9004 Final Project

Gansen Deng, Duo Xu, Chufan Wu

December 18, 2019

1 Introduction

1.1 Project objectives

Colon cancer is a type of cancer that begins in the large intestine (colon). It usually begins as small, noncancerous (benign) clumps of cells called polyps that form on the inside of the colon. Over time some of these polyps can become colon cancers. There are many treatments available to help control colon cancer[1]. Some of the treatments have relatively high toxicity and probably better curative effect. To explore the curative effect of different treatments on the control of colon cancer, we are going to utilize the built-in R dataset **Colon** to do survival analysis.

In this project, we are going to fit both AFT model and Cox PH model to find out the important covariates that influence the recurrence time and death time of colon cancer and to see if those important covariates are the same for two types of events. Besides, we also plan to make inference based on the models we fit, to compute median survival time for some patients with given covariates. In the end, we are also going to compare different models and to see which model is more appropriate for this dataset.

To sum up, the main goals of this project is to

- Detect the effect of different treatments on the recurrence time and death time
- Find and compare other important covariates for the event time of recurrence and death respectively
- After model selection, make inference about the survival probability of colon cancer based on the optimized model

1.2 Data description

The dataset we use is the built-in dataset **colon** in the **survival** package, which contains 16 variables in total. The dataset comes from one of the first successful trials of adjuvant chemotherapy for colon cancer. There are two records for each individual, one is for recurrence event and the other one is for death event. The brief description of this dataset is shown below:

	Description
id	id
study	1 for all patients
rx	Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
sex	0 for female, 1 for male
obstruct	obstruction of colon by tumour
perfor	perforation of colon
adhere	adherence to nearby organs
nodes	number of lymph nodes with detectable cancer
time	days until event or censoring
status	censoring status
differ	differentiation of tumour (1=well, 2=moderate, 3=poor)
extent	Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures)
surg	time from surgery to registration (0=short, 1=long)
node4	more than 4 positive lymph nodes
etype	event type: 1=recurrence,2=death

Table 1: Descrption for variables

1.3 Data preprocessing

Before fitting any model, we first remove all the observations with missing values to get a complete data for analysis, since there are only about 4% of observations with missing values. After that, we split the dataset into two parts by the event type so that we can dig in these two types of event separately. What's more, we also remove the variable **id**, **study** and **etype**, since they are irrelevant with our further analysis.

1.4 Division of work

- Introduction: Collaboration
- Exploratory data analysis: Chufan Wu
- AFT model analysis: Duo Xu
- Cox PH model analysis: Gansen Deng
- Summary: Chufan Wu

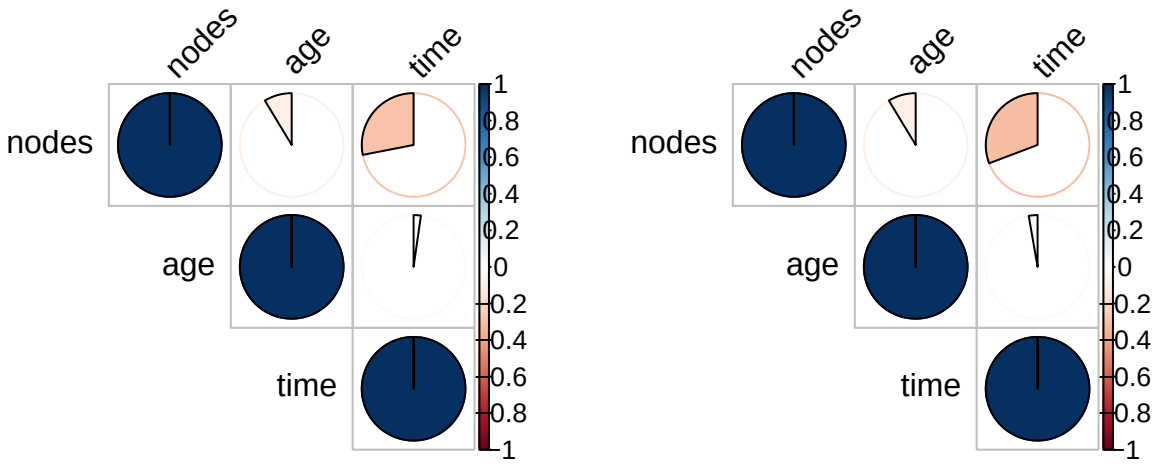
2 Exploratory data analysis

2.1 Correlation table

In the first step, we try to examine the relationship of the numeric variables in our dataset. If there exist high-correlated covariates, then we need to consider removing some of them from models or reducing the dimension of our dataset by PCA.

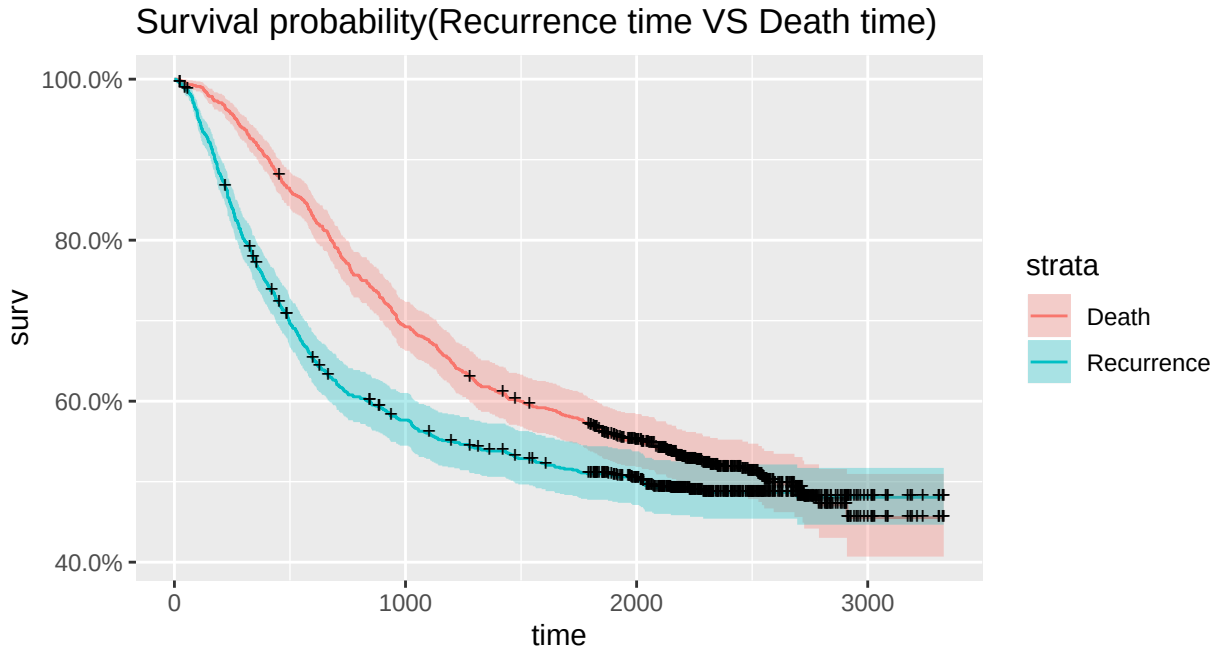
For realism sake, we have only considered 3 variables to be numeric, even though all variables except the **rx(treatment)** are numeric in the original dataset. They are **age**, **nodes(number of lymph nodes with detectable cancer)** and the response variable **time**. Since we have divided the original data set into two subsets(recurrence set and death set), here we need to plot two correlation tables for these two events respectively.

The two correlation tables we get are shown below. First we find that these two tables are almost the same, indicating the relationships between numeric variables in two subsets are highly similar. In each correlation table, only one pair of variables(**nodes** and **time**) show high correlation, which means multicollinearity doesn't exist in two subsets and **nodes** is probably an influential factor to the event time.



2.2 Recurrence VS Death

Next we explore whether the survival functions(probability) of recurrence event and death event are different. The plot below shows the Keplan-Meiev estimates $\hat{S}(t)$ for two types of events. From the plot we can see that the survival probability in death experiment is higher than that in recurrence experiment at the beginning. But then these two survival functions become closer to each other gradually and they overlap after about 2750 days. After that the survival probability in recurrence experiment keeps stable while the S(t) in death experiment goes on decreasing.



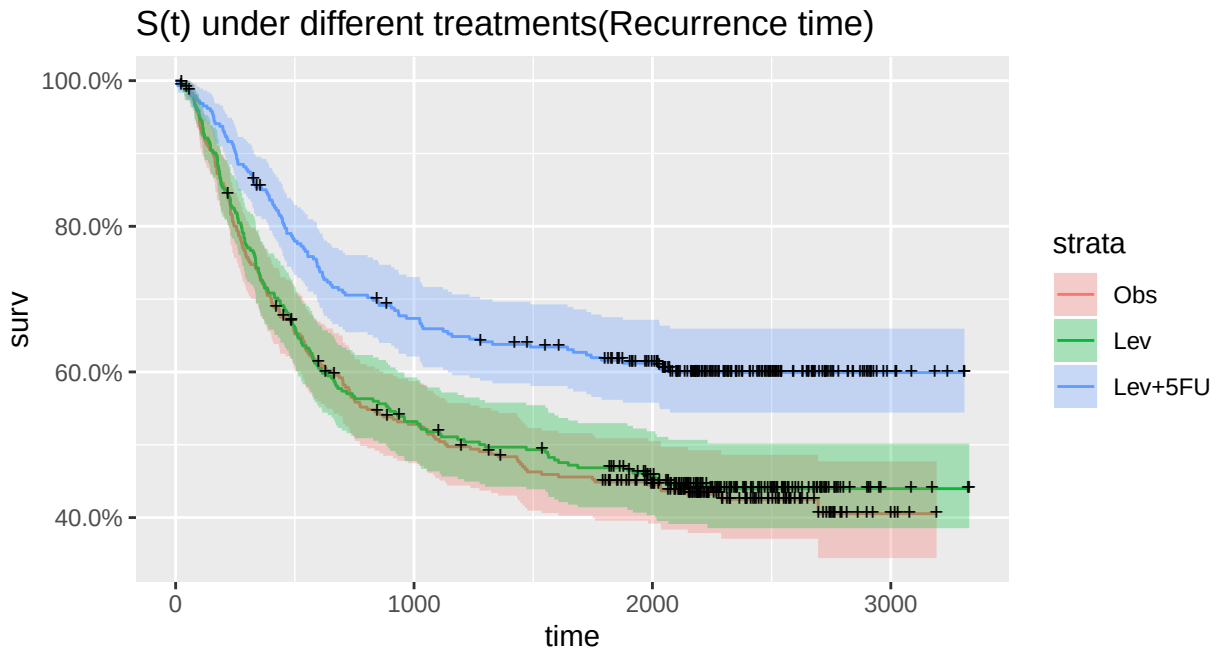
The pattern of these two estimated curves has provided plenty of information. Here we divide the time into two intervals. The first interval is from 0 to 2750 days, in which the survival probability in death experiment is higher than that in the recurrence experiment. This pattern is pretty consistent with the reality since except for censored data, people will face the threat of death only due to the recurrence of colon cancer. So it is almost sure that patients will not die before the recurrence of colon cancer, that is to say, the survival probability in death experiment is almost surely higher than the survival probability in recurrence experiment.

