BST 222, Project 1

Chemotherapy for Stage B/C colon cancer
Survival Analysis

Bohao Zou, 917796070
MS student
`bhzou@ucdavis.edu`

DEPARTMENT OF STATISTIC,
BIOSTATISTICS

December 19, 2020

Contents

1 Define Questions & Data Explore

The dataset which I used comes from the "Survival" package of R. The name of this dataset calls "Chemotherapy for Stage B/C colon cancer". Doctors used three different approaches to treat colon cancer. Those three treatments are: *1.Obs(ervation)*, *2.Lev(amisole)*, *3.Lev(amisole)+5-FU*. Beside those three different treatment, researchers also collected other information which relates with the survival time of colon cancer like *if patients have obstruction of colon by tumour*, *if colon cancer adhered to nearby organs*. The researchers not only recorded the death time of each patient. but also wrote the recurrence time of each patient. The recurrence time will be as the same as the death time if this patient didn't recur during this survey. I defined the death as event and do not care the time of recurrence.

There are two primary questions that we need to answer by analysis this data. The first question is *if the hazard rates of those three different treatments are the same in the statistical significant level of 0.05; if different, which hazard rate is higher*? The second question we need to give answer is *which covariates are increase the risk of death and which are decrease the risk of death*. For answering those question, we need to give a first glance to this dataset.

There are two ways to explore the data. One is histogram plot and the other is box plot. Histogram can give us the distribution of death time of those three treatments and box plot can help us to compare the death time of those three treatments. Excluding the censoring data, I drew the histogram of time with three different treatments.

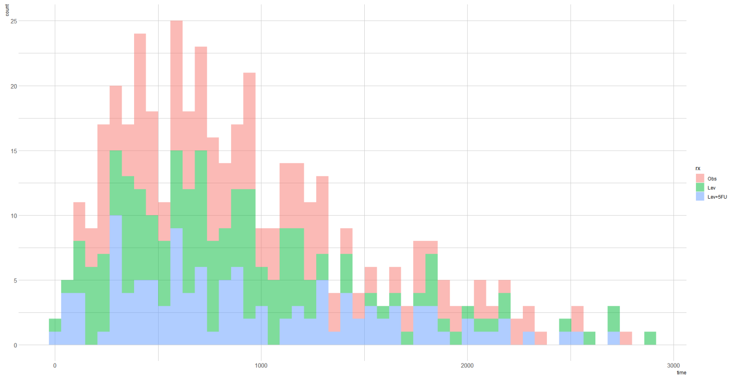


Figure 1: The histogram of time with different treatments. Red is the distribution of Obs, Green is the distribution of Lev(amisole) and Blue is the distribution of Lev(amisole)+5-FU.

From this plot we can know that the distributions of those three treatments are roughly same. Most of patients dead before 1500 days. Small number of patients dead after 2000 days. This may indicate that the hazard rates between those three treatments have no difference. For proving this, we need a further explore. So, I drew the box plot to compare the survival time of those methods.

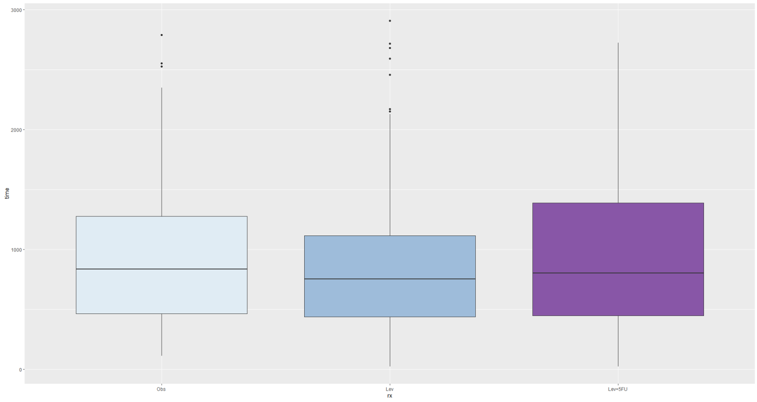


Figure 2: The box plot of treatments vs. death time.

