Undergraduate Journal of Mathematical Modeling: One + Two

Volume 6 | 2014 Fall Article 4

2014

Survival Analysis - Breast Cancer

Minh Hoang Pham
University of South Florida

Advisors:
Arcadii Grinshpan, Mathematics and Statistics
Ram C. Kafle, Mathematics and Statistics
Yu Chen, Molecular Medicine

Problem Suggested By: Ram C. Kafle

Field of Study for Problem Suggester: Mathematics and Statistics

Recommended Citation
Pham, Minh Hoang (2014) "Survival Analysis - Breast Cancer," *Undergraduate Journal of Mathematical Modeling: One + Two*: Vol. 6: Iss. 1, Article 4.
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860
Available at: https://digitalcommons.usf.edu/ujmm/vol6/iss1/4

Survival Analysis - Breast Cancer

Abstract

In this study we used the parametric survival approach to analyze the survival time of African American breast cancer patients. In research about the survival time of patients, the Cox Proportional Hazard model, a semi-parametric method, is primarily used, but this approach does not rely on the distributional assumptions. The parametric method is more consistent with a theoretical approach compared to a semi-parametric approach. We observe that the survival time of African American breast cancer patients follows the Weibull Probability Distribution. First, we fitted the model to represent the survival ability of the general population (African American women diagnosed with breast cancer), treating everyone the same. Second, we incorporate other patient-specific covariate factors affecting the survival time to predict lifetime of a particular patient, given that her information is fully known.

Keywords

breast cancer, survival analysis, Cox Hazard Model

MOTIVATION

Breast cancer is one of the most dangerous diseases and is the most commonly contracted cancer among women. The American Cancer Society estimated that in the year 2013, about 232,340 new cases of invasive breast cancer would be diagnosed; and about 39,620 women would die because of this deadly disease (American Cancer Society).

African Americans are among the races most affected by breast cancer disease. According to the statistics from the National Cancer Institute, African American women have the second highest rate of breast cancer, only slightly less than white Americans (121.4 compared to 127.4 cases per 100,000 persons). In addition, African American women have the highest death rate of those diagnosed with breast cancer. The five-year survival rate for African American breast cancer is 78% (Sisters Network). This means that for every 100 African American women diagnosed with breast cancer, there are on average 22 people that die within the first 5 years of being diagnosed. Previous studies have also shown that African- American women are more likely to have late-stage breast cancer at diagnosis and shortened survival (Bradley).

Statistics about breast cancer provided by the National Cancer Institute, such as five-year survival rate, death rates by age group, etc. cannot satisfy human need. Those numbers cannot help us answer essential questions such as:

1. How is survival time of breast cancer patients distributed among different races, age ranges, stages of the disease, etc.?
2. How do factors such as age, stage of cancer, and type of treatment affect survival time?

3. Which treatments are effective to cure breast cancer of a patient with particular age, race, and stage? How much can the treatments lengthen the survival time?

The main purpose of our study is to answer the questions above using information about African American breast cancer patients from the Surveillance Epidemiology and End Results (SEER) database of National Cancer Institute (NCI) from 1973 to 2010. In other words, we evaluate the potential risk factors and determine the best treatment methods that are suitable for the patients to increase the survival time based on their information such as age, tumor size, stage of the disease etc.

According to the National Cancer Institute, important factors affecting the survival of cancer patients are type and location of the cancer, stage of cancer, patient's age, race and overall general health, etc. (National Cancer Institute). The information included in this study includes survival time, cause-specific death classification, age, stage of cancer, types of treatment (radiation, surgery, both radiation and surgery). Specific description of data is presented in Appendix A.

We observe that the survival times of African American breast cancer patients are best characterized by the Weibull probability distributions. The data distribution was fitted using both STATISTICA and SPSS software. The results and validation of those fits are presented in Appendix B and Appendix C.

In the next section we derive the model to estimate the survival time based on the Weibull probability distribution and fit the data using SAS statistical software.

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

MATHEMATICAL DESCRIPTION AND SOLUTION APPROACH

The Weibull probability distribution function is given by: $y = f(x) = \lambda\alpha t^{\alpha-1}e^{-\lambda t^{\alpha}}$ ($\lambda, \alpha>0$, $t\geq 0$) with the parameter being $\theta = (\lambda, \alpha)$ (Klein).

Therefore, the survival function is:

$$S(t) = \int_t^{\infty} f(t) = \int_t^{\infty} \lambda\alpha t^{\alpha-1}e^{-\lambda t^{\alpha}} \quad (1)$$

Let $u = e^{-\lambda t^{\alpha}}$ then $du = -\lambda\alpha t^{\alpha-1}e^{-\lambda t^{\alpha}}dt$. Hence,

$$S(x) = -\int_u^0 du = \int_0^u du = u = e^{-\lambda t^{\alpha}} \quad (2)$$

The hazard rate is

$$h(t) = \frac{P(t\leq T\leq t+dt|T\geq t)}{dt} = \frac{f(t)}{S(t)} = \lambda\alpha t^{\alpha-1} \quad (3)$$

We want to estimate these two parameters so that the distribution can fit the best to the data, we use the maximum likelihood method (Evans):

Given the independent and identically distributed sample $t_1, t_2 \ldots t_n$ ($t_1, t_2 \ldots t_n \geq 0$), the joint density function is

$$L = f(t_1, t_2, \ldots, t_n|\theta) = f(t_1|\theta).f(t_2|\theta) \ldots f(t_n|\theta) = \prod_{i=1}^{n}(\lambda\alpha {t_i}^{\alpha-1}e^{-\lambda {t_i}^{\alpha}}) \quad (4)$$

However, because the sample is given (from the data), we can treat the function above as a function of θ. Our job is to define the value of θ that maximize the probability to experience the given sample $(t_1, t_2, \ldots, t_n)$, i.e. maximize L.

