https://doi.org/10.52131/joe.2023.0502.0147

*i*RASD Journal of Economics

Volume 5, Number 2, 2023, Pages 577 – 588

Journal Home Page:
https://journals.internationalrasd.org/index.php/joe

# Outlier and Time-Dependent Covariate in Survival Analysis, A Simulation Based Study

Nauman Ahmad[1], Amena Urooj[2]

[1] PhD Scholar, School of Economics, Pakistan Institute of Development Economics Islamabad, Pakistan.
Email: naumanpide@gmail.com

[2] Assistant Professor, School of Economics, Pakistan Institute of Development Economics Islamabad, Pakistan.
Email: amna@pide.org.pk

**ARTICLE INFO**

***Article History:***

Received: April 28, 2023
Revised: June 29, 2023
Accepted: June 30, 2023
Available Online: June 30, 2023

***Keywords:***

Survival Analysis
Outlier
Time-dependent
Schoenfield Residuals
Hazard Ratio
Cox Regression

***JEL Classification Codes:***

C01, E24, J24

***Funding:***

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**ABSTRACT**

The Cox regression model is widely used in Survival Analysis, also related to medical fields. The Cox regression model is further extended to tackle problems such as non-proportionality and time-dependent covariates. This paper focuses on the behavior of the Cox proportional hazard model in the presence of outliers and time-dependent covariates. To compare the performance of widely used existing time-to-event models in the presence of Outliers and time-dependent covariates and propose a modified Cox model in case of outliers and time-dependent covariates. The algorithm used in the Cox and time-dependent Cox model is extended to tackle the problem of outlier and time-dependent covariates jointly in a model. The estimated model's betas, RMSE, MAE, and MAPE, were compared among the different models. The study concluded that the modified Cox model outperformed the existing time-to-event methodology if the model simultaneously has an Outlier and time-dependent covariates problem.



**Corresponding Author's Email:** naumanpide@gmail.com

**Citation:** Ahmad, N., & Urooj, A. (2023). Outlier and Time-Dependent Covariate in Survival Analysis, A Simulation Based Study. *IRASD Journal of Economics*, *5*(2), 577–588.
https://doi.org/10.52131/joe.2023.0502.0147

## 1. Introduction

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. A unique feature of survival data is that, typically, not all patients experience the event (e.g., death) by the end of the observation period. Hence, the actual survival times for some patients are unknown. This phenomenon, called censoring, must be accounted for in the analysis to make valid inferences. The statistical methods for appropriately analyzing time-to-event data include non-parametric 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 & Vetter, 2018). The modelling of survival data analysis of health-related data is an important area for econometricians because doctors are

interested in seeing if a patient has diabetes, breast cancer, HIV/AIDS, lung cancer, COVID-19, or other lethal diseases that cause death. Hence, doctors and patients are interested in whether patients will recover safely or the chance of survival of that specific type of disease (Aalen, Borgan, & Gjessing, 2008).

The unbiased estimates, magnitude, accuracy, consistency, and reliability findings remain interesting in exploring which variable is more important and which is less. To know the importance of each covariate on a 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 working on survival analysis related models with innovation and improvement come in the shape of cox regression. The Cox regression is the most used conventional survival analysis model, invented for time-to-event study (Cox, 1972). Time-to-event data analysis is primarily used in biomedical and engineering-related science. The event related to death, brain tumor, the recovery from any specific disease after diagnosis, and some examples related to the engineering field or the survival of tube light after being made, or how many years, months, days, or minutes any electric machine or device will work, for such kind of analysis 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., comes under survival data analysis.

Survival Analysis is also used in other 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 needs to know that policymakers know the rate of frictionally unemployed or the duration from completion of a degree to getting a job, so policymakers are interested in that duration from degree completion to getting a job (Stock, Siegfried, & Finegan, 2011). 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, Bryant, May, & Perry, 2013). What will happen? If such data have an outlier and time-dependent covariate problem, the conventional survival analysis models failed to estimate unbiased and efficient results from the data. The results obtained using traditional survival analysis models from such data are spurious because of the outliers and time-dependent covariate problem. Due to these problems in time-to-event data, we need to modify our model per the data structure and requirements, such as outlier and time-dependent covariates, which can handle all the problems jointly and give unbiased, efficient estimates accurate survival time (Ghasemi, Fekri, Larizadeh, Dabiri, & Jahani, 2023).

