交大統計碩一 邱泓儒 309657009

倖存分析期末報告

目錄

[壹、 Introduction 2](#_Toc75015008)

[貳、 Exploratory data analysis 4](#_Toc75015009)

[I. Kaplan-Meier  Method 4](#_Toc75015010)

[II. Descriptive statistics of variables 5](#_Toc75015011)

[III. Log-rank test and Gehan’s test 7](#_Toc75015012)

[參、 Method and testing 13](#_Toc75015017)

[I. Survival function 13](#_Toc75015018)

[II. Hazard function 13](#_Toc75015019)

[III. Cox model 13](#_Toc75015020)

[IV. AFT model 14](#_Toc75015021)

[肆、 Data analysis 15](#_Toc75015022)

[I. Cox model 15](#_Toc75015023)

[II. AFT model 17](#_Toc75015024)

[伍、 Discussion 19](#_Toc75015025)

# Introduction

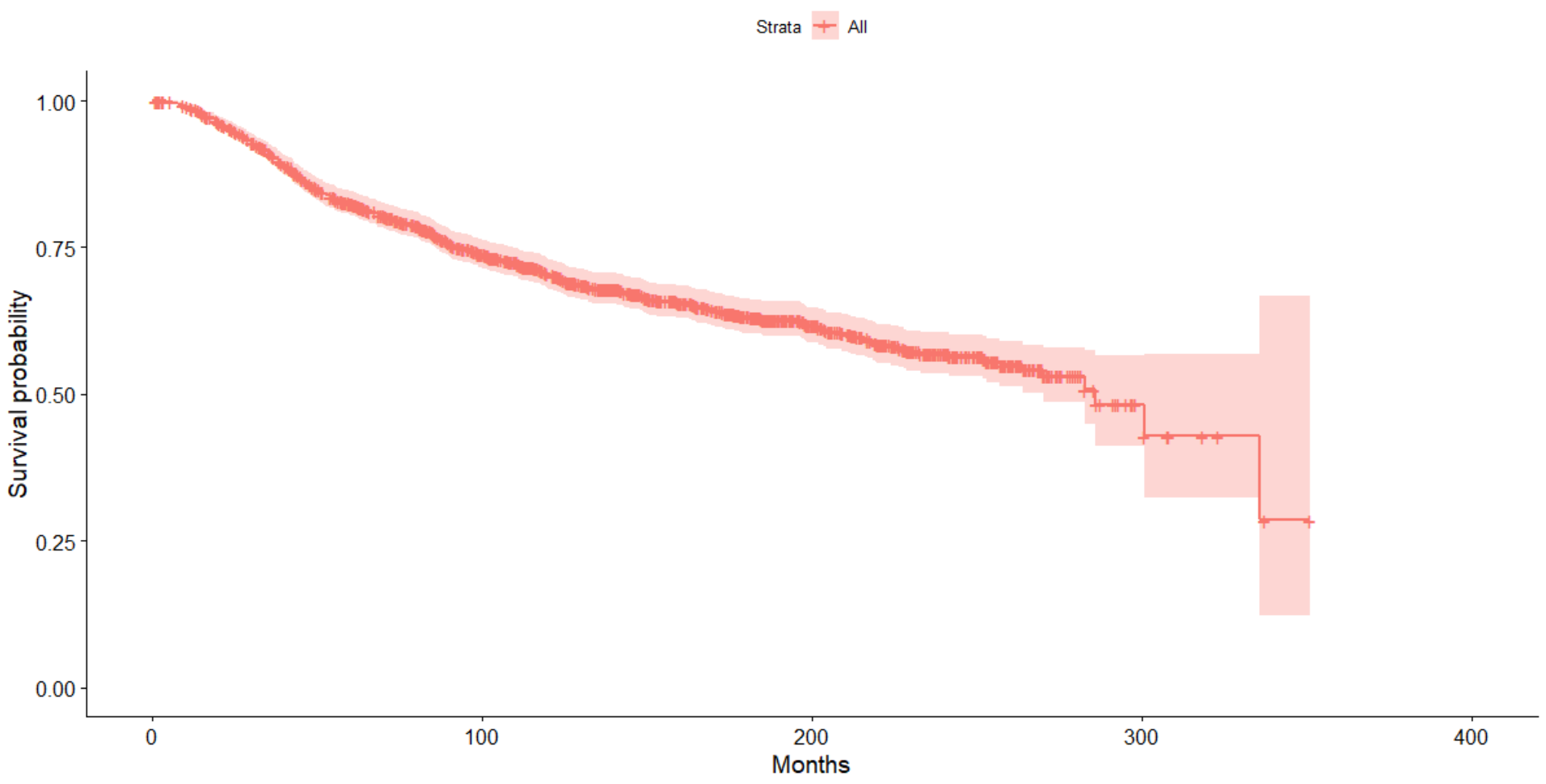
我所分析的資料為The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database，是一個來自加拿大及英國的研究計畫。此資料集收集了1,980個乳癌病人的病理及基因數據，我的報告將只針對病理數據分析，尤其著重於探討治療乳癌的三種主要方式-化學治療(chemotherapy)、放射治療(radio\_therapy)及賀爾蒙治療(hormone therapy)對於乳癌病人生存曲線的影響。下表為本報告會使用的變數及變數描述。

|  |  |
| --- | --- |
| 變數名稱 | 描述 |
| patient\_id | Patient ID |
| age\_at\_diagnosis | Age of the patient at diagnosis time |
| chemotherapy | Whether or not the patient had chemotherapy as a treatment (yes/no) |
| hormone\_therapy | Whether or not the patient had hormonal as a treatment (yes/no) |
| radio\_therapy | Whether or not the patient had radio as a treatment (yes/no) |
| overall\_survival\_months | Duration from the time of the intervention to death |
| overall\_survival | Target variable wether the patient is alive of dead. |
| tumor\_size | Tumor size measured by imaging techniques |
| tumor\_stage | Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread |
| death\_from\_cancer | Wether the patient's death was due to cancer or not(Living/Died of Disease/Died of Other Causes) |

