BAREKENG: Journal of Mathematics and Its Applications
March 2025 Volume 19 Issue 1 Page 0629–0642
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss1pp0629-0642

COMPARISON OF SURVIVAL ANALYSIS USING ACCELERATED FAILURE TIME MODEL AND COX MODEL FOR RECIDIVIST CASE

Nuraziza Arfan[1*], Asrul Irfanullah[2], Muhammad Rozzaq Hamidi[3], Utriweni Mukhaiyar[4]

[1,2,3] *Actuarial Master Study Program, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung*
[4] *Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung*
Jln. Ganesha No. 10, Bandung, 40131, Indonesia

Corresponding author's e-mail: * **zizarfan@gmail.com**

ABSTRACT

Article History:

Received: 21st August 2024
Revised: 28th November 2024
Accepted: 28th November 2024
Published: 13th January 2025

Keywords:

Accelerated Failure Time;
AIC;
BIC;
Cox PH;
Recidivist;
Regression;
Stratified Cox.

Recidivists, or ex-prisoners who commit crimes after serving a prior sentence, pose a critical challenge to the criminal justice system. This study examines social and economic factors that may reduce the likelihood of recidivists being re-arrested. Using survival analysis, the probability that a recidivist could survive in society without being re-arrested could be estimated. The purpose of this work is to compare the AFT and Cox models to determine which provides a better fit to identify factors affecting the likelihood of re-arrest within one year after release and to statistically assess the impact of these factors. This study utilizes a stratified Cox model to address variables that violate the proportional hazards (PH) assumption. The analysis is limited to four types of AFT models: Weibull, log-normal, log-logistic, and exponential. Results show that the stratified Cox model provides the best fit, based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). This demonstrates the Cox model's robustness in analyzing survival data, accurately approximating the distribution of survival times without restrictive assumptions, unlike AFT models. The study reveals that recidivists who received financial aid upon release have a 0.66644 *lower risk of re-arrest compared to those who did not, and each additional prior theft arrest increased the risk of re-arrest by 1.09193 times.*

How to cite this article:

N. Arfan, A. Irfanullah, M. R. Hamidi and U. Mukhaiyar., "COMPARISON OF SURVIVAL ANALYSIS USING ACCELERATED FAILURE TIME MODEL AND COX MODEL FOR RECIDIVIST CASE," *BAREKENG: J. Math. & App.*, vol. 19, iss. 1, pp. 0629-0642, March, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal A-mail: *Barrekeng.math@yahoo.com;Barekeng.Journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

It is believed that poverty contributes to crime, as crime rates are often higher in poorer neighborhoods, and prisons are predominantly occupied by individuals from the lowest economic classes [1]. However, poverty is linked to various individual characteristics associated with criminal behavior; therefore, it is challenging for researchers to identify the personal factors that directly contributed to the likelihood of these individuals being able to survive in society after release.

Survival analysis in general is an analytical method that observes the time span of events expected to occur [2]. The Cox model and Accelerated Failure Time (AFT) are some of the models widely used to test covariate effects at failure time. In the Cox Proportional Hazard (PH) model, the covariates affect the hazard function, while in the AFT model, covariates affect directly the estimated survival time [3].

The Cox model is a model in which the data distribution is unknown; therefore, the form of baseline hazard is unknown. Nonetheless, the Cox PH model produces estimates close to the actual data distribution [2]. Although this model offers certain advantages, it is recognized that the Cox model has some drawbacks, one of which is the requirement to examine the PH assumption [4]. If the PH assumption is not satisfied by the predictor variables, Stratified Cox is used to overcome the issue [2]. The stratified Cox regression model deals with variables that do not satisfy the assumption by stratifying the data into subgroups/strata. Despite not being incorporated into the model, they contribute through strata.

Another alternative is to use AFT models when the hazard is not proportional [5]. In the AFT model, covariate effects are assumed to be constant and multiplicative on a time scale, modeled by an acceleration factor [6]. In addition, the AFT model assumes survival time follows a specific distribution to estimate parameters. Since it is a parametric model, it is expected to give a better analysis than the nonparametric or semiparametric one [7]. Failure time distribution is assumed to have a homogenous population. Some specific distributions are frequently used, such as Exponential, Weibull, Log-normal, and Gamma [8].

Survival analysis using the Cox model and the AFT model has been carried out, such as Cox PH and AFT analysis on under-five mortality data in Uttar Pradesh, where the log-normal model of AFT gave the smallest AIC [9]. According to [10], AFT gave an optimal outcome compared to Cox PH in modeling the survival time among melioidosis patients. In addition, the recent study by the AFT model gave a better fit than the Cox PH model in identifying factors of malaria reinfection in Congo [11]. Another previous research conducted by [12] compared the application of Cox and AFT models and revealed that log-normal AFT was chosen as the best model to estimate student performance at universities due to the violation of the PH assumption. In another study by [13], the violation of the PH assumption is extended using stratified Cox in modeling second-line antiretroviral therapy data in South Wollo Zone Public Hospitals.

This study utilized experimental data from The Baltimore LIFE Project (Living Insurance for Ex-Prisoners), a project funded by the U.S. Department of Labor in 1971, which was conducted to determine whether offering a certain amount of financial aid to released prisoners would reduce recidivism and support them to survive later in society. The data involved 432 inmates released from prison in Maryland and observed for one year after release. Research subjects were limited to male recidivists under 45 years with several punishments, one of which was related to property crime—crimes committed to enrich the perpetrators, such as theft, robbery, and property-related crime [14].

