TIME-TO-EVENT MODELS: COMPARISON, MODIFICATION AND APPLICATION

By
Nauman Ahmad
PIDE2017FPHDETS03

Supervised By
Dr Amena Urooj
Assistant Professor
PIDE Islamabad

Co-Supervised By
Dr Saud Ahmed Khan
Assistant Professor
PIDE Islamabad

PIDE School of Economics
Pakistan Institute of Development Economics, Islamabad
2025

Author's Declaration

I Mr. Nauman Ahmad hereby state that my PhD thesis titled **"Time-To-Event Models: Comparison, Modification and Application"** is my own work and has not been submitted previously by me for taking any degree from **Pakistan Institute of Development Economics, Islamabad'** or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my PhD degree.

Mr. Nauman Ahmad

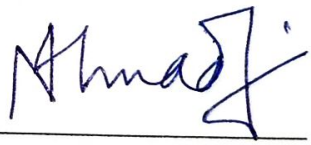
PIDE2017FPHDETS03

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled **"Time-To-Event Models: Modification and Application"** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the **HEC** and **Pakistan Institute of Development Economics, Islamabad** towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the rights to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Students/Author Signature: 

Mr. Nauman Ahmad

PIDE2017FPHDETS03

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled: **"Time-To-Event Models: Comparison, Modification and Application"** was conducted by **Mr. Nauman Ahmad** under the supervision of **Dr. Amena Urooj and Dr. Saud Ahmed Khan.** No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Econometrics from **Pakistan Institute of Development Economics, Islamabad.**

Student Name: Mr. Nauman Ahmad
PIDE2017FPHDETS03

Signature: _____

Examination Committee:

a) **External Examiner: Dr. Eatzaz Ahmed**
Professor/ Dean
Iqra University, Islamabad

Signature: _____

b) **Internal Examiner: Dr. Iftikhar Hussain Adil**
Associate Professor
NUST, Islamabad

Signature: _____

c) **Internal Examiner: Dr. Ahsan ul Haq**
Assistant Professor
PIDE, Islamabad

Signature: _____

Supervisor: Dr. Amena Urooj
Assistant Professor,
PIDE, Islamabad

Signature: _____

Co-Supervisor: Dr. Saud Ahmed Khan
Assistant Professor,
PIDE, Islamabad

Signature: _____

Dr. Iftikhar Ahmad
Head, PIDE School of Economics (PSE)
PIDE, Islamabad

Signature: _____

DEDICATION

Dedicated to my beloved father Mukhtiar Ghani (Retired Head Master), whose unwavering support and guidance inspired me daily, To my loving mother, whose boundless support, prayers and endless sacrifices have been the foundation of my journey. To my brother and sisters, for their definite support and companionship throughout this journey.

ACKNOWLEDGMENT

In the name of Allah, the Most Gracious, the Most Merciful. May peace and blessings be upon the Prophet Muhammad (PBUH).

All praise is due to **Allah Almighty**, whose endless blessings, guidance, and mercy enabled me to undertake and complete this research. Without His grace, this journey would not have been possible.

Over the course of my class work and thesis at the **Pakistan Institute of Development Economics (PIDE)**, I have been indebted to many professors, mentors, and colleagues whose support and guidance have enriched my academic journey. Among them, I am especially grateful to my supervisor, **Dr. Amena Urooj**, for his invaluable mentorship, continuous encouragement, and constructive feedback. His insights have been instrumental in shaping my research. I am also profoundly thankful to my co-supervisor, **Dr. Saud Ahmed Khan**, for his expert advice, thoughtful guidance, and constant motivation, which significantly contributed to the refinement of this work.

I would like to acknowledge my **external supervisor, Dr. Iftikhar Hussain Adil NUST**, whose guidance and expertise played a vital role in the development of this study. My sincere thanks also go to internal examiner **Dr. Ahsan Ul Haq Satti** for his valuable insights and academic support. I am also deeply thankful to the two "**Anonymous foreign evaluators"** for their critical review and valuable feedback, which contributed to enhancing the quality and rigor of my research. My sincere appreciation also extends to **Dr. Eatzaz Ahmed**, who served as the **final viva examiner**, for his insightful comments and constructive suggestions that helped refine the final presentation of my work. Their

evaluations played a crucial role in strengthening my research and ensuring its academic integrity.

I am grateful to the **faculty at the Pakistan Institute of Development Economics (PIDE)** for providing an excellent academic environment that facilitated my research. I sincerely appreciate the support and encouragement of **Dr. Ateeq, Dr. Hafsa Hina** and **Dr. Zahid,** whose guidance and contributions enriched my academic experience. I would also like to express my heartfelt appreciation to my friends at PIDE, particularly **Dr. Haseen Shah** my classfellow **Dr. Sara** and my junior **Tariq Khan and Asad Shahbaz** for their steady support, encouragement, and insightful discussions, which have been a source of both intellectual enrichment and emotional strength throughout this journey.

My **beloved parents** untiring support, sacrifices, and prayers have been the cornerstone of both my academic and personal accomplishments.

Finally, I extend my heartfelt gratitude to my **brother and sisters**, whose constant support and love have always been a source of encouragement. This work is dedicated to all of you, in recognition of your sacrifices, patience, and steadfast belief in my abilities.

Dr. Nauman Ahmad

ABSTRACT

The Cox model is primarily used in time-to-events models for the specific period of study, from the origin to the event of interest. The Cox model has a certain limitation: no outlier, no heteroscedasticity, and no time-dependent covariates are assumed in the data. In case of the existence of any of the above in data, the Cox model fails to estimate the true effect of covariates. A specific model is used, in the literature, for each problem. If there is an outlier in the data, robust Cox is mostly used. If there is heteroscedasticity in the data, the WLS model can be used. If there is a time-dependent covariate in the data, the time-dependent Cox model can be used. However, there is no unique model for handling three problems simultaneously.

This study modifies the existing Cox model that simultaneously solves the problems of handling outliers, heteroscedasticity, and time-dependent covariates in censored time-to-event data. It compares the Cox model, robust Cox, weighted least square, time-dependent Cox, and modified Cox models. For simulation experiments, four scenarios are considered, allowing outliers, heteroscedasticity, and time-dependent Cox with varying sample sizes, outlier quantities, and magnitudes. The first scenario deals with the presence of the outlier and heteroscedasticity case. We evaluate the performance of the modified Cox model and other existing models and find that the robust Cox outperformed other models. The increase in sample size results in a slightly improved performance of all models. The outlier position doesn't affect the model performance due to cross-sectional behavior. Performance indicator RMSE slightly improves in all models if the parameter theta of exponential distribution increases from one to two. In the second scenario, we evaluate the performance of the modified Cox model in the presence of outlier, heteroscedasticity, and time-dependent covariates. The results show that the RMSE, MAE, and MAPE of modified Cox are smaller than other survival analysis models. Outlier quantity and magnitude decrease the performance of all models, and with the increase in sample size, all models' performance increases; the RMSE value for all models improves slightly with the increase in the exponential distribution theta parameter value from one to two and increase in the sample size from 100 to 500.

In the third scenario, the existence of heteroscedasticity and time-dependent covariates problem is taken into account and again, the modified Cox model performed better among the family of survival analysis models. The decision is based on RMSE, MAE, and MAPE. In the final scenario, the outlier, heteroscedasticity, and time-dependent covariates problem is taken, and modified Cox performed better among all existing survival analysis models.

After completing the simulation exercise for different scenarios, we also conducted the empirical application of the suggested modified cox model on a national dataset in Pakistan known as the Labour Force Survey (LFS, 2021). The empirical application shows that education, monthly income, gender, and marital status significantly increase the likelihood

of returning to work after a major injury or disease, while age, region, and government employed significantly decrease the likelihood of returning to work.
Although, in the first scenario, the existence of the outlier and heteroscedasticity, the robust Cox model performed better. But, In the other three scenarios and for the real-data applications, the suggested modified Cox model performed better. The increase in sample size and exponential distribution parameter theta improves the RMSE of all models slightly. However, the outlier quantity and magnitude affect all model performance. The exponential distribution parameter theta increase positively affects all models significantly, meaning that an increase in the exponential distribution parameter theta value improves all models slightly and the model's loss function.
Key words: Survival Analysis; Outlier; Heteroscedasticity; Time-dependent; Schoenfield residuals; Hazard Ratio.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

AFT	Accelerated Failure Time
AIDS	Acquired Immunodeficiency Syndrome
ANN	Artificial Neural Network
CPHM	Cox Proportional Hazard Model
CR	Cox Regression
DA	Duration Analysis
DGP	Data Generating Process
DGP	Data Generating Process
DL	Deep Learning
ECPHM	Extended Cox Proportional Hazard Model
EHA	Event History Analysis
ENC	Elastic Net Cox
FTA	Failure Time Analysis
HCC	Hepatocellular Carcinoma
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio
IC	Interval Censoring
KM	Kaplan Meier
LASSO	Least Absolute Shrinkage and Selection Operator
LC	Lasso Cox
LC	Left Censoring
MCPHM	Modified Cox Proportional Hazard Model
ML	Machine Learning
MLE	Maximum Likelihood Estimation
NPMSE	Non-Parametric Model for Survival Analysis
OC	Oscar Cox
OR	Odd Ratio
OSCAR	Octagonal Shrinkage and Clustering Algorithm for Regression
PH	Proportional Hazard
PMSE	Parametric Model for Survival Analysis
RA	Reliability Analysis
RC	Ridge Cox
RC	Right Censoring
RF	Random Forest
RSF	Random Survival Forest
SA	Survival Analysis
SAT	Scholastic Assessment Test
SPMSE	Semiparametric Model for Survival Analysis
SSVM	Survival Support Vector Machine
TEA	Time-to-event Analysis
WLS	Weighted Least Square

Tab 3. 1 Notation use with Descriptions.

Notations	Descriptions
p	Number of features
N	Number of Instances
X	R^{N*P} feature
X_i	R^{1*P} covariate vector of instance i
y	R^{N*1} vector of observed time, which is equal to min T and C
δ	N*1 binary vector for event status
β	R^{P*1} coefficients vector
f(t)	Event density function
F(t)	The cumulative event probability function
S(t)	Survival Probability function
h(t)	Hazard function
$h_o(t)$	Baseline hazard function
H(t)	Cumulative hazard function
$h_o s(t)$	baseline hazard function in every stratum
β_{ai}	a coefficient vector as the fixed effect of *a*-covariate in *i*-individual
b_i	Coefficient vector of time-dependent covariate at *b* time in *i*-individual.
$xb_i(t_j)$	time-dependent covariate at *b* time in *i*- individual at the time of *tj*
$\sigma^2 I$	The case of homoscedasticity
$\sigma^2 i$	The case of heteroscedasticity

Chapter 1

Introduction

1.1 Background of Study

The term "survival analysis" (SA) is derived from the medical vocabulary. The main objective of survival analysis is to estimate the survival rate of patients until death and the recovery time from any lethal, specific disease. In survival analysis, we further talk about the effect of each predictor on a particular time event, not the event probability. Event or time-to-event data analysis is primarily used in biomedical and engineering-related science. events related to death, brain tumors, and the recovery from any specific disease after diagnosis are examples from public health and some examples related to the engineering field are the survival of a tube light after being made or how many years, months, days, or minutes any electric machine or device will work. For all such kinds of research questions, we use survival analysis. Different health organizations use survival analysis for insurance. If a person falls within specific criteria of the median benchmark, they will be considered for insurance; otherwise, not, etc., for such kind of research question and probability finding analysis of an event, we do survival data analysis. Survival analysis is also used in different fields, such as economics and public policy. In economics, how long does it take to get a job if a person completes the degree? This information helps policymakers to know the rate of being frictionally unemployed, or the duration from completion of a degree to getting a job. Which may guide them in designing employment-related policies. Different name of Survival Analysis are: Reliability analysis which is primarily used in engineering-related fields, In sociology and history we called it event history analysis, in medical field we called

it survival analysis but In economics we called it duration analysis (Rama and Andrews, 2013).

The modelling of survival data analysis for health-related data is an essential area for econometricians because doctors are interested in seeing if a patient has diabetes, breast cancer, HIV/AIDS, lung cancer, COVID-19, or any other lethal disease that causes death. Hence, doctors and patients are interested in whether patients will recover safely or the chance of survival from that specific type of disease (Aalen *et al.,* 2008).

The unbiased estimates, accurate magnitude, high accuracy, consistency, and true findings remain of interest. Further, exploring and identifying the relevancy of regressors is also of great interest. To know the importance of each covariate along with the time-dependent variable in models the earlier model before the Cox regression, used in survival analysis (Kaplan and Meier, 1958), is a graph-related approach that shows the probability of an event at a different time. Further work on survival analysis-related models with innovation and improvement comes from Cox regression. The Cox regression is the most used conventional survival analysis model, invented for time-to-event studies (Cox, 1972).

Survival analysis, or generally, time-to-event models, refers to methods for analyzing the length of time until a well-defined endpoint of interest occurs. Survival analysis is a statistical branch that analyses the average duration of time until a well-defined event occurs. The study of that specific period from origin to the event of interest is called the time-to-event study (Guo, 2010). The time-to-event study is also known as a survival analysis. In other words, from a time-to-event study, some of the events we are taking as events of interest have not occurred yet. Such observation is called censorship and must be accounted for in the analysis to make valid inferences. The statistical methods for appropriately analyzing time-to-event data include nonparametric and semi-

parametric methods, specifically the Kaplan-Meier estimator, log-rank test, and Cox proportional hazards model. These methods are the most used for such data in the medical, social, and natural science fields (Schober *et al.,* 2018).

The event of interest will be clearly defined and well-specified, so there is no confusion about whether it will happen or not. The event can be anything that we are studying or doing research. For example, we are interested in studying COVID-19 patients, and the event of interest for the study is back to work or not; whether an affected person comes back to daily routine or work or study or not, patient of COVID-19 back to work or normal routine or not, this is our event of interest, so that the answer will be in yes/no. The time origin in the previous COVID-19 case will be the time when a person's COVID-19 test becomes positive. From there, the time origin or count should be started when the person knows that they are COVID-19 positive. Time to event will be how long does it take for that event? For example, if a person is COVID-19 positive, how many days or months does an individual return to work? Time-to-event in this COVID-19 example is the days or months the patient took back to work. Censoring is another concept used in survival analysis, if an observation has free of the event, not fully observed and not providing complete information is called censoring. we have total three types of censoring: Right censored, left censored and interval censored.

In public policy, the survival data analysis is used to find the survival and growth of the firm using different covariates such as firm size, experience, age, and other demographic factors, which are combined to see the survival and development of the firm (Solomon *et al.,* 2013). Suppose such data have an outlier, heteroscedasticity, and time-dependent covariate problem. In that case, the conventional survival analysis model, such as the Cox model, failed to estimate an efficient and unbiased model, and the results obtained using traditional survival analysis models from time-to-event data

are spurious because of the outlier, heteroscedasticity, and time-dependent covariate characteristics. Due to these problems in time-to-event data, we need to modify our model, as per the data structure and requirements, such as outlier, heteroscedasticity, and time-dependent covariates, which can handle all the problems simultaneously and give unbiased, efficient estimates and accurate survival time.

Furthermore time-to-event models are also used by different researcher such as (Agampodi *et al.,* 2007 Min *et al.,* 2011; Ediebah *et al.,* 2014) in health sciences, economics, psychology, and sociology. (Efficace *et al.,* 2006) studied graduate engineering students from time until graduation who got admission to engineering universities. A detailed explanation of the use of survival analysis in economics is explained in Section 2.3. The data used for time-to-event models is censored and uncensored. The censored data means that the event of interest has not occurred yet, and the uncensored data means that the event of interest has occurred for the specific event of interest. The dependent variables in the time-to-event study are time variables. It can be either hours, days, months, or years, depending on the nature of the study.

1.2 Motivation of the Study

The traditional Cox regression model often faces limitations such as sensitivity to outliers, inefficiency in handling heteroscedasticity, and potential issues with time-dependent covariates. Despite these advancements in cox model, challenges remain in achieving optimal model performance, particularly concerning standard error reduction, model consistency, and overall predictability. Therefore, there is a need to further refine these models to better meet the demands of survival analysis in various research and applied contexts.

1.3 Research GAP and Problem statement

Distinct time to event models are available to handle only one problem at a time. Still, the main problem in time to event data series is an outlier, heteroscedasticity, and time-dependent covariate in the data. Due to these problems in the data, we can't efficiently estimate the coefficients of the models, as results standard error inflates, which may lead to insignificant parameters, hence model's results are invalid. Therefore, we must modify our models as per the data structure, such as outliers, heteroscedasticity, and time-dependent covariates, which can handle all the problems simultaneously. To meet this gap, we proposed a modified Cox model, which have addressed these problems simultaneously and give unbiased and efficient estimates as compared to the existing time-to-event model in the literature. In previous literature, most authors, such as Robust Cox (Carrasquinha *et al.,* 2018), addressed the problem of an outlier, Weighted Least Square (Mustefa and Chen, 2021) addressed the problem of heteroscedasticity, and time-dependent Cox (Therneau *et al.,* 2017) addressed the problem of time-dependent covariates in the data. These methodologies fail if two or three problems come simultaneously into the data, which is possible in economics and medicine data. (Therneau *et al.,* 2017; Carrasquinha *et al.,* 2018; Mustefa and Chen, 2021). addressed only one problem in the model at a time, such as outlier, heteroscedasticity, or time-dependent covariate. Still, we haven't seen any work in literature that has jointly addressed all three problems such as outlier, heteroscedasticity, and time-dependent covariate problems.

1.4 Research Questions

- Do the existing time-to-event models handle the problem of outliers, heteroscedasticity, and time-dependent covariates in the data simultaneously?

- Whether the proposed modified Cox regression is improved as compared to the existing time-to-event models: Robust Cox, weighted least squares, and Time-Dependent Cox?

1.5 Objectives of the study

The study's main objective is to derive modified Cox model and compare it with the existing time-to-event models using Monte Carlo simulation and then real data Application. The specific objectives of the study are as follows:

There are specifically four objectives of the study, which are:

- To analyze the performance of widely used existing time-to-event models in the presence of outliers, heteroscedasticity, and time-dependent covariates.
- To propose a modified Cox model in case of outliers, heteroscedasticity, and time-dependent covariates.
- To compare the performance of widely used existing time-to-event models in the presence of outliers, heteroscedasticity, and time-dependent covariates.
- To evaluate the effectiveness of the proposed modified Cox regression model in a real-world data application.

1.6 Contribution and Significance of the Study

This thesis has significantly improved the existing literature in the following context. First, this study has analyzed the existing time-to-event models on simulated and real data. Secondly, this study has modified the existing time-to-event models and proposed a modified Cox model for time-to-event data in case of outlier, heteroscedasticity, and time-dependent covariate. Thirdly, this study has identified which model is better for survival analysis, either the conventional Cox Proportional Hazard Model (CPHM) or the Modified Cox Regression Model (MCRM). Lastly, this study has tested the

performance of the suggested model and identified the event of interest, recovering from a major injury or disease in the context of Pakistan using real data from the Labour Force Survey (LFS, 2020–21).

The modified Cox model for survival analysis can be significantly utilized by researchers and practitioners in fields where data is prone to outliers, heteroscedasticity, and time-dependent covariates. This includes medical research for patient survival rates, financial risk assessment, and engineering reliability studies, ensuring more reliable and unbiased estimates.

1.7 Organization of the Thesis

The remaining part of the thesis is organized as follows.

Chapter 2 is based on the literature overview of survival analysis and used in a different field.

Chapter 3 discussed the considered methods for the detection of outliers, heteroscedasticity, and time-dependent covariates, including influence plot for outliers, white test for heteroscedasticity, and Shenfield residual test for time-dependent covariate and different survival analysis models, including Cox regression, Robust Cox, Weighted Least Square, time-dependent Cox and modified Cox regression.

Chapter 4 is about the algorithm of the proposed modified cox regression and derivation of the model.

Chapter 5 illustrates the data-generating process and the result of the simulation experiment for an outlier, heteroscedasticity, and time-dependent covariates. The results of the simulation experiment are subsections according to the different scenarios.

Chapter 6 is based on the real data analysis in this section, each of the real data analyses is linked to the objective of the study and simulation experiment.

Chapter 6 is also based on the in-depth and summarized discussion of real data analysis and simulation experiments. This section also discusses the research's limits and future directions.

Chapter 2

Literature Review

2.1. Introduction

Reviewing previous studies is one of the initial steps for understanding, evaluating, and solving a research problem. It provides theoretical and empirical background and efficient knowledge to understand the depth and importance. Previous studies on survival analysis and all the existing models used in survival analysis have been reviewed in subsequent sections.

The literature review chapter includes detailed literature based on time-to-event models, whether it's related to the use of survival analysis in economics, studies related to epidemiology and biomedical sciences, or studies based on COVID-19 and HIV.AIDS, relevant studies based in Pakistan, and other relevant literature related to outliers, heteroscedasticity, and time-dependent covariates. We divided the literature review chapter into two different sections. The first section includes detailed literature on the use of survival analysis. The second section includes summary of literature review and the last section includes the conclusion of the literature.

2.2 Literature on Survival Analysis (SA)

Survival Analysis (SA) is a branch of statistics used to analyze the expected duration until that specific event occurs and further focus on time-to-event data and their analysis.

2.3 Studies related to Economics using Survival Analysis

In economics, survival analysis is known as duration analysis. The term survival analysis was early used in biomedical research. The advantage of survival analysis using economics and finance-related subjects is that it handles the censored observation

in the data. Gémar *et al.,* (2016) studied the Survival rate of the Spanish country hotel industry and the important covariates that affect the survival rate of a hotel in Spanish. The study results indicate that location is the most important covariate for the survival of a hotel. The survival rate will be higher near the airport and other picnic spots. (Burton *et al.,* 2003) studied Modelling the adoption of organic horticultural technology in the UK using survival analysis. The author collected the cross-sectional primary survey data from 237 farmers in the UK, including 151 non-organic farmers and 86 organic farmers. The authors use the (Kaplan and Meier, 1958; Lanczky and Gyorffy 2021) and (Cox, 1972) models. The result concluded that farm size, education, household size, and agricultural sources of finance are significant determinants for adopting advanced technology.

Using survival analysis, (Min *et al.,* 2011) studied graduate engineering students' success. The author studied the students who got admission to engineering universities. What is the chance of success, whether students would survive or not? The author estimated (Kaplan and Meier, 1958) life tables and concluded that female is highly riskier than male. The author further concluded that students whose Scholastic Assessment Test (Ameri *et al., 2016*) score are between 500 and 600 is likely to be more secure or survive as compared to students who score between 300 and 400 on the SAT score. Survival analysis is vast and can be used in engineering and social sciences.

2.4 Studies related to health and epidemiology using Survival Analysis

Salerno and Li (2023) studied time to event data taken from Boston Lung caner survival cohort study, the study proposed novel approaches for feature selection in case of high dimensional data, where number of covariates are higher then number of observation, the study suggested new machine learning model for feature selection in high dimensional data in the context of survival analysis (Emmerson and Brown, 2021).

Ediebah *et al.,* (2014) estimated lung cancer disease using Kaplan and Meier and Cox regression; the number of patients was 391, and the country for this study was Belgian, (Agampodi *et al.,* 2007) studied Breastfeeding mothers in Sri Lanka in 2006, the sample of the study was 219 mother who was breastfeeding to an infant, the statistical model was Kaplan and Meier and CPHM, the study concluded that the average time of infant breastfeeding a Sri Lanka is four months, the feeding to the infant was high in families, where the parental education is low (Pawar et al., 2022).

Efficace *et al.,* (2006) Studied whether either patient can predict survival rate in advance using a different indicator. This study was conducted in Boston, the USA, and the study sample was 299 patients. The author estimated Cox regression and concluded that white blood cells, alkaline phosphatase, and the patient's scale on the social functioning scale could positively affect the survival rate of patients.

Survival analysis is further used in various fields, such as COVID-19, HIV/AIDS, and Breast cancer. (Altonen *et al.,* 2020) discuss the characteristics, comorbidities and survival analysis of young adults hospitalized with COVID-19 in the capital city of New York. The total number of young adults who studied was 395. The study concluded that 57% of patients had at least one major comorbidity[1]. The author used the Kaplan and Meier model, and the study further reveals that COVID-19 infects adults but not more than age. (Panjer, 1987) estimated the survival rate for 543 patients at different stages. If it is the initial stage, the average life expectancy is 9.6 years. Suppose the second stage is 7.3 years. If it is the third stage, then 6.2 years. If the fourth stage is 4.3 years, and if the last stage is 0.93 years. (Gyorffy *et al.,* 2010) studied the survival

[1] At one time more than one disease is called comorbidities, i.e a person have COVID-19 and diabetes.

rate of breast cancer patients at 1,809. The author estimated Kaplan and Meier's graphical analysis and Cox Proportional Hazard Model. The result suggests that breast cancer females' average survival rate is 6.43 years.

2.4.1 Studies related to COVID-19 using Survival Analysis

Ershadi *et al.,* (2023) studied the COVID-19 affected patients using Kaplan Meier and Cox regression to find the predictors of survival patient with COVID-19. The COVID-19 pandemic is a challenge for every field of expertise worldwide, statisticians and econometricians, to predict the correct survival rate if a patient is affected by COVID-19. Mathematics, Statistics, and Econometrics have contributed to their expertise in this regard (Salinas *et al.,* 2020) estimated the survival rate of COVID-19 affected patients in Mexico, the time origin of the study is all the confirmed cases, the author estimated Kaplan Meier curves and Cox model, concluded results that the death rate for men is higher. Also, the total number of survival individuals for older patients was 58%. The death rate from COVID-19 remains at 42%.

Altonen *et al.,* (2020) discuss the adults affected by COVID-19 in the capital city of New York. The total number of young adults studied is 395, and 57% of patients had at least one major comorbidity. The author used the Kaplan and Meier model. The study further reveals that COVID-19 affects adults relatively more than teenage children.

2.4.2 Studies related to HIV/AIDS using Survival Analysis

Lelisho *et al.,* (2023) studied Human Immune deficiency virus (HIV) and tuberculosis (TB) Cox regression and Kaplan Meier plot were used to find the survival time of infected patients, that older age are significantly associated shorter survival time to death of TB/HIV. Human Immunodeficiency Virus (Mustefa and Chen, 2021) and Acquired Immunodeficiency Syndrome (Panjer, 1987) is one of the lethal diseases worldwide, so doctors and patients are interested in the different stages of HIV what is

the survival rate of the patients at this stage, the Modelling of this kind of patient and health-related study is essential, (Panjer, 1987) estimated survival rate for 543 patients at a different stage, if it is initial stage than average life expectancy is 9.6 years, if the second stage than 7.3 years, if it is the third stage then 6.2 years if the fourth stage than 4.3 years and if last stage than 0.93 years. (Matida *et al.,* 2007) Studying HIV AIDS, which transfers from mother to child at birth, the study was conducted from 1983 to 2002 in Brazil, and the total number of cases selected for the study was 914. They concluded that the average survival before 1988 was 20 months, which was increased to 24 months if cases were diagnosed between 1900 and 1992, and a further increase in average survival if cases were diagnosed between 1993 and 1994, and so on. The male survival rate is higher than that of males, with a 51.2% chance and a 49.8% chance of females (Pearce et al., 2022).

2.4.3 Studies related to Breast Cancer using Survival Analysis

Breast cancer is another type of cancer and a lethal disease worldwide (Konishi *et al.,* 2023) studied risk factors associated with breast cancer in Japan using Cox proportional hazard model, the study concluded that these factors such as smoking, young age, advanced cancer stage and obesity are the risk factors in breast cancer disease. (Gyorffy *et al.,* 2010) also studied the survival rate of breast cancer patients by 1,809. The author estimated Kaplan and Meier's graphical analysis and Cox Proportional Hazard Model. The result suggests that the average survival rate of breast cancer females is 6.43 years. (Prentice and Gloeckler, 1978) also estimated the survival rate of breast cancer patients in the USA using the Cox regression model: 62% survival rate in white women and 47% survival rate in black women. (Lauss *et al.,* 2008; Gyorffy and Schafer, 2009) also studied the breast cancer of 1,079 and 379 patients using survival analysis models, such

as Kaplan and Meier, and Left one out cross-validation (LOOCV), (Lauss *et al.,* 2008) further concluded that tumor size has significantly improved survival rate.

2.5 Studies related to engineering using Survival Analysis

Event or time-to-event data analysis is mainly used in biomedical and engineering-related science, the events related to death, brain tumors, and the recovery from any specific disease after diagnosis. Some examples related to the engineering field are tube light's survival after making or how many years, months, days, or minutes any electric machine or device will work. For such a kind of analysis, we use survival analysis. Electronic devices, whether mobile phones, laptops, or any other electronic device, are essential to our daily lives. Modelling electronic device survival rates or reliability analysis is an exciting and challenging part of daily life. Ali *et al.,* (2020) further discuss the reliability of different voltage electronic devices, i.e., 80V, 100V, and 120V, using the different methodology of lognormal distribution and Weibull distribution and modified generalized exponential, the results conclude that modified generalized exponential outperforms other distributions by looking at the value of AIC and BIC.

(Krivtsov, 2023) evaluated the reliability modeling for survival analysis including engineering, medical and economics and suggest that most of the sophisticated model for the class fo reliability are referred to as deep survival model to find the complex relations between the time and the covariates.