# Exploratory data analysis

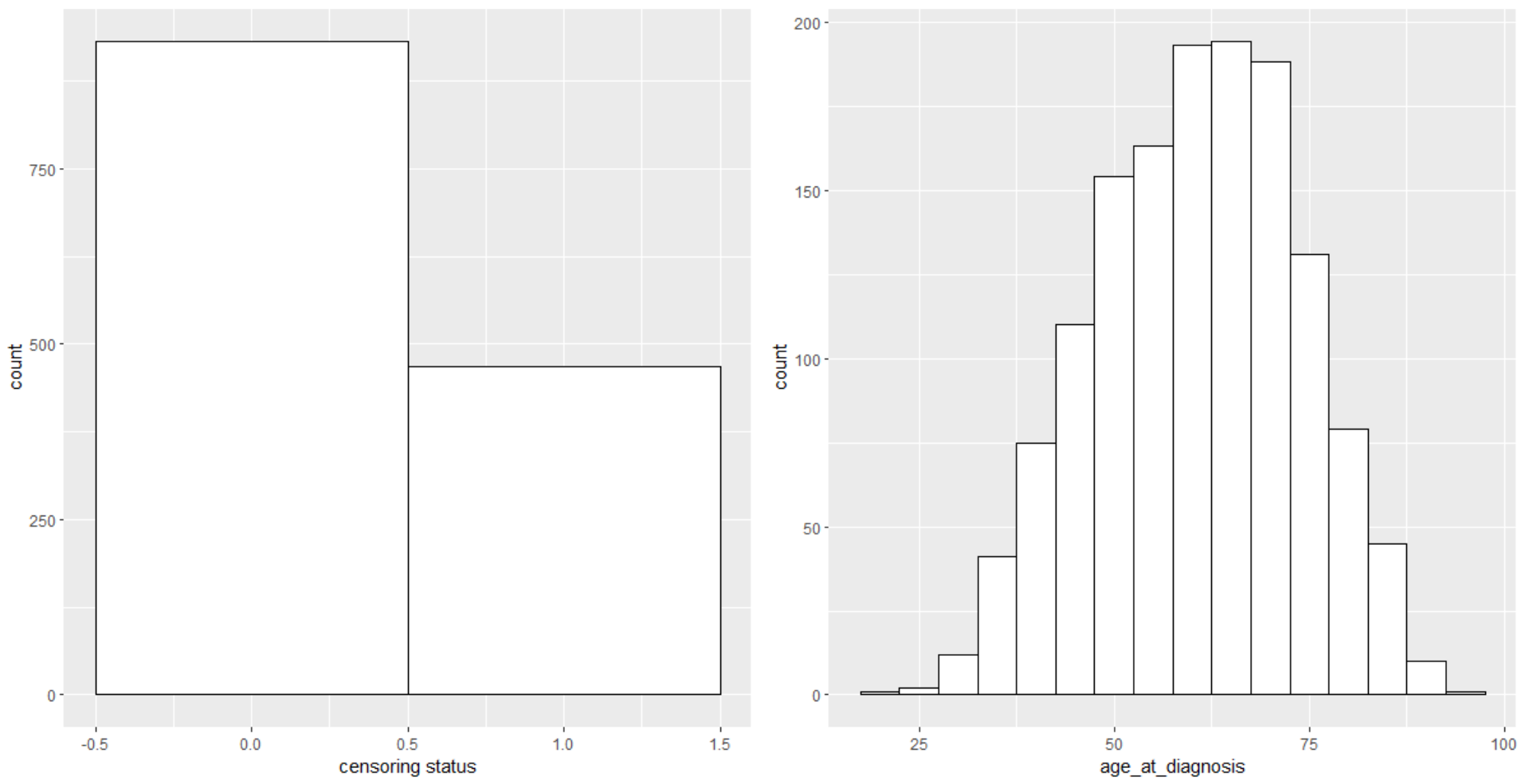
