Modeling the survivorship and the hazard functions of lognormal distribution used to predict risk factors for stroke

Sudarno¹, A Prahutama¹

¹Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia
E-mail: dsghani@gmail.com

Abstract. Model of data distribution used for predict by the general of the events. Survival times are data that measure by the time to a certain period. The survivorship function is used to predict probability that an individual survives, while the hazard function is used to predict the probability of individual sick during a time. The data are assumed lognormal distribution. This matter took a random sample of participants from a study of the incidence and risk factors for stroke. The covariates are body mass index, smoking, alcoholic, diabetes mellitus, and albuminuria. This paper discusses comparing between models of the survivorship and the hazard functions by without covariate and with covariate. Determining the survivorship and the hazard functions with covariate use Accelerated Failure Time (AFT) model. The result of this research is body mass index, smoking, and diabetes mellitus cause stroke. Body mass index is positive effect, while smoking and diabetes mellitus are negative effect. Alcoholic and albuminuria are not significant. Measure of prediction accuracy uses the Mean Absolute Percentage Error (MAPE). The MAPE is 0.8729%, this value means that 99.1271% of models are good for prediction.

1. Introduction

Let $T$ denote the survival time. By [1] that distribution of random variable can use the goodness of fit distribution. The distribution of $T$ can be characterized by the survivorship or survival and hazard functions. The survivorship function $S(t)$ is defined as the probability that an individual survives longer than $t$. The hazard function $h(t)$ of survival time $T$ gives the conditional failure rate. It is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval [2, 3, 4]. The hazard function can also be defined in term of the cumulative distribution, $S(t)$ and the probability density function $f(t)$ by ratio, that is $f(t)/S(t)$ [6].

Survival analysis is the analysis of statistical data in which the outcome variable of interest that time until an event occurs. In statistical literature, it is observed that a good number of models have been developed for analyzing survival data or lifetime data. By [2] that Accelerated Failure Time (AFT) model is a linear regression of covariate. It can be considered as a good alternative of Cox Proportional Hazard model in analyzing survival data. A review on accelerated failure time model discusses the historical developments, technical developments and past research on AFT models [7].

Bivariate failure time data is widely used in survival analysis, for example, in twins study. This article presents a class of χ²-type tests for independence between pairs of failure times after adjusting for covariates. A bivariate accelerated failure time model is proposed for the joint distribution of bivariate failure times while leaving the dependence structures for related failure times completely unspecified. Theoretical properties of the proposed tests are derived and variance estimates of the test
statistics are obtained using a resampling technique. Simulation studies show that the proposed tests are appropriate for practical use. Two examples including the study of infection in catheters for patients on dialysis and the diabetic retinopathy study are given to illustrate the methodology [5].

Truncated data are commonly seen in studies of biomedicine, epidemiology, astronomy, and econometrics. Existing regression methods for analyzing left-truncated and right-censored data have been developed under the assumption that the lifetime variable of interest is independent of both truncation and censoring variables. In that article, proposed a semi-parametric accelerated failure time model that incorporates both covariates and truncation variable as the independent variable. The proposed model utilizes the truncation information in statistical modeling and hence allows for dependent truncation. For estimation, developed a set of estimating equations constructed from the log-rank and quasi-independence test statistics. The resulting estimators are consistent and asymptotically normal. It proposed an explicit formula for variance estimation based on a kernel method. Finite-sample performances of the estimators are studied by simulations. The proposed methodology is applied to analyze real data for illustration [8]. By [9] that a better measure of prediction accuracy for models is used the mean absolute percentage error (MAPE). The problem in this paper is how to determine the survivorship and the hazard functions of lognormal distribution to without and with covariate. The functions with covariate determine some coefficients of covariates, and then they could be used to count MAPE value by error.

The objectives of this paper are verifying data lognormal distribution, construct the survivorship and the hazard models without covariate, and determine the survivorship and the hazard models with covariate on data of participants of the incidence and risk factors for stroke. Finally, measuring of prediction accuracy of models by MAPE.

