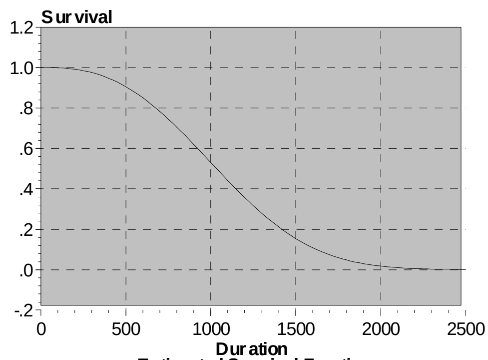
```
              +---------------------------------------------+
              | Loglinear survival model: LOGISTIC          |
              | Maximum Likelihood Estimates                |
              | Dependent variable                 LTIME    |
              | Weighting variable                   ONE    |
              | Number of observations                34    |
              | Iterations completed                  51    |
              | Log likelihood function        -27.10484    |
              +---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.280636470      .36995347      16.977   .0000
 COST      .6337165696E-03  .45023024E-03    1.408   .1593   665.10209
 LOCATION  .1666271500E-02  .15408575E-02    1.081   .2795   102.93176
          Ancillary parameters for survival
 Sigma     .4502595485      .99521050E-01    4.524   .0000
```

Matrix: Las
[4,4]

```
+-----------------------------------------------------------------+
| Parameters of underlying density at data means:                 |
| Parameter    Estimate    Std. Error     Confidence Interval     |
| ------------------------------------------------------------    |
| Lambda         .00103       .00022       .0006 to       .0015   |
| P             2.22094       .49090      1.2588 to      3.1831   |
| Median      966.45520    209.76647    555.3129 to   1377.5975   |
| Percentiles  of  survival  distribution:                        |
| Survival        .25        .50        .75        .95            |
| Time        1584.93     966.46     589.32     256.69            |
+-----------------------------------------------------------------+
```



Estimated Hazard Function



Estimated Survival Function

DELAY MODELS

```
--> RESET
--> sample;1-53$
--> read;nvar=3;nobs=53;file=C:\Delay.txt$
--> reject;x1=0$
--> reject;x2=0$
--> create;ltime=LOG(x1)$
--> create;Cost=x2$
--> create;Risk=x3$
--> survival;lhs=ltime;rhs=one,Cost,Risk;model=weibull;plot$
```

```
+-----------------------------------------------------------------------
+
| Log-linear survival regression model: WEIBULL
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.390918480    , S.D.=   .6377766651
|
| Model size: Observations =      53, Parameters =   3, Deg.Fr.=     50
|
| Residuals:  Sum of squares= .1421417386D+02, Std.Dev.=         .53318
|
| Fit:        R-squared=  .327982, Adjusted R-squared =          .30110
|
| Model test: F[  2,     50] =   12.20,    Prob value =          .00005
|
| Diagnostic: Log-L =    -40.3284, Restricted(b=0) Log-L =     -50.8613
|
|             LogAmemiyaPrCrt.=   -1.203, Akaike Info. Crt.=      1.635
|
+-----------------------------------------------------------------------
+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
 Constant  6.183351969      .26719953       23.141   .0000
 COST      .3572044608E-03  .76282997E-04    4.683   .0000  1298.3443
 RISK     -.2828962000      .25168375       -1.124   .2610  .90566038

Normal exit from iterations. Exit status=0.

              +---------------------------------------------+
              | Loglinear survival model: WEIBULL           |
              | Maximum Likelihood Estimates                |
              | Dependent variable                LTIME     |
              | Weighting variable                  ONE     |
              | Number of observations               53     |
              | Iterations completed                 10     |
              | Log likelihood function       -35.11948     |
              +---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.455428252      .24882923       25.943   .0000
```

```
 COST      .3033127962E-03  .67340665E-04    4.504   .0000  1298.3443
 RISK     -.2452470378      .22828893       -1.074   .2827  .90566038
          Ancillary parameters for survival
 Sigma     .3886040235      .55225011E-01    7.037   .0000
```

Matrix: Las
[4,4]

```
    +------------------------------------------------------------+
    | Parameters of underlying density at data means:             |
    | Parameter   Estimate   Std. Error     Confidence Interval   |
    | ------------------------------------------------------------ |
    | Lambda        .00132       .00008       .0012 to       .0015 |
    | P            2.57331       .36570      1.8565 to      3.2901 |
    | Median     655.03835     41.23349    574.2207 to    735.8560 |
    | Percentiles  of  survival  distribution:                    |
    | Survival       .25        .50        .75        .95         |
    | Time        857.53     655.04     465.43     238.15         |
    +------------------------------------------------------------+
```

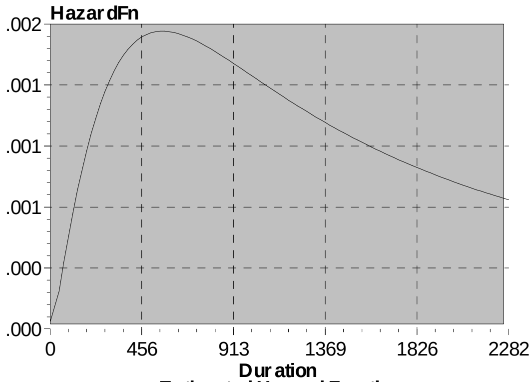
--> survival;lhs=ltime;rhs=one,Cost,Risk;model=logistic;plot$

```
+-----------------------------------------------------------------------
+
| Log-linear survival regression model: LOGISTIC
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.390918480    , S.D.=   .6377766651
|
| Model size: Observations =      53, Parameters =   3, Deg.Fr.=     50
|
| Residuals:  Sum of squares= .1421417386D+02, Std.Dev.=         .53318
|
| Fit:        R-squared=  .327982, Adjusted R-squared =          .30110
|
| Model test: F[  2,     50] =   12.20,    Prob value =          .00005
|
| Diagnostic: Log-L =    -40.3284, Restricted(b=0) Log-L =     -50.8613
|
|             LogAmemiyaPrCrt.=   -1.203, Akaike Info. Crt.=      1.635
|
+-----------------------------------------------------------------------
+
+---------+-------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+----------------+--------+---------+----------
+
 Constant  6.183351969      .26719953       23.141   .0000
 COST      .3572044608E-03  .76282997E-04    4.683   .0000  1298.3443
 RISK     -.2828962000      .25168375       -1.124   .2610  .90566038

Maximum iterations reached. Exit iterations with status=1.
Abnormal exit from iterations. If current results are shown
check convergence values shown below. This may not be a
solution value (especially if initial iterations stopped).
Gradient value: Tolerance= .1000D-05, current value= .1716D+02
Function chg. : Tolerance= .0000D+00, current value= .3527D-03
Parameters chg: Tolerance= .0000D+00, current value= .9120D+01
Smallest abs. parameter change from start value = .4716D-05
```

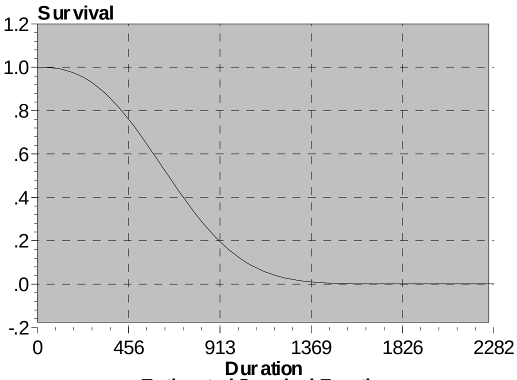
```
Note:  At least one parameter did not leave start value.

              +---------------------------------------------+
              | Loglinear survival model: LOGISTIC           |
              | Maximum Likelihood Estimates                 |
              | Dependent variable                 LTIME     |
              | Weighting variable                   ONE     |
              | Number of observations                53     |
              | Iterations completed                  51     |
              | Log likelihood function        -51.45583     |
              +---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.183411606      .75627312       8.176   .0000
 COST       .3619206197E-03 .21258129E-03   1.703   .0887  1298.3443
 RISK      -.2828656162     .68855125       -.411   .6812  .90566038
          Ancillary parameters for survival
 Sigma      .5327742696     .10264329       5.191   .0000
```

Matrix: Las
[4,4]

```
    +-----------------------------------------------------------------+
    | Parameters of underlying density at data means:                 |
    | Parameter    Estimate    Std. Error     Confidence Interval     |
    | -------------------------------------------------------------   |
    | Lambda         .00167        .00031        .0011 to       .0023 |
    | P             1.87697        .36161       1.1682 to      2.5857 |
    | Median      600.11963     112.60416     379.4155 to    820.8238 |
    | Percentiles  of  survival  distribution:                        |
    | Survival        .25         .50         .75         .95         |
    | Time        1077.55     600.12      334.23      125.01          |
    +-----------------------------------------------------------------+
```