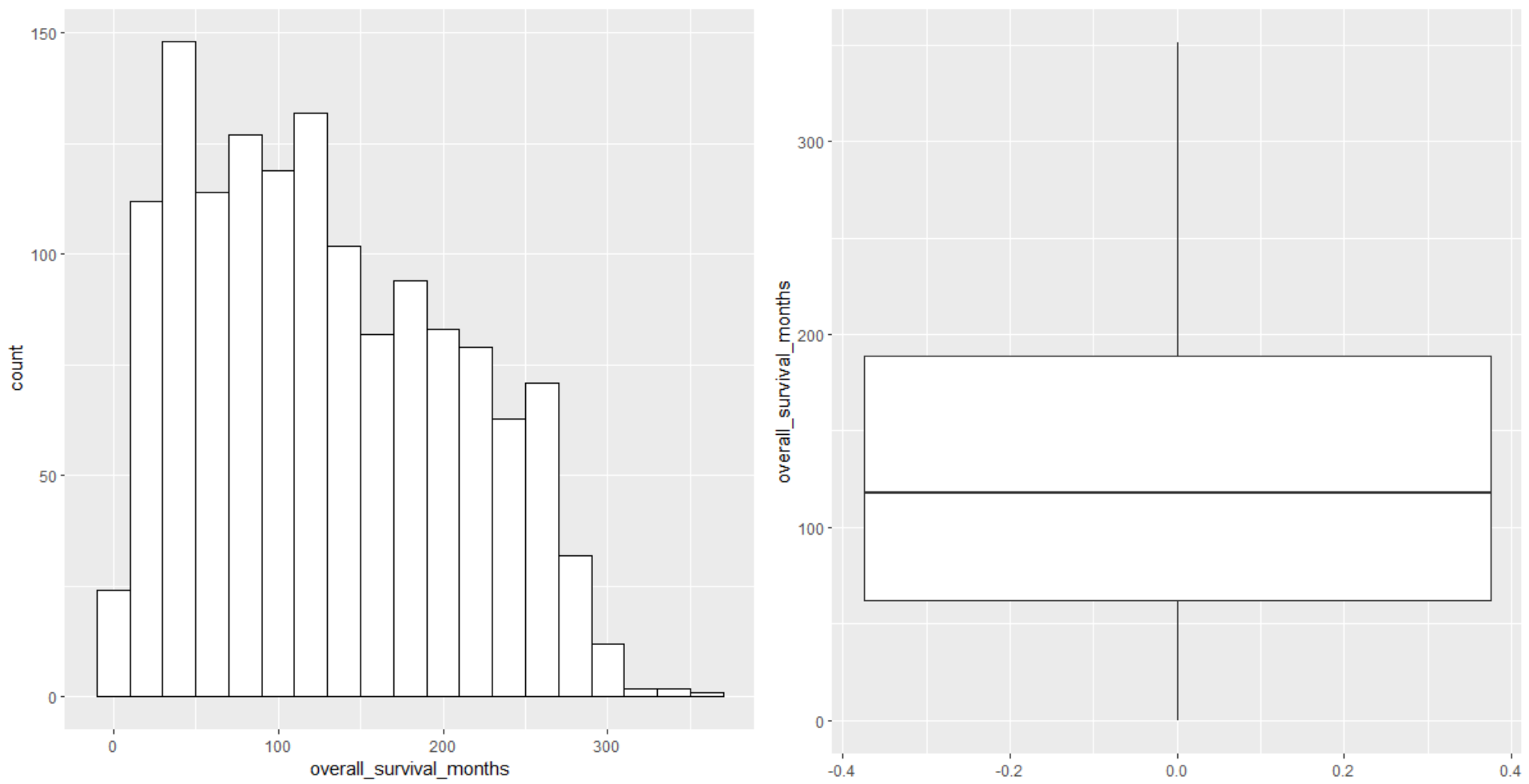
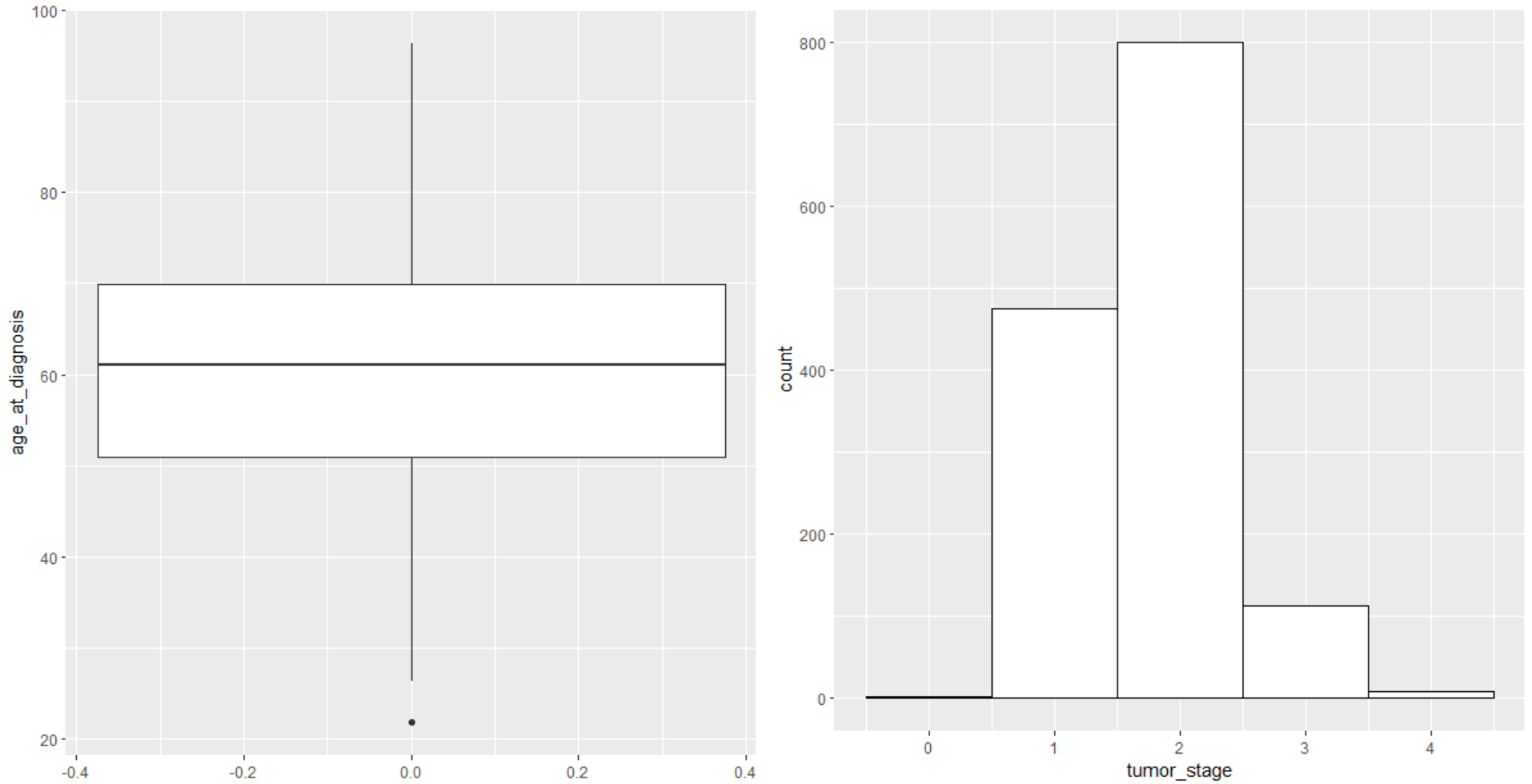
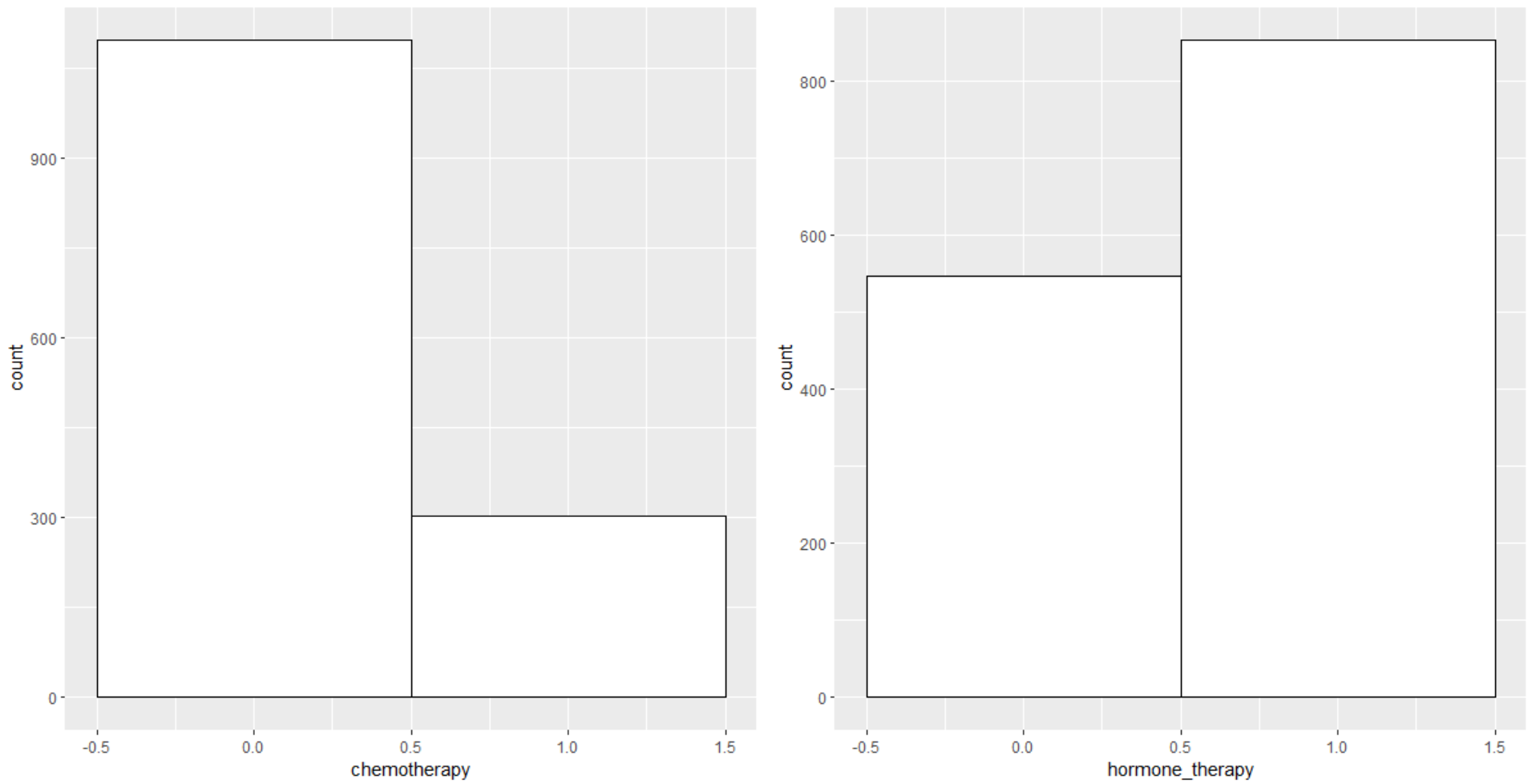