By comparing the 50-th percentile of each treatment, we can know that the 50-th percentile of each group are roughly same but the survival time of Obs treatment is the highest, the Lev(amisole)+5-FU is the second and Lev(amisole) is the lowest. This plot can give an evidence that there may be no difference between three treatments.

I want to know if other covariates can influence the survival time. For solving this question, I selected the differentiation of tumour as one covariate and drew the box plot of survival time. In the differentiation of tumour, 1 represents well condition, 2 represents moderate condition and 3 represents poor condition.

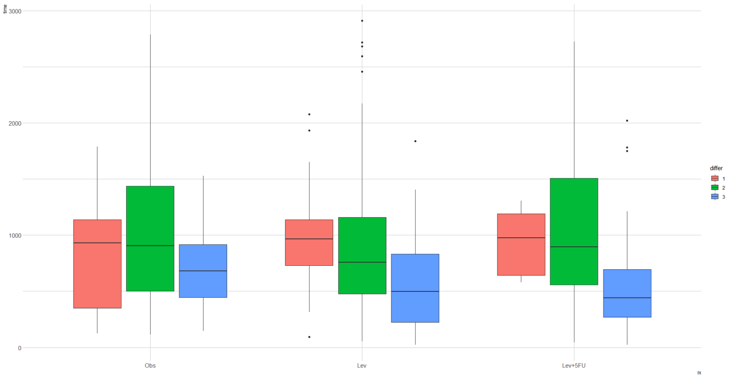


Figure 3: The box plot of time vs. treatments for differentiation of tumour.

From this plot, we can know that the differentiation of tumour has huge influence with the death time. In all of three treatments, the well condition has the longest survival time and the poor condition has the shortest survival time. This may indicate the differentiation of tumour has significant effect to the survival time of colon cancer.

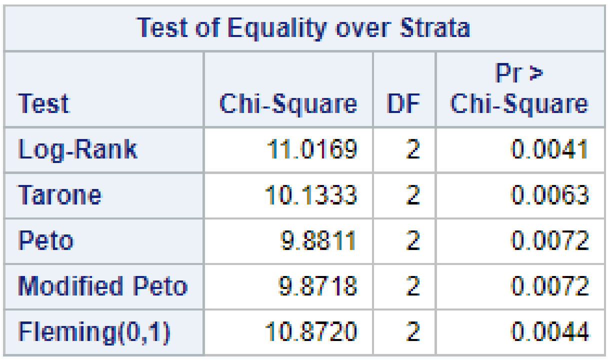
2 Methods

2.1 Log-Rank Test

For answering the first question, *if the hazard rates of those three different treatments are the same in the statistical significant level of 0.05; if different, which hazard rate is higher?* we need to use log-rank test to test if the hazard rate of those three treatments has difference. The null hypothesis and alternative hypothesis of log-rank test are:

$$\begin{aligned} &H_0 : h_1(t) = h_2(t) = \cdots = h_k(t) \quad for \quad all \quad t \leq \tau \\ &H_A : At \quad least \quad one \quad of \quad the \quad h_j(t) \quad is \quad different \quad for \quad some \quad t \leq \tau \end{aligned}$$

where k means the groups number, $h_j(t)$ represents the hazard rate of $j-th$ group and τ means the largest death time in this dataset. By using this test, we can know if the hazard rates of those three treatments have difference. I also used others K-Sample Test like Tarone, Peto and Modified Peto. The result shows below:



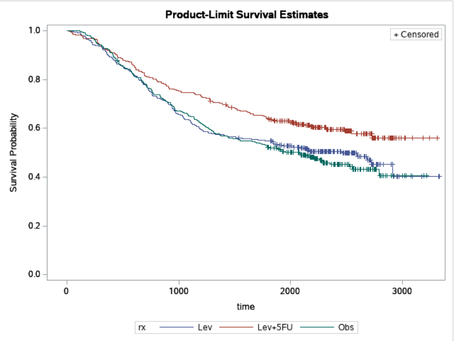
Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	11.0169	2	0.0041
Tarone	10.1333	2	0.0063
Peto	9.8811	2	0.0072
Modified Peto	9.8718	2	0.0072
Fleming(0,1)	10.8720	2	0.0044

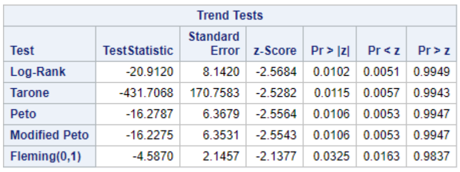
From this table we can know that all of these tests show a conclusion that the hazard rates are not the same in these three different treatments for treating colon cancer under the significant level of 0.05. So, we can know at least one of $h_j(t)$ is different for some t.

