PREDICTION OF SURVIVAL TIME FOR BREAST CANCER PATIENTS

by

Jie Yu

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Statistics, Honours

at

Dalhousie University
Halifax, Nova Scotia
April 2022

Table of Contents

List of Tables

List of Figures

Abstract

The prevalence of breast cancer has made studies to predict survival time more familiar. Not only linear regression and machine learning but also survival analysis have produced a significant impact in the analysis of breast cancer clinical data. This thesis introduces fundamental concepts about linear regression, random forests in machine learning, and survival analysis. This thesis aims to predict the survival time based on linear regression and random forest and predict the expected survival time through parametric survival regression. We then analyze the reasons for using the parametric survival regression model, AFT, instead of the prevalent semi-parametric COX model. We choose two types of AFT models, AFT with log-logistic and AFT with Weibull, as candidates for survival analysis. We applied these three models to breast cancer survival data to accomplish our goals. In the final results, we conclude that the regression model is more accurate in predicting survival times in this breast cancer dataset containing only observable death event data compared to the random forest. Because the residuals of this data set are log-logistic distribution, the AFT model with log-logistic distribution is more precise in predicting the expected survival time than the AFT model with Weibull distribution. We show a drawback of the AFT model regarding the distribution of the residuals and point out the potential way to improve the AFT model to make it more accurate in predicting average survival time.

Acknowledgements

This thesis came about as a result of a research topic that interested me, proposed by my honor advisor, Dr. Lam Ho. I am deeply grateful to Dr. Ho, along with his help, I began to understand more about the application of survival analysis to clinical data. During this time, he not only guided me in learning about survival analysis and machine learning, but also gave me valuable advice and patience when I encountered blockages in my research ideas. In addition, I would like to thank him for reading my thesis and making valuable suggestions, and it is because of his continuous support that I was able to complete my thesis. Last but not least, I would like to thank my family and friends for their support!

Chapter 1

Introduction

Breast cancer is the most frequently diagnosed cancer in 154 countries and is the leading cause of cancer death in over 100 countries [1]. From the American Cancer Society, approximately 1 in 8 women (13%) will be diagnosed with invasive breast cancer in their lifetime, and 1 in 39 women (3%) will die from breast cancer [9].

The prevalence of breast cancer has led to many studies examining breast cancer factors and predicting patient survival time more generally. Survival analysis can focus on the time to the occurrence of an event (death), and an increasing number of studies are using survival analysis to address more areas where mortality can be applied. Along with survival analysis characteristic that allows studying the expected duration of an event of interest, survival analysis is a valid statistical method to analyze breast cancer patients' mortality and survival time.

However, the general case in survival studies is that, at the end of follow-up, some individuals have not had the event of interest, and thus their true time to event is unknown [2]. Reasons behind this may include that the patients did not experience the event during the study time and potentially withdrew from the experiment during the study time. This is called *right* censoring. In general, the feature of censoring means that special methods of analysis are needed, and standard graphical methods of data exploration and presentation, notably scatter diagrams, cannot be used [2].

The standard methods for modelling the right-censored survival data are semi-parametric and parametric survival regression. The popularity of the Cox proportional hazards (PH) model as a semi-parametric is due to fewer assumptions on parameters. The PH model assumes that the underlying hazard rate is a function of the independent covariates, but no assumptions are made about the nature or shape of the hazard function [6]. Because of the simplicity, the PH model cannot have a direct interpretation of estimates in terms of the survival time for a subject but is helpful for testing and hazard ratio estimation [7]. Compared to the PH model, the parametric proportional model requires assumptions about the unknown parameters, such as the shape of the hazard function, which makes it a sufficient way to analyze the average time to the occurrence of an event.

Both linear regression and random forest techniques can evaluate survival time predictions on exact event time data. The linear regression builds the linear relationship between the response variable (survival time) and dependent variables that potentially affect the model, and the random forest build criterion to split the predictor variables for the "tree growth." The difference between the two methods is that the flexibility of random forest makes it more accurate in predicting results in non-linear data, while the linearity assumption of linear regression makes the prediction less precise if the assumption of linearity is violated [8].