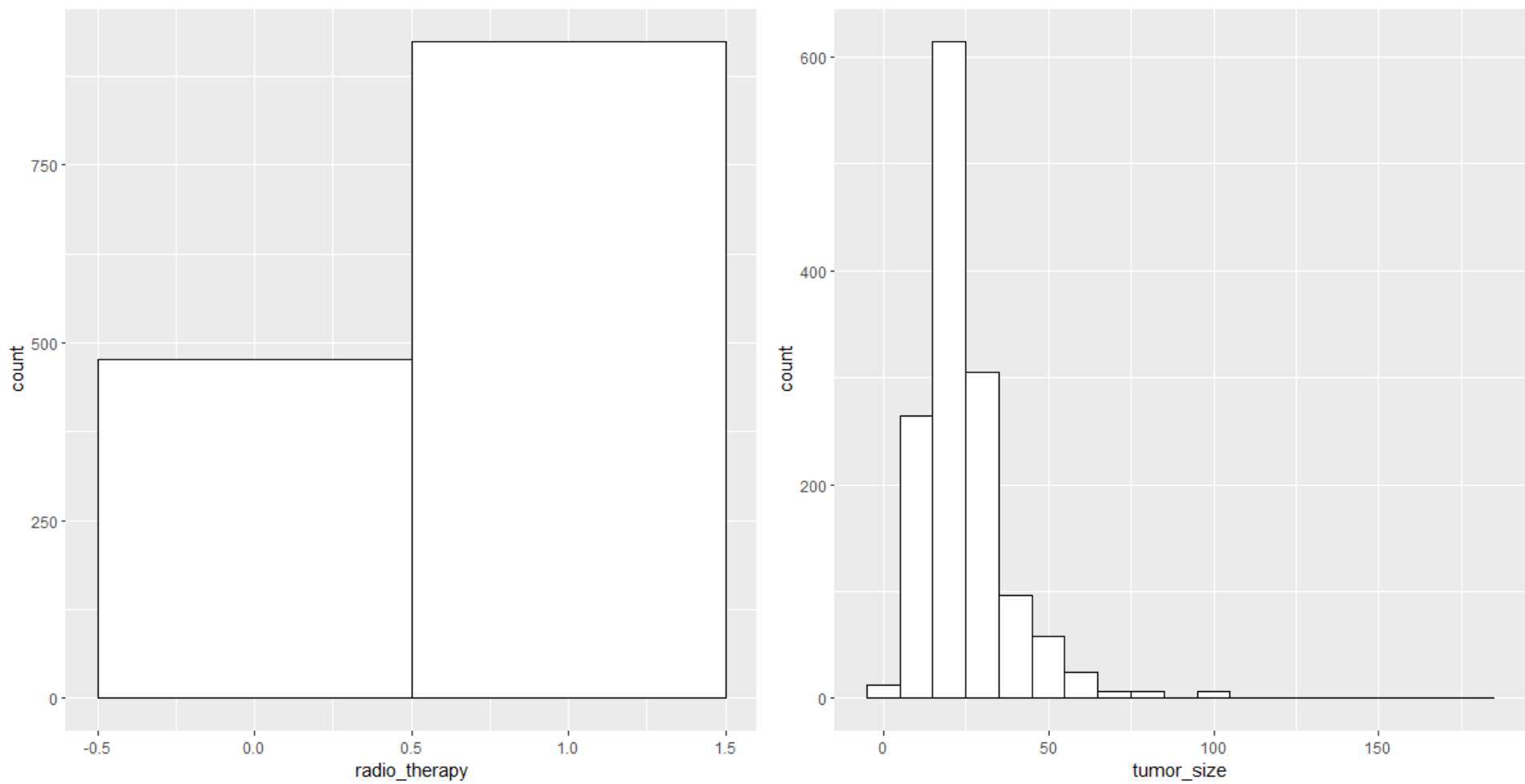
## Kaplan-Meier  Method