Since the function ln(x) is increasing on (0, ∞), maximizing ln L is the same as maximizing L

$$\ln L = \ln \prod_{i=1}^{n} (\lambda \alpha {t_i}^{\alpha-1} e^{-\lambda {t_i}^{\alpha}}) = \sum_{i=1}^{n} (\ln \lambda + \ln \alpha + (\alpha - 1) \ln t_i - \lambda {t_i}^{\alpha}) \quad (5)$$

We can maximize this function by setting $\frac{\partial \ln L}{\partial \lambda} = 0$ and $\frac{\partial \ln L}{\partial \alpha} = 0$

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^{n} (\frac{1}{\lambda} - {t_i}^{\alpha}) = \frac{n}{\lambda} - \sum_{i=1}^{n} {t_i}^{\alpha} \quad (6)$$

So $\frac{\partial \ln L}{\partial \lambda} = 0$ means

$$\frac{n}{\lambda} - \sum_{i=1}^{n} {t_i}^{\alpha} = 0 \quad (7)$$

So,

$$\lambda = \frac{n}{\sum_{i=1}^{n} {t_i}^{\alpha}} \quad (8)$$

Similar to $\frac{\partial \ln L}{\partial \alpha}$,

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} (\frac{1}{\alpha} + \ln t_i - \lambda {t_i}^{\alpha} \ln t_i) = 0 \quad (9)$$

$$\frac{n}{\alpha} + \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} ({t_i}^{\alpha} \ln t_i) = 0 \quad (10)$$

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

From (8), replace $\lambda = \frac{n}{\sum_{i=1}^{n} {t_i}^{\alpha}}$ to (10) to get

$$\frac{n}{\alpha} + \sum_{i=1}^{n} \ln t_i - \frac{n}{\sum_{i=1}^{n} {t_i}^{\alpha}} \sum_{i=1}^{n} ({t_i}^{\alpha} \ln t_i) = 0 \tag{11}$$

It is difficult to solve this equation normally. However, we can prove that equation (11) has at most one solution of α for any sample of x_i (Balakrishnan).We can use software (SAS) to estimate one solution of α and that solution is the only one.

Consider the LHS a function of α

$$g(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n} \ln t_i - \frac{n}{\sum_{i=1}^{n} {t_i}^{\alpha}} \sum_{i=1}^{n} (t^{\alpha} \ln t_i) \tag{12}$$

$$g'(\alpha) = \frac{-n}{\alpha^2} - n \frac{\sum_{i=1}^{n} (x^{\alpha} t_i)' \sum_{i=1}^{n} {t_i}^{\alpha} - \sum_{i=1}^{n} {t_i}^{\alpha} lnt \sum_{i=1}^{n} ({t_i}^{\alpha})'}{\left(\sum_{i=1}^{n} {x_i}^{\alpha}\right)^2}$$

$$= \frac{-n}{\alpha^2} - n \frac{\sum_{i=1}^{n} {x_i}^{\alpha} (\ln t_i)^2 \sum_{i=1}^{n} {t_i}^{\alpha} - (\sum_{i=1}^{n} {t_i}^{\alpha} lnt_i)^2}{\left(\sum_{i=1}^{n} {t_i}^{\alpha}\right)^2} \tag{13}$$

Using the Cauchy–Schwarz inequality we obtain:

$$\sum_{i=1}^{n} {t_i}^{\alpha} (\ln t_i)^2 \sum_{i=1}^{n} {t_i}^{\alpha} = \sum_{i=1}^{n} (\sqrt{{t_i}^{\alpha}} \ln t_i)^2 \sum_{i=1}^{n} \sqrt{{t_i}^{\alpha}}^2 \geq (\sum_{i=1}^{n} {t_i}^{\alpha} lnt_i)^2 \tag{14}$$

Therefore,

$$\sum_{i=1}^{n} {t_i}^{\alpha} (\ln t_i)^2 \sum_{i=1}^{n} {t_i}^{\alpha} - (\sum_{i=1}^{n} {t_i}^{\alpha} lnt_i)^2 \geq 0 \tag{15}$$

Hence, $g'(\alpha) \leq 0$. This means that $g(\alpha)$ is a decreasing function and if we can find a solution of (3), that is the only solution.

Once we have the result of α from software (SAS), the value of λ is easily determined from (8).

Nevertheless, the true value of t_i is not always available. In the study of breast cancer, there are events such as patients dropping their treatments, patients dying due to causes different from the breast cancer, etc. These events cause the value of t_i to be censored, but we are still sure that $x_i$is definitely larger than a certain number (the time between the beginning and the censoring events). This is the right-censored data.

Therefore, we would like to make likelihood estimations for the right-censored data (Klein). Let δ denote the status of the data (δ = 0 when lifetime T is censored and δ = 1 when the lifetime X is observed). And T is equal to T if the lifetime is observed.

The likelihood function is:

$$L = \prod_{i=1}^{n}\left[f(t_i)^{\delta}\right]\left[S(t_i)^{1-\delta}\right] = \prod_{i=1}^{n}\left(\lambda\alpha {t_i}^{\alpha-1}e^{-\lambda {t_i}^{\alpha}}\right)^{\delta}\left(e^{-\lambda {t_i}^{\alpha}}\right)^{1-\delta}$$

$$= \prod_{i=1}^{n}(\lambda\alpha {t_i}^{\alpha-1})^{\delta}.\, e^{-\lambda {t_i}^{\alpha}} \tag{16}$$

$$lnL = \sum_{i=1}^{n}(\delta ln\lambda - \delta ln\alpha + \delta(\alpha-1)lnt_i - \lambda {t_i}^{\alpha}) \tag{17}$$

Similar to the non-censor case, we set $\frac{\partial \ln L}{\partial \lambda} = 0$ and $\frac{\partial \ln L}{\partial \alpha} = 0$