This paper will divide the breast cancer data into data with the deceased event and right-censored data to predict survival time. This paper uses the conventional method, linear regression and random forest, to predict the expected-to-event time for the uncensored data. Since the parametric proportional model can predict the survival time by expectation values, the right-censored data are analyzed using the parametric proportional model with different distributions of hazard functions to predict survival time for breast cancer patients. Then, the paper compares the estimated results modelling from training data sets with the actual results from validation sets. The results will be discussed to indicate which method is more effective for the uncensored data and right-censored used in this paper. The mathematical formulas and terms used are explained in the methods section, and the relevant results analyzed in R will be presented in the results section.

Chapter 2

Methods

2.1 Dataset

The data set contains 2509 observations and 36 variables, downloaded from the cBio Cancer Genomics Portal `https://www.cbioportal.org/`. According to relevant background information on breast cancer, 24 variables are selected to estimate the expected lifetime for patients (explained in Appendix). First, the dataset is randomly split into training and validation data sets: 80% of the data in the training set is used for building models, and the rest of the data is used for the models' performances. Based on each model used, the training and testing datasets are also partitioned into datasets that only include deceased events or not: right censoring is applied to datasets that contain deceased events, and uncensored data is formed by deleting data on patients who have not experienced a death event at the end of the recording.

2.2 Software

All data is manipulated and analyzed in R, and the functions used to analyze the data are derived from existing packages: tidyr, mice, randomForest, reprtree, survival, survminer, flexsurv, ggplot2, Hmisc, and ciTools.

2.3 Missing data Type

Missing data are typical in clinical research but may affect the accuracy of the models. Reasons behind missing data include 'missing completely at random (MCAR), 'missing at random (MAR), and 'missing not at random (MNAR). MCAR means that the missing values of participants are independent of the other information in the data. For example, participants with factor A have the same probability of missing data as those with factor B. MCAR may include improper data collection or improperly processing the data. In these instances, the missing data reduce the analyzable population of the study and, consequently, the statistical power, but do not introduce bias [4]. MAR is another assumption of missing data, and this assumption is often used for statistical purposes. The missing data under

the MAR assumption are related to the observed data but not the unobserved data. In this way, statistical methods such as the expectation-maximization (EM) algorithm and multiple imputation (MI) can be applied to the unobserved data prediction predicted by observed data. The MNAR is the most complicated case for data analysis; it means the missing values are related to both the unobserved data and observed data, indicating that if the missing data is missing not at random, any prediction based on observed data will introduce the bias in the analysis. It is not easy to know what happened during data collection or data transmission; we usually assume that the data is MCAR or MAR for statistical analysis purposes.

This study uses the MI method to yield unbiased results for unobserved data, dependent on assumptions that the missing variables of this research are MAR or MCAR. The general steps for the MI method are that the missing values are imputed multiple times using the selected model and yield multiply imputed datasets. Each of these datasets is analyzed through the models the data selected (i.e. linear regression), and the results from the selected model are then pooled (averaged) into a final study result. This study uses the MI method with classification and regression trees (CART) method, developed by Burgette and Reiter in 2010, Shah et al. in 2014 and Doove, Van Buuren, and Dusseldorp in 2014, whose utilization is also proven to produce more efficient estimates than standard procedures [10]. The CART method is implemented in the second step of MI, which replaces the missing values with imputed values predicted by the CART method. The advantages of the CART method include that they are robust against outliers, can deal with multicollinearity and skewed distributions, and are flexible enough to fit interactions and nonlinear relations [10]. It is helpful when the data set has both continuous and categorical variables.

2.4 Linear Regression

One of the methods used in this analysis for the data with the deceased event is the multiple linear regression model, a primary tool for predicting the response variable. A model is hypothesized as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_p x_{ip} + \varepsilon_i, for\ i = 1, \ldots, n$$

where the model assumes $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. The matrix form is:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{X}$ is the n $\times$ (p + 1) design matrix, containing a $\mathbf{1}$ column vector and covariates columns. The goal of the general linear model is to predict the survival time for breast cancer

patients through the fitted linear model built by training dataset without living cases, which has survival months as the response variable and remaining variables as the regressors. The Gauss Markov assumptions can help validate the estimation of regression coefficients: these assumptions are the zero of the expected value of the error term, collinearity, exogeneity, and homoscedasticity. The assumptions can be tested by residual versus fitted plot, normal Q-Q plot, scale-location plot, and residual versus leverage plot in R, respectively.