2.6 Literature Review According to Pakistan

Ahmed *et al.,* (2017) studied survival analysis of heart failure patients who were admitted at Institute of Cardiology hospital Faisalabad Pakistan from April to December twenty fifteen, the cox regression and Kaplan meier plot were used to find the pattern of survival of heart patient, the renal dysfunction, ejection fraction, blood

pressure and age were considered as significant risk factors for mortality among heart failure patients.

Chaudhry *et al.,* (2018) studied the survival rate of dengue patient, the data taken from Lahore hospital for year 2013 to 2016, A sample size of patients are 708 is included in the study, Kaplan meier, (1958) and (Cox, 1972) model were used in the study, the study concluded that estimated time for male and female are same, whereas it is different among children, adults and elderly, (Hanif *et al.,* 2015) also studied dengue patients and severity to liver dysfunction in Pakistan, 60 patients data is taken from Lahore hospital, Kaplan Meier plot curve were used to find the different survival time, study concluded that the survival rate for patients with mild and moderate liver dysfunction were found compared to severe type of liver dysnfunction.

Patel *et al.,* (2019) studied the infant mortality rate and the survival of infant in Pakistan, data is taken from demographic health survey, cox model were used to account the potential risk factor and covariates, the results concluded that women having low education or poor economic condition are having higher infant death rates compare to high education and good economic condition.

Yusuf *et al.,* (2007) studied Hepatocellular Carcinoma (HCC) cancer disease, the most common type of primary liver cancer. The author mentioned that HCC is the 5th most common cancer disease worldwide, affecting more than one million individuals annually. The author estimated (Kaplan and Meier, 1958) model for the average survival time of the affected patients. The author further concluded that patients' average or median survival time is 10.5 months from the lethal hepatocellular carcinoma cancer disease. (Akram *et al.,* 2007) worked on cancer disease patients and estimated the median survival time of patients using parametric and nonparametric approaches. The data are taken from Nishtar Hospital Multan, one of the big city

hospitals in Pakistan. The author used the Kaplan and Meier model and concluded that the male gender has a different survival rate than the female gender, concluding that the female survivor rate is higher than male cancer patients.

Table 2. 1 Summary of Literature Related to Real-world Applications

Studies related to Economics using Survival Analysis				
Event Origin	**Event of Interest**	**Estimation of Time Scale, i.e. (Years, Months, or Days)**	**SA Model**	**Features/Covariates and References**
New Hotel	Survive or not	Likelihood of survival time of the hotel.	Cox Model	Location and Picnic Spot (Gémar *et al.,* 2016)
Completion of economics degree	Getting a Job or not	The likelihood of an economics graduate getting a job within time months	Cox Model	People: gender, age, education, experience, and city. (Kiefer, 1988)
Student Enrolled	Student Dropout or not	Likelihood of student dropout days.	Cox Model Time-Dependent (TD)	family background, financial status, last exam score, and previous college information (Ameri *et al.,* 2016)
Studies related to health using Survival Analysis				
Event Origin	Event of Interest	Time Scale, i.e. (Years, Months, or Days)	SA Model	Features/Covariates and References
Baby birth to mother	Stop breastfeeding an infant or not	Likelihood of babe's breastfeeding months.	Kaplan and Meier, and Cox regression	Parental education, location, and Muslim ethnicity are used (Agampodi *et al.,* 2007).
Lung Cancer Diagnose	Survive or not	Likelihood of patient's survival months.	Kaplan and Meier, and Cox regression	Age, gender, and Physical health (Ediebah *et al.,* 2014)
Studies related to COVID-19 using Survival Analysis.				
Event Origin	Event of Interest	Time Scale, i.e. (Years, Months, or Days)	SA Model	Features/Covariates and References
Covid19 Positive Diagnose	Survive or not	Likelihood of patient's survival months.	Kaplan and Meier, and Cox regression	Age, gender, and Physical health (Ediebah *et al.,* 2014)
Covid19 Positive Diagnose	Survive Back to work or not	Likelihood of patient's survival months.	Kaplan and Meier, and Cox regression	Age, gender (Kundu *et al.,* 2021)
Studies related to HIV/AIDS using Survival Analysis.				
Event Origin	Event of Interest	Time Scale, i.e. (Years, Months, or Days)	SA Model	Features/Covariates and References

HIV/AIDS Diagnosis	Survive or not	Likelihood of patient's average survival months.	Cox regression	Age, gender, and physical health status (Panjer, 1987)
HIV/AIDS Diagnosis	Survive or not	Likelihood of patient's average survival months.	Kaplan and Meier, and Cox regression	Age, gender, and Physical health (Matida *et al.,* 2007)
Studies related to Breast Cancer using Survival Analysis				
Event Origin	Event of Interest	Time Scale, i.e. (Years, Months, Days, or Hours)	SA Model	Features/Covariates
Breast Cancer Diagnose	Survive or not	Likelihood of patient's average survival months.	Kaplan Meier	Age, gender, and Physical health (Györffy *et al.,* 2010)
Breast Cancer Diagnose	Survive or not	Likelihood of patient's average survival months.	Kaplan Meier and LOOCV	Age, gender, and Physical health (Lauss *et al.,* 2008; Györffy and Schäfer, 2009)
Studies related to engineering using Survival Analysis				
Event Origin	Event of Interest	Time Scale, i.e. (Years, Months, or Days)	SA Model	Features/Covariates and References
The invention of the electronic device	Defective or not	Likelihood of Electronic device average month uses till defectively.	Generalized Exponential Functional	Different combinations of voltage (Ali *et al.,* 2020).
Invention of Laptop	Defective or not	Likelihood of Laptop device average month use till defectively.	Exponential Distribution	Different combinations of voltage (Mendez-Gonzalez *et al.,* 2016).
Different Studies on Survival Analysis According to Pakistan				
Event Origin	Event of Interest	Time Scale, i.e. (Years, Months)	SA Model	Features/Covariates and References
Hepatocellular Carcinoma (HCC) cancer diagnose	Survive or not	Likelihood of patient's survival months.	Kaplan and Meier	Age, gender, and Physical health (Yusuf *et al.,* 2007).
Cancer Diagnose	Survive or not	Likelihood of patient's survival months.	Weibull function	Age, gender, and Physical health (Akram *et al.,* 2007)

2.7 Summary and Literature Gap

The conventional survival Cox model has certain assumptions about the model covariates, such that there will be no outliers, no heteroscedasticity, and no time-dependent covariates in the model, which can be possible simultaneously at one moment. The literature on survival analysis suggests that different estimation techniques have various advantages regarding their strength and ability to capture the effects of different covariates. But all of these cannot capture everything simultaneously. Therefore, there is a need to make an improved modified model for survival analysis by taking advantage of their specific abilities. Since Cox regression was introduced in 1972, it has improved over time. The improved version of the Cox model is Robust Cox by (Carrasquinha *et al.,* 2018), Weighted Least Square by (Mustefa and Chen, 2021), and time-dependent Cox by (Therneau *et al.,* 2017). These methodologies fail if two or three problems come simultaneously in the data, which is possible in economics and medicine. There is also a possibility of masking[2] and swamping[3] effect in our model because of the same characteristic presence of another adjacent one, making the result biased, inconsistent, and inefficient. That's why there is a need for a modified Cox model, which can solve these problems simultaneously in the model so that we can get unbiased, consistent, and efficient estimates in the presence of outlier, heteroscedasticity, and time-dependent covariate in the data.

[2] Masking effect means if outlier is not detected due to presence of others large outlier.
Or in other words Outlier are not absolute larger outliers mask smaller one, if larger one is removed others evolve.
[3] Swamping effects means if good observation is considered as outlier due to presence of other clean data sets.

As discussed in detail in the literature, survival analysis is used in many science fields, such as medicine, economics, engineering, and social sciences. So, modification in such areas will be excellent work, and all the fields mentioned above will benefit.

Chapter 3

Review of related Econometric Methodology

3.1 Introduction

The methodology section has two stages; the first stage is the detection stage, whether the data have an outlier problem, heteroscedasticity, and time-dependent covariate. In the next stage, we have addressed the three issues jointly in the proposed modified Cox regression. To detect an outlier, heteroscedasticity, and time-dependent covariates, we have used an influence plot by (Riu and Bro, 2003) for outlier detection. For heteroscedasticity detection, we used the Breusch Pagan test (Breusch and Pagan, 1979; White, 1980); for the time-dependent covariate detection, we have used the Schoenfeld residuals test (Fisher and Lin, 1999). We have estimated five methodologies, such as Cox regression, Robust Cox, Weighted Least Square, and Time-Dependent Cox, and proposed modified Cox regression. We compared them for different scenarios in the presence of an outlier, heteroscedasticity, and time-dependent covariates. The details of these methodologies are given below.

3.2 Survival Analysis Different Approaches

The survival analysis approach is mainly divided into three different groups: parametric models, non-parametric models, and the semiparametric model.

The main conventional survival analysis model is cox model, the shortcomings of Cox regression are that it fails to estimate when we have high dimensional data, and it's unable to estimate when we have a non-linear function with a different covariate. Also, it's powerless to give unbiased results in the case of heteroscedasticity and multicollinearity in the data, which are often present in the data out of the box because of the cross-sectional behavior of different individuals. Also, if the primary assumption

of proportionality is violated, then the Cox proportional hazard model fails to give unbiased estimates and efficient standard error (Moncada *et al.,* 2021).

3.2.1 Parametric Model for Survival Analysis (PMSA)

This approach uses distributions such as Weibull, exponential, and log distribution. Maximum likelihood estimation is the method to estimate such distribution (Efficace *et al.,* 2006).

3.2.2 Non-parametric Model for Survival Analysis (NPMSA)

This approach does not assume the following underlying distribution: Weibull, exponential, and log distribution. The method to estimate such non-parametric models is (Kaplan and Meier, 1958). In this method, we don't make any underlying assumption and plot survival probability as a function of time.

Kaplan Meier function can be written in equation.

$S(t)=Prob(\tau > t)$, Where $\tau \geq 0$ is a random variable, which is the event of interest. The goal is to estimate the survival function, where t is the time, t=0,1,2,3 ….

3.2.3 Semiparametric Model for Survival Analysis (SPMSA)

In this approach, we use different distributions such as the name also suggested semiparametric, so this group of models makes very few assumptions compared to parametric and nonparametric. We do not assume the shape of the distribution, and the most popular model of survival analysis belongs to this category, namely Cox regression (Cox, 1972). The three methods mentioned are also used in literature. Still, many more methods are used for survival analysis, such as deep learning (DL), Artificial Neural Network (ANN), and machine learning (Efficace *et al.,* 2006) based method. At the start of survival analysis, the Cox regression or time-to-event analysis was made for the medical treatment of different disease patients to know the average survival time in different clinical trials. For example, if patients have cancer or any

other lethal disease and by doing survival analysis, they try to find the survival time of that patient by giving the specific type of treatment to the patient. That was the primary theme behind the survival analysis. Later on it is being used in social science and engineering for different electronic devices' accuracy.

3.2.3.1 LASSO Cox

Tibshirani (1997) proposed a new shrinkage method for covariate selection in survival analysis named Least Absolute Shrinkage and Selection Operator (LASSO) Cox for variable selection. Lung cancer patients are the targeted sample. The study concludes that the LASSO Cox model is better than the Cox model and Stepwise Regression.

Shahraki *et al.,* (2015) estimated the Least Absolute Shrinkage and Selection Operator (LASSO) Cox as another alternative to Cox regression. The Cox model failed to give a valid estimate if there was a small sample size. The author estimated LASSO-Cox and compared results with conventional Cox regression and concluded that LASSO-Cox performed better in case of low sample size.

3.2.3.2 Ridge Cox

Horel (1962) proposed a new technique, ridge regression, for the first time. If there is collinearity in covariates or the number of covariates is greater than the number of observations in the model, then in such a situation, literature suggest to use ridge regression through cross-validation. The ridge regression was further modified by (Hoerl and Kennard, 1970), which was additionally used by (Verweij and Van, 1994; Wang *et al.,* 2019) as an alternative to Cox regression. Its main strength is to incorporate collinear covariates and shrink their value towards each other.

3.2.3.3 Elastic Net (EN) Cox

Elastic Net (EN) Cox is another alternative for Cox regression if there is high dimensionality and collinear covariates (Vinzamuri and Reddy, 2013) discussed the survival rate of patients who are readmitted to a hospital within 30 days after leaving the hospital bed, the author considered the elastic net Cox model as a better alternative if there are covariates features and high dimensionality in data, in such situation Cox regression failed, the number of patients is 8,000. The study concluded that the proposed Elastic Net (EN) algorithm performs better than Cox regression.

Simon *et al.,* (2011) considered the survival rate of patients using the Elastic net and Cox proportional hazard model on real data and survival data and compared results, so the study concludes that the elastic net performs better than Cox regression.

3.2.3.4 OSCAR Cox

Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) Cox proposed for the first time for high dimensional data in linear regression by (Petry and Tutz, 2011), then later on, it was used for survival analysis as OSCAR Cox by (Vinzamuri and Reddy, 2013; Wang *et al.,* 2019) to find the covariates for the event of interest, the main advantage of OSCAR Cox is to estimate the likelihood of event time when there are highly covariates features in the censored data.

3.3 Machine Learning Approach to Cox Regression

Moncada *et al.,* (2021) explained machine learning models such as Survival Support Vector Machines (SSVM), Random Survival Forests (RSF), and Extreme Gradient Boosting (XGB) as an alternative for Cox regression if there is a problem of high dimensionality in data, the author compares all model and concludes that machine learning model outperforms then conventional Cox regression.

3.3.1 Survival Support Vector Machine (SSVM)

The Survival Support Vector Machine (SSVM) is a machine learning technique that is used for survival analysis as an alternative to conventional Cox regression by many authors such as (Jung *et al.,* 2019; Roshanaei *et al.,* 2020) for the disease of breast cancer and colon cancer patients survival rate is very low. The earlier model used for high dimensional data and nonlinear relations with different covariates is the Support Vector Machine (SVM). Still, the model that is used in censored data is called SSVM. Widodo and Yang (2011) studied the survival rate of the machine, using the SSVM model and Proportional Hazard model, and compared the results with each and concluded that the SSVM model performs better than the conventional Cox proportional Hazard model using censored data.

3.3.2 Random Survival Forest (RSF)

Jung *et al.,* (2019) Studied the breast cancer patients of non-Hispanic and white women, the author concluded that alcohol and obesity are hazardous for breast cancer. The study also suggests that decreasing body weight can reduce the risk of breast cancer. (Roshanaei *et al.,* 2020) also studied Colon cancer as one of the most lethal diseases. (Greten *et al.,* 2005) took the data from 317 cancer patients. The study aimed to estimate the patient survival rate to death from the time of cancer diagnosis, using the Random Survival Forest (RSF) model to see the effect of different covariates on the survival rate of patients. The median survival rate of patients was calculated at 53 months. The most important covariates are White Blood Cell (WBC) count, disease stage, and metastasis to other organs.

3.3.3 Extreme Gradient Boosting (XGB)

This machine learning technique was addressed for the first time by (Chen *et al.,* 2013), who estimated the survival rate in the country USA for Breast cancer gene data of 1981

patients using the Gradient Boosting Algorithm and Cox Proportional Hazard Model. A study suggests that gradient boosting performs better than Cox regression. (Nemati *et al.,* 2020) also used the gradient-boosting model on COVID-19 patients. The sample of COVID-19 patients was 1182, and results indicate that gradient boosting performs better than other conventional survival analysis models. (Mayr *et al.,* 2017) also work on gradient boosting model the survival rate of a specific event (Nemati *et al.,* 2020) considered that gradient boosting is a very flexible algorithm, also performs better in case of gene high dimensional data and gives the smallest residual sum of the square as compared to an existing model of survival analysis.

3.4 Stratified Cox (SC)

Stratified Cox is another alternative survival analysis model proposed for the first time (Ata and Sozer, 2007). The stratified Cox regression is a modified form of Cox regression. If there is a violation of the proportional hazard assumption by different covariates in such a situation, stratified Cox performs better than the existing conventional Cox model. The stratified Cox model contains coefficients that do not vary over the strata, and SC Cox is further used by (Wang *et al.,* 2019).

3.5 Extended Cox Regression

After too many modifications in the existing survival analysis model, such as Cox regression, after the violation of the different assumptions of Cox regression, there is an additional alternative model, so if there is any violation of proportionality of assumption (Husain *et al.,* 2018) proposed an alternative version of Cox regression model namely the Extended Cox Proportional Hazard Model (ECPHM), if the primary assumption of Cox regression is violated such as proportionality then Cox regression is no more valid than in such a situation ECPHM model perform better as compare to Cox regression. The Cox model was further criticized by different authors, namely (Anjullo,

2018; Arsyad *et al.,* 2019; Muhammad and Yuwaningsih, 2019; Emoru *et al.,* 2020; Baik *et al.*, 2021) considered the ECPHM model as a better alternative to Cox regression when the constant proportionality assumption is violated.

3.6 Stratified Extended (SE) Cox Regression

The Stratified Extended (SE) Cox is another alternative for Cox regression by joining the stratification term and extended-term (Ratnaningsih *et al.,* 2019) proposed a new model: Stratified Extended Cox, which covers jointly the problem of non-proportionality, time-independent and time-dependent covariates. Further, SE Cox was used by many authors in survival analysis (Ratnaningsih *et al.,* 2020; Ratnaningsih *et al.,* 2021). All papers on Extended Cox regression performed better than conventional Cox regression.

3.7 Why can't We Use OLS instead of Cox Regression?

We can't use OLS instead of Cox regression because of censored observation in the data. OLS aims to minimize the residual sum of squares (RSS). In censored data, the error term is unknown; therefore, we cannot minimize the RSS. The censored observation makes the OLS model biased. That's why we shift to censored regression or survival analysis models such as Cox regression, Extended Cox regression, Survival Support Vector Machine (SSVM), and Random Survival Forest (RSF), depending on the nature of data, assumption, and other properties.

3.8 Criticism on Cox Regression

The first model used in survival analysis was suggested by (Kaplan and Meier, 1958) and later (Cox, 1972) developed Cox regression for time-to-event analysis. Cox regression received major criticism. If the Cox regression assumption is violated, then Cox regression fails to give unbiased estimates. The major criticism of Cox regression includes, for instance, not giving the correct standard error if heteroscedasticity or

multicollinearity problems in data. Secondly, it fails to estimate if there is a problem of high dimensionality in the data, as genes and medical-related data are almost highly dimensional[4] . Third, if there is a small sample size, it's unable to give consistent and reliable results if the proportionality assumption is violated.

Mustefa and Chen (2021) criticize the (Cox, 1972) failure to give the correct standard error of the heteroscedasticity problem in the data. Further, the author suggested a new model in the case of heteroscedasticity that Weighted Least Square (WLS) outperforms the existing Cox regression by comparing Root Mean Square Error (RMSE) and Akaike Information Criteria (AIC). (Moncada-Torres *et al.,* 2021) explained machine learning models such as Survival Support Vector Machines (SSVM), Random Survival Forests (RSF), and Extreme Gradient Boosting (XGB) as an alternative for Cox regression if there is a problem of high dimensionality in data, the author compares all model and concludes that machine learning model outperforms then conventional Cox regression. Shahraki *et al.,* (2015) estimated the Least Absolute Shrinkage and Selection Operator (LASSO) Cox as another alternative to Cox regression. The Cox model failed to give a valid estimate if there was a small sample size. The author estimated LASSO-Cox and compared results with conventional Cox regression and concluded that LASSO-Cox performed better in case of low sample size. (Husain *et al.,* 2018) proposed an alternative Cox regression model, the Extended Cox Proportional Hazard Model (ECPHM). Suppose the primary assumption of Cox regression is violated, such as proportionality. In that case, Cox regression is no more valid than the ECPHM model,

[4] High Dimensional is a data type, In which the number of covariates is higher than the number of observation (Collyer *et al.*, 2015)

which performs better than Cox regression in such a situation. The Cox model was further criticized by different authors, namely (Anjullo, 2018; Emoru *et al.,* 2020; Baik *et al.,* 2021) (Arsyad *et al.,* 2019; Muhammad and Yuwaningsih, 2019) considered the ECPCHM model as a better alternative to Cox regression, when the constant proportionality assumption is violated.

3.9 Robust Cox as Alternative to Cox Regression

Cox proportional hazard regression is a commonly used statistical technique to investigate the relationship between covariates and survival outcomes. However, in real-world applications, the assumption of proportional hazards may not hold, and the Cox model may produce biased or inaccurate results. A robust model has been developed for Cox survival analysis to validate the outlier issue in the model. One approach is to use robust standard errors and weighted least squares, allowing for model outlier and heteroscedasticity. This method can improve the accuracy of the estimated hazard ratios and confidence intervals when the assumption of outlier is violated. Another approach is to use the weighted Cox model, which assigns higher weights to observations with more reliable survival times. This method can reduce the impact of outliers and improve estimation efficiency (Farcomeni and Viviani, 2011).

Several studies have compared the performance of robust Cox models with the standard Cox model under various scenarios, such as small sample sizes, skewed distributions, and non-proportional hazards. The results showed that the robust methods could provide more accurate and reliable estimates, especially in the presence of outliers or non-proportional hazards (Bednarski, 1993; Farcomeni and Viviani, 2011; Xie and Zheng, 2016). In summary, robust Cox survival analysis methods can improve the accuracy and reliability of survival analysis results, especially when the assumption of

proportional hazards is not met. Researchers should consider the potential impact of violating this assumption and use robust methods as an alternative.

3.10 Weighted Least Square as Alternative to Cox Regression

Pang et al., (2015) used Weighted Least Square (WLS). They found a better alternative to Cox regression because the conventional Cox regression assumes that the covariates are Independent. Identically distributed (IID) means no heteroscedasticity, but it is present in the data, so Cox regression failed to estimate the coefficient with minimum standard error. The author found weighted least squares as a good alternative for survival analysis. (Yu *et al.,* 2019) first proposed the Novel Quasi-Likelihood Ratio test for the homoscedasticity assumption in the survival analysis. The author said that if there is a violation of the homoscedasticity assumption in the survival analysis, then the model leads to an invalid conclusion, so to test the assumption of homoscedasticity in the time-to-event analysis is necessary and proposed a new test of quasi-likelihood ratio test if there is no heteroscedasticity problem than used conventional Cox regression if there is heteroscedasticity in the error term than used WLS by (Mustefa and Chen, 2021). (Pang *et al.,* 2015; Yu *et al.,* 2019; Mustefa and Chen, 2021) Considered WLS as a better alternative for Cox regression in the case of heteroscedasticity, Weighted Least Square (WLS) outperforms the existing Cox regression by comparing Root Mean Square Error (RMSE), Akaike Information Criteria (AIC) and using simulation studies.

3.11 Time-Dependent Cox as Alternative to Cox Regression

The covariate that changes values with time is known as the time-dependent covariate. (Therneau *et al.,* 2017) Used an alternative model known as time-dependent Cox to handle such covariates in the model. (Fisher and Lin, 1999) also studied the time-dependent covariate in the model, which can be held very carefully in the model. Otherwise, time-dependent covariates lead to biased results. The time-dependent Cox

model is a better alternative for time-dependent covariates in such cases. (Zhang *et al.,* 2018) discussed the time-varying covariates in the study. The author suggested testing the proportional hazard model assumption using model residuals. Results of the study show that the time-dependent Cox model performed better in the case of the time-varying coefficient than Cox regression. So, it is finally concluded from the literature that uses time-dependent Cox if there are time-varying coefficient covariates in the model compared to Cox regression.

3.12 Consequences of outlier, heteroscedasticity, and time-dependent covariates

Outliers in regression analysis can distort parameter estimates, leading to biased results. Heteroscedasticity, where the variance of errors varies across levels of predictors, violates model assumptions and can result in inefficient parameter estimates and biased standard errors. Time-dependent covariates can introduce serial correlation, impacting the independence assumption in regression models and potentially leading to biased coefficient estimates. These issues collectively jeopardize the model's predictive accuracy and the validity of statistical inferences. To address outliers in the data (Farcomeni and Viviani, 2011) consider robust Cox. for heteroscedasticity (Pang *et al.,* 2015; Yu *et al.,* 2019; Mustefa and Chen, 2021) considered weighted least squares or transformations can be applied. For Time-dependent covariates (Therneau *et al.,* 2017; Zhang *et al.,* 2018) require specialized time-series methods to appropriately model temporal dependencies and maintain the integrity of regression analyses. The existence of outliers, heteroscedasticity and time dependency may have swamping and masking effect therefore, Proactive identification and mitigation of these issues are essential for robust and reliable statistical modeling.

3.13 Detection of outlier, heteroscedasticity, and time-dependent covariates

We have used the methodology of (Riu and Bro, 2003; Adil, 2015; Therneau *et al.,* 2017; Urooj and Asghar, 2017) to detect outliers in the data. To detect heteroscedasticity in the error term, we have used the (Breusch and Pagan, 1979) and (White, 1980) tests. For time-dependent covariate detection, we have used the methodology (Therneau *et al.,* 2017). So, to solve these problems simultaneously, we have proposed a modified Cox regression, which is discussed in detail in this chapter.

For outlier detection, we have used graphical methods such as Influence Plot used by (Riu and Bro, 2003). In this method we plot the residuals of the fitted model, and get the outliers in the graphical form, if the value of influence plot greater than 3SD, the value is considered to be outlier by influence plot and the iterative procedure (Urooj and Asghar, 2017). The iterative outlier procedure follows the likelihood ratio criteria, which are further based on outer and inner loops, which can be used to detect outliers using standard test statistics. The iterative procedure (Urooj and Asghar, 2017) suggested a better alternative for detecting the outliers in the data, but our work focuses more on the solution of the outlier, not the detection.

For heteroscedasticity detection, we used (Breusch and Pagan, 1979) and (White, 1980). The popular tests for heteroscedasticity detection methodology are as follows:

Suppose we have a model below

$$Y_i = a + bx_i + cz_i + \mu_i \tag{3.1}$$

Estimate model 3.1 and find $\widehat{\mu_i}$

Now estimate an auxiliary regression in which $\widehat{\mu_i}^2$ is regressed on the independent variable of the model, their squares, and cross terms.

$$\widehat{\mu_i}^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i + \gamma_4 {x_i}^2 + \gamma_4 {z_i}^2 + \gamma_5 x_i z_i + e_i \tag{3.2}$$

After estimating the auxiliary regression, the next step is to test the hypothesis.

Test below hypothesis

$H_o; \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0$ Or $\gamma_i = 0$, i=1,2, …, 5

$H_A; Atleast\ one\ of\ \gamma_i\ is\ non\ zero$

We know that for a single restriction, we use t statistics, but here, we have multiple restrictions, so we have used F statistics to check the null hypothesis. If the F calculated value exceeds the F tabulated value, we have rejected H_0 and conclude that the data has a heteroscedasticity problem. In other cases, If all $\gamma_i = 0$ then $\hat{\mu}_i^2$ will become constant then it will become a homoscedasticity case. If any of the γ_i Become non-zero, and it will be a heteroscedastic case. Time-dependent covariates occur when the independent variable is time-varying, which changes over a given period. The different examples of time-varying covariates are smoking status and C reactive protein (CRP) in the blood, which can be measured at different times and can change from time to time. The path of the covariates, either internal or external, changing in path is also a time-varying factor. In such cases, the time-dependent Cox model is preferred. But we are looking for three different problems simultaneously, such as outlier, heteroscedasticity, and time-dependent covariates, so the Cramer von Mises test and ZPH test have been used for testing time-dependent covariates of the independent variables, which was used earlier by (Scheike and Zhang, 2011; Therneau *et al.,* 2017; Zhang *et al.,* 2018). In the next section, we are explaining Cox regression with step-by-step equation and it's a comparison with the logit and tobit model, that how Cox regression is different with these methodologies.

3.14 Cox Regression Methodology

A Cox regression model is a statistical model used for censored data in survival analysis or time-to-event analysis. Cox regression is a handier technique than logistic regression (LR) as the former incorporates more information about survival and censored data (Annesi *et al.,* 1989). Furthermore, Cox regression is also known as the Cox Proportional Hazard Model (CPHM) because the primary assumption of Cox regression is proportionality, which means that the two individuals or two patients will be independent. Their ratio over time will be constant and proportional. Some other assumptions for using this model are that there will be no multicollinearity among the different covariates, no heteroscedasticity, and the last assumption is that there will be no interaction effects among the different covariates (Moncada *et al.,* 2021).

The dependent variable in the Cox regression is the time of individual taken to achieve the event of interest, the hazard ratio, denoted by lambda time $\lambda(t)$, which gives the probability of an event of interest according to the nature of the study. If it is related to survival, then the event of interest is recovered, which occurred before time t., and β is the Cox regression coefficient. Suppose we have X_1, X_2, X_3…. X_p covariates and the parameters for Cox regression are β_0, β_1, β_2, β_3…. β_p parameters for Cox regression, the hazard function for the Simple Cox model with one independent variable can be written as:

$$h(t) = h_0(t) \exp(\beta_1 X_1) \tag{3.3}$$

The Cox model with multiple covariates can be written as

$$h(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p) \tag{3.4}$$

Sometimes, we need the model in hazard ratio form with different covariates to write the Cox regression model.

$$\frac{h(t)}{h_0(t)} = \exp(\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p) \quad (3.5)$$

In logarithm form, we can write equation 3.3 below.

$$\ln\left\{\frac{h(t)}{h_0(t)}\right\} = \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p \quad (3.6)$$

The general form of Cox regression can be written as follows.

$$h(t) = h_0(t) \exp(\sum_{i=1}^{p} \beta_i x_i) \quad (3.7)$$

Where t denotes the survival time, i.e., years, months, and days, *h(t)*= is the expected hazard at time t, ho(t)= is the baseline hazard function at that specific time if all the covariates are equal to zero, exp(β_i) is the hazard ratio, β_i= is the vector of Cox regression parameters, X_i is the number of a predictor, covariates of Cox regression can be estimated using the maximum likelihood procedure.

