Comparison of Different Underlying Distributions in The Accelerated Failure Time (AFT) Model on Mortality of Covid-19 Patients

Astri Meiza*, Asep Solih Awaluddin1, Nurapni Oktapia Hidayah1, and Adnan Taha2
1Department of Mathematics, UIN Sunan Gunung Djati Bandung, Indonesia
2Oxford Business College, Oxford, United Kingdom
Email: asti.meiza@uinsgd.ac.id

Abstract

In 2022, the COVID-19 virus is still making headlines in various mass media because it is a virus that is very dangerous to health. The world health organization, WHO, explained that the virus caused a global pandemic that infected the whole world. The condition of a pandemic has not yet turned into an endemic. Based on the total confirmed COVID-19 positive cases, Indonesia ranks 18th in the world out of 222 infected countries. To determine the influence factors on COVID-19 cases, survival analysis is one of the techniques that could be applied. One of the most commonly used models in survival analysis is Accelerated Failure Time (AFT) model. In the AFT model, it is required to check assumptions regarding the feasibility of the distribution form. In this study, the distributions used are Weibull, Exponential, Log-normal, and Log-logistics distributions. We compare each distribution to get the best model to analyze death cases due to COVID-19. Comparisons are made by comparing the AIC values of each distribution. The best model is selected based on the smallest AIC value. The AFT model with a log-normal distribution is selected as the best model with an AIC value of 142.763. The AIC value for this log-normal distribution is the smallest compared to the AIC value for other distributions.

Keywords: accelerated failure time model; COVID-19; mortality analysis; survival analysis.

Abstrak


Kata Kunci: analisis mortalitas; analisis survival; COVID-19; model accelerated failure time.

2020MSC: 62P10
1. INTRODUCTION

On December 31, 2019, the World Health Organization (WHO) or the World Health Organization China Country Office reported pneumonia in Wuhan City, Hubei Province, China, whose cause was still unknown. This disease is Coronavirus Disease 2019 (COVID-19). COVID-19 comes from the beta-coronavirus genus with the same genus as the causative agents of Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) [1]. The initial symptoms of COVID-19 are non-specific, starting with the appearance of fever and cough, which can then resolve spontaneously or develop into shortness of breath, dyspnea, and pneumonia which can cause Acute Respiratory Distress Syndrome (ARDS). In addition, it can cause kidney damage, coagulation or blood clotting dysfunction, and death [2] [3].

China stated that COVID-19 is a new type of pneumonia case of coronavirus on January 7, 2020 [4]. In Indonesia, confirmed cases of COVID-19 are shown in Figure 1 [5]. Based on this figure, it can be seen that in June, the confirmed positive cases of COVID-19 increased after the January 2021 period. On June 5, 2021, confirmed positive COVID-19 reached 1,843,612 people with 51,296 deaths. Of the total confirmed positive for COVID-19, Indonesia ranks 18th in the world out of 222 infected countries. Several studies have been developed to determine the factors that influence the deaths of COVID-19 patients, including [6] [7]. In this article, several significant factors that affect the mortality of COVID-19 patients will be determined using the survival analysis with the Accelerated Failure Time (AFT) model.

![Figure 1. Daily graph of confirmed COVID-19 in 2021](image)

The AFT model is one of the parametric models based on the data distribution that can predict the time of an event in the observation data. The distributions used are Exponential, Weibull, Log-normal, and Log-logistics distributions [8]. The AFT model can describe the relationship between the probability of survival and the set of covariates, where covariates can affect the survival time of an acceleration factor called the accelerated factor.

Recently, the AFT model has been widely applied to several problems. Jal-Uiman et al. [9] identified covariates associated with the recovery time of Covid-19 patients using the AFT model. Khan and Howlader [10] implement the AFT to evaluate the stability in the selection process of some models with data sets containing many covariates. Granville and Fan [11] study the Buckley-James
estimator of accelerated failure time models with auxiliary covariates. They use a local polynomial approximation method to accommodate the covariates into the Buckley-James estimating equations. Shi-aq Qi et al. [12] describe the predictions of survival models to help professionals make decisions about patient conditions and resource allocation. They used a gradient-boosting Cox machine.

