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

### 目錄

[壹、 目錄 2](#_Toc11700912)

[貳、 摘要 3](#_Toc11700913)

[參、 前言 3](#_Toc11700914)

[肆、 資料背景及變數介紹 4](#_Toc11700915)

[伍、 研究方法與進行步驟 5](#_Toc11700916)

[陸、 結論 21](#_Toc11700917)

[柒、 參考資料 22](#_Toc11700918)

### 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。

### 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。

### 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。

1. 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 |

### 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：

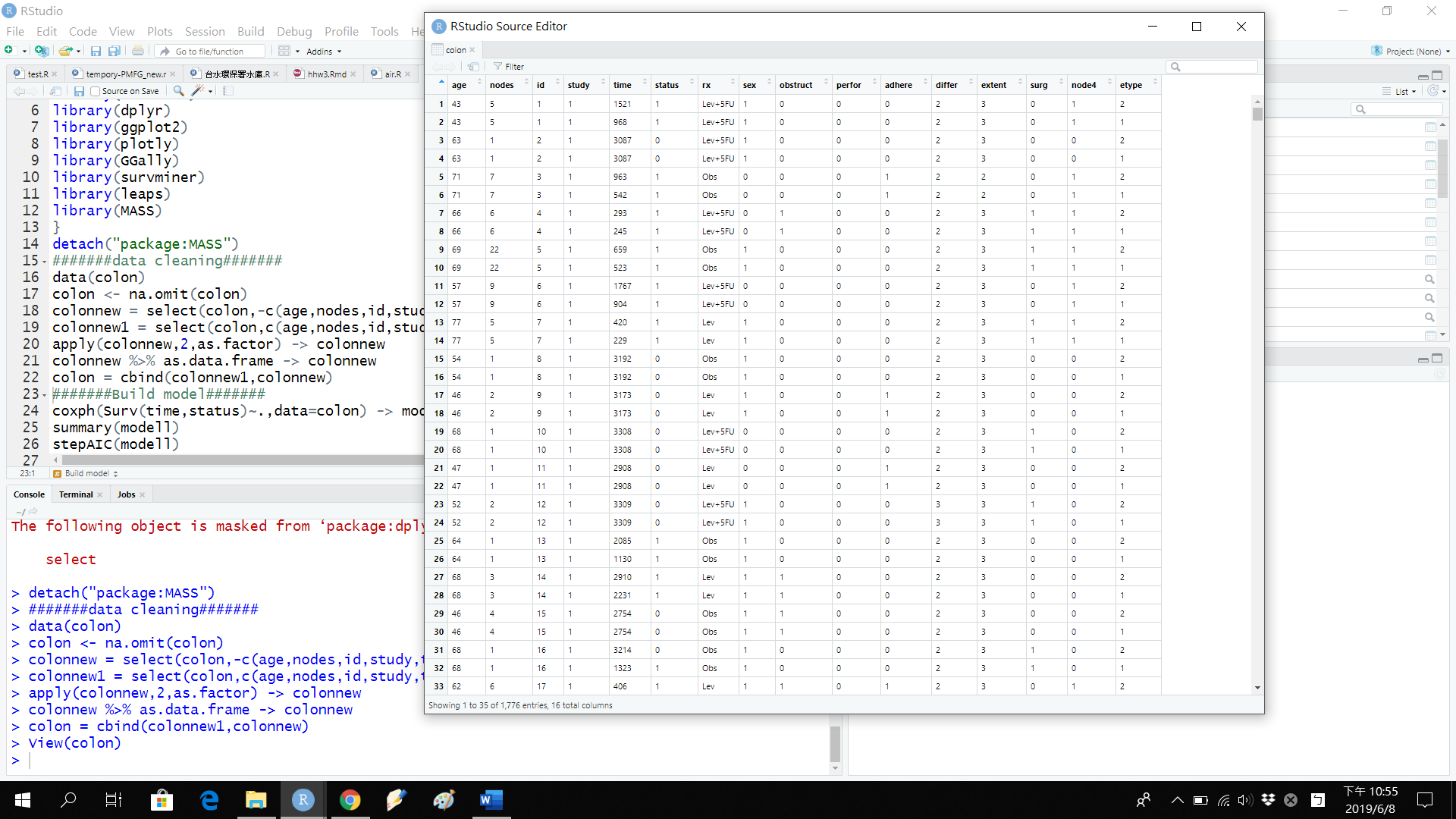
1. 一般模型資料樣貌：

Figure1：一般資料形式

1. 參數選取(利用AIC挑選參數)---一般模型

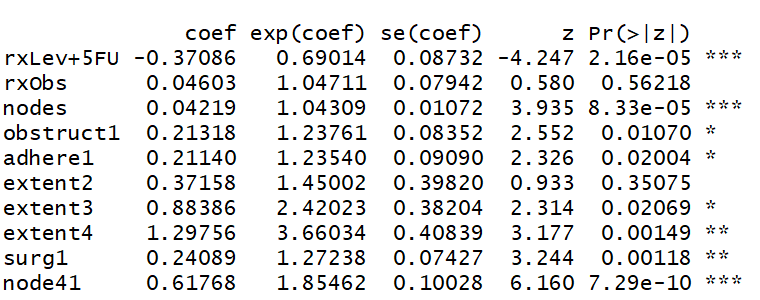
 觀察參數背後意義以及此資料目的，挑選rx、 obstruct、 adhere、 extent、 surg、 node4為主要變數，因為主要為療法為主要變數加以結腸是否堵塞、依從行與否、擴散範圍、手術完來登記於data上的時間間隔及是否超過四個顯示陽性的淋巴結，以上皆為常理認知影響重要的變數，之後對這些變數化出K-M curve及其log-rank test，期望看出此項變數在不同類別的差異性。