2.2 Test for Trend

We want to figure out which treatment is distinguished and if the hazard rate of it is higher. For giving a first impression, I drew the K-M survival curve for those three treatments:



From this K-M survival curve plot we can roughly know that the survival probability of patients who received Lev(amisole)+5-FU treatment is higher than the other two treatments. The survival probability of Obs(ervation) and Lev(amisole) are roughly same and it seems Lev(amisole) is a little bit higher than Obs(ervation). For answering this question, we need to do a two tests for trend and a Bonferroni correction to control the FWER(Familywise Error Rate) of those hypothesis testing. The first trend test is used for testing the hazard rates of Lev(amisole)+5-FU and Lev(amisole). The second trend test is used for testing Lev(amisole) and Obs(ervation). The result of the first trend tests are showed below:

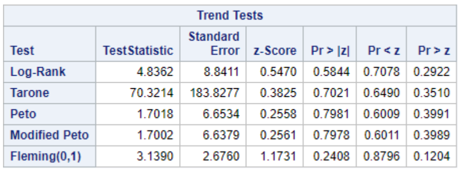


Trend Tests

Test	TestStatistic	Standard Error	z-Score	Pr > |z|	Pr < z	Pr > z
Log-Rank	-20.9120	8.1420	-2.5684	0.0102	0.0051	0.9949
Tarone	-431.7068	170.7583	-2.5282	0.0115	0.0057	0.9943
Peto	-16.2787	6.3679	-2.5564	0.0106	0.0053	0.9947
Modified Peto	-16.2275	6.3531	-2.5543	0.0106	0.0053	0.9947
Fleming(0,1)	-4.5870	2.1457	-2.1377	0.0325	0.0163	0.9837

Figure 4: The trend test for Lev(amisole)+5-FU and Lev(amisole)

The result of the second trend test are showed below:



Trend Tests

Test	TestStatistic	Standard Error	z-Score	Pr > |z|	Pr < z	Pr > z
Log-Rank	4.8362	8.8411	0.5470	0.5844	0.7078	0.2922
Tarone	70.3214	183.8277	0.3825	0.7021	0.6490	0.3510
Peto	1.7018	6.6534	0.2558	0.7981	0.6009	0.3991
Modified Peto	1.7002	6.6379	0.2561	0.7978	0.6011	0.3989
Fleming(0,1)	3.1390	2.6760	1.1731	0.2408	0.8796	0.1204

Figure 5: The trend test for Lev(amisole) and Obs(ervation)

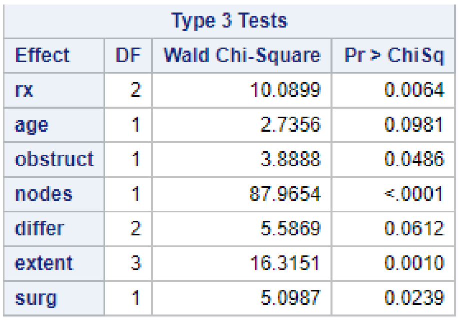
From the result of first trend test we can know that there is a significant difference between the treatment of Lev(amisole)+5-FU and Lev(amisole) under the significant level of 0.05 with Bonferroni correction. The result of the second trend test shows there is no difference between the treatment of Lev(amisole) and Obs(ervation) under the significant level of 0.05 with Bonferroni correction. Because its are trend test, so we can get a conclusion that the patient who treated with Lev(amisole)+5-FU have the highest survival probability but the survival probabilities have no difference between the patients who received Lev(amisole) or Obs(ervation).