2. Literature Review

2.1. The Goodness of Fit Distribution
Distribution of statistics data could be known by the goodness of fit distribution. This way is done in order to follow with scientific method. The result is the right distribution or a good model. Let there is a continue random variable. Distribution of the random variable known by using the goodness of fit distribution. The goodness of fit uses Kolmogorov-Smirnov test. The steps of goodness of fit distribution to know data is lognormal distribution as follow:

1. Hypothesis
   \[ H_0 : F(t) = F_0(t) \] (Data is lognormal distribution)
   \[ H_1 : F(t) \neq F_0(t) \] (Data is not lognormal distribution)

2. Level of significant
   Level of significance is given \( \alpha \)

3. Test of statistics
   \[ D_n = \sup_{-\infty < c < \infty} \left| \hat{F}_n(t) - F_0(t) \right| \] (1)
   where:
   \( \hat{F}_n(t) \): empiric cumulative distribution of data
   \( F_0(t) \): lognormal cumulative distribution

4. Decision
   If \( D_n \) is greater than Kolmogorov-Smirnov table then \( H_0 \) is rejected, \( H_0 \) is accepted elsewhere.

2.2. The Survivorship and Hazard Functions of Lognormal Distribution
Let \( T \) denote a nonnegative random variable representing the failure time of a subject. The survivorship function of \( T \) is defined as the probability that \( T \) exceeds a time \( t \). The hazard function of \( T \) is defined as the probability that a subject fails at time \( t \) given that the subject has not failed before
time \( t \). The survival time \( T \) has a lognormal distribution if \( \log(T) \) has a normal distribution. The lognormal distribution has the probability density and survivorship functions, respectively,

\[
f(t) = \frac{1}{t \sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} \left(\log(t) - \mu\right)^2\right]
\]

(2)

where \( t > 0 \), \( \mu \) is mean of \( T \), \( \sigma \) is standard deviation of \( T \), and \( a = \exp(-\mu) \), and relations about the probability density and the survivorship result that

\[
S(t) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\log(t)} \exp\left[-\frac{1}{2\sigma^2} (\log(x) - \mu)^2\right] dx = 1 - G\left(\frac{\log(at)}{\sigma}\right)
\]

(3)

where \( G(.) \) is the cumulative distribution function of a standard normal variable.

The relationships between the survivorship and the hazard functions can be written as

\[
h(t) = \frac{f(t)}{S(t)}, \quad 0 < t < \infty
\]

Then the hazard function has the form

\[
h(t) = \left(\frac{1}{(t \sigma \sqrt{2\pi})}\right)\exp\left[-\left(\frac{\log(at)}{\sigma}\right)^2\right] / \left(1 - G(\log(at)/\sigma)\right)
\]

(4)

2.3. Lognormal AFT Model

The accelerated failure time model of survival time \( T \) is linear regression model constructed the relationship between natural logarithm of survival \( T \) and the covariates. The accelerated failure time model could be abbreviated by AFT model. The formula of AFT model is written by:

\[
\log(T) = a_0 + \sum_{j=1}^{p} a_j x_j + \sigma \varepsilon
\]

(5)

where \( x_j, j = 1, K, p \), are the covariates, \( a_j, j = 0, 1, K, p \) the coefficients, \( \sigma \) is an unknown scale parameter with positive value, and \( \varepsilon \) is the error term. The error \( \varepsilon \) is a random variable with has density function \( g(\varepsilon) \) and survivorship function \( G(\varepsilon) \).