Earlier studies on recidivism behavior have been explored by [15] through several statistical approaches, such as the logistic regression model, the Cox regression model, and the cure rate model. Another analysis by [16] compares different techniques to predict recidivism in young offenders, utilizing logistic regression, random forest, and boosted classification trees. Additionally, a separate study conducted by [17] investigates the impact of demographic, psychosocial, and driving characteristics of Spanish drivers on their risk perception related to recidivist traffic offenders. Furthermore, [18] conducted a quantitative analysis using machine learning to examine topics related to the risk of habitual offenders.

Several studies have utilized experimental data from The Baltimore LIFE Project, including [19], who applied the Cox PH model with both time-independent and time-dependent covariates, and [20], who employed the Stratified Cox method. In this research, the analyzed covariates included financial aid, age, race, work experience before arrest, marital status at the time of release, parole status, and the number of previous theft arrests.

The purpose of this study is to model the length of time the observed recidivists were re-arrested for one year after release and identify the factors that affect recidivists being re-arrested within one year after release using hazard ratios from the Stratified Cox model and time ratios from the AFT model. The considered AFT models in this study are limited to four commonly used distributions as identified in prior studies, such as Weibull, log-normal, log-logistics, and exponential. Additionally, the data used in this research is limited to observations from a single year following the release of prisoners.

The rest of the paper is constructed as follows. In Section 2, the research methods and procedure are stated. In Section 3, the analysis and comparison are demonstrated through the AFT model and Cox model, and issues that did not meet the PH assumption are discussed. Finally, Section 4 concludes the paper with a brief discussion of further research.

2. RESEARCH METHODS

2.1 Cox PH Model

One of the models widely used for survival data analysis is Cox model. The selection of this model was motivated by the ease of understanding the interpretation, the existence of some empirical evidence to support Proportional Hazard assumptions used in certain fields and censoring that is accommodated by model. According to **[21]**, Cox PH regression is used to determine the relationship between response and predictor variables, utilizing survival time data for individuals. In general, Cox regression model is faced with a situation where the probability of an individual's death at a time is influenced by one or more predictor variables. The Cox Regression model is **[2]**:

$$\begin{aligned} h_i(t|\boldsymbol{X}) &= h_0(t) * exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) \\ &= h_0(t)\exp\left(\sum_{j=1}^{p} \beta_j x_{ji}\right) \end{aligned} \tag{1}$$

where

$h_i(t\|\boldsymbol{X})$	:	$i - th$ individual failure function
$h_0(t)$	:	baseline hazard function (hazard function at $x_{ji} = 0$)
x_{ji}	:	the value of the $j - th$ independent variable from the $i - th$ individual, $j = 1,2,\dots,p$ and $i = 1,2,\dots,n$
β_j	:	$j - th$ regression coefficient

The failure function in this model can take any form, but the failure functions of different individuals are assumed to always be proportional.

Hazard ratio is defined as the ratio of hazard of an individual to another individual, namely [2]:

$$\widehat{HR} = \frac{h_A(t|X^*)}{h_B(t|X)} = \frac{h_0(t)\exp\{\sum_{i=1}^{p}\beta_i X_i^*\}}{h_0(t)\exp\{\sum_{i=1}^{p}\beta_i X_i\}} = \exp\left\{\beta_i\left(\sum_{i=1}^{p} X_i^* - X_i\right)\right\} \tag{2}$$

If the hazard ratio values between the two groups being compared remains constant over time, then the predictor variable $X_1, X_2, \dots, X_p$ have satisfied the PH assumption. This PH assumption test is performed with global test statistics based on Schoenfeld residuals formulated by:

$$R_{ij} = \delta_i\left(x_{ji} - \frac{\sum_{q\in R(t_i)} x_{qj}\exp(\hat{\beta}' x_q)}{\sum_{q\in R(t_i)}\exp(\hat{\beta}' x_q)}\right) \tag{3}$$

with

R_{ij} : Schoenfeld residuals for individuals who experience *events* at the time t_i on the $j - th$ predictor variables

δ_i : censorship indicator

$R(t_i)$: the set of individuals at risk at the time t_i

$\hat{\beta}$: estimator partial maximum likelihood of β

The hypotheses testing is:

H_0 : proportional hazard assumption is satisfied

H_1 : proportional hazard assumption is not satisfied

The statistics used is

$$z = r\sqrt{n-1} \tag{4}$$

With r is defined as

$$r = \frac{\sum_{i=1}^{r}\left(R_{ij} - \bar{R}_{ij}\right)\left(RT_i - \overline{RT}_i\right)}{\sqrt{\sum_{i=1}^{r}\left(R_{ij} - \bar{R}_{ij}\right)}\sqrt{\sum_{i=1}^{r}\left(RT_i - \overline{RT}_i\right)}} \tag{5}$$

$\bar{R}_{ij}$: Schoenfeld residual mean for individuals who experienced an event at time t_i on the $j-th$ independent variable

RT_i : survival time ranking for the time of the $i-th$ event

$\overline{RT}_i$: Average survival time ranking for the time of the $i-th$ event

If the value $|Z| > Z_{\alpha/2}$ or $p-value \leq \alpha$, then H_0 is rejected. In other words, the PH assumption is not satisfied, indicating a correlation between Schoenfeld's residual and the survival time ranking.