3.14.1 Cox Regression Explanation

The Cox regression model is also known as the Cox Proportional Hazard Model (CPHM) because the primary assumption of Cox regression is proportionality, which means that the two individuals or two patients will be independent and their ratio over time will be constant and proportional, some other assumptions for using this model are: there will be no multicollinearity among the different covariates, there will be no heteroscedasticity, there will be no non-linear function between the time-dependent variable and different covariates, and the last assumption is that there will be no interaction effects among the different covariates (Moncada *et al.,* 2021).

3.14.2 Comparison of Cox Model VS Logit Model

The main difference between the Cox regression and logit model is that in the logit model, the dependent variable has probability or odds ratios. In contrast, the dependent variable in Cox regression is the hazard ratio. Further, CPHM is chosen over the logit model, in that in logit, we have zero or one, and we have two categories. Still, in Cox

regression, if we have two categories less than or greater than, and if an observation lies in the less than category, it is further categorized as much less than. For example, suppose we study early age marriage of girls with different indicators. In that case, the dependent variable for the logit model is either early marriage=1 or not=0. Still, in the case of Cox regression, early marriage further divides that a girl marrying at age 14 or 17 or what specific age, so both the cases come under the early marriage, logit model deals both cases at early age marriage category. So that's why Cox regression is preferred over the logit model in the case of censored data (Kumchulesi *et al.,* 2011). (Ghadimi *et al.,* 2010; Hashemian *et al.,* 2017) Compare the parametric logistic regression and semiparametric Cox regression to estimate the survival rate of colorectal cancer patients. Study results indicate that each parametric logistic regression outperforms semiparametric Cox regression.

3.14.3 Comparison of Cox Model with Tobit Model

To handle time-to-event data, few authors (Anastasopoulos *et al.,* 2012; Gong and Schaubel, 2018) use Tobit regression, while some researchers (Verweij and Van Houwelingen, 1994; Gémar *et al.,* 2016; Husain *et al.,* 2018) use parametric Cox regression for time-to-event data. The Cox regression and Tobit are different models from a mathematics algorithm perspective. Some pros and cons of the Cox model are that the model doesn't need the underlying distribution. The Cox model is semiparametric and works better for the general class of all kinds of semiparametric models to provide the hazard ratio. The cons of the Cox model are that it fails if the proportionality assumption violates non-linearity, high dimensionality, heteroscedasticity, and dependent covariates in data. (Gong and Schaubel, 2018) Estimated patients' survival rates and studied the pros and cons of the Tobit model used in survival analysis. The pros model is that the Tobit model can be used through

weighted maximum likelihood when each subject's censoring time is fixed or known. The Tobit model cannot be used when the censoring time is not fixed, or we can say random censoring. So, in most cases, censoring is random. Cox regression is preferred over the Tobit model (Tobin, 1958).

3.14.4 Strength of Cox Regression

The strength of the Cox regression model is it is easy to estimate the hazard ratio. In biomedical sciences, it is easily used to identify the prognostic factors, which further estimate the average survival rate of patients after a specific lethal disease or the average chance of recovery after a diagnosis of any specific disease. It is also easy to estimate and easy to interpret.

3.14.5 Shortcomings of Cox Regression

The shortcomings of Cox regression are that it fails to estimate when we have high dimensional data, and it's unable to estimate when we have a non-linear function with a different covariate. Also, it's powerless to give unbiased results in the case of heteroscedasticity and multicollinearity in the data, which are often present in the data out of the box because of the cross-sectional behavior of different individuals. Also, if the primary assumption of proportionality is violated, then the Cox proportional hazard model fails to give unbiased estimates (Moncada *et al.,* 2021).

3.14.6 Available Solution for the shortcomings of the Cox model.

To overcome this problem and shortcoming of Cox regression, in the recent literature related to survival analysis, the different authors have overcome this problem with various remedies, i.e. if we have high dimensional data or non-linear function, then in such cases, Machine Learning (Efficace *et al.,* 2006) technique such as Survival Support Vector Machine (SSVM), Random Survival Forest (RSF) and Extreme Gradient Boosting (XGB) is outperforming as compared to conventional Cox

regression which is used by (Moncada *et al.,* 2021) if we have heteroscedasticity problem in data, then in such situation Weighted Least Square (WLS) model is better to perform as compared to Cox regression (Mustefa and Chen, 2021) if we violate the main assumption of proportionality of two individual and patients, than in such case the extended Cox regression model outperforms than Cox regression (Husain *et al.,* 2018). To solve the problem of outlier in the next section we are going to explain robust cox with details equation step-by-step.

3.15 Robust Cox

Lin and Wei (1989) introduced robust Cox as a better alternative to the standard Cox model if there is an outlier in the data. Several studies have also compared the performance of robust Cox models with the standard Cox model under various scenarios, such as small sample sizes, skewed distributions, and in case of outliers. The results showed that the robust methods could provide more accurate and reliable estimates, especially in the presence of outliers, so the improved version of the Cox model is Robust Cox in case of an outlier (Bednarski 1993; Xiao *et al.,* 2016; Wand and Song, 2016; Carrasquinha *et al.,* 2018).

The robust Cox model can be written in equation form as given below.

Recall equation (3.7), the general form of Cox regression can be written as follows.

$$\prod_{i=1}^{N} \left[\frac{h\{t_i|x_i(t_i)\}}{\sum Y_j h\{t_i\{x_j(t_i)\}}\right]^{\delta_i} \tag{3.8}$$

Sum range is from i=1 upto N, t_i is the failure time, Value of sigma equal =1 if the time is observed, and equal =0, if the event time is not observed and censored value.

$$Y_i(t) = \begin{Bmatrix} 1 & (t \leq t_i) \\ 0 & (t > t_i) \end{Bmatrix} \quad (3.9)$$

To maximize the partial likelihood function, we determine B.

$$\sum_{i=1}^{N} y_i \left[x_i(t_i) \frac{S^1(t_i,B)}{S^0(t_i,B)} \right] = 0 \quad (3.10)$$

$$S^1(t_i, B) = \frac{1}{N} \sum_{i=1}^{N} Y_i(t) x_i(\mathrm{t}) \exp \{x_i(t)' B\} \quad (3.11)$$

$$S^0(t_i, B) = \frac{1}{N} \sum_{i=1}^{N} Y_i(t) \exp \{x_i(t)' B\} \quad (3.12)$$

In the below equation w_i is the weights which is multiplied to the standard error of the model, higher values given low weights and average values given high weight, to minimize the effect of outlier in the model.

$$\sum_{i=1}^{N} w_i y_i \left[x_i(t_i) \frac{S^1(t_i,\hat{B})}{S^0(t_i,\hat{B})} \right] = 0 \quad (3.13)$$

$$S^1\left(t_i, \hat{B}\right) = \frac{1}{N} \sum_{i=1}^{N} w_i y_i(t) x_i(\mathrm{t}) \exp \{x_i(t)' \hat{B}\} \quad (3.14)$$

$$S^0\left(t_i, \hat{B}\right) = \frac{1}{N} \sum_{i=1}^{N} w_i y_i(t) \exp \{x_i(t)' \hat{B}\} \quad (3.15)$$

$$\sum_{i=1}^{N} w_i y_i \left[x_i - \frac{w_i Y_i(t) x_i(\mathrm{t}) \exp \{x_i(t)' \hat{B}\}}{w_i Y_i(t) \exp \{x_i(t)' \hat{B}\}} \right] = 0 \quad (3.16)$$

In the above equation 3.16, w_i is the weights which is given low to the highest data points and while δ_i is the standard deviation of the error term. In which we assign the lower weights to the outlying observation, so the influence on the estimated parameter is decreased. The robust Cox method is more reliable if there is outlying observation in data (Binder, 1992; Carrasquinha *et al.,* 2018). In summary, robust Cox survival analysis methods can improve the accuracy and reliability of survival analysis results,

especially when the assumption of proportional hazards is not met. Researchers should consider the potential impact of violating the outlier assumption and consider using robust methods as an alternative.

3.16 Weighted Least Square for Time-to-event Study.

Weighted Least Square (WLS) is used to find a better alternative to Cox regression because the conventional Cox regression assumes that the covariates are Independently and Identically distributed (IID), meaning that there is no heteroscedasticity (Pang *et al.,* 2015). Still, its presence in the data, so Cox regression failed to estimate the coefficient with minimum standard error. The author found weighted least squares as a good alternative for survival analysis. (Yu *et al.,* 2019) proposed a Novel Quasi-Likelihood Ratio test for the homoscedasticity assumption in the survival analysis. The author said that if there is a violation of the homoscedasticity assumption in the survival analysis, then the Cox model leads to an invalid conclusion, so to test the assumption of homoscedasticity in the time-to-event analysis is necessary and proposed a new test of quasi-likelihood ratio test, if there is no heteroscedasticity problem than used conventional Cox regression, if there is heteroscedasticity in the error term than used WLS by (Mustefa and Chen, 2021).

Mustefa and Chen (2021) Considered WLS as a better alternative for Cox regression in the case of heteroscedasticity, Weighted Least Square (WLS) outperforms the existing Cox regression by comparing Root Mean Square Error (RMSE) Akaike Information Criteria (AIC) and using simulation studies (Pang *et al.,* 2015; Yu *et al.,* 2019).

Below equation 3.17 T_i is the dependent variable, which is time, α_o is the intercept, β_o is the hazard ratio of WLS of model, $1/\sigma$ is the weights of the auxiliary regression,

higher values given higher weights and average value given lower weights to balance the effect of outlier and heteroscedasticity in the model.

$$T_i = \alpha_o + \beta_o X_i + 1/\sigma(\mu_i)e_i \quad (3.17)$$

In next equation 3.18 we expand the general equation to WLS model with 2 covariates x_{1i} and x_{2i}, also $\mu_i = T_i - \hat{T}_i => \mu_i = \ T_i - \beta_1 x_{1i} - \beta_2 x_{2i}$

$$T_i = \alpha_o + \beta_1 x_{1i} + \beta_2 x_{2i} + 1/\sigma(T_i - \beta_1 x_{1i} - \beta_2 x_{2i})e_i \quad (3.18)$$

The above equation, 3.17 shows that equation 3.18 is divided by the residuals' standard error to adjust the model's heteroscedasticity problem (Mustefa and Chen, 2021).

3.17 Time-Dependent Cox

Time-dependent covariates in survival analysis are variables that can change their value over the course of the study period. For Example: Smoking Pattern, which change from day to day, week to week and month to month, Working Hours of daily wager: Which change from week to week. Blood pressure, weight, smoking status, and treatment exposure are all possible time-dependent covariates in survival analysis.

(Therneau *et al.,* 2017) Used an alternative model known as time-dependent Cox to handle such covariates in the model. Fisher and Lin (1999) studied the time-dependent covariate in the model and found that time-dependent covariate can be handled very carefully in the model. Otherwise, time-dependent covariates lead to biased results. The time-dependent Cox model is a better alternative for time-dependent covariates (Zhang *et al.,* 2018). The time-dependent Cox model can be written in equation form as given below. In equation (3.7) page#33, the general form of Cox regression can be written as follows.

we can include the time-dependent term, which resolved the problem of the time dependence between the covariates, and the above equation (3.) can be re-written as

$$h(t) = h_0(t)\exp\left[\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p + \gamma_1 z_1(t) + \gamma_2 z_2(t) + \cdots + \gamma_k z_k(t)\right] \quad (3.19)$$

In general form, the equation 3.19 can be written.

$$h(t) = h_0(t)\exp\left[\sum_{a=1}^{p1} \beta_i x_i \; + \; \sum_{b=1}^{p2} \gamma_i z_i(t_j)\right] + \mu_i \quad (3.20)$$

$$ln\left[\frac{h(t)}{h_0(t)}\right] = \exp\left[\sum_{a=1}^{p1} \beta_i x_i \; + \; \sum_{b=1}^{p2} \gamma_i z_i(t_j)\right] + \mu_i \quad (3.21)$$

In summary, the time-dependent Cox model can improve the accuracy and reliability of survival analysis results, especially when the assumption of time-dependent covariate is violated. Researchers should consider the potential impact of violating the time-dependent covariate assumption and use the time-dependent Cox model as a better alternative to time-dependent covariates (Fisher and Lin, 1999; Zhang *et al.,* 2018). Below table 3.1 shows the advantages and disadvantages of survival analysis in different nonparametric, semi-parametric, and parametric fields. The semiparametric model includes Kaplan Meier and Life tables, Cox regression, robust Cox, and time-dependent Cox. Parametric models include the logistic model.

Table 3. 1 Advantages and Disadvantages of Survival Analysis Models

Type	**Specific Methods**	**Disadvantages**	**Advantages**
Nonparametric	Kaplan Meier Life Tables	• Inaccurate Estimates • Difficult to estimates	Efficient when no suitable theoretical distribution is known
Semiparametric	Cox Regression Robust Cox Time-Dependent Cox	• Distribution is unknown. • Difficult to interpret	There is no need for an underlying distribution
Parametric	Logistic regression	• Inconsistent if distribution assumption is violated.	More efficient when survival times follow a particular distribution and easy to interpret

The table below 3.2 shows the advantages of different survival analysis models and how it's better in different scenarios. The models include Cox regression, machine learning models, LASSO, Ridge, EN, OSCAR, WLS, Time-dependent Cox, Extended Cox, and Stratified Extended Cox.

Table 3. 2 Efficiency of Different Statistical Models for Survival Analysis

Specific Methods	**Advantages**	**References**
Cox Regression	Efficient if **all the assumptions**[5] are satisfied.	(Cox, 1972)
Machine learning Model	Efficient if **non-linearity** in covariates.	(Moncada *et al.,* 2021)
Regularized Model: LASSO, Ridge, EN, OSCAR	Efficient if **high dimensionality** problem.	(Witten and Tibshirani, 2010; Wang *et al.,* 2019;)
Robust Cox	Efficient if **only Outlier in data**.	(Lin and Wei, 1989)
Weighted Least Square	Efficient if there is a **heteroscedasticity** problem in data.	(Mustefa and Chen, 2021)
Time-Dependent Cox	Efficient if **time-dependent covariates.**	(Akram *et al.,* 2007; Ata and Sözer, 2007)
Extended Cox Regression	Efficient if the **non-proportionality** assumption is violated.	(Husain *et al.,* 2018; Emoru *et al.,* 2020)
Stratified Extended Cox Regression	Efficient if **non-proportionality, time-independent, and time-dependent covariates.**	(Ratnaningsih *et al.,* 2019; Ratnaningsih *et al.,* 2021)
Modified Cox Regression	is efficient if there are **Outliers, heteroscedasticity, and time-dependent covariates**.	Nauman Ahmad PhD Thesis

[5] Assumption are Proportionality, linearity, No high dimensionality, no heteroscedasticity and no multicollinearity.

Table 3.3 shows the different models' final equations to see the differences among these models, such as Cox regression, WLS, Stratified Cox, extended Cox, stratified extended Cox, and modified Cox regression.

Table 3. 3 Different Types of Models with a Final Equation for Survival Analysis

Specific Models	**Final Equation and Methodology**	**References**
Cox Regression	$h(t) = h_0(t)\exp(\sum_{i=1}^{p}\beta_i x_i)$	(Cox, 1972)
Robust Cox	$\sum_{i=1}^{N} w_i y_i \left[x_i - \frac{w_i y_i(t) x_i(\mathrm{t})\exp\{x_i(t)'\hat{B}\}}{w_i y_i(t)\exp\{x_i(t)'\hat{B}\}}\right] = 0$	(Lin and Wei, 1989; Binder, 1992)
Weighted Least Square	$T_i^* = \alpha_0 x_i + \beta_o^T x_i + e_i$ Or $T_i^* = \alpha_o + \beta_1 x_{1i} + \beta_2 x_{2i} + 1/\sigma(\alpha_o + \beta_1 x_{1i} + \beta_2 x_{2i}) e_i$	(Mustefa and Chen, 2021)
Time-Dependent Cox	$h(t, x(t)) = h_o(t) exp\,[\sum_{i=1}^{p1}\beta_i x_i + \sum_{i=1}^{p2}\delta_j x_j(t)]$	(Husain *et al.,* 2018; Moreno *et al.,* 2018; Emoru *et al.,* 2020)
Stratified Cox Regression	$h_g(t,x) = h_{0g}(t)\exp[\beta_{1g}x_1 + \beta_{2g}x_2 \ldots\ldots + \beta_{pg}x_p]$ Or $h_g(t,x) = h_{0g}(t)\exp[\beta_1^* x_1 + \beta_2^* x_2 \ldots + \beta_p^* x_p + \beta_{p+1}^*(x_1 * z) + \beta_{p+2}^*(x_2 * z)\ldots + \beta_{p+1}^*(x_p * z)]$	(Akram *et al.,* 2007; Ata and Sözer, 2007)
Modified Cox Regression	$T_{(i\|j)}^* = \Gamma X_{(i\|j)}^* + \Lambda Y_{D(i\|j)}^* + \nu_{(i\|j)}^*$ ==> Proposed Modified Cox 2024 For more details, please see equation 4.17, page 46.	Nauman Ahmad PhD Thesis

Next chapters insight about data generating process, simulation study results, and real data analysis. Covering real data application of the proposed model in chapter 5.

Chapter 4

Modified Cox Proportional Hazard Model

4.1 Introduction

The Survival Analysis is vast field in which there are different kind of models such as parametric, semi parametric, and nonparametric models, further machine learning technique is also used such as Random Survival Forest (RSF), Extreme Gradient Boosting (XGB) and Artificial Neural Networks (ANN) but our area of interest is semi parametric, and we have proposed a modified model in the presence of outlier, heteroscedasticity and time-dependent covariates.

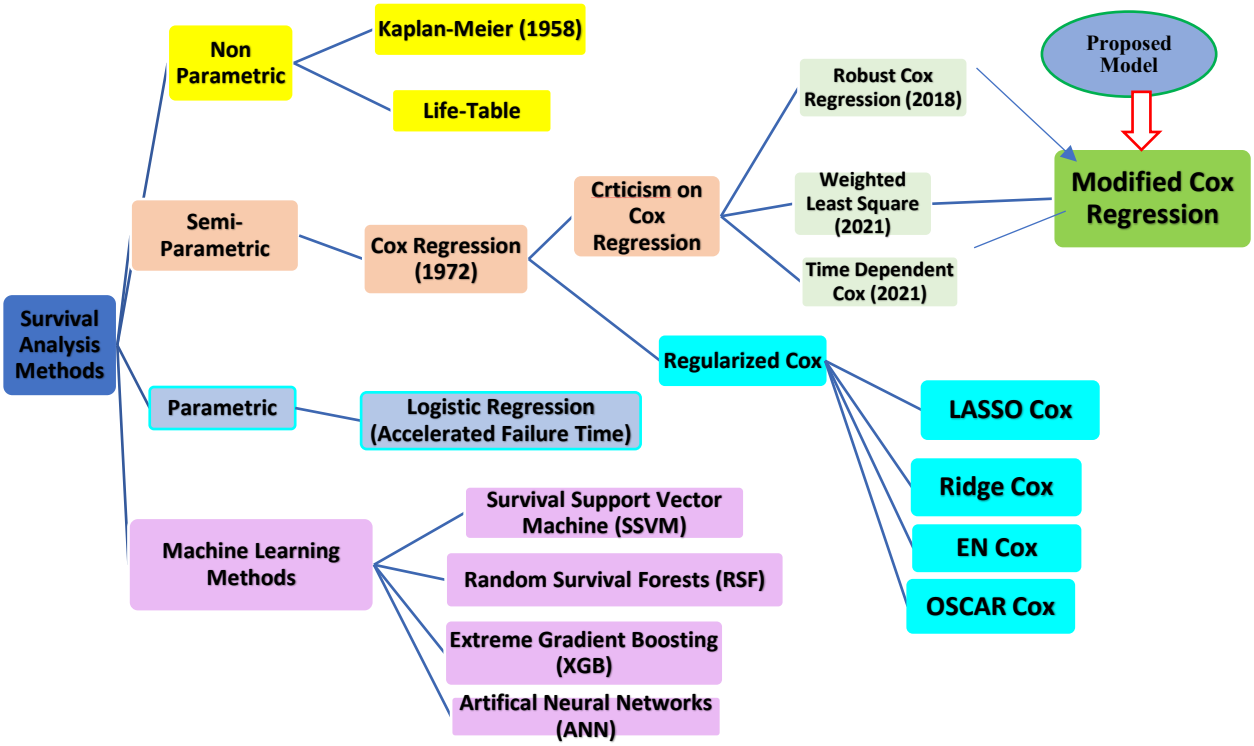


Figure 4. 1 Flow Chart of the proposed model

4.2 Explanation of the Proposed Model.

Our area of target is semi parametric area, the primary survival analysis first model by (Cox, 1972) is also belong to semi parametric category, which is further criticized by many researchers, due to the violation of Cox regression model assumption, if there is outlier in the data, Robust Cox outperformed, if there is heteroscedasticity problem then weighted least square performed better, and if there is time dependent covariate then time dependent Cox model is better, so the question arises in the mind, that what to do if these all three problem comes jointly in the data, what to do in such situation, that's why we proposed a new modified Cox regression model, which solved the problem of outlier, heteroscedasticity and time dependent covariates jointly and give better significant, reliable, efficient, consistent and unbiased results as compare to earlier advance survival analysis models.

4.3 Proposed Modified Cox Proportional Hazard Model

The Cox regression with one variable:

The Simple Cox regression with one independent variable can be written as:

$$h(t) = h_0(t)\exp(\beta_1 X_1) \tag{4.1}$$

The Cox regression with multiple predictors can be written as a recall equation can be written as given below.

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p) \tag{4.2}$$

Sometimes, we need the model in hazard ratio form with different covariates to write the Cox regression model.

$$\frac{h(t)}{h_0(t)} = \exp(\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p) \tag{4.3}$$

In logarithm form, we can write equation 3.3 below.

$$\mathrm{Ln}\left\{\frac{h(t)}{h_0(t)}\right\} = \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p \tag{4.4}$$

The general form of Cox regression can be written as follows.

$$h(t) = h_0(t)\exp(\sum_{i=1}^{p}\beta_i x_i) \quad (4.5)$$

To incorporate the time-dependent covariate term in the model 4.5 equation becomes like this:

$$h(t) = h_0(t)\exp\left[\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p + \gamma_1 z_1(t) + \gamma_2 z_2(t) + \cdots + \gamma_k z_k(t)\right] \quad (4.6)$$

In general form, the equation 4.6 can be written.

$$h(t) = h_0(t)\exp\left[\sum_{a=1}^{p1}\beta_i x_i \; + \; \sum_{b=1}^{p2}\gamma_i z_i(t_j)\right] \quad (4.7)$$

$$ln\left[\frac{h(t)}{h_0(t)}\right] = \left[\sum_{a=1}^{p1}\beta_i x_i \; + \; \sum_{b=1}^{p2}\gamma_i z_i(t_j)\right] \quad (4.8)$$

In the presence of hetero we have to further modify the model and divide equation 4.8 by weights, will assign lower weight to bigger value and bigger weight to average value. So to remove heteroscedasticity from the model, equation 4.8 is divided by w_i

$$\frac{ln\left[\frac{h(t)}{h_0(t)}\right]}{w_i} = \left[\frac{\sum_{a=1}^{p1}\beta_i x_i}{w_i} \; + \; \frac{\sum_{b=1}^{p2}\gamma_i z_i(t_j)]}{w_i} + \mu_i \right. \quad (4.9)$$

Where t denotes the survival time, i.e., years, months, and days, *h(t)*= is the expected hazard at time t, ho(t)= is the baseline hazard function at that specific time if all the covariates are equal to zero, σi is the weights which solved the problem of outliers and heteroscedasticity, $ln\left[\frac{h(t)}{h_0(t)}\right]$ the hazard ratio, β_i= is the vector of Cox regression parameters, x_i is the number of a covariates, z_i is the number of a time dependent covariates. Covariates of the model can be estimated using the maximum likelihood procedure.

To write the model in more compact format:

Let $\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p = \Gamma X_i$ (4.10)

$$\gamma_1 z_1(t) + \gamma_2 z_2(t) + \cdots + \gamma_k z_k(t) = \Lambda Y_{Di} \quad (4.11)$$

$$\frac{ln[\frac{h(t)}{h_0(t)}]}{w_i} = T_i^* \,, \tag{4.12}$$

$$\frac{X_i}{w_i} = X_i^* \,, \tag{4.13}$$

T_i^* , X_i^* are individual values, which are incorporated into final equation 4.16.

$$\frac{Y_{Di}}{w_i} = Y_{Di}^* \,, \tag{4.14}$$

$$\frac{\mu_i}{w_i} = \nu_i^* \tag{4.15}$$

$$T_i^* = X_i^* + Y_{Di}^* + \nu_i^* \tag{4.16}$$

Winsorization[6] is a statistical technique used to reduce the effect of possibly spurious outliers by limiting extreme values in the data. Instead of removing outliers, winsorization transforms them to a specified percentile value. For example, values above the 95th percentile might be set to the 95th percentile value, and values below the 5th percentile might be set to the 5th percentile value. This method helps maintain the overall structure of the data while reducing the impact of outliers.

To handle outliers in cross-sectional data, we apply winsorization. This involves limiting extreme values to reduce their impact. In statistical models, winsorization is implemented using specific commands to adjust data appropriately. In the case of cross-sectional data, the existence of an outlier can be identified as extreme observation or either the upper or lower tail of data.

To introduce the winsorization effect in the equation, we have used winsorization on a different level such as 2% or 5%, depending on the number of the outlier. Urooj and

[6] Winsorization is a procedure to minimize the influence of outliers in your data by replacing extreme value on average value.

Asghar (2017) point out five different types of an outlier: additive outlier, level shift outlier, innovative outlier, transitory change, and seasonal level shift/seasonal outlier. From equation 4.10 to 4.15, combining all equation in more compact econometrics form the equation 4.16 become like that:

$$T^*_{(i|j)} = \Gamma X^*_{(i|j)} + \Lambda Y^*_{D(i|j)} + \nu^*_{(i|j)} \tag{4.17}$$

Where i represents the number in ordered form from (i) =i=1,2,3....n and where j= 0.01,…. 0.05 represent the winsorization range depending upon the quantity of outliers. So, estimate your equation using equation 4.17 to avoid the problem of an outlier, hetero, and time-dependent covariates in the model and given unbiased, efficient and consistent estimates.

4.4 Proof of proposed modified Cox model showing variance is constant

Recall equation (4.16) $T^*_{(i)} = \alpha X^*_i + \beta Y^*_{Di} + \nu^*_i$, we have to check the heteroscedasticity of the error term of the model, either it's constant or vary.

$$\nu^*_i = \frac{\mu_i}{w_i} \tag{4.18}$$

Where z_i is the weights of the model, given smaller weights to larger values and high weights to average values.

$$var(\mu_i) = \sigma^2_i \tag{4.19}$$

4.4.1 Constant Variance.

The above equation (4.19) is case of heteroscedasticity, we can see that sigma square has subscript I, which shows that the variance is not constant.

$$Suppose\ \sigma^2_i = \sigma^2 w^2_i \tag{4.20}$$

Now we will check the variance of the ν^*_i.

$$var(\nu^*_i) = var(\frac{\mu_i}{w_i}) \tag{4.21}$$

$$var(v_i^*) = \frac{1}{w_i^2} var(\mu_i) \quad (4.22)$$

$$var(v_i^*) = \frac{1}{w_i^2} \sigma_i^2 \quad (4.23)$$

$$var(v_i^*) = \frac{1}{w_i^2} \sigma^2 w_i^2 \quad (4.24)$$

$$var(v_i^*) = \sigma^2 \quad (4.25)$$

Now variance of $v_i^* = \sigma^2$ which is constant, so estimate parameter from equation (4.17) instead of (4.1) to avoid the problem of heteroscedasticity in the model.

4.4.3 Time Dependency

Next, we have to check whether the parameters are time independent. In first stage, we have identified that which variables are time dependent, suppose the variable z_1 has time dependent covariate, So we will multiply that variable with a time variable to adjust the time dependency in that variable z_1(t)

$$\gamma_1 z_1 \quad (4.26)$$

So γ_1 is the coefficient of variable and z_1 is the variable having time dependency in below equation 4.26.

$$\gamma_1 z_1(t) \quad (4.27)$$

If more than one variable has time-dependent covariates which is possible in the model, the general form is given below.

$$Y_{Di} = \sum_{b=1}^{zk} \gamma_i z_i(t_j) \quad (4.28)$$

Incorporate this form of covariate to solve the problem of time-dependent covariate in the model (Husain *et al.,* 2018; Moreno *et al.,* 2018; Emoru *et al.,* 2020).

$$Y_{Di} = \gamma_1 z_1(t) + \gamma_2 z_2(t) + \cdots + \gamma_k z_k(t) \quad (4.29)$$

If we estimate our parameter from equation (4.17) the estimates of the model will be BLUE.

4.5 Difference between WLS and Proposed Modified Cox

The WLS model only fix the issue of heteroscedasticity, while modified Cox model fixed for outlier, heteroscedasticity and time-dependent covariate simultaneously, by allowing time-dependent covariates in the model, and winsorization effect in the model based on the outlier magnitude, if the quantity and magnitude of outliers are higher, we included higher percentage of winsorization above 5%, if lower magnitude and quantity of outliers, we included below 5% winsorization level to handle the effect of outliers in the model, the modified cox model is the customized model, which can fixed all the three issues in the model simultaneously and give unbiased and efficient estimates.