## 1.1. Importance of Survival Analysis in Economics, Public Policy

Time-to-event models are primarily used in health-related data. It can be used in economics, public policy, psychology, and sociology. The data used for time-to-event models are mostly censored, and the dependent variable is the time variable, either hours, days, months, or years, depending on the nature of the study. The main problem in this kind of data series is an outlier and time-dependent covariate in the data. Due to these problems in data, we can't efficiently estimate the coefficients of the models. Our standard error increases, leading to spurious estimates. Therefore, we need to modify our models as per data structure, such as outliers, variation, and time-dependent covariates, which can handle all the problems jointly. To meet this gap, we have proposed a modified Cox regression model to solve these problems and give unbiased and efficient estimates compared to the existing time-to-event model in the literature (Akram, Ullah, & Taj, 2007).

Different authors addressed only one problem in the model at a time, such as an outlier and time-dependent covariate (Carrasquinha, Veríssimo, & Vinga, 2018). Still, we haven't seen

any work in the literature that has addressed the outlier[1] and time-dependent [2]covariate problem jointly, which can be possible in the data at one time. What to do in such a situation is the big question mark. In this paper, we are going to fill this gap. How do we estimate the hazard ratio if the data has an outlier and time-dependent covariate problem? Finding an unbiased, efficient, and consistent average survival time in a different field is challenging, whether it is a health issue, insurance companies, or any other engineering field (Husain, Thamrin, Tahir, Mukhlisin, & Apriani, 2018).

- Do the existing time-to-event models handle the problem of outliers and time-dependent covariates in the data?
- The proposed modified Cox regression is an improved model compared to the existing time-to-event models: Robust Cox and Time-Dependent Cox.

To compare the performance of widely used existing time-to-event models in the presence of Outliers and time-dependent covariates, either modified Cox model performed better in the case of outliers and time-dependent covariates.

## 2. Literature review

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. Gemar, Moniche, and Morales (2016) studied the Survival rate of the Spanish country hotel industry and the essential 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, Rigby, and Young (2003) studied modeling 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) 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. The Cox proportional hazards model was used to estimate age's hazard ratio (HR) on survival, adjusted for stage of disease, histology, residual tumor after surgery, and chemotherapy (Bender, Augustin, & Blettner, 2005). The assumption of proportional hazards was checked using Schoenfeld residuals and found valid. The results showed that older age was associated with worse survival outcomes, with an HR of 1.02 per year increase in age. This association remained significant after adjusting for other covariates. However, the study also found that patients with advanced-stage disease, non-epithelial histology, and residual tumor after surgery had worse survival outcomes (Bewick, Cheek, & Ball, 2004). Using survival analysis, Min, Zhang, Long, Anderson, and Ohland (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 & Meier, 1958). life tables and concluded that female is highly riskier than male. The author further concluded that student whose Scholastic Assessment Test 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.

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, Agampodi, and Piyaseeli (2007) studied Breastfeeding mothers in Sri Lanka in 2006, the sample

[1] An extreme abnormal negative or positive value in the data is called outlier.

[2] A predictor variable that varies over time are called time-dependent covariates. Such as body mass index, treatment dose, smoking status, depression, blood pressure, and socioeconomic status.

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. 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, Arreglado, Leroux, Murray-Ramcharan, and Engdahl (2020) discuss the adults affected by COVID-19 in the capital city of New York. The total number of young adults who were studied was 395. The study concluded that 57% of patients had at least one major comorbidity[3]. 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. Györffy et al. (2010) studied the survival 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.

## 2.1. Literature Review According to Pakistan

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.

## 2.2. Cox Regression

A Cox regression model is a statistical model used for censored data for 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, Moreau, & Lellouch, 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 additional covariates (Moncada-Torres, van Maaren, Hendriks, Siesling, & Geleijnse, 2021).