2.2 Stratified Cox

The Stratified Cox model is a modification of Cox PH model that accommodates multiple strata. The strata form the variables into disjoint groups, containing unique baseline hazard function [22]. In this model, variables that satisfy the PH assumption are included in the model, while variables that do not meet the proportional hazard assumption are stratified. The stratified Cox model identifies variables that increase the likelihood of certain events while accounting for the effects of variables that violate PH assumptions.

Suppose there is a p covariates, Cox PH model formed as in **Equation (2)**. If there are m covariates that meet the PH assumption and k covariates that violate the PH assumption, that is $k = p - m$ then the variables that do not meet the PH assumption will then be denote as $Z_1, Z_2, \dots, Z_k$, and the variables satisfying PH assumption will be denoted as $X_1, X_2, \dots, X_m$. If the model does not account for interactions between covariate variables and variables that violate PH assumption, it is referred to as a stratified Cox model without interaction:

$$h_g(t|\boldsymbol{X}) = h_{0g}(t) * \exp\{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\} \tag{6}$$

with $g = 1,2,\dots,k^*$, where k^* is the total number of combination or strata formed after categorizing each Z. Then define a single new variable Z^*, representing new variable with k^* categories stratify on Z^*. For each different stratum g, the baseline hazard function expressed by $h_{0g}(t)$ from each model is also different. However, the coefficients β_j remain the same across all strata [2].

2.3 Accelerated Failure Time (AFT) Model

The AFT model outlines the relationship between the survival time and the response variable. This model assumes that the covariate effect is multiplicative with respect to the survival time [23]. For covariates $(X_1, X_2, \dots, X_p)$, the AFT model mathematically is [7]:

$$S(t|X) = S_0(t|\eta(x)) \tag{7}$$

where $S_0(t|\eta(x))$ is the baseline survival function with η is an acceleration factor formulated by

$$\eta(x) = \exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\} \tag{8}$$

Based on the relationship between the survival function and the hazard function with the covariate of the hazard function $(X_1, X_2, \dots, X_p)$, the hazard function is obtained as follows [24]:

$$h(t|X) = \left[\frac{1}{\eta(x)}\right] h_0\left(\frac{t}{\eta(x)}\right) \tag{9}$$

In an AFT model, the effects of covariates are assumed to be constant and multiplicative on the time scale, meaning that covariates affect survival by a fixed factor.

a. Weibull AFT Model

Let the survival time $T \sim W(\lambda, \gamma)$, under AFT model, the hazard function for the i^{th} individual is

$$h_i(t) = \left[\frac{1}{(\eta_i(x))^{\gamma}} \lambda\gamma t^{\gamma-1}\right] \tag{10}$$

with γ representing the shape parameter of the Weibull distribution [5].

b. Log-logistic AFT Model

Let the survival time $T \sim Log - logistic(\theta, k), k > 0$, under AFT model, the hazard function for the i^{th} individual is [5]

$$h_i(t) = \frac{e^{\theta - \kappa \ln(\eta_i)} \kappa t^{k-1}}{1 + e^{\theta - \kappa \ln(\eta_i)} t^k} \tag{11}$$

c. Log-normal AFT Model

Let the survival time assumes to have a log-normal distribution, the hazard function is given by [5].

$$h_0(t) = \frac{\phi\left(\frac{\ln(t)}{\sigma}\right)}{\left[1 - \Phi\left(\frac{\ln(t)}{\sigma}\right)\right]\sigma t} \tag{12}$$

Where σ is parameter, $\phi(x)$ is the probability density function and $\Phi(x)$ is the cumulative density function of the standard normal distribution [5].

d. Exponential AFT Model

Let the survival time assumed to have exponential distribution, the hazard function is $\lambda(t) = \lambda, \lambda > 0$. From **Equation (5)**, the hazard function of Exponential AFT model is given by [5]

$$\lambda(t; x) = \lambda e^{\beta' x} \tag{13}$$

2.4 Parameter Estimation with Maximum Partial Likelihood Estimator

Suppose $X_i(t) = (X_{1i}(t), X_{2i}(t), \dots, X(t)_{pi})'$ be the $p \times 1$ covariate vector of the $i - th$ individual at time t. Let $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ denote the $p \times 1$ vector of regression coefficient and let D_{obs} denote the observed data. Let $R(t_i)$ be the risk set at time t_i which consist of all individuals who are at risk at time t_i [25]. The partial likelihood of the Cox regression is:

$$L(\beta|D_{obs}) = \prod_{i \in D} \frac{exp\left\{\sum_{j=1}^{p} \beta_i x_{jl}\right\}}{\sum_{q \in R(t_i)} exp\left\{\sum_{j=1}^{p} \beta_j x_{jq}\right\}} \tag{14}$$

While the Partial Likelihood of the AFT model is

$$L(\beta|D_{obs}) = \prod_{i=1}^{n} \frac{f(t_i|x_i)}{\sum_{q \in R(t_i)} f(t_i|x_i)} \tag{15}$$

Where $f(t_i|x_i)$ is a density function of the distribution used.

