Simulation of parametric model towards the fixed covariate of right censored lung cancer data

Siti Afiqah Muhamad Jamil, M. Asrul Affendi Abdullah, Kek Sie Long, Oyebayo Ridwan Olaniran and Syahila Enera Amran

Faculty of Science, Technology and Human Development, Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor

Email: afendi@uthm.edu.my

Abstract. In this study, simulation procedure was applied to measure the fixed covariate of right censored data by using parametric survival model. The scale and shape parameter were modified to differentiate the analysis of parametric regression survival model. Statistically, the biases, mean biases and the coverage probability were used in this analysis. Consequently, different sample sizes were employed to distinguish the impact of parametric regression model towards right censored data with 50, 100, 150 and 200 number of sample. R-statistical software was utilised to develop the coding simulation with right censored data. Besides, the final model of right censored simulation was compared with the right censored lung cancer data in Malaysia. It was found that different values of shape and scale parameter with different sample size, help to improve the simulation strategy for right censored data and Weibull regression survival model is suitable fit towards the simulation of survival of lung cancer patients data in Malaysia.

1. Introduction
The application of survival analysis are varied according to the studies conducted by the researchers as the analysis normally considering the censored observation. When a study involve certain duration of time, the observation of survival could be censored. Before that, the event that occurs when conducting the analysis of survival can be the death of patient, relapse of disease, or the recovery of disease. In other words, it involves the time until an event occur [2]. Besides, as the observation is censored, the data will be not symmetrically distributed and could be either positively skewed or negatively skewed. The analysis comprises parametric, non-parametric and semi-parametric method of analysis. The application of analysis depends on the data of study for example, as the observation of data is not censored, then, a general or standard regression analysis can be applied to the study. Otherwise, if the observation of dependent variable is in binary form, the analysis applied the logistic regression analysis. Besides, count data can apply the generalized linear model (GLM) to conduct the analysis. Many forms of analysis have been proven and significant to be used depend on certain situations. The example of parametric method in survival analysis are the Weibull model, Exponential method Log-logistic model, lognormal model, and generalized gamma model while for the non-parametric method is the Kaplan-Meier data analysis and the example of semi-parametric method is the Cox proportional hazard model. In addition, the partly-interval censored (PIC) data is the combination of the left censored observation, right censored observation, interval censored...
observation and exact observation but it is not necessary to have all types of censored and depend of the situation [1].

Cancer has been a major public health burden that leads to the first cause of death globally. From the report of second National Cancer Registry of Cancer Incidence in Malaysia 2007, lung cancer is one of the five most leading cancers among the men and women. Besides, compare to other cancers that cause the death, lung cancer is known to be the majority in worldwide health issue [8]. The symptoms of lung cancer varied depending upon where and how widespread the tumour is and the sign involving lung cancer is not always present. Furthermore, there are four types of lung cancer which are the non-small cell lung cancer (NSCLC) and the small cell lung cancer (SCLC). NSCLC consists of three other types of lung cancer which are the adenocarcinoma that is peripherally located, squamous cell carcinoma that is centrally located, and large cell carcinoma that peripherally located and can be anywhere in the body. Consequently, SCLC normally related to smoking and is more aggressive than the other type which is the NSCLC. Moreover, the treatment of lung cancer disease can be seen as the following Figure 1.

![Figure 1: Staging and treatment of lung cancer.](image_url)

Adequately, in terms of method used by the previous researcher, [5] conducted a study involved the Gompertz model towards the right censored data and the interval censored data with time varying covariate. The interval estimation with right censored data used Wald and Jackknife technique to accurately measure the performances of the analysis. It was found that Jackknife method gives better estimation than Wald method towards the right censored data. Besides, Expectation Maximization (EM) algorithm is utilised to statistically consider the measurement by either real data or through simulation study [4]. Continuously, [6] observed the impact of the general right censoring on the survival analysis of data. The study claimed that, misinterpretation of the informative and non-informative involving the right censored observation could be biased. Based on the cases of lung cancer, [3] reviewed the increasing incidence of non-smoking lung cancer in United Kingdom. Besides, according to [9], Malaysia has been ranked 60th among 172 countries of worldwide statistics of the lung cancer disease.

As there is no proper procedure in analysing the right censored data with application to the lung cancer study, thus, the aims of this study are to propose parametric regression survival model into right censored of lung cancer data in measuring the parameter with fixed covariate (FC) and compared with the simulation of right censored data.