Let \( \varepsilon \) in (5) be the standard normal random variable with the density function \( g(\varepsilon) \) and survivorship function \( G(\varepsilon) \), that is

\[
g(\varepsilon) = \frac{\exp(-\varepsilon^2/2)}{\sqrt{2\pi}}
\]

and

\[
G(\varepsilon) = 1 - \Phi(\varepsilon) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\varepsilon}^{\varepsilon} \exp(-x^2/2) dx
\]

where \( \Phi \) is the cumulative distribution function of the standard normal distribution. The lognormal AFT model for the survival time \( T \) is

\[
\log(T) = a_0 + \sum_{k=1}^{p} a_k x_k + \sigma \varepsilon
\]

(6)

where \( T \) must be shown that it has lognormal distribution. If \( \mu = a_0 + \sum_{k=1}^{p} a_k x_k \), then lognormal AFT model presented by

\[
\log(T) = \mu + \sigma \varepsilon
\]

(7)

where \( \mu \) is mean of \( T \) and \( \sigma \) is standard deviation of \( T \).

Then, the survival time \( T \) of the lognormal distribution has the density function
\[ f(t, \mu, \sigma^2) = \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right) \sqrt{2\pi\sigma} \] (8)

the survivorship function

\[ S(t, \mu, \sigma^2) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \] (9)

And the hazard function has the form

\[ h(t, \mu, \sigma^2) = \frac{\exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right)}{1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)} \] (10)

where \( \mu \) is linear regression function of covariate and \( \sigma \) is scale parameter.

2.4. Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics. It expresses accuracy a percentage, and is defined by the formula:

\[ \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|A_i - F_i|}{A_i} \] (11)

where \( A_i \) is the actual value, and \( F_i \) is the forecast value.

3. Results and Discussions

Let data of a random sample of 20 participants from a study of the incidence and risk factor for stroke. They have variables are ID is participant identity number; T is stroke event time (stroke-free time); CENS is censoring status, CENS = 0 is censored and CENS = 1 is uncensored; BMI is body mass index; SMK is smoking status, SMK = 1 if current smoking and SMK = 0 otherwise; ACH is alcoholic drinking status, ACH = 1 is current drinking and ACH = 0 is not; DM is diabetes mellitus status, DM = 1 if yes that is fasting plasma glucose \( \geq 126 \text{ mg/dl} \) or under diabetes treatments, and DM = 0 if no; ALBU is albuminuria status, ALBU = 1 if normal (the ratio of urinary albumin and creatinine (UACR) \( < 30 \)), ALBU = 2 if micro albuminuria (30 \( \leq \) UACR \( < 300 \)), and ALBU = 3 if macro albuminuria (300 \( \leq \) UACR). The data is presented in table 1.

The covariates of the data are BMI as \( X_1 \), SMK as \( X_2 \), ACH as \( X_3 \), DM as \( X_4 \), and ALBU as \( X_5 \). The variable T must be tested that T has lognormal distribution. Distribution test by one-sample Kolmogorov-Smirnov test result that mean = 1.9851; Std. Deviation = 0.86261. Data analysis by counting statistics and table statistics, respectively, presented at table 2 and table 3.

The steps of goodness of fit distribution as follow:

1. Hypothesis
   \[ H_0 : F(t) = F_0(t) \text{ (Data is lognormal distribution)} \]
   \[ H_1 : F(t) \neq F_0(t) \text{ (Data is not lognormal distribution)} \]

2. Level of significant
   Level of significance is given 5%

3. Test of statistics
   \[ D_n = \sup_{t < \infty} \left| \hat{F}_n(t) - F_0(t) \right| = 0.170 \]
   where:
   \( \hat{F}_n(t) \) : empiric cumulative distribution of data
   \( F_0(t) \) : lognormal cumulative distribution

4. Decision
Comparing in table 2 and table 3, these results that (0.170 < 0.294); (0.759 < 1.96) or (0.613 > 0.05). They can be concluded that reject $H_1$, the meaning is the data have lognormal distribution.