The Maximum Partial Likelihood Estimator (MPLE) $\hat{\beta}$ of β is obtained by solving the equation:

$$\frac{\partial}{\partial \beta} \ln \left(L(\beta | D_{obs}) \right) = 0 \tag{16}$$

2.5 Parameter Significance Testing

Parameter testing is used to determine the significant differences in parameters on both model. Testing the significance of the β parameter is carried out simultaneously with the partial likelihood ratio test and individually with the Wald test.

a. Simultaneous Parameter Testing

Simultaneous testing aims to determine whether predictor variables simultaneously affect response variables in Cox and AFT regression models. The hypotheses underlying this test are:

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1 : at least there is a $j = 1,2, \ldots, p$ so that $\beta_j \neq 0$

The test statistics are carried out with a partial likelihood ratio test which is notated by G, namely:

$$G = 2\left[\ln L(\hat{\beta}) - \ln L(0)\right] \tag{17}$$

with

$\ln L(0)$: log partial likelihood of the Cox/AFT regression model without predictor variables, i.e. $\beta_j = 0$

$\ln L(\hat{\beta})$: log partial likelihood of the Cox/AFT regression model containing predictor variables

If the value $G > \chi^2_{p,\alpha}$ or $p - value \leq \alpha$, then $H_0$is rejected, in other words there is at least one predictor variable significant to the model fit.

b. Individual Parameter Testing

This test was carried out to determine the effect of individual predictor variables on response variables. The hypotheses underlying this test are:

H_0 : $\beta_j = 0$

H_1 : $\beta_j \neq 0$

The test statistics are carried out with the *Wald* Test which is denotated by W, namely:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{18}$$

If the value $|W| > Z_{\frac{\alpha}{2}}$ or $p - value \leq \alpha$, then H_0 is rejected, in other words the predictor β_j significantly improves model fit.

2.6 Research Data and Variables

This study used secondary data from The Baltimore LIFE (Living Insurance for Ex-Prisoners) Project 1971 [14]. The data was originated from a study of 432 prisoners related to property/theft crime who had been released and limited to male sex with a maximum age of 45 years, and 50% of the study subjects were given financial aid. The variables used in this study are as follows.

Table 1. Research Variables

Variable	Variable name	Category	Scale	Unit
Y	Survival time	-	Ratio	Week
d	Status	0: Not arrested 1: Arrested	Nominal	-
X_1	Financial Aid	0: No 1: Yes	Nominal	-
X_2	Age	-	Ratio	Year
X_3	Race	0: Other 1: Blacks	Nominal	-

Variable	Variable name	Category	Scale	Unit
X_4	Experience	0: No 1: Yes	Nominal	
X_5	Married	0: No 1: Yes	Nominal	-
X_6	Parole	0: No 1: Yes	Nominal	-
X_7	Prior	-	Ratio	

2.7 Research Procedure

The procedure in this study is as follows:

a. Describe the characteristics of research data.

b. Construct an Accelerated Failure Time (AFT) model:

   i. Estimate AFT parameters from Weibull, exponential, log-normal, and log-logistic distributions.

   ii. Choose the best AFT model based on the smallest AIC and BIC value.

   iii. Create the best AFT model.

c. Construct a Cox Regression Model:

d. Test proportional hazard assumptions.

e. Handle predictor variables that do not meet proportional hazard assumptions.

f. Estimate Cox regression parameters.

g. Select the best Cox regression model based on the smallest AIC and BIC value.

h. Create the best Cox regression model.

i. Compare the AFT and Cox model based on the AIC and BIC value.

j. Interpret the best model.

All statistical procedures are performed using R and Matlab Program.

3. RESULTS AND DISCUSSION

3.1 Data Description

An overview of the data used in this study is presented in **Figure 1** below. Among 432 observations, 318 (approximately 73.6%) were censored, meaning that these recidivists were not re-arrested within a 52-week observation period. Furthermore, the proportion of recidivists receiving and not receiving financial aid is equal, each at 50%. However, recidivists who did not receive financial aid tend to return to prison more frequently within 52 weeks, with 66 cases compared to 48 cases for those receiving financial aid. Regarding work experience, recidivists without work experience (42.82%) are re-arrested more often (62 cases) than those with work experience (57.18%).

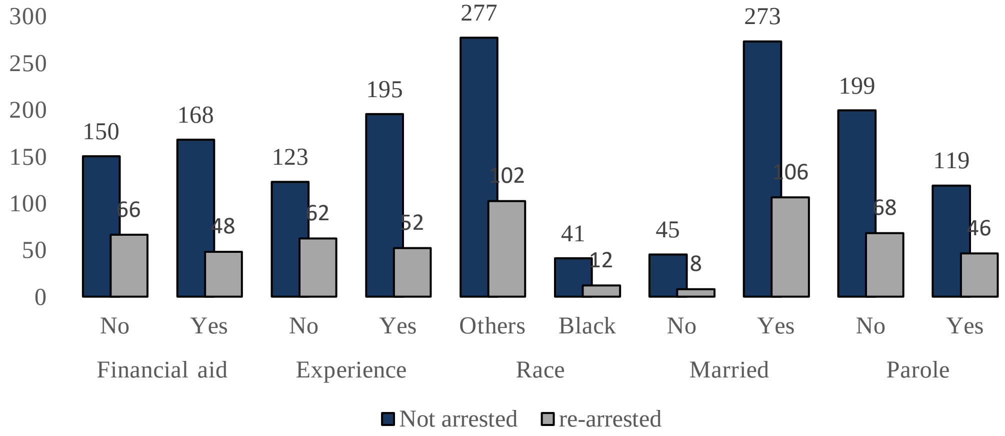


Figure 1. Descriptive Analysis of Financial Aid, Work Experience, Race, Married and Parole

The distribution shows that most recidivists are non-black (87.73%) and married (87.73%). Within 52 weeks, there were 102 cases of non-black recidivists who were re-arrested and 106 cases of married recidivists, compared to 12 cases of black recidivists and 8 cases of unmarried recidivists. Additionally, recidivists on parole (38.19%) are re-arrested less frequently than those released without conditions (61.81%), which is 46 and 68 cases, respectively. The imbalance in variables of work experience, race, marital status, and parole status increases the risk of recidivists being re-arrested within 52 weeks. The following table presents a numerical description of the data used.