2.5 Random Forest Model

Random Forest, developed by Leo Breiman and Adele Cutler, is a machine learning algorithm for classification and regression problems. The decision trees that help build random forest starts with the prediction of survival length, and then the decision trees are built by different variables, which become the decision nodes in the tree to split the training data set shown by Fig.2.1

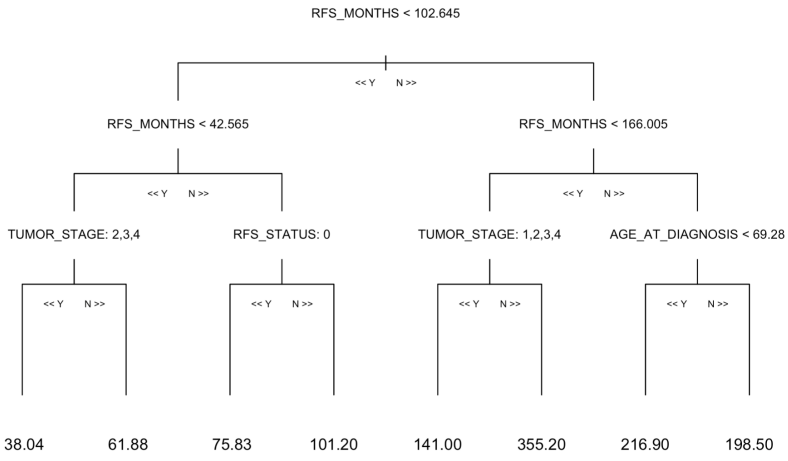

Figure 2.1: Part of a Tree Extracted From a Random Forest Model

Breiman extended the random-subspace method and formally proposed it as a random forest in 2001. The random forest model is a tree-based ensemble learning algorithm; the algorithm averages predictions over many individual trees. The individual trees are built on bootstrap samples rather than the original ones. This is called bootstrap aggregating or simple bagging, and it reduces over-fitting. The algorithm is as Fig.2.2. Each tree comprises data samples with replacements taken from the training set, called bootstrap samples. About one-third of the observations in the bootstrap sample is randomly left as

```
for i ← 1 to B do
    Draw a bootstrap sample of size N from the training data;
    while node size != minimum node size do
        randomly select a subset of m predictor variables from total p;
        for j ← 1 to m do
            if jth predictor optimizes splitting criterion then
                split internal node into two child nodes;
                break;
            end
        end
    end
end
return the ensemble tree of all B subtrees generated in the outer for loop;
```

Figure 2.2: Random forest algorithm [8]

test data for a given tree, which is called an out-of-bag (OOB) sample. Finding parameters that would produce a low OOB error is often a key consideration in model selection and parameter tuning. M, the number of variables randomly sampled as candidates at each split, is a parameter that needs to be adjusted during the model selection process and is critical to the final depth of the control tree [8]. M has a significant impact on the random forest results, but the OOB error rate can adjust M to enhance the accuracy of the random forest by adjusting the appropriate M range value. If M is too small, it will reduce the correlation between variables and the strength of the random forest and vice versa. The large M number will overestimate the results from the random forest algorithm, so it is crucial to find the optimal range of M values through OOB error.

2.6 Survival regression

The common modelling techniques used in the survival analysis are semi-parametric and parametric survival regression. These two methods help deal with right-censored data, and handling censored observations is not feasible with linear regression and ordinary random forests.

2.6.1 Cox proportional hazards model (cox model)

The Cox Proportional Hazards model, the most well-known semi-parametric survival regression proposed by Cox, does not have assumptions on the shape of the hazard function,

written as:

$$h(t|\mathbf{X}) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) = h_0(t)exp(\beta^T \mathbf{X})$$

where $h_0(t)$ is the baseline hazard function, $X = (x_1, x_2, \ldots, x_p)'$ is the vector of covariates for the individual p, and $\beta^T = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients. The time t indicates the hazard is dependent with time, the values of β_i negative correlates to the survival time (the positive β_i means the increases in hazard and then the declines in the survival length), and the hazard ratio of two individuals is estimated by:

$$\widehat{HR} = \frac{h_0(t)exp(\hat{\beta}'\mathbf{X})}{h_0(t)exp(\hat{\beta}'\mathbf{X}^*)}$$

indicating that it works well when evaluating the effects of factors on the rate of events: 1 means no effect, less than 1 means the decreases, and large than 1 means the increases. The primary assumption of the Cox, proportional hazards model, is the hazard ratio is constant over time [6]. Three tests need to be done to satisfy the assumption, and all of them can be tested in R: (1) testing the proportional hazards assumption function by commend cox.zph() (2) examining influential observations (3) detecting nonlinearity (checked by martingale residuals against continuous variables).The main assumption is violated if the true β_t is not a horizontal line, if the pattern of outliers looks non-symmetric around 0, and if there is no linear relationship for the continuous explanatory variable against the martingale residuals of the null cox model. Since the cox model does not have the assumption on their survival distribution, it is useful when analyzing the factors in clinical data. However, the violation of assumptions for the cox model will cause inaccurate conclusions on the factor influences.

2.6.2 Accelerated failure time model (AFT model)

The other way to analyze the right-censored data is Accelerated failure time model (AFT model) when the proportional hazards assumption of Cox model fails. Unlike the cox model, the AFT model allows the evaluation of the statistically significant covariates on the survival time of patients. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariates on the mean survival time [6]. In general, the AFT model can be specified as:

$$h(t|\mathbf{X}) = exp(-[\beta_1 x_1 + \ldots + \beta_p x_p])h_0\left(exp(-[\beta_1 x_1 + \ldots + \beta_p x_p]) \times t\right) = \frac{1}{\eta(x)} h_0\left(\frac{t}{\eta(x)}\right)$$

with the acceleration factor $\eta(x) = exp([\beta_1 x_1 + \ldots + \beta_p x_p])$, and the corresponding log-transformed response variable Y_i written in linear form as [6]:

$$Y_i = log(T_i) = a + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \sigma\epsilon_i$$

where T_i is a random variable denoting the survival time, a is the intercept, σ is the scale parameter, and ϵ_i is a random variable with a particular distribution. We can assume different distributions for the disturbance term $\epsilon_i$in linear form model. For example, if the assumption of $\epsilon_1$is that it is I.I.d and follows normal distribution with $\mu = 0$ and $\sigma = 1$, then the T_i has log-normal distribution conditioning on the covariates X [11]. The distributions used in this paper include Weibull and log-logistic distributions, and the relationship between ϵ_i and T_i shown as Table 2.1

Distribution of ϵ	Distribution of T	R in survreg function
extreme values (2 parameters)	Weibull	dist = weibull
logistic	log-logistic	dist = loglogistic

Table 2.1: Distribution of AFT models

Because the assumptions on survival time distribution for T_i, AFT model can obtain expected survival time estimates for a specific time point, $E(T|X)$. For Weibull distribution, $\epsilon \sim$ SEV with $F_{SEV}(\epsilon) = 1 - exp(-exp(\epsilon))$ and T is Weibull distribution with scale parameter σ and location parameter $exp(x\beta)$ so the expected survival time is [3]:

$$E(T|X) = exp(X\beta)\Gamma(1+\sigma)$$

and the survival function equals to[13]:

$$S(T|X) = exp(-exp(\frac{log(T) - X\beta}{\sigma}))$$

For log-logistic distribution, $\epsilon \sim Logistic$ with standard logistic distribution $F(\epsilon) = \frac{exp(\epsilon)}{1+exp(\epsilon)}$, and the T is log-logistic distribution with scale parameter σ and location parameter $X\beta$ according to the equation $log(T) = X\beta + \sigma\epsilon$. The corresponding expected survival time is [3]:

$$E(T|X) = exp(X\beta)\Gamma(1+\sigma)\Gamma(1-\sigma)$$

where the survival function equals to [3]:

$$S(T|X) = 1 - F_{Logistic}(\frac{log(T) - X\beta}{\sigma})$$

2.7 Akaike information criterion (AIC)

Linear regression, semi-parametric, and parametric survival regression all use backward elimination to select statistically significant covariates via AIC developed by Akaike in 1973.