Figure2：step AIC挑選變數結果

原始模型為：

**factor(rx)，nodes， obstruct ，：adhere ，**

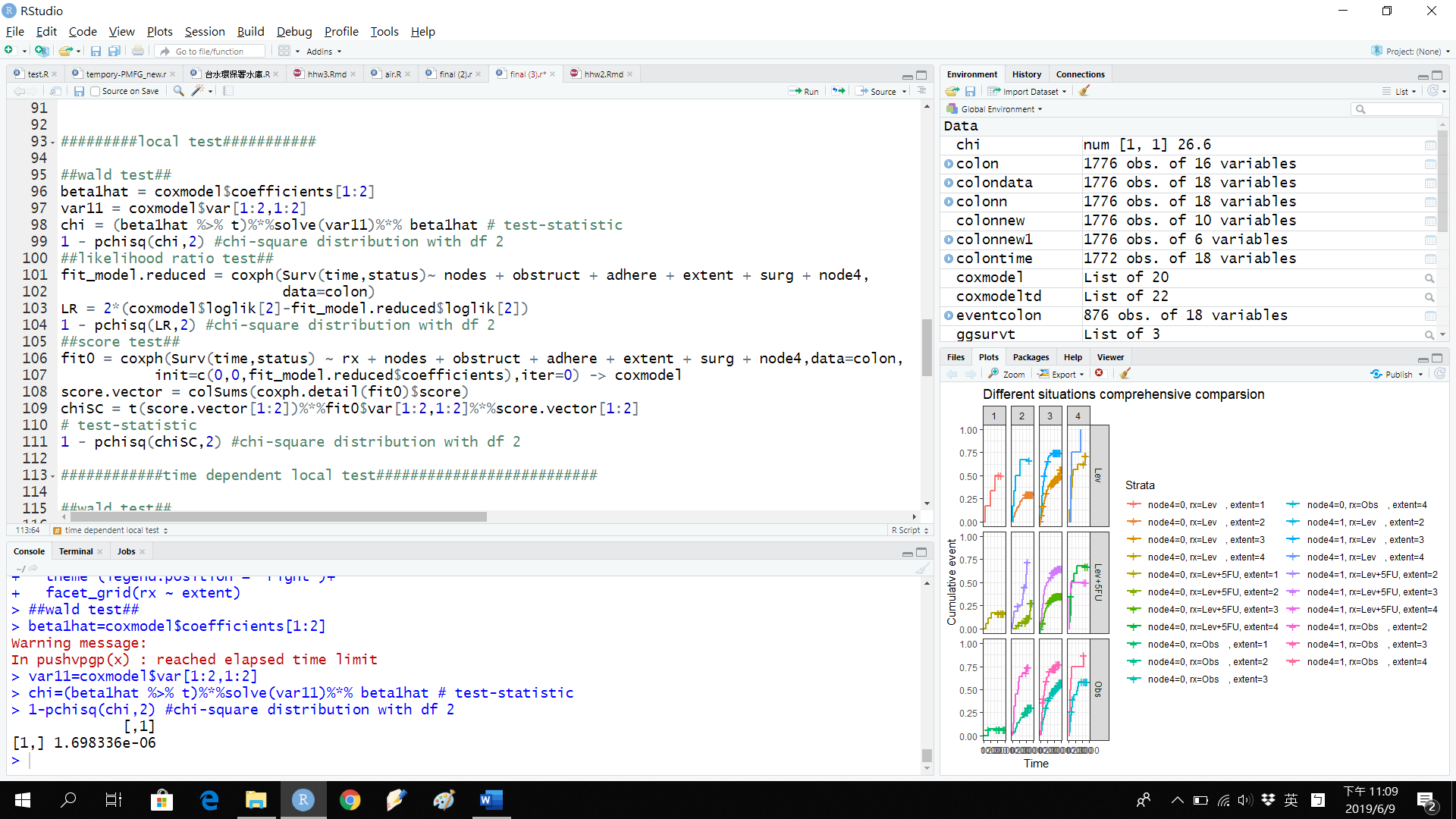
**factor(extent) ，surg， node4**

舉出兩個變數來解釋模型：

**，**意即若使用 Lev+5-FU 兩種藥物進行治療，則風險降低近 30%。，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 | |
|  | 一張含有 地圖, 文字 的圖片  自動產生的描述 |
| 一張含有 文字, 地圖 的圖片  自動產生的描述 | 一張含有 文字, 地圖 的圖片  自動產生的描述 |
| 一張含有 地圖, 文字 的圖片  自動產生的描述 | 一張含有 文字, 地圖 的圖片  自動產生的描述 |
| 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 | |
|  | |
| 一張含有 文字, 地圖 的圖片  自動產生的描述 | |
|  | |
| 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 | 95% CI of Hazard ratio |
|  |  |
| 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。 | |

1. 檢測一般資料模型下的local test(檢測rx的顯著性)

|  |  |
| --- | --- |
| 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 | |

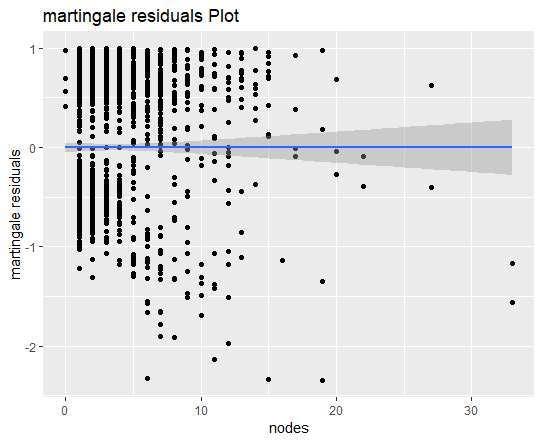
1. 以cox snell residuals plot及cox martingale residuals plot，還有shoenfeld residuals檢驗一般模型的cox PH model。
   * + 1. cox martingale residuals plot