In the other interval(from day 2750 to the end of the trials), the survival probability in death experiment is clearly lower than that in recurrence experiment. This means patients may die even if without the recurrence of colon cancer. Therefore, other risk factors like viral infection should be taken into consideration to postpone the death time.

2.3 The effect of different treatments for recurrence and death

From the analysis above we know there is a difference between the survival probabilities of recurrence and death. Now we are going to detect the effect of treatments(**Obs**, **Lev** and **Lev+5FU**) on the two survival probabilities.

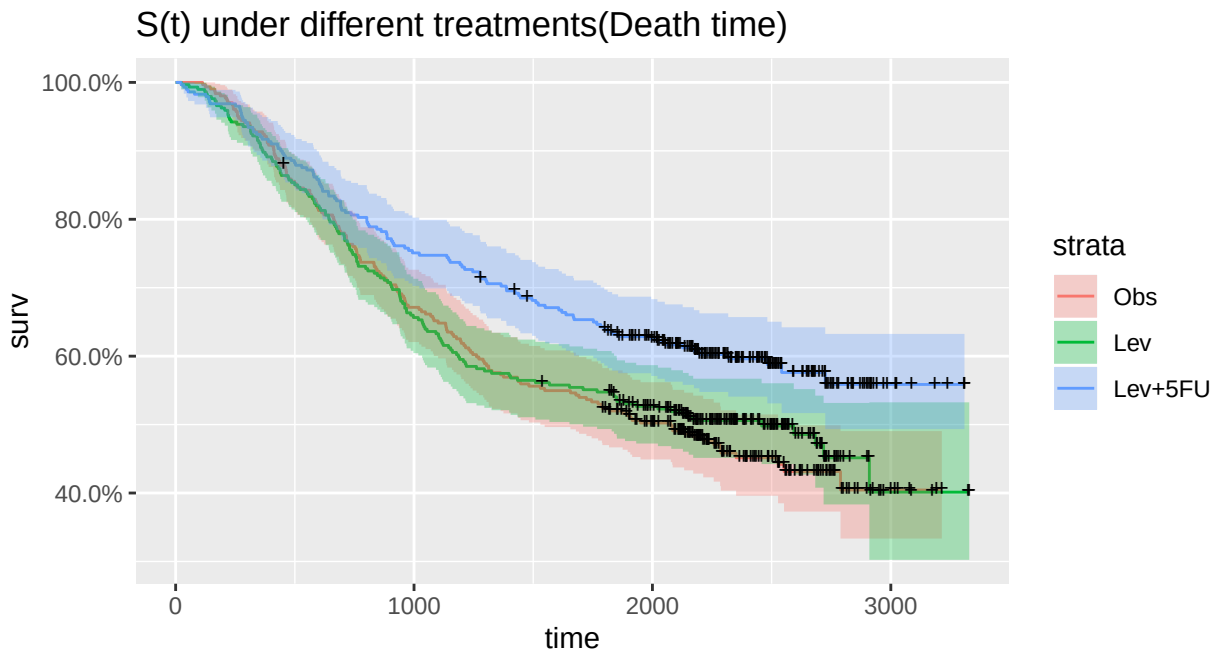
First, for the event of recurrence, we plot the estimated survival probabilities under three different treatments. The estimated S(t)'s along with their confidence intervals are almost the same for treatment **Obs** and **Lev**, which means the treatment **Lev** hardly affects the recurrence of colon cancer. On the contrary, treatment **Lev+5FU** has a significant positive influence on the colon cancer recurrence. The group of patients treated with **Lev+5FU** appear to have a 20% higher probability to survive, compared to the **Obs** group conditioning on the lifetime(3000 days).



Thus, we conclude that the combination of **Lev** and **5-FU** has a positive effect on the treatment of colon cancer while the single **Lev** treatment is not effective.

In addition, we also have the survival probability curves under three treatments for death event as follow. The information provided by this plot is slightly different from the result of the previous plot. In this plot, the survival probability under **Lev** treatment is a little bit higher than that of **Obs** treatment, indicating **Lev** helps postpone the death time to some extend. The **Lev+5FU** still has a positive effect on lowing down the probability of death. However, its effect is not so strong as that for the recurrence event. Under the **Lev+5FU** treatment, the survival probability has only increased by 15% compared to the **Obs** group,

which is lower than the increment(nearly 20%) for the recurrence event.

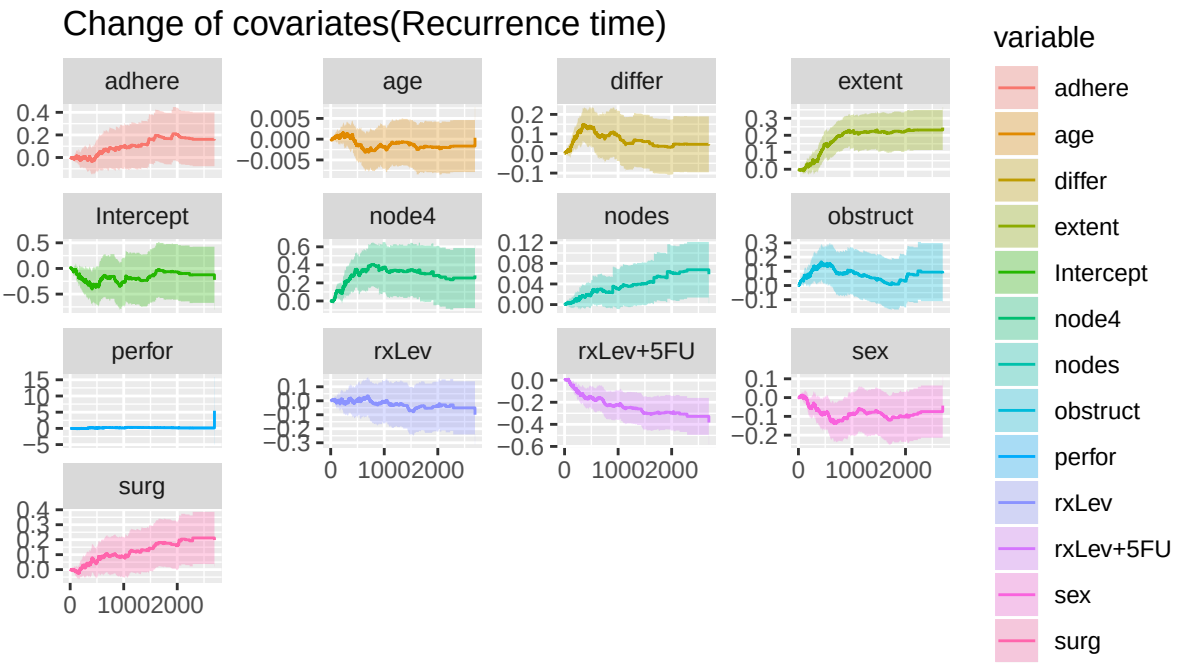


Thus, we concluded that for the death event, both **Lev** and **Lev+5FU** treatment could have a positive effect on the treatment of colon cancer in terms of preventing death caused by the cancer. The effect of **Lev** is relatively weak while the effect of **Lev+5FU** is rather strong but a little bit weaker than its effect in the recurrence event.

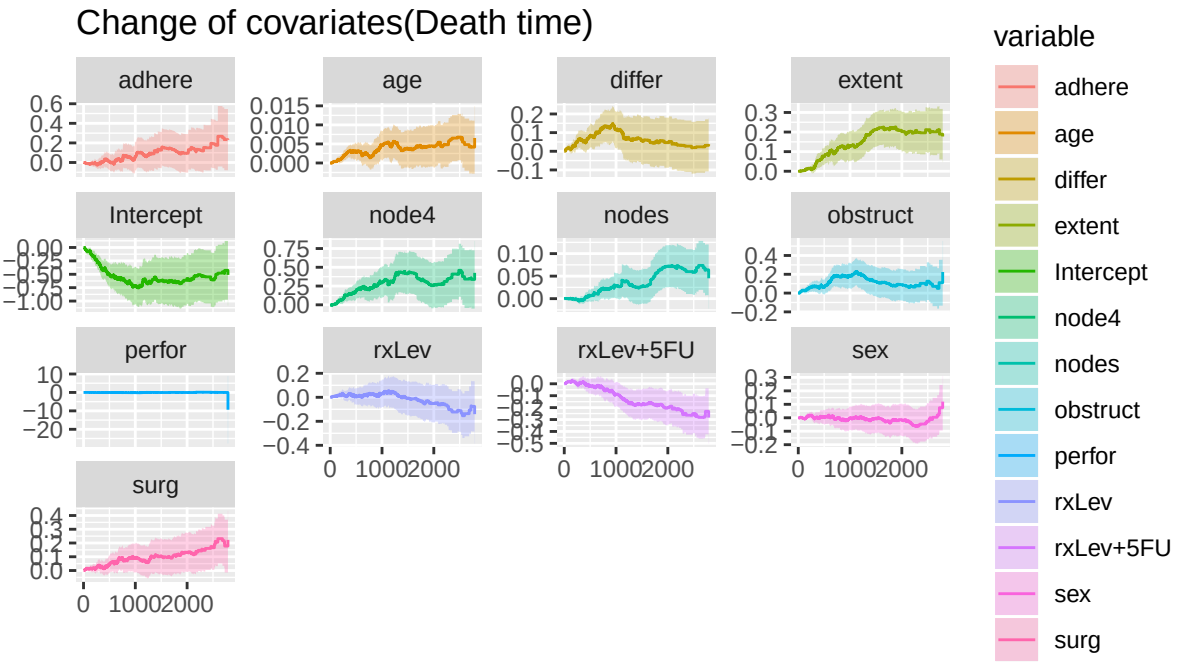
2.4 Covariates change

The last part of the EDA is to check whether the covariates in our data set are time-dependent. This analysis is specifically used to check whether the assumption of the Cox-PH model is satisfied. The key assumption of the Cox-PH model is that all covariates do not depend on time. So if this assumption is not valid for our dataset, then we should be careful about applying the normal(default) Cox-PH model.

The graph below shows the change of covariates with recurrence as the event. We need to pay attention to those time series whose slope has a steep change within a short period of time, which implies that the corresponding covariate changes dramatically and instantly. We notice that there are 4 plots in which the slope has high variation, which are **differ**, **extent**, **obstruct** and **sex**. Therefore, we think that these 4 covariates are time-dependent, and we had better remove these 4 variables or take some transformation before applying Cox-PH model.