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^{n}\left(\frac{\delta}{\lambda} - {t_i}^{\alpha}\right) = \frac{n\delta}{\lambda} - \sum_{i=1}^{n}({t_i}^{\alpha}) = 0 \tag{18}$$

$$\lambda = \frac{n\delta}{\sum_{i=1}^{n} {t_i}^{\alpha}} \tag{19}$$

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

And:

$$\frac{\partial \ln L}{\partial \alpha} =$$

$$= \frac{n\delta}{\alpha} + \delta \sum_{i=1}^{n} \ln t_i - \lambda \sum_{i=1}^{n} ({t_i}^{\alpha} \ln t_i)$$

$$= \frac{n\delta}{\alpha} + \delta \sum_{i=1}^{n} \ln t_i - \frac{n\delta}{\sum_{i=1}^{n} {t_i}^{\alpha}} . \sum_{i=1}^{n} ({t_i}^{\alpha} \ln t_i)$$

$$= 0 \quad (20)$$

So

$$\frac{n}{\alpha} + \sum_{i=1}^{n} \ln t_i - \frac{n}{\sum_{i=1}^{n} {t_i}^{\alpha}} . \sum_{i=1}^{n} ({t_i}^{\alpha} \ln t_i) = 0 \quad (21)$$

This is the same as (11). Therefore, in case we have censored values, we still have the same α. λ is determined by (19).

By using SAS, we obtain the estimate of α = 0.746, the estimate of λ = 0.059. SAS outputs are presented in Appendix D

So our probability density function is:

$$f(t) = 0.044 t^{-0.254} e^{-0.059 t^{0.746}} \quad (22)$$

Survival function:

$$S(t) = e^{-0.059 t^{0.746}} \quad (23)$$

Hazard function:

$$h(t) = 0.044 t^{-0.254} \quad (24)$$

Estimation of Variance-covariance matrix for $\hat{\lambda}$ and $\hat{\alpha}$

We need to find the three values: $Var(\hat{\alpha}), Var(\hat{\lambda}), Cov(\hat{\alpha}, \hat{\lambda})$

The estimators of λ and α are given by $\hat{\alpha} = \frac{1}{\hat{\sigma}}$, and $\hat{\lambda} = e^{-\frac{\hat{\mu}}{\hat{\sigma}}}$ (Klein).

Applying the delta method (Papanicolaou):

Using the univariate delta method for $\hat{\alpha}$, we get

$$Var(\hat{\alpha}) = Var(\hat{\sigma}).\left(\left(\frac{1}{\hat{\sigma}}\right)'\right)^2 = Var(\hat{\sigma}).\frac{1}{\hat{\sigma}^4} \tag{25}$$

Using the multivariate delta method for $\hat{\lambda}$, let $\varphi = (\mu, \sigma)$:

According to the central limit theorem:

$$\sqrt{n}(\hat{\varphi}_n - \varphi) \xrightarrow{D} N_k(0, \Sigma) \tag{26}$$

where $\Sigma = \begin{pmatrix} Var(\hat{\mu}) & Cov(\hat{\mu}, \hat{\sigma}) \\ Cov(\hat{\mu}, \hat{\sigma}) & Var(\hat{\sigma}) \end{pmatrix}$

Let g($\hat{\varphi}$) = $e^{-\frac{\hat{\mu}}{\hat{\sigma}}}$

$$\nabla g(\hat{\varphi}) = \begin{pmatrix} \frac{\partial g(\hat{\varphi})}{\partial \hat{\mu}} \\ \frac{\partial g(\hat{\varphi})}{\partial \hat{\sigma}} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\hat{\sigma}} e^{-\frac{\hat{\mu}}{\hat{\sigma}}} \\ \frac{\hat{\mu}}{\hat{\sigma}^2} e^{-\frac{\hat{\mu}}{\hat{\sigma}}} \end{pmatrix} \tag{27}$$

$$(\nabla g(\hat{\varphi}))^T = \left(-\frac{1}{\hat{\sigma}} e^{-\frac{\hat{\mu}}{\hat{\sigma}}} \quad \frac{\hat{\mu}}{\hat{\sigma}^2} e^{-\frac{\hat{\mu}}{\hat{\sigma}}}\right) \tag{28}$$

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

Then,

$$\sqrt{n}(g(\hat{\varphi}_n) - g(\varphi)) \xrightarrow{D} N_k(0,\Sigma) = (\nabla g(\hat{\varphi}))^T N_k(0,\Sigma) \tag{29}$$

So,

$$Var(\hat{\lambda}) = Var(g(\hat{\varphi}_n))$$

$$= \nabla g(\hat{\alpha},\hat{\sigma}).\Sigma.\nabla g(\hat{\alpha},\hat{\sigma})$$

$$= \left(-\frac{1}{\hat{\sigma}}e^{-\frac{\hat{\mu}}{\hat{\sigma}}} \quad \frac{\hat{\mu}}{\hat{\sigma}^2}e^{-\frac{\hat{\mu}}{\hat{\sigma}}}\right)\begin{pmatrix} Var(\hat{\mu}) & Cov(\hat{\mu},\hat{\sigma}) \\ Cov(\hat{\mu},\hat{\sigma}) & Var(\hat{\sigma}) \end{pmatrix}\begin{pmatrix} -\frac{1}{\hat{\sigma}}e^{-\frac{\hat{\mu}}{\hat{\sigma}}} \\ \frac{\hat{\mu}}{\hat{\sigma}^2}e^{-\frac{\hat{\mu}}{\hat{\sigma}}} \end{pmatrix}$$

$$= e^{-\frac{2\hat{\mu}}{\hat{\sigma}}}\left[\frac{Var(\hat{\mu})}{\hat{\sigma}^2} + \frac{\hat{\mu}^2}{\hat{\sigma}^4}Var(\hat{\sigma}) - 2\frac{\hat{\mu}}{\hat{\sigma}^3}Cov(\hat{\mu},\hat{\sigma})\right] \tag{30}$$

$$Var(\hat{\alpha}+\hat{\lambda}) = Var\left(e^{-\frac{\hat{\mu}}{\hat{\sigma}}} + \frac{1}{\hat{\sigma}}\right) = Var(\hat{\alpha}) + Var(\hat{\lambda}) + 2Cov(\hat{\alpha},\hat{\lambda}) \tag{31}$$