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。

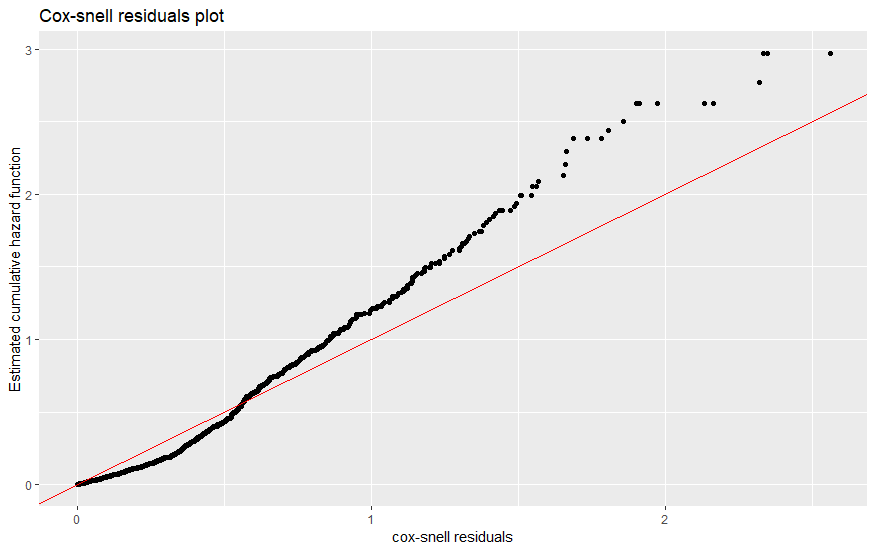
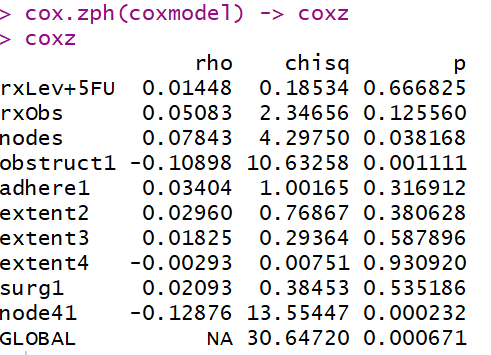
* + - 1. cox snell residuals plot

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。

* + - 1. Schoenfeld residuals



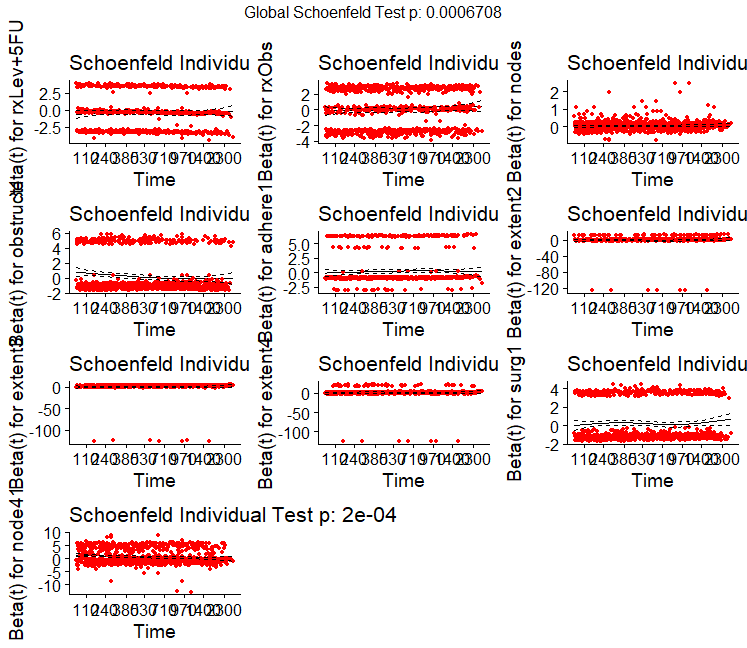
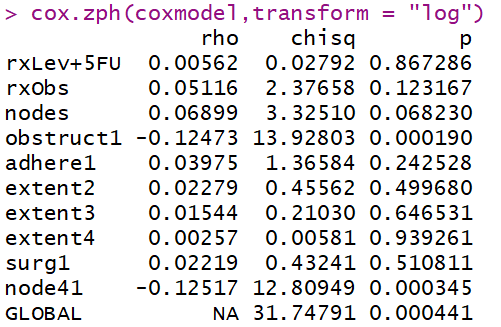
Figure 6：Table of cox Schoenfeld residuals