Estimated Hazard Function



Estimated Survival Function

DURATION MODELS

```
--> sample;1-62$
--> read;nvar=3;nobs=62;file=C:\Duration.txt$
--> reject;x1=0$
--> reject;x2=0$
--> create;ltime=LOG(x1)$


First 5 errors: Variable Subcommand    Row      Error
                   LTIME      1         51      Logminus
            Error      Count
            Logminus       1
--> create;Cost=x2$
--> create;Location=x3$
--> survival;lhs=ltime;rhs=one,Cost,Location;model=weibull;plot$

+-----------------------------------------------------------------------
+
| Log-linear survival regression model: WEIBULL
|
| Least squares is used to obtain starting values for MLE.
|
```

```
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.847668184    , S.D.=   .9923794455
|
| Model size: Observations =      62, Parameters =   3, Deg.Fr.=     59
|
| Residuals:  Sum of squares= .4783242587D+02, Std.Dev.=         .90040
|
| Fit:        R-squared=  .203773, Adjusted R-squared =          .17678
|
| Model test: F[  2,     59] =    7.55,    Prob value =          .00120
|
| Diagnostic: Log-L =    -79.9318, Restricted(b=0) Log-L =     -86.9958
|
|             LogAmemiyaPrCrt.=    -.163, Akaike Info. Crt.=      2.675
|
+------------------------------------------------------------------------
+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
 Constant  6.008342587      .26398966       22.760   .0000
 COST      .3565708968E-03  .92275378E-04    3.864   .0001  1661.9591
 LOCATION  .4635333165      .24835466        1.866   .0620  .53225806

Normal exit from iterations. Exit status=0.

              +---------------------------------------------+
              | Loglinear survival model: WEIBULL            |
              | Maximum Likelihood Estimates                 |
              | Dependent variable                  LTIME    |
              | Weighting variable                    ONE    |
              | Number of observations                 62    |
              | Iterations completed                   13    |
              | Log likelihood function        -38.68624    |
              +---------------------------------------------+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.641942049      .12829473       51.771   .0000
 COST      .2527531653E-03  .46622277E-04    5.421   .0000  1661.9591
 LOCATION  .3588211516E-01  .13772405         .261   .7945  .53225806
          Ancillary parameters for survival
 Sigma     .3571068661      .18555285E-01   19.246   .0000
```

Matrix: Las
[4,4]

```
    +------------------------------------------------------------------+
    | Parameters of underlying density at data means:                  |
    | Parameter   Estimate   Std. Error     Confidence Interval        |
    | ------------------------------------------------------------     |
    | Lambda        .00084       .00005       .0008 to       .0009     |
    | P            2.80028       .14550      2.5151 to      3.0855     |
    | Median    1043.38106     57.25488    931.1615 to   1155.6006     |
    | Percentiles  of  survival  distribution:                         |
```

```
    | Survival       .25       .50       .75       .95              |
    | Time       1336.42   1043.38    762.18    411.76              |
    +---------------------------------------------------------------+
```

--> survival;lhs=ltime;rhs=one,Cost,Location;model=logistic;plot$

```
+------------------------------------------------------------------------
+
| Log-linear survival regression model: LOGISTIC
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.847668184    , S.D.=   .9923794455
|
| Model size: Observations =      62, Parameters =   3, Deg.Fr.=     59
|
| Residuals:  Sum of squares= .4783242587D+02, Std.Dev.=         .90040
|
| Fit:        R-squared=  .203773, Adjusted R-squared =          .17678
|
| Model test: F[  2,     59] =    7.55,    Prob value =          .00120
|
| Diagnostic: Log-L =    -79.9318, Restricted(b=0) Log-L =     -86.9958
|
|             LogAmemiyaPrCrt.=    -.163, Akaike Info. Crt.=      2.675
|
+------------------------------------------------------------------------
+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
 Constant  6.008342587      .26398966       22.760   .0000
 COST      .3565708968E-03  .92275378E-04    3.864   .0001  1661.9591
 LOCATION  .4635333165      .24835466        1.866   .0620  .53225806

Line search does not improve fn. Exit iterations. Status=3
Abnormal exit from iterations. If current results are shown
check convergence values shown below. This may not be a
solution value (especially if initial iterations stopped).
Gradient value: Tolerance= .1000D-05, current value= .4611D+00
Function chg. : Tolerance= .0000D+00, current value= .3793D-04
Parameters chg: Tolerance= .0000D+00, current value= .1752D+04
      Smallest abs. parameter change from start value = .9334D-04
```

Improvement Models

--> RESET
--> sample;1-21$
--> read;nvar=3;nobs=21;file=C:\Improvement.txt$
--> reject;x1=0$
--> reject;x2=0$
--> create;ltime=LOG(x1)$
--> create;Cost=x2$
--> create;Location=x3$
--> survival;lhs=ltime;rhs=one,Cost,Location;model=weibull;plot$

```
+------------------------------------------------------------------------
+
```

```
| Log-linear survival regression model: WEIBULL
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   7.046879640    , S.D.=   .5294820676
|
| Model size: Observations =      21, Parameters =   3, Deg.Fr.=    18
|
| Residuals:  Sum of squares= .4096657697D+01, Std.Dev.=         .47707
|
| Fit:        R-squared=  .269371, Adjusted R-squared =          .18819
|
| Model test: F[  2,     18] =    3.32,    Prob value =          .05933
|
| Diagnostic: Log-L =    -12.6370, Restricted(b=0) Log-L =     -15.9324
|
|             LogAmemiyaPrCrt.=   -1.347, Akaike Info. Crt.=      1.489
|
+-----------------------------------------------------------------------
+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
 Constant  6.670942467      .17939952       37.185   .0000
 COST      .2584148379E-03  .10633046E-03    2.430   .0151  995.17989
 LOCATION  .1766378471E-02  .92929346E-03    1.901   .0573  67.238095

Normal exit from iterations. Exit status=0.

              +---------------------------------------------+
              | Loglinear survival model: WEIBULL           |
              | Maximum Likelihood Estimates                |
              | Dependent variable               LTIME      |
              | Weighting variable                 ONE      |
              | Number of observations              21      |
              | Iterations completed                11      |
              | Log likelihood function      -12.42479      |
              +---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  7.057219542      .24654376       28.625   .0000
 COST      .1658808002E-03  .22522760E-03     .737   .4614  995.17989
 LOCATION  .5155223931E-03  .90059705E-02     .057   .9544  67.238095
          Ancillary parameters for survival
 Sigma     .3854204476      .70487765E-01    5.468   .0000
```

Matrix: Las
[4,4]

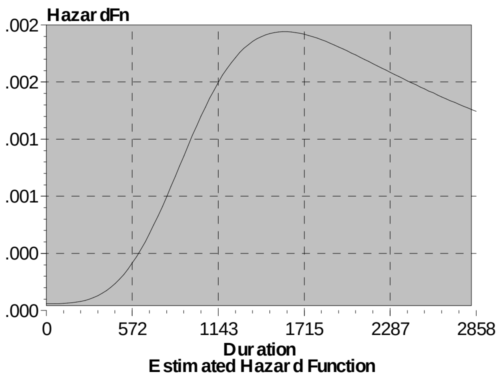
```
    +-----------------------------------------------------------------+
    | Parameters of underlying density at data means:                 |
    | Parameter   Estimate   Std. Error     Confidence Interval       |
    | ------------------------------------------------------------    |
```

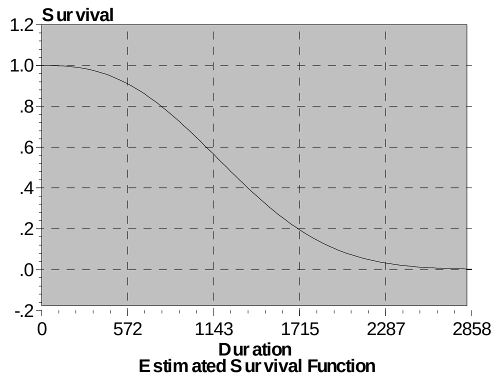
```
    | Lambda         .00071        .00043      -.0001 to       .0015   |
    | P             2.59457        .47451      1.6645 to      3.5246   |
    | Median     1231.14751     747.49740   -233.9474 to   2696.2424   |
    | Percentiles  of  survival  distribution:                         |
    | Survival        .25        .50       .75       .95               |
    | Time       1608.17   1231.15    877.23    451.33                 |
    +------------------------------------------------------------------+
--> survival;lhs=ltime;rhs=one,Cost,Location;model=logistic;plot$

+-----------------------------------------------------------------------
+
| Log-linear survival regression model: LOGISTIC
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   7.046879640    , S.D.=   .5294820676
|
| Model size: Observations =      21, Parameters =   3, Deg.Fr.=     18
|
| Residuals:  Sum of squares= .4096657697D+01, Std.Dev.=         .47707
|
| Fit:        R-squared=  .269371, Adjusted R-squared =          .18819
|
| Model test: F[  2,     18] =    3.32,    Prob value =          .05933
|
| Diagnostic: Log-L =    -12.6370, Restricted(b=0) Log-L =     -15.9324
|
|             LogAmemiyaPrCrt.=   -1.347, Akaike Info. Crt.=      1.489
|
+-----------------------------------------------------------------------
+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
 Constant  6.670942467      .17939952       37.185   .0000
 COST      .2584148379E-03  .10633046E-03    2.430   .0151  995.17989
 LOCATION  .1766378471E-02  .92929346E-03    1.901   .0573  67.238095

Normal exit from iterations. Exit status=0.

              +----------------------------------------------+
              | Loglinear survival model: LOGISTIC           |
              | Maximum Likelihood Estimates                 |
              | Dependent variable                 LTIME     |
              | Weighting variable                   ONE     |
              | Number of observations                21     |
              | Iterations completed                  11     |
              | Log likelihood function        -12.25724     |
              +----------------------------------------------+
+---------+-------------+---------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+---------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.796504583      .22111153       30.738   .0000
 COST      .2032710775E-03  .20073599E-03    1.013   .3112  995.17989
 LOCATION  .1353757504E-02  .26248103E-02     .516   .6060  67.238095
```

```
          Ancillary parameters for survival
 Sigma     .2354214744      .49530796E-01    4.753   .0000
```