**Table 1.** Data of 20 participants of the incidence and risk factors for stroke

| ID | T  | CENS | BMI  | SMK | ACH | DM  | ALBU |
|----|----|------|------|-----|-----|-----|------|
| 1  | 2.7| 1    | 28.1 | 1   | 1   | 1   | 2    |
| 2  | 6.0| 1    | 40.1 | 1   | 0   | 1   | 2    |
| 3  | 11.4| 1    | 26.6 | 0   | 0   | 0   | 1    |
| 4  | 15.3| 1    | 38.0 | 1   | 0   | 1   | 1    |
| 5  | 2.3| 1    | 27.7 | 1   | 0   | 1   | 3    |
| 6  | 16.6| 1    | 29.0 | 1   | 0   | 0   | 1    |
| 7  | 7.3| 1    | 27.7 | 0   | 0   | 1   | 1    |
| 8  | 14.3| 1    | 45.7 | 0   | 0   | 1   | 2    |
| 9  | 0.8| 1    | 27.9 | 1   | 0   | 1   | 1    |
| 10 | 10.3| 1    | 33.3 | 0   | 0   | 1   | 2    |
| 11 | 16.3| 1    | 25.6 | 0   | 0   | 0   | 1    |
| 12 | 3.8| 1    | 36.8 | 0   | 0   | 1   | 1    |
| 13 | 1.8| 1    | 25.9 | 0   | 1   | 1   | 2    |
| 14 | 15.9| 1    | 27.9 | 0   | 1   | 0   | 1    |
| 15 | 6.6| 1    | 32.7 | 0   | 1   | 1   | 3    |
| 16 | 7.9| 1    | 28.3 | 0   | 0   | 1   | 3    |
| 17 | 15.5| 1    | 32.0 | 1   | 1   | 0   | 1    |
| 18 | 9.8| 1    | 31.2 | 0   | 0   | 1   | 1    |
| 19 | 8.7| 0    | 37.2 | 0   | 0   | 1   | 3    |
| 20 | 15.4| 0    | 23.5 | 0   | 0   | 1   | 1    |

Source: Courtesy of the Southeast Oklahoma Inter-Tribal Health Board.

**Table 2.** Counting statistics

| Most Extreme Differences | Absolute $D_n$ | Kolmogorov-Smirnov Z | Asymp. Sig. (2-tailed) |
|--------------------------|---------------|----------------------|------------------------|
|                          | 0.170         | 0.759                | 0.613                  |

**Table 3.** Table statistics

| Table statistics $D_n$ | 0.294 |
|------------------------|-------|
| Normal Distribution Z  | 1.96  |
| Significance level      | 0.05  |

The other variables to complete patient stroke condition are age, SBP is systole blood pressure, DPB is diastolic blood pressure, HDL is high density lipoprotein, and LDL is low density lipoprotein. They could be used to know about patient stroke condition. Description of patient stroke condition is written at table 4.

**Table 4.** Descriptive statistics of patient stroke conditions

|          | T      | AGE    | BMI    | SBP    | DPB    | HDL    | LDL    |
|----------|--------|--------|--------|--------|--------|--------|--------|
| Mean     | 9.435  | 61.320 | 31.260 | 134.10 | 82.40  | 40.10  | 126.10 |
| Mode     | 0.8    | 45.7   | 27.7   | 126    | 77     | 32     | 136    |
| Minimum  | 0.8    | 45.7   | 23.5   | 97     | 58     | 20     | 57     |
| Maximum  | 16.6   | 74.6   | 45.7   | 176    | 110    | 59     | 198    |
This paper discusses about the survivorship and the hazard functions without covariate and the survivorship and the hazard functions with covariate. For computation of data stroke event time above without covariate given that the mean $\mu$ is 1.985 and by table 5, scale $\sigma$ is 0.8930.

Table 5. Estimate of parameters stroke event time without covariate

| Parameters | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
|------------|----------|----------------|------------|------------|
| Intercept  | 2.1012   | 0.2085         | 101.56     | < 0.0001   |
| Scale      | 0.8930   | 0.1548         |            |            |

So the survivorship function without covariate is

$$S(t) = 1 - G\left(\log \frac{at}{0.893}\right)$$  \hspace{1cm} (12)

where $a = \exp(-1.985)$.