### 2. The Simulation of Weibull Regression Model with Right Censored Data

According to [2] and [7], Weibull regression model can be classified by the censored observation. Observation of survival data may or may not involve the censored especially when the study used certain types of health disease that need the follow up time in order to survive. By using the survival function, hazard function, density function and the likelihood function, the Weibull functions follow [1],

\[
S(t) = \exp \left[ - \left( \frac{t}{b} \right)^a \right]
\]

\[
h(t) = \left( \frac{a}{b} \right) \left( \frac{t}{b} \right)^{a-1}, H(t) = \int_0^t h_0(t) \exp(x'\beta) \]

\[
f(t) = \frac{dP(t)}{dt} = \left( \frac{a}{b} \right) \left( \frac{t}{b} \right)^{a-1} \exp \left[ - \left( \frac{t}{b} \right)^a \right]
\]
\[ L(\theta) = \prod_{i=1}^{n} \left\{ \left( \frac{a}{b} \right)^{t_{i}^{a}-1} \exp \left[ - \left( \frac{t_{i}^{a}}{b} \right) \right] \right\} \delta_{i} \left\{ \exp \left[ - \left( \frac{t_{i}^{a}}{b} \right) \right] \right\}^{1-\delta_{i}} \] (4)

Where the \( S(t) \) represent the survival function, \( h(t) \) is the hazard function, \( f(t) \) is the density function while \( L(\theta) \) is the likelihood function. Besides, \( a \) is the shape parameter and \( b \) is the scale parameter with \( t_{i} \) as the survival time with \( i^{th} \) observations.

Before that, Weibull regression model is selected to perform the analysis as it is a reliable and one of the simplest method to be compared with censored observation of survival data. There are four approaches of modelling the survival model which consist of the parametric families, proportional hazard, accelerated failure time and the proportional odds. Weibull regression model is one of the simple analysis in parametric families while the hardest considering the accelerated failure time model.

\[ \delta_{i} = \begin{cases} 1, & \text{if observation is not censored (exact observation)} \\ 0, & \text{if observation is right censored} \end{cases} \]

As the observation follows the right censored data, there is some modification on the likelihood function and this study need to obtain the derivation of survival time to perform a complete procedure in comparing the function. The likelihood function of right censored data follow,

\[ L(\theta) = \prod_{i=1}^{n} \{ f(t) \delta_{i} \{ S(t_{i}) \}^{1-\delta_{i}} \] (5)

\[ \log(L(\theta)) = \sum_{i=1}^{n} (\delta_{i}) \log \left( \left( \frac{a}{b} \right)^{t_{i}^{a}-1} \exp \left[ - \left( \frac{t_{i}^{a}}{b} \right) \right] \right) + (1 - \delta_{i}) \log \left\{ \exp \left[ - \left( \frac{t_{i}^{a}}{b} \right) \right] \right\} \]

\[ \frac{\partial l(\theta)}{\partial a} = \sum_{i=1}^{n} (\delta_{i}) \left\{ \left( \frac{b^{(t_{i}^{a})}}{t_{i}^{a}} \log \left( \frac{t_{i}^{a}}{b^{a}} \right) + \frac{1}{t_{i}^{a}} \right) - \left( \frac{t_{i}^{a}}{b^{a}} \right) \log \left( \frac{t_{i}^{a}}{b^{a}} \right) \right\} + \sum_{i=1}^{n} (1 - \delta_{i}) \left\{ \left( \frac{t_{i}^{a}}{b^{a}} \right) \log \left( \frac{t_{i}^{a}}{b^{a}} \right) \right\} \] (6)

\[ \frac{\partial^{2} l(\theta)}{\partial a^{2}} = \sum_{i=1}^{n} (\delta_{i}) \left\{ \left( \frac{b^{(t_{i}^{a})}}{t_{i}^{a}} \log^{2} \left( \frac{t_{i}^{a}}{b^{a}} \right) + \frac{2 \log \left( \frac{t_{i}^{a}}{b^{a}} \right) + \frac{2 \log \left( \frac{t_{i}^{a}}{b^{a}} \right) + 1}{t_{i}^{a}} \right) - \left( \frac{t_{i}^{a}}{b^{a}} \right) \log^{2} \left( \frac{t_{i}^{a}}{b^{a}} \right) \right\} + \sum_{i=1}^{n} (1 - \delta_{i}) \left\{ \left( \frac{t_{i}^{a}}{b^{a}} \right) \log^{2} \left( \frac{t_{i}^{a}}{b^{a}} \right) \right\} \] (7)

