PREDICTION OF SCOPE CREEP FACTORS IN SOFTWARE PROJECTS USING LOGISTIC REGRESSION ANALYSIS

BY

SHARON J. LIMO

SUPERVISOR:

DR. SIMON MWENDIA

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN INFORMATION SYSTEMS MANAGEMENT IN THE SCHOOL OF TECHNOLOGY AT KCA UNIVERSITY

NOVEMBER, 2023

DECLARATION

I declare that this research dissertation is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that it contains no material written or published by other people except where due reference is made, and author duly acknowledged.

Student Name: Sharon J. Limo Reg. No:22/01006

Sign:  Date: 7 Nov 2023

This research dissertation has been submitted for examination with my approval as a University Supervisor

Sign: _____ Date:

Name:

ABSTRACT

Scope creep is a persistent challenge inherent in software project management that significantly affects the success of software projects. This study aimed to establish a logistic regression modelling strategy for assessing, the impact of scope creep factors on successful software project management. The objective was to identify and quantify the key factors that contribute to scope creep and determine their influence on the likelihood of a project's success or failure. To achieve this, a comprehensive dataset was collected which encompassed different software development projects across different countries, domains and industries. The dataset included information on project scope, project timelines, development team size, budget allocation, user involvement, and other relevant factors. Additionally, data was collected on scope creep events, such as changes in project requirements, feature additions, and uncontrolled expansion of the software project scope. Logistic regression analysis was employed to assess the relationships between scope creep factors and project success. Several factors such as human factors, measurement factors and method factors were found to be statistically significant at the 0.001 level. Model validation using F-statistic, R-squared and residual plots established goodness of fit of the developed models. By identifying statistically significant factors and their impact on project outcomes, the outcomes of this study can help project managers and stakeholders make more informed decisions regarding scope management and risk mitigation strategies.

This research therefore makes a significant contribution to the field of software project management by providing a data-driven approach to understanding and managing scope creep. The study findings also inform best practices for scope management that will help organizations improve their project success rates in an increasingly dynamic and evolving software development landscape.

ACKNOWLEDGEMENT

The success of this work has left me indebted to several people. First of all, I want to express my gratitude to the Almighty God, who has bestowed upon me innumerable advantages, wisdom, and opportunities to complete this task. Secondly, I would like to thank my supervisor for his positive criticism and suggestion for necessary corrections to perfect the document. I would like also to express my gratitude to the KCA University administration and faculty for providing us with the required resources and opportunities to undertake this project.

Finally, I acknowledge the effort of my family for their support during this course. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

SME	Small and Medium Enterprises
COVID-19	Corona Virus Disease of 2019
IS	Information System
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis
EFA	Exploratory factor analysis

DEFINITION ON TERMS

Small and Medium Enterprises: small businesses with annual sales that are lower than KSh. 1,000,000

Scope creep: A situation that arises when the project deviates from the defined project scope, or there is pressure to deliver more than was initially agreed between the customer and the vendor.

Scope creep Management: the management or prevention of scope creeps within the project timeline

Machine learning: is a branch of artificial intelligence that places focus on the use of data and algorithms to imitate the way that humans learn to gradually improve its accuracy.

Regression: a statistical technique that relates a dependent variable to one or more independent (explanatory) variables

Logistic Regression: a regression technique that estimates the probability of an event occurring, such as succeeded or did not succeed, based on a given dataset of independent variables.

Exploratory factor analysis: statistical method for reducing data to a smaller set of summary variables and for exploring the underlying theoretical structure of observed phenomena.

CHAPTER ONE

INTRODUCTION

1.1 Background

Effective management of software projects is an important concern for software companies. Research on software project management has focused on documenting best practices that software industries can follow to successfully complete software projects (Hamid et al, 2019). Nevertheless, even though software project management literature is widely available, about 70 percent of software projects per year fail to be completed successfully. Failure of software projects has negatively affected the industry by reducing job opportunities and the general revenue, reducing the motivation of software managers and the employees involved in software projects, and increasing fatigue and stress in the development team.

According to recent figures, tech startups in Kenya raised a total of 574 million dollars in 2022. This was 14.4% of the total raised funds in Africa, and it was a 97% increase in income from the previous year (Kamer, 2023). The number of tech SMEs and startups has increased significantly during and after the COVID-19 period because of the need to support digital needs for companies. However, according to project management statistics the ratio of challenging projects has increased to 43% (Komal et al., 2020). The reasons include over-budgeting, lateness or having fewer features or functions in the developed software. Several studies have been conducted to try and explain the failures experienced in software projects, and most of the studies list scope creep as one of the most pronounced causes (Cobb, 2023).

1.1.1 Failure of Software Projects

Over the recent years, studies have extensively assessed and classified project failure (Hamid et al., 2019) A software project is termed as a failed project if it does not deliver the required outputs within the allocated time, quality or budget. A number of reasons for the failure of software projects have been identified, including improper estimation (Arora et al.,

2020), incomplete requirements (Montgomery et al., 2022), insufficient human resource (Salamzadeh et al., 2023), limited involvement of users (Lalmas, O'Brien & Yom-Tov, 2022) and scope creeping (Riaz & Gilani, 2022).

Evolutions in the software industry have led towards the emergence of large amounts of variable software. With software gaining more prominent use in the digital age, software development companies are always striving to achieve complete satisfaction from customers in their software projects. Therefore, project success is the ultimate goal of every software development project. Furthermore, it has been established that project quality and customer satisfaction are two of the most critical factors used to evaluate the success of a software development project (Eftekhari et al., 2022). Because effective project management is becoming an important aspect of project success, project managers are playing an important role in ensuring project success. This means that they must take into account the different techniques and strategies for achieving this. According to the project management triangle that has in the past been popularly used by software companies to measure project success, scope constitutes one of the cornerstones for defining the quality of a software project (Toor & Ogunlana, 2010).

1.1.2 Scope Creep in Software Projects

Scope creep has been defined as the incorporation of more requirements than were originally specified which results in greater project costs, and possibly a variation in the previously estimated project time. Software companies and their customers usually establish a long-term relationship, mutual acceptance, and an understanding that the benefits to be gained by each party are at least partly dependent on the other party (Chakrabarthy et al., 2008). Because of this understanding, most requests for changes by customers are often viewed as acceptable and are not recognized to be part of scope creep (Mathuri et al., 2018). Indeed, in some situations when change requests are properly controlled and monitored, they do not

negatively impact the project success. However, in most cases scope creep risk is not accurately anticipated by project managers. The Project Management Institute (PMI) states that scope management is an important step in project management (Eftekhari et al., 2022). Scope management usually includes a number of deliverables, a budget and the expected completion time for the project (Komal et al., 2020). Because of this, the ICT industry sees uncontrolled creep as a critical risk to the success of a project. These reasons are what motivate the current research to assess the influence of scope creep factors on the success of a project.

Scope creep is the inability of a project to maintain the initial agreed scope (Kurkovsky, 2022). Scope creep arises when the project deviates from the defined project scope, or there is pressure to deliver more than was initially agreed between the customer and the vendor Sahadevan (2023). Scope creep plays an important role for determining the future of a project in terms of success or failure. For problem identification, three things are usually considered: scope, budget, and time (Eftekhari et al., 2022). How to manage scope creep management is therefore an important characteristic of project manager. Scope creep management involves managing or preventing scope creeps within the project timeline. Success or failure of software projects can be greatly determined by the ability of project managers to control scope creeps. A survey of sixty project managers found out that 92% of software projects failed because of poor scope creep management (Komal et al., 2020). This implication of how project fails or success as a result of the identified factors calls for thorough examination of factors which are associated with scope creep, and which will ultimately cause a project to deteriorate in terms of success. The assessment of these factors using real-world data is a gap in research studies that have been done so far.

Logistic regression is attractive for research such as the current study because of its strength in estimating qualitative outcomes such as project success. Standard linear regression models such as ordinary least squares (OLS) cannot guarantee that predicted probabilities will

be within the 0 and 1 interval (Das, 2021). This is because non-linearity of variables is needed in order to guarantee sensibility of values for predicted variables. With the logistic regression function, all conditional probabilities are non-linearly associated with the independent variables. The function has the default characteristic of approaching 0 and 1 asymptotically, hence the predicted probabilities can sensibly be interpreted. Additionally, logistic models can estimate complex models using a large number of independent variables against the dependent variable (Bisong & Bisong, 2019). However, despite the advantages of the logistic regression modeling approach, it has been rarely used within the domain of predicting success or failure of software projects.

1.2 Problem Statement

Software companies, such as start-ups have a high failure rate. Statistics published by Galal (2023) show that start-ups failure rate in Africa stands at 50% of the entire world. Kenya had a failure rate of 24%, as shown in Figure 1.1

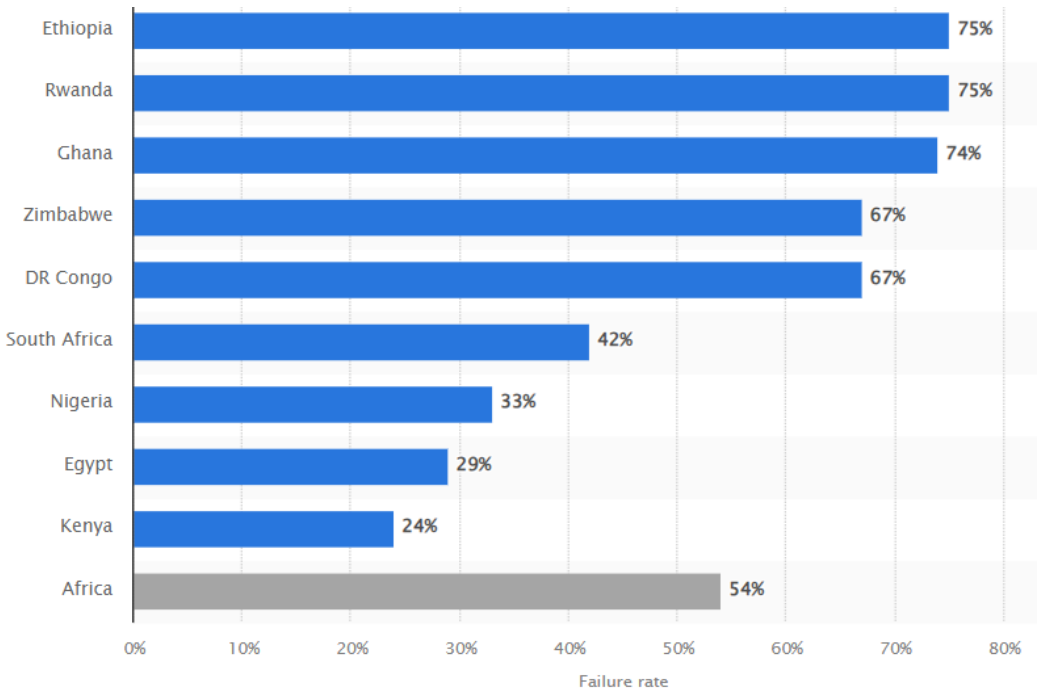


FIGURE 1.1: Start-up failure rate in selected African countries (Source: Galal, 2023)

The statistics of Figure 1.1 show that it is necessary to effectively predict scope creep factors that directly influence the failure of software projects and their ability to generate money for survival of startups.

The expectations of users are always growing as they become more informed. Therefore, user preferences are constantly changing as driven by advancements in technology, diversity of market trends and availability of many options for software to use. This means that users advance new requirements and scope in the software project lifecycle. Project success is one of the most common research topics because of its implications in evaluating projects in terms of costs, quality and time schedule, however, although scope creep is emerging as one of the most common reasons why software projects fail, there is very limited research that has been done in this domain. It seems that scope management is the most neglected topic in software project management. For example, in a survey by Schoonwinkel (2016), every project manager pointed out that scope creep was one of the three most likely risks in software projects. However, only 6% of project managers listed scope creep prevention as a method for preventing risk.

Existing project management techniques have not effectively measured or predicted the scope creep. Because of insufficient analysis of software scope creep factors and their impacts on successful projects, organizations, project managers and tech startups are unaware of the critical scope creep factors to watch out for when conducting software projects, which can result in serious budget and timeline overruns. Furthermore, the limited studies which conducted analysis of scope creep factors have limited their analysis to descriptive perspectives such as historical averages and graphical tables of survey responses. For example, Madhuri et al. (2018) conducted a triangulation of scope creep influence by purposively sampling project managers and triangulating the means of the survey responses. Similarly, Komal et al. (2020) compiled results of systematic literature review and qualitative analysis of survey responses to

identify the influential scope creep factors. Although machine learning approaches can reveal more patterns because they deeply assess the underlying trends, they have been rarely used for understanding the scope creep risk.

1.3 Research Objectives

1.3.1 Main Objective

The general objective of this study was to predict scope creep factors in software projects using logistic regression analysis.

1.3.2 Specific Objectives

i. To assess and identify the attributes that influence scope creep in software projects.

ii. To develop logistic regression models with the identified factors to predict scope creep of software projects.

iii. To test and validate the developed models.

1.3.3 Research Questions

i. Which attributes influence scope creep in software projects?

ii. How can predictive machine learning models be developed through logistic regression using the identified attributes?

iii. What is the validity of the developed models for application in predicting scope creep of software projects?

1.4 Motivation of the study

Although according to research such as Eftekhari et al. (2022), it is clear that project scope is one of the three most important determinants of software project success (the other two being budget and time), scope is rarely modeled as an important influence in software project management. The proper management of project scope has not been identified as a key

characteristic of project managers, and project schedules do not consider the scope as a critical component. Consequently, this impacts on a large number of failures for seemingly promising software projects. Researchers such as Komal et al. (2020) recently discovered that poor prediction of project scope creep resulted in many failed projects. According to Kissflow project platform guide (2022), the two most common reasons why projects fail are change in organization's priorities (at 39%) and change in project objectives (at 37%). The statistics are provided in Figure 1.2. Previous research studies have discovered that when these two factors occur, a project will experience scope creep, meaning that it will deviate from the initial scope.

FIGURE 1.2: Project ratings of features affecting project failure

Scope creep can lead to a lot of negative consequences from reworking, and even project failure. This can affect both the project manager and the entire project team. Therefore, the study has identified proper strategies for dealing with scope creep to avoid missing the delivery dates for software projects, exceeding the project budget and affecting.

Different machine learning techniques have been applied within the area of software project management and prediction, but logistic regression analysis has not been widely used.

The motivation for adopting this approach is the unique advantages that logistic regression offers, such as the ability to estimate complex models that have many different variables, and the fact that unlike approaches such as ordinary least squares (OLS) regression which assume the data to be normally distributed, logistic regression analysis makes no such assumptions. In general, analysis using logistic regression equation does not make many of the primary assumptions of linear regression or generalized linear regression models, which have a basis on OLS algorithms, such as linearity, normality, homoscedasticity (uniform variance) or measurement level, as continuous (Das, 2021). This makes the logistic regression analysis approach a useful prediction tool for project managers.