2.3 Cox-PH Regression Model

For answering the second question, we need to build a Cox-PH regression model and check the hazard ratio for p-th covariates. At the beginning, we need to select covariates.

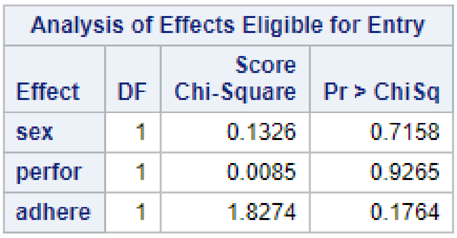
2.3.1 Model Selection

There are 10 variables needed to select to participate into the model. 8 of variables are categories variable and 2 of its are continuous variables. In this part, we need to select some appropriate covariates into Cox-PH model. I used forward selection and set the entry P-Value as 0.1. The result of selection shows below:



Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
rx	2	10.0899	0.0064
age	1	2.7356	0.0981
obstruct	1	3.8888	0.0486
nodes	1	87.9654	<.0001
differ	2	5.5869	0.0612
extent	3	16.3151	0.0010
surg	1	5.0987	0.0239

Figure 6: The selected variables by the forward selection procedure.



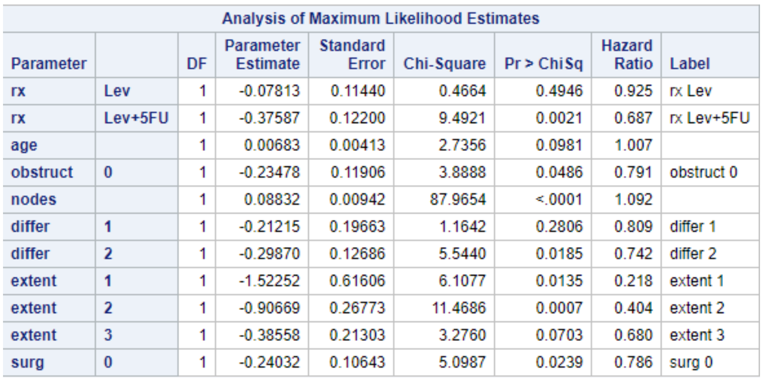
Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
sex	1	0.1326	0.7158
perfor	1	0.0085	0.9265
adhere	1	1.8274	0.1764

Figure 7: The dropped variables by the forward selection procedure.

From the result of selection we can know that the variables *rx: Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU, age, obstruct: obstruction of colon by tumour, nodes: number of lymph nodes with detectable cancer, 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)* are all included in the model. The variables *sex: 1 = male, perfor: perforation of colon, adhere: adherence to nearby organs*, are not included in the model.

2.3.2 Model Interpretation

We build a Cox-PH regression model by using those covariates which seleceted by forward selection with entry P-Value 0.1. By using the MLE to estimate the coefficients of the model, we have the following result:



Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
rx	Lev	1	-0.07813	0.11440	0.4664	0.4946	0.925	rx Lev
rx	Lev+5FU	1	-0.37587	0.12200	9.4921	0.0021	0.687	rx Lev+5FU
age		1	0.00683	0.00413	2.7356	0.0981	1.007	
obstruct	0	1	-0.23478	0.11906	3.8888	0.0486	0.791	obstruct 0
nodes		1	0.08832	0.00942	87.9654	<.0001	1.092	
differ	1	1	-0.21215	0.19663	1.1642	0.2806	0.809	differ 1
differ	2	1	-0.29870	0.12686	5.5440	0.0185	0.742	differ 2
extent	1	1	-1.52252	0.61606	6.1077	0.0135	0.218	extent 1
extent	2	1	-0.90669	0.26773	11.4686	0.0007	0.404	extent 2
extent	3	1	-0.38558	0.21303	3.2760	0.0703	0.680	extent 3
surg	0	1	-0.24032	0.10643	5.0987	0.0239	0.786	surg 0

Figure 8: The result of Cox-PH regression model.