4.6 Distribution of the modified Cox Remain the same

(Lin and Wei, 1989) discussed robust methods for Cox regression, highlighting that even when dealing with outliers or other model violations, the fundamental distributional assumptions of the proportional hazards model remain unchanged. (Therneau and Grambsch, 2000) covered various extensions of the Cox model, including time-dependent covariates and robust methods. Also discussed how these modifications are still grounded in the same distributional assumptions as the original Cox model, the central role of distributional assumptions like the exponential distribution in survival models remains unchanged (Wei, 1992).

4.7 Research Design

The (Cox, 1972) model is based on different assumptions, such as proportionality of individuals, linearity of variables, homoscedasticity, and no multicollinearity. If any of these assumptions are violated, we can't estimate the Cox regression hazard ratio efficiently. In literature, to find which model performs better for outliers, heteroscedasticity, and time-dependent covariates jointly. We did not find any work addressing this issue before, which shows how to solve the problem of outliers,

heteroscedasticity, and time-dependent covariates jointly in the same data sets. This problem motivates me to find the solution and proposes a model that can simultaneously cover all the problems.

4.8 Analytical Framework

The evaluation of the proposed model is based on two-fold analyses. Firstly, we have designed and conducted simulation experiments evaluating different scenarios, which encompassing the existence of outlier, heteroscedasticity, and time-dependent covariates. Secondly, the proposed model is applied on real data for the case of Pakistan.

4.8.1 Analysis Application

Analysis software used in the study is RStudio version 2023/3.3.0 for the proposed modified Cox model and other conventional econometrics models used in survival analysis. For both simulated data and real data application is performed in RStudio.

4.9 Estimation and Model Selection Criteria

A vast amount of literature was reviewed for model selection criteria. The essential to highlight some critical criteria that are frequently used in empirical are: Root Mean Squared Error (RMSE), Mean Absolute Error (Kumchulesi *et al., 2021*), and Mean Absolute Percentage Error (MAPE). We have adopted the Monte Carlo simulation to find the frequency of minimum RMSE, MAE, and MAPE of the different models such as Cox regression, Robust Cox, Weighted Least Square, Time-Dependent Cox, and modified Cox regression, minimum value of RMSE, MAE and MAPE better the model.

4.9.1 Root Mean Squared Error (RMSE)

The root means square error (RMSE) has been used as a standard statistical, econometric technique to measure the two-model performance in the case of prediction and forecasting. The RMSE is more appropriate to represent model performance, when

the error distribution is expected to be Gaussian. Chai and Draxler (2014) found that the RMSE is better and more advantageous than other model selection criteria to illustrate the error. The model with fewer values of RMSE shows better forecasting performance than the other. The formulae for RMSE are given below.

$$RMSE = [n^{-1} \sum_{i=1}^{n} e_i{}^2]^{1/2} \quad (4.18)$$

Where n is the number of observations e_i Is the error term, which is equal to $e_i = (y_i - \hat{y})$ and taking the square root of MSE is called RMSE.

4.9.2 Mean Absolute Error

The Mean Absolute Error has been used as a standard statistical, econometric technique to measure the two-model performance in case of prediction. One of the best advantages of Mean Absolute Error and Root Mean Squared Error (RMSE) is that it identifies a model with the best prediction ability. The model with less value of MAE is considered the better forecasting model compared to the other. The formulae for MAE are given below (Kumchulesi *et al., 2021*).

$$MAE = [n^{-1} \sum_{i=1}^{n} |e_i|] \quad (4.19)$$

Where n is the number of observations e_i Is the error term, which is equal to $e_i = (y_i - \hat{y})$. In the first step of MAE, we find the error term; in the second step, we take absolute; in the final step, we take mean absolute error, which is called MAE.

4.9.3 Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) has been used as a standard statistical, econometric technique to measure the two-model performance in case of prediction. One of the best advantages of Mean Absolute Percentage Error (Kumchulesi *et al.,* 2021) is that it identifies a model with the best prediction ability in percentage. The

model with less value of MAPE is considered the better forecasting model compared to the other. The formulae for MAPE are given below.

$$MAPE = [n^{-1} \sum_{i=1}^{n} |e_i| * 100] \quad (4.20)$$

Where n is the number of observations e_i Is the error term, which is equal to $e_i = (y_i - \hat{y})$. In the first step, we find the error term; in the second step, we take absolute; in the next step, we take mean absolute error. In the final step, we multiply by 100 to get MAPE, called MAPE.

Next chapter 5 insight about simulation study and chapter 6 real data application of the proposed model, and final chapter 6 is conclusion and policy recommendation of the study.

Chapter 5

Simulation Analysis

5.1 Introduction

Model selection is one of the crucial steps of empirical research throughout all disciplines. In survival analysis, selecting the relevant model for the data is essential to achieve the study's objective. The current chapter is divided into the following subsections: the first includes the data-generating process, the second consists of the different scenarios, the third contains the results, and the last has the comparison and conclusion based on simulation.

We have done a Monte Carlo simulation of the existing conventional model and proposed a modified Cox regression. To find accurate estimates, validate the proposed model's consistency, unbiased, and efficiency. Furthermore, this study also investigates modified Cox regression performance and how best fit, unbiased, consistent, and efficient estimates have been given by using simulated and real data, and this study has filled this gap.

5.2 Data Generating Process (DGP)

The question we are trying to answer is which model is best in time-to-event studies if there is existence of outlier, heteroscedasticity, and time-dependent covariates in the data simultaneously. First, we generate the exponential distribution series and later, we introduce the outliers, heteroscedasticity, and time-dependent covariates step by step.

The model structure is defined as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots . \beta_k x_k + \varepsilon_i \qquad (5.1)$$

Where $x_1, x_2, x_3, x_4, \ldots, x_k$ are regressors following varying distributions such as exponential, binomial and normal distribution, while ε_i follows an exponential distribution with parameter θ such that $\varepsilon_i \sim \exp(\theta)$.

5.2.1 Data Generating Process for Outlier

For Introducing an outlier in the independent variable, in exponential distribution, identifying outliers can be less straightforward compared to some other distributions due to its inherent right-skewness. Outliers in an exponential distribution are often considered to be values that are much larger than the majority of the data and deviate significantly from the typical behavior of the distribution. In exponential distribution the standard outlier to be considered is 4SD, due to right skewed behavior and the 6SD is taken to see the effect of increase in outlier magnitude on the comparing of different survival analysis model (Jiao, 2019).

$$X_i = X_1, X_2, X_3, \ldots . X_n \tag{5.2}$$

Where X~exp(θ), where θ > 0.

For 5% outlier and 4 standard deviations, we replace the value of average multiply by 4SD on 5% of the data. Same process we did for 10% outlier and 6SD (Muhammad, 2022).

which is generated in the data by multiplying four standard deviations to the average value of the series. If we want to give a high magnitude to the outlier i.e 6SD, then the six standard deviation is multiplied by the average value of the series, and the X_i is generated from the exponential distribution and the variance covariance of the data matrix are:

$$\text{Population Var-Cov} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix} \tag{5.3}$$

Variance- covariance matrix with outlier σ_{ii}^{*} shows 5% outlier with 4SD case 1, and σ_{ii}^{**}shows 10% observation outlier with 6SD case.

case 1 (5% observation outliers at 4SD)

$$\sum_1^* = \begin{bmatrix} \sigma_{11}^* & \sigma_{12}^* & \dots & \sigma_{1k}^* \\ \sigma_{21}^* & \sigma_{22}^* & \dots & \sigma_{2k}^* \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1}^* & \sigma_{k2}^* & \dots & \sigma_{kk}^* \end{bmatrix} \tag{5.4}$$

case 2 (10% observation outliers at 6SD)

$$\sum_2^* = \begin{bmatrix} \sigma_{11}^{**} & \sigma_{12}^{**} & \dots & \sigma_{1k}^{**} \\ \sigma_{21}^{**} & \sigma_{22}^{**} & \dots & \sigma_{2k}^{**} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1}^{**} & \sigma_{k2}^{**} & \dots & \sigma_{kk}^{**} \end{bmatrix} \tag{5.5}$$

$$\text{Data Matrix } \underset{(n\ x\ p)}{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \tag{5.6}$$

In such a way, an outlier of 4SD and 6SD magnitude case is generated in the series.

when I check observation above 99.7%, it's 3 times of standard normal, value above 3SD

Frost (2021) and Chebyshev (1882) suggested an alternative method for outlier in non-normal distribution ±10SD contains 99.7% area in any non- normal distribution, ±4.47SD contains 95% central area in any non-normal distribution, suggested the formula to find SD.

Area with KSD~ $(1-1/k^2)$ => ±2SD= 1-1/4 => =0.75 => ±2SD contains 75% area

Now for 99% data

=$(1-1/k^2)$=0.99, solve for k => $1-0.99=1/k^2$ => $0.01=1/k^2$ => $K^2=100$ => So k=10SD

Now for 95% data => $(1-1/k^2)$=0.95, solve for k => $1-0.95=1/k^2$ => $0.05=1/k^2$ => K^2=20, So k=4.41

So, k=4SD approximately. The below table 5.1 shows the outlier decision in exponential distribution case, either to take 3SD or 4SD, or higher SD. After experiment we decide to take 4SD for considering 0.4% data to be outlier, and 6SD for considering 0.3% data to be outlier.

Table 5.1 Outlier decision either 4SD or 6SD Exponential Distribution Experiments

Distribution	**Sample, n=1000**	**SD**	**Max value**	**Statement**	**SD Conclusion**
Standard Normal Distribution	.3% data considered to be outlier	1	3-3.5 range	So outside the central 99.7% interval of data, is considered to be outlier.	Three values are above 3SD
Exponential Distribution Case A	.3% data considered to be outlier	0.52	3.5-4 range	So above the 99.7% data, is considered to be outlier	Three values are above **6SD**
Exponential Distribution Case B	.4% data considered to be outlier	0.52	3.5-4 range	So above the 99.7% data is considered to be outlier	Three values are above 4SD
	Now Sample n=50,000				
Exponential Distribution Case C	.3% data considered to be outlier	0.33	3.5-4.5 range	So above the 99.7% data, is considered to be outlier	Three values are above **6SD**
	Now Sample n=100,000				
Exponential Distribution Case D	.3% data considered to be outlier	0.33	4-4.8 range	So above the 99.7% data, is considered to be outlier	Three values are above **6SD**

Source: Author own Calculation SD: Standard Deviation

5.2.2 Data Generating Process for Heteroscedasticity

The next step is to generate heteroscedasticity in the error term. On the principal diagonal, we have variances, and off the diagonal, we have covariances. If the variance is not constant, such a case is known as heteroscedasticity, so to generate heteroscedasticity in the error term, we follow the following steps.

In the first step we generate Xi variables from exponential distribution with parameter theta because the Cox model follows exponential distribution second step, we generate the Y variable which is the function of all Xi variables, all X_i is linear function of Y, in the third step we multiply a random variable with the error term variance to generate heteroscedastic error term.

The case below is homoscedasticity because the variance is sigma square and variance is constant on a diagonal axis which is $\sigma^2\ In$, Var-Cov(u u')= σ^2 ,

The X_i is generated from the exponential distribution and the variance-covariance matrix of the errors are:

$$E(uu') = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\ In \quad \rightarrow \text{Homoscedasticity Case} \qquad (5.7)$$

The case below is heteroscedasticity because the variance is not sigma square but varies which is $\sigma_i^2\ In$, Var-Cov(u u')= σ_i^2 , Where $_i$ = heterogeneous matrix of errors.

On diagonal we have variances, while off diagonal we have covariance's, the above matrix shows that variances are constant, which is σ^2, while the below matrix shows the variances are not constant, which is σ_i^2, below case is heteroscedasticity.

$$E(uu') = K_i \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma_i^2 \quad \rightarrow \text{Homoscedasticity Case} \qquad (5.8)$$

Where K_i is the random variable with mean 0 and variance σ_i. $K_i \sim N(0, \sigma_i)$, So that error become heteroscedastic.

$$E(uu') = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \Omega \quad \rightarrow \text{Heteroscedasticity Case} \qquad (5.9)$$

σ_i^2 This means there is a heteroscedasticity problem, σ_i^2 Shows that variance of the error term varies, the matrix is called Omega matrix Ω. In such a way the case of heteroscedasticity is generated, Xi is a random variable generated from the exponential distribution, $X_1 \sim \exp(\theta)$, taking the initial value of theta equal to one. In such a way the heteroscedasticity in the error term is generated.

5.2.3 Data Generating Process for Time-Dependent Covariate

The last step is to generate time-dependent covariates in the independent variable. This means the variables that change are not constant, such as smoking patterns, which may increase or decrease with time. For introducing a time-dependent covariate on the right-hand side of the model, we have used the methodology of (Thernao, 2022). In which the time-dependent covariate is generated random series from exponential distribution and multiplied by the time trend variable. i.e (1,2,3….).

In the case of the time-dependent covariate, the variable age= age_i means that age varies across the study, which shows that variable age is a time-dependent covariate. To multiply time (t) 1,2,3… With a variable generated with exponential distribution, that variable is a time-dependent covariate in the model. For the time-dependent covariate,

the methodology of (Thernao, 2022) is used to generate the time-dependent covariate in the model.

Suppose X_5 is the time-dependent covariate.

X_{5A}~exp(θ) , where θ > 0.

$X_5 = X_{5A}$*t, where t is 1,2,3, …. N.

X_5 is considered to be an age variable, which changes with time, that's why multiplied with time to generate the time dependent covariate feature in that variable and the final equation looks like below equation 5.10:

$$T^*_{(i|j)} = \Gamma X^*_{(i|j)} + \Lambda Y^*_{D(i|j)} + \nu^*_{(i|j)} \quad (5.10)$$

5.3 Simulation Design and Different Scenarios

For simulation experiments, four scenarios are considered by allowing the outlier, heteroscedasticity, and time-dependent covariates with varying sample sizes and various covariates. To be more specific, the random finite samples of sizes 100 and 500 are drawn from the exponential distribution with 50,000 times replication. Moreover, the first scenario is about the outlier and time-dependent covariates, scenario 2 is about outlier and heteroscedasticity, scenario three is about heteroscedasticity and time-dependent covariates, and final scenario is about outlier, heteroscedasticity and time dependent covariates as presented below in Fig 5.1.

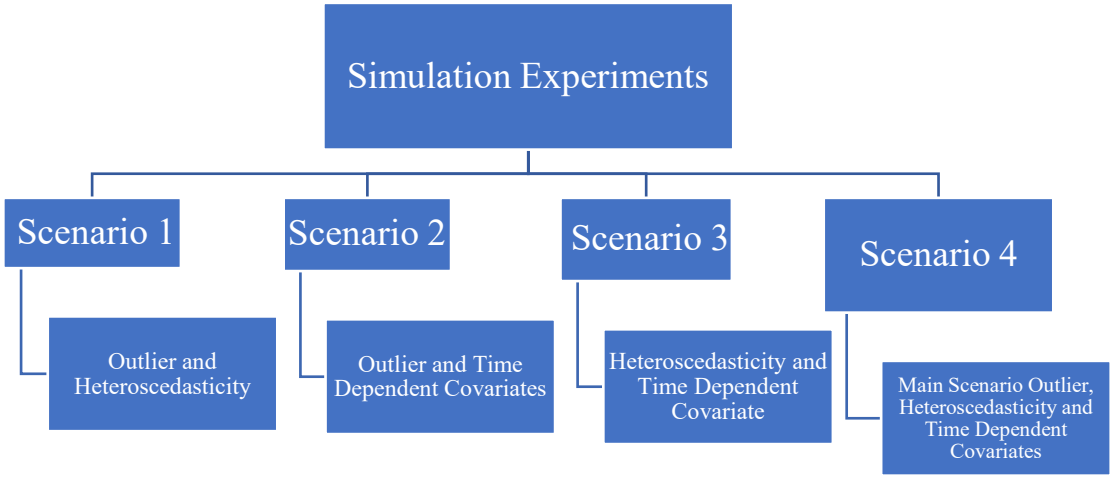


Figure 5. 1 Different Scenario for Simulation

Here, we take six regressors, so that we have both continuous and categorical variables, and the dependent are generated from an exponential distribution. Epsilon alpha is generated from an exponential distribution. A detailed explanation of DGP is provided in the preceding chapter. The RMSE, MAE, and MAPE are taken as performance indicators to evaluate the performance of all models.

5.4 Reason for Varying Different Scenarios

The reason for varying different scenario in the simulation design is important, to see the performance of proposed modified Cox in different scenarios, that either it's best in all scenario and one can used for general purpose or best in some specific scenarios, that's why need for varying in different scenario is important, increase in sample size, quantity of outliers from 5% to 10%, magnitude of outliers from 4SD to 6SD, and exponential distribution parameter theta increase from one to two.

When the underlying distribution is normal, the value is considered to be outlier if it is above the range of 3SD, while exponential distribution is rightly skewed, that's why

initial magnitude for outlier is taken 4SD and later on increase to 6SD, also increase the quantity of outlier from 5% to 10% to see the performance of different models.

5.5 Simulation Results and Discussion

The total number of main scenarios is four: I, II, III, IV; Scenario one is about Heteroscedasticity and Outlier, Scenario II is about outlier and time-dependent case, Scenario III is about heteroscedasticity and time-dependent case, the final scenario is about Outlier, heteroscedasticity, and time-dependent covariate. Each scenario is divided into sub-scenarios like the quantity of outlier 5%, 10%, and 20%, magnitude of the outlier 4SD and 6SD. The main scenario is further varied according to sample size, i.e., small, and large, such as 100 and 500. The third parameter, which is varied, is about theta because this model class follows an exponential distribution, so the exponential distribution parameter is theta, so theta is also varied from 1 to 2.

5.6 Scenario-I Handling Outlier and Heteroscedasticity Covariate

The first scenario considers the DGP with different cases such as 5% outlier, 10% outlier, and 20% outlier cases with a magnitude of 4 standard deviations and 6 standard deviations. The heteroscedasticity case is natural. The heteroscedasticity in the error term is observed using the brush pagan test, and the outlier in the data is observed using the influence plot. In such circumstances, it is crucial to estimate the desired effect of an independent variable through Cox regression. Thus, in different simulation scenarios and experiments, we understand the problem better in the detailed DGP given in the previous chapter; below Table 5.1 results suggest that the robust Cox model performed better in all cases, either if 5% outlier or 10% or 20% outlier case.

Table 5.1 below shows the simulation finding in the outlier and heteroscedasticity case. For three cases of an outlier, such as a 5%, 10%, and 20% outlier quantity, all models' performances slightly decreased due to the increase in the quantity of outlier. Still, the

DGP used for variable generation is from an exponential distribution. Hence, an increase in theta makes all the slightly improved. The RMSE of the model is decreasing and converging towards zero, as we can see in below table 5.1 the modified Cox model value is 34.6, if the theta value is one and if theta value is two, the modified Cox model RMSE is 19.3, a significant decrease in RMSE.

Figure 5.2 below shows that the robust Cox model has a low RMSE of 32.0, low MAE of 13.5, and low MAPE of 69.1 compared to Cox regression, Weighted Least Square, and modified Cox. This means that in the case of outlier and heteroscedasticity, the best model is the robust Cox model, which handles the problem of outlier and heteroscedasticity. The results are stable even if there is an increase or decrease in sample size, an increase or decrease in the outlier quantity, or an increase or decrease of exponential distribution parameter theta doesn't affect model performance position but slightly improved all models, from scenario one, it is concluded that robust Cox is better in the case of outlier and heteroscedasticity case.

Table 5. 1 Outlier and Heteroscedasticity Case

N=100, Sims=50,000, Parameter=1, Outlier with 4 SD						
5% Outlier						
	Theta=1			**Theta=2**		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	51.7	20.4	104.7	28.4	11.2	57.6
Robust Cox	**32.0**	**13.5**	**69.1**	**18.8**	**7.4**	**38.0**
WLS	42.9	16.9	86.9	23.6	9.3	47.8
Modified Cox	34.6	13.9	71.2	19.3	7.6	39.2
10% Outlier						
	Theta=1			**Theta=2**		
Cox	62.6	24.7	126.7	34.4	13.6	69.7
Robust Cox	**38.7**	**16.3**	**83.6**	**22.7**	**9.0**	**46.0**
WLS	51.9	20.5	105.2	28.6	11.3	57.8
Modified Cox	41.9	16.8	86.1	23.4	9.2	47.4
20% Outlier						
	Theta=1			**Theta=2**		
Cox	80.1	31.6	162.3	44.1	17.4	89.3
Robust Cox	**49.6**	**20.9**	**107.1**	**29.1**	**11.5**	**58.9**
WLS	66.5	26.2	134.7	36.6	14.4	74.1
Modified Cox	53.6	21.5	110.4	30.0	11.8	60.7

Note: Authors Own Calculation, Sims stands for Simulation.

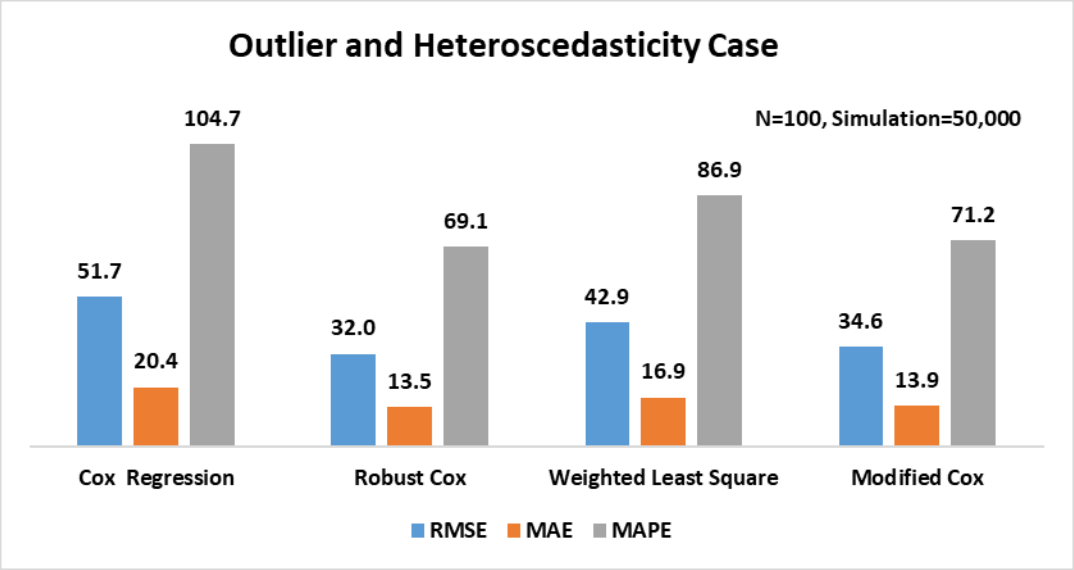


Figure 5. 2 Handling Outlier and Heteroscedasticity

5.7 Scenario-II Handling Outlier and Time-Dependent Covariate

The second scenario considers the DGP with different cases, such as a 5%, 10%, and 20% outlier with a time-dependent covariate, the outlier in the data is observed using the influence plot, and the time-dependent covariate is observed using the Shenfield residual ZPH test (Husain *et al.,* 2018; Moreno *et al.,* 2018; Emoru *et al.,* 2020). In outlier and time-dependent covariate cases, it is crucial to estimate the desired effect of an independent variable through Cox regression. Thus, in different simulation scenarios and experiments, we understand the problem better. The detailed DGP is given in the previous chapter. Table 5.2 results suggest that modified Cox performed better in all cases of outliers, either 5%, 10%, or 20 %.

Table 5.2 below further shows the outlier and time-dependent covariate case simulation finding for three quantities of outliers, such as 5%, 10%, and 20 %. The DGP used for variable generation is from an exponential distribution. Hence, an increase in theta makes all the models slightly better. The RMSE of the model is decreasing and converging towards zero. As shown in Table 5.2, in the modified Cox model, the RMSE

value is 42.4. If the theta value is one and theta value increases to two, the modified Cox model RMSE is 23.3.

Figure 5.3 below shows that the modified Cox model has a low RMSE of 42.4, a low MAE of 17.7, and a low MAPE of 81.4 compared to Cox regression, Robust Cox, and Weighted Least Square. which means that in the case of outlier and time-dependent covariates, the best model is the modified Cox model, which handles the problem of outlier and time-dependent covariates jointly. The results are stable even if there is increases or decrease in sample size or changing quantity of outliers. Either the exponential distribution theta parameter increases or decreases doesn't affect the performance of the modified Cox model.

Table 5. 2 Outlier and Time-Dependent Covariate Case

N=100, Sims=50,000, Parameter=1, Outlier with 4SD						
5% Outlier						
	Theta=1			**Theta=2**		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox Model	61.4	25.7	117.9	33.8	14.1	64.8
Robust Cox	45.3	19.0	87.0	24.9	10.4	47.9
Time-Dependent	51.0	21.4	98.0	28.1	11.7	53.9
Modified Cox	**42.4**	**17.7**	**81.4**	**23.3**	**9.8**	**44.7**
10% Outlier						
	Theta=1			**Theta=2**		
Cox Model	74.3	31.1	142.7	40.9	17.1	78.5
Robust Cox	54.8	22.9	105.3	30.9	12.9	59.3
Time-Dependent	61.7	25.8	118.5	34.8	14.6	66.8
Modified Cox	**51.3**	**21.5**	**98.4**	**28.9**	**12.1**	**55.5**
20% Outlier						
	Theta=1			**Theta=2**		
Cox Model	90.3	37.8	173.3	49.6	20.8	95.3
Robust Cox	70.2	29.4	134.9	38.6	16.2	74.2
Time-Dependent	75.0	31.4	144.0	41.3	17.3	79.2
Modified Cox	**65.7**	**27.5**	**126.1**	**36.1**	**15.1**	**69.4**

Note: *Authors Own Calculation, Sims stands for Simulation.*

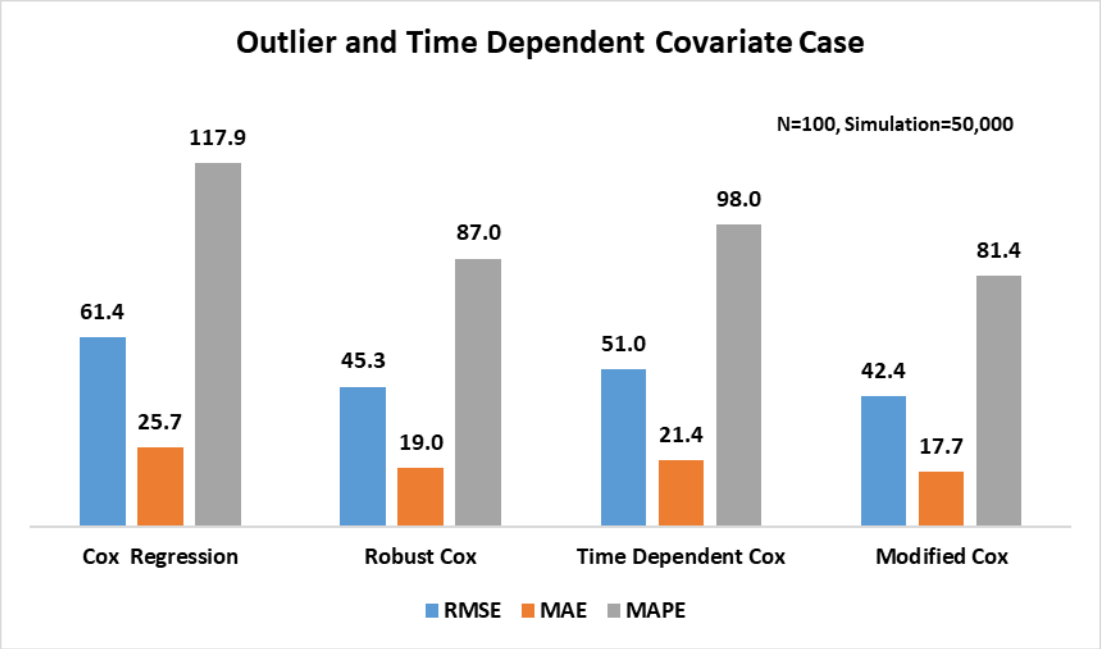


Figure 5. 3 Handling Outlier and Time-Dependent Covariate

5.8 Scenario-III Handling Heteroscedasticity and TD Covariate

The third scenario considers the DGP with different cases, such as heteroscedasticity and time-dependent covariate cases. The heteroscedasticity in the model error term is observed using the (Breusch and Pagan, 1979) test, while the time-dependent covariate is observed using the Shenfield residual ZPH test. In such circumstances, heteroscedasticity and time-dependent covariate cases, it is crucial to estimate the desired effect of an independent variable through Cox regression. Thus, in different simulation scenarios and experiments, we understand the problem better. Detailed DGP is given in the previous chapter.

Table 5.3 below shows the simulation finding in the outlier and time-dependent covariate case cases. The outlier quantity, such as 5%, 10%, and 20%, also doesn't affect the results of any model or RMSE of the model. Still, the DGP used for variable generation is from an exponential distribution. Hence, an increase in theta makes all the models slightly better. The RMSE of the model is decreasing and converging towards

zero, as we can see in the modified Cox model 38.7. If the theta value is one and theta value is 2, the modified Cox model RMSE is 21.4, with a significant decrease in RMSE, same for sample size increase slightly improved the performance of all model RMSE. Figure 5.4 below shows that the modified Cox model has a low RMSE of 38.7, low MAE of 17.0, and low MAPE of 82.2 compared to Cox regression, Robust Cox, and Weighted Least Square. which means that in the case of heteroscedasticity and time-dependent covariates, the best model is the modified Cox model, which handles the problem of heteroscedasticity and time-dependent covariates jointly. The results are stable even if there is an increase or decrease in sample size or theta parameter of the exponential distribution or outlier quantity and magnitude. Also, the modified Cox model performance is stable. It doesn't affect parameter stability, performance, or unbiased.