1.5 Significance of the study

Scope creep can lead to a lot of negative consequences from reworking, and even project failure. This can affect both the project manager and the entire project team. Therefore, the study has identified proper strategies for dealing with scope creep to avoid missing the delivery dates for software projects, exceeding the project budget and affecting the quality for delivered products. Such predictions of project creep factors will therefore greatly help project managers.

The study will benefit software developers and software development companies such as start-ups since they can use the model to properly manage their software development projects. They will be able to complete all defined tasks without bad multi-tasking where they are forced to deliver more tasks, but the quality suffers. Through the developed model, the software developers will be able to prioritize tasks and features to properly manage project requirements.

Users on the other hand will benefit by obtaining quality software on time, which can increase their satisfaction with the software and the software projects. Predicting scope creep factors that impact badly on project success will also improve communication between the users and the project team. It will promote trust-based relationships and make it easier to discuss and address any rising issue. Finally, it is expected that this study will benefit future

researchers as they will be able to use the developed model as a case study to build better models in future.

1.6 Scope of the study

The study was limited to factors that affect scope creep and hinder scope creep management, thereby affecting the overall success of software projects. It did not specifically consider multi-projects or mega software projects, but it instead looked at small, medium and large sized projects. The factors to be identified were also limited to agile, traditional and hybrid project models. In terms of area scope, the study was not a countrywide project limited to Kenya. Because of COVID-19, the infrastructure for working remotely has been advanced since 2020, and many projects can now be conducted without limitations of country borders. Therefore, this study utilizes dataset comprising survey responses from software project managers located in different countries across the world.

1.7 Chapter Summary

The chapter has outlined the background of the study and the problem statement. It has provided the project objectives and significance of the study. The next chapter is an extensive literature review. The chapter will examine the literature on project scope creep and its factors, as well as machine learning techniques that have been applied in past studies to predict scope creep in software projects.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter will review the available literature in the domain of prediction of scope creep in software projects using machine learning technologies to understand how they relate to the current study. The chapter structure will be on the theoretical review of the literature, a review of different machine learning prediction models as well as an empirical assessment of the factors influencing scope creep in software projects. The chapter seeks to identify knowledge and technological gaps that the study will address. The general objective is to provide a comprehensive understanding of the theoretical underpinnings while highlighting the diverse models employed in prediction in the field of software project management coupled with the underlying dataset considerations.

2.2 Theoretical Review

Many past studies have tried to seek possible explanations for success or failure of software projects. Some of these theories are described as follows:

2.2.1 Agency Theory

This theory has suggested that project monitoring can reduce privately held information, which in turn increases the possibilities for project success (Mahaney & Ledere, 2011). According to this theory, the agents or employees have knowledge (advantage) over the principal or supervisor since the agents are aware of the performance goals and what they are able to achieve. The agency theory is relevant for understanding the interaction between agents and principal. An agent works on behalf of the principal. The agent is also responsible for acting in the interest of the principal without considering his own selfish interests. In this context, the Agency theory purports that deviations in the project will happen when the

principal (e.g., client) and the agent (e.g., the project manager) have differing interests when executing the project. Although this theory offers some explanation regarding the fate of software projects, it is one sided because it ignores dimensions like political, and the role of other stakeholders, and it only focuses on the economic dimension.

2.2.2 Information Systems (IS) Success Theory

This theory has six interrelated components to explain the success of information systems: Systems quality, information quality, user's satisfaction, individual's impact and organization's impact (DeLone& McLean, 2003). The IS success theory has been widely adopted to test success or failure of different IT technologies such as web quality, perceived value and satisfaction (Chang, 2013). Software project management can consider how scope creep factors can impact on the components used to measure IS success in order to effectively manage software projects.

2.2.3 Satisfaction-Loyalty Theory

According to this theory, satisfaction is the overall attitude of users towards the software or service, and it represents an emotional reaction to the difference between the user's expectation and perception, while considering the user's goals and desires (Yachin, 2018). Satisfaction can be quantified as an individual's evaluation and how he/she emotionally responds to the entire experience of interacting with the system or software, both during the development and after implementation. This experience usually remains in memory alongside the user's expectations and perceptions, leading to a level of satisfaction or dissatisfaction (de Freitas et al., 2018). This research considers satisfaction as the cumulative feeling generated throughout a user's experience with software applications.

User loyalty creates a long-lasting competitive advantage for a software or service. Therefore, software producers compete for user loyalty especially since a lot of software is

freely available and users have many choices. This can explain why project managers allow scope to creep in trying to fulfill unreasonable user demands. Soltani-Nejad et al. (2020) have defined loyalty as a feeling of attachment or interest by the user. If the users feel too attached to the project, they can end up dominating the requirements and interfering with the project scope.

In seeking to address these gaps the current research relied on these theories as a basis for explaining the success or failure of software projects and to understand scope creep and its impact factors.

2.3 Machine Learning Techniques for Predicting Scope Creep and Project Success

Different machine learning models are applicable for measuring project risk factors such as scope creep and the influence on project success. The most popular machine learning approaches are discussed below.

2.3.1 Support Vector Machine (SVM)

The SVM is widely used in classification problems for image, text and hypertext. SVM is a supervised machine learning algorithm that is meant for classification and regression problems. Its aim is to form the finest and most suitable decision boundary called a hyperplane to separate the n-dimension space to different classes that ease the placement of different points in their correct categories (Pisner & Schnyer, 2020). SVM has been mostly used for face detection and image classification problems.

2.3.2 Exploratory Factor Analysis (EFA)

EFA is a highly complex multivariate analytical technique that allows the measured variables or factors to be linked with other multiple latent factors with the aim of identifying a

data structure. EFA is important for describing shared variation among factors, and for assessing the potentially unknown points (Luo et al., 2019). However, EFA involves many linear and sequential steps which can make the technique computationally intensive.

2.3.3 Regression

Regression analysis is an approach for fitting straight lines to data patterns. For a regression mode, the main variable, called the dependent variable, is predicted using k other variables known as the independent variables via a linear equation (Madhuri et al., 2018). Since Y represents the dependent variable and X_1 to Xk represent the independent variables, we assume that the value of Y at time t (i.e., on row t) in the sampled data is determined by the multiple-regression equation. However, one of the basic assumptions of the standard regression analysis is that the data follows a normal distribution, which is not always the case.

2.3.4 Logistic Regression

Logistic regression is based on the odds of a two-level outcome of interest, 0 and 1. The odd of an outcome is the ratio of the probability of the outcome happening, divided by the probability of the outcome not happening (Bisong & Bisong, 2019). Even odds are usually associated with 1 and correspond to the probability of an outcome happening half of the time.

When building the logistic regression model, one uses the natural log of the odds as a regression function of the predictors. With a single predictor, X, this takes the form:

$$\ln[odds(Y = 1)] = \beta_0 + \beta_1 X,$$

where *ln* represents the natural logarithm and *Y* is the outcome. *Y*=1 when the event happens, and *Y*=0 when it does not. β_0 is the intercept term, and β_1 represents the regression coefficient, the amount of change in the logarithm of the odds of the event in the event of a 1-unit change in the predictor variable X. The difference in the logarithms of 2 values is equal

to the logarithm of the ratio of the 2 values, hence taking the exponential of β_1 allows achieving of the odds ratio corresponding to a 1-unit change in X (Das, 2021).

One of the major strengths of logistic regression as compared to other related techniques such as probit regression is the convenience of interpreting the exponentiated slope of the logistic regression coefficient (e^b) as the odds ratio (Schober & Vetter, 2021). This ratio is an indicator of how much the odds of a certain outcome under observation can change for a one-unit increase in the independent variable when dealing with continuous) or versus a reference category, when the independent variable has categorical values. This ease of interpreting the outcomes makes logistic regression a preferred approach.

2.4 Application of Machine Learning in Project Success Analysis

Selecting the most influencing factors to improve project performance is a difficult task for project managers. Practical data evaluation methods using machine learning algorithms can be used to predict such influential factors. Machine learning is used to discover patterns in large datasets which are used for different objectives, for example to assess project success, or to cost projects based on a set of independent variables (Sheikhalishahi et al., 2022).

Several studies have used different machine learning tools and techniques to understand different aspects of software development projects. Linares-Vásquez et al. (2014) used support vector machines (SVM) algorithms provided in the WEKA tool to categorize different software applications into domains categories. The output was to assist stakeholders in realizing software requirements and help to predict maintenance problems in software development projects. Rathore and Kumar (2021) used the Dynamic Selection Learning (DSL) technique to predict the performance of different machine learning techniques to model fault prediction. Mehta and Patnaik (2021) combined regression analysis, Partial Least Square analysis (PLS), and Recursive Feature Elimination analysis (RFE) to classify the modules within software as

whether they were prone to defects or not. These authors also found out that stacking Ensemble and XGBoost techniques revealed the best results for data sets where the defect prediction was more than 90% accurate. Pace (2019) has applied correlation analysis and linear regression of 367 project managers data in the USA to find out the effect of moderating variables for project management. The results were not aligned with previous findings and the authors recommended further studies to measure success of different project management approaches.

Clustering techniques have also been used in evaluating project success. Lalic et al. (2022) applied exploratory factor analysis (EFA) combined with K-means clustering on a sample size of 227 project managers to test whether traditional projects, agile projects and hybrid projects had any differences regarding the likelihood of project success. The authors found that agile projects had much more chances of success. They concluded that their findings using K-means cluster estimations could not be termed as conclusive.

Simulation techniques for machine learning have also been used to effectiveness of different factors associated with software projects. For example, Li et al. (2021) applied system dynamics (SD) modelling and simulation methods to monitor and predict unexpected changes and errors in the software requirements. The analysis outcomes were used to inform a knowledge management model for controlling the process of developing software. Komal et al. (2020) used partial least squares structural equation modelling (PLS-SEM) to analyse primary data collected through a survey. The authors wanted to find out the factors that undermine software projects and inflate project costs in Pakistan. The study found that PLS-SEM lacked sufficient accuracy estimate the influential factors and can be used by managers to control project success rates.

Most of the research that has been conducted using machine learning has focused on nondeterministic approaches for predicting project success instead of empirical approaches that use data to measure the effects of different variables. Some of those that used data relied on

primary data collects and sometimes had inconclusive results. To date there is still no specific framework that clearly outlined which algorithms can be used to determine which specific scope creep factors affect project success.

2.5 Factors Influencing Scope Creep in Software Projects

Although the assessment of scope creep factors is minimal in research, general research is being conducted in all areas which can potentially influence project success. Most of the focus is in the areas of project management process, the role and importance of project managers, digital maturity of software companies, different process models and other project-influencing parameters (Madhuri et al., 2018). In a risk identification study of agile software projects, Elkhatib et al. (2022) discovered that risks related to the organization, technology, process, monitoring and analysis were the most influential underlying risks for agile project management. Ahmadi et al. (2022) used 21 semi-structured interviews to conduct a survey of senior project management practitioners involved in complex projects. They discovered that the experience and competencies of project managers was the most essential factor for successful projects. They also found out that leadership was a critical competency factor for a project manager.

Lalic et al. (2022) explored whether the approach of software project management (as traditional, agile or hybrid) had any impact on project management success. They used five success parameters to compare the three project management approaches: Proficiency in the project, project impact on the team, impact on the customer, business success factors and effort to prepare for the future. They found out that correct selection of a project management approach does not automatically to project success, but other aspect like implementation effort and efficiency must be considered. Pace (2019) conducted a non-experimental correlation study to test the influence of project management methodologies adopted (traditional or agile).

The project success was the dependent variable, project management method was the independent variable, and industry and experience were the moderating variables. They did not find any strong relationship between the success of a project and the type of methodology being applied.

An extensive literature review has been done by Iriarte and Bayona (2020) to establish success factors of IT projects. The authors found out that the most cited critical factors are involvement of users, support, communication and commitment. These factors have also been discovered to be associated with scope creep in software projects (Komal, 2020). These findings show that soft skills of a project's team members are highly relevant for project success.

2.6 Conceptual Framework

This study adopts Project Success as the dependent variable. The independent variables have been sourced from Section 2.5 where the factors influencing scope creep have been identified from previous studies. Based on the factors identified in the literature review, the study has established the following conceptual framework.

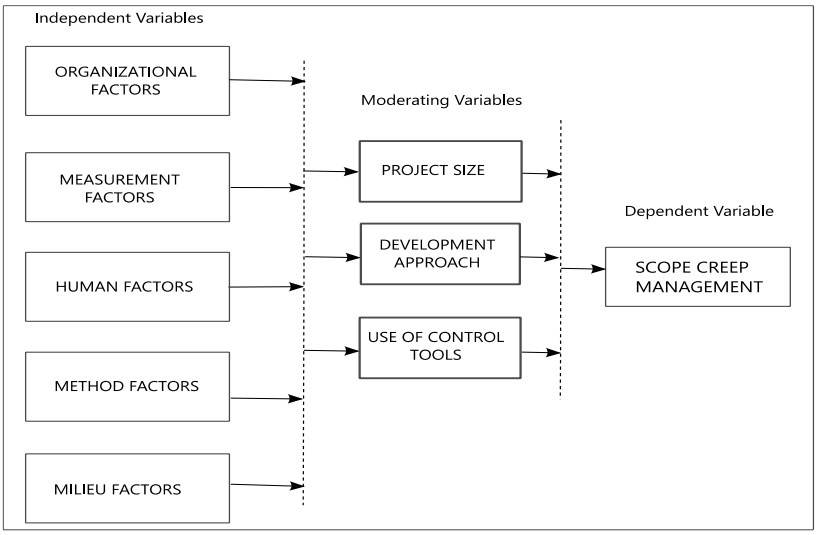


FIGURE 2.1: Conceptual Framework of Scope Creep Factors

2.7 Operationalization of Variables

The operationalization information can be found in Table 2.1 below.