The dependent variable in the Cox regression is the time taken to the event of interest, denoted by lambda time $\lambda(t)$, hazard ratio gives the probability of an event of interest according

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

to the nature of the study. If it is related to survival, then the event of interest is recovery, which occurred before time t., and $\beta$ is the Cox regression coefficient.

Suppose we have $X_1$, $X_2$, $X_3$…. Xp covariates and the parameters for Cox regression is $\beta_1$, $\beta_2$, $\beta_3$…. $\beta_p$ parameters for Cox regression, such situation our Cox regression hazard function will be;

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

$$h(t) = h_0(t) \exp\{\beta_1 X_1\} \quad (1)$$

The Cox regression with multiple predictors can be written as

$$h(t) = h_0(t) \exp\{\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p\} \quad (2)$$

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

$$\frac{h(t)}{h_0(t)} = \exp\{\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p\} \quad (3)$$

In logarithm form, we can write equation 3 below.

$$\ln\left\{\frac{h(t)}{h_0(t)}\right\} = \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p \quad (4)$$

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

$$h(t) = h_0(t) \exp\{\sum_{i=1}^{p} \beta_i x_i\} \quad (5)$$

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{$\beta_i$} is the hazard ratio, $\beta_i$= is the vector of Cox regression parameters. At the same time, $X_i$ is the number of a predictor, and Covariates of Cox regression are estimated using the maximum likelihood procedure.

## 3. Methodology

The methodology section has two stages; the first stage is the detection stage, whether the data have an outlier and time-dependent covariate problem or not. In the next stage, we solved the two issues jointly in the proposed modified Cox regression model. The following test is used as a graphical and iterative procedure for detecting outliers in the data Urooj and Asghar (2017) for the time-dependent covariate detection, the Schoenfeld residuals test by Fisher and Lin (1999) is used. The four methodologies are estimated: Cox regression, Robust Cox, Time-Dependent Cox, and modified Cox regression. They are compared using simulated data in the presence of an outlier and time-dependent covariates. The detail of these methodologies is given below.

### 3.1. Detection of Outlier and time-dependent covariates

We have used an influence plot to detect an outlier in the data, and for time-dependent covariate detection, we have used the methodology of (Therneau, Crowson, & Atkinson, 2017). ZPH Cox function, if any variable P value is less than 5%, will be considered a time-dependent covariate, and we will take it as a function of time in the modified Cox model.

## 3.2. An Algorithm based on the Modified Cox Proportional Hazard Model (MCPHM)

The general form of Cox regression can be written as follows, recall equation (5).

$$h(t) = h_0(t)\exp\{\sum_{i=1}^{p}\beta_i x_i\} \quad (5)$$

It can be written as below to incorporate the time-dependent covariate term in equation 5.

$$h(t) = h_0(t)\exp^{\{\beta_1 X_1+\beta_2 X_2 \ldots\ldots+\beta_p X_p+\gamma_1 z_1(t)+\gamma_2 z_2(t)+\cdots+\gamma_k z_k(t)\}.} \quad (6)$$

In general form, equation 6 can be written.

$$h(t) = h_0(t)\exp\left[\sum_{a=1}^{p1}\beta_{ai}x_{ai} + \sum_{b=1}^{p2}\gamma_{bi}z_{bi}(t_j)\right] + \mu_i \quad (7)$$

$$ln[\frac{h(t)}{h_0(t)}] = \exp\left[\sum_{a=1}^{p1}\beta_{ai}x_{ai} + \sum_{b=1}^{p2}\gamma_{bi}z_{bi}(t_j)\right] + \mu_i \quad (8)$$

Let $\beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p = X_i$ (9)

$\gamma_1 z_1(t) + \gamma_2 z_2(t) + \cdots + \gamma_k z_k(t) = Y_D$ (10)

$$ln[\frac{h(t)}{h_0(t)}] = T_i \quad (11)$$

$$\mu_i = \nu_i \quad (12)$$

$$T_i = X_i + Y_{Di} + \nu_i \quad (13)$$

To introduce the winsorization[4] effect in the equation, we will use winsorization on a different level, such as 5% or 10%, depending on the distinct nature of the outlier. Urooj and 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.

$$T_{(.10)} = X_i + Y_{Di} + \nu_i \quad (14)$$

For the estimation of the final model, equation 13 is used to avoid the problem of an outlier and time-dependent covariates in the model.

