國立中山大學應用數學系

存活分析期末報告

存活資料分析

Chemotherapy for Stage B/C Colon Cancer

M072040002 孫浩哲

M072040006 吳強銘

M072040036 吳政霖

指導教授：張中 教授

壹、 目錄

貳、 摘要

報告分成三個大部分，第一部分是先從 R 中 survival 套件中下載我們需要的存活分析資料，並做資料整理以達到分析需求，名為 colon 之首次成功的結腸癌輔助化療試驗數據，紀錄結腸癌手術後，使用不同藥劑來做輔助治療，以及其他變數來確定藥劑以外變數是否造成復發或是死亡等關聯性的探討。

第二部分是建立模型分別為 Cox proportional hazards regression model，並利用 AIC 來挑選變數，何項變數對於模型較為顯著，以及往後統計推論，加上視覺化呈現，更加了解在各項 covariate 影響下，存活函數曲線及風險函數的變化呈現。

第三部分是使用多種檢定，像是 log rank test 或是 local test，並配合 K-M curves 來檢驗模型是否符合假設等探討，加上殘差的檢驗更深入了解模型適切性以及是否變數上有時間相依的問題，是否該對模型做出改善等推論。

參、 前言

利用此項存活資料分析，利用 Cox proportional hazards regression model 的實際應用推論，了解結腸癌輔助化療試驗數據在三種輔助藥物(observation/ levamisole/ levamisole + 5-FU)下的治療成效，了解在不論是男女、年齡以及其他醫學變數下(結腸穿孔、病人依從性、癌細胞數量等)，是否有效壓制病情復發抑或是嚴重到造成死亡，藉此運用各項視覺化以及檢定殘差等方式，了解在何項影響之下風險較高，存活函數亦受到何種影響，為我們主要探討對象。

肆、 資料背景及變數介紹

1. 資料背景

R 中 survival 套件中下載的存活分析資料，名為 colon 之首次成功的結腸癌輔助化療試驗數據，該研究最初在 Laurie（1989）中描述；主要報告見 Moertel（1990）；該數據集在 Moertel（1991）的最終報告；Lin（1994）在論文中也使用了。紀錄結腸癌手術後，使用在三種輔助藥物(observation/ levamisole/ levamisole + 5-FU)來做輔助治療，以及其他影響變數下，是否造成病情復發抑或嚴重至死亡。

2. 變數介紹

變數	型態
ID	病患編號
STUDY	對所有病患為 1
RX	三種輔助藥物 (observation/ levamisole/ levamisole+5-FU) 記為(Obs/ Lev/ Lev+5-FU)
SEX	1：男性、0：女性
AGE	年齡
OBSTRUCT	是否有阻塞(0/ 1)
PERFOR	是否有結腸穿孔(0/ 1)
ADHERE	病人依從性(0/ 1)
NODES	含有癌細胞的淋巴結數量
TIME	觀察時間(到事件發生或 censoring)
STATUS	Delta(是否 censoring)
DIFFER	腫瘤好壞(好到壞=1 到 3)
EXTENT	局部擴散程度(1：黏膜下層、2：肌肉、3：漿膜、4：連續結構)
SURG	手術完來登記於 data 上的時間間隔 (0：短、1：長)
NODE4	是否超過四個顯示陽性的淋巴結(0/ 1)
ETYPE	1：復發、2：死亡

伍、 研究方法與進行步驟

- 進行步驟：

1. 清理資料，轉換資料型態
2. 建構全模型，再利用 AIC 選取參數，最終選得 nodes、 rx、 obstruct、 adhere、 differ、 extent、 surg、 etype、 node4
3. 再利用背後參數意義以及顯著程度，加上 K-M curves 視覺化圖判定佐以 log-rank test 判斷，決定最後參數 nodes、 rx、 obstruct、 adhere、 extent、 surg、 node4
4. 進行參數估計及了解 Hazard ratio(95% confidence interval)
5. 再進行 local test，檢測 rx 參數是否在模型中顯著
6. 檢驗其 cox PH model，以 cox snell residuals plot 及 cox martingale residuals plot，還有 shoenfeld residuals，了解推論結果並做修正改善
7. 之後轉換資料型態成 time-dependent data(依 etype)
8. 建構全模型，再利用 AIC 選取參數，最終選得 nodes、 rx、 obstruct、 adhere、 extent、 surg、 etype、 node4(再依照一般模型選擇相同變數)
9. 進行參數估計及了解 Hazard ratio(95% confidence interval)
10. 再進行 local test，檢測 rx 參數是否在模型中顯著
11. 檢驗其 cox PH model，以 cox snell residuals plot 及 cox martingale residuals plot，還有 shoenfeld residuals，了解推論結果並做修正改善

➢ 研究結果：

1. 一般模型資料樣貌：

	age	nodes	id	study	time	status	rx	sex	obstruct	perfor	adhere	differ	extent	surg	node4	etype
1	43	5	1	1	1521	1	Lev+5FU	1	0	0	0	2	3	0	1	2
2	43	5	1	1	968	1	Lev+5FU	1	0	0	0	2	3	0	1	1
3	63	1	2	1	3087	0	Lev+5FU	1	0	0	0	2	3	0	0	2
4	63	1	2	1	3087	0	Lev+5FU	1	0	0	0	2	3	0	0	1
5	71	7	3	1	963	1	Obs	0	0	0	1	2	2	0	1	2
6	71	7	3	1	542	1	Obs	0	0	0	1	2	2	0	1	1
7	66	6	4	1	293	1	Lev+5FU	0	1	0	0	2	3	1	1	2
8	66	6	4	1	245	1	Lev+5FU	0	1	0	0	2	3	1	1	1
9	69	22	5	1	659	1	Obs	1	0	0	0	2	3	1	1	2
10	69	22	5	1	523	1	Obs	1	0	0	0	2	3	1	1	1
11	57	9	6	1	1767	1	Lev+5FU	0	0	0	0	2	3	0	1	2
12	57	9	6	1	904	1	Lev+5FU	0	0	0	0	2	3	0	1	1
13	77	5	7	1	420	1	Lev	1	0	0	0	2	3	1	1	2
14	77	5	7	1	229	1	Lev	1	0	0	0	2	3	1	1	1
15	54	1	8	1	3192	0	Obs	1	0	0	0	2	3	0	0	2
16	54	1	8	1	3192	0	Obs	1	0	0	0	2	3	0	0	1
17	46	2	9	1	3173	0	Lev	1	0	0	1	2	3	0	0	2
18	46	2	9	1	3173	0	Lev	1	0	0	1	2	3	0	0	1
19	66	1	10	1	3308	0	Lev+5FU	0	0	0	0	2	3	1	0	2
20	66	1	10	1	3308	0	Lev+5FU	0	0	0	0	2	3	1	0	1
21	47	1	11	1	2908	0	Lev	0	0	0	1	2	3	0	0	2
22	47	1	11	1	2908	0	Lev	0	0	0	1	2	3	0	0	1
23	52	2	12	1	3309	0	Lev+5FU	1	0	0	0	3	3	1	0	2
24	52	2	12	1	3309	0	Lev+5FU	1	0	0	0	3	3	1	0	1
25	64	1	13	1	2085	1	Obs	1	0	0	0	2	3	0	0	2
26	64	1	13	1	1130	1	Obs	1	0	0	0	2	3	0	0	1
27	68	3	14	1	2910	1	Lev	1	1	0	0	2	3	0	0	2
28	68	3	14	1	2231	1	Lev	1	1	0	0	2	3	0	0	1
29	46	4	15	1	2754	0	Obs	1	1	0	0	2	3	0	0	2
30	46	4	15	1	2754	0	Obs	1	1	0	0	2	3	0	0	1
31	68	1	16	1	3214	0	Obs	1	0	0	0	2	3	1	0	2
32	68	1	16	1	1323	1	Obs	1	0	0	0	2	3	1	0	1