TABLE 2.1: **Operationalization of Study variables**

Variable	Indicators (Methods)	Values
ORGANIZATION	Includes 5 sub variables (Company size, Project size, Team capability, Standards and policies, Project personnel). The sum of 5 features measured using 5-point Likert Scale	Numeric (0 – 25)
MEASUREMENT	Includes 5 sub variables (Methodology, Tools, Approach, Quality, and Scope Management). The sum of 5 features measured using 5-point Likert Scale	Numeric (0 – 25)
HUMAN FACTORS	Includes 4 sub variables (Communication, Client knowledge, Expectations, Developer experience). The sum of 4 features measured using 5-point Likert Scale	Numeric (0 – 20)
METHOD FACTORS	Includes 4 sub variables (Quality control, Scope management, Requirements change, Goals clarity).	Numeric (0 – 20)

	The sum of 4 features measured using 5-point Likert Scale	
MILIEU FACTORS	Includes 4 sub variables (User requirements change, Budget control, Uncertainty Management, Market Needs Change). The sum of 4 features measured using 5-point Likert Scale	Numeric (0 – 20)
SCOPE CREEP MANAGEMENT	One feature measured using 5-point Likert Scale	Numeric (0 – 5)

2.8 Chapter Summary

As per the literature review presented in this chapter, this study has recognized the factors linked with scope creep and went along to propose a conceptual framework that has formed the benchmark for evaluating impact of the identified factors influencing success or failure of a project.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the strategy that has been used to achieve the study objectives. The main objective of the study was to establish a machine learning model that was used to assess the influence of scope creep factors on software project success. This chapter therefore outlines which steps were used to meet the specific study objectives. It addresses the study design, the data that has been used to conduct this study, and the data analysis method. At the end of this chapter, the research considerations for the study analysis and presentation of the results have been outlined.

3.2 Research Design

The research has adopted a secondary research design. Secondary research is also known as desk research and involves the use of existing data found in different sources, such as internal sources but more commonly from external sources (Sileyew, 2019). This means that the study has applied existing data and literature to build the study model. The process was iterative and involved steps such as data collection and cleaning, model building and model refinement after testing, in order to achieve results that are stable, easy to interpret and meaningful. All this was conducted under a set of ethical guidelines and data privacy regulations. This research design is convenient because it saves time and allows the researcher to collect a great amount of data without the need for sampling. The research design followed the structure shown in Figure 3.1.

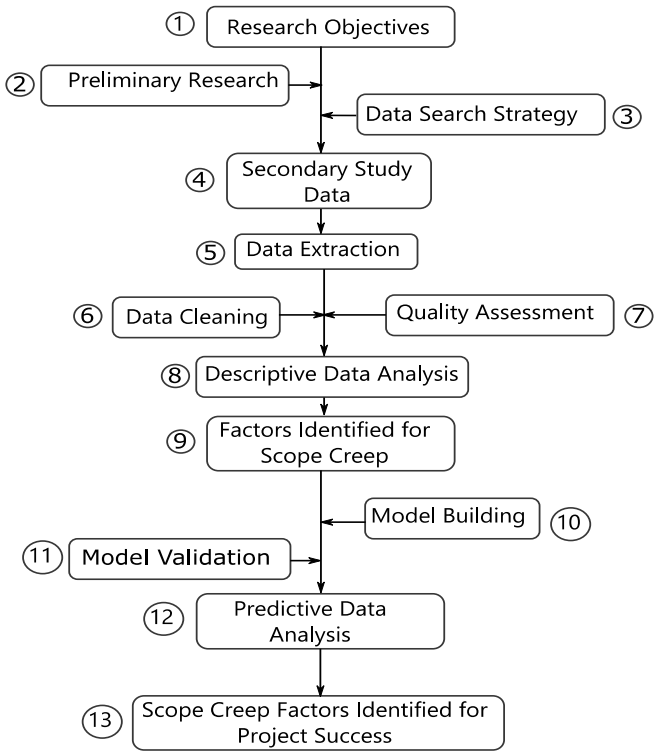


FIGURE 3.1: Flowchart of secondary research design activities

In the description of Figure 3.1 above, the study began by establishing a set of research objectives to guide the research process. These objectives were used to identify a correct study data set through conducting preliminary research that helped to identify a set of study hypotheses, as well as identifying the optimal strategy for data search. The identified data was extracted through a download process with the necessary permissions. It was cleaned and assessed for quality, where it was found suitable for achieving the study objectives. This is a very important step since it allows the description of a problem within a certain contextual framework that allows for interpretability of the study findings. After confirming the conceptual framework using descriptive analysis of secondary data and identifying the key variables that are relevant for this study, a logistic regression approach was used to develop the five models for assessing the effects of human, measurement, method, milieu and organizational factors of the scope creep in software projects. The main advantage of logistic

regression method is that it can test multiple dependent variables and several independent variables at the same time. It can also be used on small or large data sets with no assumption of normal distribution for data. The model validation and predictive data analysis steps have been extensively discussed in this chapter, together with the findings, and their interpretation and discussion.

3.2.1Target Population and Data

The target population for this study were the project managers that are involved in small, medium, and large projects in 50 countries, including the ones working remotely. The study considered in this population the managers who manage agile projects, traditional projects, and hybrid projects. This is because the literature review found out that the project methodology applied is an important determinant of project success (Lalic et al., 2022).

The data that was used in this study was obtained from Mendeley datasets (https://data.mendeley.com). The Mendeley data is a dataset that has been previously collected through survey research of 306 project managers who are involved in managing traditional, agile and hybrid software projects. Each of these managers had 3 years or more of experience at the time of the research survey. The responses were measured and recorded using a 5-point Likert scale. The data contains 42 variables. Nine of these variables represent demographic information about the participants and their organizations. Data about the organization includes the country and business approach of the organization, the approach used to manage projects, size, the main business approach, and the tools used to control scope change. Data about the participants includes experience in managing software projects. The remaining variables (33) measure qualitative responses about factors that can potentially influence scope creep, and scope creep factors that can influence project success or failure. These include factors like

communication, project complexity, customer expectations, time constraints, project size, etc. **FIGURE 3.2** shows a header of the sample data set.

3.2.2 Sampling Technique

Since the research analysis employs a secondary data source, there was no need for sampling and the entire dataset was used for the analysis. However, this data was cleaned in order to remove incomplete and inconsistent data. As a result, therefore, that this process reduced the size of the collected data.

3.3 Data Collection

Data collection was done through desk research. The study collected secondary data that is available online from Mendeley data sets. The data consists of responses from 306 participants based on a set of 42 questions. Among the questions that the managers were asked are the specific questions about factors influencing scope creep, as well as descriptive questions such as Name of the organization, Country, Primary business, the respondent's designation, respondent's experience, size of the organization, type of project, and whether the organization uses any tool to manage scope changes. The responses from this survey study were used to build the study variables. In addition to descriptive analysis to discover the trends in the data set in terms of mean, median min, max and standard deviation, a correlation analysis was conducted for all the study variables to determine if there were inherent relationships between the independent variables used for the study.

The specific considerations that were made to this dataset were regarding the currency of the data since the survey study was conducted in 2021.This reflected a faithful representation of the opinions of current software project managers. Additionally, the project managers were

sourced from different parts of the world and reflected a good representation of the global scope creep problem. However, there were restrictions to the dataset, including the limited representation of Kenya and other countries in Africa, as well as the small sample size.

3.4 Data Analysis

The study 42 variables used in the study were categorized according to similarity and combined to construct the 5 variables to be used in the study (Organization, Measurement, Human, Method and Milieu). Descriptive analysis, model development and model validation were conducted using the R statistical analysis software.

3.4.1 Methodology for Achieving Study Objective 1

The first objective of this study was to assess and identify the attributes that influence scope creep in software projects. The study performed descriptive analysis using mean, median, min, max and standard deviation. These metrics were used to discover the trends in the data and understand how it was distributed. The Pearson product-moment *r* correlation was then used to assess the linear relationships between each two variables. Pearson correlation is indicated with "r" and its values can vary between -1 and +1. The r value of -1 indicates a perfect negative correlation. 0 indicates the complete absence of a linear relation, while +1 shows a perfectly positive correlation.

3.4.2 Methodology for Achieving Study Objective 2

The second objective of this study was to develop logistic regression models with the identified factors to predict scope creep of software projects. After descriptive analysis, a set of five logistic regression models were built from variables in the study that were found to be significant using the R statistical analysis software. The regression models were: 1) Model 1 measuring the effect of nine HUMAN factors (X5 – X13) on scope creep management, 2) Model 2 measuring the effect of five MEASUREMENT factors (X14 – X18) on scope creep

management, 3) Model 3 measuring the effect of six ORGANIZATIONAL factors (X19 – X24) on scope creep management, Model 4 measuring the effect of seven MILIEU factors (X25 – X31) on scope creep management, and 5) Model 5 measuring the effect of four METHOD factors (X32 – X35) on scope creep management. The model equations are presented as follows:

$$logit(pX) = log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_i X_i + .. + \beta_n X_n$$

Where the outcome represents the log odds of project success, β_0 is the model intercept, β_i ..β_n represent the model coefficients for each of the five developed models, and X_i – X_n are model variables for the Human, Organization, Measurement, Milieu and Method groups of factors.

The data was divided into 70-30 partitions where 70 percent of the data was used to train the model and 30 percent was used for testing and validating the model. Xu and Goodacre (2018) have found out that when the training set contains too few samples, this can have a negative effect on the outcomes. They therefore recommend the 70 / 30 percent splitting that has been adopted in this study. The most common method for splitting the data set is to randomly identify a proportion of samples and to retain them as the validation set, and thereafter to use the rest of the sample as a training dataset. The process was done repetitively, and the final estimates of model performance represented the mean of performance on the validation data sets across all the iterations. The most recommended method used to repartition the data is the bootstrap method by Efron (2000).

3.4.3 Methodology for Achieving Study Objective 3

The third objective was to test and validate the developed models. Metrics such log likelihood, chi-square, corrected Akaike information criterion (AICc) and Bayesian information criterion (BIC) were applied in this study to test and validate the logistic regression

models. The negative log-likelihood (-Log Likelihood) test was used to compare the goodness of fit of whole model (i.e., the model that lists all the labels) to the reduced model (i.e., the model with omitted regression parameters and containing only the intercept parameters). This test is equivalent to the sums of squares test and is specified as -LogLikelihood = - (log(ŷ) * y + log (1 – ŷ) * (1 – y)), where y represents the model labels and $\hat{y}$ represents the predicted probabilities. Smaller values indicate improvements in the model's goodness of fit.

Another test that was adopted in this study is called the chi-square (χ^2) test. This test is similar to the F-test and is represented as

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed value and E_i is the expected value. A smaller chi-square value represents a better model. The tests also used the corrected Akaike's Information Criterion (AICc). This is defined as

$$AICc = -2LogLikelihood + 2k + 2k(k+1)/(n-k-1)$$

where k is the number of parameters that the model estimates, and n is the number of observations. A smaller value for the AICc indicates an improved model.

The BIC test is written as BIC = -2LogLikelihood + $k$ln(n), where k represents the number of estimated parameters and n is the number of model observations. A smaller value indicates a better model.

3.5 Ethics

Proper permissions were sought from the relevant authorities before collecting data and conducting analysis. All unique identifiers that could link the results to particular software

company's participants were removed. Additionally, no information about a single participant was disclosed, since results are to be reported as a whole. This means that it was impossible to derive information about a particular user from the collection of the data. Finally, the analysis and interpretation of the result adopted a realistic approach without falsifying the outcomes.

3.6 Chapter Summary

This chapter has extensively outlined the methodology for developing the study model that was used to assess the scope creep factors influencing project success using secondary data. The next chapter (Chapter 4) presents results from the achieved objectives and the interpretation of the key results.

CHAPTER FOUR

FINDINGS AND DISCUSSIONS

4.1 Introduction

This chapter offers a presentation of the results that have been achieved from the study objectives. The main objective of the study was to establish a machine learning model that was used to assess the influence of scope creep factors on software project success. This chapter therefore outlines the key outcomes from the developed model and offers an interpretation of these results. It addresses the descriptive analysis of the study data set, model outcomes and validation results. The purpose of this chapter is to provide a comprehensive understanding of how the research problem has been addressed to establish an adequate model for understanding the factors that directly influence scope creep and its management in software projects also, a detailed comprehensive analysis of the logistic regression model used is provided, including the methods used for data analysis, the results obtained, and the interpretation of the findings.

Additionally, this chapter presents a discussion of the key observed findings. The discussion encompasses the application of logistic regression for modeling Scope creep management in software projects in relation to factors that are related to the organization, human, measurement, milieu and methods

4.2 Results from Descriptive Analysis

Descriptive statistics have the primary role of providing key information about a data set. Therefore, descriptive analysis was adopted as the initial analysis in this study with the aim of assessing the trends in the study data set. This was done through identification of general trends in the data. This exercise helped to determine the suitability of logistic regression analysis for assessing the effect of different types of variables using the study data set. The measures that

were used to compare the study variables through the descriptive process are minimum value, median, mean (average), maximum value and standard deviation. These are the standard metrics that generally summarize the data and provide critical information about the distribution.

TABLE 4.1: **Descriptive statistics of study variables**

	Variable		Min	Median	Mean	Max	SD
Dependent	Scope creep management (Y)	105	1	4	3.858	5	1.261
Moderating	Organization Size (X1)	105	1	6	2.574	3	0.613
	Project size (X2)	105	1	3	2.865	5	1.168
Independent Variables	Development approach (X3)	105	1	3	2.877	3	0.384
	Control tools	105	1	2	1.594	2	0.493

		used/not used (X4)							
HUMAN		Unrealistic Expectations (X5)	05	1	.219	2	2	.265	5	1
		Lack of knowledge (X6)	05	1	.361	2	2	.289	5	1
		Poor Communication (X7)	05	1	.071	2	2	.217	5	1
		Stakeholder involvement (X85)	05	1	.452	2	2	.244	5	1
		Lack of timely feedback (X9)	05	1	.355	2	2	.288	5	1
		Manager Experience (X10)	05	.5	.33	0	2	.409	5	1

	Unclear goals (X11)	105	1	2	2	5	1
				.232		.268	
	Ego (X12)	105	1	3	3	5	1
				.013		.344	
	Inexperienced Staff (X13)	105	1	2	2	5	1
				.581		.227	
MEASUREMENT	Unauthorized Changes (X14)	105	1	2	2	5	1
				.548		.094	
	Underestimating Change (X15)	105	1	2	2	5	1
				.497		.261	
	Client's Knowledge (X16)	105	1	2	2	5	1
				.303		.281	
	Developer's Inexperience (X17)	105	1	3	2	5	1
				.652		.204	
	Over-accommodati	105	1	2	2	5	1
				.065		.121	

	ng Client (X18)							
ORGANIZATION	Budget constraint (X19)	205	1	2	.426	2	5	1.304
	Project Complexity (X20)	205	1	3	.794	2	5	1.293
	Organizational capabilities (X21)	205	1	3	.591	2	5	1.317
	Lack of resources (X22)	205	1	3	.697	2	5	1.250
	Fixed cost (X23)	205	1	3	.729	2	5	1.296
	Standards and policies (X24)	205	1	3	.903	2	5	1.127
MILIEU	Changing market (X25)	205	1	3	.774	2	5	1.312

		Personnel (X26)	05	1	2	.09	3	5	.047	1
		Requirement Volatility (X27)	05	1	2	.348	2	5	.177	1
		Time constraint (X28)	05	1	2	.548	2	5	.191	1
		Uncertainty (X29)	05	1	3	.587	2	5	.104	1
		Changing requirements (X30)	05	1	2	.174	2	5	.300	1
		Unforeseen Risk (X31)	05	1	2	.439	2	5	.140	1
	METHOD	Changing market needs (X32)	05	1	3	.684	2	5	.161	1
		Poor initial requirements (X33)	05	1	2	.219	2	5	.369	1

Quality issues (X34)	205	1	2	2.561	5	1.223
Informal Approval of Changes (X35)	205	1	2	2.252	5	1.287

Notes: N = sample size; Min = minimum; Max = maximum; SD = Standard deviation

The descriptive analysis included a total of 36 variables. Out of these, there was one independent variable, *scope creep management* and a set of moderating variables (i.e., Organization size, project size, development approach used (i.e., agile, traditional or Hybrid) and whether the company employed control tools to manage scope creep. The rest of the study variables (i.e., 32 variables) were independent variables. From Table 4.1 results above, it is evident that most of the study variables employed a 5-point Likert scale of responses, and therefore have a minimum value of 1 and maximum value of 5. Nevertheless, there was significant variation in the mean and standard deviation values. The highest mean of 8.33 was observed in the variable *Manager Experience* since this value was recorded in years. The dependent variable, *Scope creep management* also had a high mean of 3.858. The lowest mean was observed in the variable Control tools used/ not used, with a value of 1.59. This variable was binary in nature (with values of Yes or No).