The AFT model is also widely used in medical research. Rachmaniyah et. al. [13] determine the chance of a faster death in people with diabetes mellitus. Crowther et al. [14] use the AFT model to accommodate time-dependent acceleration factors in breast cancer. Sulantari and Hariadi [15] model the time recovery of COVID-19 patients using the Kaplan-Meier method and the Log-Rank test on COVID-19 patients. The application of the AFT model in the insurance sector [16]. In this paper, we determine the factors that significantly affect the mortality of COVID-19 patients, determine a suitable distribution for COVID-19 cases, and determine the AFT model of the best distribution in cases of COVID-19 [17].

2. METHODS

Survival analysis is a statistical method used to analyze data about the time until an event occurs [18]. Survival time is the primary variable in survival analysis. The survival time can be interpreted as a variable that can measure the observation time from the initial point of observation to the end of the observation.

Definition 1. Suppose $T$ is a non-negative random variable of the individual in survival time which states the time until a failure occurs. Survival function $S(t)$, define as the probability that the random variable $T$ beyond time $t$, $(t > 0)$ (experiencing events after time $t$), which can be written as

$$S(t) = P(T > t).$$  \hspace{1cm} (1)

Definition 2. Hazard function $h(t)$ is commonly referred to as the force of mortality or the conditional failure rate. This hazard function is the probability of failure at a time interval [20], which can be expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$  \hspace{1cm} (2)

2.1. Accelerated Failure Time (AFT) Model

The AFT model is a parametric model based on the data distribution that can predict the time of an event in the observation data. The Accelerated Failure Time model is an alternative to the Cox Proportional Hazards (CPH) model [7]. The AFT method can describe the relationship between the probability of survival and the set of covariates, which can affect the survival time of an accelerated factor.

Hazard function of accelerated failure time with covariate $x_1, x_2, ..., x_p$ can be written as

$$h_1(t) = \left(\frac{1}{\eta(x)}\right) h_0 \left(\frac{t}{\eta(x)}\right),$$  \hspace{1cm} (3)

where $h_0(t)$ is the baseline hazard function, $\eta(x)$ is the accelerated factor, and $x$ is an independent variable. Research using the AFT model has been carried out by several researchers, including Khanal...
et al. 2012 who examined the use of AFT to see the resilience of Acute Liver Failure Patients in India [21].

2.2. Distributions for the AFT Model

As stated in the previous section, the AFT is a parametric model. In this section, we will briefly discuss several distributions expressed in the AFT model, namely the Weibull distribution, Exponential Distribution, Log-normal distribution, and Log-logistic distribution. One distribution that best fits the data will be selected from these four distributions to be expressed in the AFT model.

The following is the hazard function for several distributions to be compared.

a) Weibull distribution

Weibull distribution hazard function can be written as

\[ h_0(t) = \lambda t^{\gamma - 1}, \text{ for } \lambda > 0 \text{ and } \gamma > 0. \]  \hspace{1cm} (4)

b) Exponential distribution

The exponential distribution hazard function can be written as

\[ h_0(t) = \lambda, \]  \hspace{1cm} (5)

with parameters \( t \geq 0 \) and \( \lambda \geq 0 \).

c) Log-normal distribution

The log-normal distribution hazard function is as follows

\[ h_0(t) = \frac{1}{\sqrt{2\pi\sigma t}} \exp \left[ -\frac{1}{2} \left( \log t - \mu \right)^2 \right]. \]  \hspace{1cm} (6)

d) Log-logistic distribution

The log-logistics distribution hazard function is as follows

\[ h_0(t) = \frac{e^{\theta k t^{-k - 1}}}{1 + e^{\theta t^k}}. \]  \hspace{1cm} (7)

The function \( h_0(t) \) is a baseline hazard function that describes the risk for individuals with \( X_i = 0 \) [22]. This function is left unspecified [23]. Unknown means it is non-parametric [15], although the AFT is a parametric model.