AIC is a statistical methods that can estimate the quality of models by comparing different possible parameters of models. Let k be the number of estimated parameters in the model. Let $\hat{l}$ be the estimates of the log-likelihood function for the model. Then the AIC value of the model is the following:

$$AIC \ = \ 2k \ - \ 2ln(\hat{l})$$

The second term in AIC is twice the negative log likelihood, which turns out to be the residual sum of squares corresponding to the model for the linear regression model with a Gaussian likelihood [5]. That is,

$$-2ln(\hat{L}) \ = \ \sum_{i \ = \ 1}^{n} \left(y_i \ - \ x_i^T \hat{\beta}\right)^2$$

where $\hat{\beta}$ is the least squares estimator for model. The first term acts as a penalty term to penalize models if more non-predictive parameters are added to the model [6], so a lower AIC score means that the model with fewer parameters can also explain the same amount of variation as the model with more parameters, the one with fewer parameters corresponding to the best fit of the model.

2.8 Concordance index (c-index)

The C-index, or concordance index introduced in Harrell et al. 1982, is used to evaluate the predictive performance of the survival model. Generally, c-index calculates the proportion of of all patient pairs in which the predicted results agrees with the actual results. Harrell's C-index can be seen as:

$$c = \frac{\#\ concordant\ pairs}{\#\ concordant\ pairs + \#\ discordant\ pairs}$$

or it can be expressed in the formula:

$$c = \frac{\sum_{i,j} \mathrm{I}(\tilde{T_i} > \tilde{T_j}) \cdot \mathrm{I}(\eta_j > \eta_i) \cdot \Delta_j}{\sum_{i,j} \mathrm{I}(\tilde{T_i} > \tilde{T_j})}$$

In this equation, the $\mathrm{T_i}, \mathrm{T_j}$, time to deceased event for i^{th}, j^{th} patient, is either the deceased time or the last time of recording. The η_i, η_j are the risk score, and Δ_j is the factor representing the censored data such that $\Delta_j = 0$ if the data are censored, $\Delta_j = 1$ if the data are not censored.

The steps of C-index calculation are: (1) randomly draw pairs with two patients from the data, (2) delete the pairs that both patients have censored data for the death event, (3) compare the η_i, η_j of uncensored data with the condition that if $\eta_i > \eta_j$ then $\mathrm{T}_i < \mathrm{T}_j$. The

pair(i, j) is concordant pair when fulfilling the previous conditions concordant pairs, and conversely pair(i, j) is discordant pair if they do not. (4) compare the η_i, η_j if one of them is censored data, because it is certain that the uncensored data patient i reached the death event first relative to the censored data patient j. For example, the η_j is censored, and η_j is not. The pair(i, j) is concordant pair if $\eta_i > \eta_j$ then $\mathrm{T}_i < \mathrm{T}_j$and conversely pair$(i, j)$ is discordant pair if they do not.

2.9 Mean squared error criterion (MSE)

Mean square error is the common methods used for testing the predictive performance of models of uncensored data. The MSE measures the performances of each fitted predictive training model, through calculating the mean squared differences between predictions of each model and observed survival time:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

In the linear regression and random forest performance tests, the estimates are predicted by the model fitted from the training set by the uncensored data. The actual values used for each prediction are selected data in the validation set based on the fact that the death event happened at the end of the observation.

Chapter 3

Results

3.1 Linear regression and random forest

After regression of all variables, the Variance Inflation Factor (VIF) is used to test whether there is multicollinearity between covariates, and the VIF is calculated by $\frac{1}{1-R_k^2}$, where R_k^2 is the proportion of the variance for a k^{th} dependent variable explained by model inputs and estimated on rest of the predictors in the model. If VIFs exceed 10, the result indicates collinearity problems between them that need to be adjusted. Fig 3.1 shows the VIF of variables exceeding 10. For simplicity, those variables are removed from the multiple linear models before the variables selection step. The variables chosen for linear model by

	Variables <chr>	Tolerance <dbl>	VIF <dbl>
37	GRADE3	0.09335645	10.71163
40	TUMOR_STAGE1	0.03668915	27.25601
41	TUMOR_STAGE2	0.02867844	34.86940
42	TUMOR_STAGE3	0.06225931	16.06185

Figure 3.1: Part of VIF table

the backward and forward step-wise selection are NPI , ER IHC, AGE AT DIAGNOSIS, RFS STATUS, RFS MONTHS, PR STATUS, and TMB NONSYNONYMOUS. Then, the plot() function in R is used to test the Gauss-Markov assumptions.

