存活分析期末報告

主題:膀胱癌分析

目錄: 一.研究動機與資料介紹

二.模型應用及變數解釋

三.模型基本假設

四.結論

一.研究動機與資料介紹

1.膀胱癌介紹: 膀胱癌是常見的泌尿系統惡性腫瘤，最常見的症狀包括血尿、頻尿和排尿困難。它通常發生在 60 歲以上的男性身上，且具有高復發率。但只要通過定期覆診，治癒的機會相對其他癌症會高一些。

2.研究動機: 膀胱是我們日常生活中很重要的器官，然而膀胱癌確切的成因尚未了解，所以想要透過簡單的分析來更了解有什麼因素會影響膀胱癌的存活。

3.資料介紹:本研究的資料是是來自於 MSK 癌症中心 2023 年的臨床數據，原始資料總共有 526 筆

4.變數介紹

變數名稱	變數解釋
Cancer.Type.Detailed	Bladder Urothelial Carcinoma(膀胱尿路上皮癌) Upper Tract Urothelial Carcinoma(上泌尿道尿路上皮癌) Urethral Urothelial Carcinoma(尿道尿路上皮癌)
Age	診斷年紀
Fraction.Genome.Altered(FGA)	拷貝數受影響的百分比
MSI.Score	微星體不穩定性的分數
Mutation.Count	基因突變數量
TMB	腫瘤突變負荷
Sex	性別
Overall.Survival..Months	存活時間(月)
Overall.Survival.Status	1:事件發生，0:censoring

Cancer type:是一個類別變數，那他會以發生部位的尿路上皮癌來做分類
分別佔資料的比例為 76.4%、21.5%、1.9%。

Age:是一個連續變數，範圍落在 19~90。

FGA:是一個連續變數，拷貝數的意思為基因或染色體片段在基因組中的個數。其範圍落在 0~0.7016。

MS.Score:是一個連續變數，微星體不穩定性的意思是 DNA 複製過程發生錯誤，導致重複序列發生改變。其範圍落在 -1~32。

Mutation.Count:是一個連續變數，除了一個樣本為 414，其他範圍落在 1~101

TMB:是一個連續變數，其為百萬鹼基對中，腫瘤基因突變的數量。除了一個樣本為 405.18995，其他範圍落在 0~ 84.16989。

Sex:是一個類別變數，男性 265 位跟女性 126 位

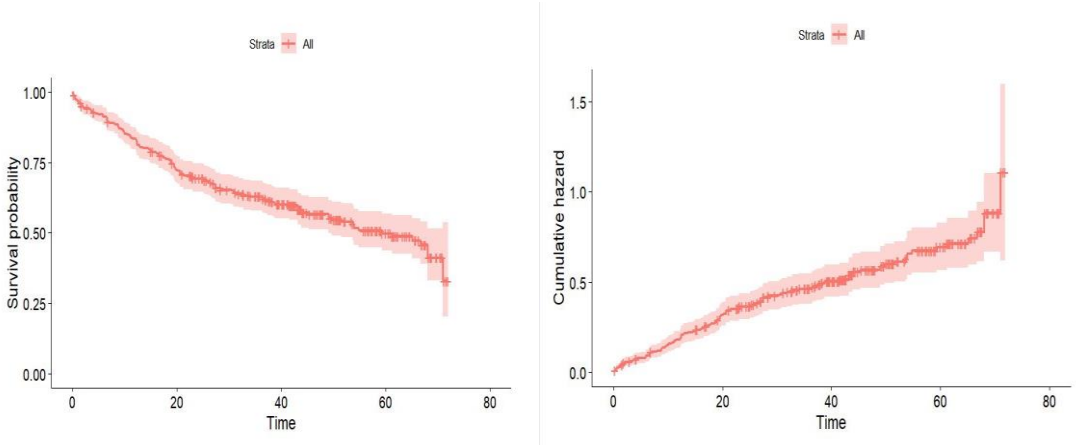
Overall.Surivival..Months:是一個連續變數，以月為單位，範圍落在 0~105.292

Overall.Survival.Status:是一個類別變數，事件發生有 280 位，censored 為 205 位。

5.變數處理:我將全部有 NA 值的病患資料直接刪除，且設一個截止時間為 72 個月，那我們的研究時間就是從病人第一次診斷到 6 年後，那經過處理後我們的資料剩餘 338 筆