From the result we can know that there are only two variables can increase the relative risk. The two variables are *age and nodes* respectively. The interpretations of those two variables are:

- *Age:* relative risk for a 50 years old patient compared to a 49 years old patient (both treated with the same treatment and other factors are all the same) is 1.007, and it is significant (p-value=0.0981, under the significant level of 0.1).
- *Nodes:* relative risk for a patient who has 3 nodes compared to a patient who has 2 nodes (both treated with the same treatment and other factors are all the same) is 1.092 and it is significant (p-value $\leq$ 0.001, under the significant level of 0.1).

The others variables can decrease the relative risk. The interpretations of those variables are:

- *rx:* The patient who treated with Lev has a probability of dying 0.925 times less than the probability of dying for a patient who treated with Obs with other same factors. But it is not significant (p-value=0.4946, under the significant level of 0.1).
- *rx:* The patient who treated with Lev+5FU has a probability of dying 0.687 times less than the probability of dying for a patient who treated with Obs with other same factors. It is significant (p-value=0.0021, under the significant level of 0.1).
- *obstruct:* The patient who does not have obstruction of colon by tumour has a probability of dying 0.791 times less than the probability of dying for a patient who have obstruction of colon by tumour with other same factors. It is significant (p-value=0.0486, under the significant level of 0.1)
- *differ:* The patient who differentiation of tumour is well has a probability of dying 0.809 times less than the probability of dying for a patient who differentiation of tumour is poor with other same factors. But it is not significant (p-value=0.2806, under the significant level of 0.1)
- *differ:* The patient who differentiation of tumour is moderate has a probability of dying 0.742 times less than the probability of dying for a patient who differentiation

of tumour is poor with other same factors. It is significant (p-value=0.0185, under the significant level of 0.1)

- *extent:* The patient who Extent of local spread is submucosa has a probability of dying 0.218 times less than the probability of dying for a patient who Extent of local spread is contiguous structures with other same factors. It is significant (p-value=0.0135, under the significant level of 0.1)
- *extent:* The patient who Extent of local spread is muscle has a probability of dying 0.404 times less than the probability of dying for a patient who Extent of local spread is contiguous structures with other same factors. It is significant (p-value=0.0007, under the significant level of 0.1)
- *extent:* The patient who Extent of local spread is serosa has a probability of dying 0.680 times less than the probability of dying for a patient who Extent of local spread is contiguous structures with other same factors. It is significant (p-value=0.0703, under the significant level of 0.1)
- *surg:* The patient who surgery time is short has a probability of dying 0.786 times less than the probability of dying for a patient who surgery time is long with other same factors. It is significant (p-value=0.0239, under the significant level of 0.1)

The other factors like *sex, perforation of colon, adherence to nearby organs* may have no effect to the survival time of those patients. This is because those factors are all not significant under significant level of 0.1.

2.3.3 Model Diagnostics

At first, we should check the overall fit of the model. The method which I used is Cox-Snell residuals. Drew a plot $\hat{H}(r_j)$ vs. r_j, where r_j means the Cox-Snell residuals for the $j-th$ observation. $\hat{H}(t)$ is Breslow's estimator of the baseline hazard rate. The plot shows below:

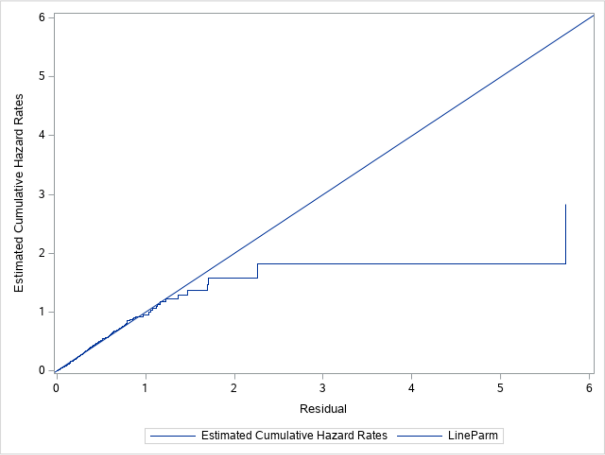


Figure 9: The Cox-Snell residual plot of this dataset and Cox PH model.

From this plot we can know this model fits excellent at the beginning but fits bad at the end. This may indicates that the model can partially fit the data.