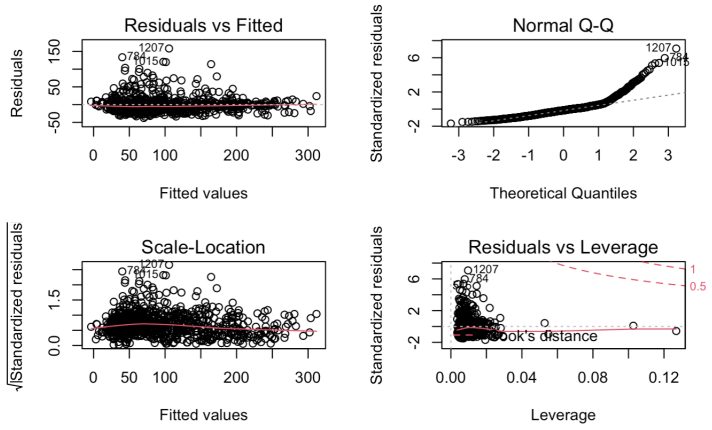

Figure 3.2: Gauss-Markov assumption test

From Fig 3.2, the plot of residuals against fitted values shows that residuals are evenly distributed, representing that the error is randomly generated. The output of normal q-q plot shows the residuals do not essentially follow a normal distribution, as they leave the straight line at the tail of the figure. Most residuals gather round 0 in the plot of externally studentized residuals against fitted values. Also, there are only a few influential points in the leverage plot against residuals.

The M is decided by the R function tuneRF() for the random forest model, which returns the optimal value, 15, for M based on OBB error. The random forest model is then built by M = 15 with all variables in the training data set of uncensored data. MSEs of predicted survival time and actual survival time were obtained from a validation dataset that included only previously occurred death events to compare the linear and random forest models' performance. The results with respect to different models are shown in Table 3.1, indicating that the linear regression model has a smaller MSE compared to the random forest model.

Models for uncensored data	MSE
Linear regression	1062.165
Random forest	1103.366

Table 3.1: MSE of Linear regression and random forest

3.2 Parametric survival regression

Both AFT models apply backward and forward step-wise selection of covariates in terms of AIC, and the final covariates selected for the model in R are shown by Fig 3.3: Then,

```
wei.fit <- survreg(formula = Surv(OS_MONTHS, as.numeric(OS_STATUS)) ~
                     RFS_MONTHS + AGE_AT_DIAGNOSIS + TUMOR_STAGE +
                     ER_STATUS + TMB_NONSYNONYMOUS +
                     GRADE + LATERALITY + PR_STATUS + NPI +
                     HER2_STATUS + BREAST_SURGERY,
                   data = train, dist = "weibull")

llog.fit <- survreg(formula = Surv(OS_MONTHS, as.numeric(OS_STATUS)) ~
                      RFS_MONTHS + AGE_AT_DIAGNOSIS + ER_IHC +
                      TUMOR_STAGE + PR_STATUS + HORMONE_THERAPY +
                      NPI + RFS_STATUS + LATERALITY + BREAST_SURGERY,
                    data = train, dist = "loglogistic")
```

Figure 3.3: Variables selected for the final AFT model

the final AFT model is computed using distribution formulas of Weibull and log-logistic to predict the expected survival time of the validation dataset for the death events that have occurred. The confidence interval for expected survival time is also calculated in the

R function add_ci from the package ciTools. Fig 3.4 shows the results and indicates the interval of mean survival time for AFT with log-logistic distribution is higher than that for AFT with Weibull distribution within $\alpha = 0.005$. The specific values of MSE are shown by