Figure1：一般資料形式

2. 參數選取(利用 AIC 挑選參數)---一般模型

觀察參數背後意義以及此資料目的，挑選 rx、 obstruct、 adhere、 extent、 surg、 node4 為主要變數，因為主要為療法為主要變數加以結腸是否堵塞、依從行與否、擴散範圍、手術完來登記於 data 上的時間間隔及是否超過四個顯示陽性的淋巴結，以上皆為常理認知影響重要的變數，之後對這些變數化出 K-M curve 及其 log-rank test，期望看出此項變數在不同類別的差異性。

```
              coef exp(coef) se(coef)      z Pr(>|z|)
rxLev+5FU -0.37086   0.69014  0.08732 -4.247 2.16e-05 ***
rxObs      0.04603   1.04711  0.07942  0.580  0.56218
nodes      0.04219   1.04309  0.01072  3.935 8.33e-05 ***
obstruct1  0.21318   1.23761  0.08352  2.552  0.01070 *
adhere1    0.21140   1.23540  0.09090  2.326  0.02004 *
extent2    0.37158   1.45002  0.39820  0.933  0.35075
extent3    0.88386   2.42023  0.38204  2.314  0.02069 *
extent4    1.29756   3.66034  0.40839  3.177  0.00149 **
surg1      0.24089   1.27238  0.07427  3.244  0.00118 **
node41     0.61768   1.85462  0.10028  6.160 7.29e-10 ***
```

Figure2：step AIC 挑選變數結果

原始模型為：

$$\mathbf{h(t|Z) = h_0(t) \times e^{\beta_1 Z_1+\beta_2 Z_2+\beta_3 Z_3+\beta_4 Z_4+\beta_5 Z_5+\beta_6 Z_6+\beta_7 Z_7}}$$

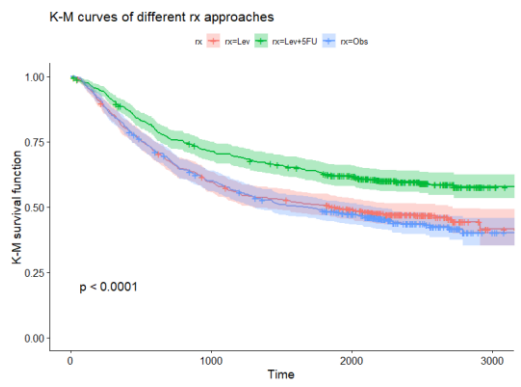
Z_1：factor(rx)，Z_2：nodes，Z_3： obstruct ，Z_4：adhere ，

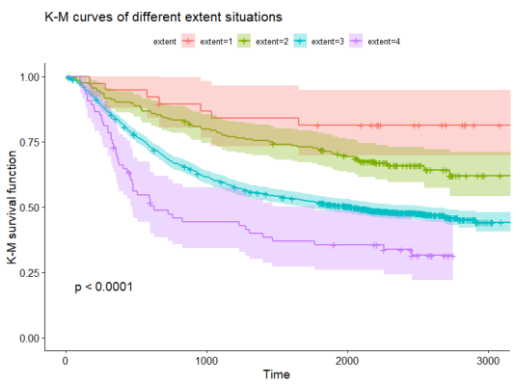
Z_5： factor(extent) ，Z_6：surg， Z_7：node4

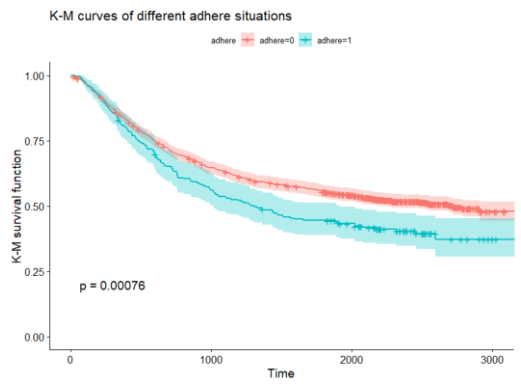
舉出兩個變數來解釋模型：

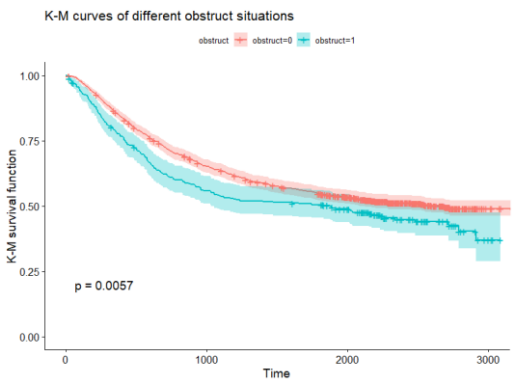
$e^{\widehat{\beta}_1} = 0.6901$ ，意即若使用 Lev+5-FU 兩種藥物進行治療，則風險降低近 30%。$e^{\widehat{\beta}_2} = 1.0431$ ，意即若含癌細胞之淋巴結個數增加一個，則風險增加近 4%。

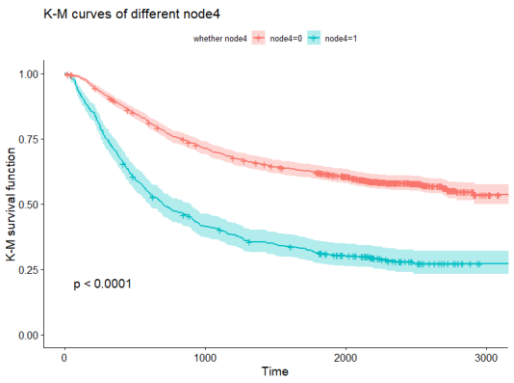
以下為個別對六項變數之 K-M curve 以及 log-rank test 的 p-value