Matrix: Las
[4,4]

```
+------------------------------------------------------------------+
| Parameters of underlying density at data means:                  |
| Parameter   Estimate   Std. Error     Confidence Interval        |
| ------------------------------------------------------------    |
| Lambda         .00083      .00016        .0005 to        .0011   |
| P             4.24770      .89368       2.4961 to       5.9993   |
| Median     1199.69177   225.80479     757.1144 to    1642.2692   |
| Percentiles  of  survival  distribution:                         |
| Survival       .25       .50       .75       .95                 |
| Time       1553.80   1199.69    926.29    599.82                 |
+------------------------------------------------------------------+
```





Rehabilitation Models

```
--> RESET
--> sample;1-20$
--> read;nvar=3;nobs=20;file=C:\Rehab.txt$
--> reject;x1=0$
--> reject;x2=0$
--> create;ltime=LOG(x1)$
--> create;Cost=x2$
--> create;Location=x3$
--> survival;lhs=ltime;rhs=one,Cost,Location;model=weibull;plot$
```

```
+------------------------------------------------------------------------
+
| Log-linear survival regression model: WEIBULL
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.906315954    , S.D.=   .3392936126
|
| Model size: Observations =      20, Parameters =   3, Deg.Fr.=     17
|
```

```
| Residuals:  Sum of squares= .1598894816D+01, Std.Dev.=         .30668
|
| Fit:        R-squared=  .269004, Adjusted R-squared =          .18300
|
| Model test: F[  2,     17] =    3.13,    Prob value =          .06971
|
| Diagnostic: Log-L =     -3.1146, Restricted(b=0) Log-L =      -6.2480
|
|             LogAmemiyaPrCrt.=   -2.224, Akaike Info. Crt.=       .611
|
+--------------------------------------------------------------------
+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
 Constant  6.582087732      .21788472      30.209   .0000
 COST      .2346815499E-03  .20239456E-03   1.160   .2462  748.00276
 LOCATION  .1319850475E-02  .57957917E-03   2.277   .0228  112.65350

Normal exit from iterations. Exit status=0.

              +---------------------------------------------+
              | Loglinear survival model: WEIBULL           |
              | Maximum Likelihood Estimates                |
              | Dependent variable                 LTIME    |
              | Weighting variable                   ONE    |
              | Number of observations                20    |
              | Iterations completed                  12    |
              | Log likelihood function       -.9170032     |
              +---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+--------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.874030895      .12788986      53.750   .0000
 COST      .1090861651E-03  .11129613E-03    .980   .3270  748.00276
 LOCATION  .7015755202E-03  .74498348E-03    .942   .3463  112.65350
          Ancillary parameters for survival
 Sigma     .2074652506      .41006718E-01   5.059   .0000
```

Matrix: Las
[4,4]

```
    +----------------------------------------------------------------+
    | Parameters of underlying density at data means:                |
    | Parameter    Estimate    Std. Error     Confidence Interval     |
    | -------------------------------------------------------------  |
    | Lambda         .00088       .00007       .0007 to       .0010  |
    | P             4.82008       .95272      2.9528 to      6.6874  |
    | Median     1052.18464     83.30799    888.9010 to   1215.4683  |
    | Percentiles  of  survival  distribution:                       |
    | Survival       .25        .50        .75       .95             |
    | Time       1214.91    1052.18     876.71    613.05             |
    +----------------------------------------------------------------+
--> survival;lhs=ltime;rhs=one,Cost,Location;model=logistic;plot$
```

```
+---------------------------------------------------------------------
+
| Log-linear survival regression model: LOGISTIC
|
| Least squares is used to obtain starting values for MLE.
|
| Ordinary    least squares regression    Weighting variable = none
|
| Dep. var. = LTIME    Mean=   6.906315954    , S.D.=   .3392936126
|
| Model size: Observations =      20, Parameters =   3, Deg.Fr.=     17
|
| Residuals:  Sum of squares= .1598894816D+01, Std.Dev.=         .30668
|
| Fit:        R-squared=  .269004, Adjusted R-squared =          .18300
|
| Model test: F[  2,     17] =    3.13,    Prob value =          .06971
|
| Diagnostic: Log-L =     -3.1146, Restricted(b=0) Log-L =      -6.2480
|
|             LogAmemiyaPrCrt.=   -2.224, Akaike Info. Crt.=       .611
|
+---------------------------------------------------------------------
+
+---------+-------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+----------------+--------+---------+----------
+
 Constant  6.582087732      .21788472       30.209   .0000
 COST      .2346815499E-03  .20239456E-03    1.160   .2462   748.00276
 LOCATION  .1319850475E-02  .57957917E-03    2.277   .0228   112.65350

Normal exit from iterations. Exit status=0.

              +---------------------------------------------+
              | Loglinear survival model: LOGISTIC          |
              | Maximum Likelihood Estimates                |
              | Dependent variable               LTIME      |
              | Weighting variable                 ONE      |
              | Number of observations              20      |
              | Iterations completed                10      |
              | Log likelihood function      -2.843067      |
              +---------------------------------------------+
+---------+-------------+----------------+--------+---------+----------
+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of
X|
+---------+-------------+----------------+--------+---------+----------
+
          RHS of hazard model
 Constant  6.637195375      .15071419       44.038   .0000
 COST      .2059035474E-03  .15338922E-03    1.342   .1795   748.00276
 LOCATION  .1192419945E-02  .77137869E-03    1.546   .1221   112.65350
          Ancillary parameters for survival
 Sigma     .1564500182      .30893588E-01    5.064   .0000
```

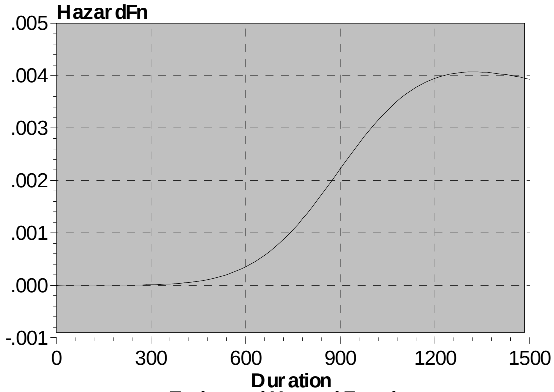
Matrix: Las
[4,4]

```
    +-----------------------------------------------------------------+
    | Parameters of underlying density at data means:                 |
```

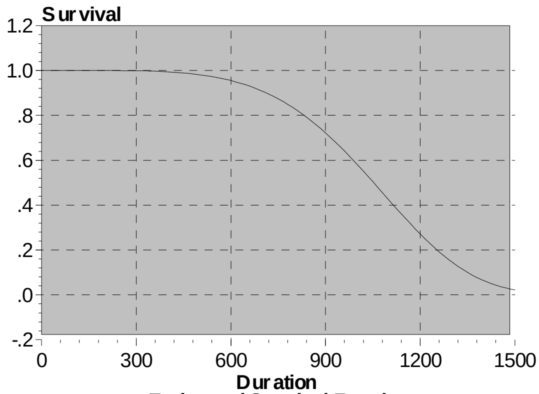
Parameter	Estimate	Std. Error	Confidence Interval		
Lambda	.00098	.00009	.0008	to	.0011
P	6.39182	1.26217	3.9180	to	8.8657
Median	1017.94592	88.33181	844.8156	to	1191.0763

Percentiles of survival distribution:

Survival	.25	.50	.75	.95
Time	1208.84	1017.95	857.19	642.19



Estimated Hazard Function



Estimated Survival Function

APPENDIX-II

Deterministic Outputs

Descriptives

			Statistic	Std. Error
FREQUENCY	Mean		11.3750	5.83658
	95% Confidence Interval for Mean	Lower Bound	-1.0654	
		Upper Bound	23.8154	
	5% Trimmed Mean		7.3056	
	Median		3.0000	
	Variance		545.050	
	Std. Deviation		23.34631	
	Minimum		1.00	
	Maximum		95.00	
	Range		94.00	
	Interquartile Range		12.00	
	Skewness		3.458	.564
	Kurtosis		12.727	1.091

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
FREQUENCY	16	100.0%	0	.0%	16	100.0%

Bridge Construction

SUMMARY OUTPUT

Regression Statistics	
Multiple R	*0.580511254*
R Square	*0.336993316*
Adjusted R Square	*0.294218691*
Standard Error	*453.6209407*
Observations	*34*

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	*2*	*3242286*	*1621143*	*7.878347*	*0.001712*
Residual	*31*	*6378931*	*205772*		
Total	*33*	*9621216*			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	*663.5724425*	*273.5146*	*2.426095*	*0.021274*	*105.7358*	*1221.409*	*105.7358*	*1221.409*
Completed Indexed Cost (X-2)	*0.496081319*	*0.125895*	*3.94045*	*0.000431*	*0.239318*	*0.752845*	*0.239318*	*0.752845*
Geographical Location	*-89.13132707*	*279.9302*	*-2.31841*	*0.752313*	*-660.053*	*481.79*	*-660.053*	*481.79*

Construction

Regression Statistics	
Multiple R	0.635357769
R Square	0.403679495
Adjusted R Square	0.37528328
Standard Error	663.1250639
Observations	45

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	*2*	*12502507.48*	*6251254*	*14.215962*	*1.92786E-05*
Residual	*42*	*18468863.72*	*439734.9*		
Total	*44*	*30971371.2*			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	*871.622942*	*205.6502375*	*4.238376*	*0.0001206*	*456.6039606*	*1286.642*	*456.604*	*1286.642*
Completed Indexed Cost (X-2)	*0.402176772*	*0.077318087*	*5.201587*	*5.52E-06*	*0.246142556*	*0.558211*	*0.246143*	*0.558211*
Geographical Dummy	*97.09058462*	*208.8320461*	*2.464922*	*0.6443889*	*-324.3495466*	*518.5307*	*-324.35*	*518.5307*

Improvement

Regression Statistics	
Multiple R	*0.528370636*
R Square	*0.279175529*
Adjusted R Square	*0.199083922*
Standard Error	*500.0534793*
Observations	*21*

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	*2*	*1743224.169*	*871612.0843*	*3.485703*	*0.052537*
Residual	*18*	*4500962.678*	*250053.4821*		
Total	*20*	*6244186.847*			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	*732.9106525*	*262.5312767*	*2.791707951*	*0.012049*	*181.3529*	*1284.468*	*181.3529*	*1284.468*
Completed Indexed Cost (X-2)	*0.325745082*	*0.123372636*	*2.640334953*	*0.016625*	*0.066549*	*0.584941*	*0.066549*	*0.584941*
Geographical Location	*246.5838176*	*237.8091171*	*2.036898083*	*0.313513*	*-253.035*	*746.2022*	*-253.035*	*746.2022*

Rehabilitation

Regression Statistics	
Multiple R	0.651217887
R Square	0.424084737
Adjusted R Square	0.35633
Standard Error	240.0204965
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	2	721173.2918	360586.6	6.259116	0.009184
Residual	17	979367.2582	57609.84		
Total	19	1700540.55			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	254.8016993	249.9675232	1.019339	0.32233	-272.584	782.1871	-272.584	782.1870689
Completed Indexed Cost (X-2)	0.620438882	0.177172634	3.501889	0.002733	0.246637	0.99424	0.246637	0.994240462
Location	86.42300991	150.3366888	2.574863	0.572914	-230.76	403.6057	-230.76	403.6056949

Delay

Regression Statistics	
Multiple R	0.726874533
R Square	0.528346587
Adjusted R Square	0.509480451
Standard Error	294.4065419
Observations	53

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4854682.613	2427341	28.00502	6.92747E-09
Residual	50	4333760.595	86675.21		
Total	52	9188443.208			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	480.9178758	147.5391691	3.259595	0.002011	184.5767334	777.259	184.5767	777.259
Cost	0.303849488	0.04212107	7.213717	2.78E-09	0.21924683	0.388452	0.219247	0.388452
R1	-181.0252991	138.9718454	-1.3026	0.198678	-460.1584655	98.10787	-460.158	98.10787

Duration

Regression Statistics	
Multiple R	0.747008265
R Square	0.558021348
Adjusted R Square	0.54303902
Standard Error	442.337543
Observations	62

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	14575019.16	7287510	37.24531	3.46E-11
Residual	59	11544087.61	195662.5		
Total	61	26119106.77			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	433.6078345	129.6897145	3.343425	0.001442	174.0993	693.1164	174.0993	693.1164
Cost	0.379406852	0.045331949	8.369524	1.3E-11	0.288698	0.470116	0.288698	0.470116
Geographical Location	156.1719118	122.0087365	1.280006	0.205553	-87.967	400.3108	-87.967	400.3108

APPENDIX III:

Research Paper

(Kaleem, S., Irfan, M., and Gabrial, H. F. (2014). Estimation of Highway Project Duration at the Planning Stage and Analysis of Risk Factors Leading To Time Overrun. Presented at the Second ASCE T&DI Congress, June 8-11, Orlando, FL.)

Estimation of Highway Project Duration at the Planning Stage and Analysis of Risk Factors Leading To Time Overrun

Sidra Kaleem[1], Muhammad Irfan[2], and Hamza F. Gabriel[3]

[1]Graduate Research Assistant, NIT/ SCEE, National University of Sciences and Technology (NUST), Islamabad, Pakistan; email: sidrakaleem29@hotmail.com
[2]Associate Professor, NIT/ SCEE, National University of Sciences and Technology (NUST), Islamabad, Pakistan; email: drirfan-nit@nust.edu.pk
[3]Associate Professor, SCEE, National University of Sciences and Technology (NUST), Islamabad, Pakistan; email: hamza.gabriel@nice.nust.edu.pk

ABSTRACT

Estimation of time duration of highway projects at the planning stage serves as a vital input for construction planning, scheduling and contract administration. However time overrun, resulting from various factors is the most cardinal issue which eventually leads to cost overrun and hence induces turbulence in the initial cost and time estimates. In this paper, highway project duration is estimated on the basis of variables such as planned cost and project type which are known at the planning phase. Data comprises of project types such as pavement construction, improvement, rehabilitation and bridge construction projects of National Highway Authority, Pakistan. A mathematical relationship between highway project duration, planned cost and project type is demonstrated in this paper by using various model specifications. Furthermore using multivariate regression analysis correlation of the time overrun with potential risk factors is investigated encompassing attributes such as project type, cost and geographical location. This paper identifies a number of significant risk variables and their severity that contributes to the extensive delays and consequently exceeds the planned time estimates. The models developed can assist the project administrators in determining improved estimates of project duration and enhancing the expected delay estimation in completion time of planned projects.

INTRODUCTION

Realistic and precise planning helps to derive maximum benefits out of investments. The key goal of every project is to achieve work completion on time and within the specified budget. Transformation of paper drawings into concrete form while ensuring quality and safety is indeed a daunting task. As time is the essence of every project, development of reliable duration estimates can help agencies to deliver optimum project schedules and thus avoid issues pertaining to time overruns.

Success rate of most of the highway projects in Pakistan is relatively unsatisfactory; the commonality among the projects is time overrun and cost overrun. These two reasons can be attributed by various risk factors arising at every stage of the project development. It's known worldwide that economic sustainability depends largely upon the expansion of the existing facilities along with their modernization to meet the growing needs. For the rapid socio-economic uplift government of Pakistan desires to execute maximum public development projects in short span of time. 75% to 80% of Pakistan's total commercial traffic is carried by National Highway Network and Motorway system. NHA's official website reports that two-third of the total road network is in relatively poor condition. This scenario indicates a number of upcoming rehabilitation, construction and improvement projects in the pipeline.

Early and reliable cost estimates are essential inputs for decision making in the initial stages of construction projects (Czarnigowska et al. 2013). Project duration is essentially needed for proper project planning and contract administration. Prior knowledge of the project expected duration can be useful in bid evaluation and life cycle cost analysis (Irfan et al. 2010). Considerable reliance on engineering judgment is made while a project is planned (Hendrickson et al. 1987) though early studies have proposed that only the completion of the project marks its true duration though the final duration is adversely affected by some aberrations (Arditi et al. 1985; Kraiem 1987; Majid and McCaffer 1998). These aberrations could include extreme weather conditions, financial delays, skilled labor scarcity, political situations, force majeure and other project related changes taking place at various phases of the project's life cycle.