Hence,

$$Cov(\hat{\alpha},\hat{\lambda}) = \frac{Var\left(e^{-\frac{\hat{\mu}}{\hat{\sigma}}+\frac{1}{\hat{\sigma}}}\right) - Var(\hat{\alpha}) - Var(\hat{\lambda})}{2} \tag{32}$$

By the same approach as (25), we get

$$Var\left(e^{-\frac{\hat{\mu}}{\hat{\sigma}}} + \frac{1}{\hat{\sigma}}\right) =$$

$$Var(\hat{\sigma}).\frac{1}{\hat{\sigma}^4} + e^{-\frac{2\hat{\mu}}{\hat{\sigma}}}\left[\frac{Var(\hat{\mu})}{\hat{\sigma}^2} + \frac{\hat{\mu}^2}{\hat{\sigma}^4}Var(\hat{\sigma}) - 2\frac{\hat{\mu}}{\hat{\sigma}^3}Cov(\hat{\mu},\hat{\sigma})\right] + 2e^{-\frac{\hat{\mu}}{\hat{\sigma}}}\left[\frac{Cov(\hat{\mu},\hat{\sigma})}{\hat{\sigma}^3} - \frac{\hat{\mu}Var(\hat{\sigma})}{\hat{\sigma}^4}\right] \tag{33}$$

Therefore,

$$Cov(\hat{\alpha},\hat{\lambda}) = e^{-\frac{\hat{\mu}}{\hat{\sigma}}}\left[\frac{Cov(\hat{\mu},\hat{\sigma})}{\hat{\sigma}^3} - \frac{\hat{\mu}Var(\hat{\sigma})}{\hat{\sigma}^4}\right] \tag{34}$$

From (25), (30), and (34), the Variance-Covariance matrix for the Weibull distribution is:

$$\begin{pmatrix} Var(\hat{\alpha}) & Cov(\hat{\alpha},\hat{\lambda}) \\ Cov(\hat{\alpha},\hat{\lambda}) & Var(\hat{\lambda}) \end{pmatrix}$$

$$= \begin{pmatrix} Var(\hat{\sigma}).\frac{1}{\hat{\sigma}^4} & e^{-\frac{\hat{\mu}}{\hat{\sigma}}}\left[\frac{Cov(\hat{\mu},\hat{\sigma})}{\hat{\sigma}^3} - \frac{\hat{\mu}Var(\hat{\sigma})}{\hat{\sigma}^4}\right] \\ e^{-\frac{\hat{\mu}}{\hat{\sigma}}}\left[\frac{Cov(\hat{\mu},\hat{\sigma})}{\hat{\sigma}^3} - \frac{\hat{\mu}Var(\hat{\sigma})}{\hat{\sigma}^4}\right] & e^{-\frac{2\hat{\mu}}{\hat{\sigma}}}\left[\frac{Var(\hat{\mu})}{\hat{\sigma}^2} + \frac{\hat{\mu}^2}{\hat{\sigma}^4}Var(\hat{\sigma}) - 2\frac{\hat{\mu}}{\hat{\sigma}^3}Cov(\hat{\mu},\hat{\sigma})\right] \end{pmatrix} \tag{35}$$

The parameters and their variations can be generated from (8), (11), (19), and (25). We rely on SAS to do the calculation.

SAS output for parameters covariance matrix is presented in Appendix D.

The survival function is given in (23). This model is conducted from the survival times of all patients, regardless of any factor. This model cannot be applied to predict survival time for new patients because it treats everyone the same. Obviously, the survival time for an 18-year-old girl who is at the first stage of cancer cannot be the same as a 40-year-old woman who is at the last stage. Therefore, using the model in (23) to make predictions in survival time can be misleading. The associated covariate factors of the breast cancer patients should have an impact on their survival time. So those variables should be incorporated into the model. The more covariates existing in the model, the more robust our model is and the less error we will experience in our predictions.

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

A model with covariates

The technique used to incorporate covariates into our survival model is the Accelerated Failure Time model (AFT).

The AFT model assumes the following relationship between two individuals i and j:

$$S_i(t) = S_j(c_{i.j}t) \tag{36}$$

where $c_{i,j}$ is a constant that is specific to the pair of individuals (i, j) (Allison)

A special case of this model is:

$$\log(T_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \sigma\varepsilon_i \tag{37}$$

$$T_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\sigma\varepsilon_i} \tag{38}$$

where $\beta_0, \beta_1, \ldots, \beta_k$, and σ are the parameters, $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the values of covariates, ε is a random variable that is independently and identically distributed. In our case, T_i follows the Weibull distribution, so ε follows the standard extreme value distribution (Allison).

The survival function of T at covariate values $(x_{i1}, x_{i2}, \ldots, x_{ik})$ is:

$$S(t|(x_{i1}, x_{i2}, \ldots, x_{ik})) = e^{-[te^{-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik})}]^{\frac{1}{\sigma}}} \tag{39}$$

(Zhang).

SAS gives us the estimates for the parameters in this model. The output is presented in Appendix D

The specific model in our case is:

$$S\left(t\middle|(x_1, x_2, x_3, x_4, x_5, x_6, x_7)\right) = e^{-\left[te^{-(5.1548-0.003x_1-1.6264x_2-3.0266x_3-4.3651x_4+0.9821x_5-0.0603x_6+0.1885x_7)}\right]^{0.8819}} \quad (40)$$

where x_1 = AGE_DX, x_2 = z1, x_3 = z2, x_4 = z3, x_5 = SURGERY, x_6 = RADIATION, x_7 = BOTH

Evaluating model fit

The use of the obtained statistical model to make a conclusion and prediction is based on the verification that the model fits the data well. To check the validity, goodness of fit, and the assumptions of the proposed distributional survival model, we perform the test suggested by Paul Allison. We observe that the plot of $\log\left[-log\left(\hat{S}(t)\right)\right]$ versus log(t) is a straight line indicating that the assumptions of Weibull model is good fit to estimate the survival times of the African American breast cancer patients. The plot is presented in Appendix D.

