Modeling survival data: Application to Larynx dataset

Heqiao Ruan SID:915490857
email:hruan@ucdavis.edu

April 9, 2019

Contents

1 Abstract

Survival Analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen. The core difference of this and traditional statistic is that the data is incomplete due to the limitation of observational studies. Our project primarily focus on applying as many as techniques and advanced methods in survival analysis in analyzing the dataset *larynx* by using R. For the first part we will apply various exploratory analysis on this and extract some certain features so that we can validate them in further modeling. Then we construct the estimated survival function through *Kaplan-Meier* estimator and cumulative hazard *Nelson-Aalen* through estimator.We guess the survival function's distribution and validate them via goodness of fit test. Furthermore,we will apply various non-parametric methods to explore the difference of survival time among different treatments and other covariates. For the second part, we will model this survival data by cox proportional hazard model and do model diagnostics via various techniques. Then we fit semi-parametric accelerated failure time model to examine the covariate effects on event times in censored data regression. as well as some parametric regression models and compare them and do model diagnostics. Finally we will compare these methods and try to identify the factors to influence the survival time and discuss some potential advanced topics such as additive hazard regression models and cox model fitting with partly timevarying effects.

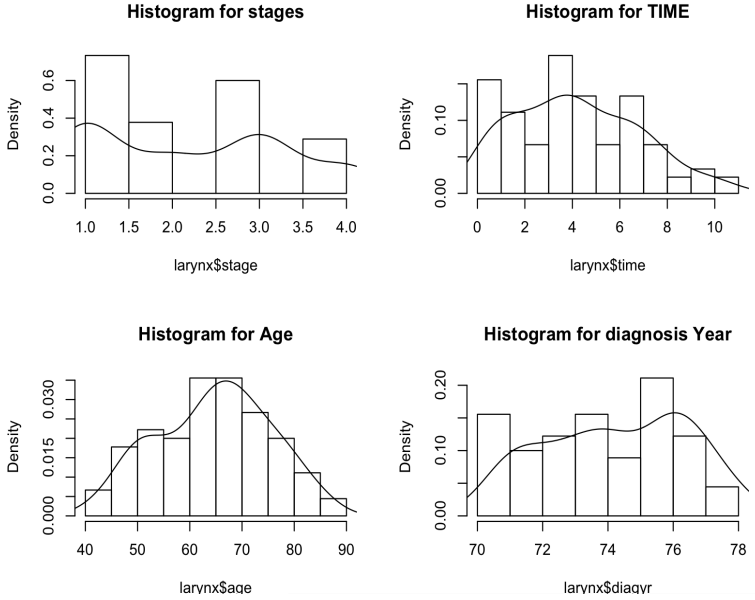
2 Preliminary Analysis

2.1 Introduction and Preprocessing

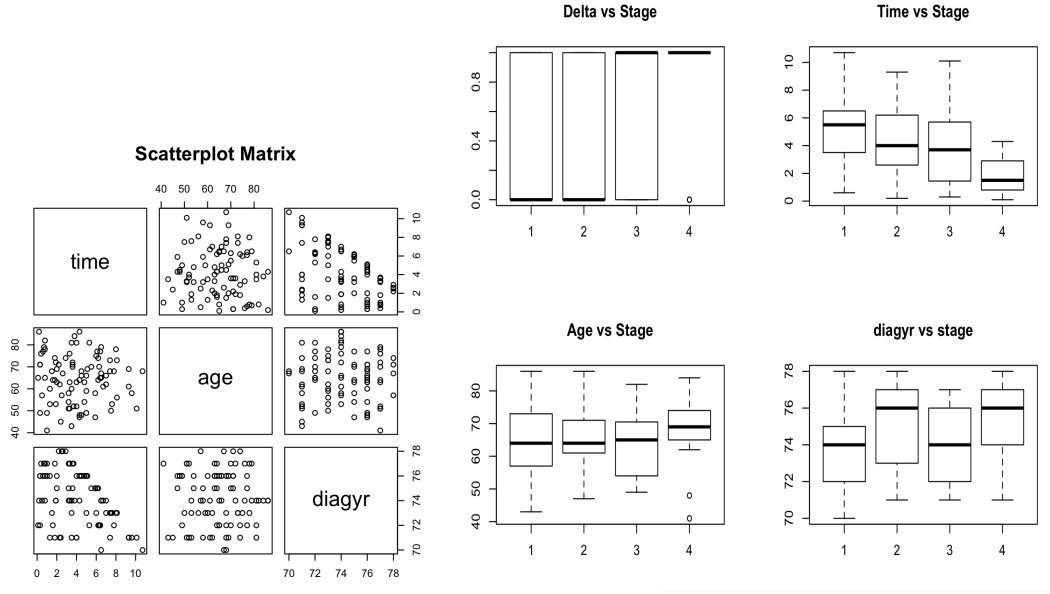
We choose the larynx dataset to conduct our analysis. This is a well-known dataset and it contains 90 rows(observations) and 5 columns(variables).This dataset originally appear in *Survival Analysis Techniques for Censored and truncated data* by Klein and Moeschberger(1997). The attributes are shown in the below table:

Stage	Stage of diseases(1,2,3,4)
Time	Time to death or on-study time,months
Age	Age at diagnosis of larynx cancer
diagyr	Year of diagnosis of larynx cancer
delta	Death indicator(0=alive,1=dead)

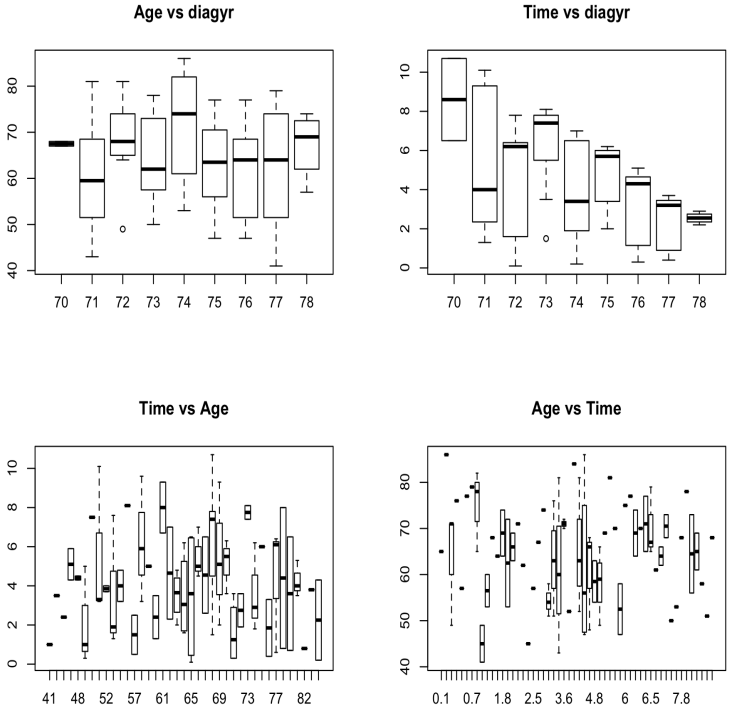
Here we model the survival data as $(\delta_i, X_i)_{i=1}^n$ and denote the delta variable as δ_i and denote the Time variable as X_i. For delta=0 means alive which means this case is censored during the observational study.
Here we can denote the variable *Age*, *Diagyr*, *Stage* as covariates. We can see the age and diagyr as qualitative variables for we can divide them into several groups so that we can examine the influence of them on survival time by using various modeling techniques.
First we are going to do some preliminary data analysis to examine the association between variables intuitively. We draw the histogram of the three covariates and time with its estimated density line:



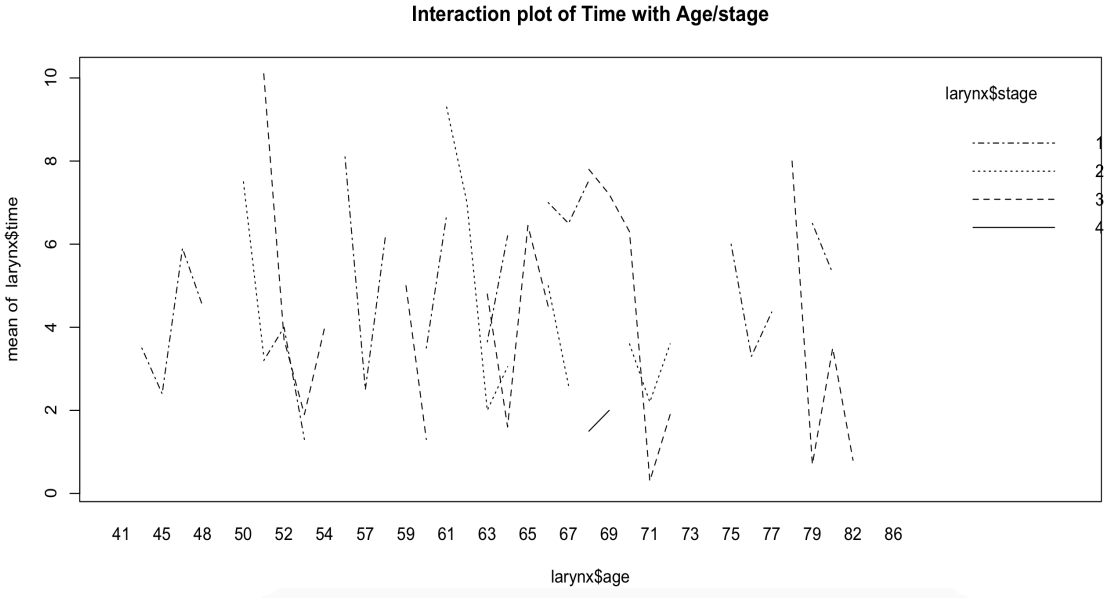
We can see that the distribution of age seems likes normal. There seems no obvious trend of the distribution of diagyr and stage. For the distribution of survival time,there's also no obvious pattern.
Then we draw the scatterplot matrix of the variables age,diagyr and censoring time:



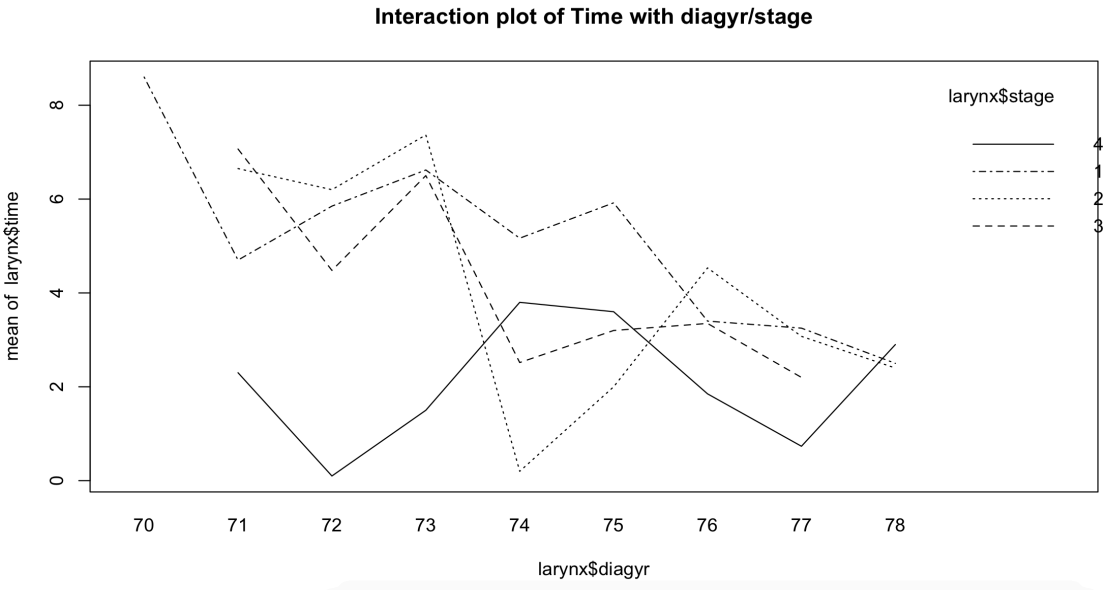
We can see that there's no obvious direct linear relation between these variables.
Then we plot the boxplot between the other variables and stages.We can see that survival time,diagyr and age are different among different stages. Observations with stage 4 have the smallest censoring time and latter the stage the shorter the censoring time. It matches our intuition and we will validate later. What's more, observations with stage 3 and stage 4 have more death cases than censoring cases which means they tend to die earlier than the other two stages. It also obeys our intuition.
Then we examine the relation between covariates age,diagyr and time:



From the plot we can see that, generally, later the cancer is diagnosed, the shorter the censoring/survival time. However,the older age when diagnosed does not necessarily mean this observed people will have shorter censoring time. It also depend on which stage the cancer is when they are diagnosed.So here we have to use the interaction plot to examine the relation between time and age diagnostics depending on different stages of cancer:



Here we still can't see obvious influence of stage1,2,3 on the censoring time of observations with different stages. However,observations diagnosed as stage 4, it tends to have shorter censoring time. What's more, the four lines(may not even continuous) are not parallel so we can conclude that there are relation between age and time depend on different diagnostic stages. Similarly we can draw the interaction plot of relation between time and diagyr's dependence on different stages.