Similarly, we repeat the analysis above with death as the event. In the following graph, we notice that the variation of slope in the plots of **extent**, **obstruct** and **sex** has become much smaller. Meanwhile, high slope variation remains in the plot of **differ**, and the slope in the plots of **nodes** and **node4** greatly fluctuates, which means these three variables are likely to be time-dependent. Therefore, we should pay attention to **differ**, **nodes** and **nodes4** when utilizing the Cox-PH model. If Cox-PH does not have a good performance, we may try to remove these three variables or use other statistical models.



In short, we can draw 4 conclusions from the exploratory data analysis above:

1. The numeric variables in our dataset are low-correlated.

2. Other risk factors except the colon cancer may lead to death of patients.

3. The **Lev+5FU** treatment has a significantly positive effect on the treatment of colon cancer, while singe **Lev** treatment does not.

4. Some covariates are probably time-dependent variables, which will influence the performance of the Cox-PH model.(Recurrence: **differ**, **extent**, **obstruct** and **sex**; Death: **differ**, **nodes** and **nodes4**)

3 AFT model

3.1 Introduction

In the statistical area of survival analysis, an accelerated failure time model (AFT model) is a parametric regression model that assumes the effect of covariates is to accelerate or decelerate the life time of an event by some constant[2]. The general expression of it is:

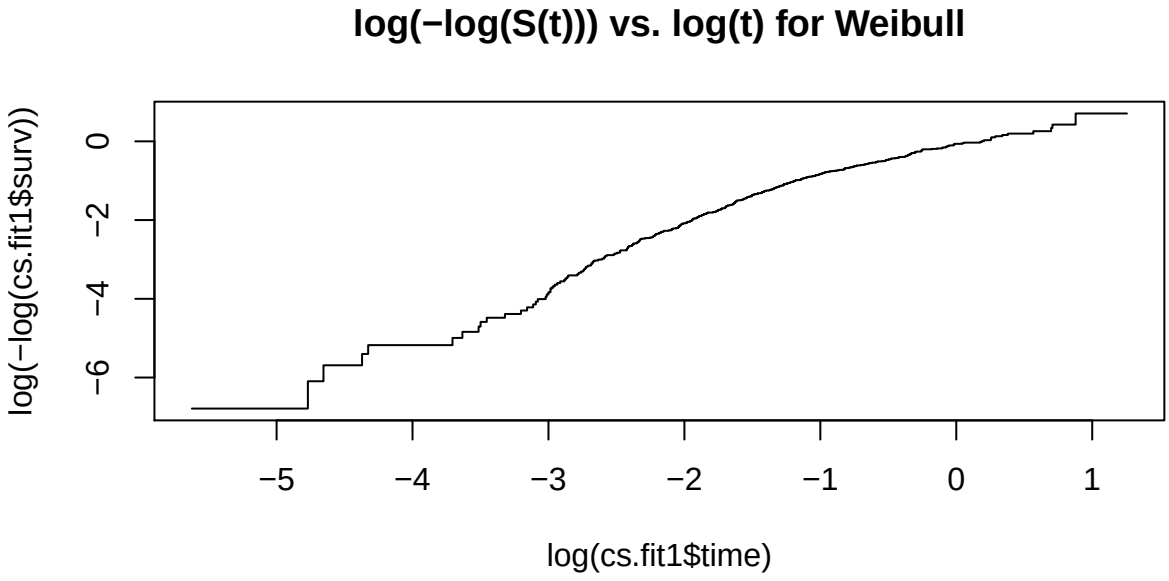
$$log(T) = X^T\beta + b\epsilon$$

X^T is the covariate vector, β is the parameter we want to estimate, which shows the effect of covariates. Besides, b is the scale and ϵ is the error term, we usually assume it follows a certain distribution.

3.2 Analysis on death events

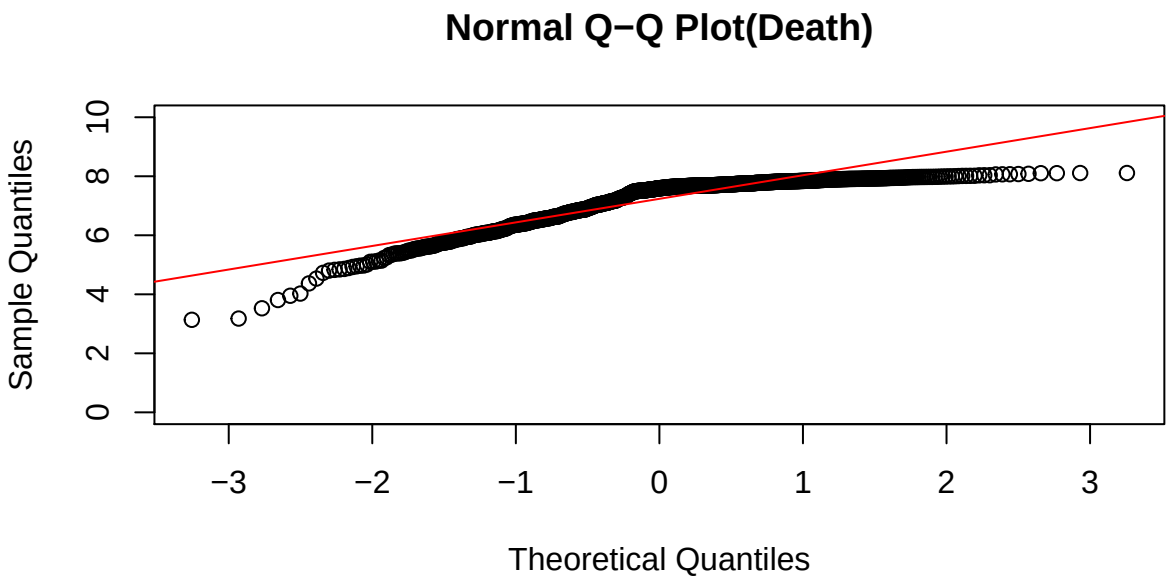
3.2.1 Select the distribution for errors

First of all, we should determine the distribution of errors ϵ. There are three kinds of distributions that often used in survival analysis, which are log-logistic distribution, weibull distribution and log-normal distribution. Firstly, we plot $log(-log(S(t))$ against $log(t)$ to test if the errors follow the weibull distribution.

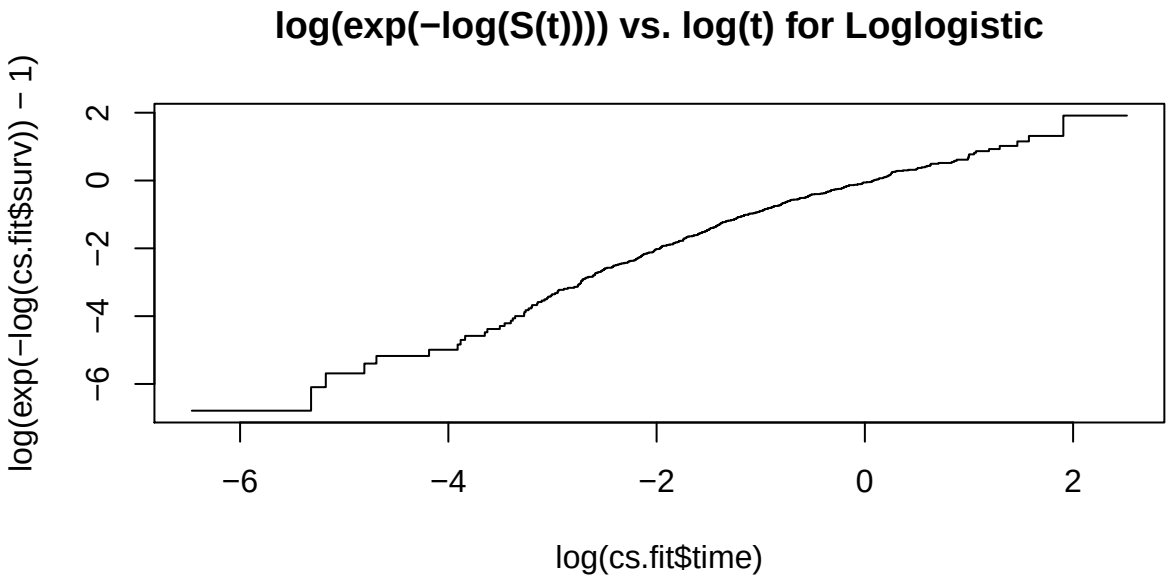


It can be seen that the linear pattern is quite evident from the plot, so we think Weibull distribution is an appropriate choice for us.

Then we check the log-normal distribution by plotting Q-Q plot for $log(T)$



It can be seen that some of the points are far away from the line, which means that log-normal is not proper for this dataset.
In the end, we test the log-logistic distribution as follow:



We can see from the plot that the linear pattern is more obvious for log-logistic distribution than weibull distribution, so we decide to choose the log-logistic distribution for the errors.