And the hazard function without covariate has the form

$$h(t) = \frac{\left(1/(\text{2.24}t)\right)\exp\left[-(\log at)^2/1.6\right]}{1 - G(\log(at/0.893))}$$  \hspace{1cm} (13)

Computation of the survivorship and the hazard functions at stroke event time $T$ without covariate presented by table 6.

Table 6. The survivorship and the hazard functions at stroke event time without covariate

| Nu. | T  | S(t)  | h(t)  |
|-----|----|-------|-------|
| 1   | 0.8| 0.9948| 0.0266|
| 2   | 1.8| 0.9475| 0.0773|
| 3   | 2.3| 0.9092| 0.0931|
| 4   | 2.7| 0.8748| 0.1022|
| 5   | 3.8| 0.7740| 0.1166|
| 6   | 6.0| 0.5875| 0.1237|
| 7   | 6.6| 0.5439| 0.1236|
| 8   | 7.3| 0.4972| 0.1230|
| 9   | 7.9| 0.4606| 0.1222|
| 10  | 8.7| 0.4164| 0.1208|
| 11  | 9.8| 0.3634| 0.1186|
| 12  | 10.3| 0.3419| 0.1176|
| 13  | 11.4| 0.2997| 0.1152|
| 14  | 14.3| 0.2151| 0.1091|
| 15  | 15.3| 0.1928| 0.1072|
| 16  | 15.4| 0.1908| 0.1070|
| 17  | 15.5| 0.1887| 0.1068|
| 18  | 15.9| 0.1808| 0.1060|
| 19  | 16.3| 0.1733| 0.1053|
| 20  | 16.6| 0.1680| 0.1047|

Computation of the survivorship and the hazard functions above can be used to predict the probability of stroke event time by survival and hazard functions.

While, computation of data stroke event time with covariate is given table 7 as follow.

By table 7, the mean model is

$$\mu = 1.4557 + 0.0587x_1 - 0.6531x_2 - 0.3675x_3 - 1.2666x_4 + 0.0033x_4$$  \hspace{1cm} (14)

Then, the lognormal AFT model for the survival time $T$ is
$$\log T = 1.4557 + 0.0587 x_1 - 0.6531 x_2 - 0.3675 x_3 - 1.2666 x_4 + 0.0033 x_5 + 0.6274 \varepsilon$$ \hspace{1cm} (15)

Table 7. Estimate of parameters stroke event time with covariate

| Parameters | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
|------------|----------|----------------|------------|------------|
| Intercept  | 1.4557   | 0.8604         | 2.86       | 0.0907     |
| BMI        | 0.0587   | 0.0276         | 4.52       | 0.0335     |
| SMK        | -0.6531  | 0.2997         | 4.75       | 0.0293     |
| ACH        | -0.3675  | 0.3491         | 1.11       | 0.2924     |
| DM         | -1.2666  | 0.3963         | 10.21      | 0.0014     |
| ALB        | 0.0033   | 0.2168         | 0.00       | 0.9877     |
| Scale      | 0.6274   | 0.1062         |            |            |

The survivor function with covariate is

$$S(t, \mu, \sigma^2) = 1 - \Phi \left( \frac{\log t - \mu}{0.6274} \right) \hspace{1cm} (16)$$

And the hazard function with covariate has the form

$$h(t, \mu, \sigma^2) = \frac{\exp \left[ -\left(\frac{\log t - \mu}{0.787}\right)^2 \right]}{(1.573 t)^{(\log t - \mu)/0.6274}} \hspace{1cm} (17)$$

If assumed that level of significance is 5%. By table 7, the significant covariate is body mass index (BMI), smoking (SMK), and diabetes mellitus (DM). Covariate of body mass index is positive effect, while covariates of smoking and diabetes mellitus are negative effect. Positive effect makes a person better his health, negative effect make a person easier has stroke. The unsignifcant covariate is alcoholic (ACH) and albuminuria (ALB). They cause small effect on stroke disease. So, a person could have stroke caused by condition of body mass index, smoking, and diabetes mellitus.