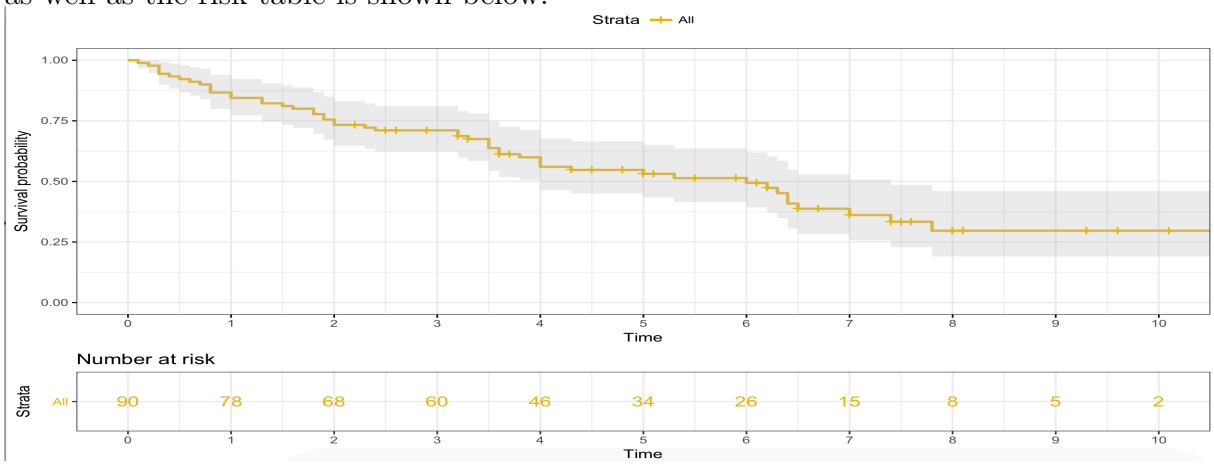
We can see that the lines are not parallel which means the relation between time and diagyr is depend on different stages. It also validate there are interaction effects among them. So there are interaction effects between these covariates.

For further model fitting and comparing survival curves between different groups,we should transform the continuous covariate to categorical variable which means we should divide them into several groups and denote them as category 1-4 and change their class as factor in r.We group them both by 4 quantiles, for variable age: if $Age \leq 55$ then denote as category 1,if $55 < Age \leq 65$,then denote as category 2,if $65 < Age \leq 75$,then denote as category 3,else denote as category 4. For variable diagyr: if $diagyr \leq 72$ then denote as category 1,if $72 < diagyr \leq 74$ then denote as category 2,if $74 < diagyr \leq 76$ then denote as category 3,else denote as category 4. By letting them as class factor,we transform them into categorical variables and add at the right side of our dataset for further survival model fitting(we maintain the original variables for we can fit a model with continuous variables).
Then we see the basic summary statistics for every group, first we see that the mean survival time for every group are different:5.25 for stage I,4.38 for stage II,3.93 for stage III and 1.83 for stage IV.

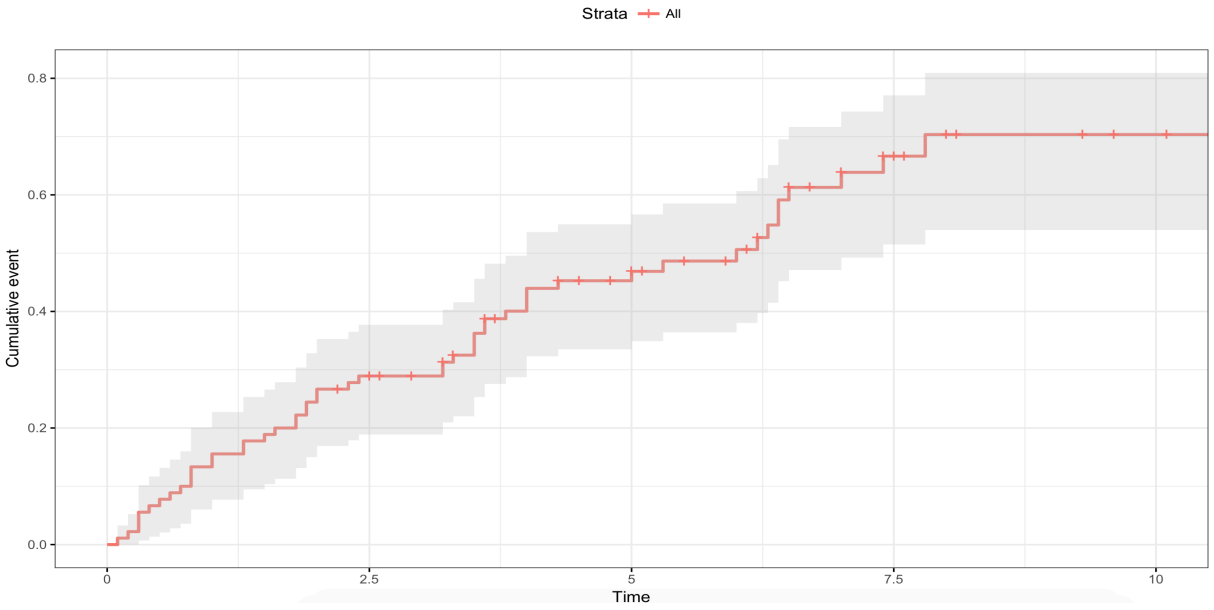
2.2 Estimating survival function and relevant exploration

One of the core task for modeling survival data is to estimate the survival function and the related statistics.First we get the risk set and the event set for every observation and then we can derive kaplan-meier estimator and Nelson-Aalen estimator based on the information in the dataset. Then we estimate the mean and median of the survival function.
As we all know, the kaplan-meier estimator is given by $\coprod_{x_i \leq t}(1-\frac{1}{n-i+1})^{\delta_i}$ and in this problem we have some ties which means the death case and censoring case for a specific observed time is not unique(tied).So the formula become: $\hat{S}_i = \coprod_{j \leq i}(1-\frac{d_j}{r_j})^{\delta_j}$ Here r_j is the risk set at time j and d_j is the number of death observations at time j.
For Nelson-Aalen estimator, the formula is $\Lambda_i = \sum_{i \leq j} \frac{d_j}{r_j}$. Then we would like to derive the interval estimation of Kaplan-Meier estimator. Here if we only use the traditional technique which indicates that $var(\hat{S}_t) = \int_0^\infty \frac{dF_u}{(1-H(S))^2}(\int_t^\infty S(s)ds)^2) = (\hat{S}_t)^2 \sum_{j:\tau_j \leq t} \frac{d_j}{(r_j-d_j)r_j}$(Greenwood formula) we will find that some of the estimators is more than 1,it's not accurate.Note that this formula is indeed derived from a method we applied many times in empirical process including matrix transformation.
So here we apply log-log technique to modify this.Here we define $L(t) = log(-log(S(t)))$,then $\hat{L(t)} = log(-log(\hat{S}(t)))$. The confidence interval for L(t) is $[\hat{L}(t)-H, \hat{L}(t)+H]$,here H is determined by $\phi(0.95)sd(\hat{L}(t))$ So by delta method, the variance become $\frac{1}{[log\hat{S}(t)]^2}\sum_{j:\tau_j \leq t}\frac{d_j}{(r_j-d_j)r_j}$. Then the confidence interval for $\hat{S}(t)$ become:$([\hat{S}(t)]^{e^H}, [\hat{S}(t)]^{e^{-H}})$. Then we can see that none of the interval estimator will contain 1. Similary,for Nelson Aalen estimator the interval estimator is becoming: $[\hat{A}(t)e^{\phi(0.95)\frac{\hat{\sigma}(t)}{\hat{A}(t)}}, \hat{A}(t)e^{-\phi(0.95)\frac{\hat{\sigma}(t)}{\hat{A}(t)}}]$ where $\hat{\sigma}(t) = \sum_{t_j \leq t}\frac{(r_j-d_j)d_j}{(r_j-1)r_j^2}$.

Then comes one of our core part,we fit the survival function by the *survfit* function in R package *survival*. Here we would like to use the function *ggsurv* and *ggsurvplot* in R package *survminer*. For we want to compare the survival time and survival function among different categories of categorical variables, we fit them with respect to *factor(stage)*,*factor(age)* *factor(diagyr)* as well as the full model. Firstly we fit the full model and the survival curve as well as the risk table is shown below:

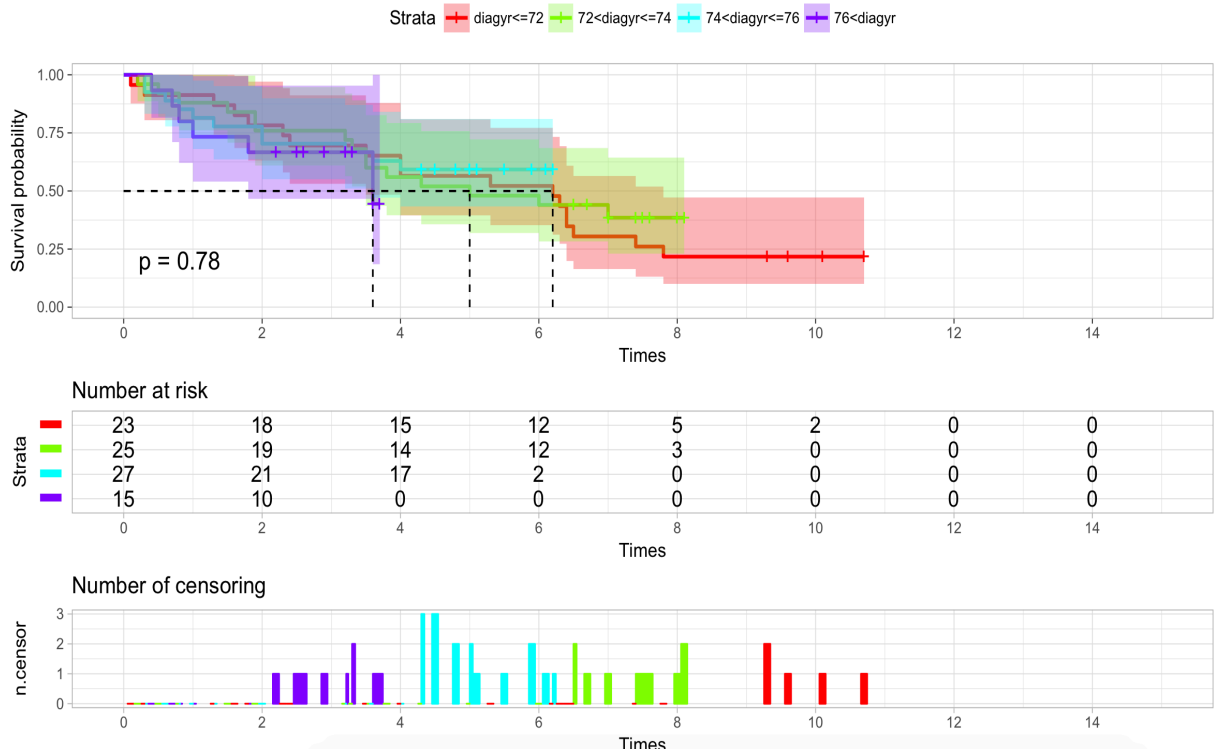


Here we may guess the distribution of the survival function as exponential or weibull distribution.We will do goodness of fit test to see which guess will have more probability to be right in the next part. Then we plot the cumulative hazard of the full model:

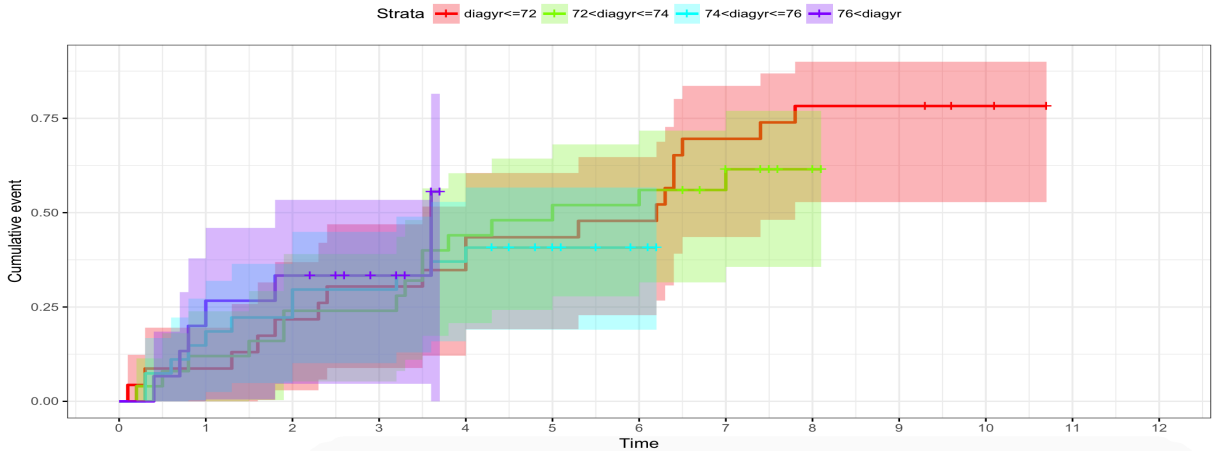


From these plots we can see that the estimated interval for the estimated cumulative hazard(Nelson-Aalen estimator) and kaplan-meier estimator are becoming much wider as the time increase.It is indeed natural from the formula.

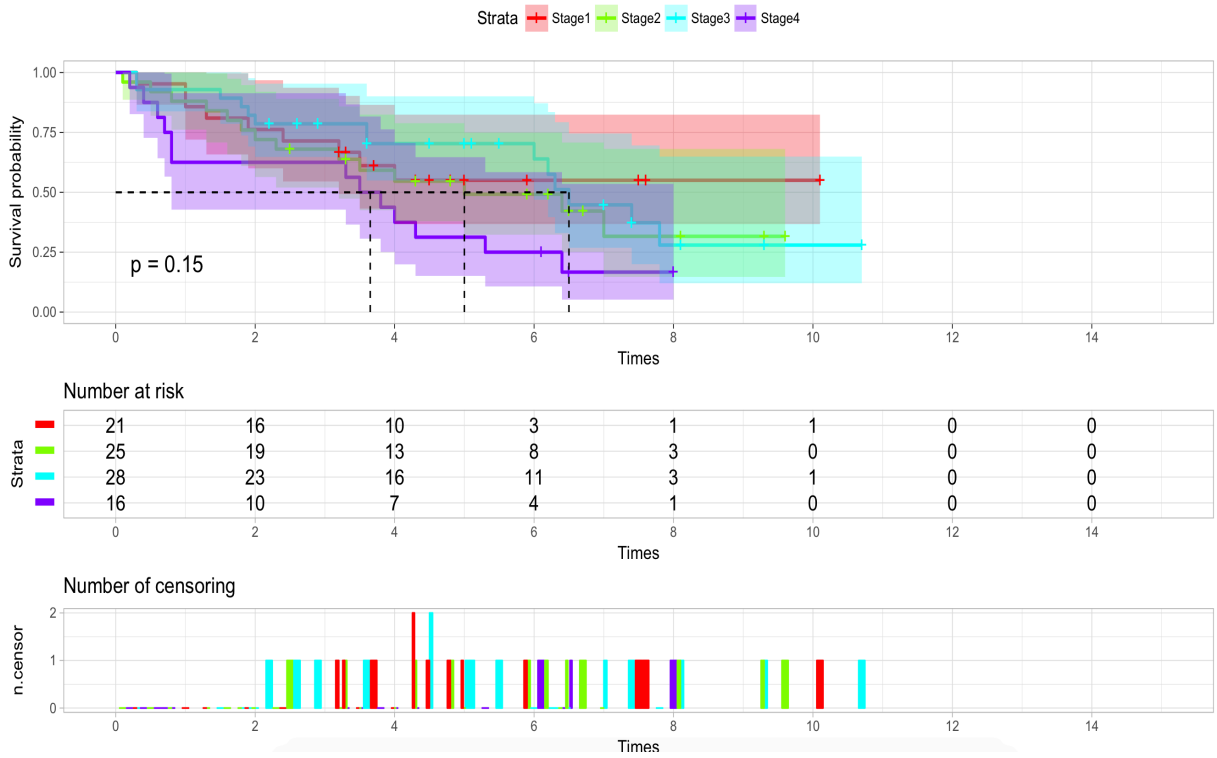
Then we fit the model with respect to the four diagyr group and the survival curve as well as the risk table and table of number of censoring is shown below:



Here we can see that the group with the youngest diagnostic year of cancer has the longest survival time and the longest median survival time and the censoring status is perfectly separable by the four groups of diagnostic year.We see that there are certain kinds of difference among these survival curves but the difference may not be very significant.
Then we plot the cumulative hazard curve for diagyr groups:

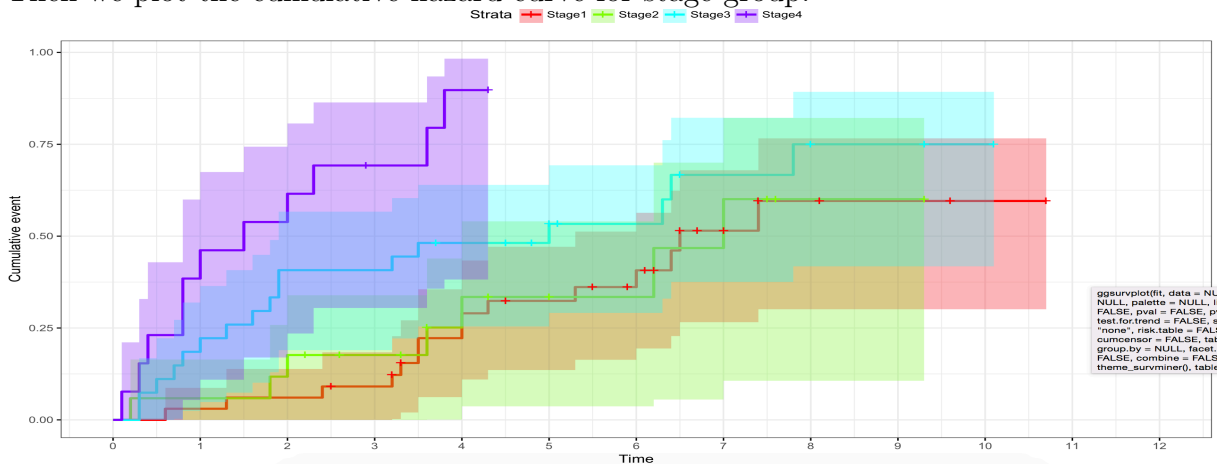


From here we can see that the four curves are fairly similar in their scope of existence and further test later may validate this. Then we fit the model with respect to the four stages of cancer and the survival curve as well as the risk table and the table of number of censoring is shown below:



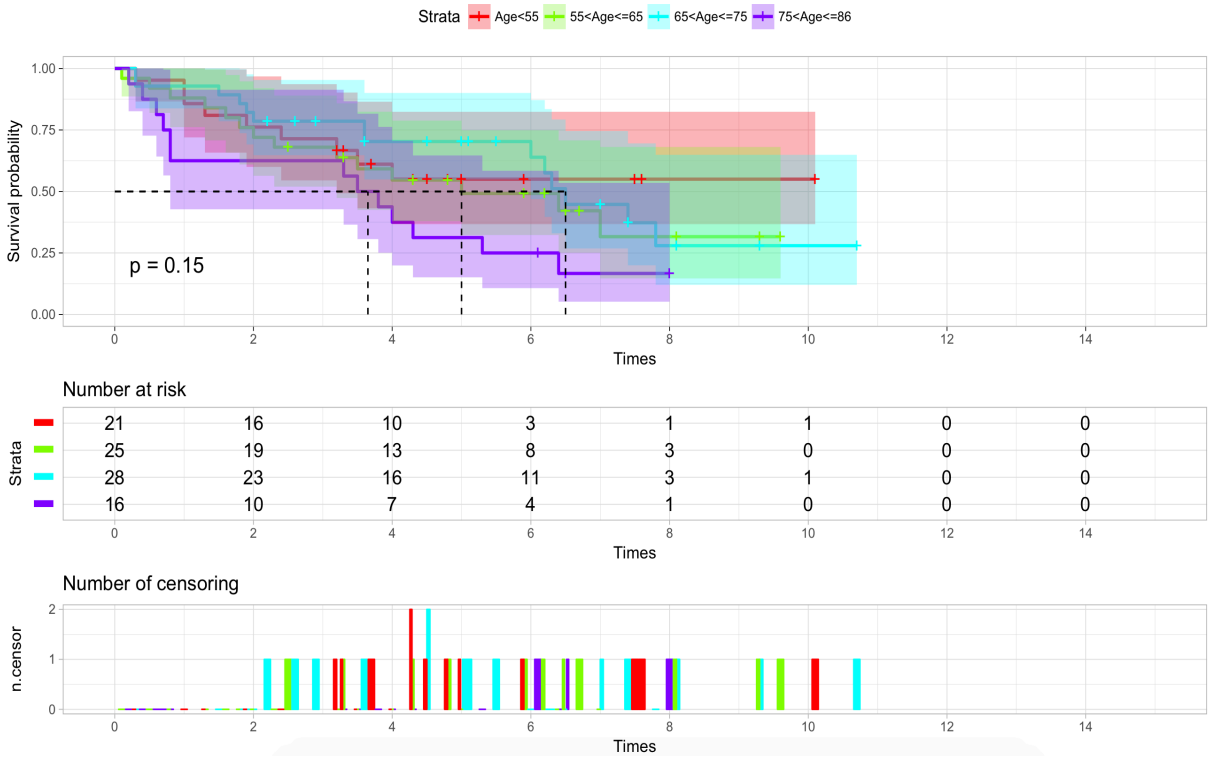
Here we can see that the difference of survival curves of stage1,stage2 and stage 3 is not as obvious as the difference between survival curves of stage4 cancer and them. What's more, the censoring observations are not so perfectly separable among different stages as among different diagnosed years. The median survival time is becoming smaller as the stage become later(stage 4 is a later stage 1 of the cancer i.e).It matches the common sense.