3.3 Model fitting

The second step is to fit our model, and remove those covariates that are not significant. We set the significance level to 0.05 and after removing those insignificant variables one by one, we get our final model as follow:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ rx + age + obstruct +
##     nodes + extent + surg + node4, data = colon2, dist = "loglogistic")
##                Value Std. Error     z       p
## (Intercept) 10.20514    0.44155 23.11 < 2e-16
## rxLev        0.03046    0.11402  0.27   0.789
## rxLev+5FU    0.29887    0.11957  2.50   0.012
## age         -0.00886    0.00412 -2.15   0.032
## obstruct    -0.31525    0.12249 -2.57   0.010
## nodes       -0.04377    0.01899 -2.30   0.021
## extent      -0.53732    0.11598 -4.63 3.6e-06
## surg        -0.23809    0.10682 -2.23   0.026
## node4       -0.76449    0.15916 -4.80 1.6e-06
## Log(scale)  -0.29593    0.04164 -7.11 1.2e-12
##
## Scale= 0.744
##
## Log logistic distribution
## Loglik(model)= -3844.4   Loglik(intercept only)= -3919.7
##  Chisq= 150.74 on 8 degrees of freedom, p= 1.4e-28
## Number of Newton-Raphson Iterations: 4
## n= 888
```

We can see from the summary output that the final model we get for the death events is:

$$\begin{aligned} log(T) = y = 10.20514 + 0.03046A + 0.29887B - 0.00886 * age - 0.31525 * obstruct - 0.04377 * nodes \\ - 0.53732 * extent - 0.23809 * surg - 0.76449 * node4 - 0.744 * \epsilon \end{aligned}$$

With $A = I_{LEV}$, $B = I_{LEV+5FU}$, and ϵ follows the standard logistic distribution.

Intuitively, only the treatment can alleviate the risk of death, so their coefficients are positive; and for those other covariates, a larger number(including 1 for binomial variables) will reduce the expected lifespan, so their coefficients for them are negative.

3.4 Inference

3.4.1 Point estimation and CI for quantile for T|X

Since we only have the variance for $\hat{u}$ and $\hat{logb}$, we need to use delta method to obtain the variance matrix for $\hat{u}$ and $\hat{b}$

$$Var\begin{pmatrix}\hat{u}\\ \hat{b}\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & \hat{b}\end{pmatrix} Var\begin{pmatrix}\hat{u}\\ \hat{logb}\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & \hat{b}\end{pmatrix}$$

First we can obtain the 95% CI for the quantiles for t|X by delta method.

$$Var(\hat{y_p}) = \begin{pmatrix}X & \epsilon_p\end{pmatrix} Var\begin{pmatrix}\hat{u}\\ \hat{b}\end{pmatrix}\begin{pmatrix}X\\ \epsilon_p\end{pmatrix}$$

Now suppose there is a 60-year-old male patient, who has just taken an examination for colon cancer. The result shows that his colon is obstructed by tumour and there are 10 lymph nodes with detectable cancer, which has spreaded to muscle. The man has already spent a lot of time in hospital from surgery to registration and now he plans to take the **Lev+5FU** treatment for the colon cancer.

Now the man wants to know his median survival time so that he can make a plan for travelling around the world in the last period of his life. Besides, he also want to know what he can do to survive longer.

So now we use the AFT model to help that patient to solve his problem and the median death time we get is 1264.084 days, with the corresponding 95% CI [833.1525, 1917.9055]. Besides, if the patient wants to survive longer, then he might had better take the **Lev+5FU** treatment and control the extent of local spread, while tackling the obstruction of colon is also important.

3.4.2 Point estimation and CI for S(t|X)

Since we have

$$\hat{\epsilon} = [log(T) - (10.20514 + 0.03046A + 0.29887B - 0.00886 * age - 0.31525 * obstruct - 0.04377 * nodes - 0.53732 * extent - 0.23809 * surg - 0.76449 * node4)]/\hat{b}$$

The variance of $\hat{\epsilon}$ can also be obtained by delta method. In our model, the residual $\hat{\epsilon}$ follows a standard logistic distribution, then we have

$$S(t) = \frac{1}{1 + e^{-\hat{\epsilon}}}$$

.

Then for 95% CI for S(t|X), we can first obtain the 95% CI for $\hat{\epsilon}$ and then just plug it into the survival function for a standard logistic distribution.

For example, the probability of that male patient living for more than 1500 days is

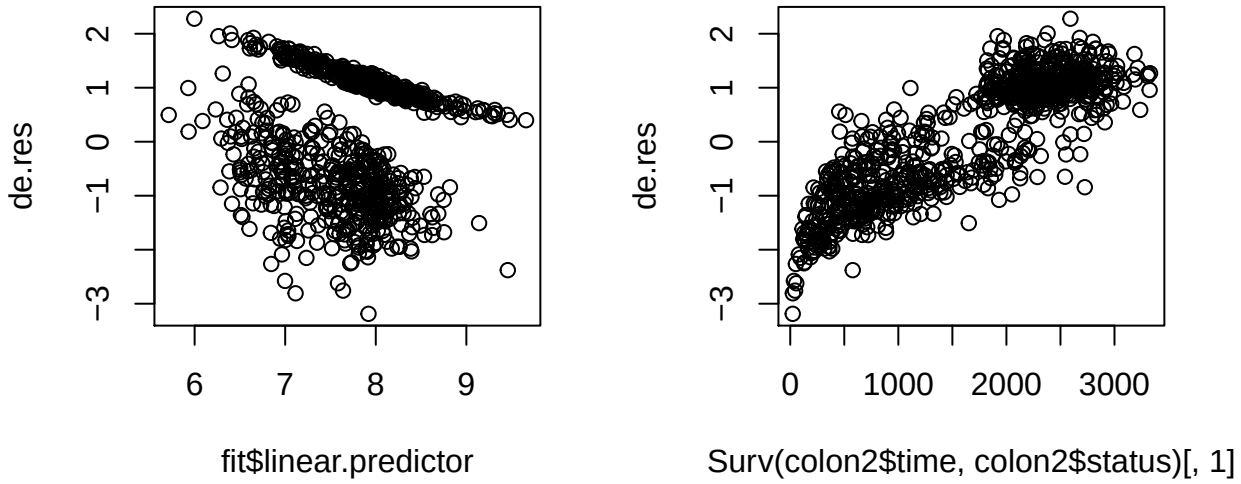
```
##           [,1]
## [1,] 0.5572595
```

```
## [1] 0.4178573 0.6881897
```

That is, the 95% CI for the probability for a patient with the above conditions will live to 1500 days is [0.4178573, 0.6881897], and the point estimation for S(1500|X) is 0.5572595.

3.4.3 Model assessment

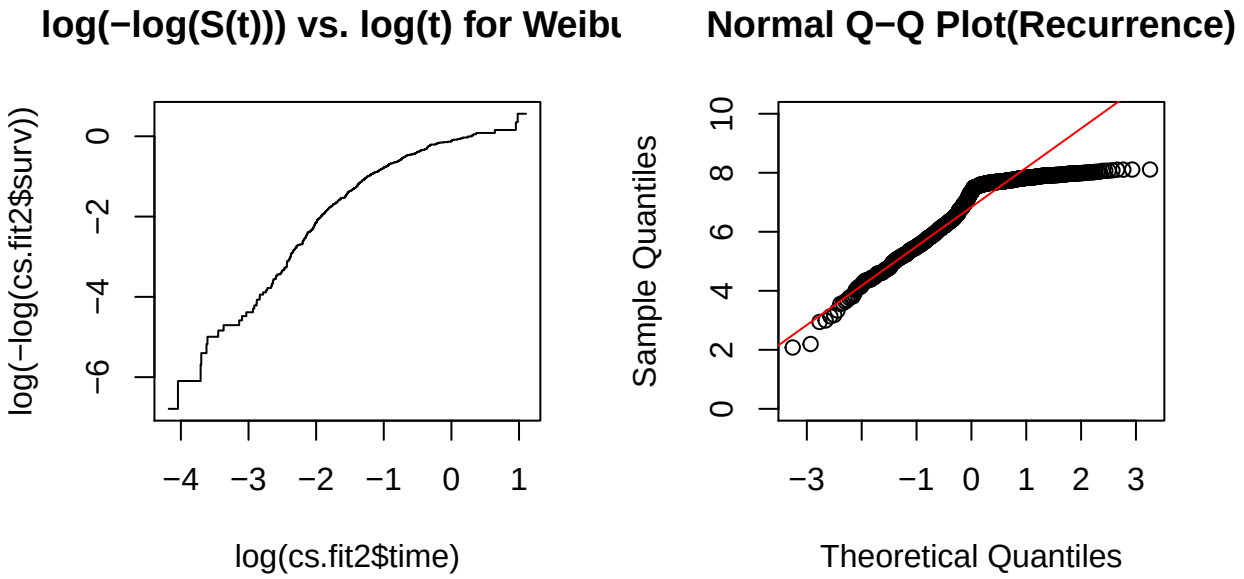
In the end, we use the deviance residual to test the randomness of our AFT model.



It can be seen that there exist linear patterns in above two plots, which means that the model is not quite appropriate for this dataset. So we think it's necessary for us to try to use Cox PH model for this dataset.

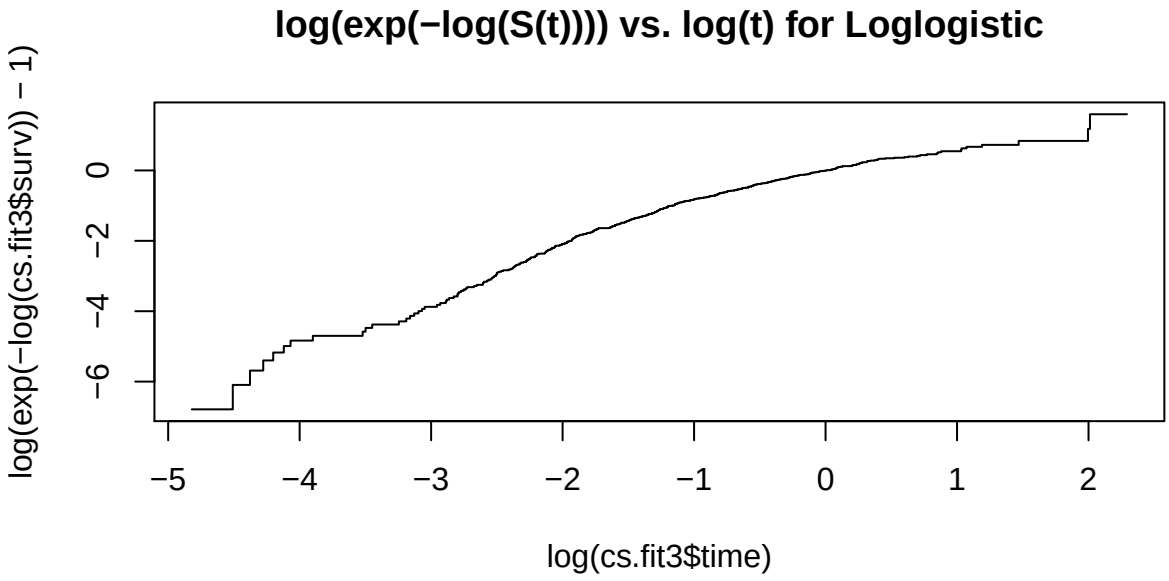
3.5 Analysis on recurrence events

For another group of data, the recurrence time for patients, we do the same test for the three distributions. The plots for weibull distribution and log-normal distributions are as follow:



It can be seen that neither of these two distributions are appropriate, though the plot of Weibull distribution looks better than that of log-normal distribution.

Thus, we also test for the log-logistic distribution as follow:



We can see the linear pattern is evident for log-logistic distribution and thus we still choose log-logistic for errors.

Similarly, we drop those insignificant covariates step by step and the final model we get is as follow:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ rx + obstruct + nodes +
##     extent + surg + node4, data = colon1, dist = "loglogistic")
##               Value Std. Error     z       p
## (Intercept) 10.1068     0.4922 20.53 < 2e-16
## rxLev        0.0621     0.1589  0.39   0.696
## rxLev+5FU    0.7185     0.1686  4.26 2.0e-05
## obstruct    -0.3687     0.1704 -2.16   0.031
## nodes       -0.0624     0.0276 -2.26   0.024
## extent      -0.7654     0.1609 -4.76 2.0e-06
## surg        -0.3512     0.1496 -2.35   0.019
## node4       -0.9550     0.2264 -4.22 2.5e-05
## Log(scale)   0.0490     0.0404  1.21   0.225
##
## Scale= 1.05
##
## Log logistic distribution
## Loglik(model)= -3837.5   Loglik(intercept only)= -3912.7
##  Chisq= 150.36 on 7 degrees of freedom, p= 3.4e-29
## Number of Newton-Raphson Iterations: 4
## n= 888
```

From the summary output, we can see that our final model is:

$$log(T) = y = 10.1068 + 0.0621A + 0.7185B - 0.3687 * obstruct - 0.0624 * nodes - 0.7654 * extent - 0.3512 * surg - 0.9550 * node4 + 1.05\epsilon$$

With $A = I_{LEV}$, $B = I_{LEV+5FU}$, and ϵ follows the standard logistic distribution.