Construction project duration is a very important factor for the client, consultant and the contractor, yet delay is a typical phenomenon which is bound to occur as construction projects are seldom on schedule, often delays are among the most critical construction disputes (Kraiem 1987). Ahmed et al (2003) regards delay as universal phenomenon which is usually accompanied by cost overrun.

Kaliba et al. (2009) identified that the duration of road construction projects in Zambia is influenced by economic problem, contract modification, material procurement, delayed payment, change in specification and drawings, construction mistakes and poor supervision and coordination on site. Similar studies have been conducted worldwide however no such problem occurs at the planning phase and the root causes for delay are unknown when a project is planned. Nevertheless knowing the probability of occurrence of a certain problem in a specific region can help the project planners in various aspects. They can act pro-actively and prepare a contingency plan keeping the historical risk factors under consideration. Hence adverse effects of potential risks can be minimized.

Planning being one of the major chunks of work sets the milestones for design and execution phases of the projects. The cardinal objective of every project is to achieve work completion on time and within the specified budget while ensuring no compromise on quality and safety. Several studies have been carried out regarding the estimation of project duration and evaluation of risk factors in highway projects. This study aims to add to the body of knowledge by estimating the project duration and identifying the potential risk factors which affect the highway projects and ultimately result into time overrun. The developed models can also act as empirical tools for the contractors who may find it useful for making appropriate project plans for equipment mobilization, material utilization and resource optimization.

LITERATURE REVIEW

Several studies on the nature of relationship between project duration and project cost annotate shifts in their underlying philosophy. Early studies assume a linear relationship between project duration and project cost (Fulkerson 1961) but subsequent studies showed flexibility by using variety of nonlinear mathematical functions that include discrete formulations (Skutella 1998; Zheng et al. 2004) convex (Foldes and Soumis 1993), concave (Falk and Horowitz 1972), hybrid of convex and concave (Moder et al., 1995) or quadratic (Deckro et al. 1995). Hierarchical rule based activity duration models were estimated by Hendrickson et al (1987). Chan (2001) carried out a study in Malaysia to estimate average project duration using a time-cost formula expressed as Duration= K x $Cost^B$, where K represents the characteristic of duration performance and B is the indicative constant of sensitivity of time performance to cost level. The possibility of having piecewise discontinuous activity time cost function has also been explained in recent past studies (Moussourakis and Haksever 2004; Yang 2005). Weibull functional form has been used for the analysis to describe the relationship between project cost and duration (Nassar et al. 2005), and contract type and project duration (Anastasopoulos 2007). Several other studies have not only sought out a relationship between cost and duration but have also proceeded using linear and integer programming techniques to investigate the trade-offs between project duration and cost (Chassiakos and Sakellaropoulous 2005). Optimization algorithm was used to develop a time-cost profile considering various mathematical forms (Yang 2007). Analyzing the data by linear regression mathematical models subsequently determined that linear forms could be used under certain conditions, like while accounting for the unique character of the empirical project data, and the restriction of Least- Square Estimation (LSE) techniques to incorporate certain project assumptions (Hosmer and Lemeshow 1999). When ordinary least square (OLS) techniques are used certain variables that are not represented by traditional explanatory variables could cause irreducible random noise (Hendrickson et al. 1987). Concept of earned value project management has also been applied by the researchers to predict the project duration (Vandevoorde and Vanhoucke 2006; Lipke et al. 2009).

This present study, structured on the findings of past research, seeks to estimate project duration by first describing the time duration data using more traditional functional forms and modeling techniques. The paper goes further to provide new insight into the potential risk factors affecting time overrun. All construction projects by their nature are economically risky undertakings. Risk is termed as an uncertain condition or event which if occurs, causes significant positive or negative effects (Project Management Institute, 2008b). Uncertain situations are characterized by the risk where actual outcome of an event or activity is deviated from the planned value (Raftery 1994) . Kwak and Stoddard (2004) termed identification of risks as the most crucial activity. Risk response measures need to be adapted to prevent the identified risk from materializing (Ropponen and Lyytinen

1997). Project duration and cost is dynamically affected by many variables at the execution stage (del Caño and de la Cruz 2002). Pakkala (2002) emphasized that better practices should be provided to ensure quicker project completion time and cost effective solutions to the owner, since he is most vulnerable to the design and construction risks.

Ibbs and Allen (1995) quantified the project changes impacts on engineering and construction project performance and concluded change as an event, which results in modification of original scope, execution time and cost of work. The problem remains the same that the future is not always predictable. Factor analysis technique was used to identify variable affecting construction time and cost overrun in Indonesia by grouping time and cost overrun variable into factors and then their relationship was determined, the study identified main causes of time delay as inadequate planning, design change and poor labor productivity (Kaming et al. 1997). According to Kaming et al. (1997) the results were specific to Indonesia but they reflected construction management problem in the developing countries. Chan and Kumaraswamy (1997) determined the significant factors causing time delays in Hong Kong and evaluated their relative importance. Their research stated poor supervision, poor site management, poor decision making, unexpected ground conditions and client initiated variations as the major causes of delay. Lo et al. (2006) found the distribution of construction delays in Hon Kong. Studies carried out in Ghana indicated time and cost overruns are related to poor contractor management, material procurement, material and cost escalation, poor technical performance and payment difficulties from agencies (Frimpong et al. 2003).

A number of studies have blazed the trail from the modeling techniques perspective, for examining the issue of cost overrun and time delays. The problem of time overrun in construction project was studied (Knight and Fayek 2002; Shaheen et al. 2007). Time and cost deviations were also investigated by Zheng and Ng (2005). This paper seeks to predict duration for various project types and identify the major risk factors for time overrun dominant in the highway projects of Pakistan. Development of reliable duration and delay estimates can help agencies to deliver optimum project schedules and thus avoid issues pertaining to time overruns that result in cost escalation.

DATA DETAILS AND ANALYSIS

To address the research objective, highway project data was collected from National Highway Authority Pakistan. This paper first describes project duration in terms of explanatory variable using traditional linear form. Separate linear models are also formed for different project types. Cognizance of past studies is taken into account to further investigate the project duration. Weibull analysis considered as a robust technique is used to yield survival curves and hazard functions. Survival analysis is used to model the time taken for an event to take place; it often involves the development of hazard function (Elandt-Johnson and Johnson, 1999). Time taken to project completion is sought to be modeled in this paper. Using historic data correlation between highway project delays and different types of projects is also calculated. Total of 120 projects, over financial years 2001-2012 were selected for estimation of highway project duration and for the analysis of potential risk factors. Projects costs were rebased to 2012 project prices. The data was related to four different project types: pavement construction, pavement rehabilitation, pavement improvement and bridge construction. Data was confirming to the regression assumptions. The residuals were random and independent. Outlying data (exceeding three standard deviations of the mean) of project duration and project delays was expunged using statistical software (SPSS). This section describes the detail on the data selection, its measurement and how we desire to elucidate it in framework of modeling results.

Planned Duration. This explanatory variable is measured as the length in calendar days allocated on the project at the preconstruction phase.

Actual Duration. It is the length in calendar days of the project estimated as a difference between the last day of work and stipulated start date of the project.

Delay Duration. It is taken as the difference between the planned duration and actual duration of the project.

Type of Projects. Four types of projects (pavement construction, pavement improvement, pavement rehabilitation and bridge construction) were selected for project duration estimation.

Project Cost. This explanatory variable depicts the size of the project and is measured in terms of total cost. Jahren and Ashe (1990) indicated size of the project a very significant predictor for time delay. Large projects are usually assumed to have greater duration. Involvement of huge number of contractors and subcontractors in large projects often leads to lapses in communication between them, thus makes them prone to longer delays.

Risk Factors. A number of time overrun factors were identified; a common time overrun variable was recorded for common time overrun factors across the projects. These variables, their frequencies and symbols are represented in Table 1. 16 risk variables were identified and were considered in multivariate delay analysis.

Geographical Location. Extent of geographic area is identified as an important factor for competitive bidding in building projects (Drew and Skitmore 1992). Construction cost and time duration are usually observed specific to the geographic areas. The data collected was from four different provinces of Pakistan: Sindh, Punjab, Balochistan and Khyber Pukhtunkhwa. Provincial data was split in the group of two depending upon the strong relationship between the risk factors and other attributes like project cost that could lead to greater time duration and time delays. Sindh and Punjab were placed in group 1 while Balochistan and Khyber Pukhtunkhwa were stationed in group 2. Two binary variables were created for each of these groups. If the information pertained to the location in the group the variable takes the value of 1 otherwise 0 is inserted.