## 3.3. Data Generating Process (DGP)

The question we are trying to answer is which model is best in time-to-event studies if there is a problem of an outlier and time-dependent covariates. First, we generate a random series and introduce the outliers and time-dependent covariates.

### 3.3.1. Data Generating Process for Outlier

For Introducing outlier in the independent variable (Lin & Wei, 1989)

$$X_i = x_1, x_2, x_3, \ldots\ldots x_n.$$

Now to introduce weights of outlier, $X_i = k*x_1, k*x_2, k*x_3, \ldots\ldots k*x_n$ , The variance-covariance matrix given below:

Var-Cov( )= X*K

Where k is the weight of the outlier

$$\text{Var-Cov}(\ ) = \begin{bmatrix} x_{11} & x_{12} & x_{1k} \\ x_{21} & x_{22} & x_{2k} \\ x_{n1} & x_{n2} & x_{nk} \end{bmatrix} n*k \begin{bmatrix} \delta_i \\ \delta_i \\ \delta_i \end{bmatrix}$$

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

In such a way, an outlier case will be generated.

### 3.3.2. 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 from time to time are not constant, such as smoking patterns, which may increase or decrease with time. We use methodology to introduce a time-dependent covariate in the independent variable (Moreno-Betancur et al., 2018). In the time-independent covariate assumption, the variable age =age, which means that age is constant throughout the study. 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. we have used the methodology of (Therneau et al., 2017) to generate the time-dependent covariate in the model.

## 4. Results

Table 1 and Figure 1 compared four methodologies using simulated data: Cox regression, robust Cox, time-dependent Cox, and modified Cox. The sample size is 100, the magnitude of the outlier is 4 SD, the exponential distribution parameter theta is taken 1 and 2, the outlier's quantity is 5%, 10%, and 20%, and the number of simulated series is 50,000, the results comparison is based on RMSE, MAE and MAPE.

**Table 1**
***Comparison of four methodologies based on RMSE, MAE and MAPE***

| 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 | 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

We simulated the four methodologies 50,000 times using simulated data and then compared the four models using 50,000 average samples RMSE, MAE and MAPE in Figure 1, the average RMSE, MAE and MAPE for the modified Cox model outperform compared to the other three models. The average value of RMSE, MAE and MAPE for the Cox model is 61.4, 25.7 and 117.9. For the robust Cox, the RMSE, MAE and MAPE value is 45.3, 19.0 and 87.0. For the time-dependent Cox model, RMSE, MAE and MAPE value is 51.0, 21.4 and 98.0. For the modified Cox model, the RMSE, MAE and MAPE value is 42.4, 17.7 and 81.4, which perform better than Cox, Robust, and Time-Dependent Cox, the modified Cox model RMSE, MAE and MAPE is lower than other survival analysis models.