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。

一張含有 文字 的圖片

自動產生的描述Figure 8：cox schoenfeld residuals with log transformation

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。

1. 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。

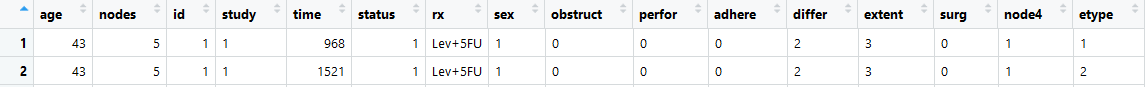
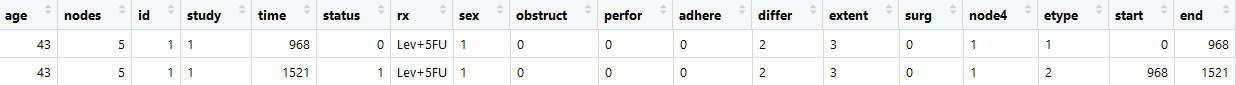
原始想法為如果此id在etype=2的資料中其status=1，表示結果為死亡，因此新增start跟end，若在etype=2的情況下其status=1，則start為同id在之etype=1之time。以下以id=1時為轉換範例：

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。

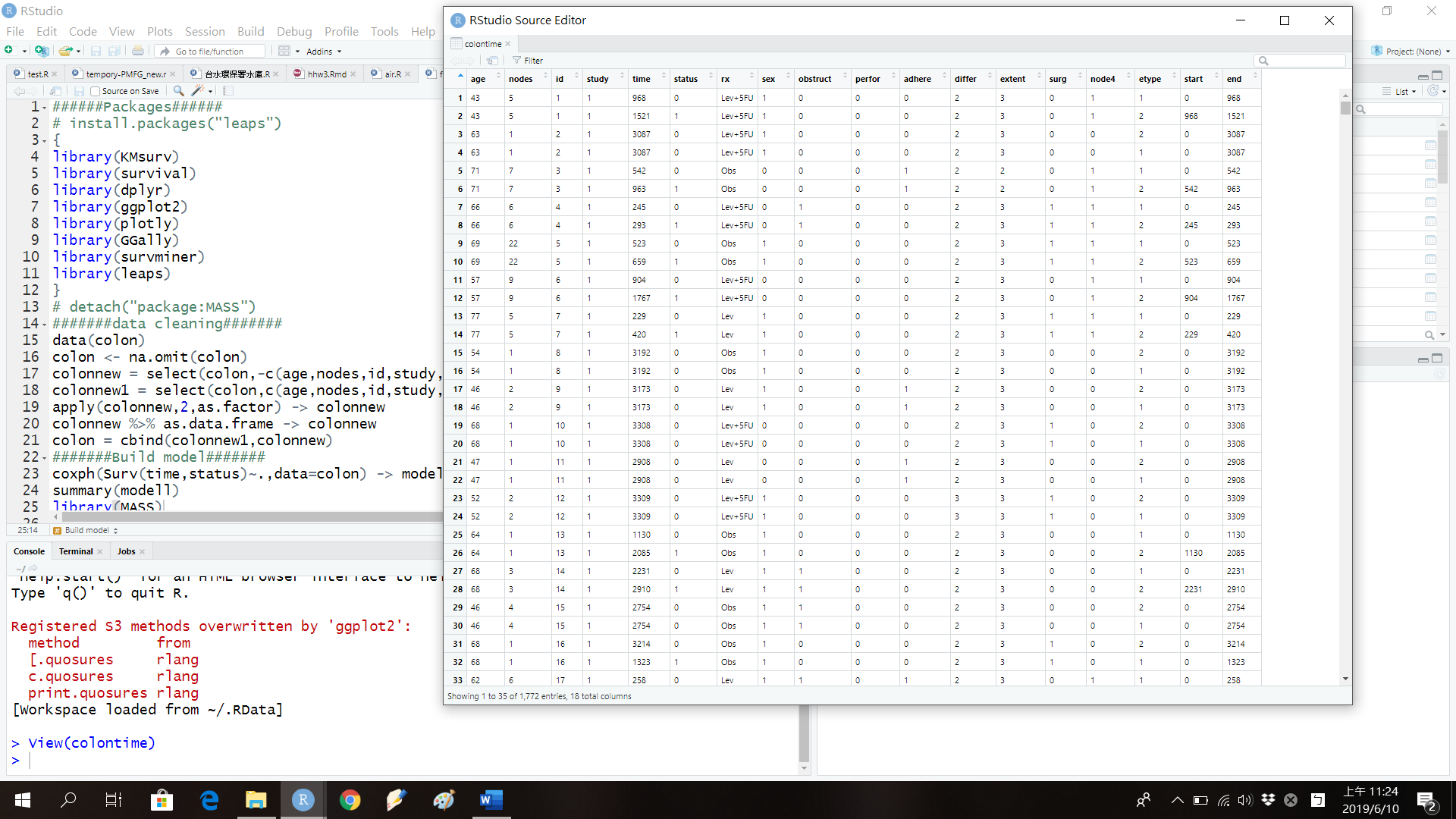
Figure 11：Time dependent data

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。