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

DISCUSSION

The result of variance – covariance matrix in (35) is presented in Appendix D. This matrix shows us how much the estimates of α and λ can vary. The values in this matrix are very small. This means our model is very accurate.

The model (23) does not provide us much information as discussed above. It estimates the general patients' survival time. From this model, most questions about the general African American breast cancer patients can be answered. For example, the probability that a random African American breast cancer patient can survive 10 years after diagnosis is S(10) = 71.98%. The probability that a patient will die at the 40^{th} month given that she has survived up to that time is h(40/12) = 86.51% (hazard function) .

The Accelerated Failure time model (40) gives us a lot of important information. All the covariates in this model are significant except for the RADIATION. This suggests that

1. Age, stage of cancer, treatment by only surgery, treatment by both surgery and radiation all affect patients' survival time.
2. Age contributes negatively to survival time and coefficients of stages are more negative when stages increase. This fits reality because younger patients can obviously resist illnesses better and a higher stage means shorter time of survival.
3. There is insufficient evidence to claim that radiation-only treatment can improve patients' survival time.

In contradiction to the model (23), model (40) helps us predict the survival time of a specific patient rather than the general population. Given a patient's specific information, we can generate a new survival function for that patient, which looks similar to model (23).

For example, a 40-year-old woman who was diagnosed at the second stage, and was treated with only surgery has survival function:

$$S(t) = e^{-0.0716t^{0.8819}} \tag{41}$$

Further prediction about this patient can be inferred easily from this model.

CONCLUSION AND RECOMMENDATIONS

Both model (23) and (40) are useful in reality. Model (23) is a survival function that helps provide a general understanding about the survival time of all African American women who are diagnosed with breast cancer. It can be used effectively in conferences and meetings about breast cancer.

Model (40) has a wider application. As discussed above, it helps predict the survivability of African American breast cancer patients. This model can be used in consulting activity. Clinics need information from this model to determine the best treatments for their patients and to provide patients with their predicted survival status. Survival distribution can also be useful information for the insurance industry. The model can certainly be adjusted easily in particular situations. For example, in the case that clinics want more information about other types of treatment, more covariates can be added to the model by the same process. If information about other races is needed, the whole data from SEER will be taken, and race becomes a new covariate in our model.

Model (40) also provides us with many interesting findings that require more investigations. The confidence interval for the coefficient for this variable is (-0.1694, 0.0489)

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

which lies mostly on the negative side. So it is suspicious that using only radiation in fact shortens patients' survival time. Two questions are raised here:

1. Should the radiation-only treatment be used for African American breast patients?
2. Do other races have the same problem with using only radiation?

The second question can be answered by using the same process in other races with the data from SEER. The second question suggests more research in the medical field.

NOMENCLATURE

Symbol	Description	Unit
t	Survival time	year
f(t)	Probability density function	N/A
S(t)	Survival function	N/A
h(t)	Hazard function	N/A
$\hat{X}$	Estimation of X, with X being α, λ, σ, etc.	N/A
α	Shape parameter of the Weibull distribution	N/A
λ	Scale parameter of the Weibull distribution	N/A

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

REFERENCES

Allison, Paul D. Survival Analysis Using the SAS system: A Practical Guide. SAS Institute, 1995.

American Cancer Society. Breast Cancer. n.d. 11 December 2013 <http://www.cancer.org/cancer/breastcancer/>.

Balakrishnan, N., and M. Kateri. "On the Maximum Likelihood Estimation of Parameters of Weibull Distribution Based on Complete and Censored Data." Statistics and Probability Letters. Vol. 78.17. National Cancer Institute, 2008. 2971-2975.

Bradley, Cathy J., Charles W. Given and Caralee Roberts. "Race, Socioeconomic Status, and Breast Cancer Treatment and Survival." Journal of the National Cancer Institute 94.7 (2002): 490-496.

Encylopedia Britannica. "Cauchy-Schwarz Inequality." 2013. Encylopedia Britannica Online:Academic Edition. 28 November 2013 <http://britannica.com/EBchecked.topic/1365910/Cauchy-schwarz-inequality>.

Evans, Addie A. "San Francisco State University." 2008. Department of Biology. <http://userwww.sfsu.edu/efc/classes/biol710/MLE/Maximum-Likelihood.pdf>.

Klein, John P., and Melvin L. Moeschberger. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer, 1997.

"National Cancer Institute." April 2013. Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1973-2010). <http://www.seer.cancer.gov>.

"National Cancer Institute." n.d. Understanding Cancer Prognosis. 2013 November 23 <http://www.cancer.gov/cancertopics/diagnosis-staging/prognosis>.

Papanicolaou, Alex. "Taylor Approximation and the Delta Method." n.d. 27 October 2013 <http://web.stanford.edu/class/cme308/OldWebsite/notes/TaylorAppDeltaMethod.pdf>.

"Sisters Network." n.d. Breast Cancer Facts. 2013.

Zhang, Daowen. "Daowen Zhang's Homepage." n.d. Modeling Survival Data with Parametric Regression Models. <http://www4.stat.ncsu.edu/~dzhang2/st745/chap5.pdf>.

