**存活分析學期報告計畫書**

**肺腺癌患者資料分析**

使用資料:lung for R

組員名單:林桓鈺、游皓淼、王熙皓、程　晨

**肺腺癌**

1. 背景

肺腺癌是肺癌的一種，屬於非小細胞肺癌。不同於鱗狀細胞肺癌，肺腺癌較容易發生於年輕女性、有抽菸史、亞洲族群。在肺部的位置常較週邊，腫瘤擴大的速度較慢（倍增時間約120天）。早期無徵兆，通常診斷出來時已經是晚期。台灣人肺癌在流行病學上部分特徵與西方國家不同，包括：女性肺癌的發生率特別高，西方國家之男性肺癌的死亡率是女性的 7~8倍，台灣則只有 1.8 倍；歐美國家的肺癌患者 85% 是吸菸者，在台灣則不及 50%；無論男女肺癌之細胞型態均以腺癌為主，根據 2018年台灣癌症登記資料統計，女性肺癌有高達 88% 是腺癌，男性肺癌也有 58% 是腺癌；肺腺癌的臨床表現以周邊型居多，早期多無症狀，容易延遲診斷。

1. 研究目的

分析存活時間,年紀,性別,ecog體能狀況karnofsky體能狀況,卡路里攝取量,體重下降等資料與肺腺癌之間的關聯

1. **分析目錄:**

1.觀察資料(變數各項值

2.觀察資料(變數各項值 )

3.觀察資料(Kaplan-Meier圖 )

4.PHA分析

5. cox regression

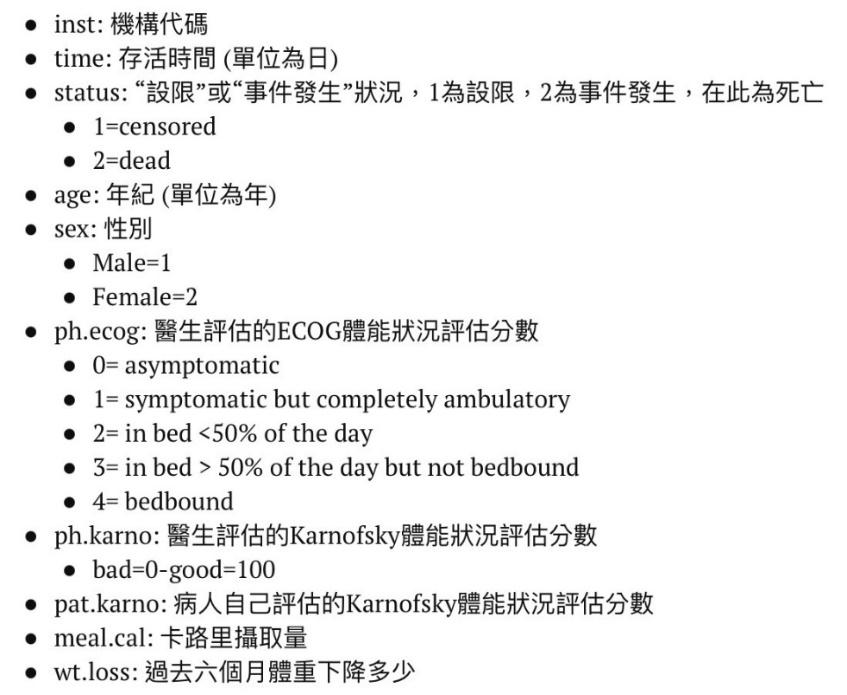
6. Parametric survival models

7.總結

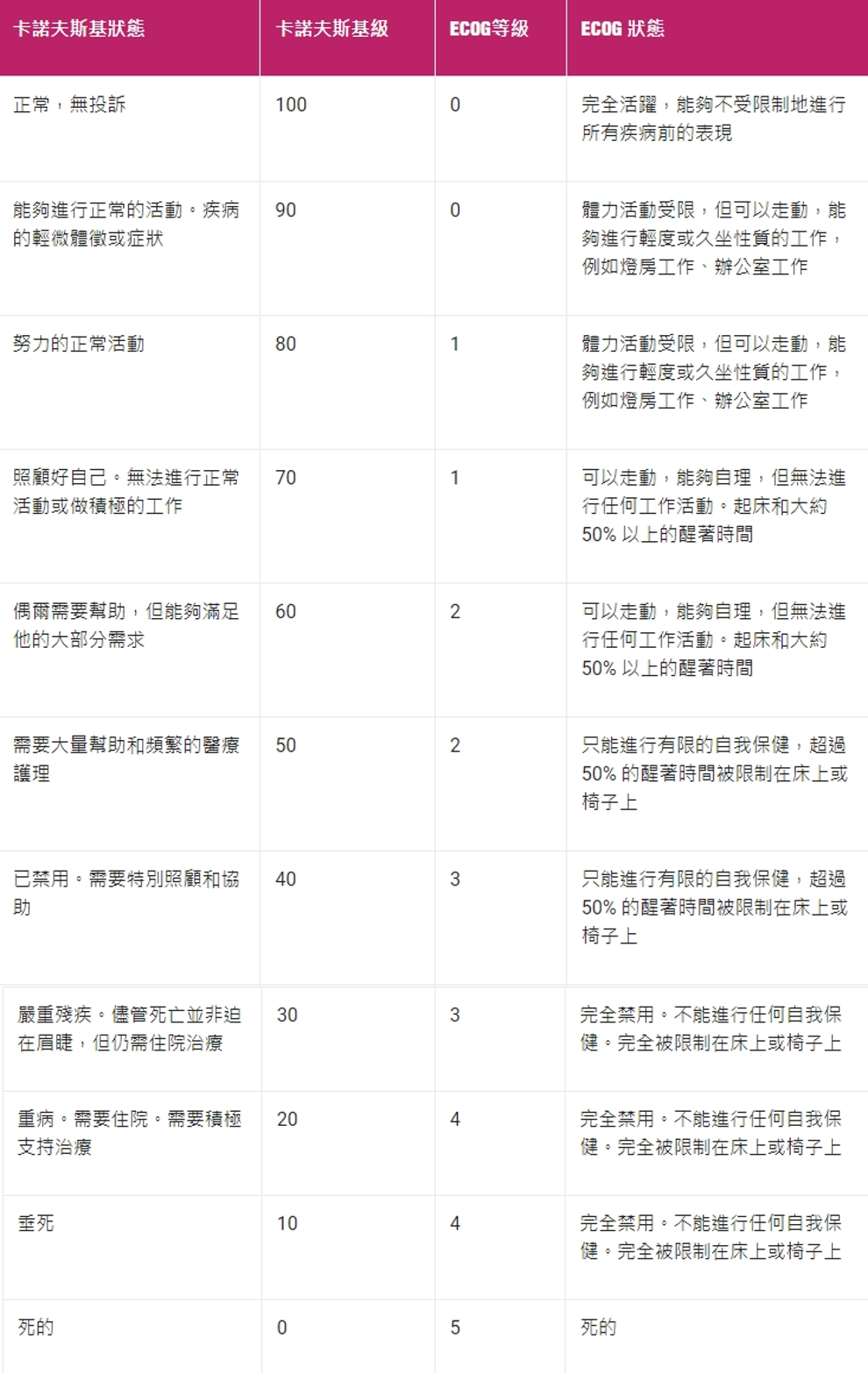
8.工作分配與程式碼

## 1.觀察資料(初步觀察 )

1. 變數:

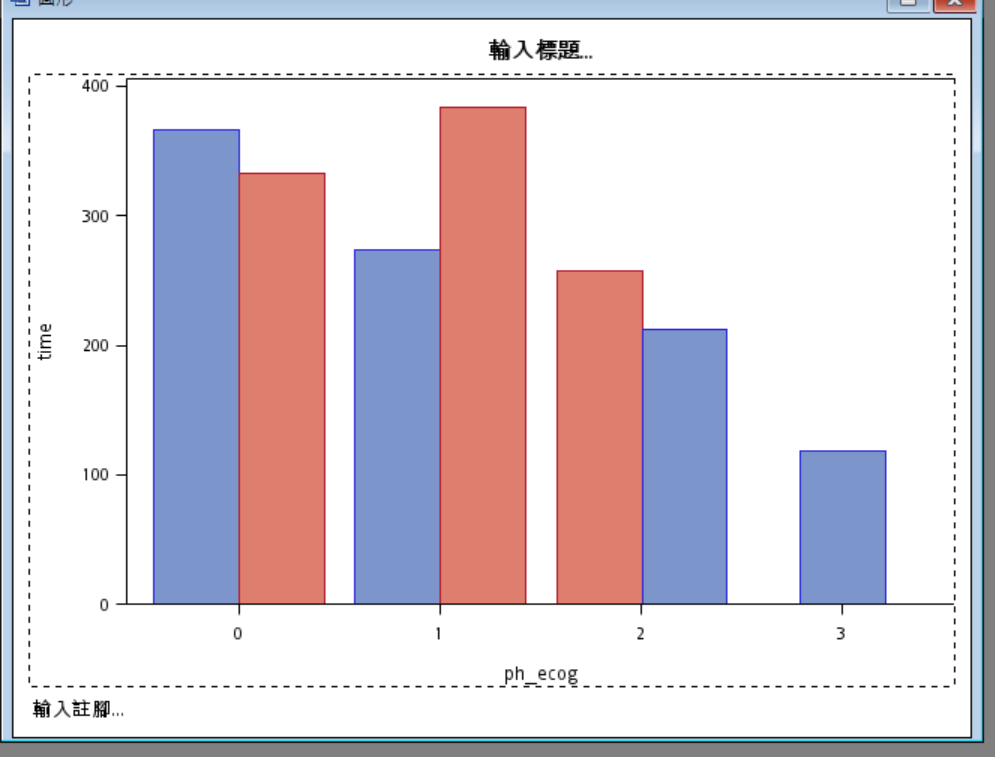


1. Karnofsky ECOG等級



## 2.觀察資料(變數各項值 )



****

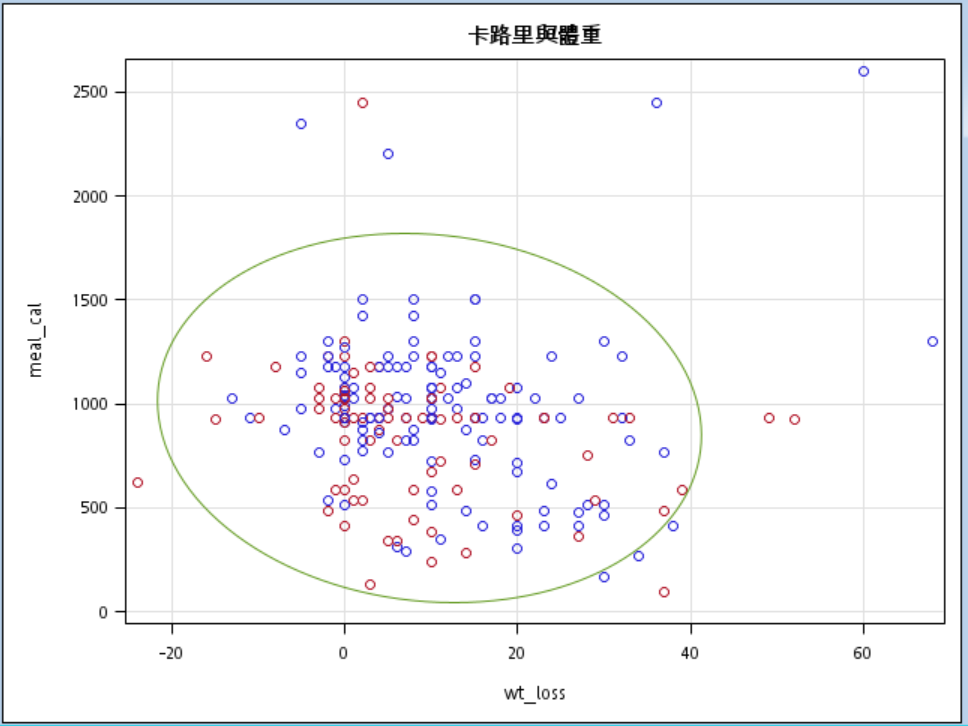
**下方為男女分組的描述:**

****

**從上圖可以看出每項都有粗略的各項變數有多少的缺失值和變數本身的最大值與最小值差異**和平均值。

**p.s對於缺失值處理我們已平均值填入**

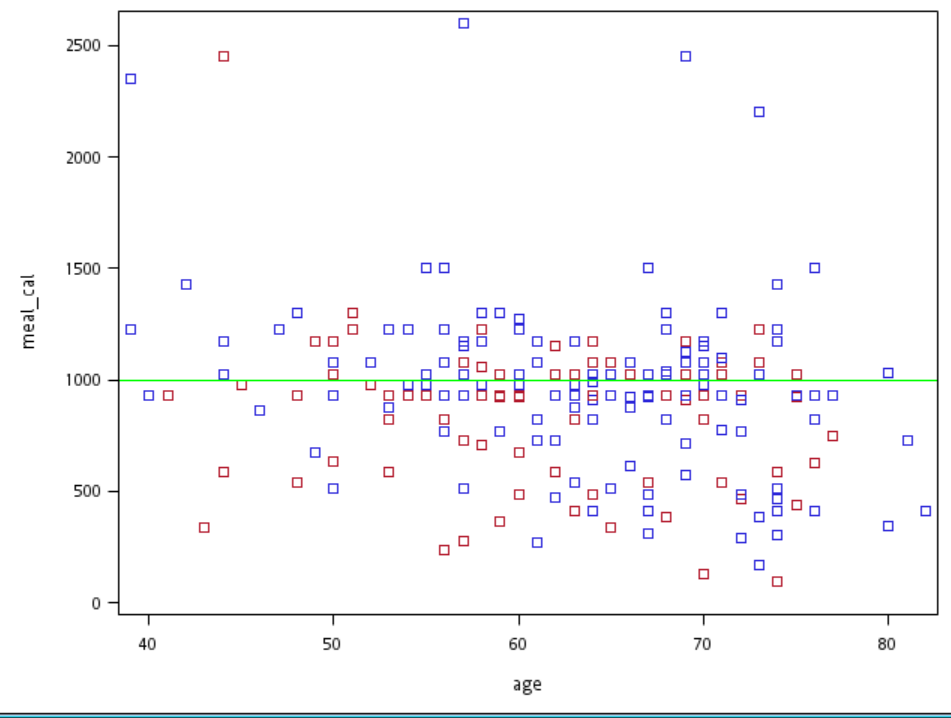
在這份資料中我們有發現一些資料應該是會有相關性例如: ECOG與Karnofsky可能會相關性較大，再者卡路里攝取量跟體重之間該也會有關係。