Table 8. The survivorship and hazard functions at stroke event time with covariate

| Nu. | T  | M    | S(t)  | h(t)  |
|-----|----|------|-------|-------|
| 1   | 0.8| 1.1736 | 0.9870 | 0.0675 |
| 2   | 1.8| 1.3451 | 0.8863 | 0.1923 |
| 3   | 2.3| 1.1685 | 0.7036 | 0.3404 |
| 4   | 2.7| 0.8212 | 0.3919 | 0.5786 |
| 5   | 3.8| 2.3492 | 0.9470 | 0.0478 |
| 6   | 6.0| 1.8931 | 0.5641 | 0.1854 |
| 7   | 6.6| 1.7476 | 0.4120 | 0.2281 |
| 8   | 7.3| 1.8150 | 0.3914 | 0.2142 |
| 9   | 7.9| 1.8568 | 0.3689 | 0.2063 |
| 10  | 8.7| 2.3792 | 0.6346 | 0.1085 |
| 11  | 9.8| 2.0204 | 0.3382 | 0.1758 |
| 12  | 10.3| 2.1470 | 0.3840 | 0.1539 |
| 13  | 11.4| 3.0204 | 0.8252 | 0.0436 |
| 14  | 14.3| 2.8749 | 0.6339 | 0.0661 |
| 15  | 15.3| 1.7665 | 0.0627 | 0.2047 |
| 16  | 15.4| 1.5685 | 0.0316 | 0.2325 |
| 17  | 15.5| 2.3168 | 0.2496 | 0.1308 |
| 18  | 15.9| 2.7292 | 0.4764 | 0.0838 |
| 19  | 16.3| 2.9617 | 0.6071 | 0.0619 |
| 20  | 16.6| 2.5082 | 0.3156 | 0.1081 |
Computation results of linear regression, the survivorship, and the hazard functions at stroke event time $T$ with covariate is presented by table 8.

Computation of linear regression, the survivorship and the hazard functions above can be used to predict stroke event time and probability of stroke event time by survival and hazard functions. This research has MAPE = 0.008729 = 0.8729%. This value means that 99.1271% of model is good to be used for prediction.

4. Conclusion
The survivorship function is used to predict probability of a person illness, while the hazard function is used to predict of person has health failure. Linear regression model can be used to predict illness event time. Risk factors which people have stroke caused by body mass index, smoking, and diabetes mellitus. Alcoholic and albuminuria as adding sick. Person has healthy should be health life. Measuring of prediction accuracy is used the mean absolute percentage error (MAPE). The result of MAPE is 0.8729%. This value means that 99.1271% of model is good for prediction.

References
[1] Conover W J 1980 Practical Nonparametric Statistics Second Edition (Canada: John Wiley & Sons, Inc).
[2] Kalbfleisch J D and Prentice R L 2002 The Statistical Analysis of Failure Time Data (Canada: John Wiley & Sons, Inc).
[3] Lawless J F 2003 Statistical Models and Methods for Lifetime Data (Canada: John Wiley & Sons, Inc).
[4] Lee E T and Wang J W 2013 Statistical Methods for Survival Data Analysis Fourth Edition (New Jersey: John Wiley & Sons, Inc).
[5] Lu W 2007 Life Data Analysis 13 75
[6] O’Connor AN, Madurres M and Mosleh A 2016 Probability Distributions Used in Reliability Engineering (Maryland: University of Maryland College Park).
[7] Saikia R and Barman P 2017 International Journal of Statistics and Systems 12 311.
[8] Takeshi E and Wang W 2016 Annals of the Institute of Statistical Mathematics 68 1073.
[9] Tofallis 2015 Journal of the Operational Research Society 66 135.