Table 5. 3 Hetero and Time-Dependent Case

Sims=50,000, Parameter=1						
N=100						
	Theta=1			**Theta=2**		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox Model	56.4	24.2	119.1	31.0	13.3	65.5
WLS	41.2	18.1	87.9	22.9	9.8	48.3
Time-Dependent	47.4	20.8	98.9	25.8	11.0	54.4
Modified Cox	**38.7**	**17.0**	**82.2**	**21.4**	**9.2**	**45.2**
N=500						
	Theta=1			**Theta=2**		
Cox Model	38.2	16.4	80.7	21.0	9.0	44.4
WLS	28.0	12.3	59.6	15.5	6.7	32.8
Time-Dependent	32.1	14.1	67.1	17.5	7.5	36.9
Modified Cox	**26.3**	**11.5**	**55.7**	**14.5**	**6.2**	**30.6**

Note: *Authors Own Calculation, Sims stands for Simulation.*

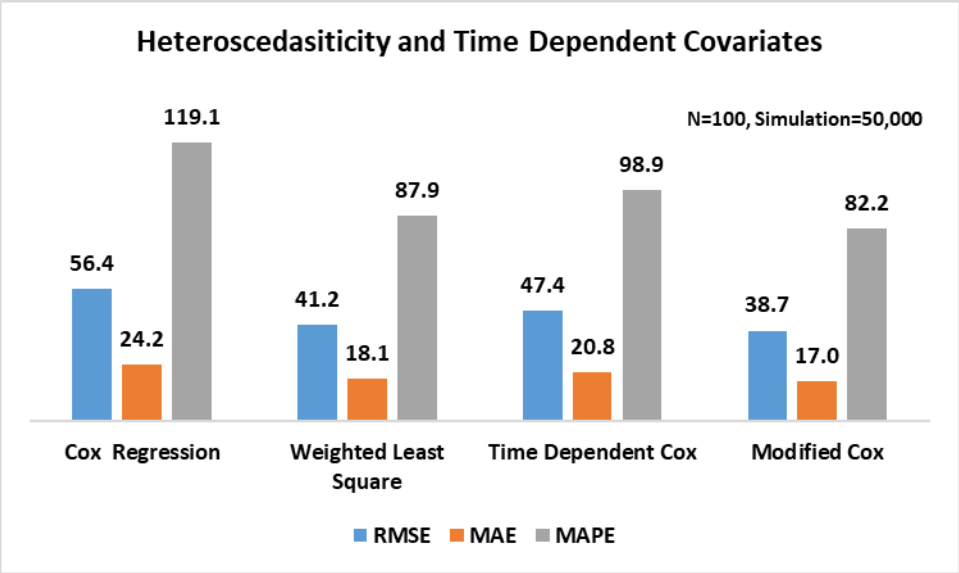


Figure 5. 4 Handling Heteroscedasticity and Time-Dependent Covariate

5.9 Final-Scenario Outlier, Heteroscedasticity, and TD covariate

The final scenario considers the DGP with different cases, such as 5%, 10%, and 20% outlier quantity, heteroscedasticity, and a time-dependent covariate case. The outlier in the data is observed using the influence plot; heteroscedasticity is observed in the model using the Brush Pagan test, and the dependent covariate is observed using the Shenfield residual ZPH test. In the outlier, heteroscedasticity, and time-dependent covariate case, it is crucial to estimate the desired effect of an independent variable through Cox regression. Thus, in different simulation scenarios and experiments, we understand the problem better. The detailed DGP is given in the previous chapter. Table 5.4 results suggest that modified Cox performed better in all cases, either 5%, 10%, or 20% outlier quantity or heteroscedasticity and time-dependent covariates.

Table 5.4 below further depicts the simulation finding in the case of an outlier, heteroscedasticity, and time-dependent covariate case for three different quantity cases. The result suggests that the outlier quantity slightly affects the results of all models, and an increase in outlier quantity increases the RMSE of all models. Still, the DGP used for variable generation is from an exponential distribution. Hence, an increase in theta makes the whole model better. The RMSE of the model is decreasing and converging towards zero, as we can see that the modified Cox model RMSE value is 49.3. If the theta value is one and theta value is two, the modified Cox model RMSE is 27.1, a significant decrease in RMSE value for the modified Cox model.

Figure 5.5 below shows that the modified Cox model has a low RMSE of 49.3, low MAE of 19.0, and low MAPE of 98.4 compared to the Cox model, Robust Cox, and Weighted Least Square. which means that in the case of an outlier, heteroscedasticity, and time-dependent covariates, the best model among the family of survival analysis is the modified Cox model, which handles the problem of an outlier, heteroscedasticity,

and time-dependent covariates jointly. The results are stable even if there is an increase or decrease in sample size or outlier quantity of 5%, 10%, or 20%; the exponential distribution theta increase slightly improved all model.

Figure 5.6 shows the parameter stability of the defined and estimated parameter differences. The modified Cox model outperformed and showed more stability and consistency toward the true parameter. The second best is the robust Cox model, and the worst model in all family of survival analysis is the Cox model regarding parameters stability and consistency.

Table 5. 4 Outlier, Hetero, and Time-Dependent Case

N=100, Sims=50,000, Parameter=1, Outlier with 4 SD						
5% Outlier						
	Theta=1			**Theta=2**		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox Model	73.5	28.3	146.7	40.4	15.6	80.7
Robust Cox	53.4	20.6	106.7	29.4	11.3	58.7
WLS	55.9	21.5	111.6	30.8	11.8	61.4
Time-Dependent	60.3	23.2	120.3	33.1	12.8	66.2
Modified Cox	**49.3**	**19.0**	**98.4**	**27.1**	**10.4**	**54.1**
10% Outlier						
Theta=1				**Theta=2**		
Cox Model	91.7	35.3	182.9	50.4	19.4	100.6
Robust Cox	66.6	25.7	133.0	36.6	14.1	73.1
WLS	69.7	26.9	139.2	38.4	14.8	76.6
Time-Dependent	75.2	28.9	150.0	41.3	15.9	82.5
Modified Cox	**61.5**	**23.7**	**122.7**	**33.8**	**13.0**	**67.5**
20% Outlier						
Theta=1				**Theta=2**		
Cox Model	129.3	49.8	258.1	71.1	27.4	142.0
Robust Cox	94.0	36.2	187.7	51.7	19.7	103.2
WLS	98.4	37.9	196.5	54.1	20.9	108.0
Time-Dependent	106.1	40.9	211.7	58.3	22.5	116.4
Modified Cox	**86.8**	**33.4**	**173.2**	**47.7**	**18.4**	**95.3**

Note: *Authors Own Calculation, Sims stands for Simulation.*

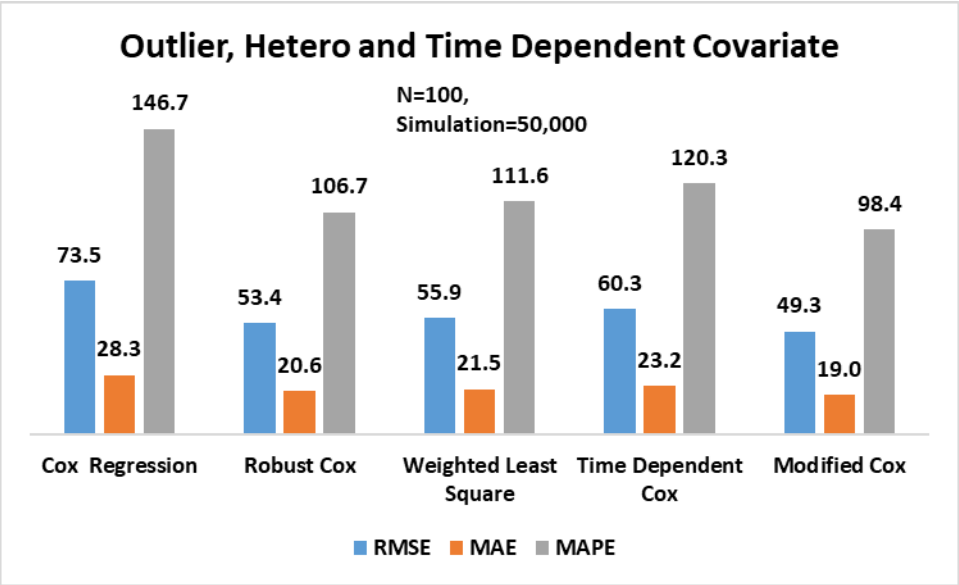


Figure 5. 5 Handling Outlier, Heteroscedasticity, and Time-Dependent covariates.

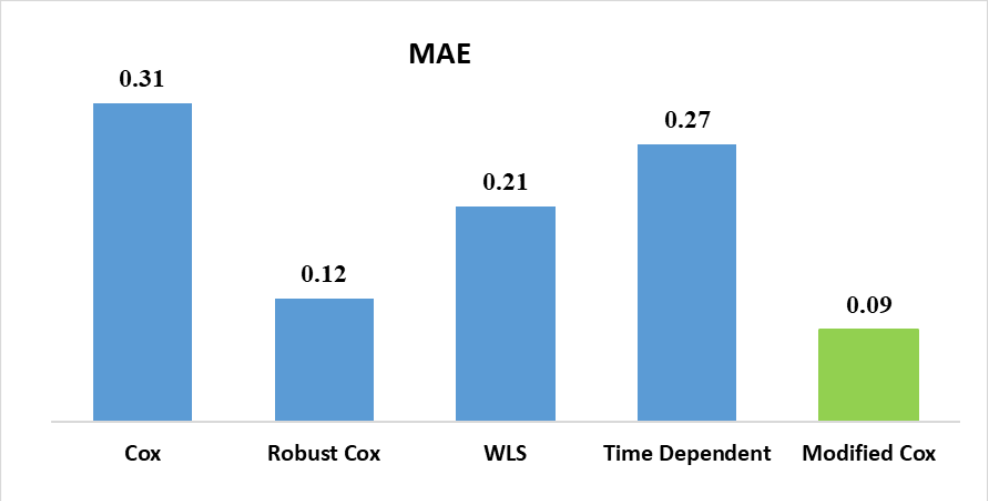


Figure 5. 6 Comparison of defined and estimated Parameter

Chapter 6

Real Data Application

Impact of major Injury on Labour Productivity and Return to Work: A case study of Pakistan.

6.1 Background

The real data analysis of the concerned variables and their results are illustrated in tables and figures. This chapter comprises the Introduction, Literature Review, Data source and variable definition, Correlation matrix, five different survival analysis results, and comparison based on RMSE, MAE, and MAPE.

6.2 Introduction

The connection between health and labor productivity is intrinsic. Healthy individuals tend to have higher productivity levels, reduced absenteeism, and improved cognitive function. Investments in employee health, including wellness programs and preventive measures, contribute to a more productive and engaged workforce, fostering a positive impact on overall economic output and prosperity. Injuries can significantly impact health and labor productivity. Workplace injuries lead to absenteeism, reduced efficiency, and increased healthcare costs. Investing in safety measures and employee well-being not only prevents injuries but also fosters a healthier workforce, enhancing overall productivity and organizational success.

Health is an important component of human capital, and its impact on economic growth is significant. Health is critical for economic growth because it has a direct impact on productivity, workforce participation, and human capital development. Absenteeism is reduced and labor-force efficiency is increased in a healthy population (Appleton and Teal, 1998). It encourages creativity and innovation, which leads to technological

advances and economic diversification. Investing in healthcare infrastructure also generates jobs and stimulates the economy (Tompa, 2022). Furthermore, healthier people are more likely to save and invest, resulting in higher savings rates and capital accumulation. By promoting well-being, societies can increase their potential for economic growth, create long-term prosperity, and reduce the burden on healthcare systems, freeing up resources for other productive investments (Strauss and Thomas, 1998). Labor productivity is important for economic growth because it increases efficiency, reduces costs, and increases output, resulting in higher incomes, innovation, and overall prosperity for nations and individuals (Hassan and Rehman, 2021; Rashid *et al.,* 2022).

The impact of serious injuries on labor productivity and return to work is a major concern around the world. Workplace injuries and accidents cause personal suffering as well as significant economic costs to individuals, families, and society (Ishaq *et al.,* 2022). Understanding the factors that influence injured workers' return-to-work outcomes is essential for developing effective policies and interventions to promote workforce productivity and social well-being (Butler *et al.,* 1995; Ghaffar *et al.,* 1999; Butler *et al.,* 2006). Shah and Ahmad (2016) investigate the effects of injuries on labor productivity as well as the barriers that people face when trying to re-enter the labor force in Pakistan.

In this study, we have investigated the effects of major injuries or diseases on labor productivity and return to work. What are the main challenges and issues that injured people in Pakistan face? What precautionary measures should be taken to reduce the time of major injuries so that they can return to work as soon as possible, increasing their productivity? Human capital will be improved, which will be benefit the

government and economy directly or indirectly because we cannot ignore the majority of injured or diseased people in society.

In addition, the researchers are also looking into how workplace factors like employment status and working hours help or hinder an injured worker's return to work (Vemer *et al.,* 2013). This case study's finding will contribute to the body of knowledge on the economic consequences of injuries in Pakistan, allowing policymakers, employers, and healthcare professionals to develop targeted interventions that promote better labor productivity and facilitate the successful reintegration of injured individuals into the workforce (Baldwin and Butler, 2006).

This research will help the government, policymakers, and health departments understand what factors must be considered to increase the likelihood of returning to work. Second, this study focuses on the proposed modified Cox model for time-to-event studies, which addresses outliers, heteroscedasticity, and time-dependent covariates simultaneously. A detailed literature review follows; Section 3 discusses data and methodology; Section 4 discusses the results and discussion; and Section 5 discusses the conclusion and policy recommendations.

6.3 Literature Review

After COVID-19, predicting the duration of illness and return to work (RTW) becomes more difficult (Aben *et al.,* 2023). However, psychosocial factors may play a role in addition to acute-phase symptoms and objective factors such as age and gender. Alimoradi *et al.,* (2021) conducted a more in-depth analysis of a subset of the data that looked at non-medical factors like job satisfaction and sleep disturbance.

Only sleep disturbance was found to be associated with delayed Return To Work (RTW), while job satisfaction did not show a significant association (Andersen *et al.,* 1995). The precise cause of these sleep-related issues is unknown, and whether they

contribute directly to the delay in RTW has not been confirmed. However, it is important to note that a significant proportion of COVID-19 patients have a notably high incidence of sleep problems (Aben *et al.,* 2023).

Age, education level, injury location, recorded origin of the injury, and department all have an impact on RTW, in addition to the type and severity of the injury. As a result, multiple interventions targeting these predictive factors are possible (Tamene *et al.,* 2022). Trauma has an impact on both the rate and timing of RTW. (Abedzadeh *et al.,* 2017) studied the transition from trauma to disability and discovered that subsequent RTW is associated with a variety of personal and clinical characteristics other than disability itself. Age, the severity of the disability, insurance coverage, and the quality of care received were all factors considered. As a result, exceptional care and indemnity coverage require extra thought and scrutiny.

Endo *et al.,* (2016) used a proportional hazard regression model to see if age, gender, and cancer site were statistically significant predictors of two outcomes: partial/full RTW and full RTW. Workers were classified into five groups: full and partial RTW, resigned, disabled, and deceased. Individuals who were absent due to illness for the entire 365-day study period are classified as disabled in this classification. The resigned and dead groups were treated as variables representing competing risks associated with RTW, implying that they could influence or hinder the likelihood of returning to work. Sickness absences are a significant financial burden on society in Finland, necessitating actions to promote long-term RTW and increased work participation. In the short term, legislative changes requiring employers to report prolonged SAs to Occupational Health Services and the national insurer improved RTW (Abedzadeh *et al.*, 2017). The impact of these legislative changes on the private sector must be assessed. Another study (Halonen *et al.,* 2016) investigated the consequences of legislative changes.

Employers are now legally required to report extended sick leave to occupational health services. Following the implementation of this legislation, both RTW and overall worker involvement increased significantly. There was also a significant effect for sudden RTW, which was greater in women than in men. Furthermore, those on sick leave due to mental syndromes have a greater impact than those with other diagnoses (Halonen *et al.,* 2018).

A longer period of absence is associated with a lower likelihood of RTW. (Vemer *et al.,* 2013) investigated the cost-effectiveness of accommodating care treatment for people on sick leave who have a major depressive disorder (MDD). Survival analyses were used to predict RTW, with both HRQoL and the severity of despair considered as factors. Female and adult patients required more time to RTW, and regular employees' decision-making latitude required additional time to RTW. Employees in higher levels of management, on the other hand, took longer to RTW. HRQoL was also discovered to be a significant predictor of how long people stayed at their jobs. In contrast, the severity of depression did not affect RTW duration.

Employee RTW programs must include appropriate procedures, clinical treatment, comfort design enhancements, healing, and social and economic support. While returning to work as soon as possible is critical, doing so too soon increases the risk of recurring and prolonged sick leave. Continuous collaboration among public health professionals, employees, and employers is required when evaluating individual cases. This collaborative approach ensures that rehabilitation is tailored to the individual's needs, resulting in successful and fluent RTW (Tamene *et al.,* 2022). Kamdar *et al* . (2020) investigated the survival of a previously employed person following a serious injury or illness. According to the study, 60% of employed laborers were out of work for one to five years after suffering a major injury or critical illness. (Kim *et al*., 2022)

investigated the socioeconomic factors that influence disability in Korea using data from the 2018 Korean Health Panel Survey. According to the study, people with higher incomes and more education are less likely to develop new disabilities.

6.3.1 Literature Gap

Firstly, the findings of a systematic review on returning to work (Kamdar *et al.,* 2020) indicate that the return to work after a major illness has been studied in many countries, but not in Pakistan, and this study is contributing to what factors may delay returning to work after major injuries. Secondly, this research helps to clarify which survival analysis models outperform in the presence of outliers, heteroscedasticity, and time-dependent covariates. Thirdly, the proposed modified Cox model for time-to-event studies simultaneously addresses outliers, heteroscedasticity, and time-dependent covariates.

6.4 Data Source for Real Data Application

The data source for understanding the importance of health for economic growth is the Pakistan Bureau of Statistics (PBS) Labour Force Survey. This survey gathers information on workforce participation, health-related research, and the overall impact of health on the nation's productivity and economic development.

6.5 Definition of Variable

The study makes use of the following Labour Force Survey variables: time in days, age, education, income, gender, region, marital status, employment status, and working hours. These variables allow researchers to examine the characteristics and productivity of the workforce across different demographics and regions in order to assess the relationship between returning to work after an accident or emergency injury.

6.5.1 Dependent Variable (Time in days)

In survival analysis, the dependent variable "time" refers to the number of days that pass between a specific starting point (e.g., diagnosis, treatment initiation) and an event of interest, such as death, relapse, or recovery. This analysis looks at the time-to-event outcomes of a survival-nature study.

In our study the origin of the event is Injury from an excess speed, the total number of observation from that specific injury in the data is 1120, and the event of interest is RTW or become normal.

6.5.2 Dummy (Censored Variable)

A dummy censored variable indicates whether an event of interest occurred if the observation is right-censored. When an event is observed, the dummy variable is set to 1; otherwise, it is set to 0 when the event is not observed but is known not to have occurred. Censoring occurs when the event time is unknown or incomplete.

6.5.3 Age

The "age of individual" or "respondent" in survival analysis refers to the time elapsed between their birth or enrollment in the study and the occurrence of an event of interest (e.g., death, disease onset). It is a critical time scale used in the context of survival to assess time-to-event outcomes (Plank *et al.,* 2008).

6.5.4 Education

Education is significant in survival analysis because it measures productivity and an individual's ability to return to work. Higher education is frequently linked to improved skills, knowledge, and adaptability, which leads to more job opportunities. People with a higher level of education are more likely to recover from medical setbacks and navigate labor-market challenges. In addition, education improves access to resources

and support systems, which influences a person's resilience and success in returning to work following a health-related disruption (Bruck *et al.,* 2011).

6.5.5 Monthly Income

Individual income is important in survival analysis because it influences their likelihood of returning to work. Higher-income levels provide financial security and access to resources, allowing for a faster recovery and reintegration into the labor force after a medical event. Lower income, on the other hand, may create barriers to an individual's ability to effectively resume work (Vemer *et al.,* 2013).

6.5.6 Gender

Because of gender-specific responsibilities and societal norms, an individual's gender is important in assessing their return to work in survival analysis. Women are frequently burdened with additional caregiving responsibilities and cultural expectations, which limit their ability to re-enter the labor force following a medical event. Following recovery, gender disparities in employment outcomes and productivity may exist (Abbas *et al.,* 2010).

6.5.7 Region

In survival analysis, the region or workplace, whether rural or urban, is a critical factor in determining an individual's ability to return to work. Urban areas typically have more job opportunities, better access to healthcare, and social support, which may aid in faster recovery and reintegration into the labor force. In contrast, rural areas may have fewer job opportunities and resources, which can affect return-to-work outcomes (Macran *et al.,* 1996; Skafida, 2012).

6.5.8 Marital Status

An individual's employment status is critical in determining their return-to-work outcomes in survival analysis. Employees may have job protection or benefits that

allow them to return to work following a medical event. Unemployed people may face additional barriers to re-employment, such as a lack of financial security and employer support (Abedzadeh *et al.,* 2017).

6.5.9 Employment Status

By employment status here we mean that either person is government fully employed or daily wagers or frictionally unemployed, the selected observation is taken for the employed person. Back-to-work outcomes in survival analysis depend on employment status. After a health event, employees may have employment protection or benefits to return to work. Unemployed people may struggle to re-enter the workforce due to financial insecurity and company assistance (Abedzadeh *et al.,* 2017).

6.5.10 Working Hour Week

In a survival analysis, a person's working hours significantly predict their ability to return to work. Individuals who work part-time or have flexible hours may benefit from resuming work gradually after a medical event. Those who work full-time or have demanding schedules, on the other hand, may struggle to balance work and recovery, affecting their return-to-work outcomes (Vemer *et al.,* 2013). The variables used in the study are Time in days, Education, Income, Gender, Age, Region, Marital Status, Employment Status, and Working Hours.

The sample size 1120 observation of those major accident who have injuries due to access speed, the origin of the event is time of accident, event of interest is return to work or back to normal situation. who have major injury due to access speed.

Table 6. 1 List of Variables and Definitions.

Variables	**Definition of Variable (Data Source: Labour Force Participation (LFS 2020-21)**
Sample Size	1120 observation for Major Injury due to excess speed, from the major injury/ disease section 8.1, (LFS 2020-21)
Time (Dependent)	The time variable in days, how much time an individual takes from injury to work, is the duration from origin to event of interest.
Event of Interest (Censored/Dummy)	The event of interest is back to work. If the individual goes back to work, equal 1. Otherwise, zero, if unknown, or continue to take rest.
Education	Education of individuals in years
Monthly Income	Monthly Income of individual in thousand rupees.
Marital Status	If Married=1, otherwise zero.
Gender	If male =1, otherwise zero.
Age (Independent)	Age of individual in years.
Region	Rural and Urban region, if Rural =1, otherwise zero
Employment Status	If fully employed=1, otherwise daily wagers or frictionally unemployed or any other=zero
Working Hour Week	Number of hours an individual works in a week.

Descriptive statistics for each variable are provided in Table 6.2. The first column contains the variable's name, the second column contains the observations, the third column contains the averages, the fourth column contains the standard deviation, the second last column contains the minimum, and the last column contains the maximum value of that variable; the total number of observations is 1120, the average time back to work is 12 days with a standard deviation of 28.4, the minimum is 1 and the maximum is 270 days, the monthly income reported below is in PKR, the income reported in the data was PKR, the same interpreter was used for both variables.

Table 6. 2 Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Time	1120	12.07	28.40	1	270
Age	1120	35.19	12.49	13	80
Education	1120	3.32	2.48	1	14
Income	1120	38044	6065	20000	50000
Work hour	1120	25.14	26.03	1	80

Table 6.3 shows detailed tabulations of each dummy variable. The first variable is gender, followed by employment status and a region dummy. The total number of observations was 1120, with 87% men and 13% women, 69% employed and 31% unemployed. The remaining 20% live in cities, while the remaining 80% live in rural areas.

Table 6. 3 Tabulation of gender Male, Employment Status, and Rural Region

	Gender Yes=Male		Emp_Status, yes=employed		Region Yes=Rural	
	Freq.	Percent	Freq.	Percent	Freq.	Percent
No	151	13.48	348	31.07	226	20.18
Yes	969	86.52	772	68.92	894	79.82
Total	1120	100.00	1120	100.00	1120	100.00

Table 6.4 displays the detailed correlation matrix for each variable. The variable's name appears in the first column, the correlation with time in the second, and the correlation with itself in the diagonal column, which equals one. Income is negatively related to time, whereas age, education, and working hours are positively related to time.

Table 6. 4 Correlation Matrix

Variables	(1)	(2)	(3)	(4)	(5)
(1) time	1.000				
(2) age	0.045	1.000			
(3) education	0.014	-0.137	1.000		
(4) Income	-0.005	-0.032	-0.022	1.000	
(5) Work hour	0.006	-0.025	0.911	-0.047	1.000

6.6 Outlier, Heteroscedasticity, and Time Dependent Detection.

Table 6.5 displays the outlier, hetero, and time-dependent covariate detection tests. An influence plot is used to identify data outliers and to detect heteroscedasticity in the model error term, we have used the Breush Pagan heteroscedasticity test. The Schoenfeld residuals ZPH test is used to detect time-dependent covariate variables. As a result, observations at point 356, 478, 503, 665, 1049, and 1099 are outliers, as shown in Figure 6.1 and table 6.5. The Breush Pagan test P value is less than 5%, indicating a heteroscedasticity problem; the Schoenfeld residuals ZPH test p-value indicates that the working hour is a time-dependent covariate that can be changed over time; and modified Cox regression is the only solution to simultaneously solve these three problems in the model of outlier, heteroscedasticity, and time-dependent covariate (Stevens, 1984).

6.6.1 Influence Plot for Outlier Detection

Influence plot is an alternative technique for detection of outlier, boxplots are effective at identifying the central tendency and spread of a distribution, including the presence of outliers. However, they may not explicitly show the influence of individual data points on statistical models. But Influence plots are specifically designed to highlight the influence of each data point on a statistical model. Influence plot further help to identify influential observations that may disproportionately impact regression coefficients, leverage, or other model diagnostics (Chernick and Murthey, 1983; Gray, 1989).

Influence Plot consists of Cook's D values, leverage hat values and Studentized residuals, the **Cook's Distance (D)** formula $D_i = \frac{r_i^2}{p \times \hat{\sigma}^2}$ Where r_i is the residual for

observation i, p is the number of predictors and $\hat{\sigma}$ is the standard error of the regression (Cooks and Weisberc, 1994).

Leverage formula $h_i = X_i(X^T X)^{-1} X_i^T$ Where X_i is the row vector for the ith observation .

Studentized Residuals: $t_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$, Where r_i is the residual for observations i, $\hat{\sigma}$ is the standard error of the residuals and h_i is the leverage $h_i = X_i(X^T X)^{-1} X_i^T$.

These formulas are combinedly visually used in the influence plot to identify observations that may disproportionally affect the regression model's results.

Table 6. 5 Outlier, Heteroscedasticity and Time-Dependent Detection.

InfluencePlot for Outlier		**Breush Pagan for Hetero**		**Schoenfeld residuals ZPH for Time Dependent**	
Observation	356,478,503, 665,1049,1099	BP Test	19.14	Hours_Week	4.088
		P Value	0.0018	P Value	0.043

Figure 6.1 below shows the influence plot for outlier detection. The bubble circle greater than three times of average value is considered an outlier because these values are three times greater the average value, so they are considered outliers. In this case, the observations 356, 478, 503, 665, 1049, and 1099 are outliers.

The size of each bubble in the influence plot is proportional to Cook's distance.

Cook's distance is a measure of how much the fitted values of the model would change if that particular observation were removed. Larger Cook's distance values indicate that the observation has a substantial influence on the regression coefficients.

Larger circles: Indicate observations with higher Cook's distance, suggesting they have a significant impact on the model's fit.