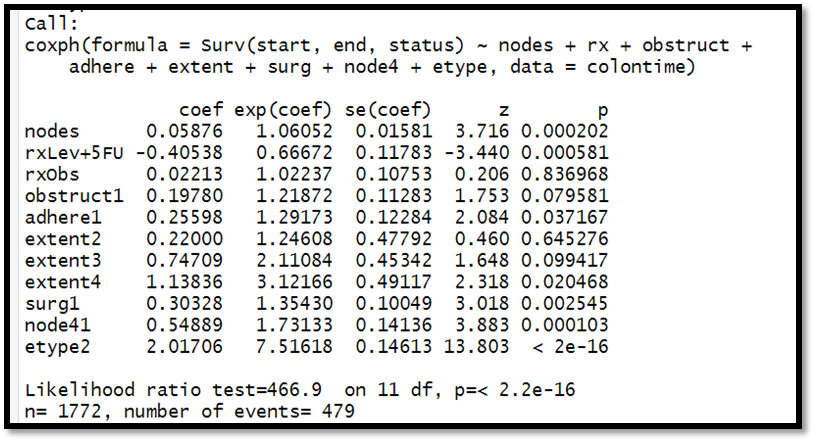
1. 建構time dependent model，並利用AIC選取變數

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：

**factor(rx)，nodes， obstruct ，：adhere ，**

**factor(extent) ，surg， node4**

1. 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 | 95% CI of Hazard ratio |
|  |  |
| Table 5：參數信賴區間 | |

1. 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 | |
| Table 6： p-value of local test | |

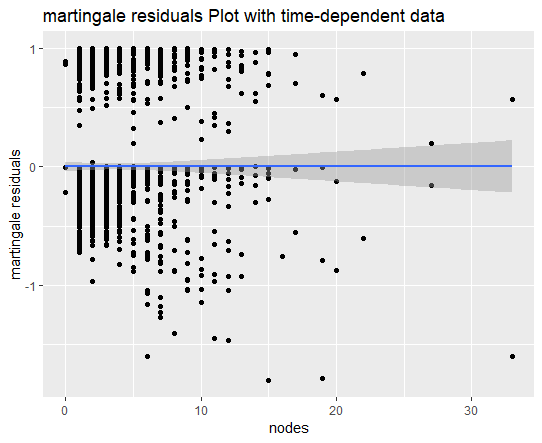
1. 以cox snell residuals plot及cox martingale residuals plot，還有shoenfeld residuals檢驗time dependent資料的cox PH model
   * + 1. cox martingale residuals plot