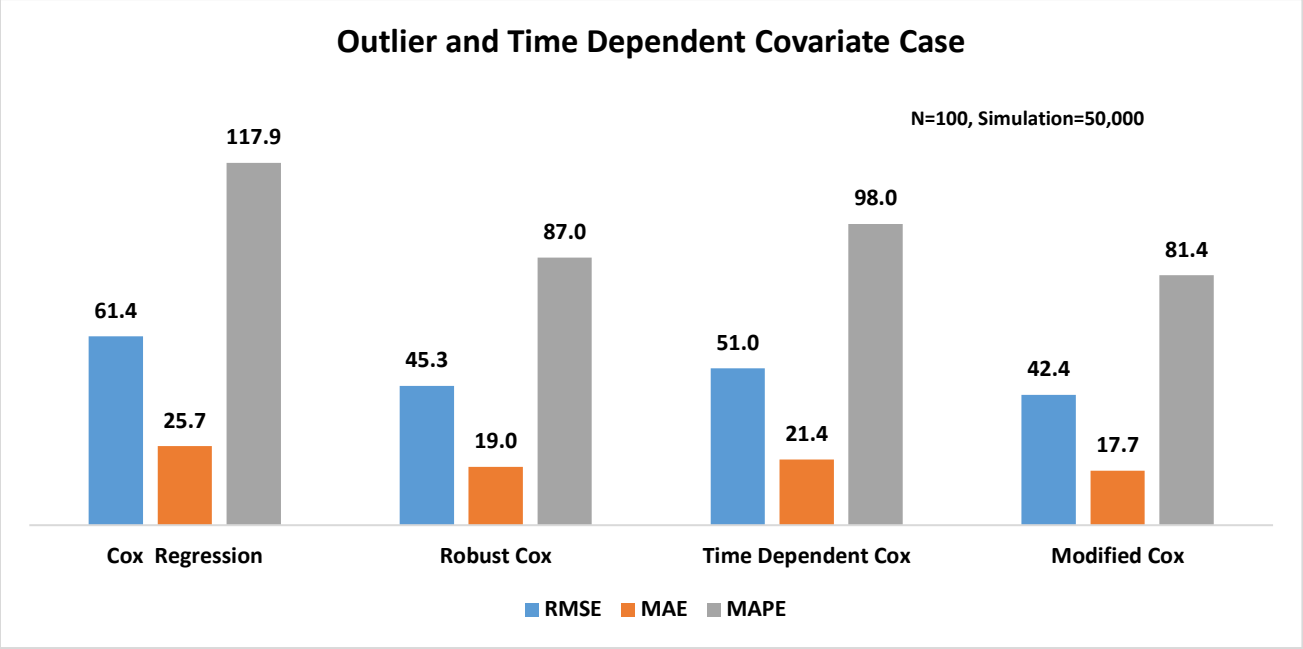


**Figure 1: Comparison of four different methodologies based on RMSE and MAE**

The modified Cox proportional hazards model extends the Cox model, allowing for time-varying covariates and outliers in the data. This model can handle changes in the effect of covariates over time and outliers, which may provide a better fit for the data than the Cox, time-dependent, and robust Cox models. The modified Cox proportional hazards model is the best choice in this comparison. This model can handle time-varying covariates and outliers, which may be essential in many applications. Additionally, the modified Cox model extends the widely used Cox model. Overall, the modified Cox model is a powerful and flexible model that can provide valuable insights into survival data, as shown in Figure 1. However, we have used modified Cox on real data of lungs in the survival analysis package of RStudio to detect whether an outlier and time-dependent covariate problem exists in the data.

## 4.1. Outlier and Time-Dependent Covariate Detection

We have used an influence plot to detect the outliers in the data. The student residual value, if greater than three, should be considered an outlier, so it is confirmed that there is an outlier in the data.

**Table 2**

***Influence Plot for Outlier Detection***

| Observation | Student Residuals |
|---|---|
| 3 | 3.6050458 |
| 9 | 0.8043773 |
| 25 | 3.0157674 |
| 150 | -1.6247218 |

The next step is to check the variable for time-dependent covariate among all. The list of independent variables in the study is age, gender, physician-rated score, patient-rated score, meat in calories, and weight in pounds, so after testing for time-dependent covariates for all variables in the model, we have used the Schoenfeld residual test (Cox ZPH test), which tell us that the variable ph. Karno, the physician-rated score, is a time-dependent covariate that changes occasionally. The table below shows that the ph. Karno variable is statistically significant at a 5% significance level.

**Table 3**
***ZPH Function using Schoenfeld residual test for Time Dependent***

| Variable | Chisq | df | p |
|---|---|---|---|
| age | 0.478 | 1 | 0.4892 |
| ph.karno | 8.017 | 1 | 0.0046 |
| sex | 3.085 | 1 | 0.0790 |
| GLOBAL | 2.35 | 3 | 0.157 |

We have addressed the problem of outliers and time-dependent covariates in the model and estimated the coefficient using the modified Cox model in Table 3 below.