Then we plot the cumulative hazard curve for stage group:

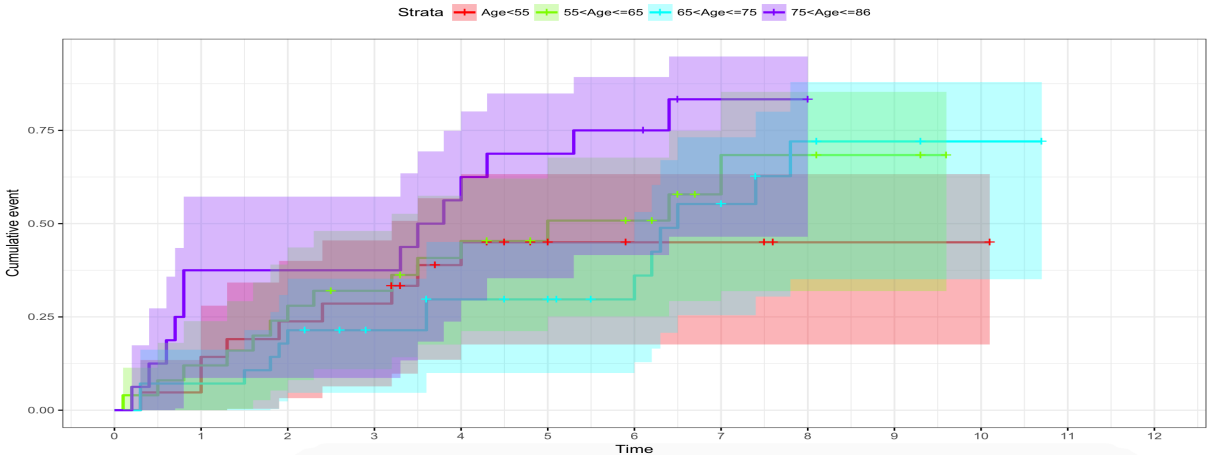


We can conclude that the four cumulative hazard are really different which means the four survival curves are also significantly different in their scope of existence and further test will validate this.

Then we fit the model with respect to the four groups of age and the survival curve as well as the risk table and the table of number of censoring is shown as below:



In terms of survival functions in different age groups, we can see that generally, older people tend to have lower median survival time and we can see that as the time flies the estimator interval become larger and larger(this is trivial for the formula of the variance).However, patients in age 65-75 years old tend to have lower risk to die as patients whose age is less than 55 years old at first and it will be validated later in the cox model.
Then we plot the cumulative hazard curve for age group:



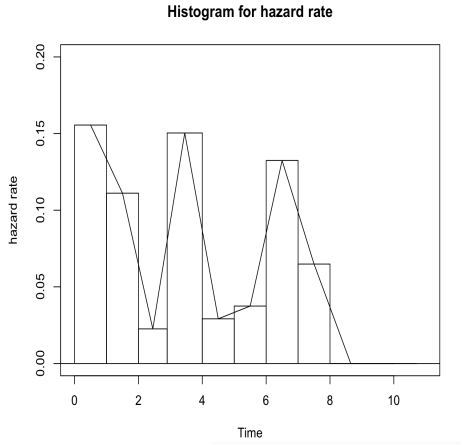
Then we estimate the mean and median of the survival time. According to the lecture,the mean survival time point estimator is $\hat{\mu} = E[\hat{T}] = \int_0^\infty S(\hat{S})ds = \sum_{i=0}^{n-1}(X_{i+1} - X_i)S(\hat{S})$ and its variance is $\sum_{i=1}^{n}[\int_{t_i}^{\tau} \hat{S}(t)dt]^2 \frac{d_i}{Y_i(Y_i - d_i)}$.This is derived by the idea of empirical process. Then we know that in asymptotic point of view it follows normal distribution. Then in the problem,we construct 95% confidence interval,its result is 5.689754 for the estimated mean survival time

and [4.827,6.552] for the interval estimate for the mean survival time. We implement this method by R without calling the survfit function in the survival package and we can check it in appendix.
Then for the median survival time: $median = \hat{S}^{-1}(\frac{1}{2})$ We see that $\hat{S}(k) > \frac{1}{2}, \hat{S}(k+1) < \frac{1}{2}$ then(for more accurate),the point estimator is $\frac{\hat{S}(k+1)+\hat{S}(k)}{2}$. What's more,we know that its variance is $\sum_{i=1}^{n} \frac{d_i}{Y_i(Y_i-d_i)}$. Similarly,we have asymptotic normal so we can get the 95% interval estimation. Its result is 5.95 for the estimated median survival time and [5.439,6.464] for the interval estimation. Notice that it's much larger than the mean survival time. This is usual for the estimator gives increasing jump sizes with increasing t and due to censored observations dropping out,the gaps between uncensored observations tend to increase with t. So $\hat{\theta}$ tends to be larger.
Then we estimate the mean survival time and median survival time of each stage:
For the different diagnosed years the median survival time of four categories are 6.2 for category 1,5.0 for category 2,3.6 for category 4 and for category 3,the survival function will never reach 0.5 so no median. We can see that the earlier the cancer diagnosed,the longer is the survival time.
For the different age of diagnosis, the median survival time of four categories are 5 for category 2,6.5 for category 3,3.65 for category 4. For category 1,the survival function never reaches 0.5,no median.
For the different stages of cancer,the median survival time of four stages are 6.5 for stage1,7.0 for stage 2,5.0 for stage 3,1.5 for stage 4. Here we can see that the median estimator of stage 2 is not that accurate,it's much larger than the sample mean.So later the stage when diagnosed,the shorter survival time.
So we can see that after comparing the median estimators of every groups,it is fairly consistent with what we conclude in preliminary analysis(boxplot). However the difference of median and mean survival time does not necessarily indicate the significant difference between different curves. This may result from the difference of their domain. There may be some smoothing techniques to improve the efficiency of the estimation of the median survival time.

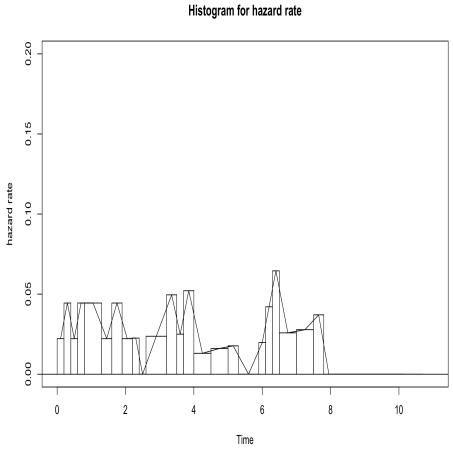
2.3 Estimating Survival Function's Distribution and Goodness of Fit test

A very important topic is to what distribution the survival function follows and if we can show that the survival function nearly follows certain distribution,we can do parametric regression models on the dataset and it will greatly help us.In doing this the core part is to do the goodness of fit test,we can first draw the histogram for the survival function and hazard rate and guess what the distribution may look like. We plot the histogram for the hazard rate and is shown as below(the breaks are almost same time interval).
However,in real life, the dataset often not do well in helping us identifying what the distribution of survival function is. Then if the survival function shows no obvious pattern(no distribution will make sense here), the minimum chi-square technique will not work. That's what I have to point out.

First after drawing the histogram to see what kind of distribution it may follows.However,it's not easy for us to identify what kinds of distribution it may follows purely from the histogram.The histogram shows no obvious pattern of the distribution of the survival function.



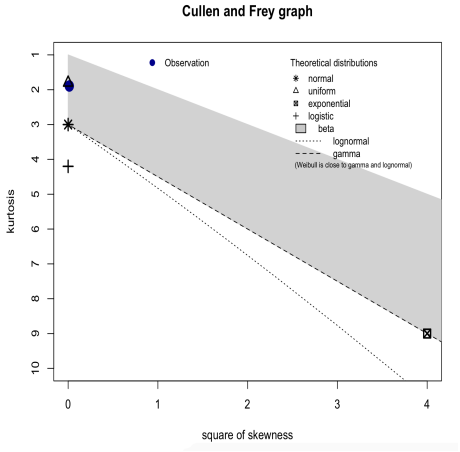
Unfortunately, we can guess any specific kind of distribution just based on this histogram, if we then choose the breaks that there are only two points in every break, the histogram is shown as below:



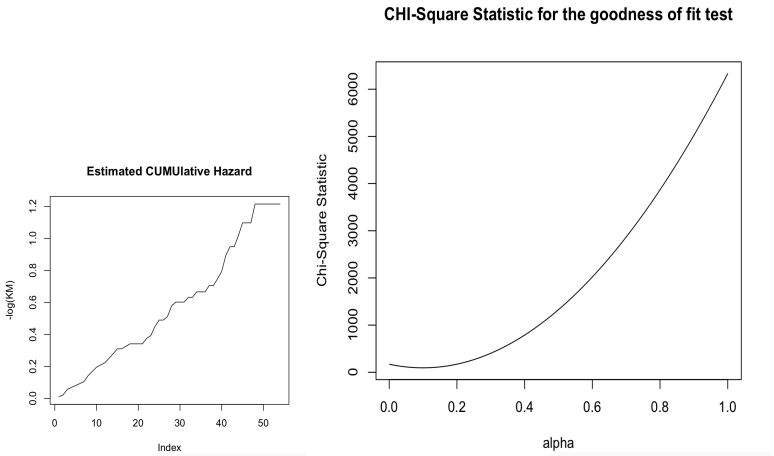
I still can't guess which specific distribution it may follows. So we do can try some parametric distributions such as exponential,weibull and even gampertz but these are all return to very negative results.

However, we should somewhat apply the methods mentioned in this course to do the goodness of fit test by using minimum chi square technique. Here the minimum chi square technique construct a gaussian process which $Z(y_1), ..., Z(y_k)$ which $Z(y_1)$ is independent to $Z(y_i) - Z(y_{i-1})$,i=2,3,...,k. Here $Z(y_i) = \hat{\Lambda}_n(y_i)$ is the cumulative hazard. Then the test statistic we construct is $\sum_{k=1}^{K} \frac{(\Delta Z_k)^2}{var(\Delta_k Z_n(y_k))}$ here $\Delta_k Z_n(y_k) = Z_n(y_k) - Z_n(y_{k-1})$ and $\Delta_1 Z_n(y_1) = Z_n(y_1)$. It is a function of the parameter of our assumed distribution. If the assumed distribution is exponential(α), $\Lambda(t) = \alpha t$.Then we do find the parameter that minimizes this statistic and the distribution of the minimized chi-square statistic is χ^2_{K-1}. Here we may need to do simulations to find the optimized parameter α. Note that this procedure

is better than the ordinary chi square test $A = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i^2}$ for here we use the independent increment property of the gaussian process and thus wipe out the covariance approximately among different terms on statistic A. Here we may assume the distribution of the survival function is exponential and do this procedure. We find that the optimized α is 0.09 and at this time the minimum chi square statistic is 94.66 and we can see that the chi-square statistic is drastically large when α increases. It's obviously an awful fit.
Then we use the *descdist* function in the fitdistrplus package in r to try to see which distribution fits the survival function $\hat{S}_t$) best based on skewness and kurtosis.



From the plot we can see that the uniform distribution fits the survival function(our KM estimator) best. It totally makes non-sense.
Then we try to see whether the cumulative hazard follows certain pattern:we draw the graph of $\hat{\Lambda}_t$ and it is shown as below:
The only thing we can see from this very graph is that the cumulative hazard may seems



to looks like a straight line so may be we can guess is that the survival function follows exponential distribution which has been proven awful fitting by the previous minimum chi-square technique. So unfortunately, from both the minimum chi-square technique and even some more advanced methods, we can't identify any kinds of parametric models based on this

dataset. So maybe all of the parametric techniques may not works very well in this case.