3.5.1 Conclusion

It can be seen from the two models for death and recurrence events that the coefficient for **Lev** only is around 0.05, which means that conditioning on other covariates, a patient with **Lev** will increase by $e^{0.05} - 1 = 0.051$ times. Compared to the coefficient with **Lev+5FU**, which is 0.29 in death data and 0.71 in recurrence data, which means given other conditions the same, a patient with **Lev+5FU** treatment will increase much more than a patient with no treatment(observation) and **Lev** only. So we can say that the **Lev+5FU** treatment will significantly increase the expected life time for a patient with colon cancer, while **Lev** can only also increase the life time slightly.

On the other hand, we can see that the covariate age is in the death model but it is not in the recurrence model, which means age may only have effect on the death time for patients with colon cancer, but have no effect on recurrence time.

We can also see that the coefficients for other covariates are all negative, which means a larger value in these covariates will lead to a larger probability of death or recurrence.

In conclusion, for the death time, a younger patient taking the **Lev+5Fu** treatment, colon not being constructed by tumour, having less lymph nodes with detectable cancer(especially less than 4), whose extent of local spread is 1(submucosa) and time from surgery to registration is short will be expected to live longer.

For the recurrence time, a patient taking the **Lev+5Fu** treatment, colon not being constructed by tumour, having less lymph nodes with detectable cancer(especially less than 4), whose extent of local spread is 1(submucosa) and time from surgery to registration is expected to have a lower risk of recurrence and longer expected recurrence time.

4 Cox PH model

4.1 Introduction

In the model assessment part of the AFT model, we find that there is a linear pattern between deviance residual and lifetime, which means that AFT model might not be appropriate for this data. So we also try to fit a Cox PH model to see whether it's more appropriate or not.

Proportional hazards(PH) models are a class of survival models in statistics, which are usually alternatives of AFT models. In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate[3]. Its general form is:

$$h(t|X) = \psi(X)h_0(t)$$

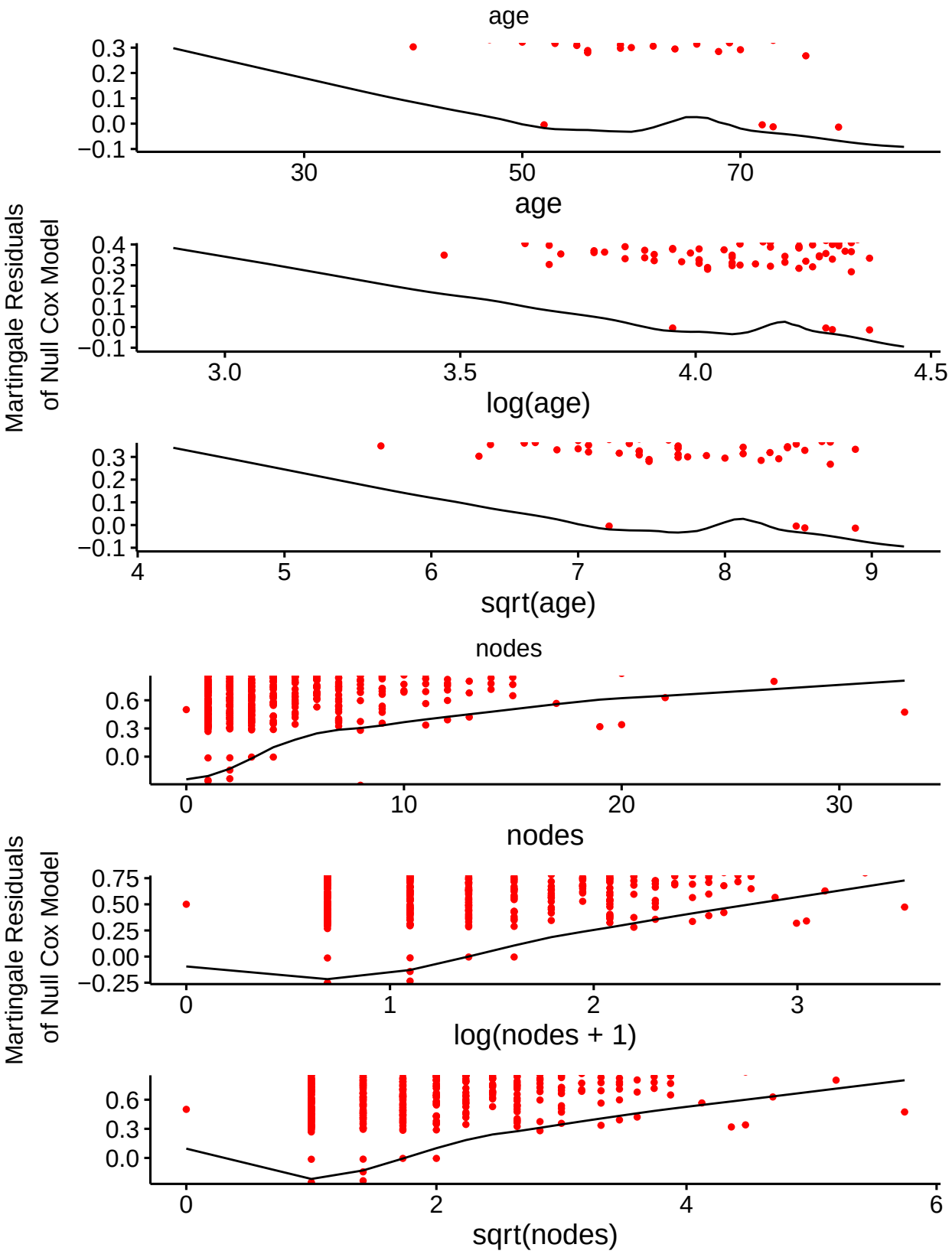
Particularly, when $\psi(X) = exp(X^T\beta)$, we call the model Cox PH model.

4.2 Analysis on recurrence events

4.2.1 Check the nonlinearity for continuous covariates

Before fitting the model, we first need to check the nonlinearity of continuous covariates to see if we need to employ any data transformation on them.

In this dataset, there are two continuous covariates **age** and **nodes**. For each covariate, we plot three different functional forms$(x, log(x), \sqrt{x})$ of it against the martingale residuals of null cox proportional hazards model[4]. Then we fit lines with those points to see if those lines are linear. The plots we obtain are shown below:



From above plots, we can see that the linear patterns in $log(age)$ and $nodes$ are most significant. Thus, we

are going to take log transformation on *age* while remain *nodes* as its original form to fit the model.

4.2.2 Use AIC for variable selection

Then we fit a model including all the covariates, which is shown below:

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + sex + log(age) + obstruct +
##     perfor + adhere + nodes + differ + extent + surg + node4,
##     data = colon1)
##
##   n= 888, number of events= 446
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.02520   0.97512  0.11094 -0.227 0.820314
## rxLev+5FU -0.49892   0.60719  0.12185 -4.095 4.23e-05 ***
## sex       -0.13844   0.87071  0.09584 -1.445 0.148593
## log(age)  -0.17630   0.83837  0.21302 -0.828 0.407882
## obstruct   0.19271   1.21253  0.11927  1.616 0.106151
## perfor     0.20593   1.22867  0.25659  0.803 0.422233
## adhere     0.16095   1.17462  0.12969  1.241 0.214588
## nodes      0.03803   1.03876  0.01508  2.521 0.011696 *
## differ     0.15340   1.16579  0.09791  1.567 0.117187
## extent     0.45202   1.57149  0.11896  3.800 0.000145 ***
## surg       0.24070   1.27215  0.10412  2.312 0.020787 *
## node4      0.59111   1.80599  0.14116  4.188 2.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9751     1.0255    0.7846     1.212
## rxLev+5FU    0.6072     1.6469    0.4782     0.771
## sex          0.8707     1.1485    0.7216     1.051
## log(age)     0.8384     1.1928    0.5522     1.273
## obstruct     1.2125     0.8247    0.9598     1.532
## perfor       1.2287     0.8139    0.7431     2.032
## adhere       1.1746     0.8513    0.9110     1.515
## nodes        1.0388     0.9627    1.0085     1.070
## differ       1.1658     0.8578    0.9622     1.412
## extent       1.5715     0.6363    1.2447     1.984
## surg         1.2721     0.7861    1.0373     1.560
## node4        1.8060     0.5537    1.3695     2.382
##
## Concordance= 0.666  (se = 0.013 )
## Likelihood ratio test= 136.9  on 12 df,   p=<2e-16
## Wald test            = 141.6  on 12 df,   p=<2e-16
## Score (logrank) test = 150  on 12 df,   p=<2e-16
```

From the summary output we can see that there are many variables with p-values greater than 0.05. Thus, we think it's necessary to use AIC to do variable selection and after that we get the following model:

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + obstruct + nodes +
##     differ + extent + surg + node4, data = colon1)
##
##   n= 888, number of events= 446
```

```
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.02711   0.97325  0.11078 -0.245   0.8067
## rxLev+5FU -0.49619   0.60885  0.12174 -4.076 4.59e-05 ***
## obstruct   0.22007   1.24616  0.11784  1.867   0.0618 .
## nodes      0.03697   1.03766  0.01492  2.478   0.0132 *
## differ     0.15889   1.17221  0.09777  1.625   0.1041
## extent     0.47829   1.61331  0.11946  4.004 6.24e-05 ***
## surg       0.23189   1.26098  0.10397  2.230   0.0257 *
## node4      0.60220   1.82613  0.13993  4.303 1.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9733     1.0275    0.7833    1.2093
## rxLev+5FU    0.6088     1.6425    0.4796    0.7729
## obstruct     1.2462     0.8025    0.9892    1.5699
## nodes        1.0377     0.9637    1.0078    1.0684
## differ       1.1722     0.8531    0.9678    1.4198
## extent       1.6133     0.6198    1.2765    2.0389
## surg         1.2610     0.7930    1.0285    1.5460
## node4        1.8261     0.5476    1.3881    2.4024
##
## Concordance= 0.666  (se = 0.013 )
## Likelihood ratio test= 131.7  on 8 df,   p=<2e-16
## Wald test            = 136.8  on 8 df,   p=<2e-16
## Score (logrank) test = 144.7  on 8 df,   p=<2e-16
```

From the model we can see that the covariates **sex**, **perfor**, **adhere** and **age** are removed from the model.