Table 1. Project Time Overrun Risk Factors

Causes of Delays	Code	Total Indices
Late funds release	R1	95
Relocation of underground utilities	R2	3
Extreme weather condition (rainfall/ snow)	R3	4
Floods in the area	R4	1
Land acquisition	R5	19
Design change	R6	2
Scope change	R7	3
Adverse law and order situation	R8	17
Weak financial position of contractors	R9	2
Traffic management problem due to high traffic	R10	1
Contractors' lack of competence	R11	5
Non-availability of bitumen in the country	R12	2
Non-availability of skilled labor	R13	2
Scarcity of water	R14	1
Cash flow problems	R15	22
Dismantling (structures, trees)	R16	3

RESEARCH METHODOLOGY

Identification of project duration and analysis of risk factors that influence the planned time estimates was the paramount objective of this paper. Data from the provinces showing similar trends was grouped together. One of the hypothesis thus in this paper is everything remains same, difference in geographical groupings may result in different project duration. For example the projects in group 1 are subjected to similar situations different than the projects in group 2. The variables that were used in the prediction of duration models are given in Table 2.

Table 2. Explanation of Variables

Variable	Description
Project Duration	Highway project duration in days
X_1= Project indexed cost	Final Cost in millions of PKR, rebased to 2012 price
X_2= Geographical Location	X_2=1 indicates that variable pertains to Sindh or Punjab X_2=0 indicates that variable pertains to Balochistan or Khyber Pakhtunkhwa

Considering the convenience for use linear normal models were formed. The general form of the duration model is presented in Equation 1:

$$\text{Project Duration} = X_1 \times \text{Cost} + X_2 \times \text{Geographical Location} + 433.6 \quad (1)$$

The regression results produced an R value of 0.56 indicating reasonably good correlation. It can be seen that marginal increase in project duration with unit increase in project cost seems to be linear, and gets high if the projects are placed in geographical locations in group 1. Table 3 represents the model estimation results for the duration model for all project types.

Table 3. Model Parameter Estimates

Variable	Coefficients (t-statistics in parentheses)
Project Duration Model	
Intercept	433.60 (3.343)
Project indexed cost	0.37 (8.369)
Geographical Location	156.17 (1.30)
Number of observations	92
R^2	0.56
Adjusted R^2	0.55

Separate models were also developed for different project types to study the impact of explanatory variables over the project duration. Table 4 represents the descriptive statics of selected variable by project type. Table 5 presents the coefficients of variables in individual project models and their corresponding t-statistics.

Table 4. Descriptive Statics of the Data by Project Type

Statistics	Construction	Improvement	Rehabilitation	Bridge
(a) Project Duration				
Mean	1585.46	1285.53	1047.85	1094.14
Std. dev	838.98	558.75	299.16	539.95
Minimum	545	333	395	250
Maximum	4474	26996.31	1500	2500
Observation Count	45	21	20	34
(b) Project Indexed cost (in millions PKR)				
Mean	1667.65	1299.98	1159.80	1031.76
Std. dev	1357.27	986.46	310.85	640.15
Minimum	300	287.29	664.5	148
Maximum	5617.73	27299.77	1616.57	3221.68
Observation Count	45	21	20	34

Probabilistic Modeling. Probabilities that change over time are generally suited for hazard function analysis. Probability plots generated by survivor function in log logistic analysis provide the likelihood of the project duration being equal or greater than some specified duration.

Log-linear functional form was used in this paper for survival and hazard models. The parameters are $\lambda > 0$ and $P > 0$. Though Weibull exhibits a flexible means of calculating duration dependence but it has a limitation of keeping the hazard to be monotonic (Washington et al., 2003). While log logistic on the other hand caters non monotonic hazard function.

This study has therefore utilized log logistic distribution, describing that the hazard increases in duration till some extent and then it starts to decrease. Limdep statistical software package (Greene, 2007) is being used to produce log-linear logistic modeling process. Probability plots are provided by the survival function of log-logistic and Weibull duration model, $S(t)$ = Prob [$T \geq t$] = exp [$-(\lambda t)^{P}$] which indicates the duration of the project to be equal or greater than the specified duration. The survival plots in Figure 1 represents the probability of survival at the vertical axis and abscissa of the plot presents the duration in days. The percentile of the survival distribution is given by the Equation 2, provided α is the probability that the project will survive up to time‘t’ or greater:

$$t = [((1.0- \alpha) / \alpha)^{1/P}] / \lambda \qquad (2)$$

Table 5. Parameter Estimates of Individual Project Types (t-statistics in parentheses)

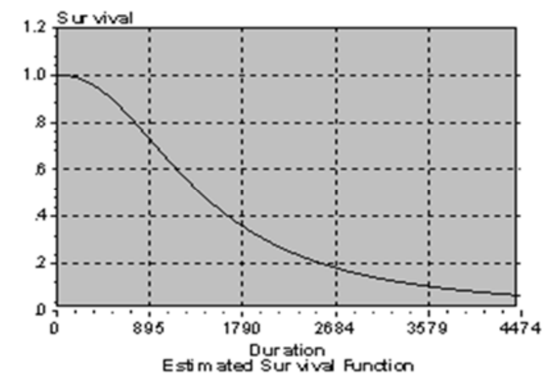
Variable	Construction	Bridge	Rehabilitation	Improvement
Project Duration Model				
Intercept	871.6 (4.23)	663.57 (2.42)	254.80 (1.01)	732.91 (2.79)
Project indexed cost	0.40 (5.20)	0.49 (3.94)	0.620 (3.50)	0.325 (2.64)
Geographical Location	97.09 (0.46)	-89.13 (-0.31)	86.42 (0.574)	246.58 (1.03)
Number of observations	45	34	20	21
R^2	0.4	0.34	0.43	0.28
Adjusted R^2	0.38	0.3	0.36	0.2

For different project types, project duration functions of the above form and respective survival distributions at percentiles 0.25, 0.50, 0.75 and 0.95 are presented in Table 6.

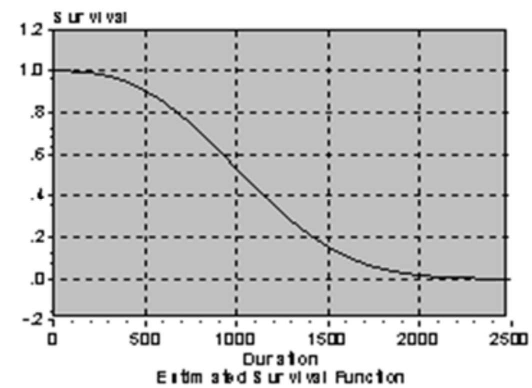
Table 6. Probabilistic Project Duration Models Survival Distribution

Project Type	Model Parameter (Log-Logistics)		Probability of Surviving to Time				Inflection Duration Point (days)
	λ	P	α =0.25	α =0.50	α =0.75	α =0.95	t^*
Pavement Construction	0.00072	2.25	2251	1382	849	374	1535
Pavement Improvement	0.00083	4.24	1554	1200	926	600	1590
Pavement Rehabilitation	0.00098	6.39	1209	1018	857	642	1328
Bridge Construction	0.00103	2.22	1585	966	589	257	1062
Delay Models	0.00167	1.87	1078	600	334	125	558

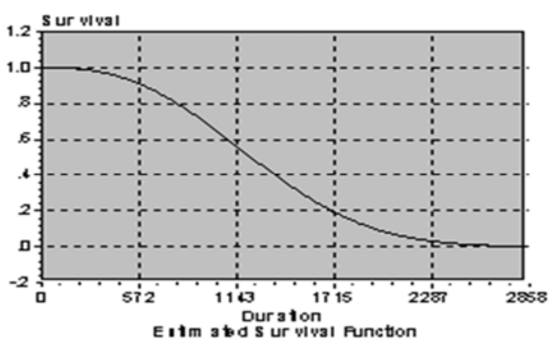
The rate at which the project durations are ending at time‘t’ are presented in hazard functions plot in Figure 2. Level of hazard is presented at the vertical axis and abscissa of the plot presents the duration of project in days. The parameter P is found to be greater than 1, and hence it indicates all the project types to be non monotonic. This exhibits the property of hazard function, $h(t) = [(\lambda P)\ (\lambda t)^{P-1}]\ /\ [1+(\lambda t)^{P}]$, increasing in duration from zero to an inflation point where duration, $t^{*} = (P\text{-}1)^{1/P} / \lambda$, is calculated in Table 6.