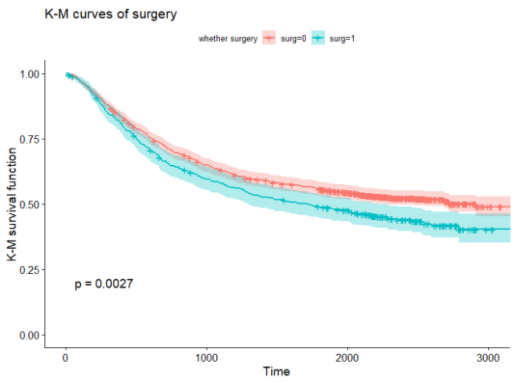








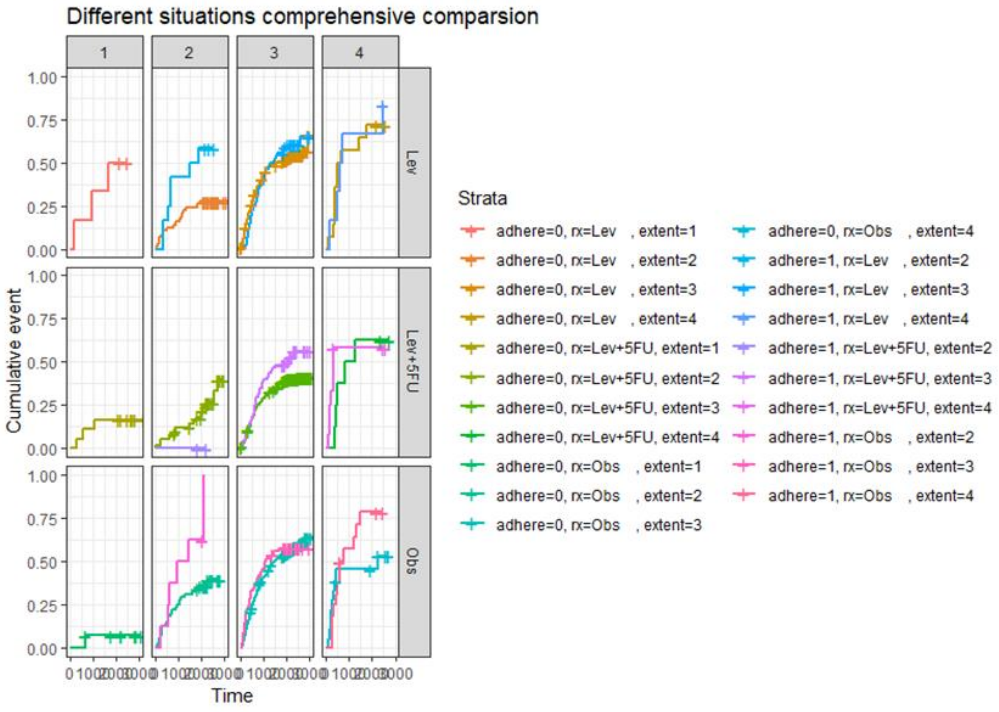


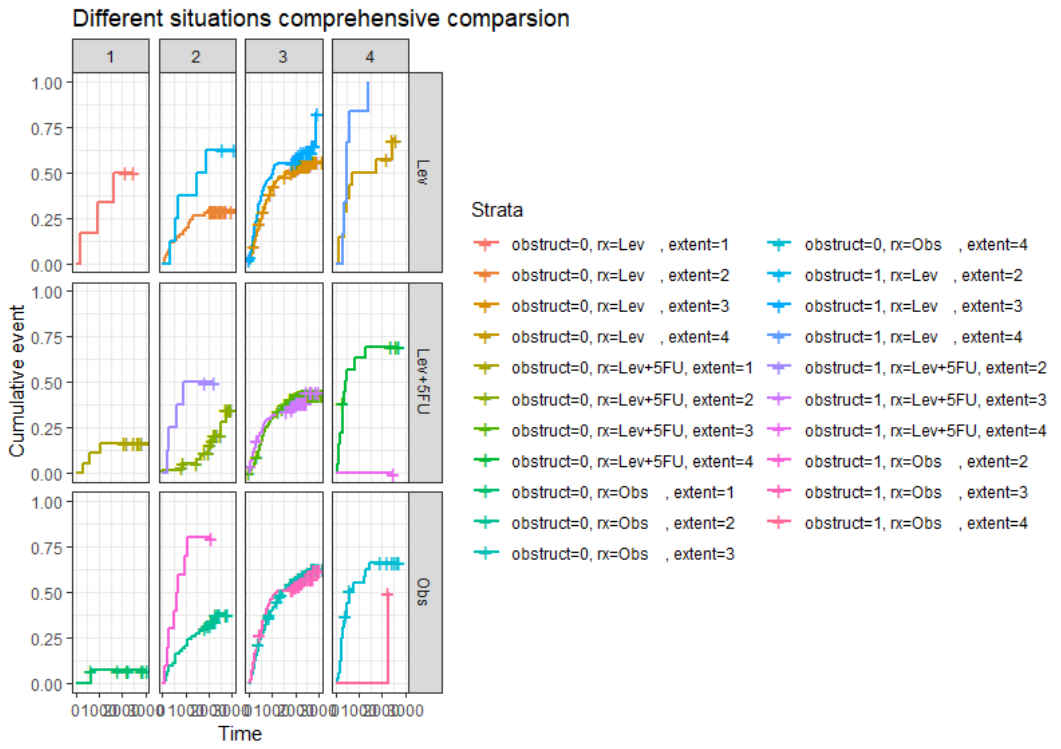


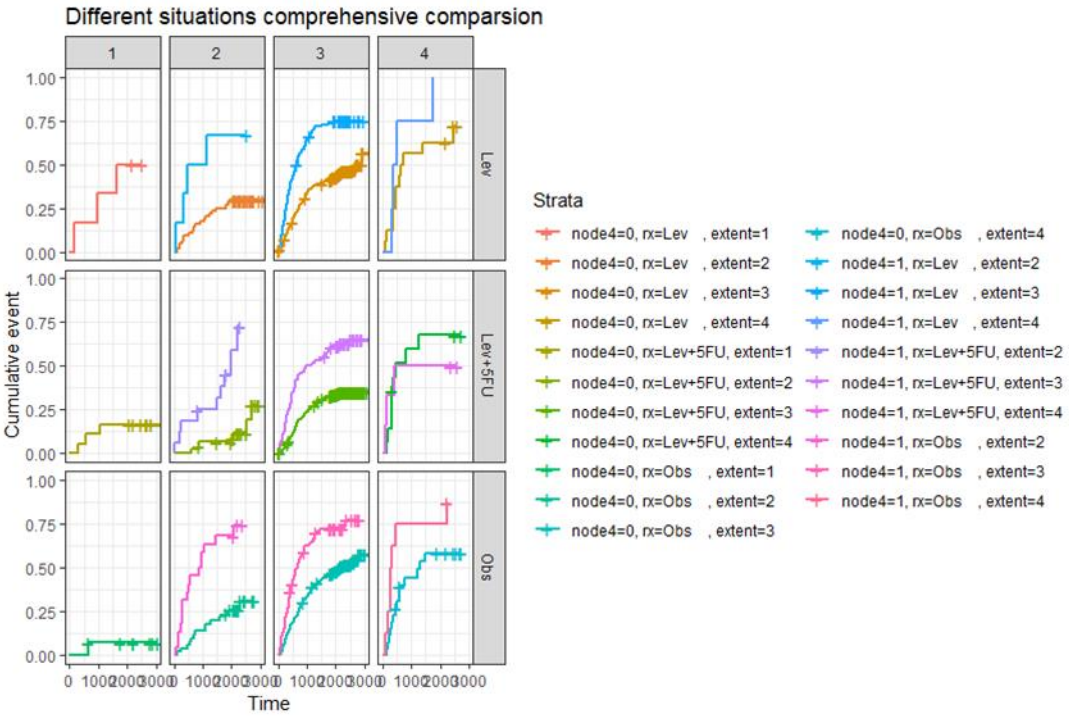
此六項變數在 K-M curve 以及 log-rank test 的 p-value 皆達到顯著水準，代表此項變數在其不同類別上有著明顯不同，因此可以拒絕虛無假設，接下來劃出綜合比較圖(以 rx 及 extent 為核心畫圖)，期望看出一些特性。