由上圖可知道卡路里攝取量跟體重相關性待商確。

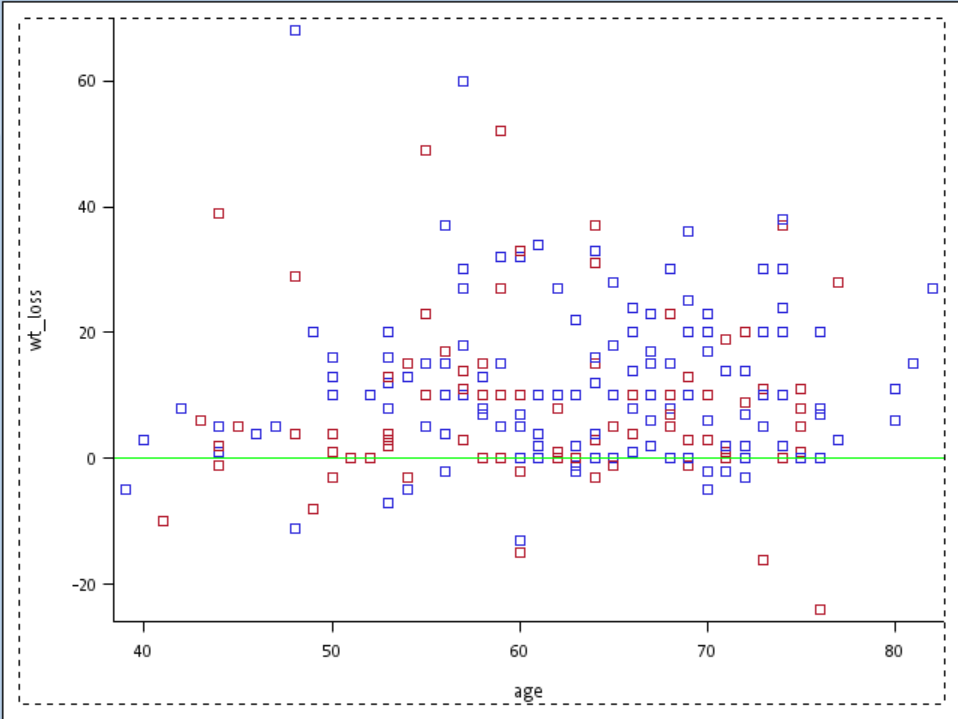
下圖為年齡與卡路里:

大部分的每日攝取量都在1000左右。



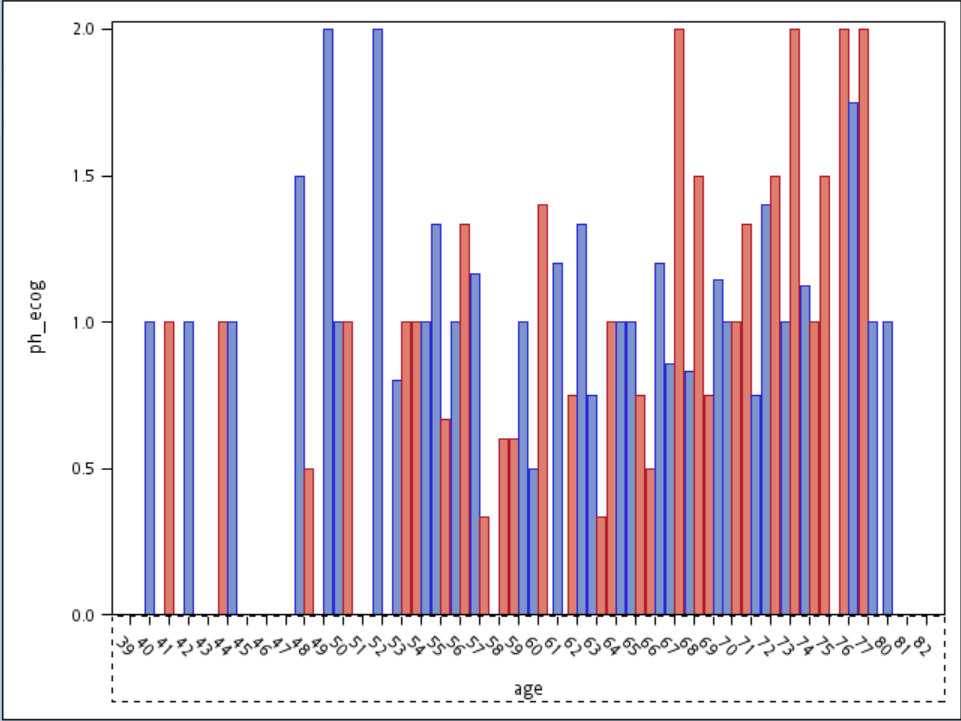
下圖為體重年齡:

大部分的人體重都增加。

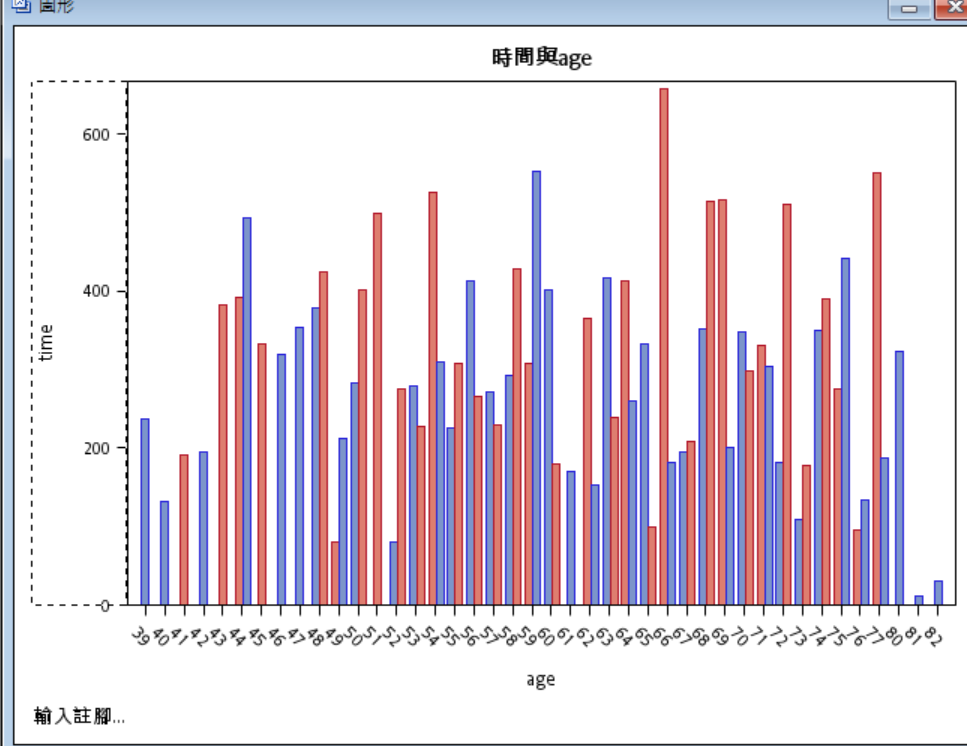


下圖年齡與ecog:

跟年齡關係比較小



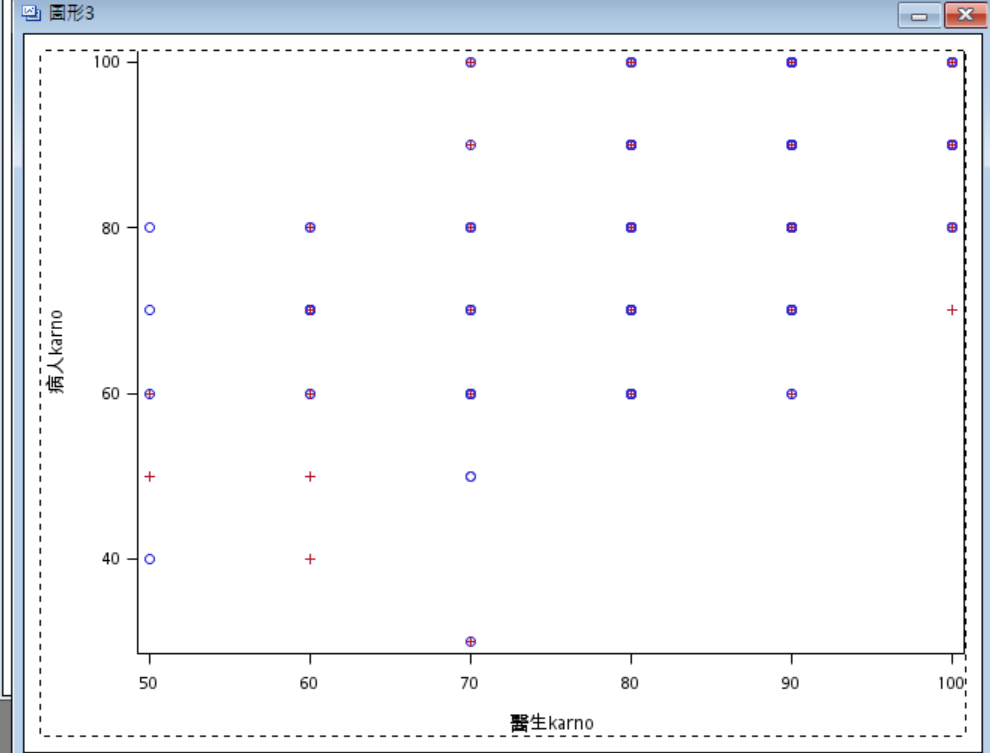
下圖年齡與存活:



倆者關係不大

病人與醫生評量:

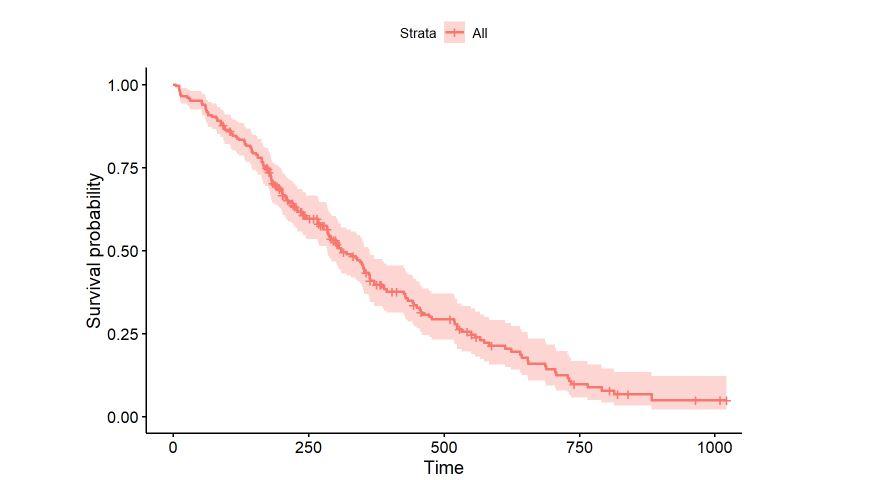
雙方都有相關性。

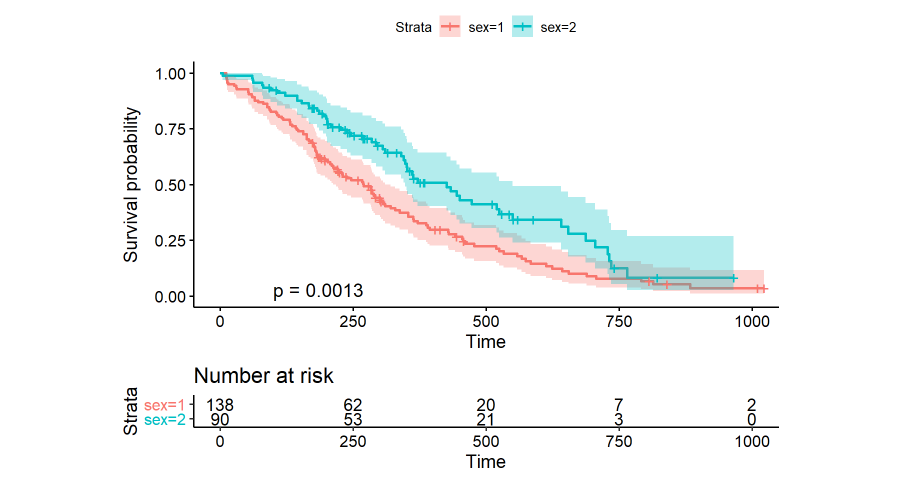


## 3.觀察資料(Kaplan-Meier圖 )

以最簡單的單一族群存活區限為例，畫個資料集中**所有病人**的存活曲線，(其實就是不分組的意思)，將**分組依據**設定成單一數字**1** 後再用性別做分組，也就是男生畫一條存活曲線，女生也畫一條，並放在同一張圖上，此時就將性別sex做為**分組依據(因為這是所有變數中差異最大的類別其他例如ECOG都較不明顯**

**)，並觀察曲線**





**估計xx年生存**

需要關注的一個數量是生存超過一定數量（xx）年的概率。

例如，要估算生存到11年的可能性

**估計中位生存時間**

關注的另一個數量是平均生存時間，我們使用中位數對其進行量化。預計生存時間不會呈正態分佈，因此平均值不是適當的總結。

**比較各組之間的生存時間 (3.PHA方法)**

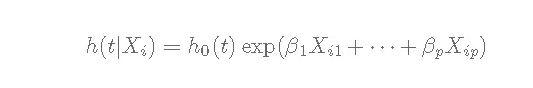
(預計會用status, age, ph.ecog, ph.karno, pat.karno, meal.cal, wt.loss七個變數與時間做比較)嘗試對資料做出多種模型解釋1.是否符合PHA 2.當成分層變數處理，假設變數可能跟時間有關，分析時間\*變數跟變數間的交互作用。