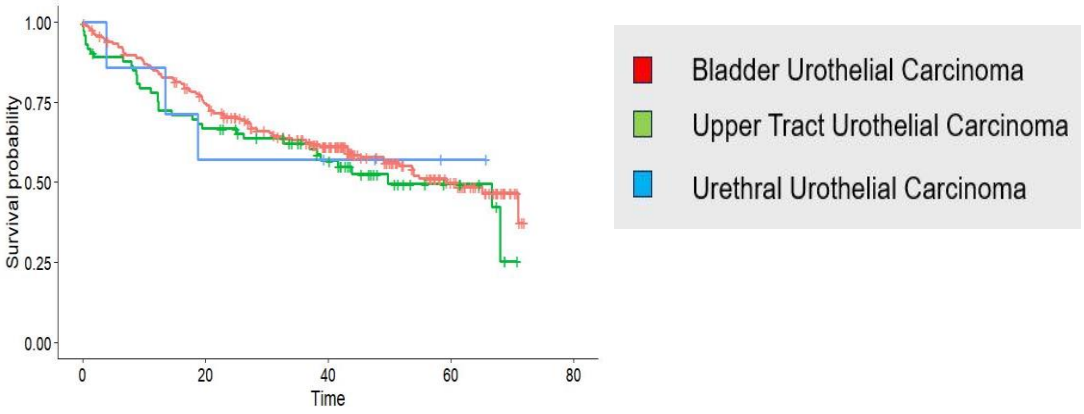
二.模型應用及變數解釋

1.Kaplan Meier S(t) and Nelson Aalen H(t)



可以看到兩種方法的估計出來的 S(t)是遞減的，H(t)是遞增的，所以是沒有太大問題的。我們可以從 survival function 的圖能得到一些資訊，因為通常癌症對於存活率都是以五年為準，那就會對應到我們 time=60 的存活率，可以看到還有 50%，就可以利用這個值去跟別的癌症做比較。

2.Log-rank test



主要想知道的是不同類別的尿路上皮癌是否會對存活的影響是有差別的，那圖中顯示三種類別基本上都是交疊在一起，所以我們會判斷這三種的影響是沒有差別的。從圖我們還能得知截止時間時，上泌尿道尿路上皮癌存活率是最低的。

$$H_0 : h_1(t) = h_2(t) = h_3(t)$$

```
Call:
survdiff(formula = Surv(event, time_new) ~ factor(cancer), data = all)

                                                N Observed Expected (O-E)^2/E (O-E)^2/V
factor(cancer)=Bladder Urothelial Carcinoma   258      110   115.47   0.25938   1.15910
factor(cancer)=Upper Tract Urothelial Carcinoma  73       36    30.46   1.00588   1.27142
factor(cancer)=Urethral Urothelial Carcinoma      7        3     3.06   0.00129   0.00132

 Chisq= 1.3  on 2 degrees of freedom, p= 0.5
```

從檢定來看，虛無假設就是三個類別的風險函數是一致的，那得出的 p value >0.05，也就是三種類別確實對風險的影響是沒有差別的。

3. Cox proportional hazards model

3.1 原始模型:

```
Call:
coxph(formula = Surv(event, time_new) ~ factor(cancer) + factor(sex) +
    age + Genome + MSI + Mutation + tmb, data = all)

  n= 338, number of events= 149

                                             coef exp(coef)  se(coef)      z Pr(>|z|)
factor(cancer)Upper Tract Urothelial Carcinoma  0.305145  1.356822  0.197696  1.544 0.122708
factor(cancer)Urethral Urothelial Carcinoma     0.373815  1.453268  0.595900  0.627 0.530456
factor(sex)Male                                 0.279171  1.322033  0.183597  1.521 0.128370
age                                             0.034195  1.034786  0.008518  4.015 5.96e-05 ***
Genome                                          2.025266  7.578124  0.562070  3.603 0.000314 ***
MSI                                            -0.128048  0.879811  0.053231 -2.406 0.016150 *
Mutation                                       -0.289414  0.748702  0.075134 -3.852 0.000117 ***
tmb                                             0.293592  1.341236  0.077716  3.778 0.000158 ***

Concordance= 0.662  (se = 0.022 )
Likelihood ratio test= 63.47  on 8 df,   p=1e-10
Wald test            = 43.57  on 8 df,   p=7e-07
Score (logrank) test = 46.74  on 8 df,   p=2e-07
```

我們可以看到一致性的分數為 0.622，且下面三個 test 的 p value 都是非常小的，代表這個模型比什麼變數都不放的模型擬合的要好。但當我們看到變數的 p value 除了類別變數以外都是很小的，那有可能會有隱藏的共線性問題

3.2 查看共線性問題:

	GVIF	Df	GVIF^(1/(2*Df))
factor(cancer)	1.092636	2	1.022395
factor(sex)	1.043605	1	1.021570
age	1.008580	1	1.004281
Genome	1.158694	1	1.076427
MSI	1.450463	1	1.204352
Mutation	241.757186	1	15.548543
tmb	247.401589	1	15.729005