In the second step, we need to use Martingale residuals to check non-linearity of a covariate. There are only two continuous variables(*age and nodes*) in my model, so, we need to check those two variables. The results are showed below:

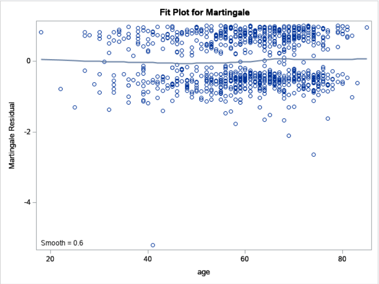


Figure 10: The Martingale Residual plot for age variable in Cox PH model.

From this plot we can know that there is no non-linearity of age in this Cox-PH model.

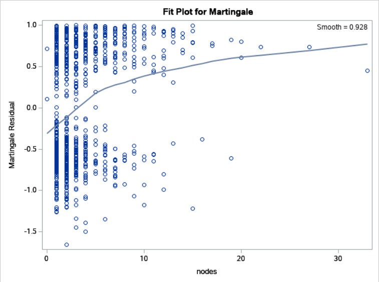


Figure 11: The Martingale Residual plot for nodes variable in Cox PH model.

From this plot we can know that there is a non-linearity of nodes variable. It likes a square root function with the form $y = \sqrt{x}$. So, I transformed the nodes variable with square root operation and then used the transformed s_nodes variable to fit the Cox-PH model. The Martingale Residual plot after transformation shows below:

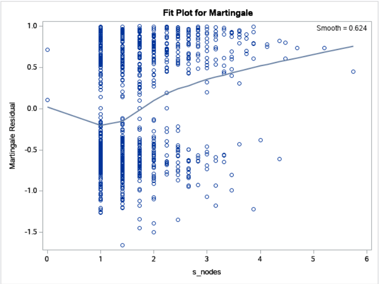


Figure 12: The Martingale Residual plot for s_nodes variable in Cox PH model.

From this plot we can know the non-linearity of s_nodes is batter than the original nodes variable. So, I will use this transformed variable s_nodes for next analysis.

In the final step, we need to check the PH assumption by using Schoenfeld residuals of *rx* variable. If PH holds, then Schoenfeld residuals are uncorrelated with time. We can draw a Schoenfeld residuals vs. time plot. If the plot shows a non-random pattern against time is the evidence of violation of the PH assumption. The plot shows below:

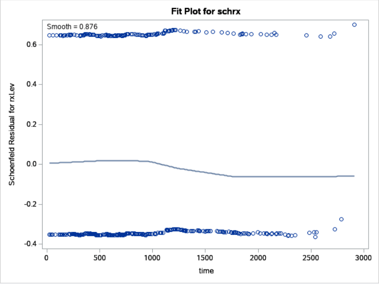
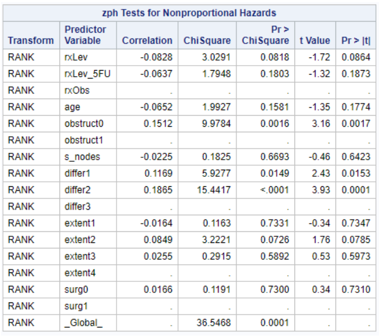


Figure 13: The Schoenfeld residual plot for *rx* variable in Cox PH model.

From this plot we can know Schoenfeld residuals have a non-random pattern against time. This indicates that we have violated the PH assumption. I also used test correlation of Schoenfeld residuals against with time to show which covariates do not fit the PH assumptions. The result shows below:

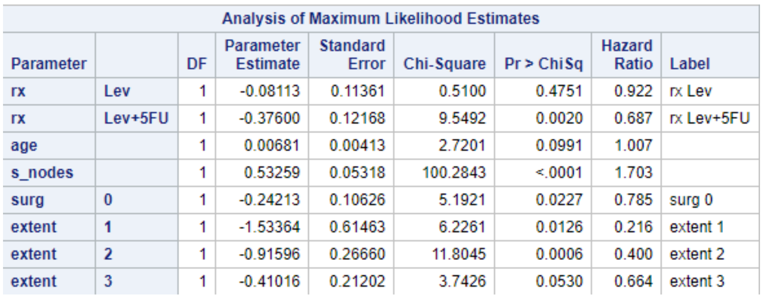


zph Tests for Nonproportional Hazards

Transform	Predictor Variable	Correlation	ChiSquare	Pr > ChiSquare	t Value	Pr > |t|
RANK	rxLev	-0.0828	3.0291	0.0818	-1.72	0.0864
RANK	rxLev_5FU	-0.0637	1.7948	0.1803	-1.32	0.1873
RANK	rxObs	.	.	.	.	.
RANK	age	-0.0652	1.9927	0.1581	-1.35	0.1774
RANK	obstruct0	0.1512	9.9784	0.0016	3.16	0.0017
RANK	obstruct1	.	.	.	.	.
RANK	s_nodes	-0.0225	0.1825	0.6693	-0.46	0.6423
RANK	differ1	0.1169	5.9277	0.0149	2.43	0.0153
RANK	differ2	0.1865	15.4417	<.0001	3.93	0.0001
RANK	differ3	.	.	.	.	.
RANK	extent1	-0.0164	0.1163	0.7331	-0.34	0.7347
RANK	extent2	0.0849	3.2221	0.0726	1.76	0.0785
RANK	extent3	0.0255	0.2915	0.5892	0.53	0.5973
RANK	extent4	.	.	.	.	.
RANK	surg0	0.0166	0.1191	0.7300	0.34	0.7310
RANK	surg1	.	.	.	.	.
RANK	_Global_	.	36.5468	0.0001	.	.

Figure 14: ZPH tests for checking the PH assumption.

From the ZPH tests, we can know that *obstruct* and *differ* do not fit the PH assumptions under the significant level of 0.05. So, we should remove them from our model. After removing two variables and doing the square root operation for *nodes* variable, we can get the final model. At the end, the final result of the final Cox-PH model is showed below:



Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
rx	Lev	1	-0.08113	0.11361	0.5100	0.4751	0.922	rx Lev
rx	Lev+5FU	1	-0.37600	0.12168	9.5492	0.0020	0.687	rx Lev+5FU
age		1	0.00681	0.00413	2.7201	0.0991	1.007	
s_nodes		1	0.53259	0.05318	100.2843	<.0001	1.703	
surg	0	1	-0.24213	0.10626	5.1921	0.0227	0.785	surg 0
extent	1	1	-1.53364	0.61463	6.2261	0.0126	0.216	extent 1
extent	2	1	-0.91596	0.26660	11.8045	0.0006	0.400	extent 2
extent	3	1	-0.41016	0.21202	3.7426	0.0530	0.664	extent 3

Figure 15: The final result of Cox-PH model.

We can check the overall fit of the new model by using Cox-Snell residuals. The plot shows below:

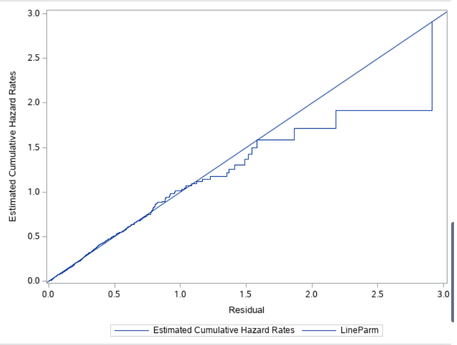


Figure 16: The cox-snell residual plot of new model.

From this plot we can know this new model fits the data much better than old model.

3 Conclusion

We have answered the two primary question that raised up previously. For the first question, the hazard rates of those three different treatments(Obs(ervation), Lev(amisole) and Lev(amisole)+5-FU) are different. The patient who treated with Lev(amisole)+5-FU have the highest survival probability but the survival probabilities have no difference between the patients who received Lev(amisole) or Obs(ervation). For the second question, the increasing of age and number of lymph nodes with detectable cancer will have a higher hazard rates. Sex, perforation of colon, adherence to nearby organs may have no effect and well or moderate differentiation of tumour, without obstruction of colon by tumour, a short surgery time and Extent of local spread to submucosa or muscle or serosa are all will have a lower hazard rates. If we want to interpret the variables, we should use the old model and if we want to do some predictions for new patients, we should use the new model.