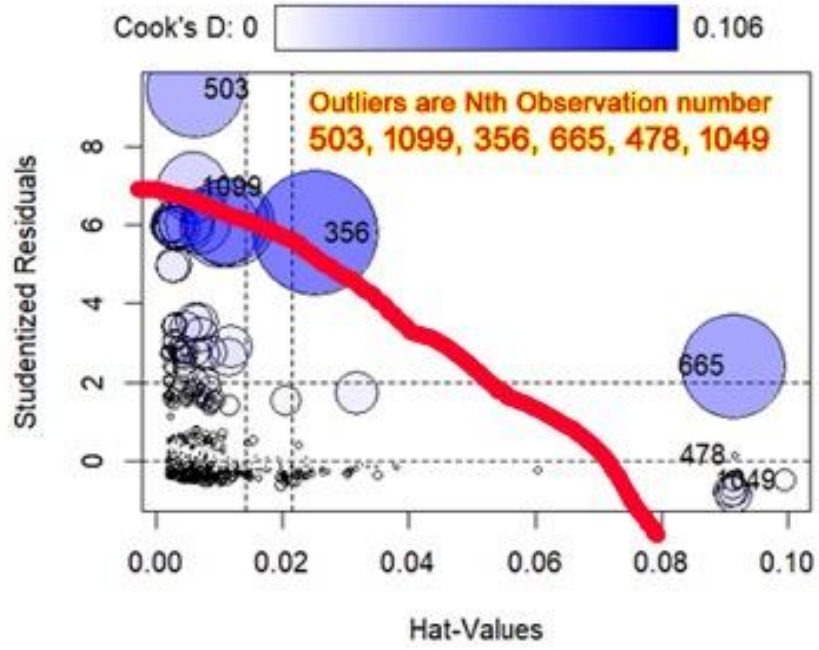


Figure 6. 1 Detection of Outlier

Figure 6.2 depicts the Kaplan-Meier plot for survival analysis. The Kaplan-Meier plot is a basic survival analysis visualization. A statistical method investigates the time it takes for an event of interest to occur. The event of interest is the return to work or recovery. In this graph, the x-axis represents analysis time and the y-axis represents estimated survival probability. The curve is depicted in Figure 6.2 as a series of step-like drops at each event occurrence.

Censored data points appear as vertical tick marks on the curve, representing observations that did not experience the event by the end of the study or were lost to follow-up. The graph may also include multiple curves, each representing a different subgroup, allowing for cohort comparisons. Interpreting Figure 6.2 entails assessing the shape, steepness of drops, and potential divergences between curves. The Kaplan-Meier graph is used to test hypotheses and model survival data, providing valuable insights

into the timing and likelihood of events in a variety of scenarios (Kaplan and Meier, 1958; Bland and Altman, 1998).

The Kaplan-Meier curve indicates the probability of an event occurring. The following study is interested in the recovery or return to work after a major injury. The vertical axis represents the likelihood of returning to work. The horizontal axis represents an individual's time spent. As a result, the likelihood of returning to work decreases over time. It could be a person losing their job or a serious injury. Another issue could arise. There is a negative relationship between time and recovery from major surgeries. As time passes, the amount of time needed to recover from work decreases.

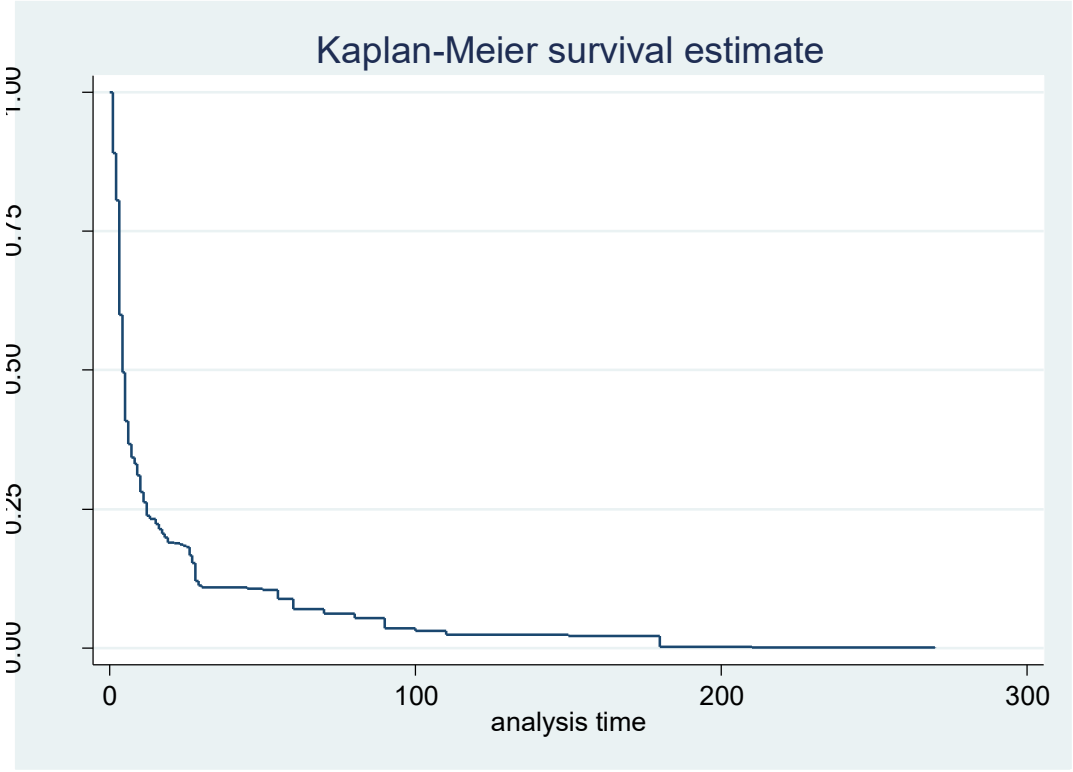


Figure 6. 2 Kaplan Meier Survival Function of Return to Work

The result below in Table 6.6 suggests that each additional year of education will increase HR= e(Coeff)=> e0.076≈1.079=>1.079-1=>0.079 =>7.9% return to work chances after a major accident, which is statistically significant at 5% in the Cox and

time-dependent model, while statistically significant at 10% in Robust Cox, WLS, and Modified Cox.

The income reported in the data was PKR, A thousand increase in monthly income will increase the chance of returning to work because he will do good medical treatment compared to a person with low income, so it might be that an increase in monthly income increases the return-to-work chances after a major accident or disease.

Marital Status If a person is married as compared to unmarried, or Male compared to female, average time back to work after a major accident or disease will increase, which is statistically significant at 5% in the Cox and time-dependent model, while statistically significant at a different level, star shows a different level of significance.

Results suggest that a year increase in age will decrease the chance of returning to work after a major accident. It might be that aged people take more time in recovery than young people.

Rural region and employment status If a person belongs to a rural area as compared to an urban, or employment status is employed as compared to daily wagers on the average chance of a return to work after the major accident will decrease because the rural region person might have low-quality medical treatment and for hospitals, which may decrease the chance of a return to work. The employed person has a stable earning source of income, so they will go back to work slowly compared to daily wagers.

The variable weekly hours is time-dependent covariates, which change from time to time, but weekly hours variable is not statistically significant.

Table 6. 6 Cox, Robust Cox, WLS Model, Time-Dependent Cox, and Modified Cox.

VARIABLES	(1) Cox Model	(2) Robust Cox	(3) WLS Model	(4) Time Dep Cox	(5) Modified Cox
Education	0.076**	0.031*	0.071*	0.067**	0.152*
	(0.033)	(0.017)	(0.039)	(0.031)	(0.079)
Monthly_Income	0.108*	0.164*	0.218*	0.015*	0.194*
	(0.061)	(0.095)	(0.124)	(0.009)	(0.101)
Marital_Status	0.155*	0.246*	0.167*	0.143	0.182**
	(0.081)	(0.146)	(0.101)	(0.102)	(0.079)
Gender(male)	0.453*	0.418*	0.447**	0.425*	0.491***
	(0.268)	(0.231)	(0.218)	(0.251)	(0.163)
Age	-0.003**	-0.005**	-0.007**	-0.004**	-0.217**
	(0.001)	(0.002)	(0.003)	(0.002)	(0.091)
Region(rural)	-0.311*	-0.352**	-0.364*	-0.314*	-0.312**
	(0.184)	(0.159)	(0.205)	(0.188)	(0.134)
Employment_Status	-0.185	-0.157	-0.171	-0.185	-0.215*
	(0.165)	(0.134)	(0.153)	(0.128)	(0.129)
Hour_week	-0.596	-0.418	-0.496	-0.581	-0.614
	(0.523)	(0.323)	(0.443)	(0.486)	(0.409)
Constant	7.220	8.458	9.401	8.246	5.951
	(7.804)	(6.157)	(7.294)	(6.861)	(5.748)
Observations	1,120	1,120	1,120	1,120	1,120
RMSE	11.1	6.2	11.3	11.2	**5.9**
MAE	9.2	4.3	9.5	9.3	**3.9**
MAPE	103.6	44.4	104.9	107.7	**41.8**

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Cox regression, robust cox, WLS model, time-dependent cox, and modified cox outperform the other four survival analysis models. This conclusion is based on lower RMSE, MAE, and MAPE values indicating better predictive accuracy. Modified Cox is the best choice among the family of survival analysis models due to its improved predictive capabilities. It is determined to be the best of the five models.

Table 6. 7 Comparison of five different Survival Analysis Models

Models	**RMSE**	**MAE**	**MAPE**	**Cumulative Sum**
Cox Model	11.1	9.2	103.6	123.9
Robust Cox	6.2	4.3	44.4	54.9
WLS	11.3	9.5	104.9	125.7
Time-Dependent	11.2	9.3	107.7	128.2
Modified Cox	**5.9**	**3.9**	**41.8**	**51.6**

When comparing Cox regression, robust Cox, WLS Model, time-dependent Cox, and modified Cox, modified Cox outperforms the others, as shown in Figure 6.3. This decision is based on the RMSE, MAE, and MAPE metrics in the figure 6.3, which indicate Modified Cox's superior predictive accuracy and make it the preferred model among the survival analysis family.

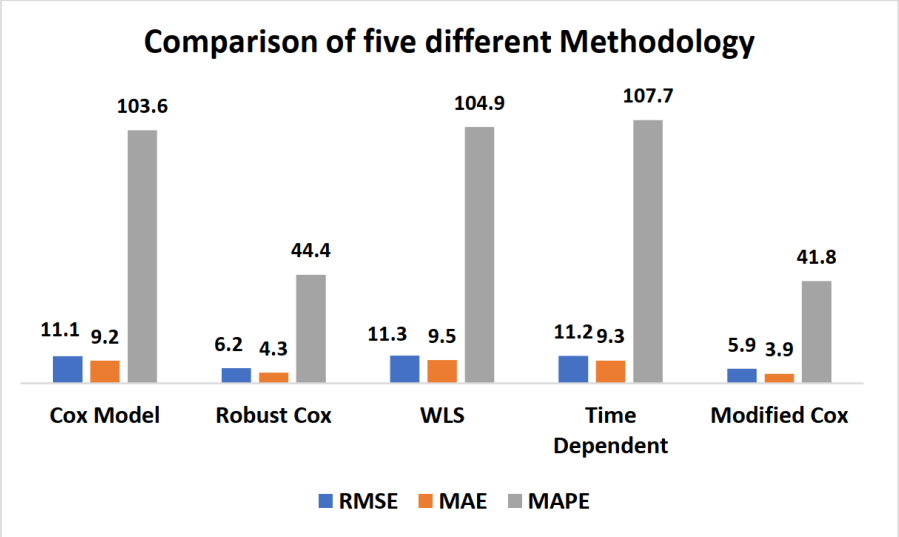


Figure 6. 3 Comparison of Different Methodology

Finally, a detailed comparison of five different survival analysis models—Cox Regression, Robust Cox, Weighted Least Squares, Time-Dependent Cox, and Modified Cox—uncovered important information about their effectiveness. A careful examination of the tables and figures provided reveals that, as demonstrated by the results, the modified Cox model outperforms the other four models. When several measures are carefully evaluated, such as the RMSE, MAE, and MAPE, the modified Cox model consistently outperforms the other four. It distinguishes itself from competitors by capturing complex variable interactions while maintaining computing efficiency. The modified Cox model can estimate survival data accurately and produce forecasts that are nearly as accurate as actual results due to its increased adaptability and responsiveness to different settings. Furthermore, the figures' visual representations contribute to comprehension of the model comparisons. Cumulative hazard plots and Kaplan-Meier curves from the Modified Cox model consistently show a better fit to the observed data and higher prediction accuracy than the other models in the outlier, heteroscedasticity, and time-dependent Cox models. This pattern is particularly visible

in different dataset subsets, confirming the Modified Cox model's functionality with its well-balanced combination of statistical rigor and real-world applicability.

The modified Cox model is a useful tool for survival analysis in a variety of research fields. Although each of the other models has advantages, such as outlier, heteroscedasticity, and the ability to include time-dependent covariate effects, the robust Cox model is better for outliers, the weighted least square is better for heteroscedasticity, the time-dependent Cox is better for time-dependent, and the modified Cox model is the best overall for outlier, heteroscedasticity, and time-dependent. Overall, the modified Cox model is the superior option, making it the preferred option for survival analysis.

Finally, in terms of predictability, adaptability, and accuracy, a thorough comparison of the five survival analysis models reveals that the modified Cox model outperforms the others. Because it is a powerful tool for gaining important insights from survival data; Education and monthly income, if the individual is male and married, would significantly increase the chance of returning to work, whereas age, rural area region, and government-employed individual would significantly decrease the chance of regaining work. The modified Cox model should be used by researchers and practitioners looking for a reliable method for survival analysis.

6.7 Study Limitation related to real life data

We don’t have survey data specifically on survival nature or time to event study nature but the Labour Force Survey (LFS) 2020-21, with a selected sample of 1,120 observations specifically focused on major injuries resulting from excessive speed. several limitations should be acknowledged: sample size, scope of variables, self-reported data, focus on major injuries only and we also claimed that we don’t have funds to collect data related to any specific disease for an ideal case.

Sample Size: The sample size of 1,120 observations, though sufficient for analysis, may not fully capture the broader population trends, especially in regions or demographics that are underrepresented in the data.

Scope of Variables: The LFS data may not include all relevant variables that could influence the occurrence of major injuries due to excessive speed, such as environmental conditions, road infrastructure, or vehicle type. This limits the study's ability to control for all potential confounders.

Self-Reported Data: The LFS data is primarily based on self-reported information, which can introduce biases such as underreporting or inaccurate recall of injury incidents. This may affect the reliability of the findings.

Focus on Major Injuries: The study specifically focuses on major injuries due to excessive speed, excluding minor injuries or other causes of accidents. This narrow focus may limit the ability to draw broader conclusions about traffic safety.

These limitations should be considered when interpreting the results of this study, as they may impact the generalizability and accuracy of the findings. Future research could address these limitations by utilizing larger, more comprehensive primary datasets to better capture the complexities of major injuries resulting from excessive speed.

Chapter 7

Conclusion, Limitations, and Future Directions

In econometrics and statistics, researchers are often interested in different potential covariates of the model, unbiased parameters, accurate survival time, and magnitude of the covariates, which are close to population parameters. The study proposed a new model for the Survival Analysis modelling, named the modified Cox model, for handling outliers, heteroscedasticity, and time-dependent covariates. Simultaneously we have a total of four major different scenarios and one real data implication. Four different scenarios are discussed in chapter five in detail. In chapter six we talk about real data application, now in chapter seven we talk about the conclusion, policy recommendation, and study limitation.

7.1 Conclusion

The proposed modified Cox model has improved the accuracy of survival predictions, unbiased estimates, and consistency of the survival analysis model is comparable to existing models in the presence of outliers, heteroscedasticity, and time-dependent covariates. The modified Cox model performed better in all scenarios except for the outlier and heteroscedasticity case. The robust Cox performs slightly better compared to the modified Cox model.

The modified Cox outperformed other models in the second scenario with the presence of outlier and time-dependent covariate case. The sample size increase slightly improves all model performance. The outlier quantity increases from 5% to 10% or 20%, slightly decreasing all model's performance. The parameter of exponential distribution theta value increases also improved all model performance.

In the third scenario, the presence of heteroscedasticity and time-dependent covariate case, the modified Cox performed better as compared to other models, the sample size and the parameter of exponential distribution theta value increased slightly improving all model performance.

In the fourth scenario, the presence of outlier, heteroscedasticity, and time-dependent covariate case, the modified Cox model performed better than other models. The sample size and the exponential distribution theta value increase slightly improved all model's performance.

After simulations of all the different scenarios, we applied the proposed modified Cox model to the real data application of labor force survey data, in which the event of interest was back to work from any major accident, in that study the modified Cox model also performed better with the RMSE value of 5.9, the Robust model RMSE is 6.2, the Cox model RMSE was 11.1, Time-Dependent Cox model RMSE was 11.2, and WLS model RMSE was 11.3, no significant difference in last three models.

Education, monthly income, gender and marital status, significantly increase the chance of a return to work because educated people take good care of themselves compared to non-educated people, also high-income individuals can take good care of themselves compared to low income. If the gender is male and the marital status is married, compared to female and unmarried individual, increase the likelihood of returning to work after a major accident or disease.

Age, rural region, and if an individual is employed significantly decrease the chance of a return to work because an increase in age might be an energy and stamina decrease. If a person belongs to a rural area, the facility is low compared to an urban area and thus decreases the chance of returning to work. Also, the employed person might be

given leave from the organization, reducing the possibility of returning to work relatively.

The conclusion from all the simulation and empirical data application studies shows that the modified Cox model performed better in all the families of the survival analysis model, in the case of outlier, heteroscedasticity, and time-dependent covariates, except the outlier and heteroscedasticity case, where the robust Cox model outperformed. The study concluded that the modified Cox model outperforms in the presence of outlier, heteroscedasticity, and time-dependent covariate. The modified Cox model is an improved version of the Cox model.

7.2 Policy Recommendation

The study suggested that a modified Cox model should be used in the presence of outlier, heteroscedasticity, and time-dependent covariate cases, whether we are estimating survival times at insurance companies, or returning to work from injury, getting a job after graduation, or any real data implications if the purpose is time to event study if there is an outlier, heteroscedasticity, and time-dependent covariate.

All health departments, organizations and practitioners need to use the latest, updated, and modified version of the Cox model in the field of time-to-event studies, which encompass the conventional existing survival analysis model.

The government needs to give proper training sessions at school, college, and university level related to road safety so that major accident decreases, and the productivity of human capital is stable or comparatively better.

The government further needs to increase the number of hospitals in rural areas to avoid any delay in the major injured individual, which will increase the chance of returning to work.

7.3 Study Limitation

The Labour Force Survey (LFS) 2020-21, with a selected sample of 1,120 observations specifically focused on major injuries resulting from excessive speed. While the data provides valuable insights, several limitations should be acknowledged: sample size, scope of variables, self-reported data and focus on major injuries only.

It can be used in many other real-world data applications due to the non-availability of such data, in which the origin and event of interest are defined, the time interval is defined (i.e., years, months, days or hours), we are limited to only one real-world application data, It will be best if it is applied to many more real-world applications such as getting a job after graduation, back to work after HIV/AIDS or any time to event study based on the availability of data.

7.4 Suggestion for Future Research

In the current dissertation, we have covered the problem of outliers, heteroscedasticity, and time-dependent covariates. Still, it can be further extended to find the optimal number of outliers to become winsorized.

The modified Cox model can be further extended to find the optimal number of covariates in Survival Analysis. Further, this research can be expanded to find the optimal number of covariates using advanced machine learning techniques in case of high dimensional data, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest (RF), and address the different problem simultaneously.

The modified Cox model algorithm exhibits a broader range of applicability in several domains, including the medical field, economics, engineering, financial, and social sciences. It enables the evaluation of the performance of both conventional approaches and the modified Cox model in real-world data applications.

References

Aalen, O. O., Borgan, Ø., Gjessing, H. K. J. S., and View, E. H. A. A. P. P. o. (2008). An introduction to survival and event history analysis. 1-39.

Lin, D. Y., & Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. Journal of the American Statistical Association, 84(408), 1074-1078.

Therneau, T. M., & Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer Science & Business Media.

Wei, L. J. (1992). The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis. Statistics in Medicine, 11(14-15), 1871-1879.

Abbas, Q., Hameed, A., and Waheed, A. (2010). Gender discrimination and its affect on employee performance/productivity. *Managerial and Entrepreneurial Developments in the Mediterranean Area, 1*(15), 170-176.

Abedzadeh-Kalahroudi, M., Razi, E., Sehat, M., and Asadi-Lari, M. (2017). Return to work after trauma: A survival analysis. *Chinese journal of traumatology, 20*(02), 67-74.

Aben, B., Kok, R. N., and de Wind, A. (2023). Return-to-work rates and predictors of absence duration after COVID-19 over the course of the pandemic. *Scand J Work Environ Health*, *4077*.

Adil, I. H. (2015). A modified approach for detection of outliers. Pakistan Journal of Statistics and Operation Research, 91-102.

Agampodi, S. B., Agampodi, T. C., and Piyaseeli, U. K. D. (2007). Breastfeeding practices in a public health field practice area in Sri Lanka: a survival analysis. *International Breastfeeding Journal, 2*(1), 1-7.

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, *12*(7), e0181001.

Akram, M., Ullah, M. A., and Taj, R. J. P. V. J. (2007). Survival analysis of cancer patients using parametric and non-parametric approaches. *27*(4), 194.

Ali, S., Ali, S., Shah, I., Siddiqui, G. F., Saba, T., and Rehman, A. J. I. A. (2020). Reliability analysis for electronic devices using a generalized exponential distribution. *8*, 108629-108644.

Alimoradi, Z., Broström, A., Tsang, H. W., Griffiths, M. D., Haghayegh, S., Ohayon, M. M., ... and Pakpour, A. H. (2021). Sleep problems during COVID-19 pandemic and its' association to psychological distress: A systematic review and meta-analysis. *EClinicalMedicine*, *36*.

Al-Kutubi, H. S., & Ibrahim, N. A. (2009). On the estimation of survival function and parameter exponential life time distribution. Journal of Mathematics and Statistics, 5(2), 130.

Altonen, B. L., Arreglado, T. M., Leroux, O., Murray-Ramcharan, M., and Engdahl, R. (2020). Characteristics, comorbidities and survival analysis of young adults hospitalized with COVID-19 in New York City. *PloS one, 15*(12), e0243343.

Ameri, S., Fard, M. J., Chinnam, R. B., and Reddy, C. K. (2016). *Survival analysis-based framework for early prediction of student dropouts.* Paper presented at the Proceedings of the 25th ACM International Conference on Information and Knowledge Management.

Anastasopoulos, P. C., Mannering, F. L., Shankar, V. N., and Haddock, J. E. (2012). A study of factors affecting highway accident rates using the random-parameters Tobit model. *Accident Analysis and Prevention, 45*, 628-633.

Andersen, P. K., and Vaeth, M. (1995). *Survival analysis*: Københavns Universitet. Department of Biostatistics.

Anjullo, B. B. (2018). A Simulation Study to Evaluate the Performance of Extended Cox model in Testing Treatment Effect with Possible Non-proportional Hazards. *International Journal of Progressive Sciences and Technologies, 10*(2), 284-293.

Annesi, I., Moreau, T., and Lellouch, J. (1989). Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Statistics in medicine, 8*(12), 1515-1521.

Appleton, S., and Teal, F. (1998). *Human capital and economic development*. African Development Bank Group.

Arsyad, R., Thamrin, S., and Jaya, A. (2019). *Extended Cox model for breast cancer survival data using Bayesian approach: A case study.* Paper presented at the Journal of Physics: Conference Series.

Ata, N., and Sözer, M. T. (2007). Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics, 36*(2), 157-167.

Baik, S. H., Fung, K.-W., and McDonald, C. J. (2021). The Mortality Risk of Proton Pump Inhibitors in 1.9 Million US Seniors: An Extended Cox Survival Analysis. *Clinical Gastroenterology and Hepatology*.

Baldwin, M. L., and Butler, R. J. (2006). Upper extremity disorders in the workplace: costs and outcomes beyond the first return to work. *Journal of occupational rehabilitation, 16*, 296-316.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, *79*(1), 139-147.

Bland, J. M., and Altman, D. G. (1998). Survival probabilities (the Kaplan-Meier method). *Bmj, 317*(7172), 1572-1580.

Bogaerts, K., Komárek, A., and Lesaffre, E. (2017). *Survival analysis with interval-censored data: A practical approach with examples in R, SAS, and BUGS*: Chapman and Hall/CRC.

Breusch, T. S., and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, 1287-1294.

Brück-Klingberg, A., Burkert, C., Garloff, A., Seibert, H., and Wapler, R. (2011). *Does higher education help immigrants find a job? A survival analysis* (No. 6/2011). IAB-Discussion Paper.

Burton, M., Rigby, D., and Young, T. (2003). Modelling the adoption of organic horticultural technology in the UK using duration analysis. *Australian Journal of Agricultural and Resource Economics, 47*(1), 29-54.

Butler, R. J., Baldwin, M. L., and Johnson, W. G. (2006). The effects of occupational injuries after returns to work: Work absences and losses of on-the-job productivity. *Journal of Risk and Insurance, 73*(2), 309-334.

Butler, R. J., Johnson, W. G., and Baldwin, M. L. (1995). Managing work disability: why first return to work is not a measure of success. *ILR Review, 48*(3), 452-469.

Carrasquinha, E., Veríssimo, A., and Vinga, S. J. b. (2018). Consensus outlier detection in survival analysis using the rank product test. 421917.

Chaudhry, K. A., Jamil, F., Razzaq, M., & Jilani, B. F. (2018). Survival analysis of dengue patients of Pakistan. *Headache*, *83*, 0-00.

Chebyshev, P. L. (1882). O priblizennyh vyrazenijah odnih integralov cerez drugie, Soobšcenija i protokoly zasedani Matemmaticeskogo obcestva pri Imperatorskom Har'kovskom Universitete, No. 2, 93–98; Polnoe sobranie socineni PL Chebyshev. *Moskva-Leningrad, 1948a*, 128-131.

Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and mathematical methods in medicine, 2013*.

Chernick, M. R., & Murthy, V. K. (1983). The use of influence functions for outlier detection and data editing. *American Journal of Mathematical and Management Sciences*, *3*(1), 47-61.

Collyer, M. L., Sekora, D. J., and Adams, D. C. J. H. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *115*(4), 357-365.

Cook, R. D., & Weisberc, S. Influence and Outliers. An Introduction to Regression Graphics, 203.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-202.

Deo, S. V., Deo, V., & Sundaram, V. (2021). Survival analysis—part 2: Cox proportional hazards model. *Indian journal of thoracic and cardiovascular surgery*, *37*, 229-233.

Ediebah, D., Coens, C., Zikos, E., Quinten, C., Ringash, J., King, M., . . . Flechtner, H. (2014). Does change in health-related quality of life score predict survival? Analysis of EORTC 08975 lung cancer trial. *British journal of cancer, 110*(10), 2427-2433.

Efficace, F., Bottomley, A., Coens, C., Van Steen, K., Conroy, T., Schöffski, P., . . . Köhne, C.-H. (2006). Does a patient's self-reported health-related quality of life predict survival beyond critical biomedical data in advanced colorectal cancer? *European Journal of Cancer, 42*(1), 42-49.

Emmerson, J., & Brown, J. M. (2021). Understanding survival analysis in clinical trials. *Clinical Oncology*, *33*(1), 12-14.

Emoru, K. E., Chelule, J. C., Imboga, H., and Ayubu, A. O. (2020). Extended Cox Model on Duration Taken to Release Cargo at Kenyan Border Entry Point: A Case Study of Malaba Osbp.

Ershadi, R., Rafieian, S., Salehi, M., Kazemizadeh, H., Amini, H., Sohrabi, M., ... & Vahedi, M. (2023). COVID-19 and spontaneous pneumothorax: a survival analysis. *Journal of Cardiothoracic Surgery*, *18*(1), 211.

Fisher, L. D., and Lin, D. Y. J. A. r. o. p. h. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *20*(1), 145-157.

Frost, J. (2021). Chebyshev's theorem in statistics. *URL https://statisticsbyjim. com/basics/chebyshevs-theorem-in-statistics*.

Gémar, G., Moniche, L., and Morales, A. J. (2016). Survival analysis of the Spanish hotel industry. *Tourism Management, 54*, 428-438.

Ghadimi, M., Mahmoodi, M., Mohammad, K., Zeraati, H., Hosseini, M., and Sheikh Fathollahi, M. (2010). Comparison of survival analysis of gastrointestinal cancer patients using parametric and Cox models. *Journal of School of Public Health and Institute of Public Health Research, 8*(2).

Ghaffar, A., Hyder, A. A., Mastoor, M. I., and Shaikh, I. (1999). Injuries in Pakistan: directions for future health policy. *Health policy and planning, 14*(1), 11-17.

Gong, Q., and Schaubel, D. E. (2018). Tobit regression for Modelling mean survival time using data subject to multiple sources of censoring. *Pharmaceutical statistics, 17*(2), 117-125.

Gray, J. B. (1989). The four-measure influence plot. *Computational Statistics & Data Analysis*, *8*(2), 179-188.

Greten, T. F., Papendorf, F., Bleck, J. S., Kirchhoff, T., Wohlberedt, T., Kubicka, S., ... & Manns, M. P. (2005). Survival rate in patients with hepatocellular carcinoma: a retrospective analysis of 389 patients. *British journal of cancer*, *92*(10), 1862-1868.

Gujarati, D. N. (2022). *Basic econometrics*. Prentice Hall.

Guo, S. (2010). *Survival analysis*. Oxford University Press.

Györffy, B., and Schäfer, R. (2009). Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. *Breast cancer research and treatment, 118*(3), 433-441.

Győrffy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., and Szallasi, Z. (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment, 123*(3), 725-731.

Halonen, J. I., Solovieva, S., Pentti, J., Kivimäki, M., Vahtera, J., and Viikari-Juntura, E. (2016). Effectiveness of legislative changes obligating notification of prolonged sickness absence and assessment of remaining work ability on return to work and work participation: a natural experiment in Finland. *Occupational and environmental medicine*, *73*(1), 42-50.

Halonen, J. I., Solovieva, S., Virta, L. J., Laaksonen, M., Martimo, K. P., Hiljanen, I., ... and Viikari-Juntura, E. (2018). Sustained return to work and work participation after a new legislation obligating employers to notify prolonged sickness absence. *Scandinavian journal of public health*, *46*(19_suppl), 65-73.