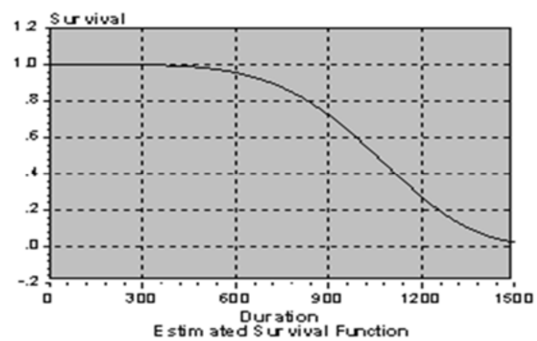
(a) Road Construction Projects



(b) Bridge Construction Projects



(c) Improvement Projects

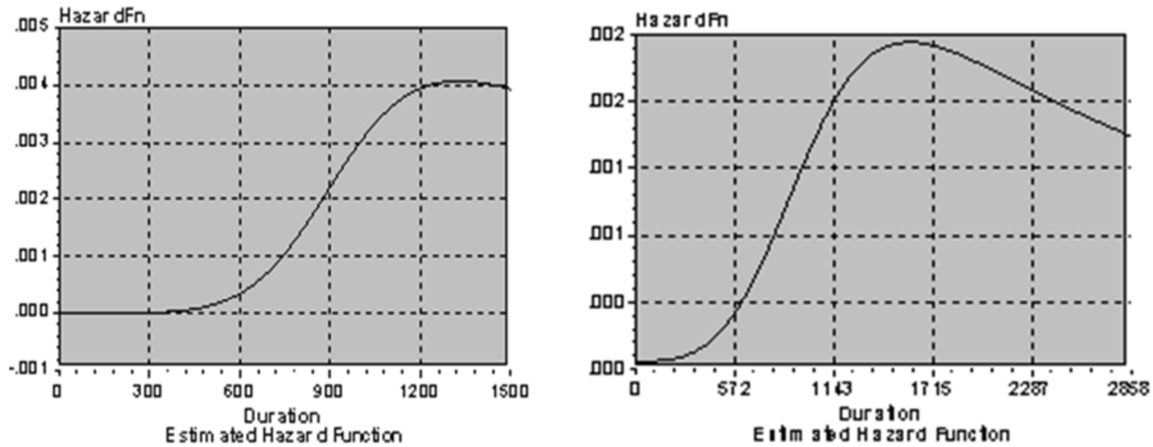


(d) Rehabilitation Projects

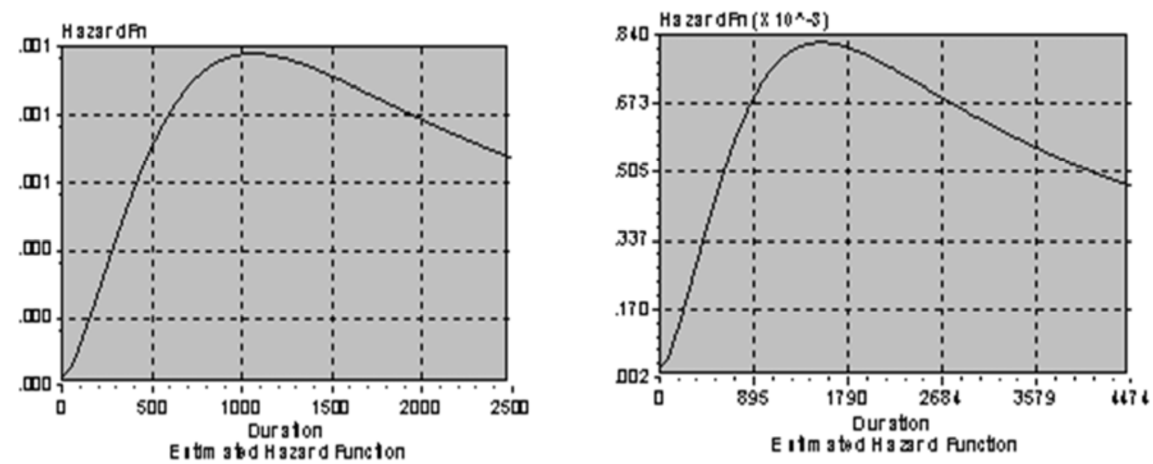
Figure 1. Survival plots of different projects

The last step was the identification of correlation between the highway projects and time overrun factors by analyzing the historic data. Multivariate regression analysis was used to investigate any correlation between the time delay and project attributes. Analytical model that correlate different project attributes to time overruns and owner project risks were created using delay duration as a dependent variable.

The highway project data was collected from National Highway Authority Pakistan's records. The available data comprised of description of the work type and reasons of delay for individual projects. As mentioned above all project costs were standardized to 2012 project prices to remove the effect of inflation.



(a) Rehabilitation Projects **(b) Improvement Projects**



(c) Bridge Projects **(d) Construction Projects**

Figure 2. Log- Logistic hazard model Plots of different projects

Statistical models explaining correlation between several relationships relating to time overrun were identified in this research. Multivariate regression analysis was used to manage the relationship between project cost, project risks and project delay.

The time overrun, adopted as a dependent variable was denoted as the difference between the programmed duration and actual duration. The correlation was identified among the following project variables:

1. Highway completed indexed cost;
2. Highway project type (construction, rehabilitation, improvement, bridge);
3. Highway geographic location (Sindh, Punjab, Balochistan, Khyber Pakhtunkhwa);
4. Risk factors (R1-R16)

Taking into account the convenience for future use linear normal models were formed. Forward, backward and stepwise multivariate regression was used in analysis for finding correlation between the variables. Null hypothesis was assumed that there is no correlation between the project cost, project type, project delays and the project risks. Identification of best models was allowed by the use of coefficients of multiple determinations (R^2 and adjusted R^2 statics).

Multivariate regression taking into account all the explanatory variables provided an R value < 0.02, indicating that the data has not fitted the model very well. Data showed R1, R5 and R15 having maximum frequency of occurrence across the country. Stepwise regression showed risk variable R1 exhibiting reasonably strong correlation then rest of the risk factors responsible for delay.

An interesting finding was the correlation between the delay and project variables like project indexed cost, risk factor R1 and geographical location. The model parameter results are represented in Table 7.

Variable representing delay was found to be statistically significant to reasonable extent with the explanatory variables, irrespective of the project type. These results thus go in the favor to reject the null hypothesis, and support the relationship between delays, project indexed cost and risk factors. One of the interesting finding was the impact of geographical location on the delay duration which shows an inverse correlation. Historical data showed risk factor R1 prevailing in most of the projects and influencing the planned duration which is also found to be inversely related with each other and statistically significant (at 80% level of confidence).

Table 7. Model Estimation Results

Variable	Coefficients (t-statistics in parentheses)
Project Delay Model	
Intercept	545.41 (3.36)
Project indexed cost	0.29 (6.86)
R1 (Risk)	-176.16 (-1.26)
Geographical Location	-85.40 (-0.96)
Number of observations	90
R^2	0.53
Adjusted R^2	0.54

System weighted R-square related to goodness of fit of the developed model indicates that statistically significant variables that were used to explain the delay showed 50% of variation in the data. It is therefore concluded that the attributes considered in the present study are not the comprehensive representatives of the delay factors and there is an evident possibility of various other common variables which can further explain the variation in time overrun.

To evaluate the accuracy of predicted models in forecasting duration and delays, mean absolute percent error (MAPE) was calculated as shown in Equation 3:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}|PE_i| \tag{3}$$

Where $PE_i = (X_i - F_i) / X_i$ is the error for the observations, the percentage error can be obtained by multiplying the above equation with 100. Here X_i is the actual duration and F_i is the predicted duration. The MAPE value closer to zero has better accuracy. The calculated MAPE values in this research suggest the extent of data over estimation or underestimation ranges from 20% to 40% which should be taken into account by the planning agencies while predicting the project durations and delays. For example MAPE value for rehabilitation duration model suggests that duration predicted is 21% over estimated or it is 21% underestimated.

DISCUSSION OF FINDINGS

Duration model represents the data pertaining to four different provinces of Pakistan exhibits some similar characteristics and is therefore grouped into two pairs. Sindh and Punjab were sited in group-1 while Balochistan and Khyber Pakhtunkhwa were placed in group-2. Each paired group was prone to the similar risk factors which affect their durations. A finding from the literature review suggested that extent of geographic location is an important factor in building projects (Drew and Skitmore 1992), in this research weak correlation between geographic location of the project and project duration was encountered. The difference could be due to the changed nature and complexities of highway projects as compared to the building projects.

Late funds release from to the funding agencies (R1), land acquisition problems (R5) and cash flow problems were the chief time overrun factors prevailing in most of project across the country. The remedial measure to overcome R1 was investigated that the funds releases on the projects should be balanced for the smooth completion of work. A number of projects are taken up simultaneously by the National Highway Authority with limited resources in hand which ultimately results into exceeding planned time estimates, it is therefore suggested that ongoing projects should accorded the priority and fresh projects should be initiated depending upon the available funds. Land acquisition, procurement and loan signing must also be synchronized to minimize the implementation delays.

The appropriateness of probabilistic model specifications is also investigated in this paper by using Weibull analysis which produced survival curves and hazard functions for the project models. Probabilistic models introduce stochastic element into the duration model and enhance the prediction of project duration. Thus prediction process is transformed to a robust stochastic description from an exact deterministic statement.

Accuracy estimation of the predicted models was carried out by using MAPE which identified the upper and lower extent of deviation of the true duration and delay values from the predicted models.