Table 2. Numerical Description of Survival Time, Age, and Prior

Statistics	Survival time	Age	Prior
Min	1	17	0
Max	52	44	18
Mean	45.85	24.6	2.98
SD	12.66	6.11	2.9
Median	52	23	2
Q1	50	20	1
Q3	52	27	4
Skewness	-1.98	1.38	2.07
Kurtosis	2.62	1.32	5.21

Table 2 shows that the average time for a recidivist to return to prison after release was 45.85 weeks, with the shortest time being 1 week and the longest time being 52 weeks, and a standard deviation of 12.66 weeks. The average age of the observed recidivists was 24.6 years old, with the youngest subject being 17 years old and the oldest subject being 44 years old. Additionally, the average number of prior theft arrests was 3 times, with the fewest being none and 18 being the highest number. Although **Table 2** indicates that the distribution of survival time is left-skewed, the survival time data will be modeled using common distributions in AFT analysis, such as Weibull, exponential, log-logistic, and log-normal distributions. However, it is important to note that the data observation is limited to one year following the release of prisoners. In reality, it is possible that recidivists would be re-arrested after 52 weeks, but such cases fall outside the scope of the recorded data. This paper aims to explore the distributional patterns of recidivists' likelihood of being re-arrested using commonly applied distribution. The best AFT model demonstrates the distribution of the survival time data.

Table 3. AIC and BIC of The Survival Time Distribution

Model	Weibull	Exponential	Log-normal	Log-logistics
AIC	3406.829	4171.203	3885.867	3662.439
BIC	3414.966	4175.271	3849.004	3670.576

Figure 2 displays the distribution of survival data. Based on **Figure 2**, it is found that the Weibull distribution is the best fit for modeling the data compared to the other three distributions. This is further supported by **Table 3**, which shows that the smallest AIC and BIC are given by the Weibull distribution.

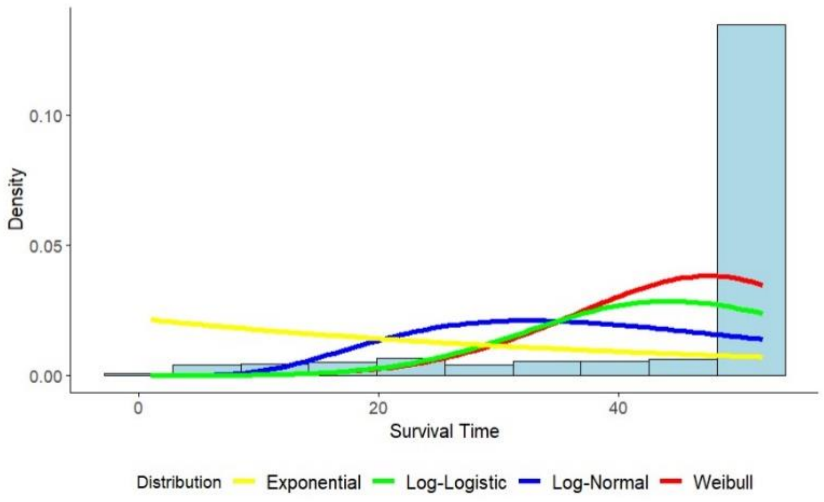


Figure 2. Distribution of Survival Time
Source: *Analysis output by RStudio*

3.2 Regression Modeling

a. AFT Modeling

In this study, the AFT model was selected by comparing the AIC and BIC values of candidate models from the Weibull, exponential, log-normal, and log-logistic distributions. The following presents estimates of AFT model parameters for each distribution.

Table 4. Multiple Distribution AFT Model Parameter Estimates

		Weibull	Exponential	Log-normal	Log logistics
Variable	X_1	0.2722	0.3663	0.3428	0.2889
	X_2	0.0407	0.0556	0.0272	0.0364
	X_3	0.2248	0.3049	023632	0.2791
	X_4	0.1066	0.1467	0.2681	0.1784
	X_5	-0.3113	-0.4270	-0.4604	-0.3473
	X_6	-0.0588	-0.0826	-0.0559	-0.0508
	X_7	-0.0658	-0.857	-0.0655	-0.0692
Metric	AIC	1377.833	1388.732	1384.469	1377.877
	BIC	1414.449	1421.279	1421.085	1414.493

Given that the Weibull AFT model shows the lowest AIC and BIC values compared to other distributions, the model selected for this study is the Weibull AFT model, incorporating significant predictor variables. This also confirms that the best AFT model aligns with the distribution of survival time. **Table 5** presents a significance test of predictor variables.