**Table 4**
***Modified Cox Model Estimates***

| Variables | Modified Cox Model Time (y) |
|---|---|
| Age | 0.011 |
| -1.701 | |
| Sex | -0.628*** |
| -31.393 | |
| ph. Ecog | 0.741*** |
| -34.321 | |
| tt(ph.jarno | -0.010 |
| -6.098 | |
| pat.karno | -0.012 |
| -1.15 | |
| wt.loss | -0.022*** |
| -1.221 | |
| Observation | 167 |
| R2 | 0.10 |
| Log Likelihood | 0.033 |
| Wald Test | 23.760*** (df = 8) |
| LR Test | 0.002 (df = 8) |
| Score (lofrank) Test | 0.001 (df = 8) |

**Data Source:** Lungs Cancer Data, Library Survival, *p<0.1; **p<0.05; ***p<0.01

## 4.2. Discussion

In general, though, it's important to interpret the coefficients of a Cox model in terms of their hazard ratio rather than their raw value. The hazard ratio represents the relative increase or decrease in the hazard (risk of an event) associated with a one-unit increase in the predictor variable. So, the age value is 0.011, and a hazard ratio of 0.011 would indicate that a one-year increase in age is the associated risk of a 1.1% increase in the hazard of the event of interest, holding all other variables constant. Here, the event of interest is death.

Sex(gender) If a person is a male, it is associated with a 62.8% decrease in the hazard of the event of interest, which is death. (Ph. ECOG) performance score as rated by the physician. With one unit increase in ph. ecog, the associated risk towards the event of interest (death) is 74.1%. (Ph. Karno) performance score as rated by the physician, Karnofsky performance score considered zero as bad and 100 as good, so one unit increase in (ph karno), the associated risk towards the event of interest (death) is decreased by 1%. (Pat Karnofsky) performance score as rated by the patient, one unit increase in pat karnofsky, the associated risk towards the event of interest (death) is decreased by 1.2%. Weight loss in the last six months in (pounds) one pound increase in weight loss, the associated risk towards the event of interest (death) is decreased by 2.2%, the results are stable with the previous study of (Fagbamigbe et al., 2019; Györffy et al., 2010; Panjer, 1987).

## 5. Conclusion

The study concluded that the modified Cox model outperforms in the case of outlier and time-dependent covariate, the modified Cox model is an improved version of the Cox model, The variable Karnofsky performance score rated by the physician is a time-dependent covariate, age, and ECOG performance score has a positive impact on the event of interest and other variables such as gender, ph. Karno, Pat. Karnofsky and weight loss negatively impact the event of interest (death). The variables gender, Ph.ecog, and weight loss are statistically significant, which means lung cancer patients need to focus more on Ph.ecog and weight loss indicators to recover successfully.

### 5.1. Future Research Direction

The work can be further extended for heteroscedasticity and time-dependent covariates.
Or three problems jointly, like outlier, heteroscedasticity, and time-dependent covariate, so that different dimensions can be studied.

**Authors Contribution**

Nauman Ahmad: Manuscript preparation, Introduction, literature, methodology and data analysis.
Amena Urooj: Study design, Concept topic idea, conclusion, supervision direction and proofreading.

**Conflict of Interests/Disclosures**

The authors declared no potential conflicts of interest w.r.t the research, authorship and/or publication of this article.

**REFERENCES**

Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). An introduction to survival and event history analysis. *Survival and Event History Analysis: A Process Point of View*, 1-39. doi:https://doi.org/10.1007/978-0-387-68560-1_1

Agampodi, S. B., Agampodi, T. C., & 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. doi:https://doi.org/10.1186/1746-4358-2-13

Akram, M., Ullah, M. A., & Taj, R. (2007). Survival analysis of cancer patients using parametric and non-parametric approaches. *Pakistan Veterinary Journal, 27*(4), 194.

Altonen, B. L., Arreglado, T. M., Leroux, O., Murray-Ramcharan, M., & Engdahl, R. (2020). Characteristics, comorbidities and survival analysis of young adults hospitalized with COVID-19 in New York City. *PloS one, 15*(12), e0243343. doi:https://doi.org/10.1371/journal.pone.0243343

Annesi, I., Moreau, T., & Lellouch, J. (1989). Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Statistics in medicine, 8*(12), 1515-1521. doi:https://doi.org/10.1002/sim.4780081211

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine, 24*(11), 1713-1723. doi:https://doi.org/10.1002/sim.2059

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *Critical care, 8*(9), 1-6. doi:https://doi.org/10.1186/cc2955