**Cox回歸模型(4. cox regression)**

我們可能想量化單個變量的效應大小，或者將多個變量包括在回歸模型中以說明多個變量的效應。

Cox回歸模型是半參數模型，可用於擬合具有生存結果的單變量和多變量回歸模型。

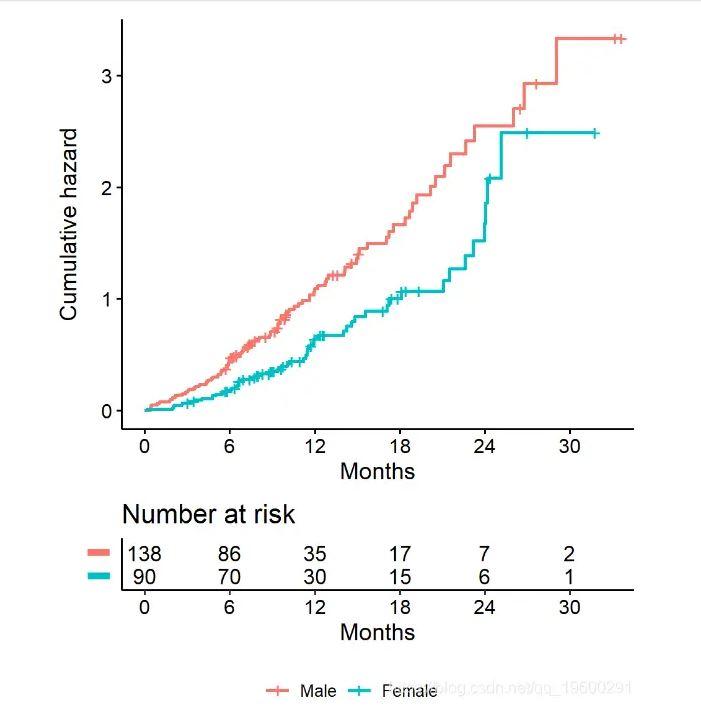
h(t)h(t)：危險或事件發生的瞬時速率h0(t)h0(t)：基本基準危險。



**危險比**

來自Cox回歸模型的關注數量是危險比（HR）。HR表示在任何特定時間點女性和男性之間的危險比率。

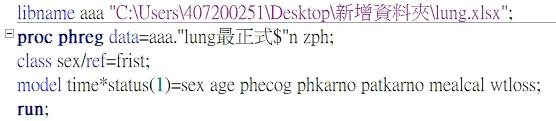
HR被解釋為感興趣事件中那些仍處於事件風險中的事件的瞬時發生率。

因此，我們的HR = 0.59意味著在任何給定時間，女性死亡的人數大約是男性的0.6倍。

## 4.PHA分析

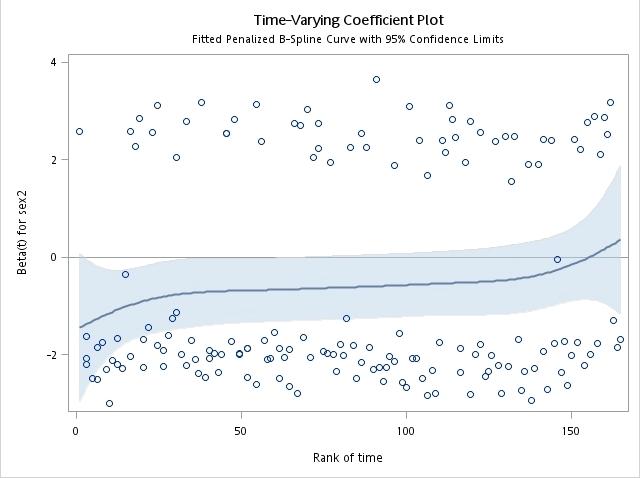
Model1

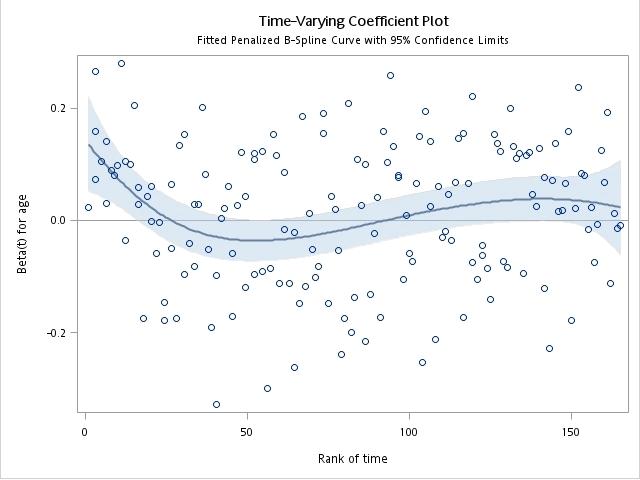
測試變數的PHA是否成立

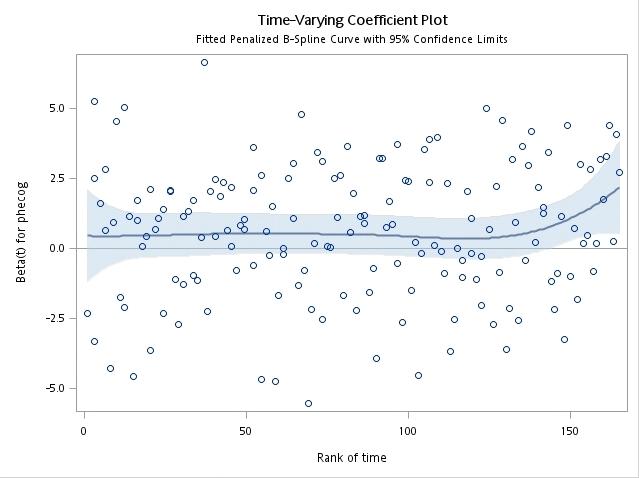


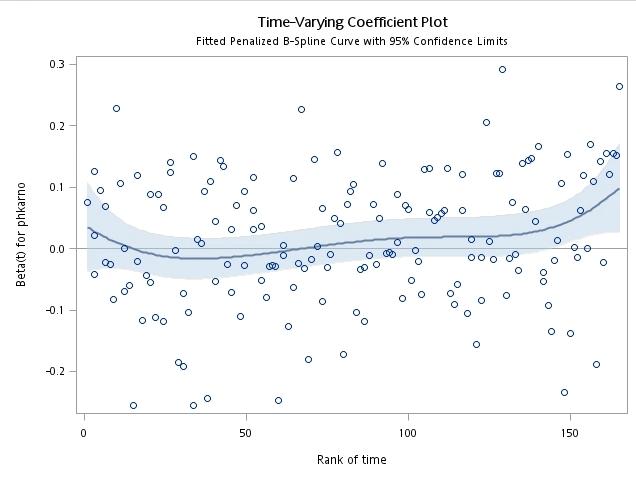


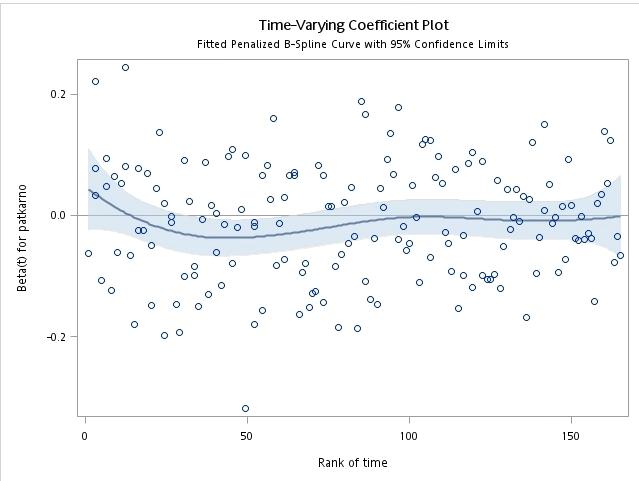
Age , phe.cog , pat.karno , wt.loss的P 值不顯著