2.4 Non-Parametric Techniques

In terms of comparing several survival functions, we don't assume any specific distribution of the data so the non-parametric techniques including tests and modeling are required. Here we apply Mantel-Haenszel test and wilcoxon rank test to this multi-sample problems which is just an natural extension of two-sample problem. Here we want to test whether the survival function under each group are the same. We have three covariates and we examine their difference with respect to different age_groups,$stages$ and $diagyr_{groups}$.
We apply the Cochran-Mantel-Haenszel test. We know that in two sample problem we have a series of contingency table and the weighted sum of the statistic constructed based on each contingency is our statistic. In the multi-sample problem we can just extend it naturally and our test statistic still follows chi-square distribution.
For convenience we obtain the risk-event table first reflecting the event set and risk set for each group in every observation.
Then if we want to compare the survival function of each stage group with the whole dataset.Now the null-hypothesis become $H_0 : F_1 = F_2 = .. = F_p$ which means there's no difference of survival function between each group. The test statistic here become: where r_j is the number of elements in the jth risk set of full dataset.$r_{i,j}$ is the number of elements in the jth risk set for stage i. Here for the multi-sample problem, let $O_j = (d_{1,j}, ..., d_{p,j})$ bet the vector of observed number of failures in group $1- > p$,$E_j = (\frac{d_j r_{i,j}}{r_j}, ..., \frac{d_j r_{p,j}}{r_j})^T$ denotes the mean of O_j. This is in fact similar to one kind of goodness of fit test $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$Then they have a covariance matrix: $V_j = \begin{bmatrix} v_{11,j} & v_{12,j} & ... & v_{1p,j} \\ ... & v_{22,j} & ... & v_{2p,j} \\ ... & ... & ... & ... \\ ... & ... & ... & v_{pp,j} \end{bmatrix}$ where diagonal element is $v_{kk,j} = \frac{r_{k,j}(r_j - r_{k,j})d_j(r_j - d_j)}{r_j^2(r_j - 1)}$ and non-diagonal element is $v_{km,j} = \frac{r_{k,j} r_{m,j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$. For all observations we sum all the v matrix and all the vector $O_j - E_j$ up. According to asymptotic theory we know that $(O - E)^T V^{-1} (O - E)$ follows $\chi^2(p - 1)$ distribution. Then after calculation the test statistic become:22.8 and the p-value is $4.53e - 05$. So we reject the null hypothesis which means we have enough evidence to state that the survival functions are different among the 4 stages.
Similarly when we conduct this test in age group and diagyr group,it doesn't works well for the p-value of the CMH test are 0.145,0.783 respectively which means we should state that the survival functions are the same among these two kinds of grouping in significance level 0.1. There are also some non-parametric tests such as Gehan-test,tarone-wane test. However, they are indeed very similar to the CMH test but only differences in the weights in each observation(each contigency table) and CMH test is the most commonly used test here.We have to point out that this kind of test is limited to location shift and in scale shift,it won't work well.

3 Main modeling and exploration

3.1 Cox Proportional Hazard Model

Cox model is one of the most fundamental part of survival analysis. The idea of proportional hazard model comes from comparing the two survival curves and getting the relative risk. $S_1(t) = (S_0(t))^\theta$. So in terms of hazard rate we have equation: $h(t|Z) = h_0(t)exp(\sum_{k=1}^p \beta_k Z_k)$,in another form we have:$\frac{h(t|Z)}{h_0(t)} = exp(\sum_{k=1}^p \beta_k Z_k)$ where $\theta = exp(\sum_{k=1}^p \beta_k Z_k)$
Here the cumulative hazard are also proportional:
Here,if we want to fit the cox model with the covariate stage(4 categories) and ages. Then the model become: $\lambda_l(t) = e^{\beta H_l}\lambda_0(t)$ where $\lambda_0(t)$ is the baseline hazard. The exponential term is $exp(\beta H_l) = exp(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_{p-1} Z_{p-1})$. In the modeling of age and stages,the covariates Z become:
Z_1:1 If stage II of cancer,0 otherwise;$H_l = (1, 0, 0, Z_4)$: If stage II of cancer
Z_2:2 If stage III of cancer,0 otherwise;$H_l = (0, 1, 0, Z_4)$: If stage II of cancer.
Z_3:3 If stage Iv of cancer,0 other wise;$H_l = (0, 0, 1, Z_4)$:If stage III of cancer.
Z_4:Age variabels.$H_l = (0, 0, 0, Z_4)$: If Stage I of cancer.
Another model we consider(called model3) is only consider the stage variable and then the model become: $\lambda_l(t) = e^{\beta H_l}\lambda_0(t)$ where $\lambda_0(t)$ is the baseline hazard. The exponential term is $exp(\beta H_l) = exp(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_{p-1} Z_{p-1})$. In the modeling of age and stages,the covariates Z become:
Z_1:1 If stage II of cancer,0 otherwise;$H_l = (1, 0, 0)$: If stage II of cancer
Z_2:2 If stage III of cancer,0 otherwise;$H_l = (0, 1, 0)$: If stage II of cancer.
Z_3:3 If stage Iv of cancer,0 other wise;$H_l = (0, 0, 1)$:If stage III of cancer.
If Stage I of cancer,$H_l = (0, 0, 0)$.
Our dataset has various ties so we should use another approach to compute the partial likelihoods. The most common one is conducted by Breslow(1974), $L(\beta) = \coprod_{i=1}^n \frac{exp^{(\beta^T \sum_{j\in D_i} z_j)}}{[\sum_{j\in R_i} exp^{\beta^T z_j}]^{d_i}}$,here d_i denotes the number of deaths at t_i and D_i be the set of all individuals who die at time t_i.Here we use this method to handle the ties. This likelihood considers each of the d_i events at a given time as distinct and construct their contribution to the likelihood function and obtains the contribution to the likelihood by multiplying over all events at time t_i. The estimator $\hat{\beta} = argmax_\beta L(\beta)$ maximize the partial likelihood. Here the log-likelihood function is $l(\beta) = \sum_{i=1}^n \delta_i(\beta^T z_i) - \sum_{i=1}^n \delta_i ln(\sum_{j\in R(t_i)} exp(\beta^T z_j))$ .So score function is $U_\beta = \frac{\partial l}{\partial \beta} = \delta^T Z - \sum_{i=1}^n \delta_i \frac{\sum_{j\in R(t_i)} exp(\beta^T z_j) z_j^T}{\sum_{j\in R(t_i)} exp(\beta^T z_j)}$.In real framework,we can solve it numerically.
Here,various tests are applied in evaluating different models and we will explain them one by one briefly. First: wald test is to test a hypothesis about a subset of the parameter β,the null hypothesis $H_0 : \beta_1 = \beta_{10}$ where $\beta = (\beta_1^T, \beta_2^T)^T$ and here β_1 is $q \times 1$ vector and β_2 is $(p - q) \times 1$ vector.For we know that $(\hat{\beta} - \beta)^T I^{-1}(\hat{\beta})(\hat{\beta} - \beta)$ follows $\chi^2(p)$ distribution, then the test statistic now for the subset become: $\chi^2_{subset} = (\hat{\beta}_1 - \beta_{10})^T [I^{11}(\hat{\beta})]^{-1}(\hat{\beta}_1 - \beta_{10})$,it follows χ^2_q distribution and we reject the null hypothesis when $\chi^2_{subset} > \chi_q(\alpha)$ in significance level α.Usually we often denote β_{10} as 0 vector and in real problems we often test the global

assumption: $\beta_1 = \beta_2 = .. = \beta_p$. Other tests include likelihood ratio test and score test. The likelihood ratio test here is also applied in evaluating difference among models. The null hypothesis is $H_0 : \beta_k = 0$,the alternative hypothesis is $H_a : \beta_k \neq 0$(the variable added in the full model from the reduced model) The test statistic is $L = -2 \times log - likelihood(reduced\ model) - (-2 \times log - likelihood(full\ model)$, which asymptotically follows χ^2_{k-1} distribution. Then if $L > \chi^2_{k-1}(1-\alpha)$ we reject the null hypothesis which means the full model may be correct.If we want to test the whole model's validity,we can also apply the same test. For the score test,it is also called the logrank test and its test statistic is $T_{SC} = U_1(\hat{\beta}_0)^T I^{11}(\hat{\beta}_0) U_1(\hat{\beta}_0)$ here $U_1(\beta)$ is the subvector of first q elements of the score function $U(\beta)$ and $\hat{\beta}_0 = (0^T, \hat{\beta}_2{}^T)$ is the mle of β under null hypothesis,and in null hypothesis,the test statistic follows χ^2_q distribution.

Here we use the *coxph* function in package *survival* to fit the model and get the coefficients and relevant testing statistics.

Now let's start to fit the model: first we have the null model which only include the intercept: β_0 in $e^{\beta H_l}$ and a full model which include all of the linear terms and the interaction terms among these covariates(age,stage,diagyr) which equivalent to the full model. Then we apply *step* function in r to choose the model,this function indeed choose AIC criterion which denoted as: $-2 \times log - likelihood + 2 \times n_{par}$ which combines the model complexity and goodness of fit on the data. We find that the model we find via *step* function including *stage,age,diagyr and age:diagyr*.We denote this as $model1.1$. Naturally we consider another model with no interaction terms which also denoted as first-order model. We denote this as $model2$.

Then we fit the full model with 2 order(including interaction terms) and then we apply model selection technique AIC(Akaike Information Criterion)(we choose not to use BIC for BIC often returns to a smaller model and may lose some importance information) to select the best submodel to fit the cox regression model.We denote our this chosen model as model1.

Then if we are interested in examine the effects of diagyr on survival time, we fit the first-order model here with age,stage(II,III,IV) and diagyr as predictor variables.However,we can see that the pvalue for the diagyr variable is 0.80,which is pretty large. So this variable is not significant and this is pretty consistent with our common sense which may because there won't be any significant change in the environment in this period of time which will impose such a great influence in the survival time of patients with all ages with various stages of the cancer diagnosed.The coefficients are shown as below:

Variables	Coefficients	P value
Age	0.01869	0.1922
factor(stage)2	0.15164	0.7442
factor(stage)3	0.64473	0.0703
factor(stage)4	1.73211	7.09e-5
diagyr	-0.01819	0.8120

This indeed coincides with our previous tests which indicates that there's no difference of survival time among diagyr(CMH test).So we just ignore this model.

Then after deleting the diagyr variable from the model, we naturally fit the model with age and stage which appear at the top of this subsection($Z_1 - Z_4$). We denote this as model3 this is a model we are really interested for we know that the most important variables influencing the survival time are age and different stages. Then another model we may interest in is only to include the stages covariate in the model which helps us explore how the stage of the cancer influence the survival time exactly. We denote this as model4 and its coefficients is shown as below:

Variables	Coefficient	Standard Error	p value
Z1:Stage II	0.06481	0.45843	0.8876
Z2:Stage III	0.6218	0.35519	0.0835
Z3:Stage IV	1.7349	0.4194	3.52e-5

From this model we can see that the relative risk of stage II over stage I is$exp(\beta_1) = e^{0.06481} = 1.06696$ while the relative risk of Stage III of the cancer over stage I is 1.8493 and the relative risk of Stage IV of the cancer over stage I is 5.688,we can see that in stage Iv of the cancer,the risk of death improves drastically than stage I while stage II 's relative risk over stage I is almost one which means there may not be large difference between this two stages.However,it depends on another covariate age according to later model interpretation. Note that we can compare this with that in AFT model.

According to the previous exploration, there are certainly interaction effects between stage and age, so we should fit the model with stage(i-iv) and age and there interaction effect terms to examine the effect of different stages on age,we denote this as *model3.int*.Then the model has 7 parameters(age,stage2,3,4,three interaction terms). After examing the summary we can see that the interaction terms:*age:factor(stage)3* and *age:factor(stage)4* are not significant(pvalue are more than 0.7). So we should drop these two interaction terms. Then the model become *model3.updates*.Here we show the coefficient of the model for sake of interpretation:

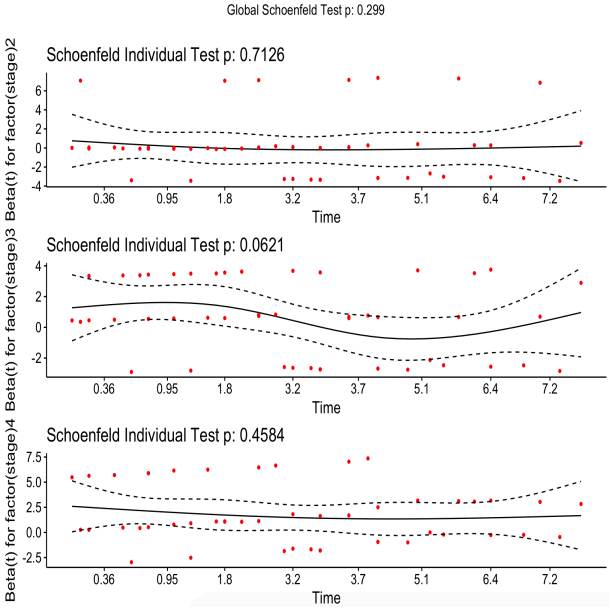
Variables	Coefficient	Standard Error	p value
Z1:Stage II	-7.3820	3.4027	0.03
Z2:Stage III	0.6218	0.3558	0.08
Z3:Stage IV	1.7534	0.4240	<0.0001
Z4:Age	0.0060	0.0149	0.69
Z5:Z1× Z4(Interaction term)	0.1117	0.0477	0.02