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

APPENDICES

Appendix A – data

Since the sample size is more than 44000, it is inconvenient to show the whole data set. Here is a part of it:

1	SRV_TIME_YEAR	DTH_CLASS	AGE_DX	HST_STGA	SURGERY	RADIATION	BOTH	RAC_RECY
44525	0.333333333	0	43	4	1	1	1	2
44526	0.25	1	55	4	0	0	0	2
44527	0.583333333	0	34	4	0	0	0	2
44528	0.416666667	0	53	0	1	1	1	2
44529	0.25	0	78	1	1	0	0	2
44530	0.416666667	0	52	0	0	0	0	2
44531	3.333333333	1	72	4	0	0	0	2
44532	0.916666667	0	51	2	1	0	0	2
44533	0.583333333	0	45	1	1	1	1	2
44534	0.833333333	1	44	4	0	1	1	2
44535	0.666666667	0	75	1	0	0	0	2
44536	0.416666667	0	55	1	1	1	1	2
44537	0.916666667	0	46	4	0	0	0	2
44538	0.083333333	0	41	2	1	1	1	2
44539	3.916666667	0	74	1	1	1	1	2
44540	0.666666667	0	45	1	0	0	0	2
44541	7.333333333	1	59	2	1	0	0	2
44542	5.166666667	1	27	2	0	0	0	2
44543	0.083333333	0	56	1	1	0	0	2
44544	6.916666667	0	30	0	1	0	0	2
44545	10.91666667	0	59	1	1	0	0	2
44546	0.5	0	60	2	0	0	0	2
44547	0.916666667	0	80	4	0	0	0	2
44548	0.25	0	57	4	0	1	0	2

Variables description by SEER:

SRV_TIME_YEAR: survival time in year. In the original dataset from SEER, this survival time was given in month (SRV_TIME_MON), which makes this variable discrete.

However, survival time should be continuous so that the whole analyzing process can be applied. Therefore, the new variable was created: SRV_TIME_YEAR = SRV_TIME_MON/12. (85)

DTH_CLASS: variable that indicates censoring status (77).

=1: patient died because of breast cancer (event)

= 0: patient is still alive or dead because of other causes (censor)

AGE_DX: age of patient at diagnosis for breast cancer (15).

HST_STGA: a simplified version of cancer's stage (0 = in situ, 1 = localized, 2 = regional, 4 = distant) (70).

HST_STGA is replaced in the AFT model by z1, z2, z3 (z1=1 when HST_STGA=0, z2=1 when HST_STGA=1, z3=1 when HST_STGA=2, otherwise z's=0).

SURGERY: indicates if patient took surgery (1 = yes, 0 = no), created from NO_SURG variable (47).

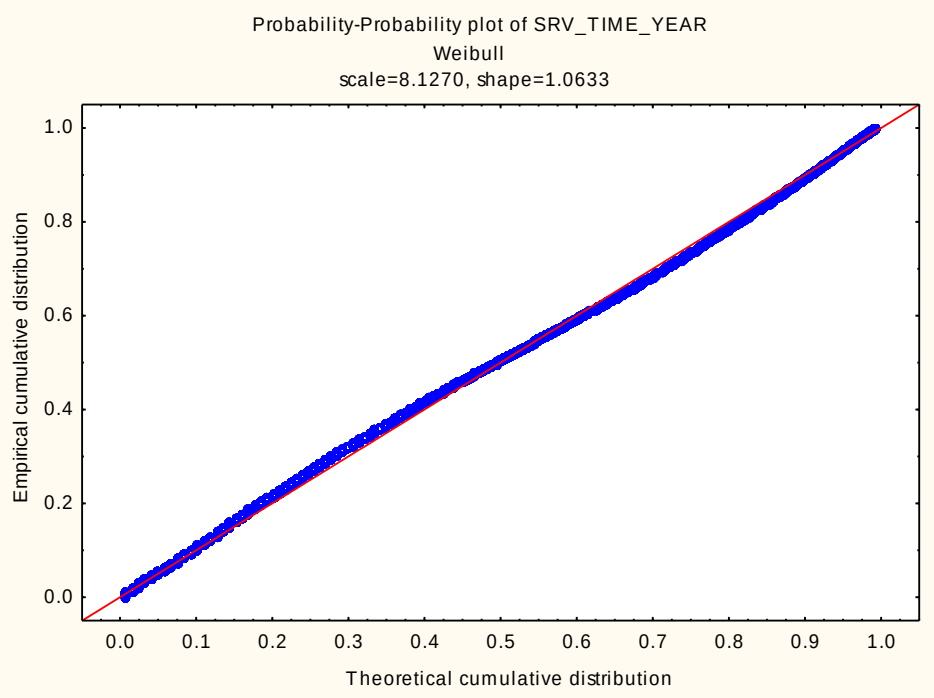
RADIATION: indicates if patient took radiation (1 = yes, 0 = no), created from RADIATN variable (47)

BOTH: indicates if patient took both surgery and radiation (1 = yes, 0 = no), created from RAD_SURG variable (49).

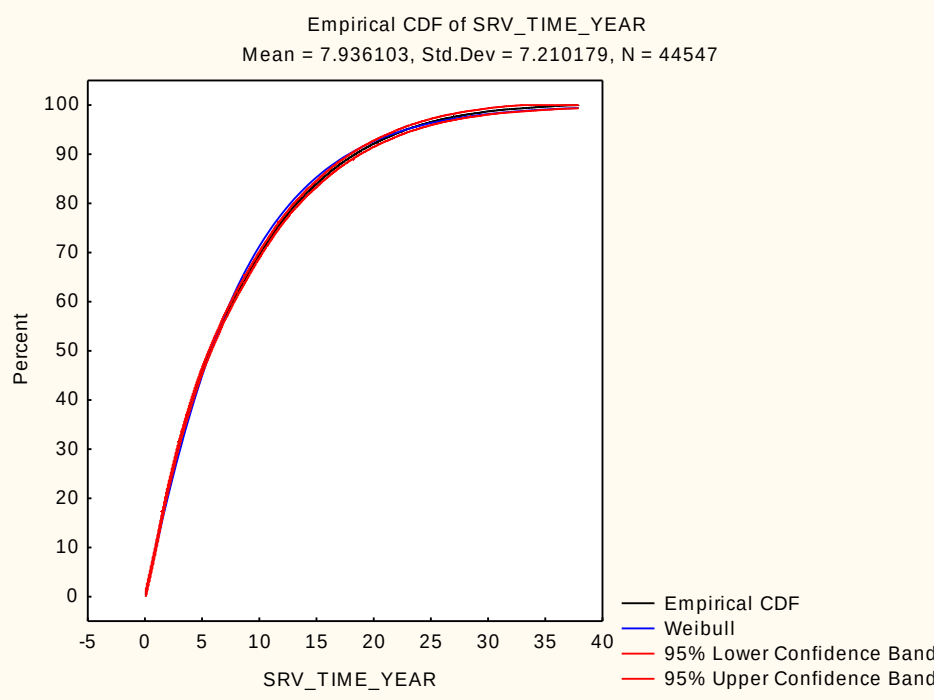
https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