4.2.3 Stratified Cox PH model

Considering that **sex**, **perfor** and **adhere** are categorical variables, we can incorporate them into our model by using the stratified Cox PH model. The "stratified Cox model" is a modification of the Cox PH model that allows for control by "stratification" of a predictor that does not satisfy the PH assumption.[5]

The stratified Cox PH model we get is as follow:

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + strata(sex) + obstruct +
##     strata(perfor) + strata(adhere) + nodes + differ + extent +
##     surg + node4, data = colon1)
##
##   n= 888, number of events= 446
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.02365   0.97663  0.11158 -0.212 0.832142
## rxLev+5FU -0.49965   0.60674  0.12231 -4.085 4.40e-05 ***
## obstruct   0.25230   1.28699  0.12051  2.094 0.036293 *
## nodes      0.03978   1.04058  0.01519  2.619 0.008831 **
## differ     0.18013   1.19737  0.09860  1.827 0.067727 .
## extent     0.45597   1.57770  0.12022  3.793 0.000149 ***
## surg       0.23905   1.27004  0.10444  2.289 0.022084 *
## node4      0.57259   1.77285  0.14199  4.033 5.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9766     1.0239    0.7848    1.2154
## rxLev+5FU    0.6067     1.6482    0.4774    0.7711
## obstruct     1.2870     0.7770    1.0162    1.6299
## nodes        1.0406     0.9610    1.0101    1.0720
## differ       1.1974     0.8352    0.9870    1.4527
## extent       1.5777     0.6338    1.2465    1.9969
## surg         1.2700     0.7874    1.0350    1.5585
## node4        1.7729     0.5641    1.3422    2.3417
##
## Concordance= 0.674  (se = 0.013 )
## Likelihood ratio test= 128.7  on 8 df,   p=<2e-16
## Wald test            = 134.5  on 8 df,   p=<2e-16
## Score (logrank) test = 142  on 8 df,   p=<2e-16
```

We can see that all the coefficients except **rxLeV** have a p-value less than 0.1. But since **rxLeV+5FU** is significant, so we cannot remove the covariate **rx** from our model.

4.2.4 Test for proportionality

However, when we use the Schonefeld residual to verify the proportionality of the model, we get the following result:

```
##                rho   chisq        p
## rxLev     -0.03092  0.4472 5.04e-01
## rxLev+5FU  0.02567  0.2966 5.86e-01
## obstruct  -0.11825  6.7041 9.62e-03
## nodes      0.06311  1.4568 2.27e-01
## differ    -0.16919 14.2525 1.60e-04
## extent     0.00775  0.0277 8.68e-01
## surg       0.05796  1.5156 2.18e-01
## node4     -0.14030  8.1248 4.37e-03
## GLOBAL          NA 31.9372 9.56e-05
```

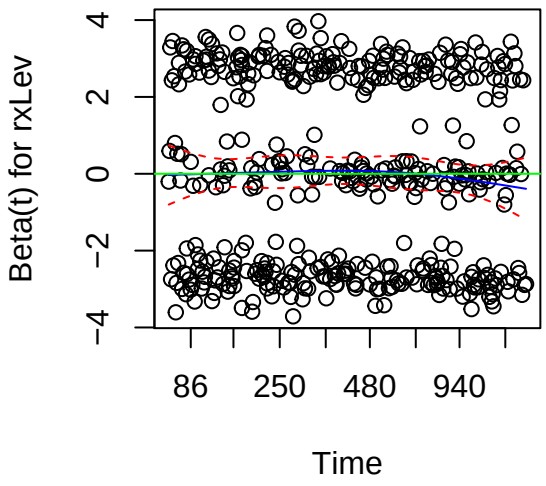
From the test result, we can see that the overall p-value of the model is $9.56e-05$, which is smaller than 0.05. It means that the assumption of proportionality is not true in this model.

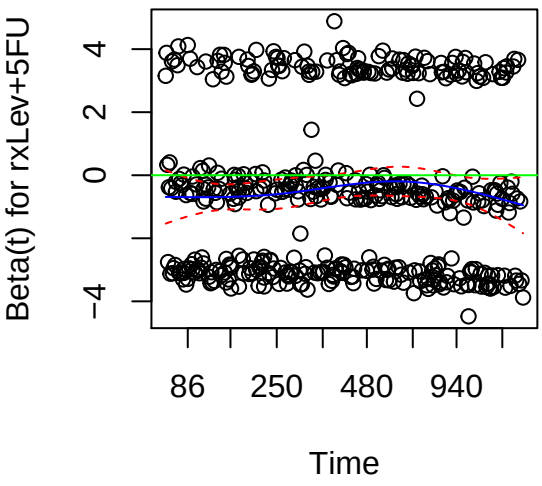
Therefore, we decide to also stratify those covariates with p-values greater than 0.05. Similarly, we do the same test on the new model to test the proportionality. The test result is shown below:

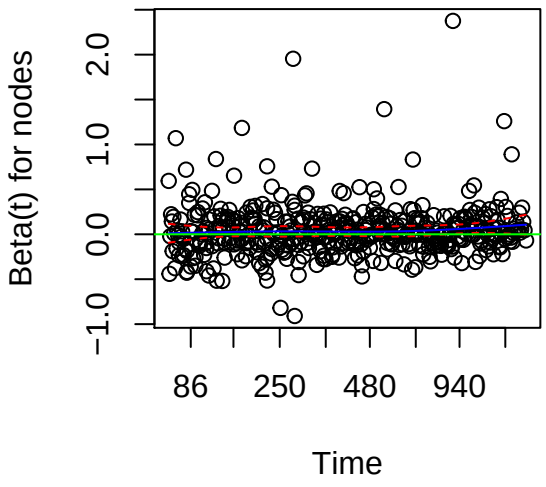
```
##               rho chisq     p
## rxLev     -0.0342 0.523 0.469
## rxLev+5FU  0.0160 0.112 0.738
## nodes      0.0309 0.337 0.561
## extent     0.0187 0.165 0.685
## surg       0.0677 2.041 0.153
## GLOBAL         NA 3.517 0.621
```

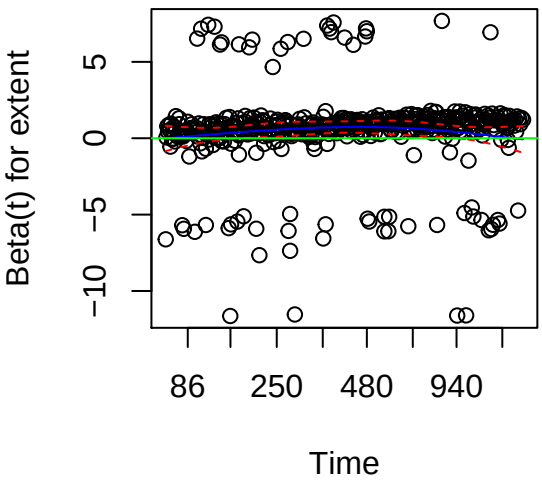
We can see from the result that the p-values of all covariates and the overall model are all greater than 0.05, which means that the proportionality assumption holds for the new model.

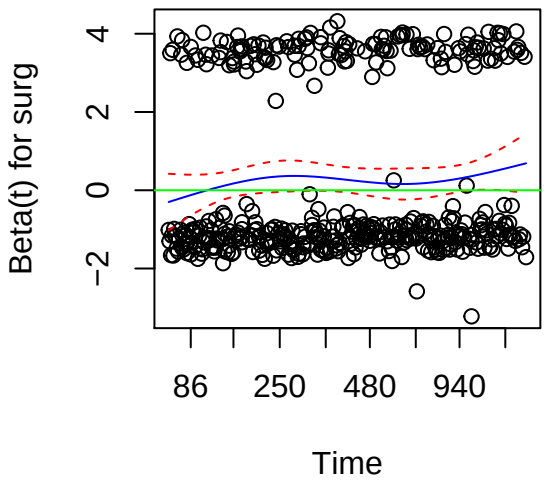
Besides, we can also plot the Schonefeld residual for each covariate:


Beta(t) for rxLev
4
2
0
−2
−4
86
250
480
940
Time


Beta(t) for rxLev+5FU
4
2
0
−4
86
250
480
940
Time


Beta(t) for nodes
2.0
1.0
0.0
−1.0
86
250
480
940
Time

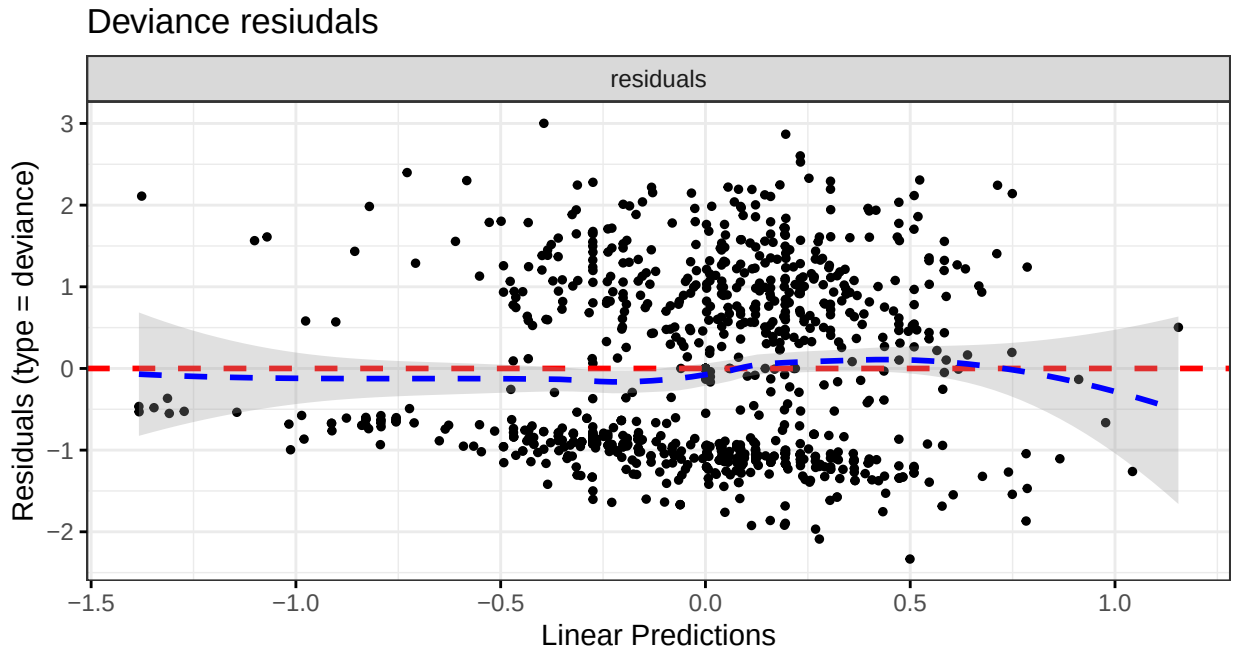

Beta(t) for extent
5
0
−5
−10
86
250
480
940
Time



In the plots above, the red dash lines are the 95% confidence limits for the fitted residuals and the green line is the $y = 0$ line. We can see that for the coefficients of **rxLev**, **nodes** and **surg**, the green line is entirely within the boundary of red dash lines, while for the other two coefficients, most parts of the green lines are in the boundary of red dash lines. Thus, these plots also show that we can accept the proportionality assumption in this model.

4.2.5 Check the exponential link

In addition, we also plot the deviance residuals against the linear predictors to check if the exponential link is true in this model[4].



In the plot we can see that there is no pattern between deviance residuals and linear predictors and the deviance residuals also have a mean close to 0. It means that the exponential link assumption is also true in this model.

4.2.6 Final model

To sum up, the final Cox PH model we get for the recurrence type event is:

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + strata(sex) + strata(obstruct) +
##     strata(perfor) + strata(adhere) + nodes + strata(differ) +
##     extent + surg + strata(node4), data = colon1)
##
##   n= 888, number of events= 446
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.10935   0.89641  0.11779 -0.928 0.353212
## rxLev+5FU -0.46988   0.62508  0.12596 -3.731 0.000191 ***
## nodes      0.03687   1.03756  0.01602  2.302 0.021333 *
## extent     0.55411   1.74039  0.12985  4.267 1.98e-05 ***
## surg       0.27761   1.31997  0.11006  2.522 0.011660 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.8964     1.1156    0.7116    1.1292
## rxLev+5FU    0.6251     1.5998    0.4883    0.8001
## nodes        1.0376     0.9638    1.0055    1.0706
## extent       1.7404     0.5746    1.3493    2.2448
## surg         1.3200     0.7576    1.0638    1.6377
##
## Concordance= 0.638  (se = 0.018 )
## Likelihood ratio test= 48.41  on 5 df,   p=3e-09
## Wald test            = 45.99  on 5 df,   p=9e-09
## Score (logrank) test = 46.22  on 5 df,   p=8e-09
```