2.3. The Parameter Significance Test

a) The Simultaneous Test

The Likelihood Ratio Test or \( G \) test is a test to see whether the independent variables as a whole have an effect or not on the model. The statistical formula for the \( G \) test used is [20]

\[ G = -2(\log L_0 - \log L_p), \]  \hspace{1cm} (8)

where \( L_0 \) is the likelihood function of the model without independent variables, and \( L_p \) is the likelihood function of the model consisting of \( p \) independent variables. Test statistics \( G \) follows a chi-square distribution with degrees of freedom \( p \). If \( G > \chi^2_{(\alpha,p)} \), then \( H_0 \) must be rejected, meaning that at least one independent variable can affect the model and vice versa.
b) The Partial Test

The partial test used is the Wald test statistic. This Wald test statistic follows a chi-square distribution with p degrees of freedom [21]. The hypothesis used is as follows:

\[ H_0: \beta_j = 0 \text{ (independent variable is not significant)} \]
\[ H_1: \beta_j \neq 0 \text{ (independent variable is significant)}. \]

The statistical formula for Wald's test is

\[ W = \left( \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)^2, \quad (9) \]

with \( \hat{\beta}_j \) are coefficients that accompany the variables, and \( SE(\hat{\beta}_j) \) is the error standard. If the significance value, \( W > \chi^2_{p} \), \( H_0 \) is rejected, the independent variable is significant to the model, and vice versa.

2.4. Model Fit Test

Akaike's Information Criterion (AIC) method is one method that can be used to determine the best model found by Akaike. According to Klein and Moeschberger [18], the value of the AIC can be written as follows

\[ AIC = -2l + 2p, \quad (10) \]

where \( l \) is log-likelihood and \( p \) is the total degrees of freedom used in the model. In determining the best model with this method, we use the smallest value of the AIC.

3. RESULTS AND DISCUSSION

In this research, we use primary data from the Bandung City Regional General Hospital in the period of January to June 2021. The sample was 100 patients, where 22 patients were uncensored, and 78 were censored. The data analysis method is divided into parameter significance and model feasibility tests. The parameter significance test is further divided into two jobs: simultaneous test and partial test.

The results of this study are divided into several parts, starting with descriptive results, then comparing the AFT model for several distributions. Finally, the model feasibility test results obtained the best distribution for the AFT model.

3.1. Descriptive Analysis

There are four variables in this study, i.e., age, gender, comorbid, and symptom. The descriptive statistics are presented in Table 1.
Table 1. Descriptive analysis of all variables

<table>
<thead>
<tr>
<th>Age ($X_1$)</th>
<th>Number of Patients</th>
<th>Mortalitas %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &lt; 25.45</td>
<td>19</td>
<td>5.26%</td>
</tr>
<tr>
<td>25.45 ≤ Age &lt; 67.27</td>
<td>70</td>
<td>24.28%</td>
</tr>
<tr>
<td>67.27 ≤ Age</td>
<td>11</td>
<td>36.36%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender ($X_2$)</th>
<th>Number of Patients</th>
<th>Mortalitas %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>49</td>
<td>16.33%</td>
</tr>
<tr>
<td>Male</td>
<td>51</td>
<td>27.45%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comorbid ($X_2$)</th>
<th>Number of Patients</th>
<th>Mortalitas %</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>77</td>
<td>6.49%</td>
</tr>
<tr>
<td>Comorbid</td>
<td>23</td>
<td>73.91%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Symptom ($X_4$)</th>
<th>Number of Patients</th>
<th>Mortalitas %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild</td>
<td>35</td>
<td>0%</td>
</tr>
<tr>
<td>Moderate</td>
<td>37</td>
<td>8.11%</td>
</tr>
<tr>
<td>Severe</td>
<td>28</td>
<td>67.86%</td>
</tr>
</tbody>
</table>

3.2. Estimation Parameter of The AFT Model

Before determining the AFT model, it is necessary to check the distribution assumptions first. Weibull, Exponential, Log-normal, and Log-logistics distributions are the distributions used in checking this assumption. Below is a provisional model of each distribution as follows:

a) Weibull Distribution

Table 2 shows the parameter estimation of the AFT Weibull model.