Table1：六項變數之 K-M curve 以及 log-rank test

以下為綜合比較圖(以 rx 及 extent 為核心畫圖)加上 adhere、obstruct 及 node4







可以發現在綜合比較圖我們固定治療方式及擴散程度，期望了解 obstruct、adhere、 node4 與他們兩者間的關係，發現符合我們期待，像是在第二期

中，同樣為 Lev 控制療法的 obstruct 有阻塞的確風險較高，也可以觀察 adhere 依從性越好，風險有明顯下降；node4 超過四個陽性淋巴結反應的也是反映出較高的風險；同樣為 Lev+5FU 控制療法的 obstruct 的確有阻塞造成風險較高，adhere 一從性的好與不好也是造成風險高低 ，node4 相同的超過四個陽性淋巴結反應的也是反映出較高的風險。因此我們推論這些變數不僅在 stepAIC 底下被保留，也滿足我們視覺化圖的解釋以及 log rank test 的假設，並且符合我們直覺上的想法，下一步我們給出一些參數在 95%信心水準下的的信賴區間。

Table2：綜合比較圖(以 rx 及 extent 為核心畫圖)

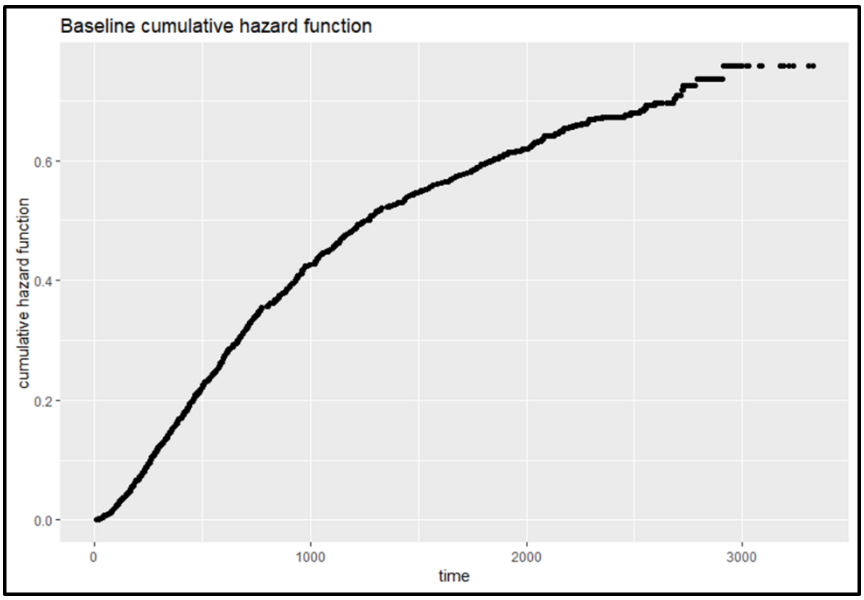


Figure3：baseline cumulative hazard function

累積風險剛開始有較快速的成長，後來漸趨緩和

以下為一般型態模型的參數估計及 Hazard ratio(95% confidence interval)

一般資料模型

95% CI of coefficient

```
> confint(coxmodel)
                  2.5 %      97.5 %
rxLev+5FU -0.54199581 -0.19971776
rxObs     -0.10962540  0.20168732
nodes      0.02117219  0.06320153
obstruct1  0.04947841  0.37688013
adhere1    0.03323464  0.38955760
extent2   -0.40888464  1.15204600
extent3    0.13508476  1.63264198
extent4    0.49712373  2.09798828
surg1      0.09533300  0.38645252
node41     0.42113647  0.81422799
```

95% CI of Hazard ratio

```
> exp(confint(coxmodel))
                2.5 %    97.5 %
rxLev+5FU 0.5815864 0.8189619
rxObs     0.8961698 1.2234654
nodes     1.0213979 1.0652415
obstruct1 1.0507229 1.4577296
adhere1   1.0337931 1.4763275
extent2   0.6643909 3.1646612
extent3   1.1446338 5.1173769
extent4   1.6439859 8.1497584
surg1     1.1000251 1.4717505
node41    1.5236922 2.2574322
```

Table3：參數信賴區間

從模型的信賴區間可以看出，使用 Lex+5-Fu 的 95%信賴區間的指數信賴區間皆小於 1，意及 Lex+5-Fu 能夠有效改善病情，增加存活時間。

3. 檢測一般資料模型下的 local test(檢測 rx 的顯著性)

```
########local test###########

##wald test##
beta1hat = coxmodel$coefficients[1:2]
var11 = coxmodel$var[1:2,1:2]
chi = (beta1hat %>% t)%*%solve(var11)%*% beta1hat # test-statistic
1 - pchisq(chi,2) #chi-square distribution with df 2
##likelihood ratio test##
fit_model.reduced = coxph(Surv(time,status)~ nodes + obstruct + adhere + extent + surg + node4,
                          data=colon)
LR = 2*(coxmodel$loglik[2]-fit_model.reduced$loglik[2])
1 - pchisq(LR,2) #chi-square distribution with df 2
##score test##
fit0 = coxph(Surv(time,status) ~ rx + nodes + obstruct + adhere + extent + surg + node4,data=colon,
           init=c(0,0,fit_model.reduced$coefficients),iter=0) -> coxmodel
score.vector = colSums(coxph.detail(fit0)$score)
chiSC = t(score.vector[1:2])%*%fit0$var[1:2,1:2]%*%score.vector[1:2]
# test-statistic
1 - pchisq(chiSC,2) #chi-square distribution with df 2
```

p-value of wald test	1.698336e-06
p-value of likelihood ratio test	1.008707e-06
p-value of score test	1.417983e-06
所以 rx 是一項顯著的變數，無法被移除	
Table 4：p-value of local test	

4. 以 cox snell residuals plot 及 cox martingale residuals plot，還有 shoenfeld residuals 檢驗一般模型的 cox PH model。