Hanif, A., Butt, A., Ahmed, A., Sajid, M. R., Ashraf, T., & Nawaz, A. A. (2015). Survival analysis of Dengue patients in relation to severity of liver dysfunction in Pakistan. *Adv Biolog Res*, *9*, 91-94.

Hashemian, A., Garshasbi, M., Pourhoseingholi, M., and Eskandari, S. (2017). A comparative study of Cox vs. log-logistic regression (with and without its frailty) in estimating patients' colorectal cancer survival time. *Journal of Medical and Biomedical Sciences, 6*(1), 35-43.

Hassan, N., and ur Rehman, A. (2021). Examining the Inter-Sectoral Relationship, Productivity and Inclusive Growth of Pakistani and Indonesian Economies. *iRASD Journal of Economics*, *3*(1), 38-57.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Horel, A. (1962). Applications of ridge analysis toregression problems. *Chem. Eng. Progress., 58*, 54-59.

Husain, H., Thamrin, S. A., Tahir, S., Mukhlisin, A., and Apriani, M. M. (2018). *The application of extended Cox proportional hazard method for estimating survival time of breast cancer.* Paper presented at the Journal of physics: Conference series.

Ishaq, A., Sadaf, M., Ali, A., and Naz, S. (2022). Imagining the Growth in Small and Medium Enterprises (SMEs) of Pakistan under COVID19 Outbreak. *iRASD Journal of Economics*, *4*(4), 583-593.

Jiao, X. (2019). Essays on asymptotics of outlier detection algorithms with applications to economics (Doctoral dissertation, University of Oxford).

Jung, S. Y., Papp, J. C., Sobel, E. M., and Zhang, Z.-F. (2019). Post Genome-Wide Gene–Environment Interaction Study Using Random Survival Forest: Insulin Resistance, Lifestyle Factors, and Colorectal Cancer Risk. *Cancer Prevention Research, 12*(12), 877-890.

Jung, S. Y., Papp, J. C., Sobel, E. M., Yu, H., and Zhang, Z.-F. (2019). Breast Cancer Risk and Insulin Resistance: Post Genome-Wide Gene–Environment Interaction Study Using a Random Survival Forest. *Cancer research, 79*(10), 2784-2794.

Kamdar, B. B., Suri, R., Suchyta, M. R., Digrande, K. F., Sherwood, K. D., Colantuoni, E., ... and Hopkins, R. O. (2020). Return to work after critical illness: a systematic review and meta-analysis. *Thorax*, *75*(1), 17-27.

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical Association, 53*(282), 457-481.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of economic literature, 26*(2), 646-679.

Kim, T., Park, S. Y., and Oh, I. H. (2022). Health-related factors leading to disabilities in Korea: Survival analysis. *Frontiers in Public Health*, *10*, 1048044.

Kojimahara, N., and Yamaguchi, N. (2016). Returning to work after sick leave due to cancer: a 365-day cohort study of Japanese cancer survivors. *Journal of Cancer Survivorship*, *10*, 320-329.

Konishi, T., Tanabe, M., Michihata, N., Matsui, H., Nishioka, K., Fushimi, K., ... & Yasunaga, H. (2023). Risk factors for arm lymphedema following breast cancer surgery: a Japanese nationwide database study of 84,022 patients. *Breast Cancer*, *30*(1), 36-45.

Krivtsov, V. Reliability Models Evolution: from Survival Regression to Deep Survival.

Kumchulesi, G., Palamuleni, M., and Kalule-Sabiti, I. (2011). *Factors affecting age at first marriage in Malawi.* Paper presented at the InSixth African Population Conference, Ouagadougou-Burkina Faso.

Kundu, S., Chauhan, K., and Mandal, D. (2021). Survival Analysis of Patients With COVID-19 in India by Demographic Factors: Quantitative Study. *JMIR Formative Research, 5*(5), e23251.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 139-156.

Lánczky, A., & Győrffy, B. (2021). Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. *Journal of medical Internet research*, *23*(7), e27633.

Lauss, M., Kriegner, A., Vierlinger, K., Visne, I., Yildiz, A., Dilaveroglu, E., and Noehammer, C. (2008). Consensus genes of the literature to predict breast cancer recurrence. *Breast cancer research and treatment, 110*(2), 235-244.

Lelisho, M. E., Teshale, B. M., Tareke, S. A., Hassen, S. S., Andargie, S. A., Merera, A. M., & Awoke, S. (2023). Modeling survival time to death among TB and HIV co-infected adult patients: An institution-based retrospective cohort study. *Journal of Racial and Ethnic Health Disparities*, *10*(4), 1616-1628.

Lin, D. Y., and Wei, L.-J. J. J. o. t. A. s. A. (1989). The robust inference for the Cox proportional hazards model. *84*(408), 1074-1078.

Macran, S., Joshi, H., and Dex, S. (1996). Employment after childbearing: a survival analysis. *Work, Employment and Society, 10*(2), 273-296.

Matida, L. H., Ramos Jr, A. N., Moncau, J. E. C., Marcopito, L. F., Marques, H. H. d. S., Succi, R. C. M., . . . Hearst, N. J. C. d. s. p. (2007). AIDS by mother-to-child transmission: survival analysis of cases followed from 1983 to 2002 in different regions of Brazil. *23*, S435-S444.

Mayr, A., Hofner, B., Waldmann, E., Hepp, T., Meyer, S., and Gefeller, O. (2017). An update on statistical boosting in biomedicine. *Computational and mathematical methods in medicine, 2017*.

Mendez-Gonzalez, L. C., Rodríguez-Borbón, M. I., Piña-Monarrez, M. R., Ambrosio, R., and Del Valle, A. (2016). Reliability analysis for laptop computer under

electrical harmonics. *Quality and Reliability Engineering International, 32*(8), 2945-2960.

Michael, R. D. (2022). Treatment profile and survival analysis acute respiratory distress syndrome (ARDS) COVID-19 patients. *International Journal of Applied Pharmaceutics*, *14*.

Min, Y., Zhang, G., Long, R. A., Anderson, T. J., and Ohland, M. W. (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering students. Journal of Engineering Education, 100(2), 349-373.

Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., and Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports, 11*(1), 1-13.

Moreno-Betancur, M., Carlin, J. B., Brilleman, S. L., Tanamas, S. K., Peeters, A., and Wolfe, R. J. B. (2018). Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modelling (MIJM). *19*(4), 479-496.

Muhammadullah, S., Urooj, A., Mengal, M. H., Khan, S. A., & Khalaj, F. (2022). Cross-Sectional Analysis of Impulse Indicator Saturation Method for Outlier Detection Estimated via Regularization Techniques with Application of COVID-19 Data. *Computational and Mathematical Methods in Medicine*, *2022*(1), 2588534.

Muhammad, M., and Yuwaningsih, D. (2019). *Estimating Survival Time of Dengue Haemorrhagic Fever Using Extended Cox Model.* Paper presented at the Journal of Physics: Conference Series.

Mustefa, Y. A., and Chen, D.-G. (2021). Accelerated failure-time model with weighted least-squares estimation: application on survival of HIV positives. *Archives of Public Health, 79*(1), 1-7.

Nemati, M., Ansary, J., and Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns, 1*(5), 100074.

Pang, L., Lu, W., and Wang, H. J. (2015). Local Buckley-James estimation for heteroscedastic accelerated failure time model. *Statistica Sinica, 25*, 863.

Panjer, H. H. J. A. (1987). AIDS: Survival analysis of persons testing HIV. *6*, 9.

Patel, K. K., Rai, R., & Rai, A. K. (2021). Determinants of infant mortality in Pakistan: evidence from Pakistan Demographic and Health Survey 2017–18. *Journal of Public Health*, *29*, 693-701.

Pawar, A., Chowdhury, O. R., & Salvi, O. (2022). A narrative review of survival analysis in oncology using R. *Cancer Research, Statistics, and Treatment*, *5*(3), 554-561.

Pearce, T., Jeong, J. H., & Zhu, J. (2022). Censored quantile regression neural networks for distribution-free survival analysis. *Advances in Neural Information Processing Systems*, *35*, 7450-7461.

Petry, S., and Tutz, G. (2011). The OSCAR for generalized linear models.

Plank, S. B., DeLuca, S., and Estacion, A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education, 81*(4), 345-370.

Prentice, R. L., and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 57-67.

Rama, D., & Andrews, J. D. (2013). A reliability analysis of railway switches. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of rail and rapid transit, 227(4), 344-363.

Rashid, R. N., Saeed, M. K., and Ali, H. (2022). Impact of human capital on poverty reduction in Pakistan. *iRASD Journal of Economics*, *4*(3), 419-428.

Ratnaningsih, D. J., Saefuddin, A., and Kurnia, A. (2021). Stratified-extended Cox with frailty model for non-proportional hazard: A statistical approach to student retention data from Universitas Terbuka in Indonesia. *Thailand Statistician, 19*(1), 209-228.

Ratnaningsih, D. J., Saefuddin, A., Kurnia, A., and Mangku, I. W. (2020). The terminology of survival Modelling: An insight and alternative Modelling of student retention.

Ratnaningsih, D., Saefuddin, A., Kurnia, A., and Mangku, I. (2019). *Stratified-extended Cox model in survival Modelling of non-proportional hazard.* Paper presented at the IOP Conference Series: Earth and Environmental Science.

Riu, J., and Bro, R. (2003). Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Labouratory Systems*, *65*(1), 35-49.

Roshanaei, G., Safari, M., Faradmal, J., Abbasi, M., and Khazaei, S. (2020). Factors affecting the survival of patients with colorectal cancer using random survival forest. *Journal of Gastrointestinal Cancer*, 1-8.

Salerno, S., & Li, Y. (2023). High-dimensional survival analysis: Methods and applications. *Annual review of statistics and its application*, *10*, 25-49.

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martínez-Valverde, S., Toledano-Toledano, F., and Garduño-Espinosa, J. (2020). A survival analysis of COVID-19 in the Mexican population. *BMC Public Health, 20*(1), 1-8.

Scheike, T. H., and Zhang, M.-J. J. J. o. s. s. (2011). Analyzing competing risk data using the R timereg package. *38*(2).

Schober, P., Vetter, T. R. J. A., and analgesia. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *127*(3), 792.

Shah, H., Ahmad, I., and Head, P. I. D. E. Impact of HIV/AIDS on Labour Productivity and Return to Work in Khyber Pakhtunkhwa Province.

Shahraki, H. R., Salehi, A., and Zare, N. (2015). Survival prognostic factors of male breast cancer in Southern Iran: a LASSO-Cox regression approach. *Asian Pacific journal of cancer prevention, 16*(15), 6773-6777.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software, 39*(5), 1.

Skafida, V. (2012). Juggling work and motherhood: the impact of employment and maternity leave on breastfeeding duration: a survival analysis on Growing Up in Scotland data. *Maternal and child health journal, 16*, 519-527.

Solomon, G. T., Bryant, A., May, K., and Perry, V. J. T. (2013). Survival of the fittest: Technical assistance, survival and growth of small businesses and implications for public policy. *33*(8-9), 292-301.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological bulletin*, *95*(2), 334.

Strauss, J., and Thomas, D. (1998). Health, nutrition, and economic development. *Journal of economic literature*, *36*(2), 766-817.

Tamene, A., Habte, A., Derilo, H. T., Endale, F., Gizachew, A., Sulamo, D., and Afework, A. (2022). Time to return to work after an occupational injury and its prognostic factors among employees of large-scale metal manufacturing facilities in Ethiopia: a retrospective cohort. *Environmental health insights*, *16*, 11786302221109372.

Therneau, T., Crowson, C., and Atkinson, E. J. S. V. (2017). Using time dependent covariates and time dependent coefficients in the Cox model. *2*, 3.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine, 16*(4), 385-395.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24-36.

Tompa, E. (2002). The impact of health on productivity: empirical evidence and policy implications.

Urooj, A., and Asghar, Z. (2017). Analysis of the performance of test statistics for detection of outliers (additive, innovative, transient, and level shift) in AR (1) processes. *46*(2), 948-979.

Vemer, P., Bouwmans, C. A., Zijlstra-Vlasveld, M. C., van der Feltz-Cornelis, C. M., and Hakkaart-van Roijen, L. (2013). Let's get back to work: survival analysis on the return-to-work after depression. *Neuropsychiatric Disease and Treatment*, 1637-1645.

Verweij, P. J., and Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in medicine, 13*(23-24), 2427-2436.

Vinzamuri, B., and Reddy, C. K. (2013). *Cox regression with correlation based regularization for electronic health records.* Paper presented at the 2013 IEEE 13th International Conference on Data Mining.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR), 51*(6), 1-36.

White, H. J. E. j. o. t. E. S. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. 817-838.

Widodo, A., and Yang, B.-S. (2011). Machine health prognostics using survival probability and support vector machine. *Expert Systems with Applications, 38*(7), 8430-8437.

Witten, D. M., and Tibshirani, R. J. S. m. i. m. r. (2010). Survival analysis with high-dimensional covariates. *19*(1), 29-51.

Yu, L., Liu, L., and Chen, D.-G. J. C. S. (2019). A homoscedasticity test for the accelerated failure time model. *34*(1), 433-446.

Yusuf, M. A., Badar, F., Meerza, F., Khokhar, R. A., Ali, F. A., Sarwar, S., and Faruqui, Z. S. J. A. P. J. o. C. P. (2007). Survival from hepatocellular carcinoma at a cancer hospital in Pakistan. *8*(2), 272.

Zaman, Q., and Pfeiffer, K. P. J. I. (2011). Survival analysis in medical research. *17*(4), 1-36.

Zhang, Z., and Sun, J. (2010). Interval censoring. *Statistical methods in medical research, 19*(1), 53-70.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., and Groothuis-Oudshoorn, C. G. J. A. o. t. m. (2018). Time-varying covariates and coefficients in Cox regression models. *6*(7).

Appendix

Appendix A Results of Other Different Scenarios

Scenario A Outlier and Hetero

Table A. 1 Outlier and Hetero Case, Sample N=100, Outlier 4 SD

N=100, Sims=50,000, Outlier with 4 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	51.7	20.4	104.7	28.4	11.2	57.6	83.2	32.8	168.6	44.9	17.7	91.0
Robust Cox	**32.0**	**13.5**	**69.1**	**18.8**	**7.4**	**38.0**	**54.9**	**21.7**	**111.3**	**29.7**	**11.7**	60.1
WLS	42.9	16.9	86.9	23.6	9.3	47.8	69.1	27.3	139.9	37.3	14.7	75.6
Modified Cox	34.6	13.9	71.2	19.3	7.6	39.2	56.6	22.3	114.6	30.6	12.1	61.9
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	62.6	24.7	126.7	34.4	13.6	69.7	100.7	39.7	204.0	54.4	21.5	110.1
Robust Cox	**38.7**	**16.3**	**83.6**	**22.7**	**9.0**	**46.0**	**66.5**	**26.2**	**134.6**	**35.9**	**14.2**	72.7
WLS	51.9	20.5	105.2	28.6	11.3	57.8	83.6	33.0	169.3	45.1	17.8	91.4
Modified Cox	41.9	16.8	86.1	23.4	9.2	47.4	68.5	27.0	138.7	37.0	14.6	74.9
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	80.1	31.6	162.3	44.1	17.4	89.3	129.0	50.9	261.3	69.7	27.5	141.1
Robust Cox	**49.6**	**20.9**	**107.1**	**29.1**	**11.5**	**58.9**	**85.2**	**33.6**	**172.4**	**46.0**	**18.1**	**93.1**
WLS	66.5	26.2	134.7	36.6	14.4	74.1	107.1	42.3	216.9	57.8	22.8	117.1
Modified Cox	53.6	21.5	110.4	30.0	11.8	60.7	87.7	34.6	177.7	47.4	18.7	95.9

Note: *Author Own Calculation*

Table A. 2 Outlier and Hetero Case, Sample N=100, Outlier 6 SD

N=100, Sims=50,000, Outlier with 6 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	76.4	26.1	110.5	42.0	14.4	60.8	123.0	42.0	177.9	66.4	22.7	96.1
Robust Cox	**53.1**	**18.1**	**72.9**	**27.7**	**9.5**	**40.1**	**81.2**	**27.7**	**117.4**	**43.8**	**15.0**	**63.4**
WLS	64.9	22.2	91.7	34.9	11.9	50.4	102.1	34.9	147.7	55.1	18.8	79.7
Modified Cox	57.3	19.6	75.1	28.6	9.8	41.3	83.6	28.6	121.0	45.2	15.4	65.3
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	92.4	31.6	133.7	50.8	17.4	73.5	148.8	50.8	215.3	80.4	27.5	116.2
Robust Cox	**64.2**	**21.9**	**88.2**	**33.6**	**11.5**	**48.5**	**98.2**	**33.6**	**142.1**	**53.0**	**18.1**	**76.7**
WLS	78.6	26.8	111.0	42.2	14.4	61.0	123.5	42.2	178.7	66.7	22.8	96.5
Modified Cox	69.3	23.7	90.9	34.6	11.8	50.0	101.2	34.6	146.4	54.7	18.7	79.0
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	40.5	171.3	65.1	22.3	94.2	190.7	65.1	275.8	103.0	35.2	148.9	40.5
Robust Cox	**28.1**	**113.0**	**43.0**	**14.7**	**62.2**	**125.8**	**43.0**	**182.0**	**67.9**	**23.2**	**98.3**	**28.1**
WLS	34.4	142.2	54.1	18.5	78.2	158.2	54.1	228.9	85.5	29.2	123.6	34.4
Modified Cox	30.3	116.5	44.3	15.1	64.1	129.6	44.3	187.5	70.0	23.9	101.3	30.3

Note: *Author Own Calculation*

Table A. 3 Outlier and Hetero Case, Sample N=500, Outlier 4 SD

N=500, Sims=50,000, Outlier with 4 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	31.5	26.1	77.4	17.3	14.4	42.6	50.8	42.0	124.6	27.4	22.7	67.3
Robust Cox	**22.4**	**18.1**	**51.1**	**11.4**	**9.5**	**28.1**	**33.5**	**27.7**	**82.2**	**18.1**	**15.0**	**44.4**
WLS	26.5	22.2	64.2	14.4	11.9	35.3	42.1	34.9	103.4	22.8	18.8	55.9
Modified Cox	23.3	19.6	52.6	11.8	9.8	28.9	34.5	28.6	84.7	18.6	15.4	45.8
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	38.2	31.6	93.7	21.0	17.4	51.5	61.4	50.8	150.8	33.2	27.5	81.4
Robust Cox	**27.1**	**21.9**	**61.8**	**13.9**	**11.5**	**34.0**	**40.5**	**33.6**	**99.5**	**21.9**	**18.1**	**53.7**
WLS	32.1	26.8	77.7	17.4	14.4	42.8	51.0	42.2	125.1	27.5	22.8	67.6
Modified Cox	28.2	23.7	63.7	14.3	11.8	35.0	41.8	34.6	102.5	22.6	18.7	55.4
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	48.9	40.5	120.0	26.9	22.3	66.0	78.7	65.1	193.2	42.5	35.2	104.3
Robust Cox	**34.7**	**28.1**	**79.2**	**17.7**	**14.7**	**43.5**	**51.9**	**43.0**	**127.5**	**28.0**	**23.2**	**68.8**
WLS	41.1	34.4	99.6	22.3	18.5	54.8	65.3	54.1	160.3	35.3	29.2	86.6
Modified Cox	36.2	30.3	81.6	18.3	15.1	44.9	53.5	44.3	131.3	28.9	23.9	70.9

Note: *Author Own Calculation*

Table A. 4 Outlier and Hetero Case, Sample N=500, Outlier 6 SD

N=500, Sims=50,000, Outlier with 6 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	47.6	26.1	104.0	26.2	14.4	57.2	76.7	42.0	167.4	41.4	22.7	90.4
Robust Cox	**33.8**	**18.1**	**68.6**	**17.3**	**9.5**	**37.8**	**50.6**	**27.7**	**110.5**	**27.3**	**15.0**	**59.7**
WLS	40.0	22.2	86.3	21.7	11.9	47.5	63.6	34.9	139.0	34.4	18.8	75.0
Modified Cox	35.2	19.6	70.7	17.8	9.8	38.9	52.1	28.6	113.9	28.2	15.4	61.5
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	57.6	31.6	125.8	31.7	17.4	69.2	92.8	50.8	202.6	50.1	27.5	109.4
Robust Cox	**40.9**	**21.9**	**83.1**	**20.9**	**11.5**	**45.7**	**61.2**	**33.6**	**133.7**	**33.1**	**18.1**	**72.2**
WLS	48.5	26.8	104.4	26.3	14.4	57.4	77.0	42.2	168.2	41.6	22.8	90.8
Modified Cox	42.6	23.7	85.6	21.6	11.8	47.1	63.1	34.6	137.8	34.1	18.7	74.4
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	73.8	40.5	153.9	40.6	22.3	80.7	108.1	59.2	236.1	58.4	32.0	127.5
Robust Cox	**52.4**	**28.1**	**101.6**	**26.8**	**14.7**	**53.2**	**71.3**	**39.1**	**155.8**	**38.5**	**21.1**	**84.1**
WLS	62.1	34.4	127.8	33.7	18.5	66.9	89.7	49.2	196.0	48.5	26.6	105.8
Modified Cox	54.6	30.3	104.7	27.6	15.1	54.8	73.5	40.3	160.5	39.7	21.8	86.7

Note: *Author Own Calculation*

Scenario B Outlier and Time-Dependent

Table A. 5 Outlier and Time-Dependent Case, Sample N=100, Outlier 4 SD

N=100, Sims=50,000, Outlier with 4 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	61.4	25.7	117.9	33.8	14.1	64.8	98.9	41.4	189.8	53.4	22.3	102.5
Robust Cox	45.3	19.0	87.0	24.9	10.4	47.9	73.0	30.5	140.1	39.4	16.5	75.6
Time-Dependent	51.0	21.4	98.0	28.1	11.7	53.9	82.1	34.4	157.7	44.4	18.6	85.2
Modified Cox	**42.4**	**17.7**	**81.4**	**23.3**	**9.8**	**44.7**	**68.2**	**28.5**	**131.0**	**36.8**	**15.4**	**70.7**
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	74.3	31.1	142.7	40.9	17.1	78.5	119.6	50.1	229.7	64.6	27.0	124.0
Robust Cox	54.8	22.9	105.3	30.9	12.9	59.3	90.5	37.9	173.7	48.9	20.4	93.8
Time-Dependent	61.7	25.8	118.5	34.8	14.6	66.8	101.9	42.6	195.6	55.0	23.0	105.6
Modified Cox	**51.3**	**21.5**	**98.4**	**28.9**	**12.1**	**55.5**	**84.6**	**35.4**	**162.4**	**45.7**	**19.1**	**87.7**
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	90.3	37.8	173.3	49.6	20.8	95.3	145.3	60.8	279.0	78.5	32.8	150.7
Robust Cox	70.2	29.4	134.9	38.6	16.2	74.2	113.1	47.3	217.1	61.1	25.6	117.3
Time-Dependent	75.0	31.4	144.0	41.3	17.3	79.2	120.8	50.5	231.9	65.2	27.3	125.2
Modified Cox	**65.7**	**27.5**	**126.1**	**36.1**	**15.1**	**69.4**	**105.7**	**44.2**	**203.0**	**57.1**	**23.9**	**109.6**

Note: *Author Own Calculation*

Table A. 6 Outlier and Time-Dependent Case, Sample N=100, Outlier 6 SD

N=100, Sims=50,000, Outlier with 6 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	82.7	36.3	157.4	45.5	19.9	86.6	134.0	58.4	253.4	72.4	31.5	136.8
Robust Cox	61.0	26.8	116.2	33.6	14.7	63.9	98.9	43.1	187.0	53.4	23.3	101.0
Time-Dependent	68.7	30.1	130.8	37.8	16.6	71.9	111.4	48.5	210.6	60.1	26.2	113.7
Modified Cox	**57.1**	**25.0**	**108.6**	**31.4**	**13.8**	**59.7**	**92.5**	**40.3**	**174.9**	**49.9**	**21.8**	**94.4**
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	101.0	44.3	192.2	55.5	24.4	105.7	163.6	71.3	309.4	88.3	38.5	167.1
Robust Cox	74.5	32.7	141.8	41.0	18.0	78.0	120.7	52.6	228.4	65.2	28.4	123.3
Time-Dependent	83.9	36.8	159.7	46.2	20.2	87.8	136.0	59.3	257.1	73.4	32.0	138.8
Modified Cox	**69.7**	**30.6**	**132.6**	**38.3**	**16.8**	**72.9**	**112.9**	**49.2**	**213.5**	**61.0**	**26.6**	**115.3**
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	124.9	54.8	237.7	68.7	30.1	130.7	202.3	88.2	382.7	109.3	47.6	206.6
Robust Cox	92.2	40.4	175.4	50.7	22.2	96.5	149.3	65.1	282.4	80.6	35.1	152.5
Time-Dependent	103.8	45.5	197.5	57.1	25.0	108.6	168.1	73.3	318.0	90.8	39.6	171.7
Modified Cox	**86.2**	**37.8**	**164.0**	**47.4**	**20.8**	**90.2**	**139.6**	**60.8**	**264.0**	**75.4**	**32.9**	**142.6**

Note: *Author Own Calculation*

Table A. 7: Outlier and Time-Dependent Case, Sample N=500, Outlier 4 SD

N=500, Sims=50,000, Outlier with 4 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	44.8	19.2	92.7	24.6	10.6	51.0	72.1	30.9	149.2	38.9	16.7	80.6
Robust Cox	33.1	14.5	68.4	18.2	7.8	37.6	53.2	22.8	110.1	28.7	12.3	59.5
Time-Dependent	37.2	16.3	77.0	20.5	8.8	42.4	59.9	25.7	124.0	32.4	13.9	67.0
Modified Cox	**30.9**	**13.6**	**64.0**	**17.0**	**7.3**	**35.2**	**49.8**	**21.3**	**103.0**	**26.9**	**11.5**	**55.6**
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	54.7	24.4	102.9	27.4	11.7	56.6	80.1	34.3	165.7	43.2	18.5	89.5
Robust Cox	40.4	18.4	75.9	20.2	8.6	41.8	59.1	25.3	122.3	31.9	13.7	66.0
Time-Dependent	45.5	20.7	85.5	22.7	9.7	47.0	66.5	28.5	137.7	35.9	15.4	74.3
Modified Cox	**37.7**	**17.2**	**71.0**	**18.9**	**8.1**	**39.0**	**55.2**	**23.7**	**114.3**	**29.8**	**12.8**	**61.7**
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	67.6	29.0	138.2	36.7	15.7	76.0	107.5	46.1	222.5	58.1	24.9	120.2
Robust Cox	49.9	22.9	102.0	27.1	11.6	56.1	79.4	34.0	164.2	42.9	18.4	88.7
Time-Dependent	56.2	24.7	114.9	30.5	13.1	63.2	89.4	38.3	184.9	48.3	20.7	99.9
Modified Cox	**46.7**	**20.5**	**95.4**	**25.3**	**10.9**	**52.5**	**74.2**	**31.8**	**153.5**	**40.1**	**17.2**	**82.9**

Note: *Author Own Calculation*

Table A. 8 Outlier and Time-Dependent Case, Sample Size N=500, Outlier 6 SD

N=500, Sims=50,000, Outlier with 6 SD

	Parameter 1						Parameter 2					
	5% Outlier											
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	65.7	28.2	127.5	36.1	15.5	70.1	105.8	45.3	205.3	57.1	24.5	110.9
Robust Cox	48.0	21.1	94.1	26.7	11.4	51.8	78.1	33.5	151.5	42.2	18.1	81.8
Time-Dependent	55.2	24.2	106.0	30.0	12.9	58.3	87.9	37.7	170.6	47.5	20.3	92.1
Modified Cox	**45.1**	**19.8**	**88.0**	**24.9**	**10.7**	**48.4**	**73.0**	**31.3**	**141.7**	**39.4**	**16.9**	**76.5**
	10% Outlier											
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	80.2	35.8	162.0	45.9	19.7	89.1	134.3	57.6	260.7	72.5	31.1	140.8
Robust Cox	58.6	26.7	119.5	33.9	14.5	65.7	99.1	42.5	192.4	53.5	22.9	103.9
Time-Dependent	67.4	30.7	134.6	38.1	16.3	74.0	111.6	47.8	216.7	60.3	25.8	117.0
Modified Cox	**55.1**	**25.1**	**111.7**	**31.7**	**13.6**	**61.5**	**92.7**	**39.7**	**179.9**	**50.1**	**21.4**	**97.2**
	20% Outlier											
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	99.2	42.5	192.6	54.6	23.4	105.9	159.7	68.4	310.0	86.3	37.0	167.4
Robust Cox	72.5	33.3	148.7	42.1	18.1	81.8	123.3	52.9	239.4	66.6	28.5	129.3
Time-Dependent	83.3	36.5	160.0	45.3	19.4	88.0	132.7	56.9	257.6	71.7	30.7	139.1
Modified Cox	**68.2**	**29.9**	**132.9**	**37.6**	**16.1**	**73.1**	**110.2**	**47.2**	**213.9**	**59.5**	**25.5**	**115.5**