In terms of standard deviation, the highest value was observed in the variable named *manager experience (X10)* followed by the dependent variable *scope creep management (Y)* Most of the study variable had a standard deviation above 1.1, except for the moderating

variables where only one of the variables had a value of 1.168 (Project size). The lowest value of standard deviation was seen in the variable *Development approach*. None of the observed study variables had a standard deviation value that was higher than its mean value. This suggested that there was no skewness in the data. There was thus confidence in applying the assumptions of logistic regression modelling to this data set.

4.3 Correlation Analysis of Study Variables

The analysis of how research variables are correlated is important since it helps to assess the dependency of variables to be included in a model. Correlation analysis is also called bivariate analysis. It is used to discover the existing association between study variables. This information can then be used to determine the nature of the relationship (whether positive or negative) as well as the strength or magnitude of the relationship.

The correlation analysis generated a heatmap that included a total of 36 variables. This heatmap is shown in Figure 4.1. Out of these, there was one independent variable, scope creep management and a set of moderating variables (i.e., Organization size, project size, development approach used (i.e., agile, traditional or Hybrid) and whether the company employed control tools to manage scope creep. The rest of the study variables (i.e., X5 to X35) were independent variables.

The correlation value is usually a ratio with values between 1- and +1, where values closer to -1 indicate negative correlation, values closer to +1 indicate positive correlation and values closer to zero indicate weak or no correlation (Gogtay & Thatte, 2017). This research assessed the correlation among a set of 36 different study variables, of which one was the dependent variable, four were moderating variables, and the remaining variables were independent variables. All variables with a positive correlation value greater than zero have an

increasing color hue that is approaching red while the variables with a negative correlation value higher than zero have an increasing color of blue.

Most of the independent variables that were under study had some form of negative relationship with the dependent variable *Scope creep management*. Apart from the diagonal of ones, most independent variables shared a very weak positive correlation between themselves. Most of the moderating variables did not appear to be correlated to the independent variables. The relationship between the independent variables themselves was almost nonexistent since no two independent variables had a correlation value higher than 0.5. The test of multicollinearity using variance inflation factors indicated the absence of the risk since the largest value was 1.18 which was far below the cut off mark of 10.

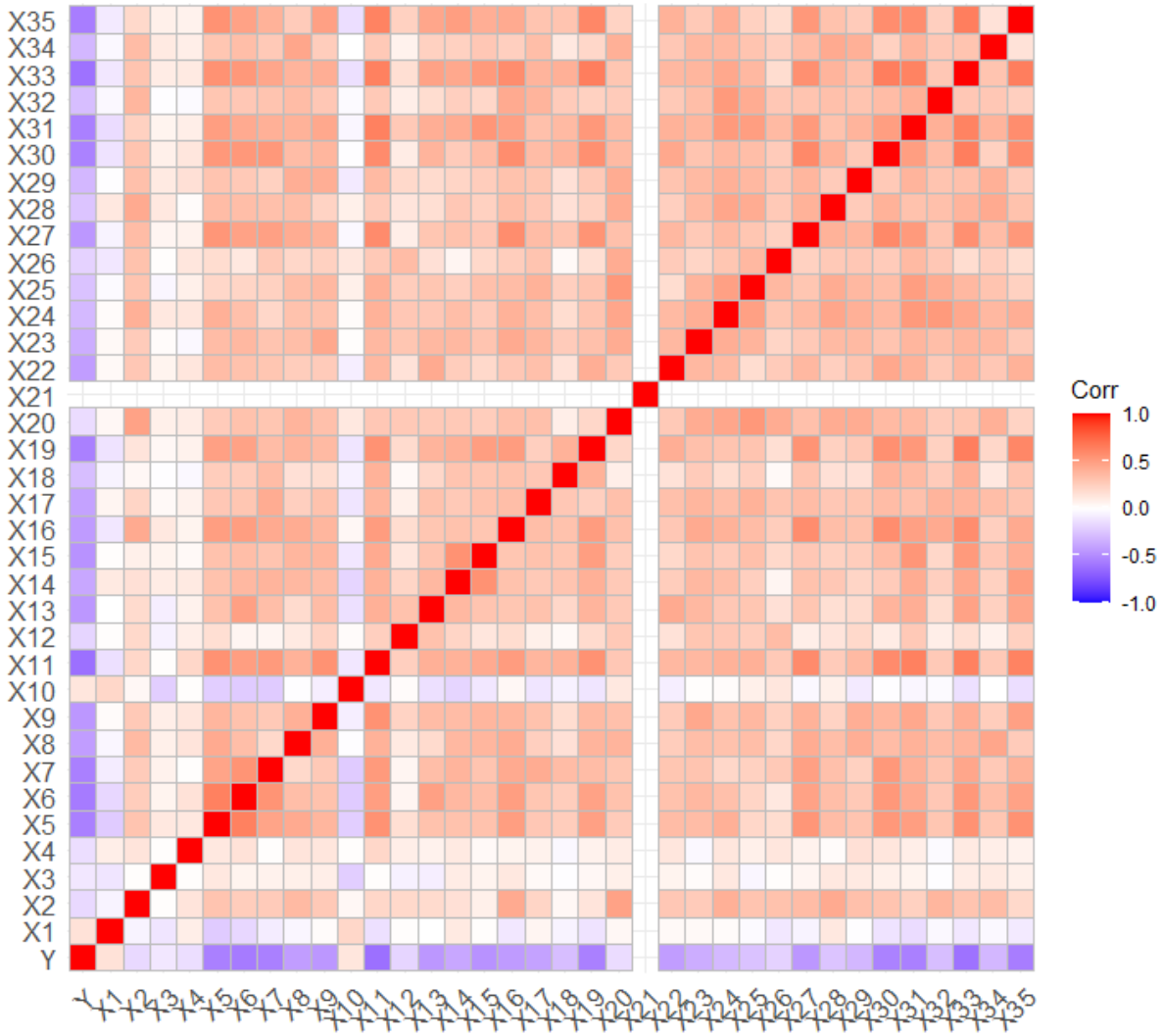


FIGURE 4.1: **Correlation matrix of study Variables.**

Dependent Variable: Y = Scope Management. **Moderating Variables**: X1= Organization Size, X2 = Project Size, X3 = Development approach, X4 = Control tools used/not used. **Independent Variables:** X5 = Unrealistic Expectations, X6 = Lack of knowledge, X7 = Poor Communication, X8 = Stakeholder involvement, X9 = Lack of timely feedback, X10 = Manager Experience, X11 = Unclear goals, X12 = Ego, X13 = Inexperienced Staff, X14 = Unauthorized Changes, X15 = Underestimating Change, X16 = Client's Knowledge, X17 = Developer's Inexperience, X18 = Over-accommodating Client, X19 = Budget constraint, X20 = Project Complexity, X21 = Organizational capabilities, X22 = Lack of resources, X23 = Fixed cost, X24 = Standards and policies, X25 = Changing market, X26 = Personnel, X27 = Requirement Volatility, X28 = Time constraint, X29 = Uncertainty, X30 = Changing requirements, X31 = Unforeseen Risk, X32 = Changing market needs, X33 = Poor initial requirements, X34 = Quality issues, X35 = Informal Approval of Changes

4.3 Logistic Regression Analysis of the Study Variables

The analysis of how the dependent variable *Scope creep management* is influenced by different factors was carried out in separate logistic regression models for each of the variable categories listed in the conceptual framework, i.e., HUMAN, MEASUREMENT, ORGANIZATION, MILIEU and METHOD. Each of these models had a different number of independent variables, where the model with HUMAN factors as the independent variables had the largest size, with 9 variables whose effects in scope creep management were being observed. The smallest model was observing the influence of METHOD factors on scope creep management, and it had 4 independent variables. The model of MEASUREMENT had 5 independent variables. The model of ORGANIZATION had six independent variables, while MILIEU had 7 variables being regressed against the dependent variable.

The different observations from the logistic regression models against the independent variables from the conceptual framework have been discussed in the next sub sections.

4.3.1 Logistic Regression Model of HUMAN Factors (Model 1)

The first model to be observed was the model of HUMAN factors, which had nine independent variables that were being assessed to determine their individual and collective influence on scope creep management. Table 4.2 presents results from this logistic regression analysis.

TABLE 4.2: Logistic regression outcomes of HUMAN influences on scope creep management

Variable	Coef.	Std, error	t - value	Signif.
(Intercept)	3.280	0.280	2.449	0.006**

X5 - Unrealistic Expectations	- 0.109	0.082	- 1.325	0.187
X6 - Lack of knowledge	- 0.140	0.084	- 1.658	0.099*
X7 - Poor Communication	- 0.227	0.075	- 3.023	0.003**
X8 - Stakeholder involvement	- 0.135	0.067	- 2.020	0.045*
X9 - Lack of timely feedback	- 0.081	0.049	- 1.864	0.046.
X10 - Manager Experience	- 0.003	0.014	- 0.201	0.842
X11 - Unclear goals	- 0.218	0.083	- 2.636	0.009**
X12 - Ego	- 0.012	0.058	- 0.198	0.843
X13 - Inexperienced Staff	- 0.122	0.073	- 2.883	0.015**
Residual standard error	0.8767 on 295 degrees of freedom			
Multiple R-squared	0.645			
Adjusted R-squared	0.617			
F-statistic	19.29 on 9 and 295 DF (p-value: < 2.2e-16)			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the results of logistic regression analysis shown in Table 4.2, all the independent variables had a negative relationship with the dependent variable, Scope creep management. This means that an increase in the value of each of the independent variables caused the value of the dependent variable to decrease by a certain amount. In other words, these variables are all detrimental to the management of scope creep in software projects.

Three of the independent variables were significant at the 0.001 level. These variables were X7 - Poor Communication, X11 - Unclear goals, and X13 - Inexperienced Staff. Two variables, i.e., X6 - Lack of knowledge and X8 - Stakeholder involvement were significant at the 0.01 level. One variable, namely X9 - Lack of timely feedback was significant at the 0.05 level. This means that this variable was at the borderline of statistical significance in its influence on the dependent variable. Although the remaining three (X5 - Unrealistic Expectations, X10 - Manager Experience, X12 – Ego) were negative to the dependent variable Scope creep management, their influence was not statistically significant.

The validation statistics of the logistic regression model measuring HUMAN influences showed the model performance to be good. The residual standard error was 0.8767 on 295 degrees of freedom. Residual degrees of freedom represent the total number of observations (records, N) of the dataset, subtracted by the number of variables under observation in the model. In this model, there were 305 observations and ten model variables (one dependent and nine independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value < 2.2e-16. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared (which accounts for the influence of multiple independent variables in the model). The values of 0.645 for multiple R-squared and 0.617 for adjusted R-squared both suggest that the independent variables explained up to 65% of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was high. Residual plots of the logistic regression model of HUMAN factors are presented in Figure 4.2.

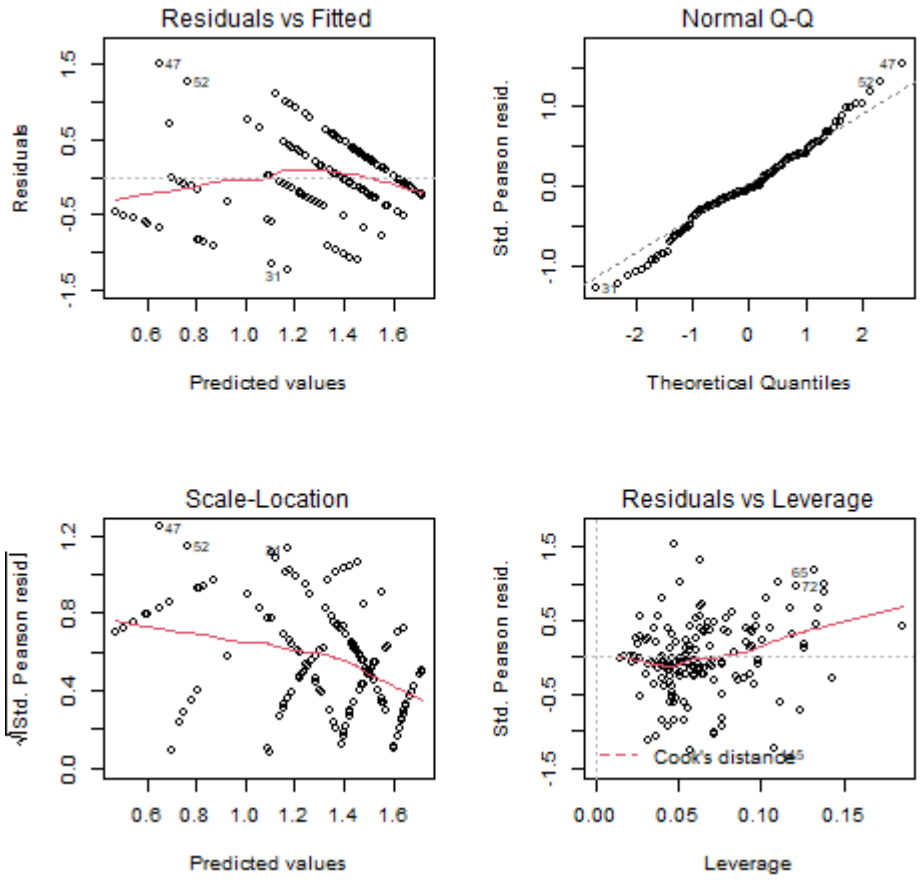


FIGURE 4.2: Residual plots of the logistic regression model of HUMAN factors

The most popular diagnosis tool for logistic regression models is the examination of residuals. This is because residuals pinpoint the difference between estimated values and observed values of the dependent variable. The plot of residual versus predicted values was

used to verify the assumption that residuals from the logistic regression model measuring the HUMAN factors were normally distributed and with a constant (uniform) variance. Based on the observation of Figure 4.2, the observation points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of HUMAN factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot displays graphically how any two quantiles of a distribution have lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal since most of the data points lie along the line y=x on the plot. Only a few points have deviated from this line, and they did not deviate too far away except for two outlier data points labelled as 52 and 47.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of HUMAN factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship n=between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that these observation points were influential observations. From the observation of

the Residuals vs Leverage plot in Figure 4.2, most of the data points were lying outside the red line and were therefore considered as influencing the variation that was observed in the dependent variable, Scope creep management.