\[ \frac{\partial l(\theta)}{\partial b} = \sum_{i=1}^{n} (\delta_{i}) \left\{ \left( \frac{-((a-1) \log \left( \frac{t_{i}^{a}}{b^{a}} \right) + 1)}{t_{i}^{a}} \right) + \left( \frac{a^{(t_{i}^{a})}}{b^{a}} \right) \right\} + \sum_{i=1}^{n} (1 - \delta_{i}) \left\{ \left( \frac{a^{(t_{i}^{a})}}{b^{a}} \right) \right\} \] (8)

\[ \frac{\partial^{2} l(\theta)}{\partial b^{2}} = \sum_{i=1}^{n} (\delta_{i}) \left\{ \left( \frac{-((a-1) \log \left( \frac{t_{i}^{a}}{b^{a}} \right) + 2a-1)}{t_{i}^{a}} \right) + \left( \frac{a^{(t_{i}^{a})}}{b^{a}} \right) \right\} + \sum_{i=1}^{n} (1 - \delta_{i}) \left\{ \left( \frac{a^{(t_{i}^{a})}}{b^{a}} \right) \right\} \] (9)

Those likelihood functions were used to obtain the parameters of shape parameter \( a \) and the scale parameter \( b \). Consequently, since this analysis used the fixed covariate of data, to obtain the survival
time for the simulation, this study assumed time follows uniform distribution. Hence the survival time, \( T \) follows,

\[
\Lambda(t) = \left( \frac{t}{b} \right)^a \exp(x' \beta)
\]

\[
S(t) = \exp\{-\Lambda(t|x_i)\}
\]

\[
S(t) \sim U(0,1)
\]

\[
U = \exp\left\{ -\left( \frac{t}{b} \right)^a \exp(x' \beta) \right\}
\]

\[
- \frac{\log U}{\exp(x' \beta)} = \left( \frac{t}{b} \right)^a
\]

\[
T = b \left( -\frac{\log U}{\exp(x' \beta)} \right)^{\frac{1}{a}}
\] (10)

3. **Assess the Simulation Procedure using R-Statistical Software**

The results of simulated data is observed based on the cumulative survival distribution of right censored simulated data within 120 days length of observation. Within duration of study, patient that is over 120 days is assigned as right censored observation as the future results is unknown. Before that, this simulation is obtained by imitating the variables of lung cancer data in Malaysia.

![Cumulative survival distribution](image)

**Figure 2.** The cumulative distribution of right censored simulated data.

**Table 1.** The bias, mean bias, and the coverage probability (where \( a \) is shape parameter and \( b \) is the scale parameter) with \( n=50, 100, 150 \) and 200.
Table 2. The bias, mean bias, and the coverage probability (where $a$ is shape parameter and $b$ is the scale parameter) with $n=50$, 100, 150 and 200.

| Model | Sample size | Variables | Estimates | $bias(\hat{\theta})$ | Mean bias | Coverage Probability |
|-------|-------------|-----------|-----------|----------------------|-----------|----------------------|
|       | n=50        | $x_1$     | 0.3289    | 0.4045              | 0.7570    | 1.000                |
|       |             | $x_2$     | 0.2123    | 0.3943              | 0.6436    | 1.000                |
|       |             | $x_3$     | 0.4156    | 0.4257              | 0.9567    | 0.960                |
|       |             | $x_4$     | 0.5318    | 0.4394              | 1.1206    | 0.980                |
|       |             | $x_5$     | 0.2108    | 0.3879              | 0.6756    | 1.000                |
|       | n=100       | $x_1$     | 0.2964    | 0.4687              | 0.9020    | 0.980                |
| $a=2$ |             | $x_2$     | 0.2080    | $0.4300$            | 0.6906    | 0.990                |
| $b=\frac{1}{0.002}$ | n=150     | $x_3$     | 0.3995    | 0.5008              | 1.2453    | 0.970                |
|       |             | $x_4$     | 0.5070    | 0.5472              | 1.5876    | 0.980                |
|       |             | $x_5$     | 0.2053    | 0.4315              | 0.6826    | 1.000                |
|       | n=200       | $x_1$     | 0.2843    | 0.5907              | 1.5350    | 0.980                |
| $a=4$ |             | $x_2$     | 0.1992    | $0.4654$            | 0.8079    | 0.987                |
| $b=\frac{1}{0.004}$ | n=300     | $x_3$     | 0.3831    | 0.5832              | 1.8106    | 0.987                |
|       |             | $x_4$     | 0.4894    | 0.6484              | 2.3013    | 0.993                |
|       |             | $x_5$     | 0.1945    | 0.4627              | 0.7693    | 1.000                |