Table 2. Parameter estimation of the AFT Weibull Model

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.220</td>
<td>0.215</td>
<td>-1.03</td>
<td>0.30491</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.405</td>
<td>0.319</td>
<td>-1.27</td>
<td>0.20377</td>
</tr>
<tr>
<td>Comorbid</td>
<td>-1.354</td>
<td>0.363</td>
<td>-3.74</td>
<td>0.00019</td>
</tr>
<tr>
<td>Symptom</td>
<td>-0.796</td>
<td>0.309</td>
<td>-2.58</td>
<td>0.00992</td>
</tr>
</tbody>
</table>

Based on Table 2, the provisional model of the AFT Weibull model is obtained as follows:

$$h(t|X) = \left(\frac{\exp(-0.220x_1-0.405x_2-1.354x_3-0.796x_4)}{\sigma}\right)\exp\left(-\frac{\mu}{\sigma}\right)\exp\left(-\frac{t}{\exp(-0.220x_1-0.405x_2-1.354x_3-0.796x_4)}\right)^{\frac{1}{\sigma}}.$$ (11)

b) Exponential Distribution

Table 3 shows the parameter estimation of the AFT Exponential model.

Table 3. Parameter estimation of the AFT Exponential model

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.482</td>
<td>0.397</td>
<td>-1.21</td>
<td>0.22463</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.835</td>
<td>0.564</td>
<td>-1.48</td>
<td>0.13896</td>
</tr>
<tr>
<td>Comorbid</td>
<td>-2.115</td>
<td>0.614</td>
<td>-3.45</td>
<td>0.00057</td>
</tr>
<tr>
<td>Symptom</td>
<td>-1.412</td>
<td>0.567</td>
<td>-2.49</td>
<td>0.01277</td>
</tr>
</tbody>
</table>
Based on Table 3, the provisional model of the AFT Exponential model is obtained as follows:

$$h(t|X) = \exp \left( -\frac{\mu}{\sigma} - (-0.482x_1 - 0.835x_2 - 2.115x_3 - 1.412x_4) \right).$$  \hspace{1cm} (12)

Remember, in an Exponential distribution, the value of the basic hazard function is the same as the rate.

c) Log-normal Distribution

Table 4 shows the parameter estimation of the AFT Log-normal model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Std.Error</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.411</td>
<td>0.274</td>
<td>-2.11</td>
<td>0.1331</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.695</td>
<td>0.330</td>
<td>-4.52</td>
<td>0.0048</td>
</tr>
<tr>
<td>Comorbid</td>
<td>-1.421</td>
<td>0.314</td>
<td>-2.88</td>
<td>6.0e-06</td>
</tr>
<tr>
<td>Symptom</td>
<td>-0.828</td>
<td>0.287</td>
<td>-2.06</td>
<td>0.0039</td>
</tr>
</tbody>
</table>