Here we can see that that there is a significant interaction between age and stage II disease. Then the relative risk of dying for a stage patient of age Z_4 as compared to a stage I patient of the same age depend on the age. The relative risk is $e^{\beta_1+\beta_5 Z_4} = e^{-7.382+0.1117\times Age}$. For a 80 years old patient, the relative risk is 4.730 while for a 65 years old patient,the relative risk is 0.886 and for a 50 years old patient, the relative risk is only 0.17. Then with this model we can also test whether the risk of dying is the same for different ages. The null hypothesis here is $H_0 : \beta_1 + \beta_5 \times Age = 0$.To test this we use a technique very similar to the local wald test: the test statistic is $\chi^2_{RISK} = \frac{(\beta_1+\beta_5\times Age)^2}{V(b_1)+age^2\times V(b_5+2\times cov(b_1,b_5))}$ Under null hypothesis it follows χ^2_1 distribution. Then the test statistic for a 60 years old patient is 0.99 with a p-value 0.32

while for a 75 years old patient the test statistic is 4.09 with pvalue 0.04. This test do suggest that for young age there's little difference in survival between stage I and II patients while for old patients the stage II greatly effects the lifetime of the patients.
According to the previous part,the survival curves between different diagyr are not significantly different due to the CMH test(which is also validated with score test and local wald test,we fit the model only consisting of age and factor(stage). Finally, we choose model3 and model4 which reflects the survival time among different stages(model4) and different ages with stages(model3).The several tests we mentioned before's statistic are shown below in the table:

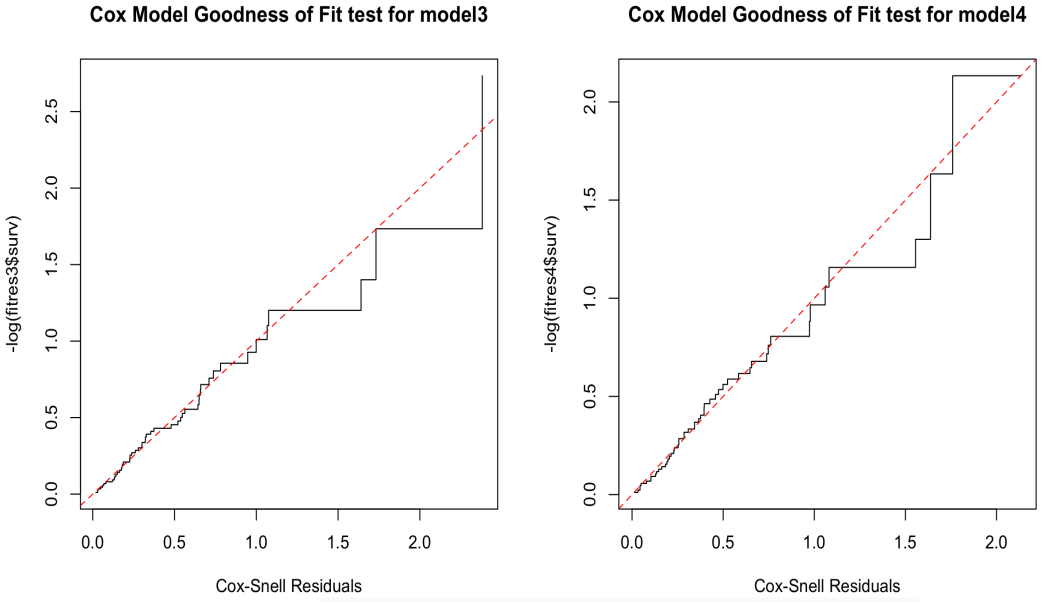
Models	Wald Test	Likelihood Ratio Test	Score test
Model 3	21.15	18.31	24.78
P value	0.0002958	0.001072	5.57e-5
Model 4	19.24	16.49	22.88
P value	0.0002433	0.0009016	4.28e-5

We can see that all of the tests indicates that the model is significant.Then we do the model diagnostics for the two models.
To diagnose the cox model we fit,we should first test whether it follows the assumption of cox-model.One assumption is the proportional hazards assumption which means the survival function of different groups are nearly proportional which means $\Lambda_l(t) = \Lambda_0(t)\theta$,it doesn't vary over time. Here we use the *ggcoxzph* function in *survminer* package to draw the residual vs time plot for the model3(with only stage).Here we use a graphical diagnostics based on scaled schoenfeld residuals. If the residual doesn't have obvious pattern over time, it may be consistent to the proportional assumption.The plot is shown as below:

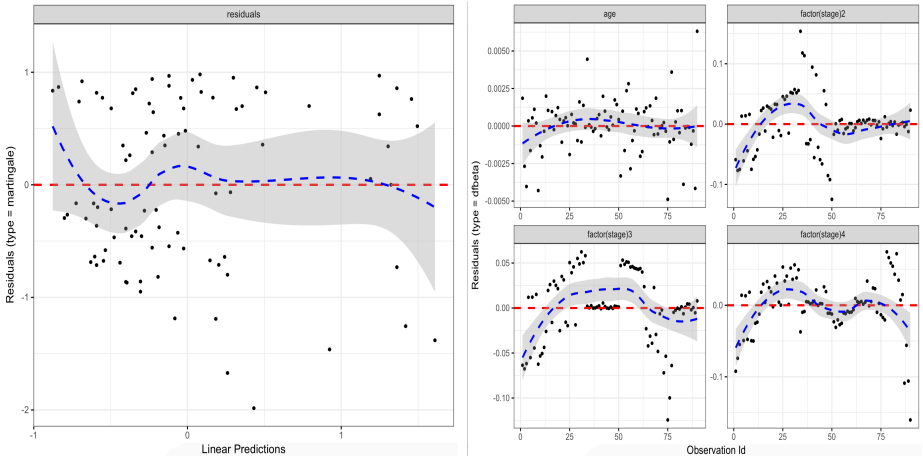


We can see that there's no obvious pattern of these residuals over time. So the model is fairly consistent with the proportion hazard assumption.
Another diagnostic method is to do the goodness of fit test: according to the lecture, $r_j =$

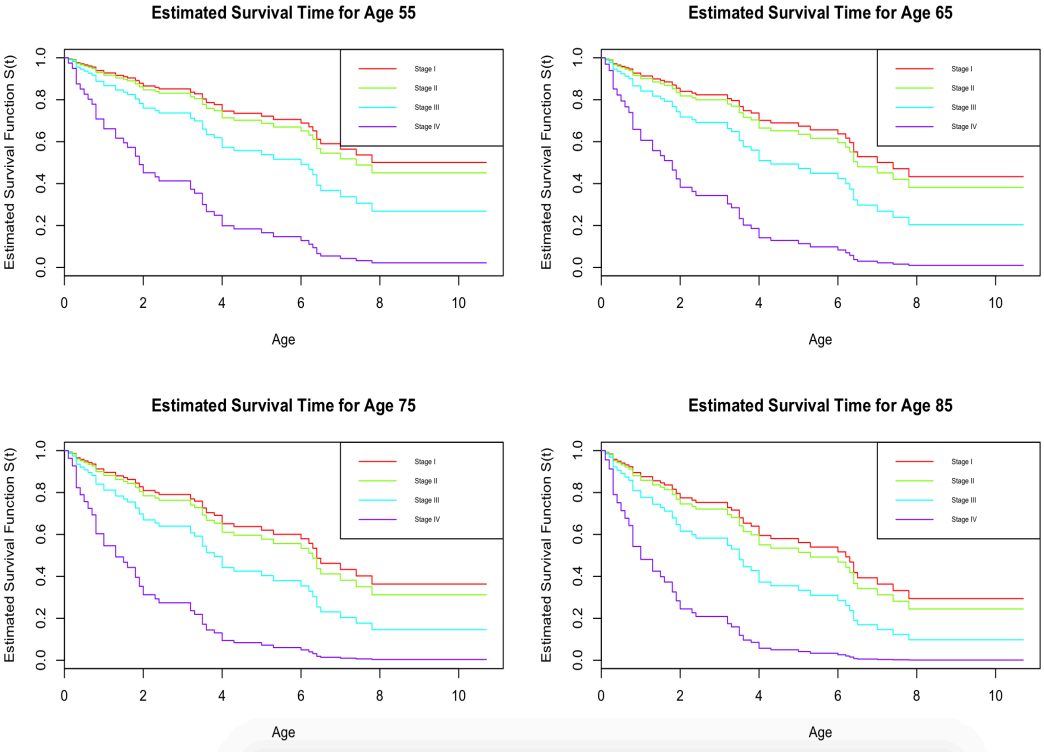
$\hat{\Lambda_0}(x_i)exp(\beta z_i)$ follows exponential distribution with parameter 1.This is because $P(\Lambda_0(x) > t) = P(x > \Lambda_0^{-1}(e^{\beta z_i}t)) = exp(-\Lambda_0(\Lambda_0^{-1}(e^{\beta z_i}t))) = e^{-t}$.Here $\hat{\Lambda_0}(x_i)$ is the baseline hazard rate. Here we use cox-snell residual to examine the fit of the model. Cox and Snell proposed a more straightforward method in 1968:to check whether the r_j behave as a sample from a unit exponential,we can compute the Nelson-Aalen estimator of the cumulative hazard rate of $r_j's$(We use (r_j, δ_j) to fit the model and find the cumulative hazard function). If it fits well, this estimator should be approximately equal to the cumulative hazard rate of the unit exponential $\lambda_t = t$,which means $\hat{H}_r(r_j)$ versus r_j plot should looks like a straight line through the origin with slope 1.So this is indeed equivalent to the method in the lecture and it will be more convenient to test intuitively by the graph.
We examine this for our chosen model model3(w.r.t age and stage) and model4(w.r.t stage) and the plots are shown as below:



From the plot we can see that it's indeed a decent fit but there are some outliers(doesn't so badly) and model4 fits slightly better than model3.
However,there are still various different methods to do the model diagnostics via residuals diagnostics: deviance residuals, martingale residuals. The deviance residuals can be used to test influential observations and we can use *ggcoxdiagnostics* function in *survminer* package to do the exploration.We do this for model3(with stage and age covariates).What's more,the martingale residual can be used to test the non-linearity and we won't implement it here.What's more,we can also visualize the dfbeta values which can plot the estimated changes in the regression coefficients upon deleting each observation in turn;Following are the two graphs of the diagnostics:
From the first plot we can see that the pattern looks not really symmetric around 0 and from the second plot we can see that the some of the observations are influential but the dispersion are not extreme which means there are some outliers(especially for stage covariate) but these won't make our modeling nonsense. Here the negative one means that the patient may live too long compared to their expected survival time and the positive residuals correspond to



individuals that died too soon compared to expected survival time. So from this plot we can see that there are some patients lives much longer than they expected which indicates that there are certain amount of outliers(can't be perfectly predicted by the model).I have to say it's very usual in the practical case for the real data often doesn't looks really like what we derive in theory.That's one of the most important that I have learned from this project.
After modeling, we can do prediction on the survival time of a individual of a certain age: here we use the model3 to do it,we predict the survival time for individual aged 55,65,75,85 and the plot is shown as below:



From the model3 we can see that the estimated survival time for stage I cancer patients is

$S_0(t)^{exp(0.019*age)}$, for stage II patient the survival time is $S_0(t)^{exp(0.019*age+0.140)}$,for stage III patient the survival time is $S_0(t)^{exp(0.019*age+0.6423)}$ while for stage IV the survival time become $S_0(t)^{exp(0.019*age+1.7059)}$. From the plot we can see that for every age, the survival time tends to decrease as the latter stage of the cancer diagnosed.What's more,the survival time for patients stage IV seems not very different among different ages,they are all much more shorter than the earlier stages. It's consistent with our guess and observation in preliminary analysis.
Then we would like to treat the age variable as a categorical variable and divide them into 4 groups(like that in part2.1) and fit the cox proportional model with covariates factor(age) and factor(stage) and we denote this model as model5. The coefficients and the test statistics are shown as the table below:

Variable	coef	pvalue	variable	coef	pvalue
age_2	0.34072	0.4312	stage II	0.16680	0.7209
age_3	-0.03841	0.9292	stage III	0.6767	0.0588
age_4	0.7619	0.0856	stage IV	1.768	2.91e-5
Wald Statistic	24.69	LRT	21.28	Score	28.44

From the coefficient we can see that this model is decent and the several tests tells us the regression relation is significant. From the model we can see that the different ages of diagnosis of the cancer do influence the survival time and we can see that the survival time tend to be even longer of age group3(65-75 years old) than age group1(<55 years old) which indicated by the factor that the relative risk of age group 2 to age group 1 is less than 1. It is consistent with our guess in part 2.2. Then the relative risk of age group2 to age group 1 is 1.40,the relative risk of age group4 to age group 1 is 2.14236.It is consistent with our guess that older people tend to have shorter survival time.
Then we try to fit the interaction model and after fitting this, we find the stage4:age_{group2} variable is significant(check it in R output) and this indicate that the fact that age group2 with smaller risk than age group1 is influenced by the fourth stage. The exponential of coefficient for the interaction term is 2.59 and it's much smaller than the exponential of coefficient of stage IV patients 4.87. This indicates that patients with age 55 to 65 years old if they are diagnosed as the latest stage cancer, will tend to have lower risk in dying as fast as other groups. This may be an reasonable explanation for the phenomena we found in the previous paragraph.