Table 5. Weibull AFT Model Parameter Significance Testing

Variable	Coefficient	$p-value$	Results
Intercept	4.1354	2×10^{-16}	Reject H_0
X_1	0.2722	0.04852	Reject H_0
X_2	0.0407	0.01096	Reject H_0
X_3	0.2248	0.30721	Failed to reject H_0
X_4	0.1066	0.48196	Failed to reject H_0
X_5	-0.3113	0.25473	Failed to reject H_0
X_6	-0.0588	0.67355	Failed to reject H_0
X_7	-0.0658	0.001678	Reject H_0
Likelihood ratio		2.2×10^{-5}	Reject H_0
$\lambda = 0.712$		$\gamma = \frac{1}{\lambda} = 1.404$	

Based on **Table 5**, p − value of the partial likelihood ratio test is smaller than 0.05, therefore it is concluded that there is at least one predictor variable significant to the model fit, namely the variable X_1, X_2, and X_7 which has a value smaller than 0.05.

The next step is to build survival function from Weibull's AFT model as follows:

$$S(t|\boldsymbol{X}) = exp(-exp(4.1354 + 0.2722X_1 + 0.0407X_2 - 0.0658X_7)t)^{1.404}) \quad (19)$$

The hazard function of the Weibull AFT model is expressed in **Equation (20)** below.

$$h(t|\boldsymbol{X}) = 250.4108t^{4.0462}ex\,p(0.2722X_1 + 0.0407X_2 - 0.0658X_7) \quad (20)$$

b. Cox Modeling

The first step in Cox regression analysis is to test whether the hazard ratios of two different groups are proportional to time or in other words will be tested whether the predictor variable $X_1, X_2, \dots, X_7$ has satisfied the PH assumption. The results of the assumption test are given by **Table 6** as follows.

Table 6. PH Assumption Testing

Variable	Base Form	Elimination X_2 and X_4	Stratification X_2 and X_4	Decision
	p − value			
X_1	0.803	0.78	0.82	Failed to reject H_0
X_2	0.015			Reject H_0
X_3	0.150	0.15	0.17	Failed to reject H_0
X_4	0.040			Reject H_0
X_5	0.315	0.33	0.43	Failed to reject H_0
X_6	0.892	0.86	0.97	Failed to reject H_0
X_7	0.471	0.50	0.78	Failed to reject H_0
Global		0.66	0.77	Failed to reject H_0

Based on **Table 6,** variables X_2 and X_4 violate the PH assumption at a significant level $\alpha = 0.05$. To address this, the violating variables were incorporated into the strata, leading to the application of the stratified Cox model. Prior to this, variable X_2 was categorized based on the average age:

$$age = \begin{cases} 1, & for\ age \leq 25 \\ 0, & for\ age > 25 \end{cases} \quad (21)$$

Table 6 also presents the results of the PH assumption test after variables X_2 and X_4 were eliminated and stratified, showing that all remaining variables satisfy the PH assumption. Parameter estimates for both the Cox PH and stratified Cox models are presented in **Table 7**.

Table 7. Parameter Estimates of Cox PH and Stratified Cox

		Cox PH	*Stratified* Cox
Variables	X_1	-0.42702	-0.40580
	X_3	-0.29396	-0.23181
	X_5	0.69371	0.345408
	X_6	0.01404	0.04998
	X_7	0.10414	0.08795
Metric	AIC	1338.619	1051.619
	BIC	1352.104	1065.300

The best Cox regression model is determined based on the smallest AIC and BIC as presented in **Table 7**. Since the AIC and BIC values of the stratified Cox model are both smaller than those of the Cox PH model, it is shown that the stratified Cox model fits the data set better than the Cox PH model. Consequently, a stratified Cox regression model is constructed by considering the independent variable significant to the model fit. Parameter significance testing results are presented in **Table 8**.

Table 8. Parameter Significance Testing of Stratified Cox Model

Variable	Coefficient	$p-value$	Results
X_1	-0.40580	0.00337	Reject H_0
X_3	-0.23181	0.45394	Failed to reject H_0
X_5	0.34541	0.23312	Failed to reject H_0
X_6	0.04998	0.79737	Failed to reject H_0
X_7	0.08795	0.00245	Reject H_0
Likelihood ratio		0.01	Reject H_0

The partial p-values from the likelihood ratio test in **Table 8**, which are both smaller than 0.05, indicate that at least one predictor is significant to the response variable. Specifically, X_1 and X_7 have p-values of 0.003373 and 0.00245, respectively. Thus, a stratified Cox regression model is obtained as shown in **Equation (22)** below.

$$h_g(t,X) = h_{0g}(t)exp\,(-0.40580X_1 + 0.08795X_7) \tag{22}$$

where $h_{0g}(t)$ represents the baseline hazard function, which varies for each possible combination of strata conditions.