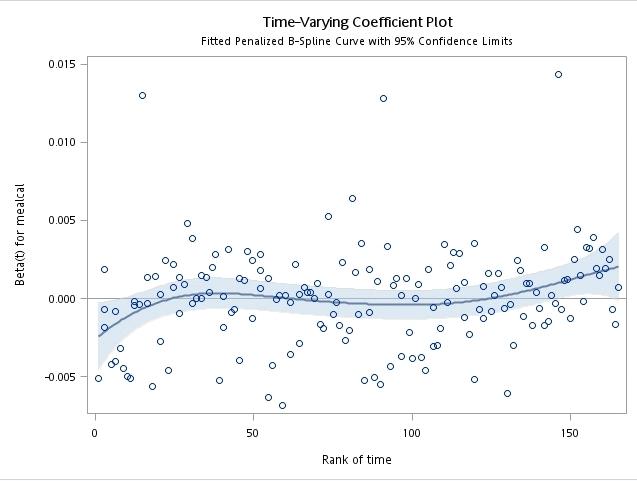


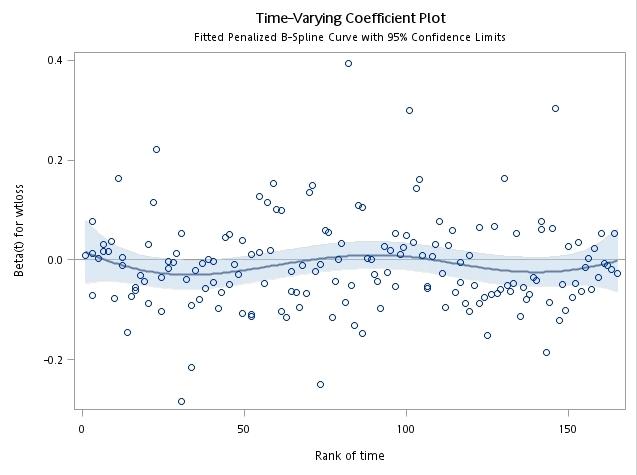






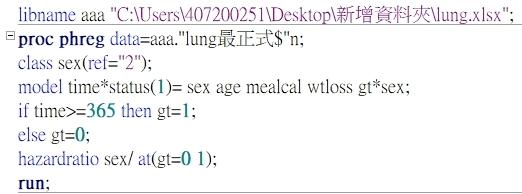






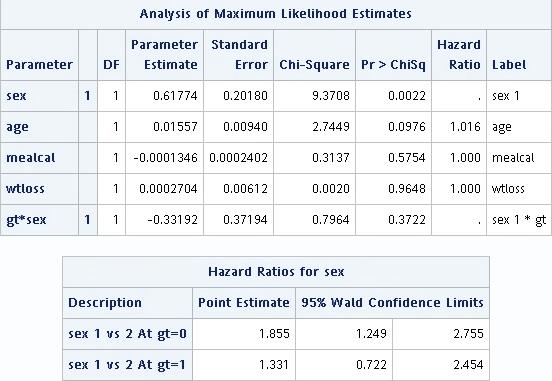
Model2

性別和存活時間有關



時間設為1 如果存活時間大於1年，如果小於1年 設成0

對sex做hazardratio 在gt=0跟1各做一次

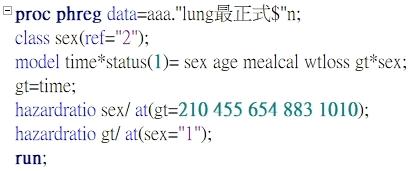


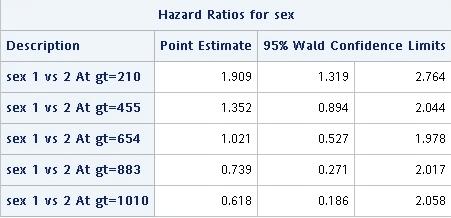
存活時間小於一年的情況下 男女性存在差異

女性的風險比是男性的1.855倍

Model3

任意時間的gt之下的hazardratio





Mode2 大於1年 男性 跟 女性 沒有差距

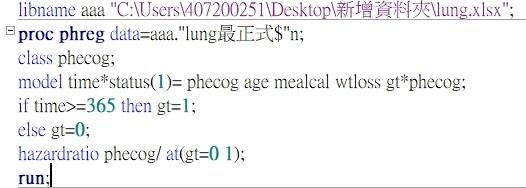
Mode3 在210天 男性 跟 女性 有差距(信賴區間沒有包含1)

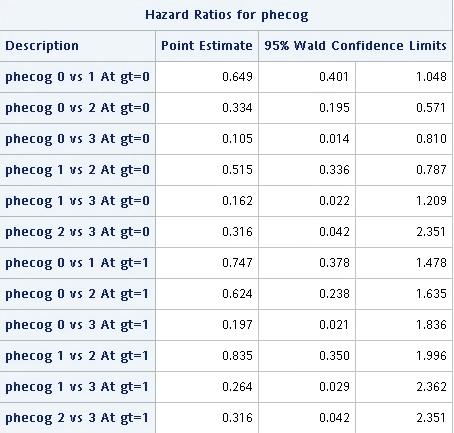
在883天 男性 跟 女性風險比剩0.739倍

在1010天 男性 跟 女性風險比0.618倍

隨著時間增加 男性 跟 女性差距縮小

Model4





Ph.ecog跟時間有關係

## 5. cox regression

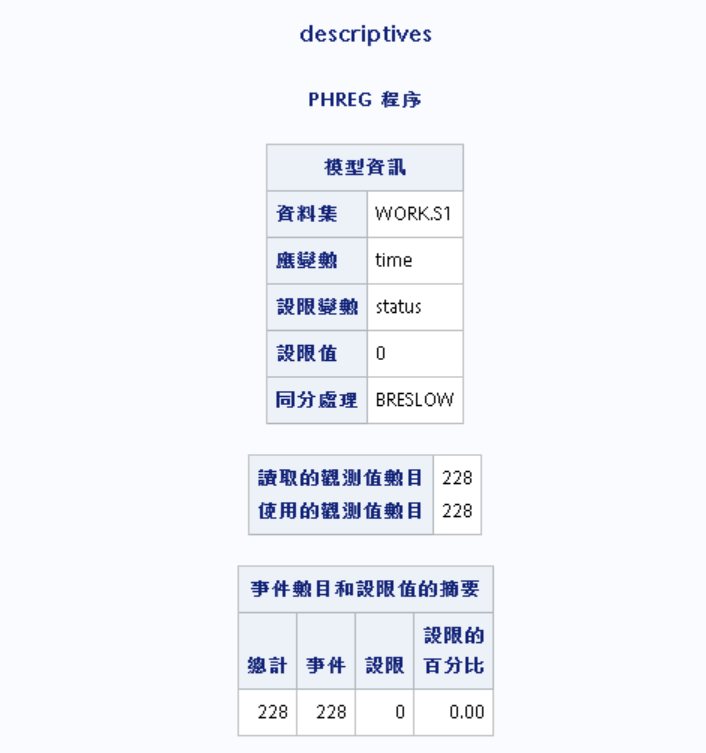
考慮到有些變數互有影響性故只挑其中幾個。

1.Ecog與時間的單變項的Cox model



跑完之後就可以看到ecog的Hazard ratio 和 95% HR CI 以及 p-value  
p-value=0.0021 <0.05 ，表示有顯著影響的風險因子

2.卡路里與時間的單變項的Cox model



p-value=0.5488 >0.05 ，表示沒有顯著影響的風險因子

3.體重與時間的單變項的Cox model



p-value=0.9795 >0.05 ，表示沒有顯著影響的風險因子

4.有class 類別資料(性別)



女生的死亡風險為男性的0,821倍

5.有class 類別資料(ecog)



Ecog數字越高危險率高。

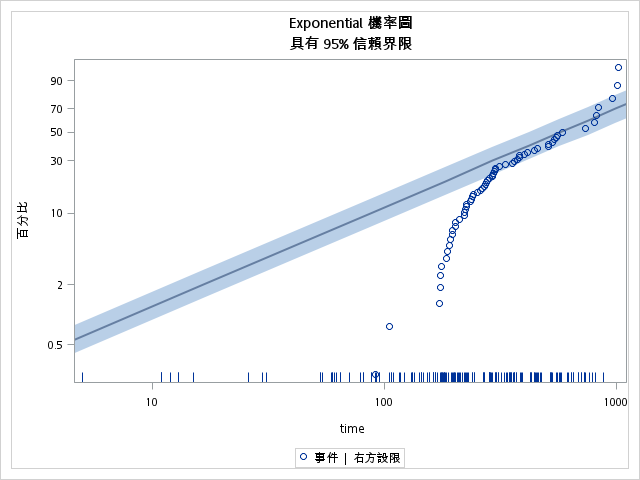
1. 小結:

我們發現對於男女之間存活率是不同的，而卡路里跟體重變化影響性不大，而ECOG指數高低對於存活率有相當的影響。

## 6. Parametric survival models

remission-exponential 

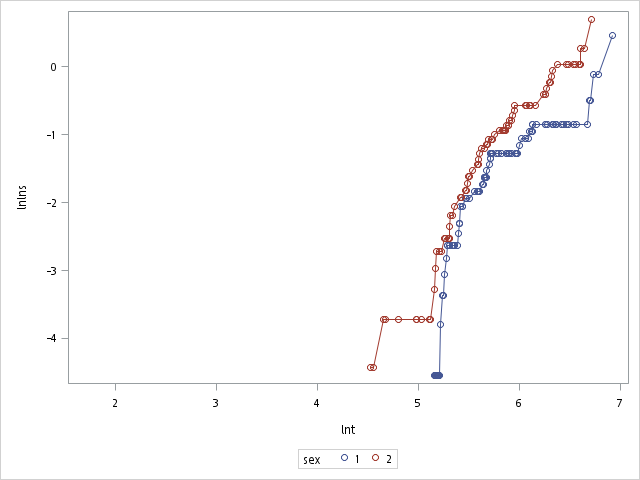
男生P值 0.0189 比截距值高



大部分的時間都在超過100

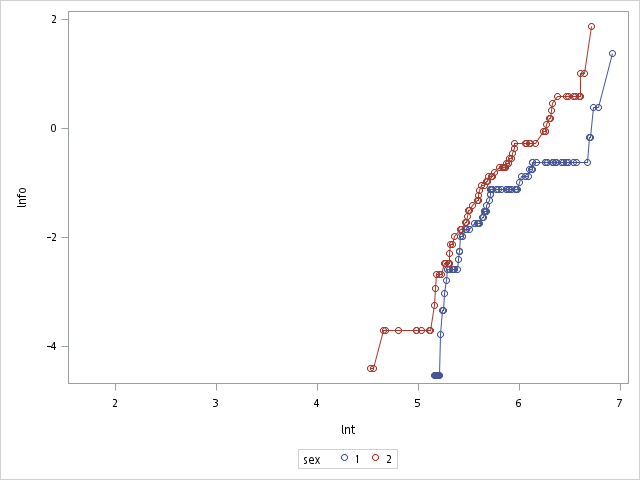
remission-Weibull

lnt= log(time)、lnlns=log(log(survive))



在同樣log(time)女性的值log(log(survive))都比男性高

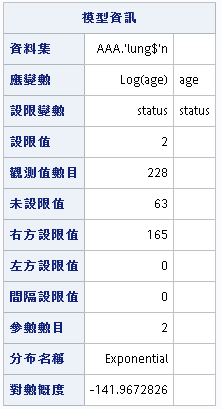
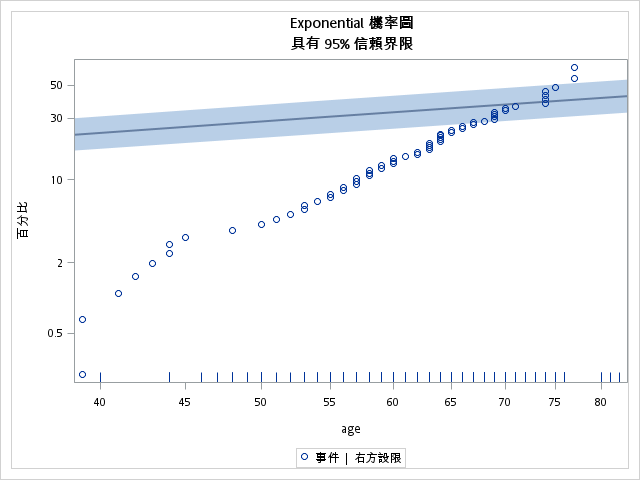
remission-Log-logistic