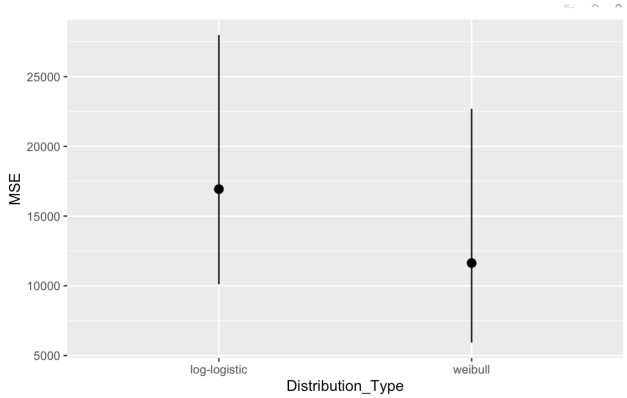

Figure 3.4: Confidence interval for mean survival time

Models for censored data	MSE
AFT with Weibull	11632.06
AFT with log-logistic	16932.09

Table 3.2: MSE of AFT models

Table 3.2, which indicates AFT with Weibull distribution has the smaller MSE, 11632.09. However, the results of the MSE do not fully explain the performance precision of the AFT model. For example, compared to Table 3.1, Table 3.2 shows the MSEs of AFT models are ten times bigger than linear regression and the random forest model, indicating the inaccuracy of the models, contradictory to the C-index performance. One reason is that the MSE for AFT models calculates the differences between the conditional expected values of different distributions with respect to $T|X$ and actual response variables. The other reason is that the MSE only predicts the events that can be observed, and an alternative method of testing model performance is explained later for the survival model.

Unlike the measurement of MSE values for each fitted training model, the C-index can measure the goodness of fit for binary outcomes: the C-index can assess fitted models with all testing data, including both living and deceased events. As previously explained in the methods section, the C-index counts concordant pairs, evaluating predicted and actual survival times. The C-index is more effective in predicting censored data than MSE, which only predicts observable events.

The results of C-index for different AFT models are shown in the Table 3.3. From

Distribution of T_i	C-index
log-logistic	0.91169
Weibull	0.89355

Table 3.3: C-index of AFT models

the calculation of C-index, it can be seen that the result of C-index will be between 0.5 and 1. If the result of C-index is 0.5, it indicates that there are as many concordant pairs as discordant pairs, which indicates that the predicted results of the model fit the actual results completely randomly. If the result of C-index is 1, the model's prediction is exactly the same as the actual result. C-index is measured in practice as moderately accurate if the value is between 0.71 and 0.90, and highly accurate if it is higher than 0.90. AFT with log-logistic distribution model has the higher values of C-index 0.91169 than that of the AFT with Weibull distribution model, which indicates that AFT with log-logistic model is more accurate than AFT with Weibull model. The result of the C-index is different from the

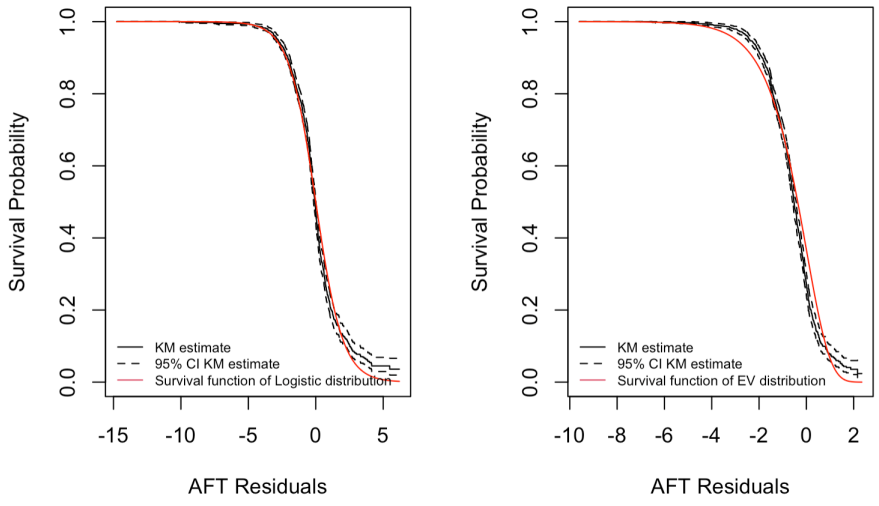

Figure 3.5: Distribution tests for ϵ_i

conclusion of MSE values; It is possible that the difference between these two results is since the AFT model uses right-censored data under the C-index assessment, which incorporates more data compared to the model that calculates MSE. However, the AFT with a log-logistic distribution model is more accurate, as indicated by the C-index is the same as a result obtained from the assumption tests for residuals ϵ_i of AFT.