a. cox martingale residuals plot

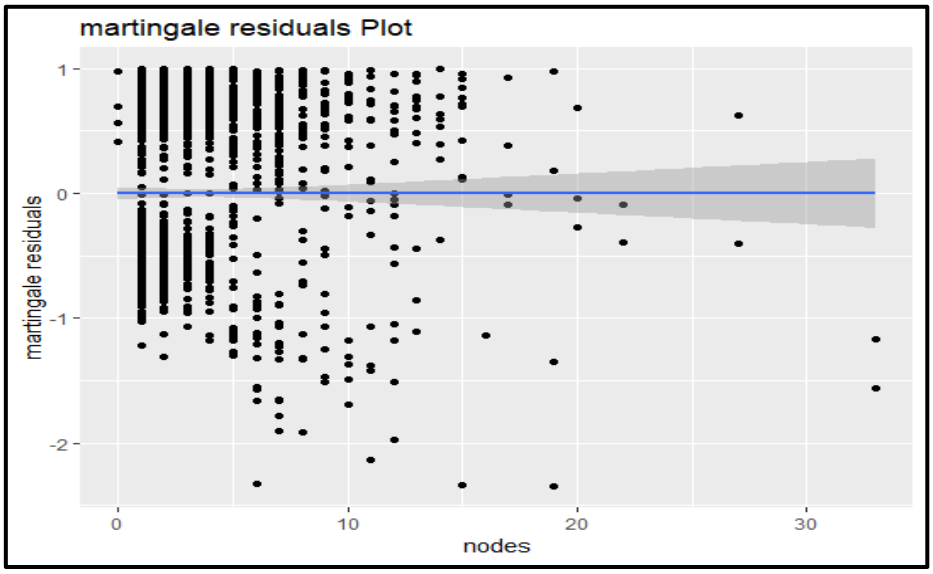


Figure 4：martingale residuals plot

從 martingale residuals plot 來看，似乎沒有任何的趨勢，滿足 cox proportional hazard model 的假設。

b. cox snell residuals plot

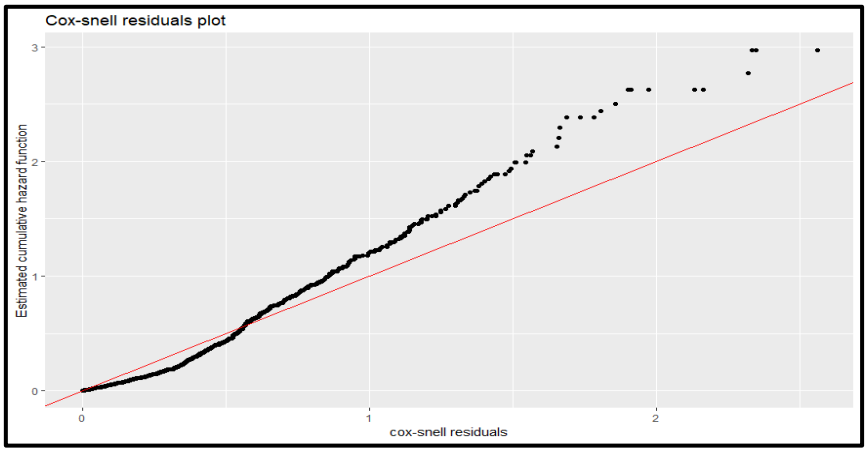


Figure 5：snell residuals plot

從以上 snell residuals plot 可以看出，隨著 cox-snell residuals 越大，偏離 45 度斜直線的幅度越大。也因為不夠符合假設，也使得我們有動機產生 time dependent data 去建構另外的模型。

c. Schoenfeld residuals

```
> cox.zph(coxmodel) -> coxz
> coxz
                rho     chisq        p
rxLev+5FU   0.01448   0.18534 0.666825
rxObs       0.05083   2.34656 0.125560
nodes       0.07843   4.29750 0.038168
obstruct1  -0.10898  10.63258 0.001111
adhere1     0.03404   1.00165 0.316912
extent2     0.02960   0.76867 0.380628
extent3     0.01825   0.29364 0.587896
extent4    -0.00293   0.00751 0.930920
surg1       0.02093   0.38453 0.535186
node41     -0.12876  13.55447 0.000232
GLOBAL           NA  30.64720 0.000671
```

Figure 6：Table of cox Schoenfeld residuals

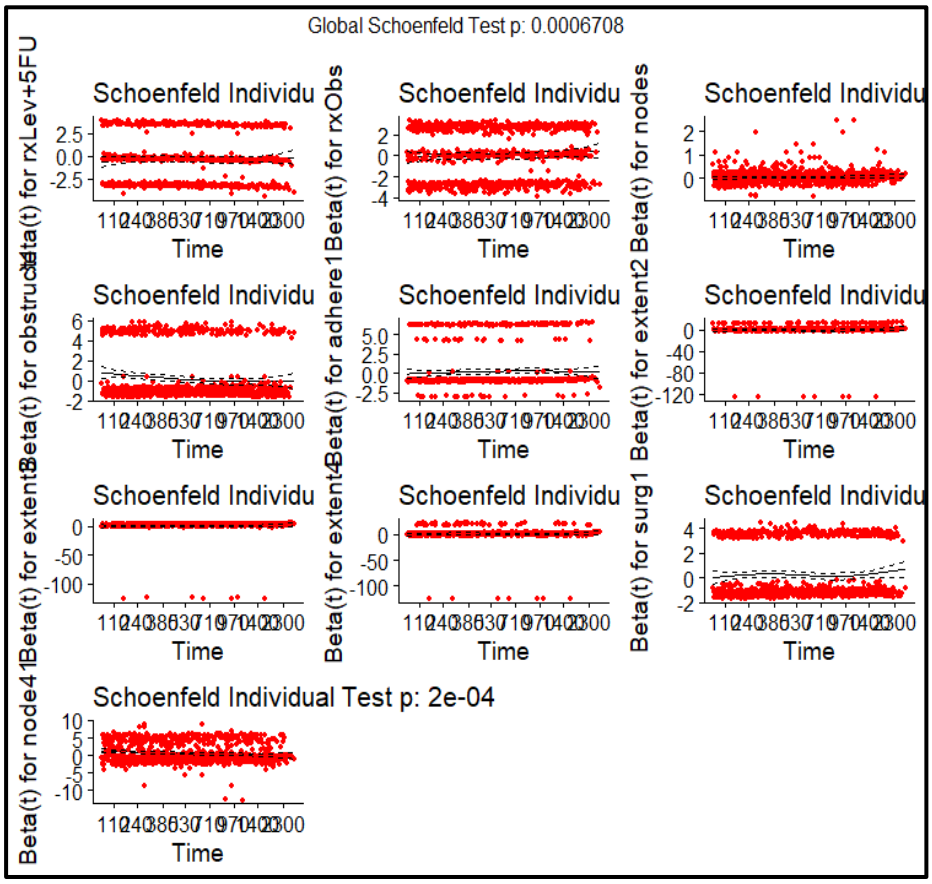


Figure 7：cox Schoenfeld residuals plot

從表格可得知，nodes、obstruct 以及 node4 變數之係數會隨著時間改變而產生變化，因此我們運用 time-dependent covariates 以及 splines 的方式改進。

```
> cox.zph(coxmodel,transform = "log")
                rho     chisq        p
rxLev+5FU   0.00562   0.02792 0.867286
rxObs       0.05116   2.37658 0.123167
nodes       0.06899   3.32510 0.068230
obstruct1 -0.12473  13.92803 0.000190
adhere1     0.03975   1.36584 0.242528
extent2     0.02279   0.45562 0.499680
extent3     0.01544   0.21030 0.646531
extent4     0.00257   0.00581 0.939261
surg1       0.02219   0.43241 0.510811
node41    -0.12517  12.80949 0.000345
GLOBAL          NA  31.74791 0.000441
```

Figure 8：cox schoenfeld residuals with log transformation

```
> coxph(Surv(time,status)~rx + bs(nodes) + obstruct + adhere
+ extent + node4,data = colon) -> modelfit2
> cox.zph(modelfit2)
                 rho    chisq        p
rxLev+5FU   0.016662 2.45e-01 0.620924
rxObs       0.052425 2.48e+00 0.115594
bs(nodes)1 -0.038961 1.34e+00 0.247156
bs(nodes)2  0.022434 4.70e-01 0.493160
bs(nodes)3  0.054585 2.15e+00 0.142531
obstruct1  -0.110183 1.10e+01 0.000899
adhere1     0.035099 1.08e+00 0.297849
extent2     0.031203 8.53e-01 0.355617
extent3     0.021792 4.18e-01 0.517757
extent4    -0.000767 5.16e-04 0.981873
node41     -0.048502 2.01e+00 0.156675
GLOBAL           NA 3.47e+01 0.000279
```