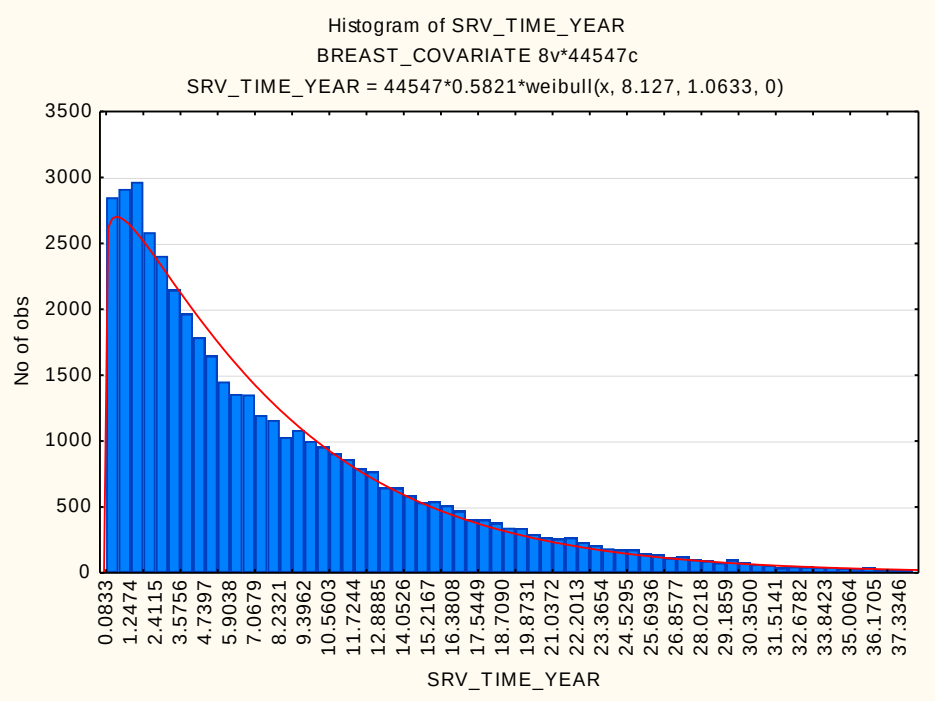
Appendix B – Result from STATISTICA:

Probability – Probability plot:

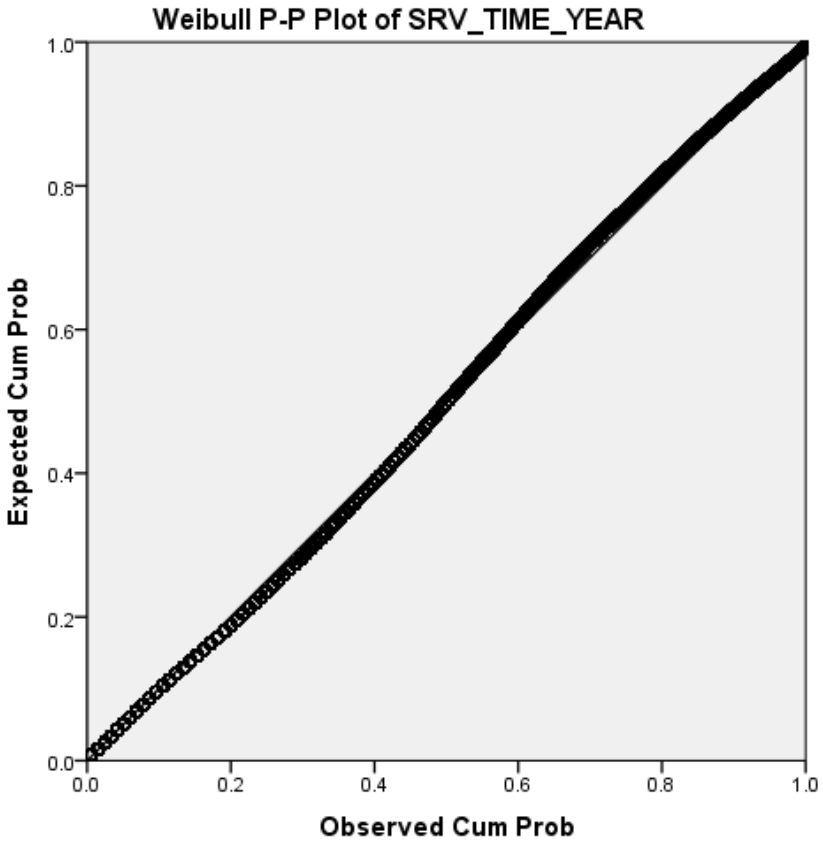


Empirical Cumulative Distribution Function Plot:





Appendix C: SPSS probability plot



DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

Appendix D: SAS outputs

Parameters estimation – no covariates:

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.7928	0.0173	3.7589	3.8267	48058.3	<.0001
Scale	1	1.3405	0.0107	1.3198	1.3616		
Weibull Scale	1	44.3795	0.7678	42.8999	45.9102		
Weibull Shape	1	0.7460	0.0059	0.7344	0.7577		

The Weibull probability distribution function used in SAS is different from our function. The pdf used in SAS is:

$$f(t) = e^{-\left(\frac{t}{\gamma}\right)^{\alpha}} . \frac{\alpha}{\gamma} . \left(\frac{t}{\gamma}\right)^{\alpha-1}$$

where α is shape parameter and γ is scale parameter.