Burton, M., Rigby, D., & 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. doi:https://doi.org/10.1111/1467-8489.00202

Carrasquinha, E., Veríssimo, A., & Vinga, S. (2018). Consensus outlier detection in survival analysis using the rank product test. *bioRxiv*, 421917. doi:https://doi.org/10.1101/421917

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-202. doi:https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

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. doi:https://doi.org/10.1038/bjc.2014.208

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 key biomedical data in advanced colorectal cancer? *European Journal of Cancer, 42*(1), 42-49. doi:https://doi.org/10.1016/j.ejca.2005.07.025

Fagbamigbe, A., Abi, R., Akinwumi, T., Ogunsuji, O., Odigwe, A., & Olowolafe, T. (2019). Survival analysis and prognostic factors associated with the timing of first forced sexual act among women in Kenya, Zimbabwe and Cote d 'Ivoire. *Scientific African, 4*, e00092. doi:https://doi.org/10.1016/j.sciaf.2019.e00092

Fisher, L. D., & Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health, 20*(1), 145-157. doi:https://doi.org/10.1146/annurev.publhealth.20.1.145

Gemar, G., Moniche, L., & Morales, A. J. (2016). Survival analysis of the Spanish hotel industry. *Tourism Management, 54*(6), 428-438. doi:https://doi.org/10.1016/j.tourman.2015.12.012

Ghasemi, J., Fekri, M. S., Larizadeh, M. H., Dabiri, S., & Jahani, Y. (2023). An Integrative Bayesian Model Analysis of Patient Characteristics and Treatment Variables to Understand Lung Cancer Survival Rates in Kerman Province, Iran. *Journal of Biostatistics and Epidemiology, 8*(4), 445-457. doi:https://doi.org/10.18502/jbe.v8i4.13357

Győrffy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., & 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*(12), 725-731. doi:https://doi.org/10.1007/s10549-009-0674-9

Husain, H., Thamrin, S. A., Tahir, S., Mukhlisin, A., & 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.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association, 53*(282), 457-481. doi:https://doi.org/10.1080/01621459.1958.10501452

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. doi:https://doi.org/10.1080/01621459.1989.10478874

Min, Y., Zhang, G., Long, R. A., Anderson, T. J., & Ohland, M. W. (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of Engineering Education, 100*(2), 349-373. doi:https://doi.org/10.1002/j.2168-9830.2011.tb00017.x

Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific reports, 11*(1), 6968. doi:https://doi.org/10.1038/s41598-021-86327-7

Moreno-Betancur, M., Carlin, J. B., Brilleman, S. L., Tanamas, S. K., Peeters, A., & Wolfe, R. (2018). Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics, 19*(4), 479-496. doi:https://doi.org/10.1093/biostatistics/kxx046

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

Schober, P., & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia and analgesia, 127*(3), 792. doi:https://doi.org/10.1213/ANE.0000000000003653

Solomon, G. T., Bryant, A., May, K., & Perry, V. (2013). Survival of the fittest: Technical assistance, survival and growth of small businesses and implications for public policy. *Technovation, 33*(8-9), 292-301. doi:https://doi.org/10.1016/j.technovation.2013.06.002

Stock, W. A., Siegfried, J. J., & Finegan, T. A. (2011). Completion rates and time-to-degree in economics PhD programs. *American Economic Review, 101*(3), 176-187. doi:https://doi.org/10.1257/aer.101.3.176

Therneau, T., Crowson, C., & Atkinson, E. (2017). Using time dependent covariates and time dependent coefficients in the cox model. *Survival Vignettes, 2*(3), 1-25.

Urooj, A., & Asghar, Z. (2017). Analysis of the performance of test statistics for detection of outliers (additive, innovative, transient, and level shift) in AR (1) processes. *Communications in Statistics-Simulation and Computation, 46*(2), 948-979. doi:https://doi.org/10.1080/03610918.2014.985383

Yusuf, M. A., Badar, F., Meerza, F., Khokhar, R. A., Ali, F. A., Sarwar, S., & Faruqui, Z. S. (2007). Survival from hepatocellular carcinoma at a cancer hospital in Pakistan. *Asian Pacific Journal of Cancer Prevention, 8*(2), 272.