Figure 9：cox schoenfeld residuals with spline function

雖然在 nodes 的 p-value 有明顯改善，但運用兩種轉換方式，無法畫出 cox snell residuals plot 以及 martingale residuals plot，這也是我們再生成 time-

dependent data，再進行建模的動機。

5. 將原始資料轉變為 time dependent model 資料型態

由於觀測到每個 id 都有兩筆資料，分別為 etype=1 及 etype=2。

原始想法為如果此 id 在 etype=2 的資料中其 status=1，表示結果為死亡，因此新增 start 跟 end，若在 etype=2 的情況下其 status=1，則 start 為同 id 在之 etype=1 之 time。以下以 id=1 時為轉換範例：

	age	nodes	id	study	time	status	rx	sex	obstruct	perfor	adhere	differ	extent	surg	node4	etype
1	43	5	1	1	968	1	Lev+5FU	1	0	0	0	2	3	0	1	1
2	43	5	1	1	1521	1	Lev+5FU	1	0	0	0	2	3	0	1	2

Figure 10：Oringinal data

在 id=1 資料中，etype=2 時對應到的 status=1，所以我們將 start 設定為 968。

age	nodes	id	study	time	status	rx	sex	obstruct	perfor	adhere	differ	extent	surg	node4	etype	start	end
43	5	1	1	968	0	Lev+5FU	1	0	0	0	2	3	0	1	1	0	968
43	5	1	1	1521	1	Lev+5FU	1	0	0	0	2	3	0	1	2	968	1521

Figure 11：Time dependent data

	age	nodes	id	study	time	status	rx	sex	obstruct	perfor	adhere	differ	extent	surg	node4	etype	start	end
1	43	5	1	1	968	0	Lev+5FU	1	0	0	0	2	3	0	1	1	0	968
2	43	5	1	1	1521	1	Lev+5FU	1	0	0	0	2	3	0	1	2	968	1521
3	63	1	2	1	3087	0	Lev+5FU	1	0	0	0	2	3	0	0	2	0	3087
4	63	1	2	1	3087	0	Lev+5FU	1	0	0	0	2	3	0	0	1	0	3087
5	71	7	3	1	542	0	Obs	0	0	0	1	2	2	0	1	1	0	542
6	71	7	3	1	963	1	Obs	0	0	0	1	2	2	0	1	2	542	963
7	66	6	4	1	245	0	Lev+5FU	0	1	0	0	2	3	1	1	1	0	245
8	66	6	4	1	293	1	Lev+5FU	0	1	0	0	2	3	1	1	2	245	293
9	69	22	5	1	523	0	Obs	1	0	0	0	2	3	1	1	1	0	523
10	69	22	5	1	659	1	Obs	1	0	0	0	2	3	1	1	2	523	659
11	57	9	6	1	904	0	Lev+5FU	0	0	0	0	2	3	0	1	1	0	904
12	57	9	6	1	1767	1	Lev+5FU	0	0	0	0	2	3	0	1	2	904	1767
13	77	5	7	1	229	0	Lev	1	0	0	0	2	3	1	1	1	0	229
14	77	5	7	1	420	1	Lev	1	0	0	0	2	3	1	1	2	229	420
15	54	1	8	1	3192	0	Obs	1	0	0	0	2	3	0	0	2	0	3192
16	54	1	8	1	3192	0	Obs	1	0	0	0	2	3	0	0	1	0	3192
17	46	2	9	1	3173	0	Lev	1	0	0	1	2	3	0	0	2	0	3173
18	46	2	9	1	3173	0	Lev	1	0	0	1	2	3	0	0	1	0	3173
19	68	1	10	1	3308	0	Lev+5FU	0	0	0	0	2	3	1	0	2	0	3308
20	68	1	10	1	3308	0	Lev+5FU	0	0	0	0	2	3	1	0	1	0	3308
21	47	1	11	1	2908	0	Lev	0	0	0	1	2	3	0	0	2	0	2908
22	47	1	11	1	2908	0	Lev	0	0	0	1	2	3	0	0	1	0	2908
23	52	2	12	1	3309	0	Lev+5FU	1	0	0	0	3	3	1	0	2	0	3309
24	52	2	12	1	3309	0	Lev+5FU	1	0	0	0	3	3	1	0	1	0	3309
25	64	1	13	1	1130	0	Obs	1	0	0	0	2	3	0	0	1	0	1130
26	64	1	13	1	2085	1	Obs	1	0	0	0	2	3	0	0	2	1130	2085
27	68	3	14	1	2231	0	Lev	1	1	0	0	2	3	0	0	1	0	2231
28	68	3	14	1	2910	1	Lev	1	1	0	0	2	3	0	0	2	2231	2910
29	46	4	15	1	2754	0	Obs	1	1	0	0	2	3	0	0	2	0	2754
30	46	4	15	1	2754	0	Obs	1	1	0	0	2	3	0	0	1	0	2754
31	68	1	16	1	3214	0	Obs	1	0	0	0	2	3	1	0	2	0	3214
32	68	1	16	1	1323	1	Obs	1	0	0	0	2	3	1	0	1	0	1323

Figure 12：Time dependent 資料形式

做完資料轉換後，就如同之前的動作與流程，建立模型後進行分析。

6. 建構 time dependent model，並利用 AIC 選取變數

```
Call:
coxph(formula = Surv(start, end, status) ~ nodes + rx + obstruct +
    adhere + extent + surg + node4 + etype, data = colontime)

              coef exp(coef) se(coef)      z        p
nodes      0.05876   1.06052  0.01581  3.716 0.000202
rxLev+5FU -0.40538   0.66672  0.11783 -3.440 0.000581
rxObs      0.02213   1.02237  0.10753  0.206 0.836968
obstruct1  0.19780   1.21872  0.11283  1.753 0.079581
adhere1    0.25598   1.29173  0.12284  2.084 0.037167
extent2    0.22000   1.24608  0.47792  0.460 0.645276
extent3    0.74709   2.11084  0.45342  1.648 0.099417
extent4    1.13836   3.12166  0.49117  2.318 0.020468
surg1      0.30328   1.35430  0.10049  3.018 0.002545
node41     0.54889   1.73133  0.14136  3.883 0.000103
etype2     2.01706   7.51618  0.14613 13.803  < 2e-16

Likelihood ratio test=466.9  on 11 df, p=< 2.2e-16
n= 1772, number of events= 479
```

Figure 13：step AIC 挑選變數結果

觀察參數背後意義以及此資料目的加上比對一般模型的變數，挑選 rx、obstruct、 adhere、 extent、 surg、 node4 的相同變數為主要變數，因為主要為療法為主要變數加以結腸是否堵塞、依從行與否、擴散範圍、手術完來登記於 data 上的時間間隔及是否超過四個顯示陽性的淋巴結，以上皆為常理認知影響重要的變數，而 etype 屬於 time dependent covariates，故不放入模型之中。。

Time dependent model 也與原始模型形式相同，只有係數有變化：

$$\mathbf{h(t|Z) = h_0(t) \times e^{\beta_1 Z_1+\beta_2 Z_2+\beta_3 Z_3+\beta_4 Z_4+\beta_5 Z_5+\beta_6 Z_6+\beta_7 Z_7}}$$