c. Comparison and Interpretation

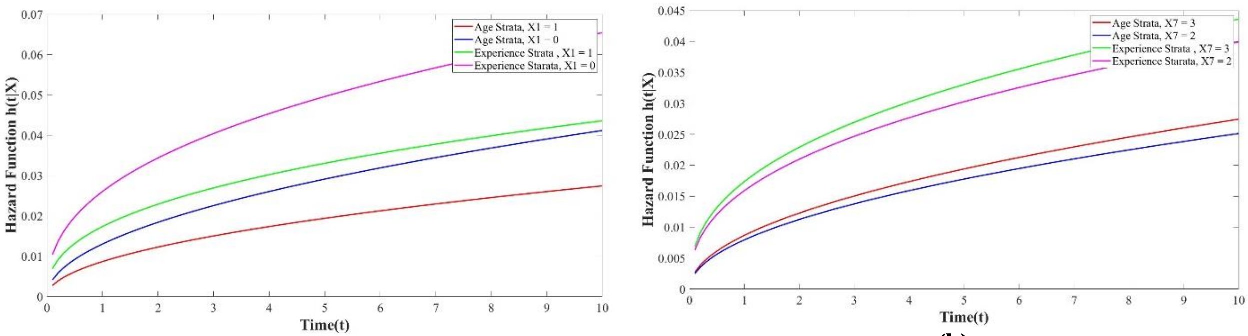
Table 4 indicates that the Weibull AFT model fits the data better than the other AFT models with AIC and BIC of 1377.833 and 1414.449, respectively. However, based on **Table 7**, the stratified Cox model was identified as the best-performing Cox model with AIC and BIC of 1036.831 and 1065.300, respectively. Additionally, from **Figure 2**, the result of the fitting distribution indicates that the Weibull distribution did not adequately capture the characteristics of the observed survival time data. This observation underscores a limitation of the Weibull AFT model, as its assumption of a specific form for the survival time distribution may not align well with the predictor data distribution.

This analysis led to the conclusion that stratified Cox model outperformed the Weibull AFT model. The Cox model's advantage lies in its capacity to closely estimate the true distribution of survival time without needing to assume a specific distribution form, along with its robustness in managing complex data variations and censored data.

Table 9. Hazard Ratio of Stratified Cox Model

Variable	Hazard ratio
Financial Aid (X_1)	0.66644
Prior (X_7))	1.09193

As shown in **Table 9**, recidivists who received financial aid after release had a 0.66644 times lower risk of re-arrest compared to those who did not receive financial aid. Each additional one-time prior theft arrest increases the risk of re-arrest by 1.09193 times. This is further supported by **Figure 3**, which indicates that receiving financial aid $X_1 = 1$ is associated with a reduced risk of re-arrest within the same time frame, while a history of prior theft arrests is linked to an increased risk of re-arrest.



(a) **(b)**

Figure 3. Simulated Hazard Function of Stratified Cox Model from Variable: (a) Financial Aid (b) Prior

Source: *Analysis output by MATLAB*

The result indicates contrasting interpretations of predictor variables between the AFT and Cox models. The AFT model, assuming survival time follows a Weibull distribution, does not fully align with the data as shown in **Figure 2**. Additionally, as illustrated in **Figure 4**, the hazard function curve for recidivists who received financial aid (blue line) is higher than that for recidivists who did not receive financial aid (red line).

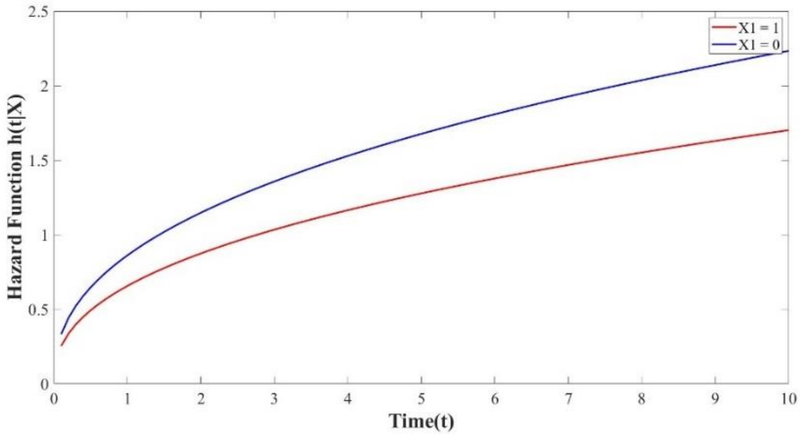


Figure 4. Simulated Hazard Function of AFT model
Source*: Analysis output by MATLAB*

It indicates that recidivists who received financial aid are more likely to return to custody sooner. This conclusion differs from the results obtained in the Cox model. The difference may occur because the Weibull distribution assumption in the AFT model does not capture all the characteristics of the empirical data.

4. CONCLUSIONS

This paper aims to explore the distributional patterns of the force of recidivists being re-arrested by conducting AFT and Cox models. Based on the results and discussion, the stratified Cox model is identified as the most effective model for estimating the survival time of recidivists in society using the 1971 LIFE project data. This conclusion is based on the comparison between the AFT and Cox models, where the stratified Cox model yielded a smaller AIC and BIC value compared to the AFT model. The result obtained highlights the robust nature of the Cox model in dealing with survival data, particularly if the parametric distribution of the survival time is unknown.

The stratified Cox model further identified significant variables affecting recidivists' survival times, specifically Financial Aid (X_1) and Prior (X_7). Recidivists who received financial aid had 0.66644 times lower the risk of being re-arrested compared to those without financial aid. Moreover, each additional one-time of prior theft arrest increased the risk of re-arrest by 1.09193 times. These findings highlight the stratified Cox model's ability to effectively capture the nuanced effects of covariates on recidivism risk, making it the preferred choice for this analysis over the AFT model.

Further development in this paper could involve exploring alternative distributions within the AFT model to obtain a better fit with the data. Analysis using spline-based models or mixture models can provide flexibility in modeling complex survival time distributions. Therefore, further research on the AFT model using these approaches could enhance the understanding of predictor variable behavior in the context of complex survival data and significant censoring.