Figure 15：martingale residual plot

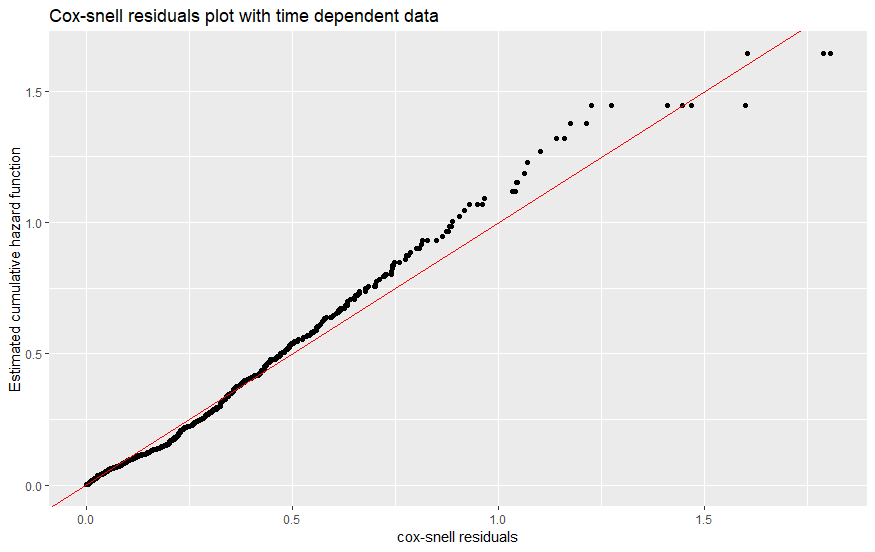
* + - 1. cox snell residuals plot

Figure 16：Cox-snell residuals plot

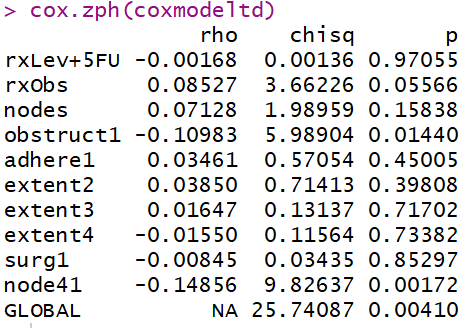
* + - 1. schoenfeld residuals

Figure 17：Table of Cox schoenfeld residuals

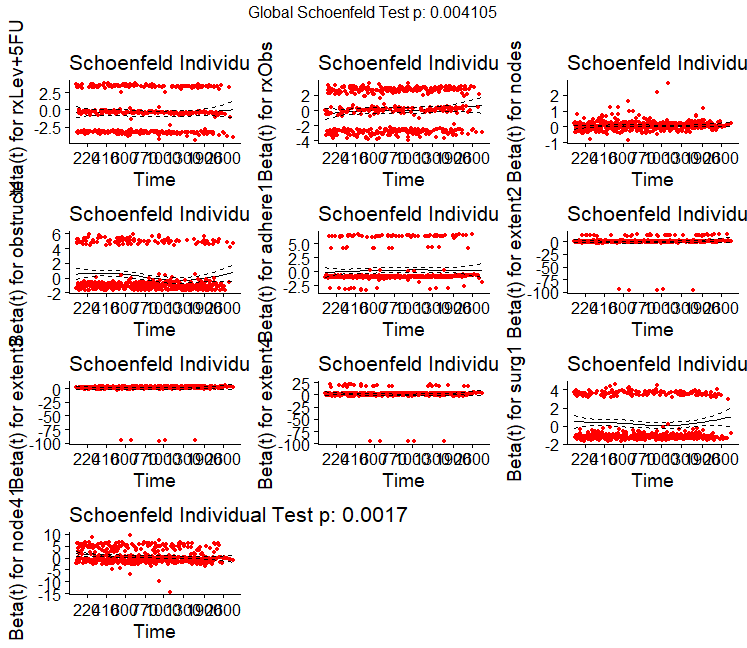


Figure 18：Cox schoenfeld residuals plot

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。

### 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。

### 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>

1. 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>

1. Residual diagnostics in Cox PH model

<https://rpubs.com/kaz_yos/resid_cox?fbclid=IwAR1lKh6b-BH93eu5nxKBT6Y60JGeBBP3Va1ItxrHJp9Z-TLaadYbN26IWxg>

1. 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>

1. K-M curve visualization

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