Note: *Author Own Calculation*

Scenario C Hetero and Time-Dependent

Table A. 9 Hetero and Time-Dependent Case, Sample N=100 and N=500

N=100, Sims=50,000												
	Parameter=1						Parameter=2					
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox Model	56.4	24.2	119.1	31.0	13.3	65.5	90.8	38.9	192.2	49.0	21.0	103.8
WLS	41.2	18.1	87.9	22.9	9.8	48.3	67.0	28.7	141.9	36.2	15.5	76.6
Time-Dependent	47.4	20.8	98.9	25.8	11.0	54.4	75.5	32.3	159.7	40.7	17.5	86.3
Modified Cox	**38.7**	**17.0**	**82.2**	**21.4**	**9.2**	**45.2**	**62.7**	**26.8**	**132.6**	**33.8**	**14.5**	**71.6**
N=500, Sims=50,000												
	Parameter=1						Parameter=2					
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox Model	38.2	16.4	80.7	21.0	9.0	44.4	61.4	26.3	135.9	33.2	14.2	73.4
WLS	28.0	12.3	59.6	15.5	6.7	32.8	45.3	19.4	100.3	24.5	10.5	54.2
Time-Dependent	32.1	14.1	67.1	17.5	7.5	36.9	51.0	21.9	112.9	27.6	11.8	61.0
Modified Cox	**26.3**	**11.5**	**55.7**	**14.5**	**6.2**	**30.6**	**42.4**	**18.2**	**93.8**	**22.9**	**9.8**	**50.6**

Note: *Author Own Calculation*

Scenario D Outlier, Heteroscedasticity, and Time-Dependent

Table A. 10 Outlier, Hetero, and Time-Dependent Case, N=100, Outlier 4 SD

N=100, Sims=50,000, Outlier with 4 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	73.5	28.3	146.7	40.4	15.6	80.7	118.3	45.6	236.2	63.9	24.6	127.5
Robust Cox	53.4	20.6	106.7	29.4	11.3	58.7	86.0	33.1	171.7	46.5	17.9	92.7
WLS	55.9	21.5	111.6	30.8	11.8	61.4	90.1	34.7	179.7	48.6	18.7	97.1
Time-Dependent	60.3	23.2	120.3	33.1	12.8	66.2	97.0	37.4	193.7	52.4	20.2	104.6
Modified Cox	**49.3**	**19.0**	**98.4**	**27.1**	**10.4**	**54.1**	**79.4**	**30.6**	**158.5**	**42.9**	**16.5**	**85.6**
10% Outlier												
Theta=1				**Theta=2**			**Theta=1**			**Theta=2**		
Cox	91.7	35.3	182.9	50.4	19.4	100.6	147.6	56.8	294.5	79.7	30.7	159.0
Robust Cox	66.6	25.7	133.0	36.6	14.1	73.1	107.3	41.3	214.1	57.9	22.3	115.6
WLS	69.7	26.9	139.2	38.4	14.8	76.6	112.3	43.3	224.1	60.6	23.4	121.0
Time-Dependent	75.2	28.9	150.0	41.3	15.9	82.5	121.0	46.6	241.5	65.3	25.2	130.4
Modified Cox	**61.5**	**23.7**	**122.7**	**33.8**	**13.0**	**67.5**	**99.0**	**38.1**	**197.6**	**53.5**	**20.6**	**106.7**
20% Outlier												
Theta=1				**Theta=2**			**Theta=1**			**Theta=2**		
Cox	129.3	49.8	258.1	71.1	27.4	142.0	208.2	80.2	415.6	112.4	43.3	224.4
Robust Cox	94.0	36.2	187.7	51.7	19.7	103.2	151.4	58.3	302.2	81.7	31.5	163.2
WLS	98.4	37.9	196.5	54.1	20.9	108.0	158.5	61.0	316.3	85.6	33.0	170.8
Time-Dependent	106.1	40.9	211.7	58.3	22.5	116.4	170.8	65.8	340.8	92.2	35.5	184.0
Modified Cox	**86.8**	**33.4**	**173.2**	**47.7**	**18.4**	**95.3**	**139.7**	**53.8**	**278.9**	**75.5**	**29.1**	**150.6**

Note: *Author Own Calculation*

Table A. 11 Outlier, Hetero, and Time-Dependent Case, N=100, Outlier 6 SD

N=100, Sims=50,000, Outlier with 6 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	96.7	37.2	188.7	53.2	20.5	103.8	155.7	60.0	303.8	84.1	32.4	164.1
Robust Cox	70.3	27.1	139.9	38.7	14.9	75.5	113.2	43.6	220.9	61.1	23.5	119.3
WLS	73.6	28.3	143.6	40.5	15.6	79.0	118.5	45.6	231.2	64.0	24.6	124.8
Time-Dependent	79.3	30.5	154.7	43.6	16.8	85.1	127.7	49.2	249.1	68.9	26.6	134.5
Modified Cox	**64.9**	**25.0**	**126.6**	**35.7**	**13.7**	**69.6**	**104.5**	**40.2**	**203.9**	**56.4**	**21.7**	**110.1**
10% Outlier												
Theta=1				**Theta=2**			**Theta=1**			**Theta=2**		
Cox	120.6	46.4	237.8	66.3	25.5	129.4	194.1	74.8	378.8	104.8	40.4	204.6
Robust Cox	87.7	33.8	176.3	48.2	18.6	94.1	141.1	54.4	275.4	76.2	29.4	148.7
WLS	91.8	35.3	180.9	50.5	19.4	98.5	147.7	56.9	288.3	79.8	30.7	155.7
Time-Dependent	98.9	38.1	195.0	54.4	20.9	106.1	159.2	61.3	310.7	86.0	33.1	167.8
Modified Cox	**80.9**	**31.2**	**159.5**	**44.5**	**17.1**	**86.8**	**130.3**	**50.2**	**254.2**	**70.3**	**27.1**	**137.3**
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	165.9	63.9	323.8	91.3	35.2	178.1	267.2	102.9	521.3	144.3	55.6	281.5
Robust Cox	120.6	46.5	240.1	66.3	25.6	129.5	194.2	74.8	379.0	104.9	40.4	204.7
WLS	126.3	48.6	246.4	69.5	26.8	135.5	203.3	78.3	396.7	109.8	42.3	214.2
Time-Dependent	136.1	52.4	265.5	74.8	28.8	146.0	219.1	84.4	427.5	118.3	45.6	230.8
Modified Cox	**111.3**	**42.9**	**217.3**	**61.2**	**23.6**	**119.5**	**179.3**	**69.0**	**349.8**	**96.8**	**37.3**	**188.9**

Note: *Author Own Calculation*

Table A. 12 Outlier, Hetero, and Time-Dependent Case, N=500, Outlier 4 SD

N=500, Sims=50,000, Outlier with 4 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	55.8	21.5	108.7	30.7	11.8	59.8	89.8	34.6	175.0	48.5	18.7	94.5
Robust Cox	40.6	15.6	79.0	22.3	8.6	43.5	65.3	25.2	127.2	35.3	13.6	68.7
WLS	42.5	16.4	82.7	23.4	9.0	45.5	68.4	26.3	133.2	36.9	14.2	71.9
Time-Dependent	45.8	17.6	89.1	25.2	9.7	49.0	73.7	28.4	143.5	39.8	15.3	77.5
Modified Cox	**37.4**	**14.4**	**72.9**	**20.6**	**7.9**	**40.1**	**60.3**	**23.2**	**117.4**	**32.6**	**12.5**	**63.4**
10% Outlier												
Theta=1				**Theta=2**			**Theta=1**			**Theta=2**		
Cox	69.6	26.8	135.5	38.3	14.7	74.6	112.0	43.2	218.2	60.5	23.3	117.8
Robust Cox	50.6	19.5	98.5	27.8	10.7	54.2	81.4	31.4	158.7	44.0	16.9	85.7
WLS	53.0	20.4	103.2	29.1	11.2	56.7	85.3	32.8	166.1	46.0	17.7	89.7
Time-Dependent	57.1	22.0	111.2	31.4	12.1	61.1	91.9	35.4	179.0	49.6	19.1	96.6
Modified Cox	**46.7**	**18.0**	**91.0**	**25.7**	**9.9**	**50.0**	**75.2**	**29.0**	**146.4**	**40.6**	**15.6**	**79.1**
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	98.2	37.8	191.3	54.0	20.8	105.2	158.1	60.9	308.0	85.4	32.9	166.3
Robust Cox	71.4	27.5	139.1	39.3	15.1	76.5	114.9	44.3	223.9	62.1	23.9	120.9
WLS	74.7	28.8	145.6	41.1	15.8	80.1	120.3	46.3	234.4	65.0	25.0	126.6
Time-Dependent	80.5	31.0	156.8	44.3	17.1	86.3	129.6	49.9	252.5	70.0	27.0	136.4
Modified Cox	**65.9**	**25.4**	**128.3**	**36.2**	**14.0**	**70.6**	**106.1**	**40.9**	**206.6**	**57.3**	**22.1**	**111.6**

Note: *Author Own Calculation*

Table A. 13 Outlier, Hetero, and Time-Dependent Case, N=500, Outlier 6 SD

N=500, Sims=50,000, Outlier with 6 SD												
	Parameter 1						Parameter 2					
5% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Cox	78.1	30.1	167.6	43.0	16.5	92.2	125.7	48.4	269.8	67.9	26.2	145.7
Robust Cox	56.8	21.9	121.8	31.2	12.0	67.0	91.4	35.2	196.2	49.4	19.0	105.9
WLS	59.4	22.9	127.5	32.7	12.6	70.1	95.7	36.9	205.3	51.7	19.9	110.9
Time-Dependent	64.0	24.7	137.4	35.2	13.6	75.6	103.1	39.7	221.3	55.7	21.4	119.5
Modified Cox	**52.4**	**20.2**	**112.5**	**28.8**	**11.1**	**61.9**	**84.4**	**32.5**	**181.1**	**45.6**	**17.5**	**97.8**
10% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	97.4	37.5	209.0	53.6	20.6	114.9	156.8	60.4	336.5	84.7	32.6	181.7
Robust Cox	70.8	27.3	151.9	38.9	15.0	83.6	114.0	43.9	244.6	61.6	23.7	132.1
WLS	74.1	28.5	159.0	40.8	15.7	87.5	119.3	46.0	256.1	64.4	24.8	138.3
Time-Dependent	79.9	30.8	171.4	43.9	16.9	94.3	128.6	49.5	275.9	69.4	26.7	149.0
Modified Cox	**65.3**	**25.2**	**140.2**	**35.9**	**13.8**	**77.1**	**105.2**	**40.5**	**225.8**	**56.8**	**21.9**	**121.9**
20% Outlier												
	Theta=1			Theta=2			Theta=1			Theta=2		
Cox	137.4	52.9	294.9	75.6	29.1	162.2	221.3	85.2	474.8	119.5	46.0	256.4
Robust Cox	99.9	38.5	214.4	55.0	21.2	117.9	160.9	62.0	345.2	86.9	33.5	186.4
WLS	104.6	40.3	224.4	57.5	22.2	123.4	168.4	64.9	361.3	90.9	35.0	195.1
Time-Dependent	112.7	43.4	241.8	62.0	23.9	133.0	181.4	69.9	389.4	98.0	37.7	210.3
Modified Cox	**92.2**	**35.5**	**197.9**	**50.7**	**19.5**	**108.8**	**148.5**	**57.2**	**318.6**	**80.2**	**30.9**	**172.1**

Note: *Author Own Calculation*

Appendix B

Main coding of Simulation

```
#33 outlier, hetero and time dependent okay
#2 best
#123 for theta1, start

set.seed(33)
#Rsquare scenario 1 all 3 problems
#Recall relevant Libraries
library(survival)   #For simple Cox regression
library(timereg)    #For time dependent Cox, need timereg pakage
library(lmtest)     #For Brush pagan test and WLS
library(robustbase) #For Robust Cox Outlier Case, need time robustbase package
#install.packages("car")
library(car) # car data for influence plot
library(dplyr)  # dplyr package for slicing
library(stargazer)
#install.packages("gtable")
library(gtable)
#results <- your_results # Replace `your_results` with your actual result object

# Create gtable object
#gt <- gtable_matrix(results)

n<-500
#theta<-1
theta<-2
#Define n number of random sample =n
#x2 gender bernaouli, binomial more experiment, bernouli one experiment
#n, gender, prob
x2<-rbinom(n,1,0.5)
#x3 generating using standard normal
x1<-rnorm(n,0,1)
#x2<-rnorm(n,0,1)
x3a<-rnorm(n,0,1)
x4a<-rnorm(n, 4,2)
x5<-rnorm(n, 0,1)
tt<-seq(1,100)
x4<-x4a*tt

x6<-rnorm(n,0,1)
#Previous study, knowlegde, theory

beta1<-1
beta2<-1
beta3<-1
beta4<-1
```

```
beta5<-1
beta6<-1

#beta1<-2
#beta2<-2.5
#beta3<- 2
#beta4<- 3
#beta5<-3
#beta6<-9
N<-1000

#Epsilon~exp(theta)
#theta<-2

#RMSE for all 5 models
RMSE_Cox<-matrix(N,1) #Matrix for RMSE Cox model
RMSE_TD<-matrix(N,1)  #Matrix for RMSE Time Dependent Cox model
RMSE_WLS<-matrix(N,1) #Matrix for RMSE Weighted Least Square model
RMSE_RC<-matrix(N,1)  #Matrix for RMSE Random Cox model
RMSE_MCOH<-matrix(N,1)  #Matrix for RMSE Modified model

#MAE for all 5 models
MAE_Cox<-matrix(N,1) #Matrix for MAE Cox model
MAE_TD<-matrix(N,1)  #Matrix for MAE Time Dependent Cox model
MAE_WLS<-matrix(N,1) #Matrix for MAE Weighted Least Square model
MAE_RC<-matrix(N,1)  #Matrix for MAE Random Cox model
MAE_MCOH<-matrix(N,1)  #Matrix for MAE Modified model

#MAE for all 5 models
MAPE_Cox<-matrix(N,1) #Matrix for MAPE Cox model
MAPE_TD<-matrix(N,1)  #Matrix for MAPE Time Dependent Cox model
MAPE_WLS<-matrix(N,1) #Matrix for MAPE Weighted Least Square model
MAPE_RC<-matrix(N,1)  #Matrix for MAPE Random Cox model
MAPE_MCOH<-matrix(N,1)  #Matrix for MAPE Modified model

#Betas
Cox_beta1<-matrix(N,1)  #initialized matrix of Cox Beta1
Cox_beta2<-matrix(N,1)
Cox_beta3<-matrix(N,1)
Cox_beta4<-matrix(N,1)
Cox_beta5<-matrix(N,1)

TD_beta1<-matrix(N,1)  #initialized matrix of TD Beta1
TD_beta2<-matrix(N,1)
TD_beta3<-matrix(N,1)
TD_beta4<-matrix(N,1)
TD_beta5<-matrix(N,1)

WLS_beta1<-matrix(N,1)  #initialized matrix of WLS Beta1
```

```
WLS_beta2<-matrix(N,1)
WLS_beta3<-matrix(N,1)
WLS_beta4<-matrix(N,1)
WLS_beta5<-matrix(N,1)

TD_beta1<-matrix(N,1)  #initialized matrix of TD Beta1
TD_beta2<-matrix(N,1)
TD_beta3<-matrix(N,1)
TD_beta4<-matrix(N,1)
TD_beta5<-matrix(N,1)

RC_beta1<-matrix(N,1)  #initialized matrix of RC Beta1
RC_beta2<-matrix(N,1)
RC_beta3<-matrix(N,1)
RC_beta4<-matrix(N,1)
RC_beta5<-matrix(N,1)

MC_beta1<-matrix(N,1)  #initialized matrix of MCox Beta1
MC_beta2<-matrix(N,1)
MC_beta3<-matrix(N,1)
MC_beta4<-matrix(N,1)
MC_beta5<-matrix(N,1)
#epsilon1<-rnorm(n, 0,2)

x3=x3a #Hetero problems comes when explanatory is correlated with error term
#
x4=x4a
#n1<- rep(1:n,1)
#
for (i in 1:N) {

 epsilon<-rexp(n,theta)*x4
 # epsilon<-rnorm(n)

 #   y<- exp(beta1*x1+beta2*x2+beta3*x3+beta4*x4+beta5*x5+epsilon) +beta6*x6

 y<- beta1*x1+beta2*x2+beta3*x3+beta4*x4+beta5*x5+epsilon #+beta6*x6

 # sd(y)

 #For 5% Outlier
 #
 #  y<- replace(y,5:9, max(y))
 # y
 #
 #y<-replace(y$entry3, y$entry5, y$entry7< max(y)*3*sd(y),max(y)*3*sd(y),max(y)*3*sd(y))
 # max(y)*3*sd(y)

 #For 10% Outlier
```

```
  # y<- replace(y,81:89, max(y))
  y<- replace(y,71:89, mean(y)*6*sd(y))
  #
  # y<-replace(y$entry3, y$entry5, y$entry7< max(y)*3*sd(y),max(y)*3*sd(y),max(y)*3*sd(y))
  # max(y)*3*sd(y)

  # status1 <- ifelse(y > mean(y), 1, 0) #status is dummy of y.
  status1 <- ifelse(y > quantile(y, 0.75), 1, 0) #status is dummy of y.
  #Model1
  Cox <- summary(Coxph(Surv(y, status1) ~ x1+x2+x3+x4+x5))
  Cox_beta1[i]<-exp(Cox$coefficients[1,1])
  Cox_beta2[i]<-exp(Cox$coefficients[2,1])
  Cox_beta3[i]<-exp(Cox$coefficients[3,1])
  Cox_beta4[i]<-exp(Cox$coefficients[4,1])
  Cox_beta5[i]<-exp(Cox$coefficients[5,1])
  #Cox_beta1[i]<-Cox$coefficients[1,1]
  #Cox_beta2[i]<-Cox$coefficients[2,1]
  #Cox_beta3[i]<-Cox$coefficients[3,1]
  #Cox_beta4[i]<-Cox$coefficients[4,1]
  # Cox_beta5[i]<-Cox$coefficients[5,1]
  Cox_predict<-
Cox$coefficients[1,1]*x1+Cox$coefficients[2,1]*x2+Cox$coefficients[3,1]*x3+Cox$coefficients
[4,1]*x4+Cox$coefficients[5,1]*x5#+Cox$coefficients[6,1]*x6


  Cox_error<-y-Cox_predict
  Cox_square_error<- Cox_error*Cox_error
  RMSE_Cox[i]<- sqrt(mean(Cox_square_error)) #RMSE
  #MAE_C0x
  Cox_absolute_error<- abs(Cox_error)
  MAE_Cox[i]<-mean(Cox_absolute_error)   #MA

  MAPE_Cox[i]<-mean(Cox_absolute_error/y)*100   #MAPE
  #MAPE = (1/n) * Σ(|(Actual - Predicted)/Actual|) * 100
  #find predicted value
  #dependent minus predicted
  #root first
  #means

  #Model2
  TD <- summary(Coxph(Surv(y, status1) ~
                      x1+x2+x3+x4+x5+tt(x6),
                    tt = function(x, t, ...) x * log(t+20)))
  TD_beta1[i]<-exp(TD$coefficients[1,1])
  TD_beta2[i]<-exp(TD$coefficients[2,1])
  TD_beta3[i]<-exp(TD$coefficients[3,1])
  TD_beta4[i]<-exp(TD$coefficients[4,1])
  TD_beta5[i]<-exp(TD$coefficients[5,1])
  #RMSE_TD
```

```
 TD_predict<-
TD$coefficients[1,1]*x1+TD$coefficients[2,1]*x2+TD$coefficients[3,1]*x3+TD$coefficients[4,1
]*x4+TD$coefficients[5,1]*x5+TD$coefficients[6,1]*x6
 TD_error<-y-TD_predict
 TD_square_error<- TD_error*TD_error
 RMSE_TD[i]<- sqrt(mean(TD_square_error))
 #MAE_TD
 TD_absolute_error<- abs(TD_error)
 MAE_TD[i]<-mean(TD_absolute_error)   #MAE
 MAPE_TD[i]<-mean(TD_absolute_error/y)*100   #MAPE
 #Model3
 hetro3 <- lm(y ~ x1+x2+x3+x4+x5)
 #bptest(hetro3)
 #influencePlot(hetro3)
 #to introduce hetero term in the model
 #define weights to us
 wt2 <- 1 / lm(abs(hetro3$residuals) ~ hetro3$fitted.values)$fitted.values^2
 length(wt2)
 WLS <- summary(Coxph(Surv(y, status1) ~ x1+x2+x3+x4+x5, weights = wt2))
 WLS
 WLS_beta1[i]<-exp(WLS$coefficients[1,1])
 WLS_beta2[i]<-exp(WLS$coefficients[2,1])
 WLS_beta3[i]<-exp(WLS$coefficients[3,1])
 WLS_beta4[i]<-exp(WLS$coefficients[4,1])
 WLS_beta5[i]<-exp(WLS$coefficients[5,1])
 #MSE_WLS
 WLS_predict<-
WLS$coefficients[1,1]*x1+WLS$coefficients[2,1]*x2+WLS$coefficients[3,1]*x3+WLS$coefficients[4,1]*x4+WLS$coefficients[5,1]*x5#+WLS$coefficients[6,1]*x6
 #WLS_predict
 WLS_error<-y-WLS_predict
 WLS_square_error<- WLS_error*WLS_error
 RMSE_WLS[i]<- sqrt(mean(WLS_square_error))
 #RMSE_WLS
 #MAE WLS
 WLS_absolute_error<- abs(WLS_error
 MAE_WLS[i]<-mean(WLS_absolute_error)   #MAE
 #MAE_WLS
 MAPE_WLS[i]<-mean(WLS_absolute_error/y)*100   #MAPE

 #Model4
 #Robust Cox
 #Model4
 RC<-summary(lmrob(y ~ x1+x2+x3+x4+x5+x6))

 #RMSE_RC
 RC_predict<-
RC$coefficients[1,1]*x1+RC$coefficients[2,1]*x2+RC$coefficients[3,1]*x3+RC$coefficients[4,1
]*x4+RC$coefficients[5,1]*x5+RC$coefficients[6,1]*x6
```

```
 RC_error<-y-RC_predict
 RC_square_error<- RC_error*RC_error
 RMSE_RC[i]<- sqrt(mean(RC_square_error))
 #MAE_RC
 RC_absolute_error<- abs(RC_error)
 MAE_RC[i]<-mean(RC_absolute_error)   #MAE
 MAPE_RC[i]<-mean(RC_absolute_error/y)*100   #MAPE

 #Model5
 #
 MC <- (Coxph(Surv(y, status1) ~ x1+x2+tt(x3)+x4+x5,  weights = wt2))
 MC
 MCOH<-summary(lmrob(y ~ x1+x2+x3+x4+x5+x6, weights = wt2))
 #RMSE_MC
 MCOH_predict<-
MCOH$coefficients[1,1]*x1+MCOH$coefficients[2,1]*x2+MCOH$coefficients[3,1]*x3+MCOH
$coefficients[4,1]*x4+MCOH$coefficients[5,1]*x5#+MC$coefficients[6,1]*x6
 #MC_predict
 MCOH_error<-y-MCOH_predict
 MCOH_square_error<- MCOH_error*MCOH_error
 RMSE_MCOH[i]<- sqrt(mean(MCOH_square_error))
 # RMSE_MC
 #MAE_MC
 MCOH_absolute_error<- abs(MCOH_error)
 MAE_MCOH[i]<-mean(MCOH_absolute_error)
 MAPE_MCOH[i]<-mean(MCOH_absolute_error/y)*100   #MAPE

}
#RMSE_Cox
#RMSE_TD
#RMSE_WLS
RMSE_WLS<-na.omit(RMSE_WLS)
#RMSE_RC
#RMSE_MCOH
#MAE_Cox
#MAE_TD
MAE_WLS<-na.omit(MAE_WLS)
#MAE_WLS
#MAE_RC
#MAE_MCOH
MAPE_WLS<-na.omit(MAPE_WLS)
#Average of RMSE, MAE and MAPE
Average_RMSE_Cox<-mean(RMSE_Cox)
Average_RMSE_TD<-mean(RMSE_TD)
Average_RMSE_WLS<-mean(RMSE_WLS)
Average_RMSE_RC<-mean(RMSE_RC)
Average_RMSE_MCOH<-mean(RMSE_MCOH)
Average_MAE_Cox<-mean(MAE_Cox)
Average_MAE_TD<-mean(MAE_TD)
```

```
Average_MAE_WLS<-mean(MAE_WLS)
Average_MAE_RC<-mean(MAE_RC)
Average_MAE_MCOH<-mean(MAE_MCOH)
Average_MAPE_Cox<-mean(MAPE_Cox)
Average_MAPE_RC<-mean(MAPE_RC)
Average_MAPE_WLS<-mean(MAPE_WLS)
Average_MAPE_TD<-mean(MAPE_TD)
Average_MAPE_MCOH<-mean(MAPE_MCOH)
#Betas Averages
ACox_beta1<-mean(Cox_beta1)
ACox_beta2<-mean(Cox_beta2)
ACox_beta3<-mean(Cox_beta3)
ACox_beta4<-mean(Cox_beta4)
ACox_beta5<-mean(Cox_beta5)
ATD_beta1<-mean(TD_beta1)
ATD_beta2<-mean(TD_beta2)
ATD_beta3<-mean(TD_beta3)
ATD_beta4<-mean(TD_beta4)
ATD_beta5<-mean(TD_beta5)
WLS_beta1<-na.omit(WLS_beta1)
WLS_beta2<-na.omit(WLS_beta2)
WLS_beta3<-na.omit(WLS_beta3)
WLS_beta4<-na.omit(WLS_beta4)
WLS_beta5<-na.omit(WLS_beta5)
AWLS_beta1<-mean(WLS_beta1)
AWLS_beta2<-mean(WLS_beta2)
AWLS_beta3<-mean(WLS_beta3)
AWLS_beta4<-mean(WLS_beta4)
AWLS_beta5<-mean(WLS_beta5)
#Averages of RMSE and MAE of N
Average_RMSE_Cox
Average_RMSE_TD
Average_RMSE_WLS
Average_RMSE_RC
Average_RMSE_MCOH
Average_MAE_Cox
Average_MAE_TD
Average_MAE_WLS
Average_MAE_RC
Average_MAE_MCOH
#Holding Betas of all Models
ACox_combine_Betas<-cbind(ACox_beta1,ACox_beta2,ACox_beta3,ACox_beta4,ACox_beta5)
#barplot(Average_combine_Betas)
head(ACox_combine_Betas)
ATD_combine_Betas<-cbind(ATD_beta1,ATD_beta2,ATD_beta3,ATD_beta4,ATD_beta5)
#barplot(ATD_combine_Betas)
head(ATD_combine_Betas)
AWLS_combine_Betas<-
cbind(AWLS_beta1,AWLS_beta2,AWLS_beta3,AWLS_beta4,AWLS_beta5)
```

```
#barplot(AWLS_combine_Betas)
head(AWLS_combine_Betas)
#Making comparison RMSE, MAE and MAPE. graph
#RMSE
#Average_combine_RMSE<-
cbind(Average_RMSE_Cox,Average_RMSE_TD,Average_RMSE_WLS,Average_RMSE_RC,
Average_RMSE_MCOH)
#barplot(Average_combine_RMSE)
#head(Average_combine_RMSE)
Average_combine_RMSE<-
cbind(Average_RMSE_Cox,Average_RMSE_RC,Average_RMSE_WLS,Average_RMSE_TD,A
verage_RMSE_MCOH)
#barplot(Average_combine_RMSE)
head(Average_combine_RMSE)
Average_combine_MAE<-
cbind(Average_MAE_Cox,Average_MAE_RC,Average_MAE_WLS,Average_MAE_TD,Averag
e_MAE_MCOH)
#barplot(Average_combine_MAE)
head(Average_combine_MAE)
Average_combine_MAPE<-
cbind(Average_MAPE_Cox,Average_MAPE_RC,Average_MAPE_WLS,Average_MAPE_TD,A
verage_MAPE_MCOH)
#barplot(Average_combine_MAPE)
head(Average_combine_MAPE)
#Average_combine_MAE<-
cbind(Average_MAE_Cox,Average_MAE_TD,Average_MAE_WLS,Average_MAE_RC,Averag
e_MAE_MCOH)
#barplot(Average_combine_MAE)
#head(Average_combine_MAE)
#Average_combine_MAPE<-
cbind(Average_MAPE_Cox,Average_MAPE_RC,Average_MAPE_WLS,Average_MAPE_MCO
H)
#barplot(Average_combine_MAPE)
#head(Average_combine_MAPE)
#bptest(hetro3)
#influencePlot(hetro3)
#fit2 <- Coxph(Surv(y, status1) ~
#           +x2+x3+x4+x5+x6)
# zph <- Cox.zph(fit2)
#zph
```

Appendix C

Algorithm

- We take the base distribution exponential distribution because the family of survival analysis follow exponential distribution, and we are going to do modification in the semi parametric distribution family.
- We have taken mixed form of distribution, such as exponential and normal distribution in the covariates, which is possible.
- The exponential distribution follow theta (Ɵ) parameter, we have varied Ɵ from 1 to 2, to see the change in the model performance.
- We have generated Y time variable from exponential distribution.
- For survival analysis (SA) we need event occurrence dummy, either the event occurred or not, so taken that dummy if the time is greater than average, equal one, otherwise zero.
- We have four major scenarios, such as outliers and hetero, outlier and time dependent covariates, and hetero and time dependent covariate, and main scenario including outliers, hetero and time dependent covariates.
- Next step we have vary sample size, Parameter theta and different issue of outliers, heteroscedasticity and time dependent covariates.
- In final stage for the model performance, we taken 3 indicators such as RMSE, MAE and MAPE.