	GVIF	Df	GVIF^(1/(2*Df))
factor(cancer)	1.069990	2	1.017056
factor(sex)	1.033671	1	1.016696
age	1.012368	1	1.006165
Genome	1.062969	1	1.031004
MSI	1.224292	1	1.106477
Mutation	1.180215	1	1.086377

這邊是利用廣義方差膨脹因子(GVIF)，那確實可以看到在 Mutation 跟 tmb 這兩個變數有嚴重的共線性關係，那我這邊處理的方式就是把 GVIF 最高的 tmb 直接移出模型。

3.3 新模型與解釋變數:

```
Call:
coxph(formula = Surv(event, time_new) ~ factor(cancer) + factor(sex) +
    age + Genome + MSI + Mutation, data = all)

  n= 338, number of events= 149

                                          coef exp(coef)  se(coef)      z Pr(>|z|)
factor(cancer)Upper Tract Urothelial Carcinoma  0.360678  1.434302  0.196445  1.836  0.06635 .
factor(cancer)Urethral Urothelial Carcinoma     0.157291  1.170336  0.593617  0.265  0.79103
factor(sex)Male                                 0.227266  1.255164  0.182692  1.244  0.21350
age                                             0.035824  1.036473  0.008699  4.118 3.82e-05 ***
Genome                                          1.497989  4.472684  0.532815  2.811  0.00493 **
MSI                                            -0.070067  0.932331  0.052785 -1.327  0.18437
Mutation                                       -0.020311  0.979894  0.010026 -2.026  0.04278 *

Concordance= 0.653  (se = 0.023 )
Likelihood ratio test= 49.62  on 7 df,   p=2e-08
Wald test            = 36.76  on 7 df,   p=5e-06
Score (logrank) test = 38  on 7 df,   p=3e-06
```

Cancer:以癌症的類別來說，baseline 為膀胱尿路上皮癌，那可以得知其他兩風險都是比它高的

Sex:以性別來說，baseline 為女性，那可以看到男性的風險是比女性高的，那這是與以往研究是符合的。

Age:年齡越高，風險是越高的，不管什麼癌症基本上都有這種情況。

Genome(FGA):可以特別注意到這個變數的 hazard ratio 為 4.472，且是顯著的，這代表他對風險的影響是非常大的。

MSI:這個變數是不顯著

Mutation:突變數增加，風險是降低的

4.Local test

變數名稱	膀胱癌類別	性別	MSI
LRT test(p 值)	0.2023	0.2071	0.1157
Wald test(p 值)	0.1842	0.2135	0.1844

```
Call:
coxph(formula = Surv(event, time_new) ~ MSI, data = all)

  n= 338, number of events= 149

       coef exp(coef) se(coef)     z Pr(>|z|)
MSI -0.08655   0.91709  0.03797 -2.28   0.0226 *
```

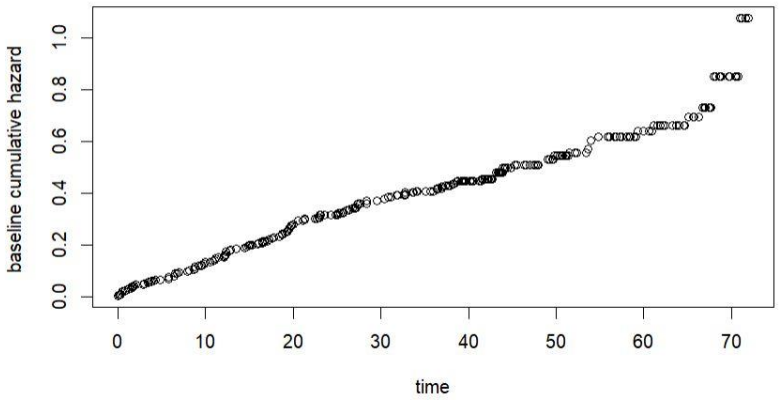
可以看到這幾個變數的 p value 都是大於 0.05 的，表示這些變數對存活是沒有什麼影響的。以 MSI 來說雖然 Local test 不顯著，但是單獨用它做 PH model 的時候，它的 p value 是顯著的，那這邊可能的原因是因為它的效應已經被其他變數解釋完了。那後續我們就會把這些不顯著的變數都拿掉來做後續的分析。

5.最終模型

```
Call:
coxph(formula = Surv(event, time_new) ~ age + Genome + Mutation,
    data = all)

  n= 338, number of events= 149

              coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.037501  1.038213  0.008493  4.416 1.01e-05 ***
Genome    1.393367  4.028389  0.518259  2.689  0.00718 **
Mutation -0.027374  0.972997  0.009022 -3.034  0.00241 **

Concordance= 0.638  (se = 0.023 )
Likelihood ratio test= 43.74  on 3 df,   p=2e-09
Wald test            = 34.29  on 3 df,   p=2e-07
Score (logrank) test = 30.28  on 3 df,   p=1e-06
```