與上圖相同。

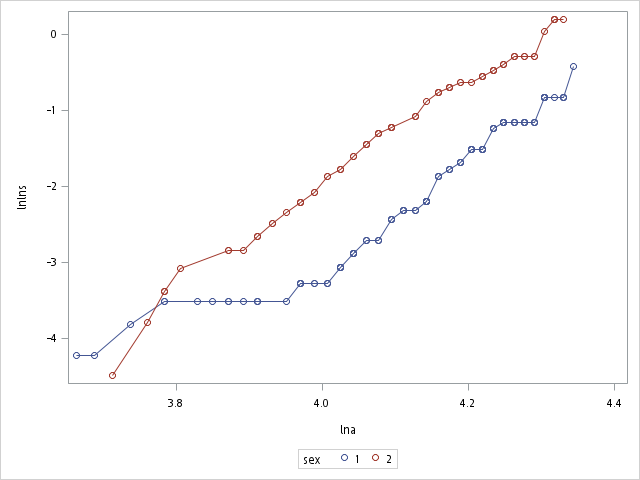
Age-status

remission-exponential

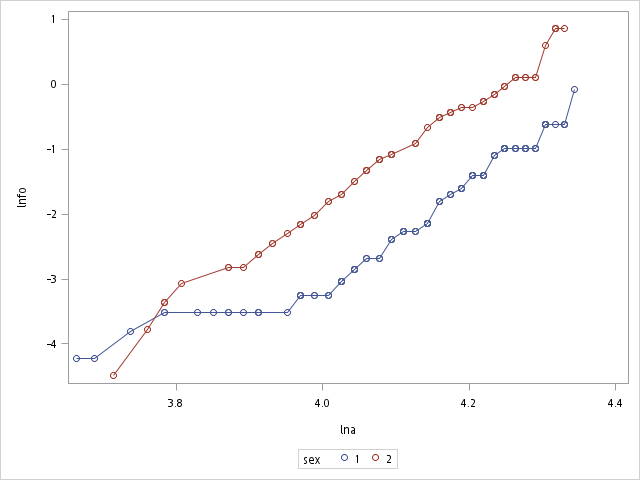
年紀越大信賴間就越高

remission-Weibull

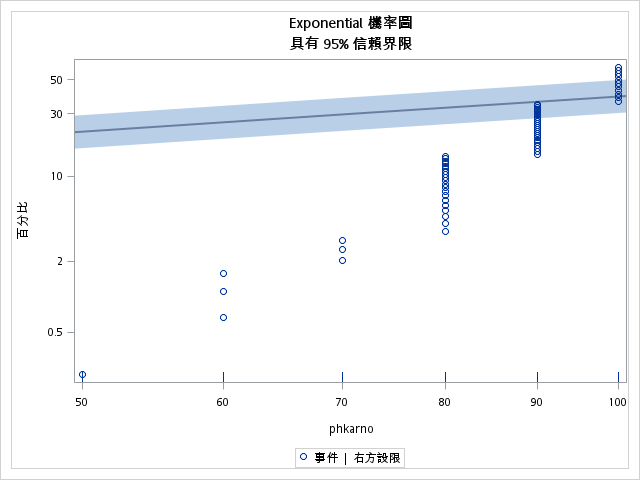


女性數值都比男性高，再40 後趨於平行

remission-Log-logistic



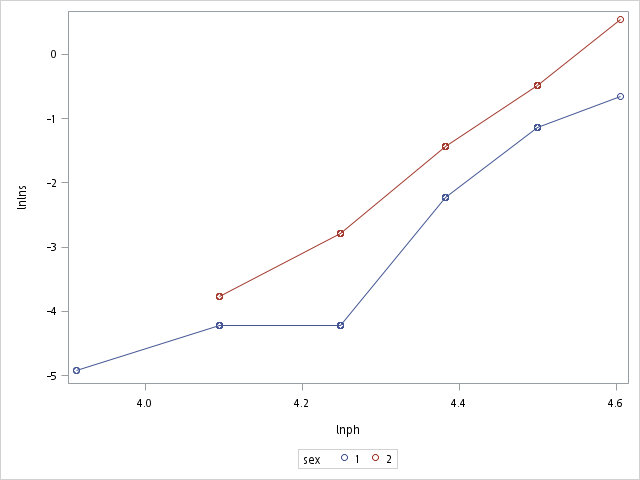
ph.karno-status

remission-exponential  

p值為0.0025<0.05

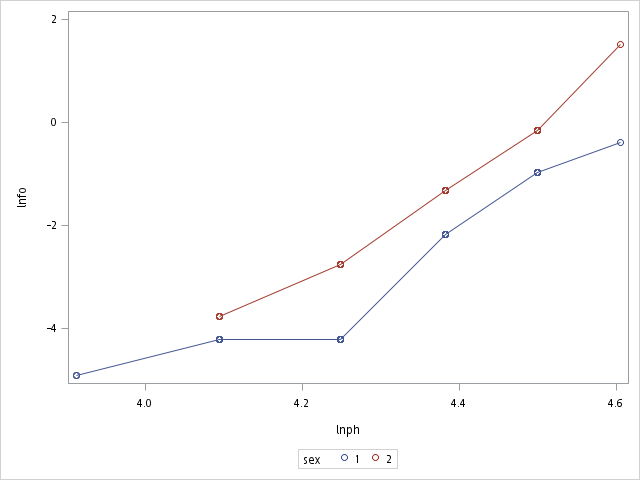


remission-Weibull



女性值都比男性高

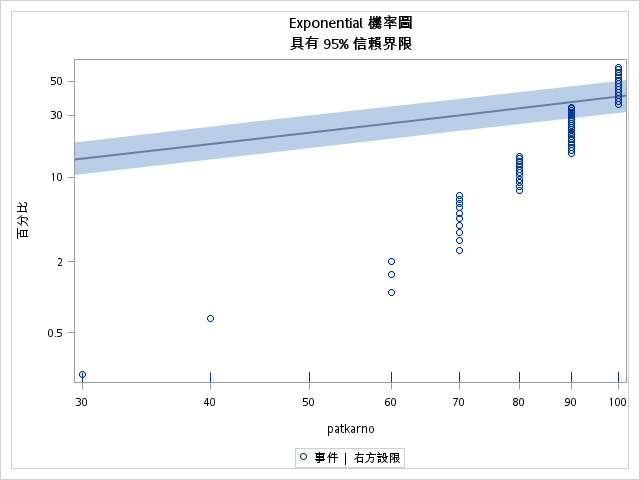
remission-Log-logistic



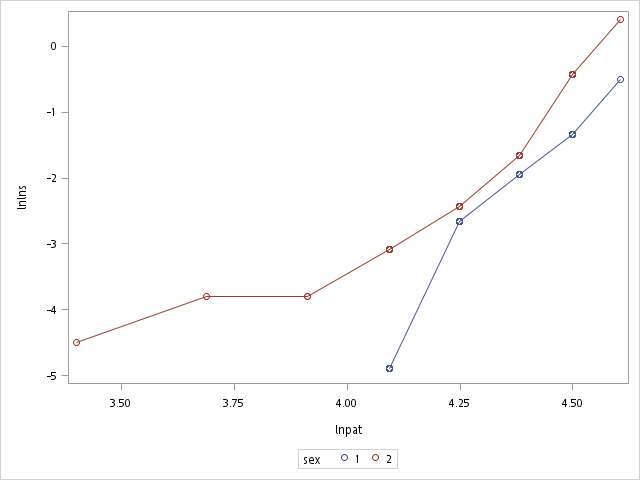
pat.karno-status

remission-exponential 

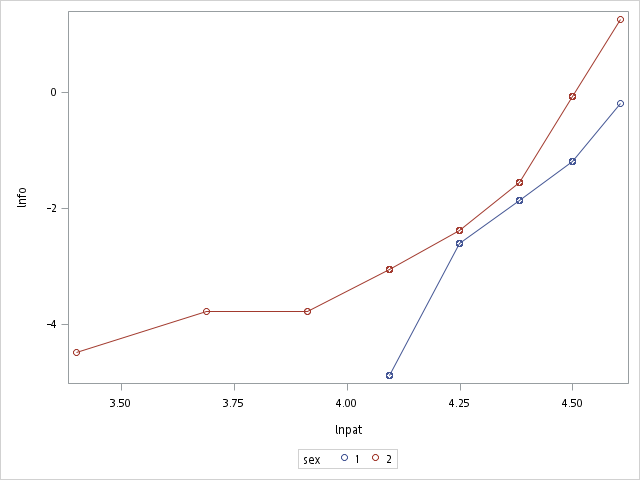
男性p值為0.0028



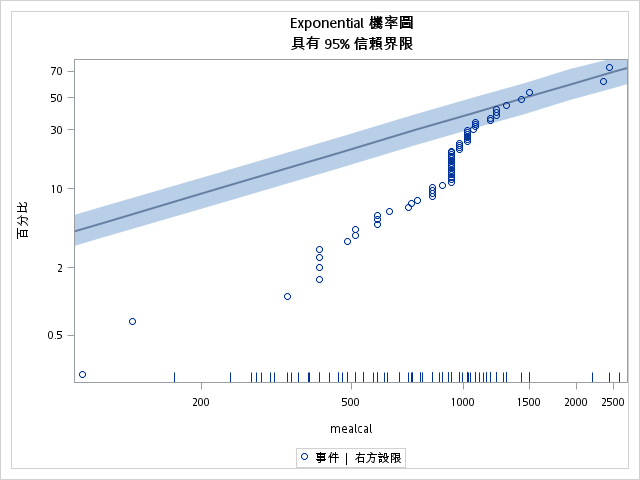
remission-Weibull



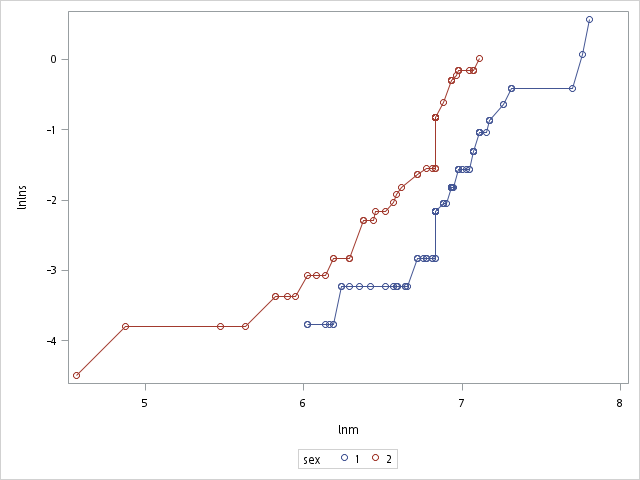
remission-Log-logistic



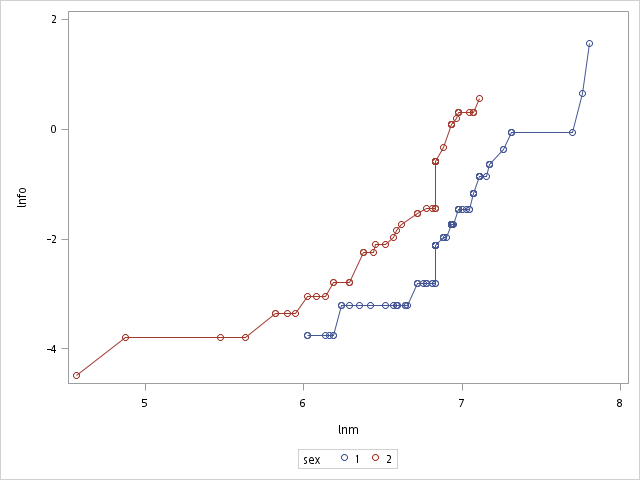
Meal.cal-status

remission-exponential  

remission-Weibull



remission-Log-logistic



## 總結

在這次報告中我們發現到ecog指標高低對於患者的存活率有一定的相關性，雖然從整體來看ecog好像相關性偏低，但是從分組來看，會發現指數越高，對發病率，死亡率都有影響，而在性別上我們發現女性比男性擁有更高的存活率，身體也更為健康大部分再ecog較低的數裡，而年齡跟時間關係並不是很大，至於卡路里跟體重變化則沒有絕對關係，連卡路里跟體重

互相之間也無絕對關係。

## 8. 工作分配與程式碼

工作分配:

報告、ppt創作:林桓鈺、游皓淼

程式: 程　晨、王熙皓

程式碼:



6. Parametric survival models