4.3.2 Logistic Regression Model of MEASUREMENT Factors (Model 2)

The second model to be observed was the model of MEASUREMENT factors, which had five independent variables that were being assessed to determine their individual and collective influence on scope creep management. Table 4.3 presents results from this logistic regression analysis.

TABLE 4.3: Logistic regression outcomes of MEASUREMENT influences on scope creep management

Variable	**Coef.**	**Std, error**	**t - value**	**Signif.**
(Intercept)	5.969	0.264	0.595	0.051.
X14 - Unauthorized Changes	-0.094	0.094	-0.998	0.319
X15 - Underestimating Change	-0.279	0.082	-3.443	0.000***
X16 - Client's Knowledge	-0.239	0.072	-3.381	0.000***
X17 - Developer's Inexperience	-0.206	0.077	-2.689	0.008**
X18 - Over-accommodating Client	-0.037	0.082	-0.452	0.652

Residual standard error	1.029 on 299 degrees of freedom
Multiple R-squared	0.575
Adjusted R-squared	0.554
F-statistic	16.43 on 5 and 295 DF (p-value: 6.799e-13)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the results of logistic regression analysis shown in Table 4.3, all the independent variables had a negative relationship with the dependent variable, Scope creep management. This means that an increase in the value of each of the independent variables caused the value of the dependent variable to decrease by a certain amount. In other words, these variables are all detrimental to the management of scope creep in software projects.

Two of the independent variables were significant at the 0.000 level. These variables were X15 - Underestimating Change and X16 - Client's Knowledge. When change occurring in a software project was underestimated by one unit, it caused a decrease of 0.279 (i.e., 28 percent decrease) in the unit value that was being used to measure scope creep management. Additionally, a one-unit increase in the client's knowledge meant a 24 percent decrease (i.e., -0.239) in the ability to manage scope creep in software projects. The variable X17 - Developer's Inexperience was significant at the 0.01 level. This inexperience caused a 21 percent decrease (i.e., -0.206) in the ability to manage scope creep in the software projects. Although the remaining two variables (namely, X15 - Underestimating Change and X18 - Over-accommodating Client) were negative to the dependent variable Scope creep management, their influence was not statistically significant. This means that their negative impact did not statistically affect scope creep management.

The validation statistics of the logistic regression model measuring MEASUREMENT influences showed the model performance to be moderately good. The residual standard error was 1.029 on 299 degrees of freedom, since in this model there were 305 observations and six model variables (one dependent and five independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 6.799e-13. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.575 for multiple R-squared and 0.534 for adjusted R-squared. Both values suggest that the independent variables explained up to 58 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was moderately high. Residual plots of the logistic regression model of MEASUREMENT factors are presented in Figure 4.3.

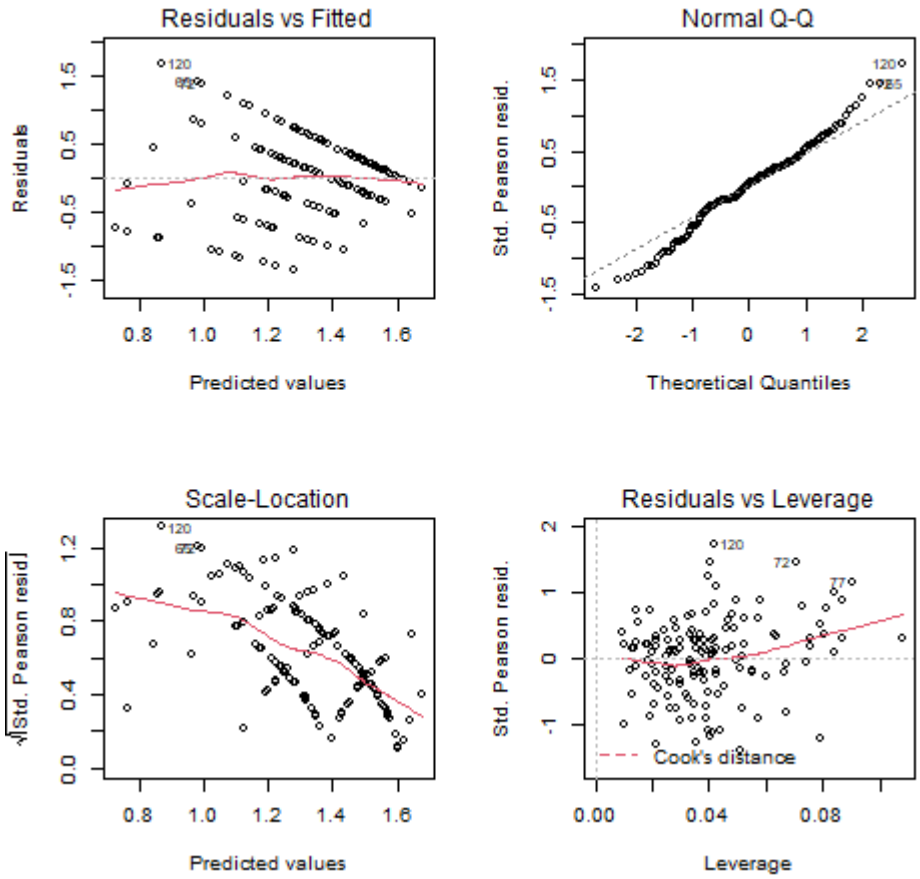


FIGURE 4.3: Residual plots of the logistic regression model of MEASUREMENT factors

The plot of residual versus predicted values shown in Figure 4.3 was used to verify the assumption that residuals from the logistic regression model assessing the influence of MEASUREMENT factors on software project scope creep management were normally distributed and with a uniform variance. Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of HUMAN factors used a quantile – quantile (Q-Q) plot to assess the degree of

normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal except for the values lying at the bottom area of the plot. Most of the remaining data points lie along the line $y = x$ on the plot. At the top of the plot only a few points have deviated from this line, and they did not deviate too far away except for three outlier data points labelled as 72, 77 and 120.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of MEASUREMENT factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot in Figure 4.3, most of the data points were lying outside the red line and were therefore considered as influencing the variation that was observed in the dependent variable, Scope creep management.

4.3.3 Logistic Regression Model of ORGANIZATION Factors (Model 3)

The third model to be observed was the model of ORGANIZATION factors, which had six independent variables that were being assessed to determine their individual and collective

influence on scope creep management. Table 4.4 presents results from this logistic regression analysis.

TABLE 4.4: **Logistic regression outcomes of ORGANIZATIONAL influences on scope creep management**

Variable	Coef.	Std, error	t - value	Signif.
(Intercept)	5.711	0.278	1.459	0.105
X19 - Budget constraint	-0.389	0.071	-5.488	0.001***
X20 - Project Complexity	0.114	0.076	1.514	0.502
X21 - Organizational capabilities	-0.179	0.082	-2.200	0.037*
X22 - Lack of resources	-0.207	0.075	-2.757	0.007**
X23 - Fixed cost	-0.124	0.075	-1.657	0.099*
X24 - Standards and policies	-0.039	0.088	-0.449	0.654
Residual standard error	0.997 on 298 degrees of freedom			
Multiple R-squared	0.602			
Adjusted R-squared	0.578			
F-statistic	16.46 on 6 and 298 DF (p-value: 1.852e-14)			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the results of logistic regression analysis shown in Table 4.4, all the independent variables had a negative relationship with the dependent variable, Scope creep management except for one variable, namely, X20 - Project Complexity. This means that except for this variable, an increase in the value of each of the independent variables caused the value of the dependent variable to decrease by a certain amount. In other words, these variables are detrimental to the management of scope creep in software projects.

One of the independent variables was significant at the 0.001 level. This variable was X19 - Budget constraint. As constraints in the project budget increased, the increase in turn caused a 39 percent decrease (i.e., -0.389) in the ability to manage project scope creep. One variable, X22 - Lack of resources, was significant to scope creep management at the level of 0.01. When a software project lacked one unit of resources, it caused a decrease of -0.207 in scope creep management. Although two model variables (namely, X20 - Project Complexity and X24 - Standards and policies) were negative to the dependent variable Scope creep management, their influence was not statistically significant. This means that their negative impact did not statistically affect scope creep management.

The validation statistics of the logistic regression model measuring ORGANIZATIONAL influences showed the model performance to be moderately good. The residual standard error was 0.997 on 298 degrees of freedom, since in this model there were 305 observations and seven model variables (one dependent and six independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 1.852e-14. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.602 for multiple R-squared and 0.578 for adjusted R-squared. Both values suggest that the independent variables explained up to 60 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was moderately high. Residual plots of the logistic regression model of ORGANIZATIONAL factors are presented in Figure 4.4.

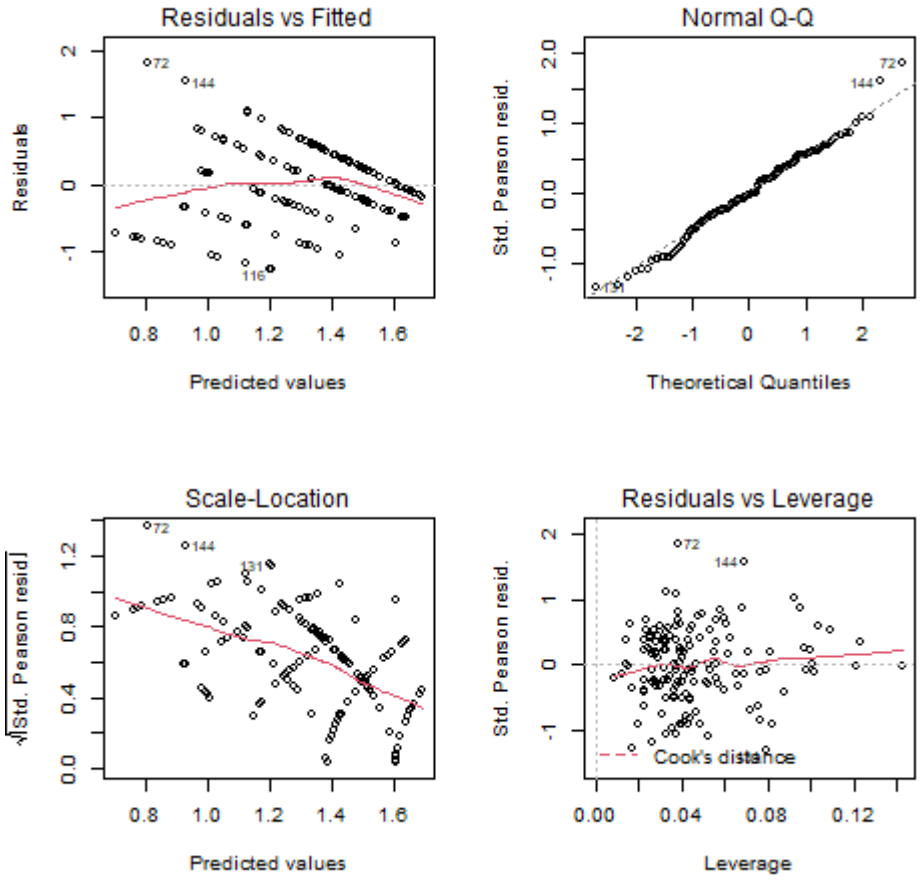


FIGURE 4.4: Residual plots of the logistic regression model of ORGANIZATIONAL factors

The plot of residual versus predicted values shown in Figure 4.4 was used to verify the assumption that residuals from the logistic regression model assessing the influence of ORGANIZATIONAL factors on software project scope creep management were normally

distributed and with a uniform variance. Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of ORGANIZATIONAL factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is almost perfectly normal, except for two data points that lie a little off the line, i.e., 72 and 144. All the remaining data points lie along the line $y = x$ on the plot.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of ORGANIZATIONAL factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot in Figure 4.4, almost all of the data points were lying outside the red line and were therefore considered as influencing the variation that was observed in the dependent variable, Scope creep management.

4.3.4 Logistic Regression Model of MILIEU Factors (Model 4)

The fourth model to be observed was the model of MILIEU factors, which had six independent variables that were being assessed to determine their individual and collective influence on scope creep management. Milieu factors are the external factors that are beyond control by the company, but which nevertheless affect the ability of software project managers to control scope creep. Table 4.5 presents results from this logistic regression analysis.

TABLE 4.5: Logistic regression outcomes of MILIEU influence on scope creep management

Variable	Coef.	Std, error	t - value	Signif.
(Intercept)	5.607	0.298	6.806	0.021**
X25 - Changing market	0.068	0.077	0.878	0.382
X26 - Personnel	0.052	0.086	0.610	0.543
X27 - Requirement Volatility	-0.064	0.093	-0.686	0.494*
X28 - Time constraint	-0.002	0.080	-0.030	0.007**
X29 - Uncertainty	-0.102	0.083	-1.224	0.223*
X30 - Changing requirements	-0.314	0.082	-3.826	0.001***

X31 - Unforeseen Risk	- 0.408	0.095 - 4.295	0.000***
Residual standard error	0.992 on 295 degrees of freedom		
Multiple R-squared	0.704		
Adjusted R-squared	0.688		
F-statistic	14.22 on 7 and 295 DF (p-value: 4.972e-14)		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the results of logistic regression analysis shown in Table 4.5, all the independent variables had a negative relationship with the dependent variable, Scope creep management except for two variables, namely, X25 - Changing market and X26 - Personnel. Although both of these variables had a positive influence of 0.068 and 0.052 respectively on the dependent variable Scope creep management, their influences were not statistically significant. For the five remaining independent variables, an increase in the value of each of the independent variables caused the value of the dependent variable to decrease by a certain amount. In other words, these variables are detrimental to the management of scope creep in software projects.