This is very important to keep in view that model suggested in this research are based on limited data so there is likely a possibility of many unobserved factors that may influence the duration and time overrun. Unforeseen events could not be considered while modeling the data. There are many other factors that are closely bonded to the project's performance and its timely completion, one of them being the competence of the management team. Construction management practices are in the state of continuous evolution and their consideration in the econometric models can produce more conclusive results.

The research is confined to the study of National Highway Authority projects in Pakistan. Data should be compared with caution with other situations.

CONCLUSION

The focus of the present study was to develop models for estimation of expected duration of highway projects on the basis of project cost, project type and geographical location. The project types included were pavement construction, pavement maintenance, pavement rehabilitation and bridge construction, based in four provinces of Pakistan. The data used in this study spans over the years 2001-2012. This paper also investigated the risk factors leading to time overrun. Consequences of various explanatory variables were studied on the duration and time overruns of the projects and statistically significant models were presented to equip the project planners beforehand for preparation of duration estimate at the planning stage. The developed models explaining time overrun as a function of variables available at the planning stage can be used as a tool for identifying projects with high time overruns and promptly remedial measures can be put in place to mitigate the risk.

Findings in this research can assist the highway agencies in forecasting project duration during planning stage and significantly improving the process of delay mitigation which can ultimately result into more competent highway project programs. Also, the projected models can help the contractors to prepare project duration estimates for making appropriate plans for labor, equipment and resource utilization which are strongly influenced by project duration.

REFERENCES

Ahmed, S., Azhar, S., Kappagantula, P., and Gollapudi, D. (2003). "Delays in construction: A brief study of the Florida construction industry." Proc., 39th Annual Conf. of the Associated Schools of Construction, Clemson Univ., Clemson, SC.

Anastasopoulos, P.Ch., (2007). "Performance-based contracting for roadway maintenance operations in Indiana". M.S. Thesis, Purdue University, W. Lafayette, Ind.

Arditi, D., Akan, G.T., Gurdamer, S., 1985. "Reasons for delays in public projects in Turkey. J. Constr. Manage." Econ. 3 (2), 171–181.

Chan, D. W. M., and Kumaraswamy, M. M. (1997). "A comparative study of causes of time overruns in Hong Kong construction projects." Int. J. Proj. Manage., 15(1), 55–63.

Chan, A.P.C., (2001). "Time–cost relationship of public sector projects in Malaysia." Int. J. Proj. Manage. 19, 223–229.

Chassiakos, A.P., Sakellaropoulos, S.P., (2005). “Time–cost optimization of construction projects with generalized activity constraints.” J. Constr.Manage. 131 (10), 1115–1124.

Czarnigowska. A, sobotka. A, (2013). “Time-cost relationship for predicting construction duration”, Original research article, Archives of Civil and Mechanical Engineering.

Deckro, R.F., Hebert, J.E., Verdini, W.A., Grimsrud, P.H., Venkateshwar,S., (1995). “Nonlinear time/cost trade-off models in project management.” Comput. Ind. Eng. 28 (2), 219–229.

del Caño, A., and Pilar de la Cruz, M. (2002). “Integrated methodology for project risk management.” J. Constr. Eng. Manage., 128(6), 473–484.

Drew, D. S., and Skitmore, M. (1992). “Competitiveness in bidding: A consultant’s perspective.” Constr. Manage. Econom., 10 (3), 227–247.

Elandt-Johnson, R., Johnson, N. (1999). “Survival Models and Data Analysis.” Wiley, New York.

Falk, J., Horowitz, J.L., (1972). “Critical path problems with concave cost time curve.” Manage. Sci. 19, 446–455.

Foldes, S., Soumis, F., (1993). “Pert and crashing revisited: mathematical generalizations.” Eur. J. Oper. Res. 64, 286–294.

Frimpong, Y., Oluwoyeb, J., and Crawford, L. (2003). “Causes of delay and cost overruns in construction of groundwater projects in a developing countries; Ghana as a case study.” Int. J. Proj. Manage., 21(5), 321–326.

Fulkerson, D.R., (1961). “A network flow computation for project cost curves.” Manage. Sci. 7, 167–178.

Greene, W., (2007). Limdep Version 9.0. Econometric Software, Inc., Plainview, NY.

Hendrickson, C., Martinelli, D., Rehak, D., (1987). “Hierarchical rule-based activity duration estimation.” J. Constr. Manage. 113 (2), 288–301.

Hosmer, D.W., Lemeshow, S., (1999). “Applied Survival Analysis: Regression Modeling of Time to Event Data.” Wiley, New York.

Ibbs, C. W., and Allen, W. E. (1995). “Quantitative impacts of project changes”, Univ. of California, Berkeley, Calif

Irfan, M., Khurshid M.B, Anastasopoulos. P ,Labi. S, Moavenzadeh. F, (2011). “Planning-stage estimation of highway project duration on the basis of anticipated project cost, project type, and contract type.” International Journal of Project Management 29(1),78–92.

Jahren, C., and Ashe, A. (1990). “Predictors of cost-overrun rates.” J. Constr. Eng. Manage., 116(3), 548–551.

Kaliba, C., Muya, M., Mumba, K., (2009). “Cost escalation and schedule delays in road construction projects in Zambia.” Int. J. Proj. Manage.27 (5), 522–531.

Kaming, P., Olomolaiye, P., Holt, G., and Harris, F. (1997). “Factors influencing construction time and cost overruns on high-rise projects in Indonesia.” Constr. Manage. Econom., 15(1), 83–94.

Knight, K., and Fayek, A. M. (2002). “Use of fuzzy logic for predicting design cost overruns on building projects.” J. Constr. Eng. Manage., 128(6), 503–512.

Kraiem, Z.M., (1987). “Concurrent delays in construction projects.” J.Constr. Manage. 113 (4), 591–601.

Kwak, Y.H., Stoddard, J., (2004). “Project risk management: lessons learned from software development environment. Technovation 24 (11), 915–920.

Lipke, W., Zwikael, O., Henderson, K., Anbari, F., (2009). “Prediction of project outcome: the application of statistical methods to earned value management and earned schedule performance indexes.” Int. J. Proj. Manage. 27, 400–407.

Lo, T. Y., Fung, I. W. H., and Tung, K. C. F. 2006. Construction delays in Hong Kong civil engineering projects.” J. Constr. Eng. Manage., 132(6), 636–649

Majid, M., McCaffer, R., (1998). “Factors of non-excusable delays that influence contractor’s performance.” J. Manage. Eng., 42–49.

Moder, J.J., Phillips, C.R., Davis, E.W., (1995). “Project Management with CPM, PERT, and Precedence Diagramming”, third ed. Blitz, Middleton, WI.

Moussourakis, J., Haksever, C., (2004). “Flexible model for time–cost tradeoff problem.” J. Constr. Manage. 130 (3), 307–314.

Nassar, K.M., Gunnarsson, Hordur G., Hegab, M.Y., (2005). “Using Weibull analysis for evaluation of cost and schedule performance.” J. Constr. Manage. 131 (12), 1257–1262.

Pakkala, P. (2002). “Innovative project delivery methods for infrastructure—An international perspective”, Finnish Road Enterprise,Helsinki, Finland.

Project Management Institute, (2008b). “The Standard for Portfolio Management - Second Edition”, 2nd ed. Project Management Institute, Newtown Square, PA.

Raftery, J. (1994). “Risk analysis in project management”, E & FN Spoon, London.

Ropponen, J., Lyytinen, K., (1997). “Can software risk management improve system development: an exploratory study.” European Journal of Information Systems 6 (1), 41–50.

Shaheen, A. A., Fayek, A. R., and AbouRizk, S. M. (2007). “Fuzzy numbers in cost range estimating.” J. Constr. Eng. Manage., 133(4), 325–334.

Skutella, M., (1998). “Approximation algorithms for the discrete time–cost trade-off problem.” Math. Oper. Res. 23 (4), 909–929.

Vandevoorde, S., Vanhoucke, M., (2006). “A comparison of different project duration forecasting methods using earned value metrics.” Int. J. Proj. Manage. 24, 289–302.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., (2003). “Statistical and Econometric Methods for Transportation Data Analysis”. Chapman & Hall/CRC.

Yang, I.T., (2005). “Chance-constrained time–cost tradeoff analysis considering funding variability. “J. Constr. Manage. 131 (9), 1002–1012.

Yang, I.T., (2007a). “Using elitist particle swarm optimization to facilitate bicriterion time–cost trade-off analysis.” J. Constr. Manage. 133 (7), 499–505.

Zheng, D.X.M., Ng, S.T., Kumaraswamy, M.M., (2004). “Applying a genetic algorithm-based multiobjective approach for time–cost optimization.” J. Constr. Manage. 130 (2), 168–176.

Zheng, D. X. M., and Ng, T. S. (2005). “Stochastic time-cost optimization model incorporating fuzzy sets theory and non replaceable front.” J. Constr. Eng. Manage., 131(2), 176–186.