Our $\alpha = \alpha_{SAS} = 0.7460$, $\lambda = \left(\frac{1}{\gamma}\right)^{\alpha} = 0.059$

Variance – Covariance matrix for Weibull model:

Variance-Covariance Matrix	
covb	
9.5352E-7	-4.779E-6
-4.779E-6	0.0000352

Parameters estimation – Accelerated failure time model:

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.1548	0.0929	4.9727	5.3370	3077.27	<.0001
AGE_DX	1	-0.0030	0.0006	-0.0041	-0.0018	26.73	<.0001
z1	1	-1.6264	0.0783	-1.7800	-1.4729	430.97	<.0001
z2	1	-3.0266	0.0789	-3.1813	-2.8719	1470.68	<.0001
z3	1	-4.3651	0.0841	-4.5300	-4.2001	2691.14	<.0001
SURGERY	1	0.9821	0.0401	0.9036	1.0606	601.20	<.0001
RADIATION	1	-0.0603	0.0557	-0.1694	0.0489	1.17	0.2790
BOTH	1	0.1885	0.0603	0.0704	0.3067	9.78	0.0018
Scale	1	1.1339	0.0085	1.1173	1.1508		
Weibull Shape	1	0.8819	0.0066	0.8690	0.8950		

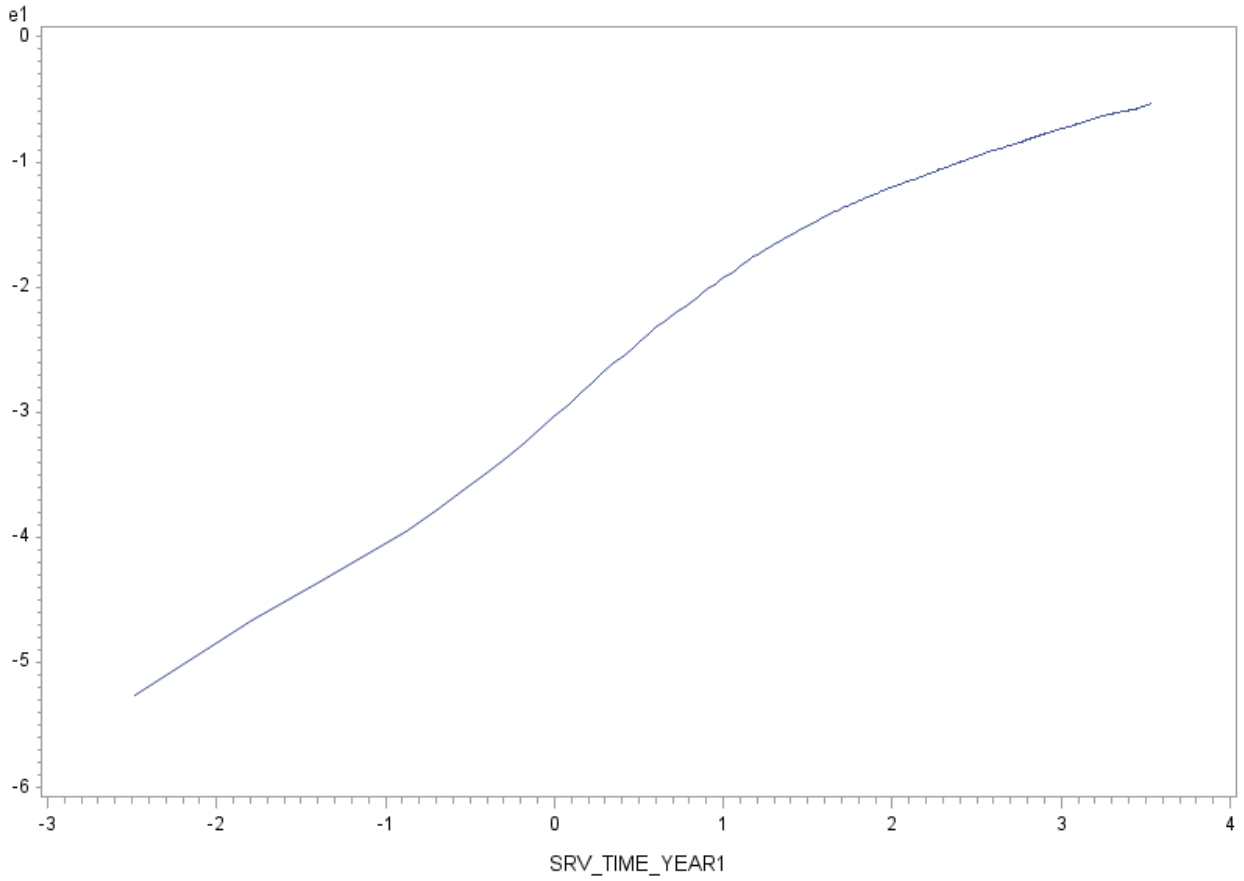
The “Estimate” column represents coefficients for the covariates in model (14).

The “Standard Error” and “95% confidence limits” show variations for these coefficients.

The “chi-square’ and “Pr>Chisq” columns represent chi-square test-statistics and p-value for the null hypothesis that the coefficient is 0.

https://digitalcommons.usf.edu/ujmm/vol6/iss1/4
DOI: http://dx.doi.org/10.5038/2326-3652.6.1.4860

Model Validation graph:



$e1 = \log\left[-log\left(\hat{S}(t)\right)\right]$

SRV_TIME_YEAR1 = log(t)