首先，定義生存時間為從開始治療到死於乳癌的時間段，即變數overall\_survival\_months。接著定義censoring status，將變數death\_from\_cancer中，Living或Died  of  Other  Causes的病人視為censored，Died  of  Disease的病人視為complete。以以上定義及Kaplan-Meier Method估計生存曲線並繪圖，如下:



可以發現約275月以後失敗點(病人因乳癌死亡)非常少，造成尾端的信賴區間非常的大。

## Descriptive statistics of variables

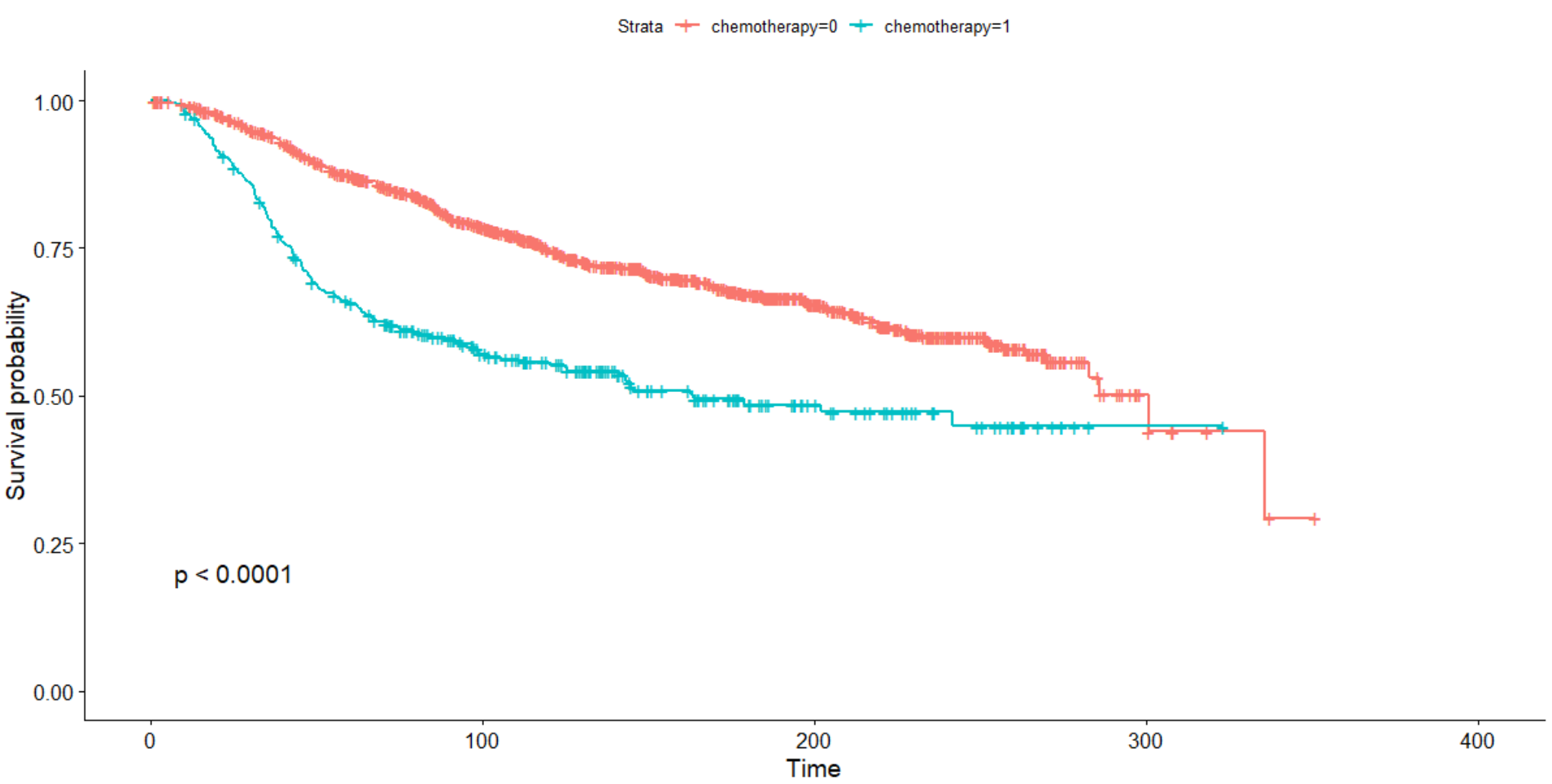
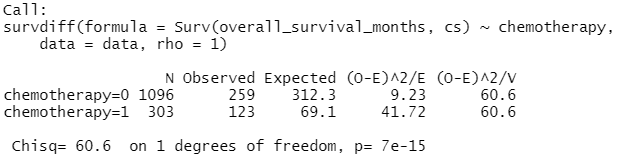
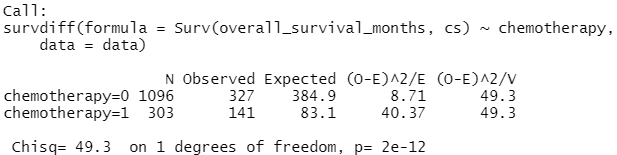
   

## Log-rank test and Gehan’s test

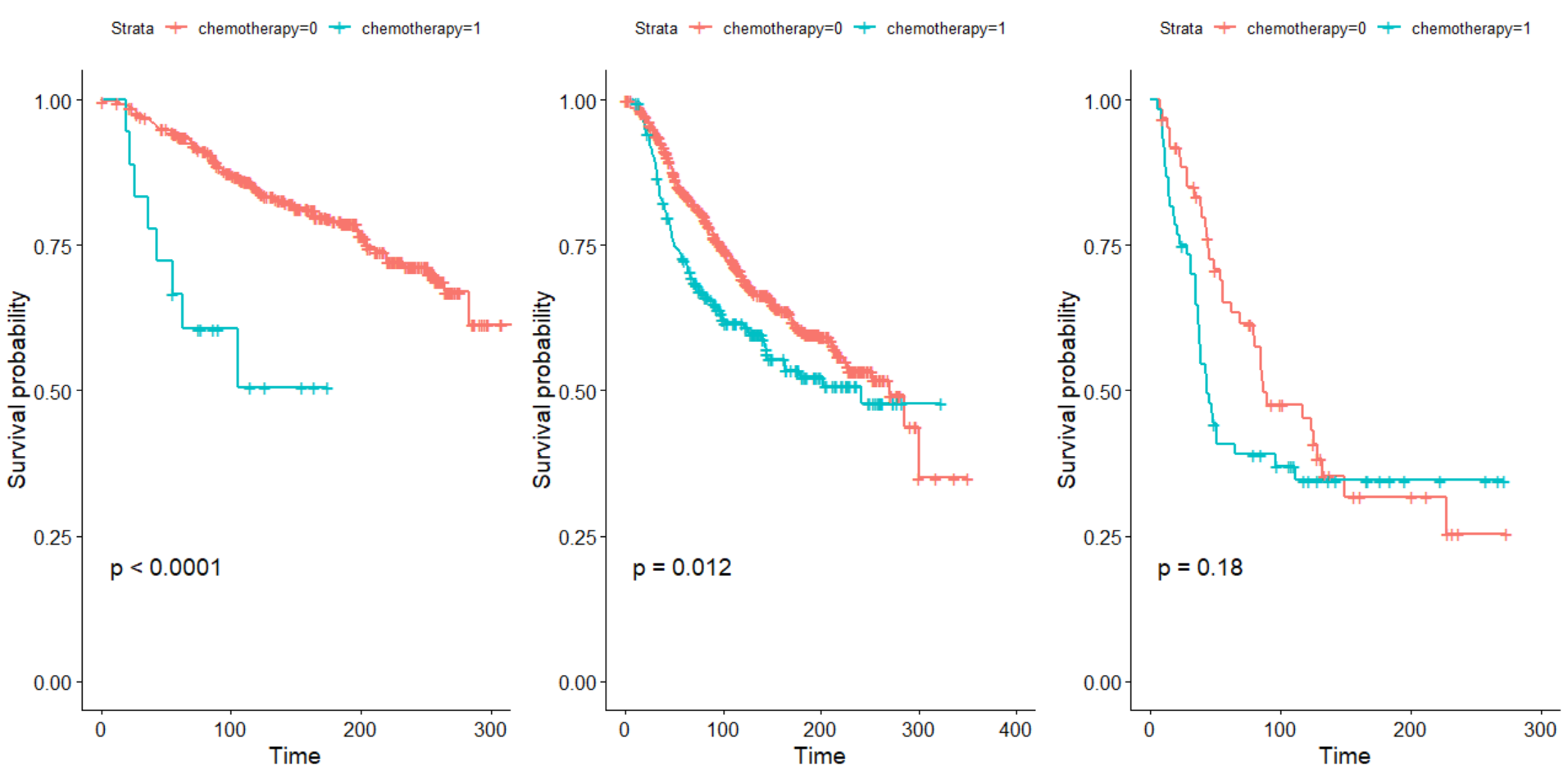
主要目的為想知道有無接受三種不同治療的病人的生存機率有無差異，進而推論三種治療法之效用。