Two of the independent variables were significant at the 0.001 level. These variables were X30 - Changing requirements and X31 - Unforeseen Risk. As requirements in the software project changed by one unit, this change caused a downward change (-0.314) in scope creep management. Similarly, an upward change in the unforeseen risk by one unit of measurement caused a reduction by 41 percent (i.e., -0.408) in scope creep management ability in software projects.

One variable, i.e., X28 - Time constraint was significant to scope creep management at the level of 0.01. When time constraints increased by a single unit, it caused a decrease of 2 percent (i.e., -0.002) in scope creep management for software projects. Two model variables were statistically significant to scope creep management at the level of 0.05. These were X27 - Requirement Volatility and X29 – Uncertainty. When requirements became more volatile, it increased the risk of poor scope creep management by -0.064 (i.e., by 6 percent). Additionally, uncertainty in a project was detrimental to scope creep management. A rise in this variable value by one caused a decrease in the value of the dependent variable by 0.102, i.e., 10 percent.

The validation statistics of the logistic regression model measuring MILIEU influences showed that the model performance was good. The residual standard error was 0.992 on 297 degrees of freedom, since in this model there were 305 observations and eight model variables (one dependent and seven independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 4.972e-14. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.704 for multiple R-squared and 0.688 for adjusted R-squared. Both values suggest that the independent variables explained up to 69 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was high. Residual plots of the logistic regression model of MILIEU factors are presented in Figure 4.5.

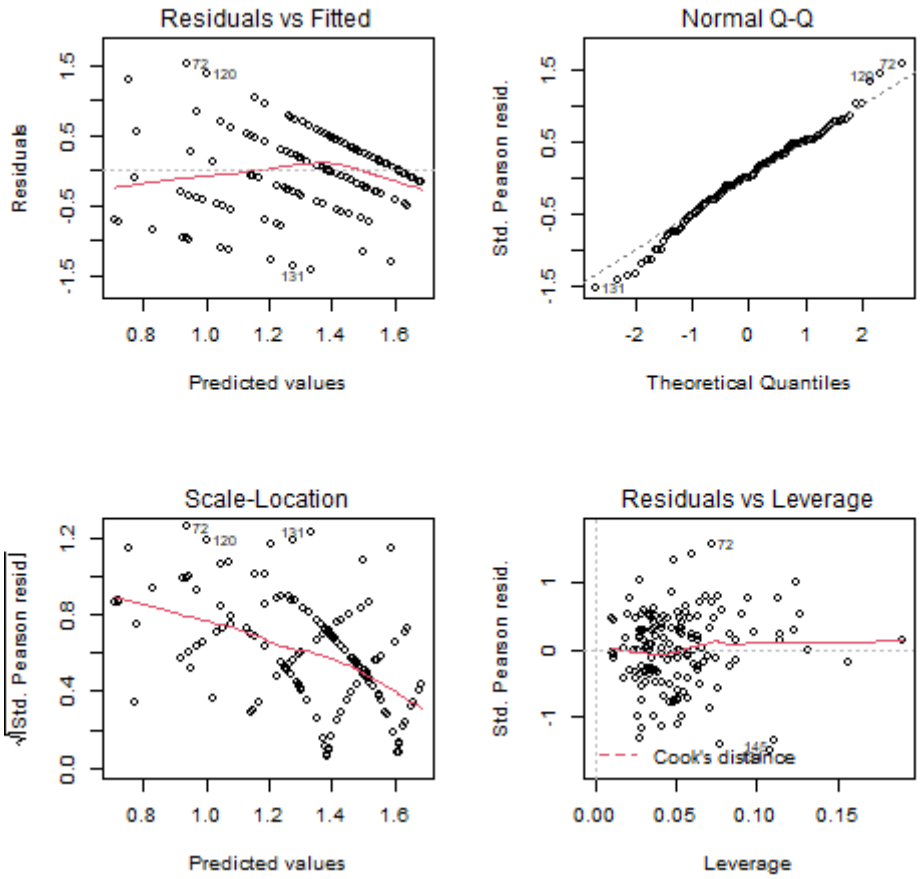


FIGURE 4.5: Residual plots of the logistic regression model of MILIEU factors

The plot of residual versus predicted values shown in Figure 4.5 was used to verify the assumption that residuals from the logistic regression model assessing the influence of MILIEU factors on software project scope creep management were normally distributed and with a uniform variance. Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of MILIEU factors used a quantile – quantile (Q-Q) plot to assess the degree of

normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal except for the values lying at the bottom area of the plot. The was one outlier data point at the bottom of the line, labelled as observation point 131. Most of the remaining data points lie along the line $y = x$ on the plot. At the top of the plot only a few points have deviated from this line, and they did not deviate too far except for three outlier data points labelled as 12 and 72.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of MILIEU factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot in Figure 4.5, most of the data points were lying outside the red line. There were three points that were highly dispersed from the regression line, i.e., observation points 72 and 145. A unique observation for this logistic regression model was that the distribution of data points on the Residual versus Leverage plot were skewed towards the right of the plot. In all the previous model plots the distribution was equal for both the left- and right-hand sides of the residual plots. Overall, the data points observed were

considered as influencing the variation that was observed in the dependent variable, Scope creep management.

4.3.5 Logistic Regression Model of METHOD Factors (Model 5)

The fifth model to be observed was the model of METHOD factors, which had four independent variables that were being assessed to determine their individual and collective influence on scope creep management. Table 4.6 presents results from this logistic regression analysis.

TABLE 4.6: Logistic regression outcomes of METHOD influence on scope creep management

Variable	Coef.	Std, error	t - value	Signif.
(Intercept)	5.556	0.239	23.194	0.000***
X32 - Changing market needs	-0.099	0.073	0.074	0.180
X33 - Poor initial requirements	-0.499	0.063	-7.950	0.000***
X34 - Quality issues	-0.126	0.070	-1.782	0.0767 *
X35 - Informal Approval of Changes	-0.328	0.013	-1.782	0.002***
Residual standard error	0.992 on 295 degrees of freedom			
Multiple R-squared	0.593			

Adjusted R-squared	0.581
F-statistic	12.62 on 4 and 295 DF (p-value: 2.59e-16)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the results of logistic regression analysis shown in Table 4.6, all the independent variables had a negative relationship with the dependent variable, Scope creep management. However, one of these variables namely, X32 - Changing market needs had no statistically significant influence on the dependent variable. Two variables, i.e., X33 - Poor initial requirements and X35 - Informal Approval of Changes were statistically significant at the 0.001 level. This means that when there were more poor initial requirements in a project, it interfered with the ability of software project managers to manage scope creep by almost half a unit of the measurement (i.e., -0.499). Similarly, an increase of Informal Approval of Changes in a software project decreased scope creep management possibilities by 0.328. Given that 3 out of 4 independent variables in this model were statistically significant, this meant that the model performance was improved.

The validation statistics of the logistic regression model measuring METHOD influences showed that the model performance was very good. The residual standard error was 0.992 on 300 degrees of freedom, since in this model there were 305 observations and five model variables (one dependent and four independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 2.59e-16. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.593 for multiple R-squared and 0.581 for adjusted R-squared. Both values suggest that the independent variables explained up to 60 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was high. Residual plots of the logistic regression model of METHOD factors are presented in Figure 4.6.

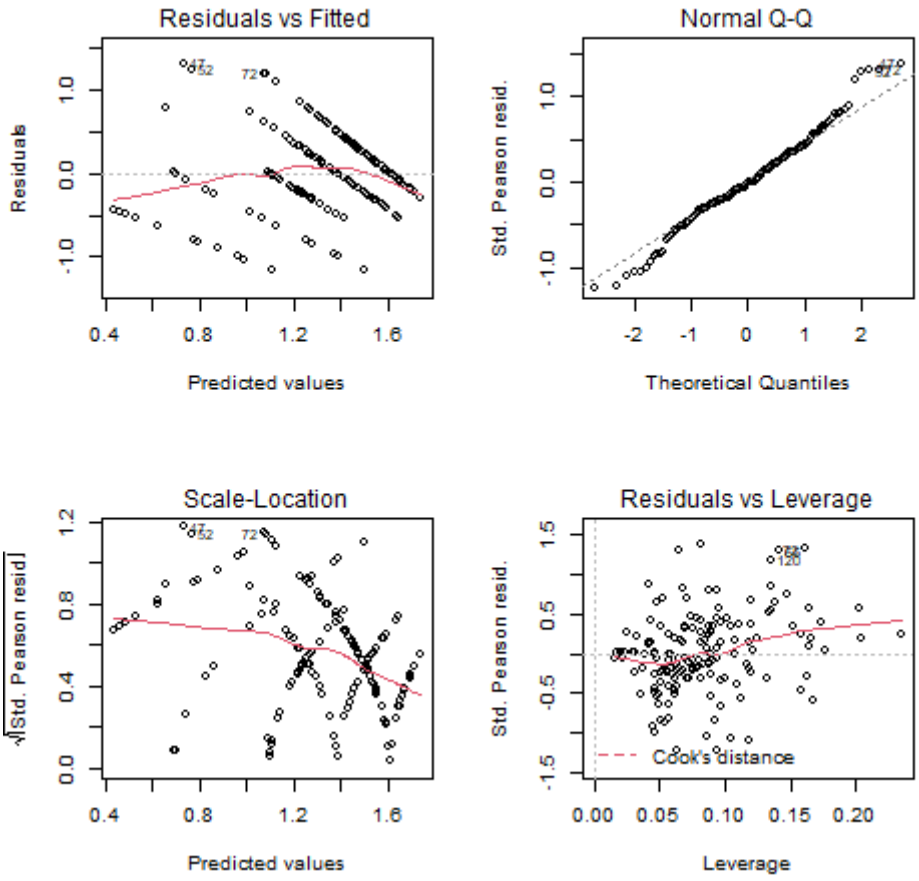


FIGURE 4.6: Residual plots of the logistic regression model of METHOD factors

The plot of residual versus predicted values shown in Figure 4.6 was used to verify the assumption that residuals from the logistic regression model assessing the influence of METHOD factors on software project scope creep management were normally distributed and

with a uniform variance. Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of METHOD factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal except for a few observation points that were located at the bottom area of the plot. Most of the remaining data points lie along the line $y = x$ on the plot. Similarly, at the top of the plot only a few points had a distribution that deviated from this line, and they did not deviate too far away except for a few outlier data points.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of METHOD factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot in Figure 4.6, most of the data points were lying

outside the red line. Only one point was highly dispersed from the regression line, i.e., observation point 122. Overall, the data points observed were considered as influencing the variation that was observed in the dependent variable, Scope creep management.

4.4 Discussion of Key Study Results

This section discusses the results that have been achieved from the study objectives. The main objective of the study was to establish a set of machine learning models that were used to assess the influence of scope creep factors on software project success using logistic a regression approach. This section therefore presents a discussion of results from each of the specific objectives in the context of results that have been found elsewhere.

4.4.1 Discussion of Results from Objective #1

The first specific objective of the study was to assess the critical factors responsible for scope creep risk in software projects. This was achieved through a descriptive analysis of the study variables. The aim of this initial analysis was to confirm or reject the hypotheses regarding the influence of study variables, as had been identified through a review of the literature. The descriptive analysis also sought to assess the trends in the study data set in order to determine the suitability of a logistic regression modelling approach for the study data set. First, the following measures were used to compare the study variables through the descriptive process: minimum value, median, mean (average), maximum value and standard deviation. The statistics were measured for a total of 36 variables. Out of these, there was one independent variable, scope creep management and a set of moderating variables (i.e., Organization size, project size, development approach used (i.e., agile, traditional or Hybrid) and whether the company employed control tools to manage scope creep. The rest of the study variables (i.e., 32 variables) were independent variables.

The highest mean of 8.33 was observed in the variable Manager Experience since this value was recorded in years. The dependent variable, Scope creep management also had a high

mean of 3.858. The lowest mean was observed in the variable Control tools used/ not used, with a value of 1.59. This variable was binary in nature (with values of Yes or No).

In terms of standard deviation, the highest value was observed in the variable named manager experience (X10) followed by the dependent variable scope creep management (Y) Most of the study variable had a standard deviation above 1.1, except for the moderating variables where only one of the variables had a value of 1.168 (Project size). The lowest value of standard deviation was seen in the variable Development approach. None of the observed study variables had a standard deviation value that was higher than its mean value. This suggested that there was no skewness in the data. There was thus confidence in applying the assumptions of a logistic regression modelling approach for this data set.

The second descriptive analysis was through a correlation analysis of the variables. The aim was to discover the existing association between study variables. This information can then be used to determine the nature of the relationship (whether positive or negative) as well as the strength or magnitude of the relationship. The correlation analysis generated a heatmap that included a total of 36 variables. From the observed results, most of the independent variables that were under study had some form of negative relationship with the dependent variable Scope creep management. Apart from the diagonal of ones, most independent variables shared a very weak positive correlation between themselves. Most of the moderating variables did not appear to be correlated to the independent variables.

4.4.2 Discussion of Results from Objective #2

The second specific objective of the study was to develop logistic regression models with the identified factors to predict scope creep of software projects. In order to achieve this objective, the study built separate logistic regression models for each of the variable categories listed in the conceptual framework, i.e., HUMAN Factors (Model 1), MEASUREMENT Factors (Model 2), ORGANIZATIONAL Factors (Model 3), MILIEU Factors (Model 4) and

METHOD Factors (Model 5). Each of these models had a different number of independent variables, where the model with HUMAN factors as the independent variables had the largest size, with 9 variables whose effects in scope creep management were being observed. The smallest model was that observing the influence of METHOD factors on scope creep management, and it had 4 independent variables. The model of MEASUREMENT had 5 independent variables. The model of ORGANIZATION had six independent variables, while MILIEU had 7 variables being regressed against the dependent variable.

For all the observed logistic regression models, most independent variables had a negative relationship with the dependent variable, Scope creep management. This means that an increase in the value of each of the independent variables caused the value of the dependent variable to decrease by a certain amount. In other words, these variables are all detrimental to the management of scope creep in software projects.

From Model #1, i.e., the model measuring HUMAN factors, three of the independent variables were significant at the 0.001 level. These variables were X7 - Poor Communication, X11 - Unclear goals, and X13 - Inexperienced Staff. Two variables, i.e., X6 - Lack of knowledge and X8 - Stakeholder involvement was significant at the 0.01 level. One variable, namely X9 - Lack of timely feedback was significant at the 0.05 level. This means that this variable was at the borderline of statistical significance in its influence on the dependent variable. Bayona (2020) has similarly discovered factors such as involvement of users, support, communication, and commitment to influence scope creep in software projects. Although the remaining three (X5 - Unrealistic Expectations, X10 - Manager Experience, X12 – Ego) were negative to the dependent variable Scope creep management, their influence was not statistically significant.