Table 1 and Table 2 show the simulation study of right censored observation based on the scenario of lung cancer data. The variables $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$ are the age of patients, gender (male and female), races (Malay, Chinese and Indian), types of lung cancer (Squamous cell (NSCLC),
adenocarcinoma (NSCLC), large cell lung cancer (NSCLC), and small cell lung cancer (SCLC) and treatment of lung cancer (chemotherapy, chemotherapy and surgery, chemoradiotherapy (CCRT) and chemotherapy and targeted therapy respectively. Overall, both Table 1 and Table 2 that are Model 1 and Model 2 were differentiated by using different shape parameter and scale parameter to observe the best model of Weibull right censored observation. Before that, the results showed that the increasing value of shape parameter from 2 to 4 and decreasing the scale parameter from 1/0.002 to 1/0.004 cause the bias and the mean bias to be increased. Both Model 1 and Model 2 are tested by using different sample size. Besides, in order to observe the correct simulation study, this study checks on the estimated result and the coverage probability. Both Model 1 and Model 2 have achieved the accuracy in simulating the data as most of the estimated result are approximately to the assume parameters that are assigned in this study. The coverage probability is said to be accepted with greater than 70% and all the simulation results are having 80% and above coverage probability.

4. Application of Lung Cancer Data.

Table 3. Overall number of lung cancer patients based on the lung cancer variables.

| Variable | Level | 0 | 1 | 2 | 3 | 4 |
|----------|-------|---|---|---|---|---|
| Gender   | 14    | 22 |   |   |   |   |
| Races    | 11    | 23 | 2 |   |   |   |
| Cell Type| 13    | 7  | 9 | 7 |   |   |
| Treatment| 5     | 7  | 21| 3 |   |   |

Based on Table 3, 22 number of male patients are having lung cancer compared to female patients. Next, 23 number of Chinese people were diagnosed with lung cancer compared to Malay and Indian. Furthermore, the highest type of lung cancer is the squamous cell lung cancer with 13 number of lung cancer patients. Additionally, the most preferred treatment that is used is the chemoradiotherapy (CCRT) to treat the cancer of lung.

Figure 3. Kaplan Meier Survival Plot.

Analysis of survival observed the survival curve and the survival model of observations in study. The graphical ways in presenting the survival showed the longest survival time for each patient. Based on
Figure 3, the curve shows that as the censoring rate increasing, the percent bias is also increasing. The result indicates increasing hazard rate as survival deceases.

5. Comparing Simulation Study and Lung Cancer data in Malaysia

The parametric methods which are Lognormal and Weibull regression survival model were used to differentiate the lung cancer data and simulation study.

| Model | Parameter | Standard Error | Log-likelihood | Z-value | P-value |
|-------|-----------|----------------|----------------|---------|---------|
| survreg | Lognormal | Intercept | 3.282 | 0.157 | -112.9 | 20.87 | 1.09e-96 |
| | | log(scale) | -0.188 | 0.139 | -1.36 | 1.75e-01 |
| | Weibull | Intercept | 3.684 | 0.146 | -114.0 | 25.26 | 7.87e-141 |
| | | log(scale) | -0.333 | 0.148 | -2.26 | 2.39e-02 |
| flexsurvreg | Lognormal | meanlog | 3.282 | 0.157 | -112.8922 | AIC=229.7844 |
| | sdlog | 0.828 | 0.115 |
| | Weibull | Shape | 1.396 | 0.206 | -114.0467 | AIC=232.0934 |
| | Scale | 39.823 | 5.808 |

As biases and the mean biases increased in Model 2 with shape, \( a = 4 \) and scale parameter, \( b = \frac{1}{0.004} \), this study would prefer Model 1 with shape, \( a = 2 \) and scale parameter, \( b = \frac{1}{0.002} \) for 200 number of observations.