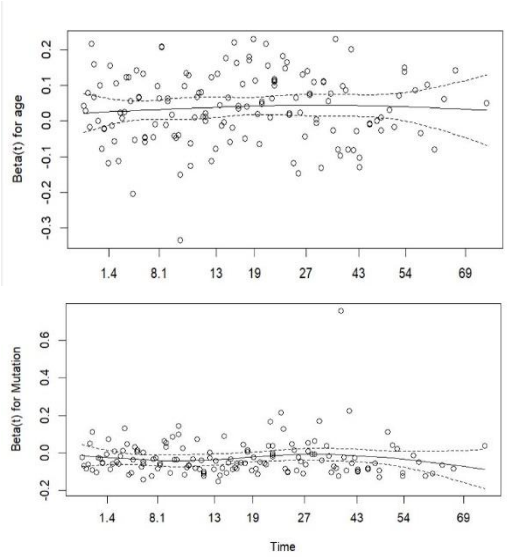
三.模型的基本假設

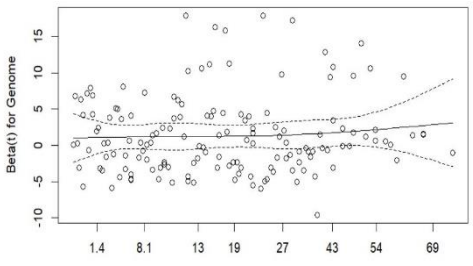
1.Breslow ‘s baseline cumulative hazard function



從圖可以看到點呈現一條直線，所以我們可以知道不同時間的的 baseline hazard function 都是一樣的。

2. Schonfeld residuals



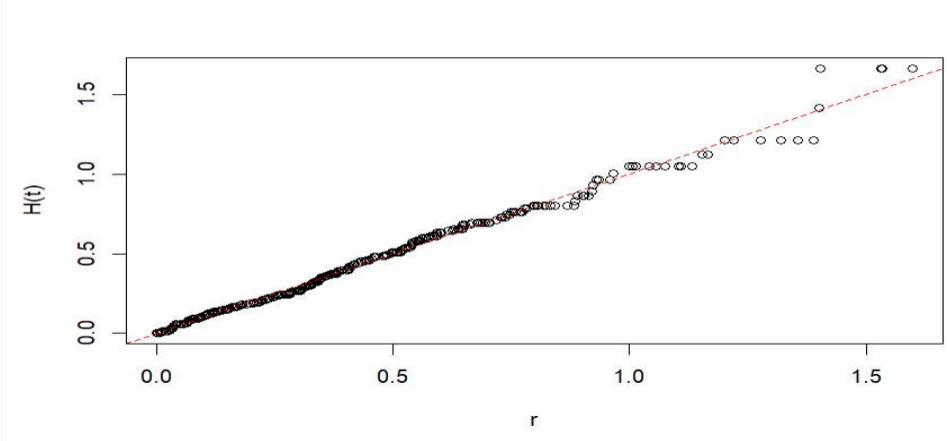


	chisq	df	p
age	0.4226	1	0.52
Genome	0.5711	1	0.45
Mutation	0.0774	1	0.78
GLOBAL	1.0511	3	0.79

我們從每個變數的圖來看，3 個變數都沒有什麼太大問題，而且從檢定的 p

value 來看，每個變數都 p 值都是很大的，那這個代表三個變數對風險的影響不會隨時間改變，不用另外去處理。

3. Cox-snell residuals



可以看到點幾乎都落在紅色虛線上(45 度角)，所以模型是有符合 PH model 的假設。

四. 結論

最終的模型:

$$\hat{h}(t|Z) = \hat{h_0}(t)e^{0.0375age+1.3934Genome+(-0.0273)Mutation}$$

1. 此模型有滿足 Cox PH model 的模型假設
2. Age 增加，風險也是增加的，大部分的疾病都會符合這個
3. FGA 越大，風險是增加的，而且這個係數是很大的，代表其對風險影響很大，後續可以針對這個變數去了解為何會影響到膀胱癌的存活。
4. 腫瘤基因突變的數量越多，風險是降低的。那這個給我們一個新資訊，可以透過其他的癌症來分析這部分是否是合理的。

參考文獻

Fraction of Genome Altered and Total Mutations added to cBioPortal Plots tab | The Hyve

拷貝數_百度百科 (baidu.hk)

腫瘤突變負荷量（TMB）低卻對免疫療法反應佳！ | GeneOnline News

【膀胱癌治療】症狀，成因及檢查 | 希愈腫瘤中心 (heal-oncology.com)

上泌尿道尿路上皮癌 (UTUC) – 患者指南 - Bladder Cancer Canada