In the assumption test for each AFT model, the residuals are estimated with censored data. Because the Kaplan-Meier estimator can take into account the censored data, the residuals of each AFT model can be implemented with the Kaplan-Meier estimator for the graphical procedure of survival probability. Moreover, this graphical procedure can then be used to verify an appropriate distribution of survival time in the AFT model, the Weibull distribution or the logistic distribution. The residuals of each AFT model are calculated by the difference between the log of true survival time and linear predictors of the fitted model divided by the scale parameter σ, and then are implemented with the survival function of each distribution, Weibull or logistic. Fig 3.5 shows the results of assumption tests. The figure clearly shows that the survival function of log-logistic distribution is nearer to the Kaplan-Meier estimator of the residuals in comparison to the survival function of the Weibull distribution. Nevertheless, both the survival functions of log-logistic and Weibull distributions deviate from the true survival probability of the residuals at the tail of the Kaplan-Meier estimator of the residuals.

Chapter 4

Conclusion

In this study, we predict the survival time using a predictive linear model and random forest, and we estimate the expected survival time through the AFT model with log-logistic and Weibull distribution. For data that only include occurred death events, we can conclude that the linear model performs more accurately than the random forest since it has smaller MSE results. According to the final linear model, variables including NPI, ER IHC, AGE AT DIAGNOSIS, RFS STATUS, RFS MONTHS, PR STATUS, and TMB NONSYNONYMOUS affect the survival time. However, we encountered the problem that residuals do not fully follow the normal distribution, and the violation of normal assumptions affects the accuracy of the linear model.

For right-censored data, both the C-index and distribution assumptions suggest better performance of the AFT model with the log-logistic distribution. However, one shortcoming of the residual distribution assumption is that in the plot of the survival function of logistic distribution, the residuals deviate at the end of the survival probability curve. This phenomenon also demonstrates that one of the disadvantages of the AFT model is that we must know the distribution of the residuals in order to develop a better AFT prediction model. The reality is that in many cases, the concrete distribution of survival times is unknown. However, if we need to predict the expected survival time, we have to know the specific distribution of the residuals. So if other studies could describe the specific distribution of survival times more precisely, the prediction of mean survival times would be more accurate.

Chapter 5

Appendix

Selected potential variables for models: Lymph nodes examined positive, NPI, cellularity, chemotherapy, ER IHC, HER2 SNP6, hormone therapy, inferred menopausal state, Int-Clust, age at diagnosis, survival months, survival status, claudin subtype, laterality, radio therapy, breast surgery, RFS status, RFS months, ER status, HER2 status, grade, PR status, tumor size, tumor stage, and TMB nonsynonymous.

Bibliography

[1] Soerjomataram I Siegel RL Torre LA Jemal A. Bray F, Ferlay J. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Love SB Altman DG. Clark TG, Bradburn MJ. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[3] John Haman. Accelerated failure time models with citools, 2020.

[4] Westreich D. Mack C, Su Z. *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide*. Agency for Healthcare Research and Quality (US), third edition edition, 2018.

[5] Naveen Naidu Narisetty. Chapter 4 - bayesian model selection for high-dimensional data. In Arni S.R. Srinivasa Rao and C.R. Rao, editors, *Principles and Methods for Data Science*, volume 43 of *Handbook of Statistics*, pages 207–248. Elsevier, 2020.

[6] Jiezhi Qi. *Comparison of Proportional Hazards and Accelerated Failure Time Models*. PhD thesis, University of Saskatchewan, 2009.

[7] Schoenfeld DA Ramchandani R, Finkelstein DM. Estimation for an accelerated failure time model with intermediate states as auxiliary information. *Lifetime Data Anal*, 26(1):1–20, 2020.

[8] Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.

[9] American Cancer Society. Breast cancer facts figures 2019–2020. Technical report, American Cancer Society, Inc., 2019.

[10] Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman Hall/CRC, second edition edition, 2018.

[11] Daowen Zhang. Modeling survival data with parametric regression models, 2000.