Z_1：factor(rx)，Z_2：nodes，Z_3： obstruct ，Z_4：adhere ，

Z_5： factor(extent) ，Z_6：surg， Z_7：node4

7. Baseline hazard 及參數估計 Hazard ratio(95% confidence interval)

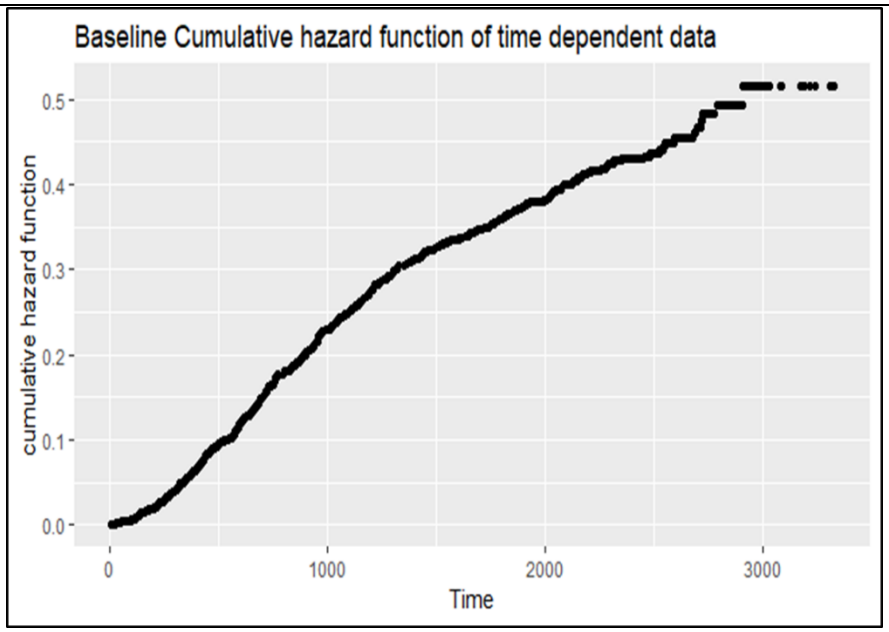


Figure 14：baseline cumulative hazard function

上圖與原始模型相同，前期 0~1000 天風險有較快的成長

Time dependent model

95% CI of coefficient

```
> confint(coxmodeltd)
                2.5 %      97.5 %
rxLev+5FU -0.659524509 -0.19697192
rxObs     -0.170313522  0.25228039
nodes      0.025206267  0.08522657
obstruct1 -0.006600834  0.43606868
adhere1   -0.002749303  0.47976222
extent2   -0.751102224  1.12273478
extent3   -0.165980239  1.61191235
extent4    0.247164066  2.17361606
surg1      0.102150864  0.49560252
node41     0.327616864  0.87260136
```

95% CI of Hazard ratio

```
> exp(confint(coxmodeltd))
                2.5 %    97.5 %
rxLev+5FU 0.5170972 0.8212137
rxObs     0.8434004 1.2869568
nodes     1.0255266 1.0889638
obstruct1 0.9934209 1.5466150
adhere1   0.9972545 1.6156902
extent2   0.4718462 3.0732474
extent3   0.8470630 5.0123875
extent4   1.2803892 8.7900119
surg1     1.1075505 1.6414870
node41    1.3876572 2.3931281
```

Table 5：參數信賴區間

8. 檢測 Time dependent model 下的 local test(檢測 rx 的顯著性)

p-value of wald test	0.0007399988
p-value of likelihood ratio test	0
p-value of score test	7.067156e-05

所以 rx 一樣是一項顯著的變數，無法被移除

```
############time dependent local test#########################

##wald test##
tdhat = modeltd$coefficients[1:2]
vartd = modeltd$var[1:2,1:2]
chitd = (tdhat %>% t) %*% solve(vartd) %*% tdhat # test-statistic
1 - pchisq(chitd,2) #chi-square distribution with df 2
##likelihood ratio test##
td.reduced = coxph(Surv(start,time,status)~ nodes + obstruct + adhere + extent + surg + node4,
                   data=colontime)
LRtd = 2*(modeltd$loglik[2]-td.reduced$loglik[2])
1 - pchisq(LRtd,2) #chi-square distribution with df 2
##score test##
fittd = coxph(Surv(start,end,status) ~ rx + nodes + obstruct + adhere + extent + surg + node4,
              data=colontime,init=c(0,0,td.reduced$coefficients),iter=0)->coxmodel
score.vector.td = colSums(coxph.detail(fittd)$score)
chiSC.td = t(score.vector.td[1:2])%*%fittd$var[1:2,1:2]%*%score.vector.td[1:2]
# test-statistic
1 - pchisq(chiSC.td,2) #chi-square distribution with df 2
```

Table 6： p-value of local test

9. 以 cox snell residuals plot 及 cox martingale residuals plot，還有 shoenfeld residuals 檢驗 time dependent 資料的 cox PH model