The model expression is

$$h(T|X) = h_0(t)exp[-0.10935*rxLev-0.46988*(rxLev+5FU)+0.03687*nodes+0.55411*extent+0.27761*surg]$$

Note that $h_0(t)$ are different for different strata.

4.3 Analysis on death events

Similarly, we can use the same procedure to build a Cox PH model for death events and the final model we get for that is:

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + strata(sex) + log(age) +
##     strata(obstruct) + strata(perfor) + strata(adhere) + log(nodes +
##     1) + strata(differ) + extent + surg + strata(node4), data = colon2)
##
##   n= 888, number of events= 430
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev          -0.1059    0.8995   0.1218 -0.870 0.384285
```

```
## rxLev+5FU      -0.3017    0.7396   0.1262 -2.391 0.016803 *
## log(age)        0.3838    1.4678   0.2408  1.594 0.111019
## log(nodes + 1)  0.4637    1.5899   0.1359  3.412 0.000645 ***
## extent          0.4811    1.6179   0.1269  3.791 0.000150 ***
## surg            0.2462    1.2792   0.1121  2.198 0.027984 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## rxLev             0.8995     1.1118    0.7085    1.1419
## rxLev+5FU         0.7396     1.3521    0.5775    0.9471
## log(age)          1.4678     0.6813    0.9156    2.3532
## log(nodes + 1)    1.5899     0.6290    1.2181    2.0752
## extent            1.6179     0.6181    1.2616    2.0747
## surg              1.2792     0.7817    1.0270    1.5934
##
## Concordance= 0.638  (se = 0.02 )
## Likelihood ratio test= 41.74  on 6 df,   p=2e-07
## Wald test            = 39.9  on 6 df,   p=5e-07
## Score (logrank) test = 40.16  on 6 df,   p=4e-07
```

The model expression is

$$h(T|X) = h_0(t)exp[-0.1059 * rxLev - 0.3017 * (rxLev + 5FU) + 0.3838 * log(age) + 0.4637 * log(nodes + 1) + 0.4811 * extent + 0.2462 * surg]$$

Note that $h_0(t)$ are different for different strata.

The proportionality checking result is as follow and the other procedures are omitted here. The codes for them are shown in the appendix.

```
##                     rho  chisq     p
## rxLev           -0.09496 3.8091 0.051
## rxLev+5FU       -0.07557 2.4040 0.121
## log(age)        -0.07223 2.5286 0.112
## log(nodes + 1)   0.02337 0.2269 0.634
## extent          -0.07461 2.2552 0.133
## surg             0.00888 0.0327 0.856
## GLOBAL                NA 9.7376 0.136
```

We can see from the test result that the p-values of all covariates and the overall model are all greater than 0.05, which means that the proportionality assumption holds for the new model.

4.4 Inference based on the models

Same as above, we are still going to predict the median survival time for that 60-year-old male patient. In order to get the median survival time, we write a function called **MST** in R, which returns the probability that a patient might live longer than a given time. And then we use the **uniroot** function in R to solve for the time such that $S(t) = 0.5$. The result we get is 960.9999, which is a bit samller than that we get from the AFT model. It means that the patient can plan a 961-day trip to travel around the world.

Besides, according to the summary output of the model on death events, if the patient want to survive longer, he can take the "reLev+5FU" treatment instead of "reLev", which can decrease his risk of death by $1 - exp(0.1059 - 0.3017) = 17.8$. In addition, it would also be helpful for him to control the extent of local spread. What's more, the decrements of lymph nodes with detectable cancer might also be a good news for the patient to live longer.

4.5 Conclusion

From above results we can see that Cox PH models can be fitted for both types of events if we stratify some covariates and do some data transformation. According to the models, we can find that the effects of **rx**, **nodes**, **extent** and **surg** are similar on the event time of both recurrence time and death time. Taking the **Lev+5FU** treatment can minimize the risk of recurrence and death, while more lymph nodes with detectable cancer means greater risk of recurrence and death and the effect of it is particularly significant on death events. It means that if the number of lymph nodes with detectable cancer is large, then the patient is more likely to die after surgery. Besides, for patients with longer time from surgery to registration, their risk for recurrence and death would also be greater.

The only difference between these two models is the effect of **age**. For older people, they are more likely to die after surgery while the risk of recurrence doesn't have much to do with age.

To sum up, the Cox PH model is an appropriate choice for describing the pattern of this dataset. We can get the effects of some covariates from the summary output of the model, while for the other stratified covariates, we can compare their baseline hazard functions to compare their effects on the event time. Therefore, using the stratified Cox PH model can help us clearly see the effect of all covariates in this problem and thus we decide to choose it as our final model.

5 Summary

5.1 Effect of covariates

According to the analysis above, we can see that there are five covariates in total that have significant effect on lowing down the risk(recurrence and death) probability. They are **rx(Lev+5FU)(Treatment)**, **nodes(number of lymph nodes with detectable cancer)**, **extent(extension of colon cancer)**, **surg(time from surgery to registration)** and **age**. The previous 4 covariates have similar influence on recurrence event and death event. The **Lev+5FU** level of **rx** has a strong positive effect on the control of colon cancer, which means taking this kind of treatment can effectively lower the probability of recurrence and death. The number of lymph nodes with detectable cancer works as an important indicator of the risk level. The more these lymph nodes are detected, the higher probability of recurrence or death after surgery is. The **extent** also plays an important role. If the colon cancer has extended to the other parts of body, then it is more likely for patients to recur colon cancer or even die. In addition, **Surg** is an influential covariate, which is reasonable since later registration will definitely lead to earlier record of recurrence or death date.

The only covariate that has a different effect for recurrence event and death event is **age**. As is shown in the Cox-PH model, older patients have higher probability of death whereas their probability of recurrence hardly changes compared to that for younger people. This result is consistent with the EDA results above. For older patients, the colon cancer is not the only risk factor that promote their death. Other risk factors like cardiovascular disease should also be taken into consideration in the survival analysis.

5.2 Model Selection

The stratified Cox PH model is selected as the final model due to its great performance. The default Cox-PH model and the AFT models(Weibull, log-logistic and log-normal) are proved to be not accurate since some of their key assumptions are not met.

5.3 Recommendation

Through the survival analysis of the **Colon** dataset, we offer the following recommendations for patients and hospitals:

1. Before surgery for colon cancer, it is essential(better) for doctors to help patients decrease the number of lymph nodes with detectable cancer and minimize the spread of cancer.
2. During surgery for colon cancer, patients should take the **Lev+5FU** treatment.

3. After surgery, Older people should be cautious about other risk factors like cardiovascular disease, to avoid the death threat imposed by these factors.

5.4 Limitation

In our survival analysis, we only compares the AFT models(Weibull, log-logistic and log-normal) with the Cox-PH model. Although the performance of stratified Cox-PH model is satisfying, we are not sure whether AFT model with other assumed distribution or time-dependent Cox-PH model will have better accuracy and interpretability. It would be better if we consider more kinds of model in the model selection part.

6 Reference

1. Colon cancer [Internet]. Mayo Clinic. Mayo Foundation for Medical Education; Research; 2019. Available from: https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669

2. Accelerated failure time model [Internet]. Wikipedia. Wikimedia Foundation; 2019. Available from: https://en.wikipedia.org/wiki/Accelerated_failure_time_model

3. Proportional hazards model [Internet]. Wikipedia. Wikimedia Foundation; 2019. Available from: https://en.wikipedia.org/wiki/Proportional_hazards_model

4. Cox model assumptions [Internet]. STHDA. Available from: http://www.sthda.com/english/wiki/cox-model-assumptions

5. Kleinbaum DG, Klein M. Survival analysis. Vol. 3. Springer; 2010.

7 Appendix

7.1 Data preprocessing

```
# Remove first two columns
colon_data = colon[ ,c(-1, -2)]

# Remove observations with NA
colon_data = colon_data[complete.cases(colon), ]

# Split the data by event type
colon1 = colon_data %>% filter(etype == 1)
colon2 = colon_data %>% filter(etype == 2)

# Remove etype
colon1 = colon1[, -14]
colon2 = colon2[, -14]
```

7.2 Exploratory data analysis

```
## Correlation table
colon_numeric_rec = colon1[, c("age", "nodes", "time")]
colon_numeric_death = colon2[, c("age", "nodes", "time")]
res_rec = cor(colon_numeric_rec)
res_death = cor(colon_numeric_death)
corrplot(res_rec, type = "upper", order = "hclust", tl.col = "black",
         tl.srt = 0)
corrplot(res_death, type = "upper", order = "hclust", tl.col = "black",
```

```
        tl.srt = 0)

## Recurrence VS Death
colon_data = colon_data %>%
  mutate(etype = replace(etype, etype == "recurrence", "Recurrence")) %>%
  mutate(etype = replace(etype, etype == 2, "Death"))
rec_death_fit = survfit(Surv(time, status) ~ etype, data = colon_data)
rec_death_plot = autoplot(rec_death_fit,
                          main = "Survival probability(Recurrence time VS Death time)")
rec_death_plot

## The effect of different treatments for recurrence and death
par(mfrow = c(1, 2))
rec_treat_fit = survfit(Surv(time, status) ~ rx, data = colon1)
rec_treat_plot = autoplot(rec_treat_fit,
                          main = "S(t) under different treatments(Recurrence time)")
rec_treat_plot

death_treat_fit = survfit(Surv(time, status) ~ rx, data = colon2)
death_treat_plot = autoplot(death_treat_fit,
                            main = "S(t) under different treatments(Death time)")
death_treat_plot