REFERENCES

[1] L. Sugiharti, R. Purwono, M. A. Esquivias and H. Rohmawati, "The Nexus between Crime Rates, Poverty, and Income Ineqality: A Case Study of Indonesia," *MDPI,* vol. 11, no. 62, pp. 1-15, 2023.

[2] D. G. Kleinbaum and M. Klein, Survival Analysis: A Self Learning Text, New York: Springer, 2012.

[3] M. J. V. Rekstad, A Comparison of Methods for Modelling Survival for Cancer Patients, Norway: NMBU, 2023.

[4] M. Parsa and I. Van Keilegom, "Accelerated Failure Time vs Cox Proportional Hazards Mixture Cure Models: David vs Goliath?," *Springer Nature,* vol. 64, pp. 835-855, 2023.

[5] J. Qi, Comparison of Proportional Hazards and Accelerated Failure Time Models, Saskatchewan: University of Saskatchewan, 2009.

[6] K. Sarkar, R. Chowdhury and A. Dasgupta, "Analysis of Survival Data: Challenges and Algorithm-Based Model Selection," *Journal of Clinical and Diagnositc Research,* pp. 14-20, 2017.

[7] D. W. Hosmer, S. Lemeshow and S. May, Applied Survival Analysis, New Jersey: John Wiley & Sons, Inc, 2008.

[8] J. D. Kalbfleisch and R. L. Prentice, The Statistical Analysis of Failure Time Data, New Jersey: John Wiley & Sons, Inc., 2002.

[9] N. Mishra and S. Misra, "Comparison of Cox Proportional Hazard Model and Accelerated Failure Time (AFT) Models: an Application to Under-five Mortality in Uttar Pradesh," *Indian Journal of Applied Research,* vol. 12, no. 12, pp. 83-87, 2022.

[10] K. Mardhiah, N. Wan-Arfah, N. N. Naing, M. R. Abu Hassan and H.-K. Chan, "Comparison of Cox proportional hazards model, Cox proportional hazards with time-varying coefficients model, and lognormal failure time model: Application in time to event analysis of melioidosis patients," *Asian Pacific Journal of Tropical Medicine,* vol. 15, no. 3, pp. 128-134, 2022.

[11] R. Mutambayi, A. Azeez, J. Ndege and Y. Qin, "Statistical Analysis of Risk factors of Malaria re-infection among Outpatients in DR, Congo: A Comparison Approach of AFT and Cox PH Models," *Journal of Communicable Diseases,* vol. 51, no. 1, pp. 22-28, 2019.

[12] A. Khamis, C. A. Che Hamat and M. A. A. Abdullah, "Modeling Students' Performance using Cox and Parametric Accelerated Failure Time Models," *Scientific Research Journal (SCIRJ),* vol. VIII, no. VII, pp. 44-49, 2020.

[13] D. Melak, S. Wedajo and R. Dewau, "Time to Viral Re-suppression and Its Predictors among Adults on Second-Line Antiretroviral Therapy in South Wollo Zone Public Hospitals: Stratified Cox Model," *HIV/AIDS - Research and Palliative Care,* vol. 15, pp. 411-421, 2023.

[14] P. H. Rossi, R. A. Berk and K. J. Lenhan, Money, Work, and Crime: Work Experimental Evidence, New York: Academic Press, 1980.

[15] R. De la Cruz, P. Oslando, M. A. V. Valle and G. A. Ruz, "Modeling Recidivism through Bayesian Regression Models and Deep Neural Networks," *MDPI,* vol. 9, no. 6, p. 639, 2021.

[16] L. Waldeck, "How to Predict Recidivism in Young Offenders: Comparing Logistic Regression, Random Forest, and Boosted Classification Trees and Examining Possible Risk Factors," Maastricht University, Maastricht, 2020.

[17] I. Lijarcio, F. J. Llamazares, E. Valle, L. Montoro and S. A. Useche, "Assessing Risk Perception over Recidivist Traffic Offenders from a Multi-group Approach: How Gendered Could It Be?," *The European Journal of Psychology Applied to Legal Context,* vol. 14, no. 1, pp. 33-41, 2022.

[18] A. Singh and S. Mohapatra, "Automatic Criminal Recidivism Risk Estimation in Recidivist Using Classification and Ensemble Techniques," *Soft Computing for Problem Solving,* pp. 279-289, 2023.

[19] J. Fox, An R and S-Plus Companion to Applied Regression, Thousand Oaks: SAGE Publications, Inc, 2011.

[20] Y. Anastasya, N. Satyahadewi and H. Perdana, "Model Stratified Cox untuk Mengatasi Nonproportional Hazard," *Buletin Ilmiah Math. Stat. dan Terapannya,* pp. 311-320, 2016.

[21] D. R. Cox and D. Oakes, Analysis of Survival Data, London: CRC Press, 1998.

[22] T. M. Therneau and P. M. Grambsch, Modeling Survival Data: Extending The Cox Model, New York: Springer, 2000.

[23] R. Saikia and M. P. Barman, "A Review on Accelerated Failure Time Models," *International Journal of Statistics and Systems,* vol. 12, no. 2, pp. 311-322, 2017.

[24] M. Mills, Introducing Survival and Event History Analysis, London: SAGE Publications Ltd, 2011.

[25] P. Allison, Survival Analysis Using SAS: A Practical Guide, Cary: SAS Institute, 2010.