   a. cox martingale residuals plot

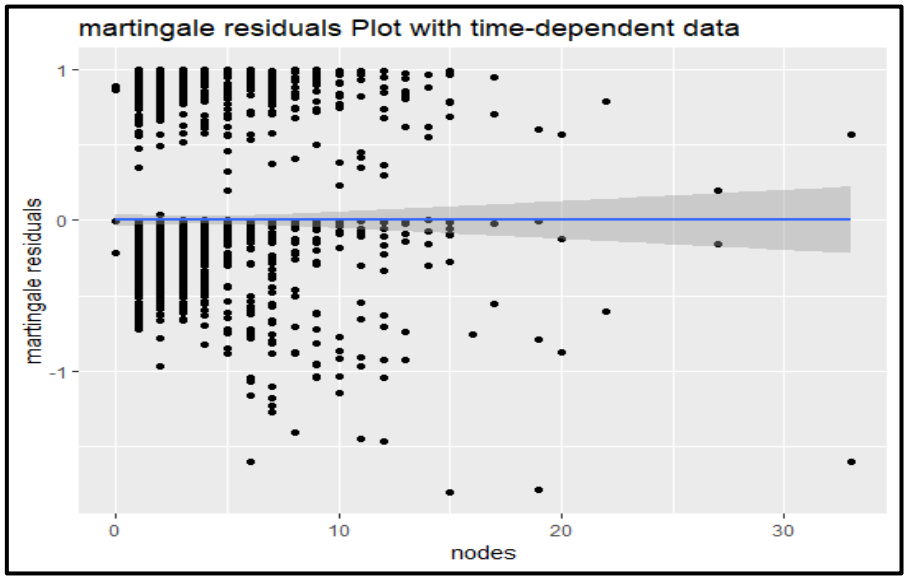


Figure 15：martingale residual plot

b. cox snell residuals plot

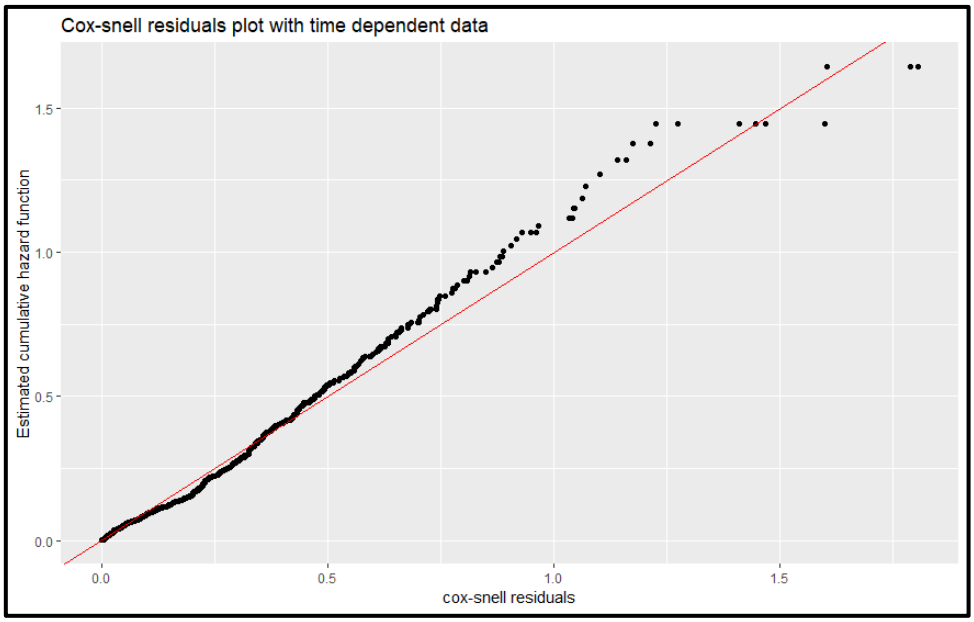


Figure 16：Cox-snell residuals plot

c. schoenfeld residuals

```
> cox.zph(coxmodeltd)
                rho    chisq        p
rxLev+5FU -0.00168  0.00136 0.97055
rxObs      0.08527  3.66226 0.05566
nodes      0.07128  1.98959 0.15838
obstruct1 -0.10983  5.98904 0.01440
adhere1    0.03461  0.57054 0.45005
extent2    0.03850  0.71413 0.39808
extent3    0.01647  0.13137 0.71702
extent4   -0.01550  0.11564 0.73382
surg1     -0.00845  0.03435 0.85297
node41    -0.14856  9.82637 0.00172
GLOBAL           NA 25.74087 0.00410
```

Figure 17：Table of Cox schoenfeld residuals

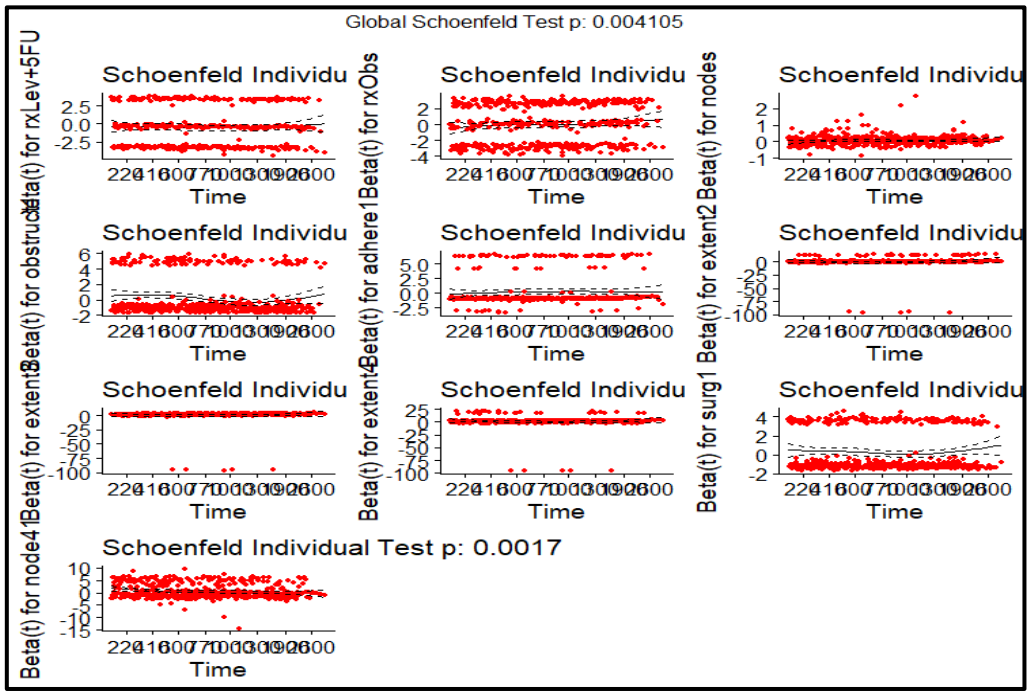


Figure 18：Cox schoenfeld residuals plot

比較原先的殘差圖，time dependent model 的 Cox snell residuals plot 以及 Cox Schoenfeld residuals plot 顯示 time dependent 有明顯的改善。因此 time dependent model 是我們的最終模型。

陸、 結論

此資料主要目的為分析兩種治療藥物是否影響存活時間，無論是原始模型還是我們後面建構的 time dependent model，從 local test 的結果來看，是否服用藥物是影響存活時間的重要因素，而不同藥物的 survival curve 也差距甚遠(p-value < 0.001)，更顯示 Lev+5-Fu 兩種藥物比起單一藥物治療方法及無使用治療方法，能更增加存活時間，也符合我們的預期結果。

此外其他變數也相同影響著我們預測的存活函數，從模型檢驗的 p-value 都可得知有明顯的顯著性，之後做出各項檢定發現原始模型，不夠完美即從 snell residuals plot 可以看出，隨著 cox-snell residuals 越大，偏離 45 度斜直線的幅度越

大。也因為不夠符合假設，加上使用 spline 改變時間相依變數雖然在 nodes 的 p-value 有明顯改善，但運用兩種轉換方式，無法畫出 cox snell residuals plot 以及 martingale residuals plot，這也是我們再生成 time-dependent data，再進行建模的動機。

建出 time dependent model 後，比較原先的殘差圖，time dependent model 的 Cox snell residuals plot 以及 Cox Schoenfeld residuals plot 顯示 time dependent 有明顯的改善，例如 Cox snell residuals plot 更貼近 45 度線及 Cox Schoenfeld residuals 檢定 p-value 有顯著改善，因此 time dependent model 是我們的最終模型。

柒、 參考資料

1. B / C 期結腸癌的化療
https://stat.ethz.ch/R-manual/R-devel/library/survival/html/colon.html?fbclid=IwAR2FPEADXaq0feNj43jfzOphYA2A1_H1WwvzGiTRmIEWhtsDL3BpjQB549Y
2. 左旋咪唑藥物
https://zh.wikipedia.org/wiki/%E5%B7%A6%E6%97%8B%E5%92%AA%E5%94%91?fbclid=IwAR2KTZhbtbxsMx033aam143MYLwoKHWlUaXINfgj-rEEdYUbgMPEIVLZaTQ
3. Residual diagnostics in Cox PH model
https://rpubs.com/kaz_yos/resid_cox?fbclid=IwAR1lKh6b-BH93eu5nxKBT6Y60JGeBBP3Va1ItxrHJp9Z-TLaadYbN26IWxg
4. The Estimation of Survival Function for Colon Cancer Data in Tehran Using Non-parametric Bayesian Model
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4142923/?fbclid=IwAR2tqA39y22-IpkK4y1VUoRQ3tUp45y5a_Dt-5inf4vYyTRrMQOhxFRNufI
5. K-M curve visualization

https://moderndata.plot.ly/pharmaceutical-survival-interactive/?fbclid=IwAR18VCw_B7mws2swl4w8xUJvNrf_M2R1SOyhqaUMIECRwsuabdWj4smA8to