3.2 AFT model

Accelerated failure time model are alternative to relative risk models which are used extensively to examine the covariate effects on event times in censored data regression. The model is: $T_i = e^{\beta z_i} U_i$ after log-transformation,here $Y_i = logT_i$ the model become $logT_i = z_i^T\beta + logU_i$,by re-parametrization,the equation become $Y_i = z_i^T\beta + e_i(\beta)$ Here β is an unknown $p\times 1$ vector of regression parameters and U_i are completely unknown and z_i is the covariate and $e_i(\beta) = logU_i$. So the cumulative hazard of T_i: $\Lambda_{T_i}(t) = \Lambda_U(e^{\beta z_i}t)$.In the presence of right censoring,the observed data are independent copies of (),where $Y_i = min(T_i, C_i), \delta_i = I(T_i < C_i)$ and $I(.)$ is

the indicator function.
For the error terms' distribution is completely unknown, we have two methods to fit this model, the first is to assume that the error term follows some distribution and do the parametric modeling. The second is to fit the semi-parametric model based on the observed data and censoring status. Now first we assume that the error term follows certain distributions. From the above part we can see that the *diagyr* variable is not significant in regression. So we try to fit the model:$Y_i = log(X_i) = \mu + \sum_{i=1}^{4} \beta_k Z_k + \epsilon_i$ Here $Z_1, .., Z_4$ are defined as:
Z_1:1 if stage II cancer,0 otherwise;
Z_2:2 if stage III cancer,0 otherwise;
Z_3:3 if stage IV cancer,0 otherwise;
Z_4:Ages variable(denote the patient's age at diagnosis).
Empirically, the common used distributions include weibull,log-logistic,log-normal,rayleigh and exponential distribution. Then we use the *survreg* function in package *survival* to fit the parametric regression model with respect to each distribution of error terms. After fitting these models we get the log-likelihood of these models,for AIC=-2*(log-likelihood)+2*(number of predictor variables+number of parameters in assumed distribution). If we want to test the whole model's validity,we apply the goodness of fit test and we will do it later.

Criterion,Distribution	Exponential	Weibull	log-logistic	log-normal	Rayleigh
log-likelihood	-141.9	-141.4	-141.6	-141.4	-155.3
AIC	293.8	294.8	295.2	294.8	322.6

Obviously,model with exponential distribution has the smallest AIC. So now our parametric regression model become: $Y = log(X) = 3.755 - 0.1456Z_1 - 0.6483Z_2 - 1.6350Z_3 - 0.0197Z_4 + \epsilon$. The chi-square test statistic is 18.44(indeed the wald test mentioned in part 3.1) and its p-value is 0.001 which means the coefficients(regression relation) are significant.In terms of interpreting this model: we can see that using the accelerated failure-time model for the exponential model, we see that the acceleration factor for stage II is $exp^{(-(-0.1456))} = 1.1567$,the acceleration factor for stage III of cancer is $exp^{(-(-0.6483))} = 1.9122$,the acceleration factor for stage IV of cancer is $exp^{(-(-1.6350))} = 5.1294$.It's a little bit similar to the relative risk in the cox model(model 4) This may suggests that the survival time for stage I patients is about 5.13 times than stage IV patients,1.91 times that of Stage III patients,1.12 times that of Stage II patients.So individuals with stage II,III,IV cancer tend to have shorter life times than individuals with stage I cancer. This is consistent with our exploration in preliminary analysis. However,there may be some limitations among this method:the exponential distribution doesn't fit our model very well, we can see this from the following plot(using *fitdist* function in r to fit the best possible distribution).We can see that the distribution fitted doesn't make any sense. So here the parametric regression model may not fit very well for this dataset although we may get similar conclusion like in the previous part of cox model fitting and preliminary exploration.
Secondly we apply the semi-parametric technique based on the observed data when we have totally no information about the residual term.We use the R package *aftgee* proposed by Sy Han Chiou and lss function in R package *lss* to help our modeling. The *aftgee* package implements the rank-based procedures and least-square estimators based on the semi-

parametric model which greatly alleviate the limitations of usage of Semi-parametric AFT model.(Based on the GEE non-parametric technique).Here the function's implementation also account for multivariate dependence through working correlation structures to improve efficiency.A well known method of estimating T_i is the Buckley James estimator(1979). $\hat{Y}_i(\beta) = \delta_i Y_i^{'} + (1-\delta_i)[\frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}_\beta(u)}{1-\hat{F}_\beta(e_i\beta)} + Z_i^T\beta]$. We do the buckley james estimator for the survival time vs age, diagyr and stage.
The coefficients are shown as below:

Model	Coefficients	P value(wald test)
Age only	-0.0167	0.1942
Stage only	-0.5832	<0.0001
diagyr only	-0.0621	0.3781

Here to examine the significance,we use wald test(similar to that in cox model).The estimator of age are only slightly different from that in the cox model and the ordinary AFT model. However,the r function bj is very limited here because we can't even plot the log(T) vs diagyr and stage. So this is what greatly hinge the use of this technique greatly.
From the plot we can see that the survival time tend to decrease as the age increases(it is consistent to our preliminary analysis). Here we plan to fit two accelerated failure time models: with stage and age covariates(like model3 in subsection 3.2),only with stage covariate(like model4 in subsection 3.2).
We apply the lss function in R package *lss* to fit the AFT model and compare this to that in *aftgee* function,here we indeed use the least square method.The model we want to fit become: $log(X_i) = \sum_{i=1}^{k} \beta_i Z_i + \epsilon_i = \sum_{i=1}^{k} \beta_i Z_i + log(U_i)$. After fitting this model we should check the model's adequacy. To check the model's adequacy,we use goodness of fit test here.We know that if the model fits well the residual term U_i follows unit exponential distribution(with parameter 1).To test whether the term U_i follows the unit exponential distribution, we can first use the formula $A = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i^2}$,k is the number of breaks we choose for convenience, we know that if the null hypothesis is true, the statistic A follows χ^2_{k-1} distribution. For the model1.1,we can see that the test statistic is 1.4 and we absolutely accept the null hypothesis so we claim it fits the unit exponential distribution well.So model1.1 is adequate. Then for model1.2,for we can see that there are several severe outliers then the test statistic is pretty large and we reject the null hypothesis thus we claim that model1.2 may not a good model.This may because some regression coefficients of model1.2 are not significant(the p value for stage 2 is 0.804) and the algorithm may not accurate enough in terms of this specific structure.However,we will still see the model's parameter and try to extract some information from them.
When we fit the model with stage only(we denote here as model 1.1),the parameter is shown as below:

Least Square estimator for model1.1	Estimate	Std.Error
Stage II	-0.247	0.4334
Stage III	-0.9466	0.369
Stage Iv	-1.89	0.43
Aftgee estimator for model1.1	Estimate	Std.Error
Intercept	2.086	0.227
Stage II	-0.247	0.441
Stage III	-0.946	0.385
Stage IV	-1.895	0.465

We can see that the estimator for both techniques are roughly the same.So the algorithm is pretty consistent in this model.(This two methods works equally well). In terms of interpretation of this model we can see that there appears to be accelerated failure(death) time on later stage(II,III,IV) compared with stage I. The acceleration factor for stage II is $e^{-(-0.247)} = 1.280$,for stage III is $e^{-(-0.946)} = 2.576$ while for stage IV is $e^{-(-1.895)} = 6.651$. It's different from that in the parametric regression model and this accelerated factor should be more accurate than the previous one. When we fit the model with stage and age(we denote here as model 1.2),the parameter is shown as below:

LSS Square estimator for model1.2	Estimate	Std.Error
Age	-0.0393	0.0384
Stage II	-0.461	1.217
Stage III	-2.571	1.091
Stage Iv	-4.138	0.907
Aftgee estimator for model1.2	Estimate	Std.Error
Age	-0.030	0.023
Stage II	-0.133	0.536
Stage III	-1.054	0.456
Stage IV	-2.070	0.523

There are some different in parameter estimation between this two methods but the difference may not extreme. We can see that the coefficient of age is larger than that in the cox proportional hazard model. What's more, we can see that the different ages will surely influence the survival time of the patients in different stages. So it naturally indicate us to add some cross-product terms to represent this kind of interaction terms.
Then we try to fit the AFT model with interaction terms between age and stage. However, both methods in r not work well. The model fitted by *aftsrr* function are not reasonable for most of the regression coefficients are not significant while the model fitted by *lss* doesn't converge in 1000 iterations. So it requires further research and investigation in actually proposing a more robust algorithm to implement this model fitting with various interaction terms or other more complex structures.

3.3 Additive Hazard model and advanced topics

Here we point out an alternative method to compare the four stages of laryngeal cancer, the additive hazard model. The additive hazard model is an alternative to the semi-parametric multiplicative hazard model.We want to express the hazard rate and the cumulative hazard as a(t)+X(B(t)) and a and X are both time-dependent. We have an event time X whose distribution depends on a vector of covariates which may be time-dependent and we denote this as $Z(t) = [Z_1(t), ..., Z_p(t)]$. We assume that the hazard rate at time t for an individual with covariate vector $Z(t)$ is a linear combination of the $Z_k(t)$'s then the model become: $h[t|Z(t)] = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) Z_k(t) + \epsilon$.We can see that this is indeed a regression model too and the error term's distribution is totally unknown which indicates that it's a semi-parametric model.

Here we consider the model with covariates same as the model3 in subsection3.1:note here we center the age covariate at its mean.We use *aalen* function in *timereg* package to do the model fitting.

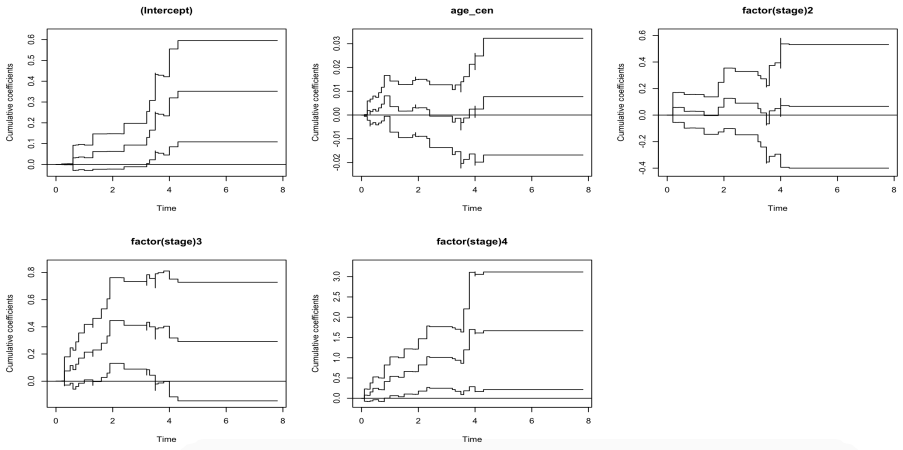
Z_1:1 If stage II of the disease,0 otherwise.

Z_2:1 If stage III of cancer,0 otherwise

Z_3:1 If stage Iv of cancer,0 other wise.