Based on Table 4, the temporary model of the AFT Log-normal model is obtained as follows:

$$h(t|X) = \left( \frac{1}{\exp(-0.411x_1 - 0.695x_2 - 1.421x_3 - 0.828x_4)} \right) \times \left[ \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{1}{2} \left( \frac{\log \left( \exp(-0.411x_1 - 0.695x_2 - 1.421x_3 - 0.828x_4) - \mu \right)}{\sigma} \right)^2 \right) \right] \left[ 1 - \Phi \left( \frac{\log \left( \exp(-0.411x_1 - 0.695x_2 - 1.421x_3 - 0.828x_4) - \mu \right)}{\sigma} \right) \right].$$  \hspace{1cm} (13)

d) Log-logistic Distribution

Table 5 shows the parameter estimation of the AFT Log-logistics model. Based on this table, the provisional model of the AFT Log-Logistics model is obtained as follows:

$$h_i(t) = \left( \frac{1}{\sigma t} \left[ 1 + \exp \left( -\frac{t - \mu + 0.435x_1 - 0.605x_2 - 1.486x_3 - 0.737x_4}{\sigma} \right) \right] \right)^{-1}.$$  \hspace{1cm} (14)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Std.Error</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.435</td>
<td>0.263</td>
<td>-1.65</td>
<td>0.098</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.605</td>
<td>0.310</td>
<td>-1.95</td>
<td>0.051</td>
</tr>
<tr>
<td>Comorbid</td>
<td>-1.486</td>
<td>0.314</td>
<td>-4.73</td>
<td>2.3e-06</td>
</tr>
<tr>
<td>Symptom</td>
<td>-0.737</td>
<td>0.273</td>
<td>-2.70</td>
<td>0.007</td>
</tr>
</tbody>
</table>
3.3. Model Fit Test

Table 6 shows that the AFT Log-normal model has the smallest AIC value compared to the other three models, which is 142.763. Hence, the best model for the length of stay for the death of COVID-19 patients at the Bandung City Regional General Hospital is the AFT Log-normal model.

Table 6. The AIC values

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull AFT</td>
<td>143.9453</td>
</tr>
<tr>
<td>Exponential AFT</td>
<td>151.7752</td>
</tr>
<tr>
<td>Log-normal AFT</td>
<td>142.7630</td>
</tr>
<tr>
<td>Log-logistic AFT</td>
<td>142.8104</td>
</tr>
</tbody>
</table>

Table 7 shows the parameter significance of the AFT Log-normal model. This table shows that the p-value for the AFT with the log-normal distribution is 5.6e-15, which is less than the 0.05 significance level. This value indicates that the log-normal distribution significantly affects the AFT model. The significant variables are gender, comorbidity, and symptoms. In contrast, the age variable is insignificant.

Table 7. Log-normal distribution

<table>
<thead>
<tr>
<th>p - value</th>
<th>Wald</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.1331</td>
<td>2.25</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0348</td>
<td>4.435</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>6.0e-06</td>
<td>20.479</td>
</tr>
<tr>
<td>Symptom</td>
<td>0.0039</td>
<td>8.323</td>
</tr>
</tbody>
</table>

p - value = 5.6e-15

From the parameter estimation, the final model of the AFT Log-normal model is

$$h(t|X) = \left( \frac{1}{\text{exp}(-0.695x_2-1.421x_3-0.828x_4)} \right) \left[ \frac{1}{\text{invnormal}(t)} \exp \left\{ -\frac{1}{2} \left( \frac{t - \log(\text{exp}(-0.695x_2-1.421x_3-0.828x_4))^\mu}{\sigma} \right)^2 \right\} \right] \left( 1 - \Phi \left( \frac{t - \log(\text{exp}(-0.695x_2-1.421x_3-0.828x_4))^\mu}{\sigma} \right) \right)$$

where $x_2$ is gender, $x_3$ is comorbidity, $x_4$ is a symptom, $t$ is age, $\mu$ is mean, and $\sigma$ is the standard deviation.

4. CONCLUSIONS

This paper compares the AFT model using the Weibull, Exponential, log-normal, and log-logistic distributions. Based on the AIC value, the AFT model with a log-normal distribution is better than other distributions. There are three significant variables, i.e., gender, comorbidity, and symptoms. In
contrast, the age variable is insignificant. These significant variables mean that gender, the presence of comorbidities, and the onset of severe symptoms can affect death in COVID-19 patients.

REFERENCES