| Model | Parameter | Estimate | Standard Error | Log-likelihood | Z-value | P-value |
|-------|-----------|----------|----------------|----------------|---------|---------|
| survreg | Lognormal | Intercept | 11.38 | 0.3136 | -1490.7 | 36.3 | 2.20e-288 |
| | | log(scale) | 1.24 | 0.0613 |
| | Weibull | Intercept | 13.76 | 0.3735 | -1519.6 | 36.8 | 5.64e-297 |
| | | log(scale) | 1.34 | 0.0561 |
| flexsurvreg | Lognormal | meanlog | 11.38 | 0.314 | -1490.68 | AIC=2985.364 |
| | sdlog | 3.452 | 0.211 |
| | Weibull | shape | 2.61e-01 | 1.47e-02 | -1519.56 | AIC=3043.12 |
| | scale | 9.43e+05 | 3.52e+05 |

For the fixed covariate, analysis can use survreg while for time varying covariate flexsurvreg will be more appropriate. In this analysis, study observed both the survreg and flexsurvreg coding of R statistical software. In Table 4, the result shows that Lognormal regression model with scale=0.828 is significant with \( P \)-value< 0.05 and the Akaike Information Matrics (AIC) value in Lognormal is lower compared to Weibull model. This result is similar when performing the simulation procedure in Table 5, this indicates that the final model follows,

For application of lung cancer data: \( Y = 11.38 + 1.24 \log(scale) \)

For the simulation data: \( Y = 11.38 + 3.452 \log(scale) \)
6. Conclusion
This study enhance the use of simulation data and application towards the real data of lung cancer to find the best, stable and reliable method for right censored data. Survival analysis have always considering the censored observation and the researchers need to effectively and accurately apply suitable method so that biases could be reduced. The aims of this study are to propose parametric regression survival model into right censored of lung cancer data in measuring the parameter with fixed covariate (FC) and compared with the simulation of right censored data. Firstly, this study used Weibull regression survival model in conducting the simulation procedure. Then, different values of shape and scale parameter were assigned to obtain the best model to be compared with real lung cancer data in Malaysia. The lung cancer data was then being analysed with Weibull regression survival model and Lognormal regression survival model. Next, the simulation data was then repeated with using the similar use of method. Both lung cancer data and simulation data found that Lognormal regression model is more stable in analysing the right censored observation. However, Weibull regression survival model is more stable to be used in conducting the simulation study as overall results of the coverage probability employed the approximate estimation of parameter used in this study. Further analysis should be focusing on the variable so that further details involving the lung cancer disease will be known by people.

Acknowledgement
We would like to thank the Ministry of Higher Education and Universiti Tun Hussien Onn Malaysia, UTHM for supporting this research project.

References
[1] Alharpy, A.M. & Ibrahim, N.A. 2013. Journal of Applied Sciences, 13(4), p. 621.
[2] Collett, D. 2015. Modelling survival data in medical research. CRC press.
[3] Cufari, M.E., Proli, C., Phull, M., Raubenheimer, H., Al Sahaf, M., Asadi, N., Perikleous, P., Allan, A., Shedden, L., Chavan, H. et al. 2015. Increasing incidence of non-smoking lung cancer: Presentation of patients with early disease to a tertiary institution in the uk. In: JOURNAL OF THORACIC ONCOLOGY, volume 10, ELSEVIER SCIENCE INC 360 PARK AVE SOUTH, NEW YORK, NY 10010-1710 USA, pp. S793–S794.
[4] Elfaki, F.A.M., Azram, M. & Usman, M. 2012. Journal of Applied Physics and Mathematics, 2(5), p. 352.
[5] Kiani, K., Arasan, J. and Midi, H. 2012. Sains Malaysiana. 41(4), 471-480.
[6] Lagakos, S.W. 1979. Biometrics, 35(1), pp. 139–156.
[7] Murthy, D. P., Bulmer, M. and Eccleston, J. A. 2004. Reliability Engineering & System Safety. 86(3), 257–267.
[8] Philip C Hoffman, Ann M Mauer, E.E.V. 2000. Lung cancer. The Lancet, 355(9202), pp. 479–485.
[9] Stewart, B., Wild, C.P. et al. 2016. World cancer report 2014. World.