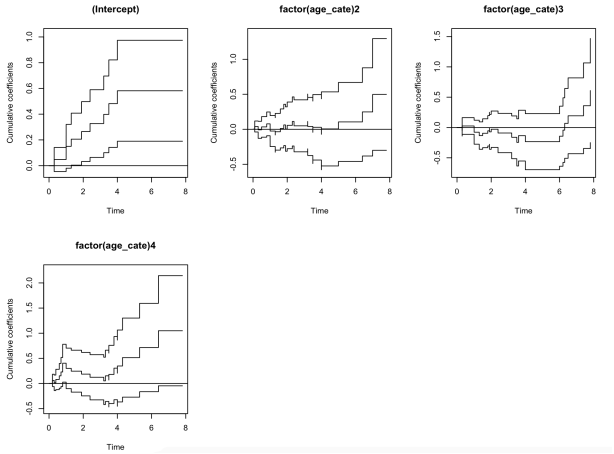
Z_4:Age at diagnosis.

Here we can get the excess risk of stage II,III,IV cancer as compared to stage I and we can also see a 95% pointwise confidence interval for the patients.The plot is shown as below:

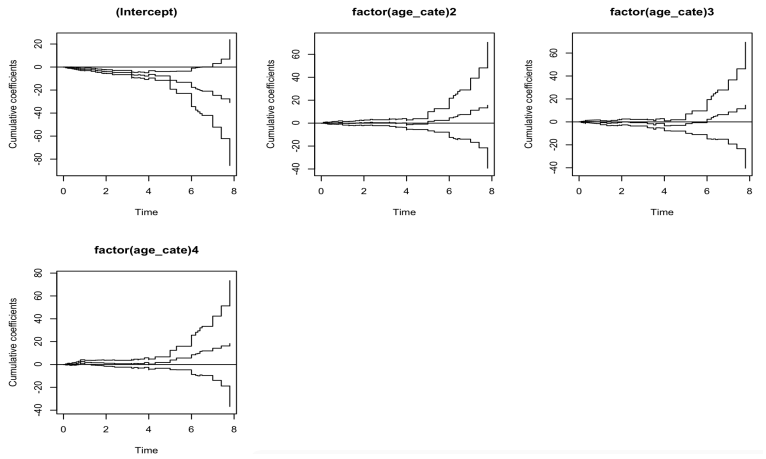


This figure show the estimate of $B_k(t) = \int_0^t \beta_k(u)du, k = 0, 1, 2, ..., p$,the first for intercept is indeed the baseline hazard. Here the baseline hazard is an estimate of the cumulative hazard rate of stage I patient in the mean age(64.1 years old). We can see from the plot that there appears little excess risk of stage II patients while for stage III and stage IV patients, they tend to have an much more risk in the first two years of diagnosis and the excess risk tends to be invariant after 2 years of diagnosis of the cancer.What's more,the age plot shows that the excess risk has relation with age but the relation seems not so significant.What's more we can also test whether there are no difference in survival time between four stages of larynx cancer adjusting for age.(Similar to wald test and we won't implement this) instead of observing the plot intuitively.

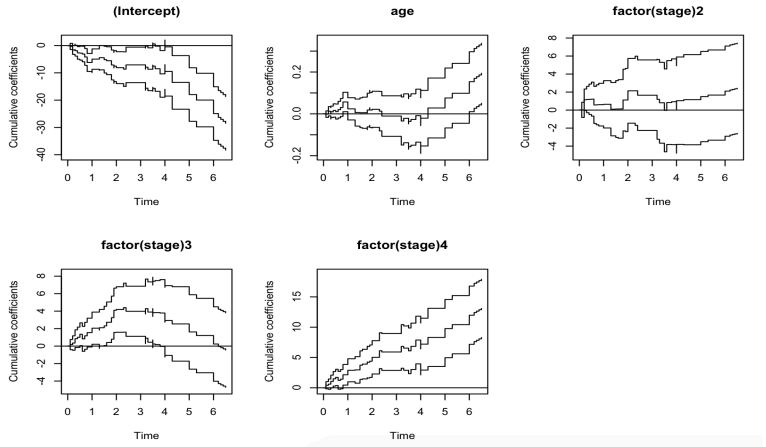
Then if we fit the model with which age variable are categorical(4 groups reflecting the relative age group). Then we can examine the excess risk of older people than younger people.The variables are shown as below:
Z_1:1 If Age group II,0 otherwise
Z_2:1 If Age group III,0 otherwise
Z_3:1 If Age group III,0 otherwise
The we plot the estimate of $B_k(t) = \int_0^t \beta_k(u)du$ for every variable and the graph is shown as below:



From the plot we can see that the excess risk of age group II and group III at first times are not significant and at shorter times,the risk of age group III is even smaller than the age group II. However as time passes,the excess risk of older people become larger and larger. So this may indicate that patients with older age when diagnosed of cancer tend to have larger risk at longer times. It matches our common sense for the patients with later stage of cancer won't live for a long time.
What's more,there are so many related topics we can do with this model. For example we can test $H_0 : \beta_k(t) = 0$ for all $t \leq \tau$ which means the covariate is zero before a certain time. Various research have been conducted among these. What's more,there is another kind of additive hazards model proposed by Lin and Ying which replace the time-varying regression coefficients in the Aalen model by constants.The further derivation of the parameter estimation and model diagnostics are somewhat based on non-parametric techniques.
Now apart from the two most important models mentioned in this course,we plan to do some advanced analysis on this dataset.
Then we use the *timecox* function in *timereg* package to try to fit cox model with partly time-varying effects. For the time-varying coefficients, the hazard of failure $\lambda(t|X) = \lambda_0(t)exp(g(\beta, t)X))$ and g is a function of time. Then after fitting this we plot the cumulative coefficients with time which indicates the difference of coefficients of one single variable among time(time-varying). We fit two models, one is like model3 in the cox model(the included variables are the same) and the other is treating age covariate as 4 group-categorical variable.The plot from the second model is shown as below:



The plot indicates that the if we treat age as several groups, we can't see significant difference of coefficients among times.
Then we draw the plot from the first model:



We conclude from the plot that the coefficients of stages are significantly varying as time passes especially for the stage IV variable while stage II and age variables' coefficients show no obvious pattern(change) as time passes.
Various tests can be implemented to test the constant effects of the coefficients:Supremum-test,Kolmogorov-Smirnov test and Cramer Von Mises test.Then we can read the test statistic from the summary and we can see that the p value for constant effect of age is solid and the constant effect to stage variables are not that solid.(check for detail in R output) So we can conclude that the effect of stage variables(categorical) are somewhat time-varying and the effect of age variables(no matter treating as categorical or numerical variable) are not significantly time-varying.We have to say that this is indeed a broad topic and can extend to many interesting models.

4 Conclusion and Discussion

In this project we use the larynx project to do survival modeling.Firstly,we observe that there are obvious difference of survival time among different stages of cancer and we conclude that later the stage of the cancer,the shorter will the patients survival.We conclude that there are interaction effects between age and stage in explaining survival time.
In terms of limitations, here the distribution of survival function is fairly weird and the parametric techniques doesn't work well. Maybe it's fairly normal in real life dataset and that's we are passionate on the non-parametric techniques with foundation of empirical process.
In cox model fitting we do the model fitting,model selection and model diagnostics and we have found that the the difference of survival time between stages are affected by ages(interaction) which indicates that for younger age there's little difference in survival between stage I and stage II but for old people, the difference will be more significant while for stageIII vs stageI and stageIV vs stageI the effect of age is not that significant. It is indeed inspiring for the scientists to do more research to extend the survival time of older patients with later stage of cancer.What's more,in the part of prediction, we see that the survival time for stage IV(last stage) is much smaller than the other stages. So more research should be conducted in regardless of improving the survival time of old people(especially older than 85 years old) with this kind of cancer and improve their quality of life. When we examine the effect of age on survival time, we can see that it does not necessarily mean patients older age must have shorter survival time.
In the AFT model fitting we can see that the accelerated factor estimator will be more accurate in using non-parametric techniques for none of the parametric techniques work well based on our exploration of the dataset.We have also found that the later the stage when diagnosed,the faster will be the patient death and the degree of accelerating are different among different stages.
From the advanced topic part we do the clustering and find that the four stages can't be perfectly distinguished just by simple clustering so more advanced methods may help. What's more we fit the cox model with partly time-varying coefficients and find that some coefficients' effects are time-varying.
In terms of limitations, here the distribution of survival function and the minimum chi-square technique not makes sense too.Alternatively,we use *fitdist* function in r to try to approximate a reasonable distribution but it still doesn't work very well which means the parametric techniques won't work well. Maybe it's very normal in the real-life data analyzing.The parametric model is limited to a very small part of cases and for most of the time,applying non-parametric techniques will help us more.What's more, the current implementation of semi-parametric AFT model and the buckley-james estimating function bj in r is not really robust to all kinds of linear models(the algorithm even not converge ever when we add a interaction term of stage and age) so it greatly limit our analysis and interpretation in AFT modeling.
Survival analysis and modeling survival data is such a broad topic and there are so many kinds of models and lots of research has been conducted in the past 40 to 50 years and although we can't exhaust all methods in one project because of the restriction of the data, I still find there are a lot to explore further in this subject.

5 Bibliography

1.Cox,D.R.(1972).Regression models and life tables(with discussion).Journal of Royal Statistical Society Serie B,34,187-220
2.Anderson,P.K.,Gill,R.D.(1982).Cox's regression model for counting process:A large sample study. Annals of Statistics,10,1100-1120.
3.Aalen,O.O.,Gjessing,H.K.(2001).Understanding the shape of the hazard rate:A process point of view.Statistical Science,16,1-22
4.L.J.Wei,D.Y.Lin,L.Weissfeld(1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association,Vol.84, No.408.(Dec,1989),pp.1065-1073.
5.David M Diez(2013).Survival Analysis in R.OpenIntro.org
6.Sy.Han Chiou,Sangwook Kang,Jun Yan(2014).Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee. Journal of Statistical Software,Vol 61,Issue 11,Nov,2014.
7.John Fox,Sanford Weisberg(2013). Cox Proportional-Hazards Regression for Survival Data in R. An Appendix to *An R Companion to Applied Regression, Second Edition*
8.Terry Therneau, Cynthia Crowson, Elizabeth Atkinson. Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. 2017
9.D.R.Cox,E.J.Snell.A General Definition of Residuals(1968). Journal of the Royal Statistical Society.Series B. Vol.30,No.2(1968),pp.248-275
10.Zhezhen Jin,D.Y.Lin,Zhiliang Ying.On least-squares regression with censored data. Biometrika(2006) , 93,1 ,pp.147-161
11.Testing Cox Model Assumptions in r:http://www.sthda.com/english/wiki/cox-model-assumptions
12.Survival Analysis in R:https://www.openintro.org/download.php?file=survival analysis in R
13.Predictions for a Cox Model:https://stat.ethz.ch/R-manual/R-devel/library/survival/html/predict.coxph.html
14.John Fox,Sanford Weisberg:Cox Proportional-Hazards Regression for Survival Data in R; url:https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf
15.Package 'timereg':https://cran.r-project.org/web/packages/timereg/timereg.pdf
16.Package 'magrittr':https://cran.r-project.org/web/packages/magrittr/magrittr.pdf
17.Package 'fitdistrplus':https://cran.r-project.org/web/packages/fitdistrplus/fitdistrplus.pdf
18.Package 'SurvMisc':https://cran.r-project.org/web/packages/SurvMisc/SurvMisc.pdf
19.Package 'ggplot2':https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
20.Package 'survminer':https://cran.r-project.org/web/packages/survminer/survminer.pdf
21.Package 'factoextra':https://cran.r-project.org/web/packages/factoextra/factoextra.pdf/
22.Package 'muhaz':https://cran.r-project.org/web/packages/muhaz/muhaz.pdf
23.Package 'survival':https://cran.r-project.org/web/packages/survival/survival.pdf
24.Package 'aftgee':https://cran.r-project.org/web/packages/aftgee/aftgee.pdf
25.L.Thomas,Eric M.Reyes:Tutorial:Survival Estimation for Cox Regression Models with Time-Varying Coefficients Using SAS and R. Journal of Statistical Software,Oct 2014,Volume 61,Code

Snippet 1.