## Covariates change
aa_fit_rec = aareg(Surv(time, status) ~. , data = colon1)
rec_covariates_plot = autoplot(aa_fit_rec,
                               main = "Change of covariates(Recurrence time)")
rec_covariates_plot

aa_fit_death = aareg(Surv(time, status) ~. , data = colon2)
death_covariates_plot = autoplot(aa_fit_death,
                                 main = "Change of covariates(Death time)")
death_covariates_plot
```

7.3 AFT model

```
## Test the exponential and weibull distribution
fit1<-survreg(Surv(time, status)~.,data=colon2)
cs.res1<-exp(-fit1$linear.predictor/fit1$scale)*
  (Surv(colon2$time, colon2$status)[,1])^(1/fit1$scale)
cs.fit1<-survfit(Surv(cs.res1,colon2$status)~1, type="fleming-harrington")
par(mfrow = c(1,2))
plot(cs.fit1$time, -log(cs.fit1$surv), type="s",
     main = "-log(S(t)) vs. t for Exponential")
plot(log(cs.fit1$time), log(-log(cs.fit1$surv)), type="s",
     main = "log(-log(S(t))) vs. log(t) for Weibull")

## Test the log-logistic distribution
fit <- survreg(Surv(time,status)~. , data = colon2 ,dist = "loglogistic")
cs.res <-  exp((log(Surv(colon2$time, colon2$status)[,1])-fit$linear.predictor)/fit$scale)
cs.fit <- survfit(Surv(cs.res,colon2$status)~1, type="fh2")
plot(log(cs.fit$time), log(exp(-log(cs.fit$surv))-1), type="s",
     main = "log(exp(-log(S(t)))) vs. log(t) for Loglogistic")
```

```
## Variable selection
fit2 <- survreg(Surv(time,status)~
                rx+age+obstruct+perfor+adhere+nodes+differ+extent+surg+node4,
                data = colon2 , dist = "loglogistic")
fit3 <- survreg(Surv(time,status)~
                rx+age+obstruct+adhere+nodes+differ+extent+surg+node4,
                data = colon2 , dist = "loglogistic")
fit4 <- survreg(Surv(time,status)~
                  rx+age+obstruct+nodes+differ+extent+surg+node4,
                data = colon2 , dist = "loglogistic")
fit5 <- survreg(Surv(time,status)~
                  rx+age+obstruct+nodes+extent+surg+node4,
                data = colon2 , dist = "loglogistic")

## Inference on T|X
coefficient <- as.matrix(fit5$coefficients, nrow=1 )
var_coef <- as.matrix(fit5$var)
var_coef[,10] <- var_coef[,10]*fit5$scale

## Inference on median death time
quan <- qlogis(0.5)
coef <- as.matrix(c(fit5$coefficients, fit5$scale), nrow = 1)
condition <- matrix(c(1,0,1,60,1,10,2,1,1), nrow = 1)
condition1 <- matrix(c(1,0,1,60,1,10,2,1,1,quan), nrow = 1)
y_median <- condition1 %*% coef
t_median <- exp(y_median)
var_y <- condition1 %*% var_coef %*% t(condition1)
CI_y <- c( y_median-1.96*sqrt(var_y),y_median+1.96*sqrt(var_y))
CI_t <- exp(CI_y)
var_coef[10,] <- var_coef[10,]*fit5$scale

## Get the point estimation and CI for S(t|X)
epsilon_hat <- (log(1500) - condition %*% coefficient)/fit5$scale
condition2 <- matrix(-c(1,0,1,60,1,10,2,1,1,epsilon_hat)/fit5$scale, nrow=1)
var_epsilon <- condition2 %*% var_coef %*% t(condition2)
CI_epsilon <- c(epsilon_hat-1.96*sqrt(var_epsilon),epsilon_hat+1.96*sqrt(var_epsilon))
CI_S <- 1/(1+exp(-CI_epsilon))
expect_S <- 1/(1+exp(-epsilon_hat))

## Model assessment
de.res<-residuals(fit, type="deviance")
par(mfrow = c(1,2))
plot(fit$linear.predictor, de.res)
plot(Surv(colon2$time, colon2$status)[,1], de.res)

## Analysis on recurrence data
### Check the error distribution
fit6<-survreg(Surv(time, status)~.,data=colon1)
cs.res2<-exp(-fit6$linear.predictor/fit6$scale)*
  (Surv(colon1$time, colon1$status)[,1])^(1/fit6$scale)
cs.fit2<-survfit(Surv(cs.res2,colon1$status)~1, type="fleming-harrington")
par(mfrow = c(1,2))
plot(cs.fit2$time, -log(cs.fit2$surv), type="s",
```

```
     main = "-log(S(t)) vs. t for Exponential")
plot(log(cs.fit2$time), log(-log(cs.fit2$surv)), type="s",
     main = "log(-log(S(t))) vs. log(t) for Weibull")

fit7 <- survreg(Surv(time,status)~. , data = colon1 ,dist = "loglogistic")
cs.res3 <-  exp((log(Surv(colon1$time,
                colon1$status)[,1])-fit7$linear.predictor)/fit7$scale)
cs.fit3 <- survfit(Surv(cs.res3,colon1$status)~1, type="fh2")
plot(log(cs.fit3$time), log(exp(-log(cs.fit3$surv))-1), type="s",
     main = "log(exp(-log(S(t)))) vs. log(t) for Loglogistic")

### Variable selection
summary(fit7)
fit8 <- survreg(Surv(time,status)~
                rx+sex+obstruct+perfor+adhere+nodes+differ+extent+surg+node4,
                data = colon1 , dist = "loglogistic")
summary(fit8)
fit9 <- survreg(Surv(time,status)~
                rx+sex+obstruct+adhere+nodes+differ+extent+surg+node4,
                data = colon1 , dist = "loglogistic")
summary(fit9)
fit10 <- survreg(Surv(time,status)~
                  rx+sex+obstruct+nodes+differ+extent+surg+node4,
                 data = colon1 , dist = "loglogistic")
summary(fit10)
fit11 <- survreg(Surv(time,status)~
                  rx+obstruct+nodes+differ+extent+surg+node4,
                 data = colon1 , dist = "loglogistic")
summary(fit11)
fit12 <- survreg(Surv(time,status)~
                   rx+obstruct+nodes+extent+surg+node4,
                 data = colon1 , dist = "loglogistic")
summary(fit12)
```

7.4 Cox PH model

```
# Cox PH model(1)
## Check the nonlinearity for continuous variable nodes
ggcoxfunctional(Surv(time, status) ~ age + log(age) + sqrt(age), data = colon1)
ggcoxfunctional(Surv(time, status) ~ nodes + log(nodes+1) + sqrt(nodes), data = colon1)

## Fit a Cox PH model with all covariates
colon_cox = coxph(Surv(time, status) ~ rx+sex+log(age)+obstruct+perfor+
                    adhere+nodes+differ+extent+surg+node4 ,
                  data = colon1)
summary(colon_cox)

## Use AIC to do variable selection
colon_AIC = step(colon_cox)
summary(colon_AIC)

## Using sex, perfor and adhere as stratum
colon_cox2 = coxph(Surv(time, status) ~ rx+strata(sex)+obstruct+strata(perfor)+
```

```
                    strata(adhere)+nodes+differ+extent+surg+node4 ,
                  data = colon1)
summary(colon_cox2)

## Test for PH
cox.zph(colon_cox2)
ggcoxzph(cox.zph(colon_cox2))

var_name = c("rxLev" , "rxLev+5FU", "obstruct" ,
             "nodes" , "differ" , "extent" , "surg" , "node4")
for(i in var_name){
  plot(cox.zph(colon_cox2), col = c('blue', 'red'), var = i)
  abline(a = 0, b = 0, col = 'green')
}

## Improve the model
colon_cox3 = coxph(Surv(time, status) ~ rx+strata(sex)+strata(obstruct)+strata(perfor)+
                     strata(adhere)+nodes+strata(differ)+extent+surg+strata(node4) ,
                   data = colon1)
summary(colon_cox3)

## Test the PH again
cox.zph(colon_cox3)
ggcoxzph(cox.zph(colon_cox3))

## Check for exponential link
ggcoxdiagnostics(colon_cox3, type = "deviance",
                 linear.predictions = T, ggtheme = theme_bw())

## Obtain nonparametric baseline hazard estimate
colon_baseh = survfit(colon_cox3, type="aalen")
colon_baseh =survfit(colon_cox3, type="breslow")
plot(colon_baseh)
#-----------------------------------------------------------------
# Cox PH model(2)
## Check the nonlinearity for continuous variable nodes
ggcoxfunctional(Surv(time, status) ~ age + log(age) + sqrt(age), data = colon2)
ggcoxfunctional(Surv(time, status) ~ nodes + log(nodes+1) + sqrt(nodes), data = colon2)

## Fit a Cox PH model with all covariates
colon_cox_b = coxph(Surv(time, status) ~ rx+sex+log(age)+obstruct+perfor+
                      adhere+nodes+differ+extent+surg+node4 ,
                    data = colon2)
summary(colon_cox_b)

## Use AIC to do variable selection
colon_AIC_b = step(colon_cox_b)
summary(colon_AIC_b)

## Using sex, perfor and adhere as stratum
colon_cox_b2 = coxph(Surv(time, status) ~
                       rx+strata(sex)+ log(age) + obstruct+strata(perfor)+
                     strata(adhere)+nodes+differ+extent+surg+node4 ,
```

```
                    data = colon2)
summary(colon_cox_b2)

## Test for PH
cox.zph(colon_cox_b2)

## Improve the model
colon_cox_b3 = coxph(Surv(time, status) ~
                 rx+strata(sex)+log(age)+strata(obstruct)+strata(perfor)+strata(adhere)+
                   log(nodes+1)+strata(differ)+extent+surg+strata(node4) ,
                    data = colon2)
summary(colon_cox_b3)

## Test the PH again
cox.zph(colon_cox_b3)
ggcoxzph(cox.zph(colon_cox_b3))

var_name_b = c("rxLev" , "rxLev+5FU", "age", "nodes" ,
                  "extent" , "surg" )
for(i in var_name_b){
  plot(cox.zph(colon_cox_b2), col = c('blue', 'red'), var = i)
  abline(a = 0, b = 0, col = 'green')
}

## Check for exponential link
ggcoxdiagnostics(colon_cox_b3, type = "deviance",
                 linear.predictions = T, ggtheme = theme_bw())

## Inference
### Solve for the mean survival time
MST = function(x){
  X_pre = data.frame(rx = "Lev", sex = 1, age = 60, obstruct = 1,
                     perfor = 0, adhere = 0, nodes = 10, differ = 2,
                     extent = 2, surg = 1, node4 = 1, status = 1, time = x)
  S = exp(- predict(colon_cox_b3, X_pre, type = "expected"))
  return (S - 0.5)
}
confint(predict(colon_cox_b3, X_pre))

uniroot(MST, interval = c(0, 3329))
```