For Model #2, i.e., the model assessing MEASUREMENT factors, two of the independent variables were significant at the 0.000 level. These variables were X15 - Underestimating Change and X16 - Client's Knowledge. When change occurring in a software project was underestimated by one unit, it caused a decrease of 0.279 (i.e., 28 percent decrease) in the unit value that was being used to measure scope creep management. Additionally, a one-unit increase in the client's knowledge meant a 24 percent decrease (i.e., -0.239) in the ability to manage scope creep in software projects. The variable X17 - Developer's Inexperience was significant at the 0.01 level. This inexperience caused a 21 percent decrease (i.e., -0.206) in the ability to manage scope creep in the software projects. Similarly, Ahmadi et al. (2022) conducted a survey of senior project management practitioners involved in complex projects and discovered that the experience and competencies of project managers were the most essential factor for successful projects. Although the remaining two variables (namely, X15 - Underestimating Change and X18 - Over-accommodating Client) were negative to the dependent variable Scope creep management, their influence was not statistically significant. This means that their negative impact did not statistically affect scope creep management.

For Model #3, i.e., the model assessing ORGANIZATION factors, one of the independent variables was significant at the 0.001 level. This variable was X19 - Budget constraint. As constraints in the project budget increased, the increase in turn caused a 39 percent decrease (i.e., -0.389) in the ability to manage project scope creep. One variable, X22 - Lack of resources, was significant to scope creep management at the level of 0.01. When a software project lacked one unit of resources, it caused a decrease of -0.207 in scope creep management. Although two model variables (namely, X20 - Project Complexity and X24 - Standards and policies) were negative to the dependent variable Scope creep management, their influence was not statistically significant. This means that their negative impact did not statistically affect scope creep management.

For Model #4, i.e., the model assessing MILEU factors, two of the independent variables were significant at the 0.001 level. These variables were X30 - Changing requirements and X31 - Unforeseen Risk. As requirements in the software project changed by one unit, this change caused a downward change (-0.314) in scope creep management. Similarly, an upward change in the unforeseen risk by one unit of measurement caused a reduction by 41 percent (i.e., -0.408) in scope creep management ability in software projects. One variable, i.e., X28 - Time constraint was significant to scope creep management at the level of 0.01. When time constraints increased by a single unit, it caused a decrease of 2 percent (i.e., -0.002) in scope creep management for software projects. Two model variables were statistically significant to scope creep management at the level of 0.05. These were X27 - Requirement Volatility and X29 – Uncertainty. When requirements became more volatile, it increased the risk of poor scope creep management by -0.064 (i.e., by 6 percent). Additionally, uncertainty in a project was detrimental to scope creep management. A rise in this variable value by one caused a decrease in the value of the dependent variable by 0.102.

For Model #5, i.e., the model assessing METHOD factors, all the independent variables had a negative relationship with the dependent variable, Scope creep management. However, one of these variables namely, X32 - Changing market needs had no statistically significant influence on the dependent variable. Two variables, i.e., X33 - Poor initial requirements and X35 - Informal Approval of Changes were statistically significant at the 0.001 level. This means that when there were more poor initial requirements in a project, it interfered with the ability of software project managers to manage scope creep by almost half a unit of the measurement (i.e., -0.499). Similarly, an increase of Informal Approval of Changes in a software project decreased scope creep management possibilities by 0.328. These observations are similar to the observations by Elkhatib et al. (2022) who discovered that risks related to the organization, technology, process, monitoring and analysis were the most influential

underlying risks for agile project management. Given that 3 out of 4 independent variables in this model were statistically significant, this meant that the model performance was improved.

4.4.3 Discussion of Results from Objective #3

The third specific objective of the study was to test and validate the developed model. This was achieved through examination of results from model validation metrics to confirm the performance of the developed models in predicting scope creep management in software projects. The validation statistics of the logistic regression model measuring HUMAN influences (Model 1) showed the model performance to be good. In this model, there were 305 observations and ten model variables (one dependent and nine independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value < 2.2e-16. This implied that the difference among the observed groups was highly statistically significant. Other model validation tests included the multiple R-squared and Adjusted R-Squared (which accounts for the influence of multiple independent variables in the model). The values of 0.645 for multiple R-squared and 0.617 for adjusted R-squared both suggest that the independent variables explained up to 65% of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was high.

The plot of residual versus predicted values was used to verify the assumption that residuals from the logistic regression model measuring the HUMAN factors were normally distributed and with a constant (uniform) variance. Based on the observation of Figure 4.2, the observation points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation

points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of HUMAN factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot displays graphically how any two quantiles of a distribution have lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal since most of the data points lie along the line y = x on the plot. Only a few points have deviated from this line, and they did not deviate too far except for two outlier data points labelled as 52 and 47.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of HUMAN factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship n=between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that these observation points were influential observations. From the observation of the Residuals vs Leverage plot, most of the data points were lying outside the red line and were therefore considered as influencing the variation that was observed in the dependent variable, Scope creep management.

The validation statistics of the logistic regression model measuring MEASUREMENT influences (Model 2) showed the model performance was moderately good. The residual standard error was 1.029 on 299 degrees of freedom, since in this model there were 305 observations and six model variables (one dependent and five independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 6.799e-13. This implied that the difference among the observed groups was highly statistically significant. Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.575 for multiple R-squared and 0.534 for adjusted R-squared. Both values suggest that the independent variables explained up to 58 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was moderately high.

The plot of residual versus predicted values was used to verify the assumption that residuals from the logistic regression model assessing the influence of MEASUREMENT factors on software project scope creep management were normally distributed and with a uniform variance. Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of HUMAN factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable

and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal except for the values lying at the bottom area of the plot. Most of the remaining data points lie along the line y = x on the plot. At the top of the plot only a few points have deviated from this line, and they did not deviate too far away except for three outlier data points labelled as 72, 77 and 120.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of MEASUREMENT factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot, most of the data points were lying outside the red line and were therefore considered as influencing the variation that was observed in the dependent variable, Scope creep management.

The validation statistics of the logistic regression model measuring ORGANIZATION influences (Model 3) showed the model performance was moderately good. The residual standard error was 0.997 on 298 degrees of freedom, since in this model there were 305 observations and seven model variables (one dependent and six independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value

of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 1.852e-14. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.602 for multiple R-squared and 0.578 for adjusted R-squared. Both values suggest that the independent variables explained up to 60 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was moderately high. Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of ORGANIZATIONAL factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is almost perfectly normal, except for two data points that lie a little off the line, i.e., 72 and 144. All the remaining data points lie along the line y = x on the plot.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of ORGANIZATIONAL factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of

data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot, almost all of the data points were lying outside the red line and were therefore considered as influencing the variation that was observed in the dependent variable, Scope creep management.

The validation statistics of the logistic regression model measuring MILIEU influences (Model 4) showed the model performance was good. The residual standard error was 0.992 on 297 degrees of freedom, since in this model there were 305 observations and eight model variables (one dependent and seven independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 4.972e-14. This implied that the difference among the observed groups was highly statistically significant.

Other model validation tests included the multiple R-squared and Adjusted R-Squared which accounts for the influence of multiple independent variables in the model. The values of 0.704 for multiple R-squared and 0.688 for adjusted R-squared. Both values suggest that the independent variables explained up to 69 percent of the variation that was observed in the scope creep management. This implied that the model's goodness of fit was high.

Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable

patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of MILIEU factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is normal except for the values lying at the bottom area of the plot. The was one outlier data point at the bottom of the line, labelled as observation point 131. Most of the remaining data points lie along the line y = x on the plot. At the top of the plot only a few points have deviated from this line, and they did not deviate too far except for three outlier data points labelled as 12 and 72.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of MILIEU factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot in Figure 4.5, most of the data points were lying outside the red line. There were three points that were highly dispersed from the regression

line, i.e., observation points 72 and 145. A unique observation for this logistic regression model was that the distribution of data points on the Residual versus Leverage plot were skewed towards the right of the plot. In all the previous model plots the distribution was equal for both the left- and right-hand sides of the residual plots. Overall, the data points observed were considered as influencing the variation that was observed in the dependent variable, Scope creep management.

The validation statistics of the logistic regression model measuring METHOD influences showed that the model performance was very good. The residual standard error was 0.992 on 300 degrees of freedom, since in this model there were 305 observations and five model variables (one dependent and four independent variables). The residual statistics showed that there were no significant errors in the estimation performed by the model. The F-statistic metric of the model was greater than the critical value, i.e., the value of F-statistic that corresponds with the alpha value of 0.05. In this model, the F-statistic was statistically significant at p-value = 2.59e-16. This implied that the difference among the observed groups was highly statistically significant.

Based on the observation results, the data points are randomly distributed on both the left and right sides of the residual line = 0 in the residual plot. There are also no recognizable patterns in the distribution of observation points. This confirmed the goodness of fit of the model that had been observed from the R-squared values.

The plot of standardized residuals versus theoretical quantiles from the logistic regression model of METHOD factors used a quantile – quantile (Q-Q) plot to assess the degree of normality. The Q-Q plot was used to display graphically how any two quantiles of a distribution were lined up. It used the theoretical distribution, i.e., the normal distribution as its x-variable and the model residuals as the y-variable. Based on the model's Q-Q plot, the distribution is

normal except for a few observation points that were located at the bottom area of the plot. Most of the remaining data points lie along the line y = x on the plot. Similarly, at the top of the plot only a few points had a distribution that deviated from this line, and they did not deviate too far away except for a few outlier data points.

The plot of standardized Pearson residuals against the predicted values was used to present the scale-location characteristic of the logistic regression model measuring the influence of METHOD factors on Scope creep management. The standardized Pearson residuals were plotted along the y axis against the predicted log odds on the x axis. The pattern of data points generated was a non-uniform one with somewhat parallel lines of data points which also deviated from the regression line. This observation showed that there was no specific relationship between the predicted values and the model residuals.

The plot of Cook's distance for the standardized Pearson residuals versus leverage. Deviation of observation points from the Cook's distance, i.e., the red dashed lines on this plot indicated that most of the observed data points were influential observations. From the observation of the Residuals vs Leverage plot in Figure 4.6, most of the data points were lying outside the red line. Only one point was highly dispersed from the regression line, i.e., observation point 122. Overall, the data points observed were considered as influencing the variation that was observed in the dependent variable, Scope creep management.

4.5 Chapter Summary

This chapter has presented results from the achieved objectives. It has outlined the results from descriptive statistics of all the model variables, as well as the correlation analysis of these variables. Furthermore, the chapter has systematically presented and offered interpretation of results from logistic regression of different factors on the dependent variables, Scope creep management. For each of the models, the chapter has presented results from validation tests which have shown that all models had acceptable goodness of fit.

This chapter has also presented an extensive discussion of the key findings observed from the analysis as well as a comparison of these findings.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter provides a conclusion based on the results that have been observed. The conclusion reflects upon the contributions made by the research to the field of software project management research and elucidates the broader significance of employing logistic regression models for scope creep analysis and prediction. Based on these results, the chapter outlines several key recommendations that can guide future research, project management interventions, and software project decisions in the scope creep control and prevention based on the findings and insights derived from the project on logistic regression modelling in multivariate analysis and forecasting of factors influencing scope creep.

These recommendations are aimed at improving the accuracy of forecasts, enhancing targeted interventions, and advancing our understanding of the complex dynamics surrounding scope management in software projects.

5.2 Key Contributions of the Study

This research sought to make significant contributions to the existing research, and in this regard the study largely has achieved this goal. One of the primary contributions of this project lies in its adoption of a multivariate approach to logistic regression modelling. Traditional models such (Komal et al., 2020) as have often overlooked the interconnectedness of various contributing factors, leading to limited accuracy in forecasting. By incorporating a diverse set of independent variables, the project acknowledges the interplay between human, measurement, organization, milieu and method factors, resulting in a more holistic understanding of how these factors jointly influence scope creep in software projects. The incorporation of these multiple variables that can potentially influence scope creep in software

projects underscores the project's practical relevance. By integrating these factors into the analysis, the research recognizes the multifaceted nature of scope creep management. This acknowledges the importance of different sources of influence which can potentially affect scope management in software projects.

Another strength and important contribution of the current research is with regards to the separation of analysis for different categories of influential factors. A set of separate logistic regression models were built to assess related factors within the human, measurement, organization, milieu and method categories. By analyzing these different domains in isolation, the study avoided the bias that comes with regressing unrelated factors. In doing this the research avoided indirect relationships among independent variables that can affect their effect in the dependent variable.

Another important contribution of this study to the existing literature and body of knowledge is the application of exploratory data analysis and regression modelling using survey data from the software project managers in multiple countries across the globe. During the COVID-19 and post-COVID-19 periods, the infrastructure for working remotely has greatly advanced and many projects can now be conducted without limitations of country borders. Therefore, the dataset comprising survey responses from software project managers across the world that this study has applied increases the generalizability of the observed findings. Additionally, the current study has expanded the scope using many survey responses (305) than what has been observed in other previous studies such as Ahmadi et al. (2022). The large size of the questionnaire also makes it possible to establish the ability to ask and answer more context relevant questions with a lower risk of misinterpretation. More in-depth analyses can be performed when a clear objective has been presented.

The utilization of a logistic regression modelling approach adds another layer of sophistication to the analysis. One of the major strengths of logistic regression as compared to other related techniques such as probit regression is the convenience of interpreting the exponentiated slope of the logistic regression coefficient (e^b) as the odds ratio (Schober & Vetter, 2021). This ratio is an indicator of how much the odds of a certain outcome under observation will change for a one-unit increase in the independent variable when dealing with continuous) or versus a reference category, when the independent variable has categorical values. This ease of interpreting the outcomes makes logistic regression a preferred approach.

The findings of this study can guide different stakeholders, such as software project managers, project developers and users of software. Accurate predictions of project creep factors will be of great greatly help project managers. The study will also benefit software developers and software development companies such as start-ups since they can use the model to properly manage their software development projects. They will be able to complete all defined tasks without bad multi-tasking where they are forced to deliver more tasks, but the quality suffers. Users on the other hand will benefit by obtaining quality software on time, which can increase their satisfaction with the software and the software projects. Predicting scope creep factors that impact badly on project success will improve communication between the users and the project team and will promote trust-based relationships to make it easier to discuss and address any rising issue. Finally, the study will benefit future researchers as they will be able to use the developed model as a case study to build better models in future.

5.3 Conclusions from the Study Results and Achieved Objective

The main objective of this study was to predict scope creep of software projects using logistic regression analysis. The following conclusions can be made based on the observed results.