4 Appendix

Codes for survival analysis:

```
proc import datafile = "/folders/myshortcuts/MyFolder/colon.csv" out=colon;

*log-rank test for different treatments;
proc lifetest data=colon;
time time*status(0);
strata rx/ test=(logrank TARONE PETO MODPETO FLEMING(0,1));
run;
* trend test for different treatments(Lev, Lev-5fu);
proc import datafile = "/folders/myshortcuts/MyFolder/lev_lev+5fu.csv" out=lev_lev5fu;
proc lifetest data=lev_lev5fu;
```

```
time time*status(0);
strata rx/trend test=(logrank TARONE PETO MODPETO FLEMING(0,1));
run;

proc import datafile = "/folders/myshortcuts/MyFolder/lev_obs.csv" out=lev_obs;
proc lifetest data=lev_obs;
time time*status(0);
strata rx/trend test=(logrank TARONE PETO MODPETO FLEMING(0,1));
run;

* cox reg;
proc phreg data=colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx;
run;
* cox forward selction ;
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx sex age obstruct perfor adhere nodes differ extent surg
/selection=forward slentry=0.1 details;
run;

* Cox-snell residuals plot for original data;
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx age obstruct nodes differ extent surg;
output out=plot1_1 logsurv=logsurv1 /method = ch;

data plot1_1;
set plot1_1;
snell = -logsurv1;
cons = 1;
```

```

proc phreg data=plot1_1;
model snell*status(0) = cons;
output out = plot1_2 logsurv= logsurv2/method=ch;

data plot1_2;
set plot1_2;
cumhaz = - logsurv2;

proc sort data=plot1_2;
by snell;

proc sgplot data= plot1_2;
step y=cumhaz x=snell /MARKERFILLATTRS=(color="red");
lineparm x=0 y=0 slope=1; /** intercept, slope **/
label cumhaz = "Estimated Cumulative Hazard Rates";
label snell = "Residual";
run;

* Martingale Residuals for age (continuous variable);
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx obstruct differ extent surg nodes;
output out=plot2_1 RESMART = Martingale;

proc loess data=plot2_1;
model Martingale =age / direct;
run;


* Martingale Residuals for nodes (continuous variable);
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx age obstruct differ extent surg;
output out=plot2_1 RESMART = Martingale;

proc loess data=plot2_1;
model Martingale = nodes / direct;
run;

* Martingale Residuals for sqrted nodes (continuous variable);
```

```
data colon;
set colon;
s_nodes = sqrt(nodes);
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx age obstruct differ extent surg;
output out=plot2_1 RESMART = Martingale;

proc loess data=plot2_1;
model Martingale = s_nodes / direct;
run;


* PH assumption test for each variables;
ods noproctitle;
ods graphics / imagemap=on;

proc phreg data=WORK.COLON zph(noplot global);
  class rx obstruct differ extent surg / param=glm;
  model time*status(0)=rx age obstruct s_nodes differ extent surg / rl;
run;

* Cox-snell residuals plot for data which have deleted some variables;
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) = rx age s_nodes surg extent;
output out=plot1_1 logsurv=logsurv1 /method = ch;

data plot1_1;
set plot1_1;
snell = -logsurv1;
cons = 1;

proc phreg data=plot1_1;
model snell*status(0) = cons;
output out = plot1_2 logsurv= logsurv2/method=ch;

data plot1_2;
set plot1_2;
cumhaz = - logsurv2;

```

```
proc sort data=plot1_2;
by snell;

proc sgplot data= plot1_2;
step y=cumhaz x=snell /MARKERFILLATTRS=(color="red");
lineparm x=0 y=0 slope=1; /** intercept, slope **/
label cumhaz = "Estimated Cumulative Hazard Rates";
label snell = "Residual";
run;

* Final model;
proc phreg data = colon;
class rx;
class sex;
class obstruct;
class perfor;
class adhere;
class differ;
class extent;
class surg;
model time*status(0) =  rx age s_nodes surg extent;
run;
```