5.3.1 Conclusions for Objective One

The first objective was to assess and identify the attributes that influence scope creep in software projects. The study achieved this objective through a descriptive analysis of the study variables. The descriptive analysis identified trends in the study data set that helped to confirm the suitability of a logistic regression modelling approach for the study data set. Descriptive measures of minimum value, median, mean (average), maximum value and standard deviation were performed for a total of 36 variables. The highest means were observed for Manager Experience with 8.33 and the dependent variable, Scope creep management with 3.858. The lowest mean was observed in the variable Control tools used/ not used, with a value of 1.59. The highest value of standard deviation was observed in the variable named Manager Experience and the lowest was seen in the variable Development Approach. None of the observed study variables had a standard deviation value that was higher than its mean value suggesting no skewness in the data. There was thus confidence in applying the assumptions of a logistic regression modelling approach for this data set. This means that objective 1 was achieved.

5.3.1 Conclusions for Objective Two

The second objective was to develop logistic regression models with the identified factors to predict scope creep of software projects. To achieve this objective, the study built separate logistic regression models for each of the variable categories listed in the conceptual framework, i.e., HUMAN Factors (Model 1), MEASUREMENT Factors (Model 2), ORGANIZATIONAL Factors (Model 3), MILIEU Factors (Model 4) and METHOD Factors

(Model 5). 17 out of 31 independent variables across the five models were found to be statistically significant in their influence on the dependent variable. To this extent objective 2 was achieved.

5.3.1 Conclusions for Objective Three

The third and final objective was to test and validate the developed models. Based on the validation results, all the five models showed good performance, with the coefficient of determination (R-square) ranging between 0.554 for the model of MEASUREMENT factors and 0.704 for the model of MILIEU factors. The five models also showed adherence to the assumption of normality since the residuals were constantly distributed with uniform variance. This means that objective 3 was achieved.

5.4 Limitations of the Study

While the current research has made important contributions, it also has several limitations. One of the limitations of this study is that findings are usually influenced by the geographic area for which the data has been collected. Scope creep patterns can vary widely across different regions due to the contextual characteristics of project managers, developers, users, technical environment and the dynamism of the software ecosystem. The lack of available sufficient time to conduct this study made it impossible to conduct a comprehensive local survey studies of software companies in Kenya. It is possible, therefore, that the exclusive usage of survey response data from project managers across the globe has prevented the observation of local patterns. Therefore, the results obtained might not be directly applicable to the local context in Kenya without proper validation. Nevertheless, it is possible to argue that the factors that have been presented in this study can be deemed as being sufficiently representative of observations that can be expected from any specific geographic region of interest. Therefore, the results observed in this study can be reasonably regarded to be relevant and reliable.

Secondly, although the study incorporated a comprehensive set of independent variables, there may be other potential covariates that could influence scope creep. Socioeconomic factors, alternative software, and management support are examples of variables that were not included but may play a significant role in scope creep management dynamics. The exclusion of these variables may have resulted in an incomplete model. Thirdly, the study employed a relatively small number of survey responses to monitor scope management in software projects. The project managers that were assessed were only 305, which is a very small proportion of the global population of software project managers.

5.5 Recommendations for Future Research

In addressing the identified limitations, this study makes several recommendations for future research. First, studies can assess scope creep factors for local software projects in Kenya by conducting a survey analysis of local project managers in the various start-ups and established software companies. A comprehensive local survey of software companies in Kenya will increase the practicability of inference from the observations.

Secondly, studies can consider a more comprehensive set of independent variables that could potentially influence scope creep. Socioeconomic factors, alternative software, and management support are examples of variables that were not included but may play a significant role in scope creep management dynamics. The inclusion of these variables may result in a more complete model. Finally, future studies can employ a larger number of observations to monitor scope management in software projects. The project managers that were assessed in this study numbered only 305, which is a very small proportion of the global population of software project managers.

5.6 Chapter Summary

This chapter has outlined several key recommendations that can guide future research, project management interventions, and software project decisions in the scope creep control and prevention. These recommendations were based on the findings and insights derived from the project on logistic regression modelling in multivariate analysis and forecasting of factors influencing scope creep. The recommendations are aimed at improving the accuracy of forecasts, enhancing targeted interventions, and advancing the general understanding of factors influencing scope creep management in software projects.

REFERENCES

Ahmadi Eftekhari, N., Mani, S., Bakhshi, J., & Mani, S. (2022). Project manager competencies for dealing with socio-technical complexity: a grounded theory construction. Systems, 10(5), 161.

Arora, M., Verma, S., Kavita, & Chopra, S. (2020). A systematic literature review of machine learning estimation approaches in scrum projects. Cognitive Informatics and Soft Computing: Proceeding of CISC 2019, 573-586.

Bisong, E., & Bisong, E. (2019). Logistic regression. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 243-250.

Cobb, C. G. (2023). The project manager's guide to mastering Agile: Principles and practices for an adaptive approach. John Wiley & Sons.

Das, A. (2021). Logistic regression. In Encyclopedia of Quality of Life and Well-Being Research (pp. 1-2). Cham: Springer International Publishing.

DeLone, W.H. and McLean, E.R. (2003), "The DeLone and McLean model of information systems success: a ten-year update", Journal of Management Information Systems, Vol. 19 No. 4, pp. 9-30.

Efron, B. (2000). The bootstrap and modern statistics. Journal of the American Statistical Association, 95(452), 1293-1296.

Elkhatib, M., Al Hosani, A., Al Hosani, I., & Albuflasa, K. (2022). Agile Project Management and Project Risks Improvements: Pros and Cons. Modern Economy, 13(9), 1157-1176.

Galal, S (2023) Startup failure rate in Africa 2020, by country. Accessed from https://www.statista.com/statistics/1295678/startup-failure-rate-in-africa-by-country/

Gogtay, N. J., & Thatte, U. M. (2017). Principles of correlation analysis. Journal of the Association of Physicians of India, 65(3), 78-81.

Kamer, L. (2023) Number of tech startups that raised funding in Kenya 2015-2022. Available at https://www.statista.com/statistics/1279467/number-of-funded-startups-in-kenya/

Komal, B., Janjua, U. I., Anwar, F., Madni, T. M., Cheema, M. F., Malik, M. N., & Shahid, A. R. (2020). The impact of scope creep on project success: An empirical investigation. IEEE Access, 8, 125755-125775.

Kurkovsky, S. (2022, July). Managing Scope in Service Learning Projects. In Proceedings of the 27th ACM Conference onon Innovation and Technology in Computer Science Education Vol. 2 (pp. 620-620).

Lalic, C. D., Lalic, B., Delić, M., Gracanin, D., & Stefanovic, D. (2022). How project management approach impact project success? From traditional to agile. International Journal of Managing Projects in Business, 15(3), 494-521.

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2022). Measuring user engagement. Springer Nature.

Linares-Vásquez, M., McMillan, C., Poshyvanyk, D., & Grechanik, M. (2014). On using machine learning to automatically classify software applications into domain categories. Empirical Software Engineering, 19, 582-618.

Luo, L., Arizmendi, C., & Gates, K. M. (2019). Exploratory factor analysis (EFA) programs in R. Structural Equation Modeling: A Multidisciplinary Journal, 26(5), 819-826.

Mahaney, R. C., & Lederer, A. L. (2011). An agency theory explanation of project success. Journal of Computer Information Systems, 51(4), 102-113.

Mehta, S., & Patnaik, K. S. (2021). Improved prediction of software defects using ensemble machine learning techniques. Neural Computing and Applications, 33, 10551-10562.

Montgomery, L., Fucci, D., Bouraffa, A., Scholz, L., & Maalej, W. (2022). Empirical research on requirements quality: a systematic mapping study. Requirements Engineering, 27(2), 183-209.

Riaz, A. R., & Gilani, S. M. (2022). Risk assessment approach for software development using cause and effect analysis. KIET Journal of Computing and Information Sciences, 5(1), 48-61.Iriarte, C., &Bayona, S. (2020). IT projects success factors: a literature review. International Journal of Information Systems and Project Management, 8(2), 49-78.

Pace, M. (2019). A correlational study on project management methodology and project success. Journal of Engineering, Project, and Production Management, 9(2), 56.

Pisner, D. A., &Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.

Rathore, S. S., & Kumar, S. (2021). Software fault prediction based on the dynamic selection of learning technique: findings from the eclipse project study. Applied Intelligence, 1-16.

Sahadevan, S. (2023). Project Management in the Era of Artificial Intelligence. European Journal of Theoretical and Applied Sciences, 1(3), 349-359.

Salamzadeh, A., Tajpour, M., Hosseini, E., & Brahmi, M. S. (2023). Human capital and the performance of Iranian Digital Startups: The moderating role of knowledge sharing behaviour. International Journal of Public Sector Performance Management, 12(1-2), 171-186.

Sheikhalishahi, M., Amani, M. A., & Behdinian, A. (2022). Evaluating Factors Affecting Project Success: An Agile Approach. Journal of Industrial Engineering International, 18(1), 79-96.

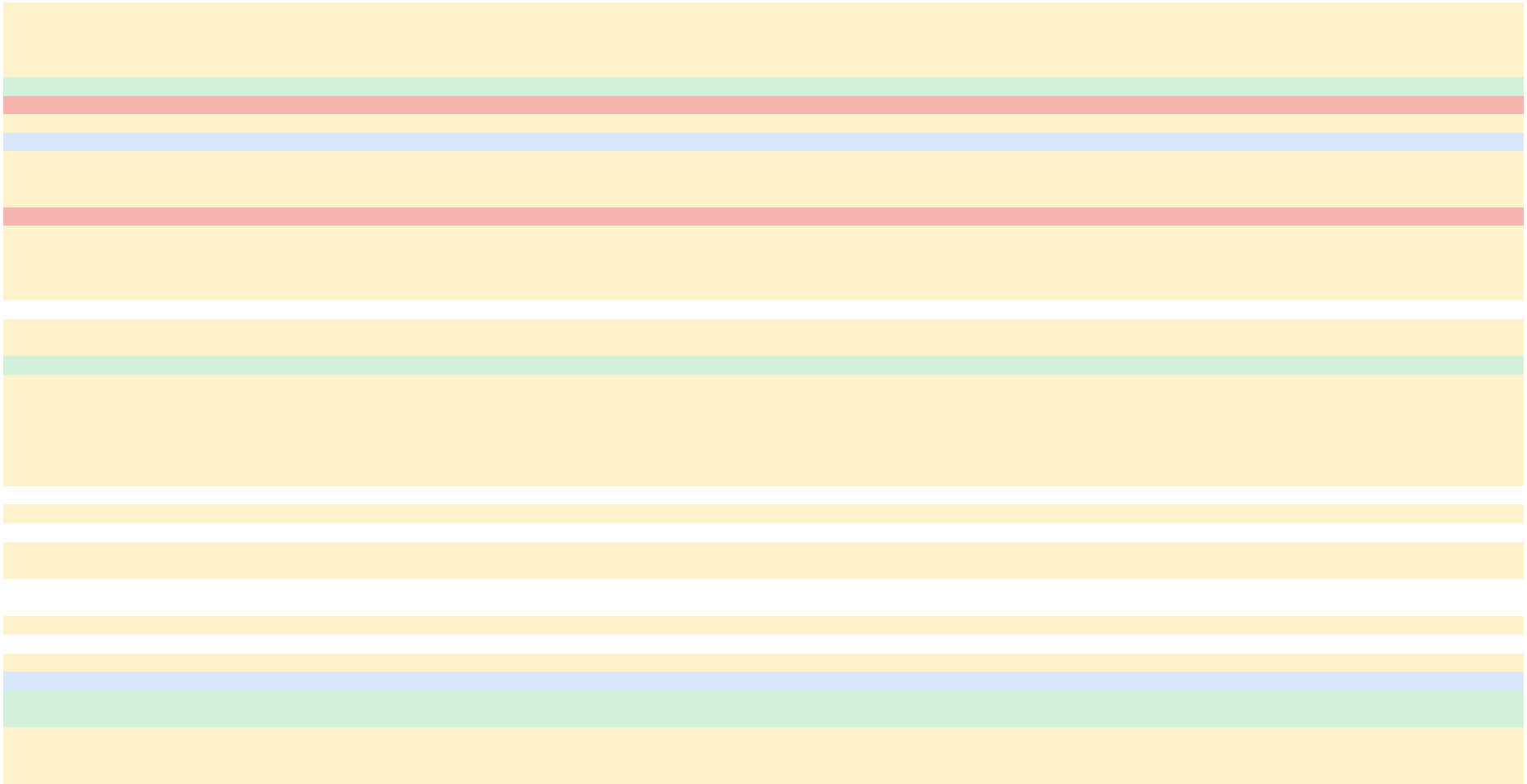
Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. Anesthesia and analgesia, 132(2), 365.

Schoonwinkel, S., Fourie, C. J., & Conradie, P. D. F. (2016). A risk and cost management analysis for changes during the construction phase of a project. Journal of the South African Institution of Civil Engineering, 58(4), 21-28.

Sileyew, K. J. (2019). Research design and methodology. Cyberspace, 1-12.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of analysis and testing, 2(3), 249-262.

APPENDIX 1: Header of the sample data set used for the study.



X1 = Development approach, X2 = tools used, X3 = name of tool, X4 = Poor Communication, X5 = Project Complexity, X6 = Lack of knowledge, X7 = Unrealistic Expectations, X8 = Time constraint, X9 = Project size, X10 = Stakeholder involvement, X11 = Requirement Volatility, X12 = Budget constraint, X13 = uncertainty, X14 = Poor scope management, X15 = Quality issues, X16 = Organizational capabilities, X17 = Changing requirements, X18 = Lack of resources, X19 = Lack of timely feedback, X20 = Fixed cost, X21 = Unclear goals,

X22 = Standards and policies, X23 = Ego, X24 = Inexperienced Staff, X25 = Poor initial requirements, X26 = No formal review, X27 = Unforeseen Risk, X28 = Changing market needs

APPENDIX 2: RESEARCH BUDGET

TABLE 7.1: Budget

No.		Item	Unit	Price	Total
1		Computer and hard disk (SSD)	1	50,000	50,000
2		Internet and data bundles	5	5,000	25,000
3		Travel cost to meet Supervisor	30	1,500	45,000
4		Data Collection and cleaning			
5		Model building and testing software	1	1,500	1,500
6		Final Dissertation preparation (i.e., printing, binding, etc.)	500	20 per page	10,000
7		Flash Disk	1	1,500	1,500
8		Miscellaneous expenses	1		5,000
9		Total			138,000

APPENDIX 3: RESEARCH SCHEDULE

TABLE 7.2: Gantt Chart

	YEAR 2023	PR	AY	UN	UL	UG	EP	CT
	Ideation							
	Draft Research Questions							
	Writing Research Proposal							
	Review of Literature							
	Proposal Presentation							
	Data Collection							
	Model Formulation							
	Data Analysis							
	Model Validation							
0	Compiling the work							
1	Final defense							
2	Document submission							