分別對變數chemotherapy、hormone\_therapy、radio\_therapy、tumor\_stage做Log-rank test及Gehan’s test並繪圖

### chemotherapy

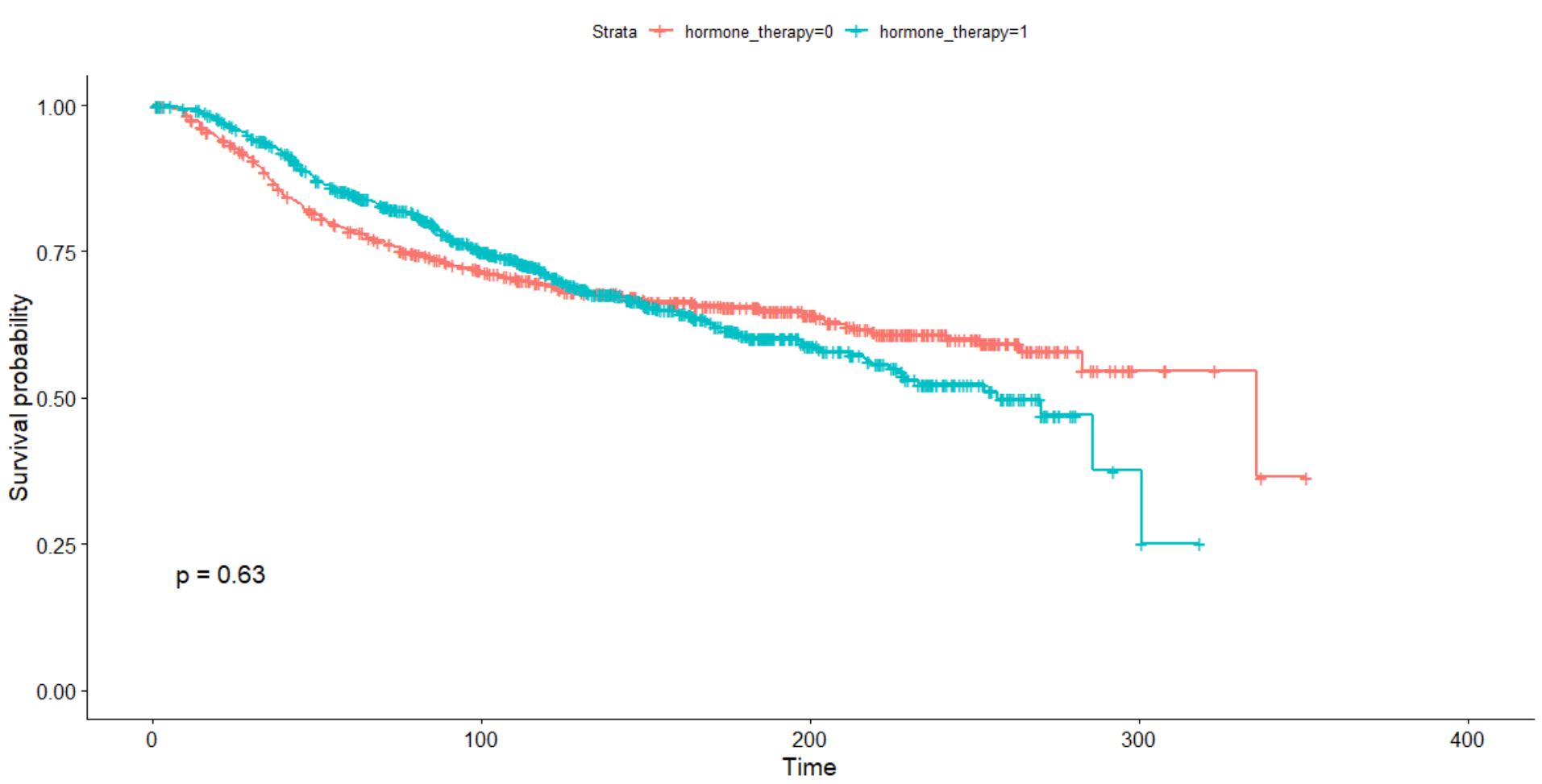
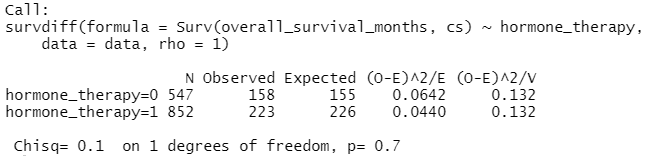
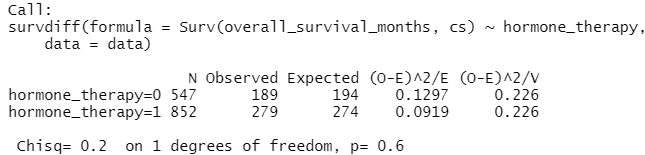


Log-rank test及Gehan’s test都說明有/無接受化學治療的患者生存曲線存在差異。值得注意的是圖中沒有接受化療的病人反而生存機率較高，我推測可能是因為會接受化療的病人本身病情就較嚴重，故此結果並不能說明化療對乳癌有反效果。為嘗試考慮病情的資訊，我以tumor stage作為病情嚴重性的指標，將病人分為三組(tumor stage=0 or 1,tumor stage=2,tumor stage=3 or 4)，再進行分析，結果如下:



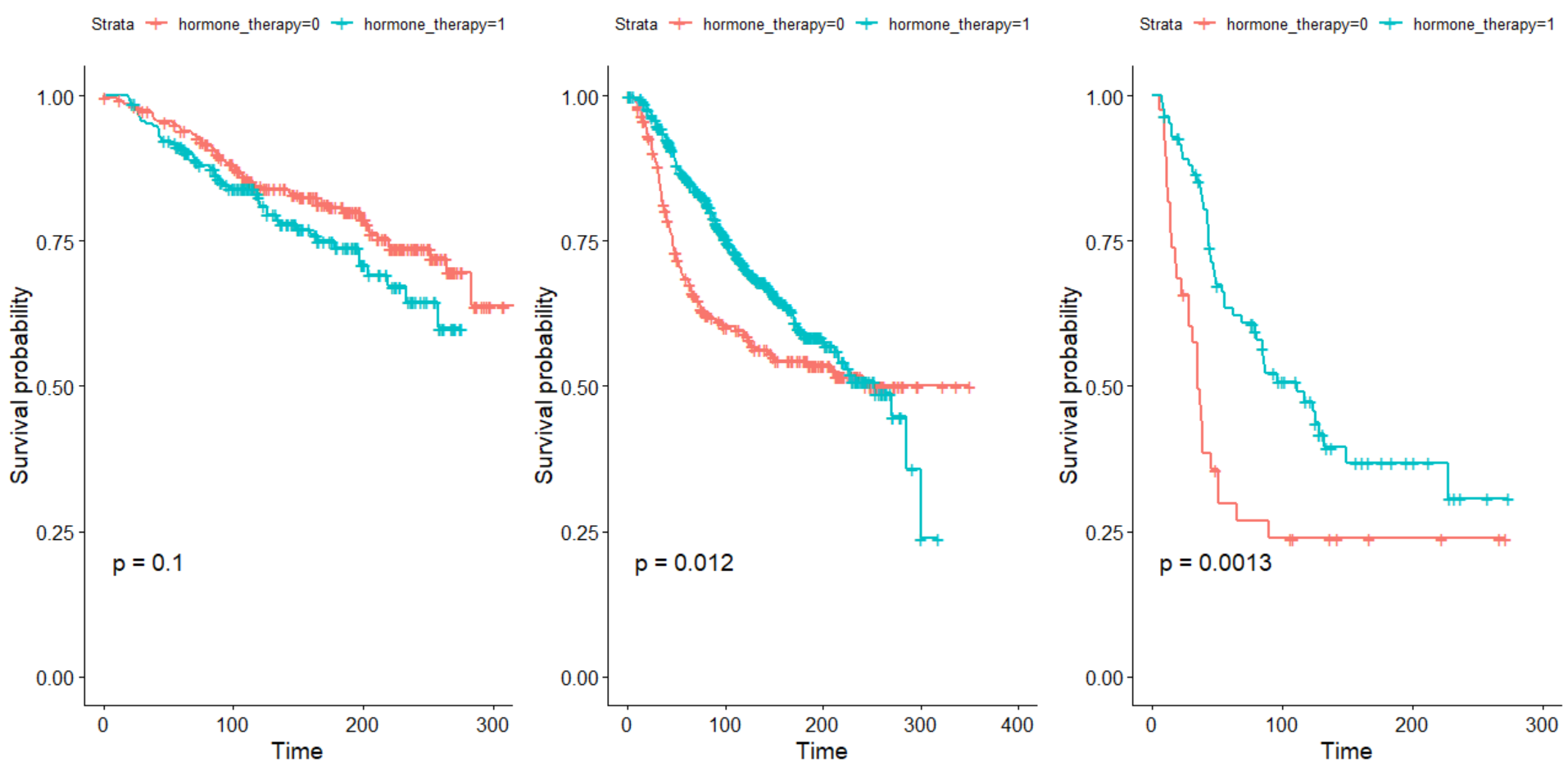
結果依然是有接受化療的病人生存機率較低，推測可能是tumor stage還是無法完全體現病人病情的嚴重性。

### hormone\_therapy



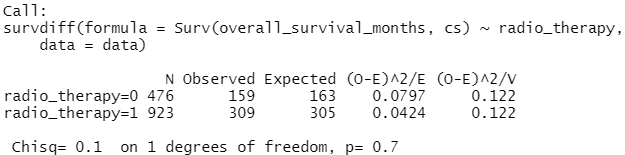
雖然Log-rank test及Gehan’s test有相同結果，即無證據說明有/無接受賀爾蒙治療的患者生存曲線有差異，但從圖中可以明顯發現，有/無接受賀爾蒙治療的患者生存曲線有相交的情況，檢定結果(do not reject，small chi-square statistics)可能是因為兩條生存曲線在交叉點前後差異相互抵消了，因此此變數不適用於Log-rank test及Gehan’s test。

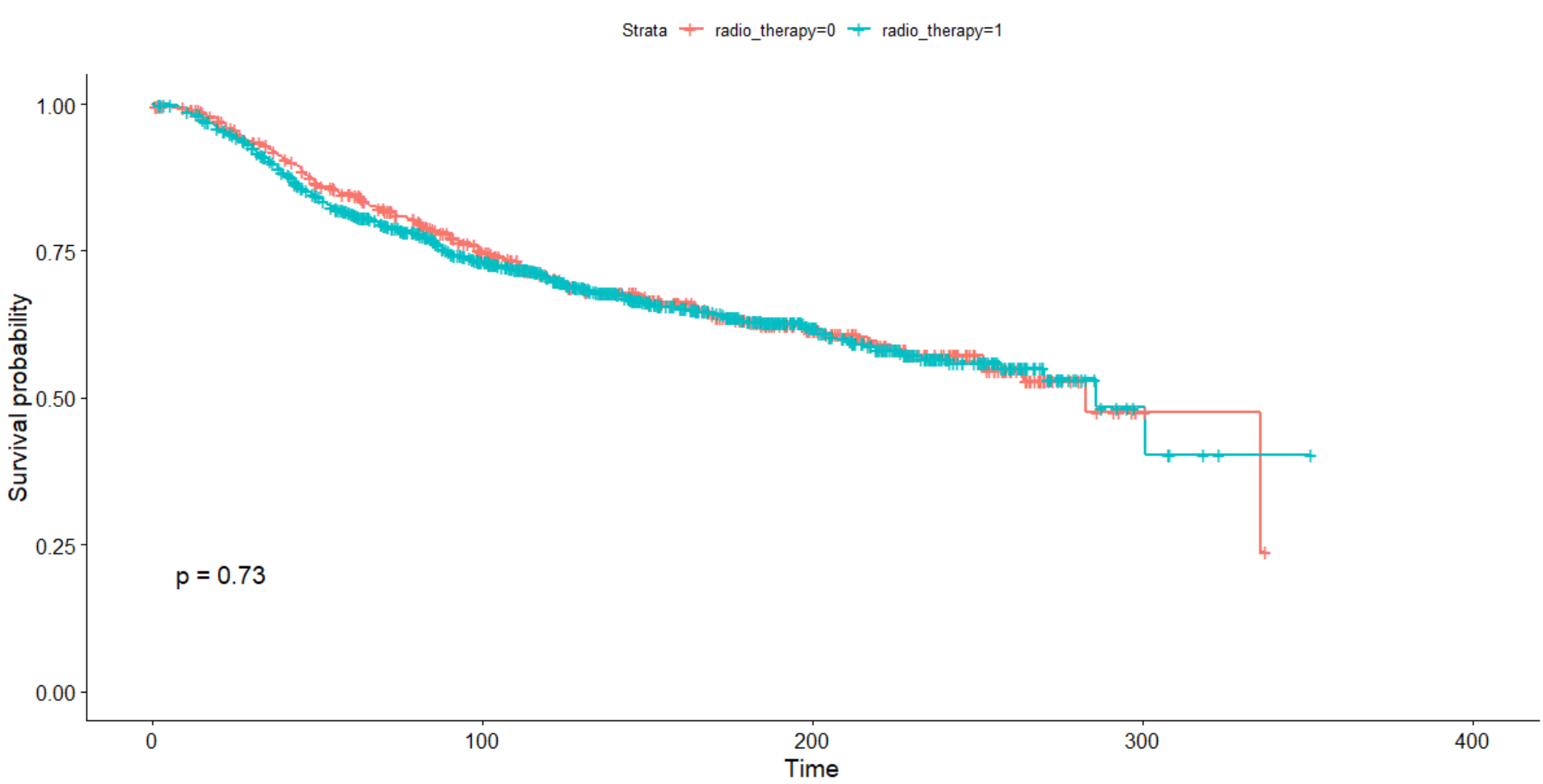
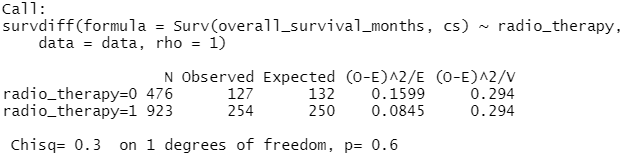
也嘗試以tumor stage將病人分為三組並分析:



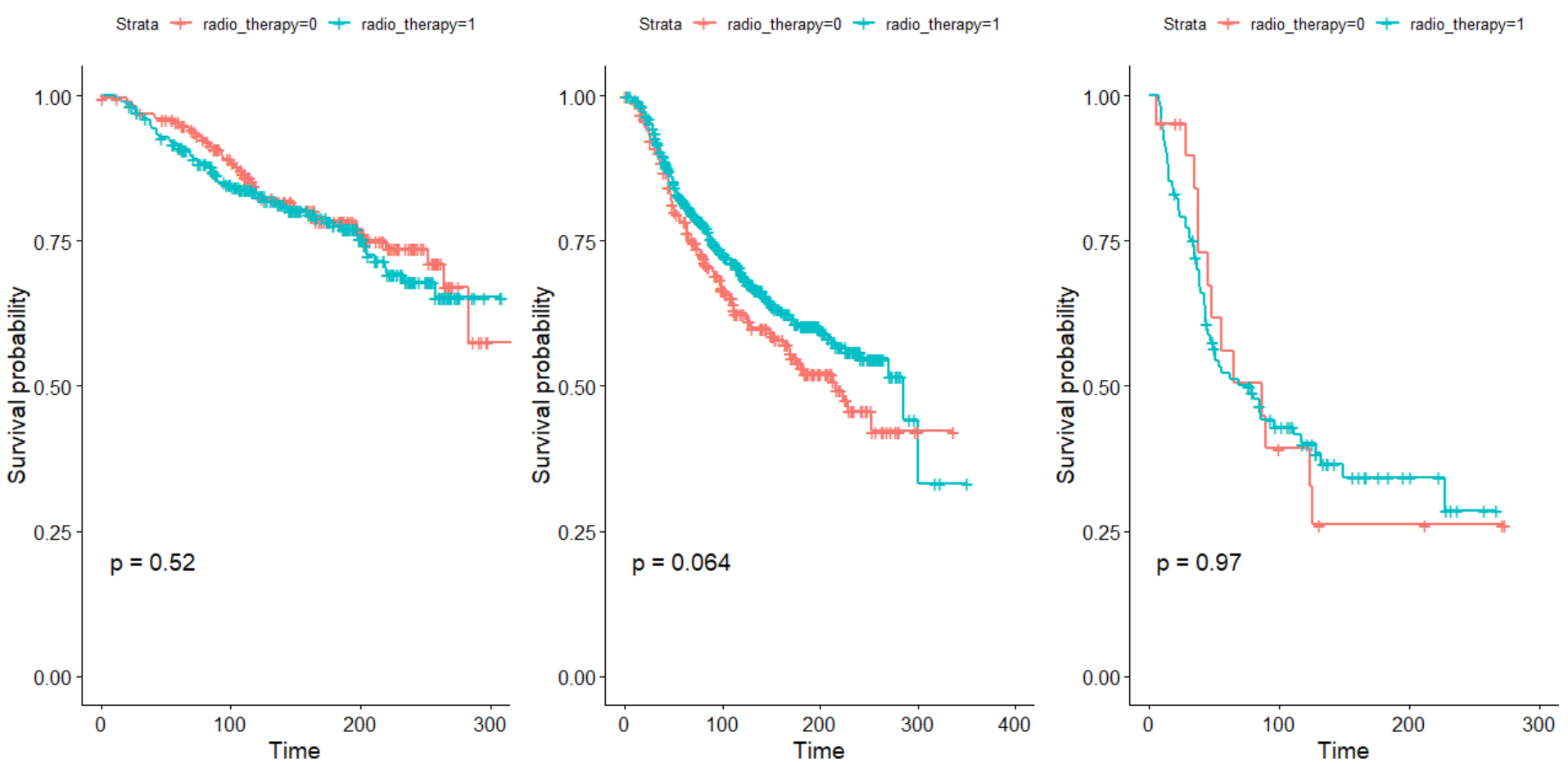
可以發現tumor stage=3 or 4的病人已經沒有生存曲線交叉的情形，且同時有顯著的結果，但tumor stage=2的病人還是有生存曲線交叉的現象。

### radio\_therapy



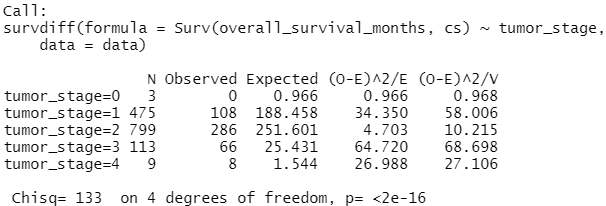
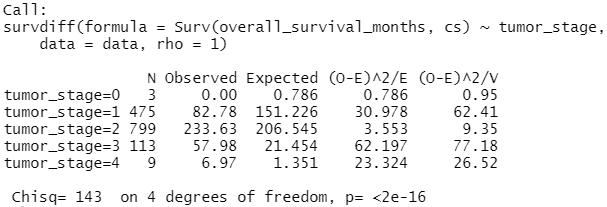


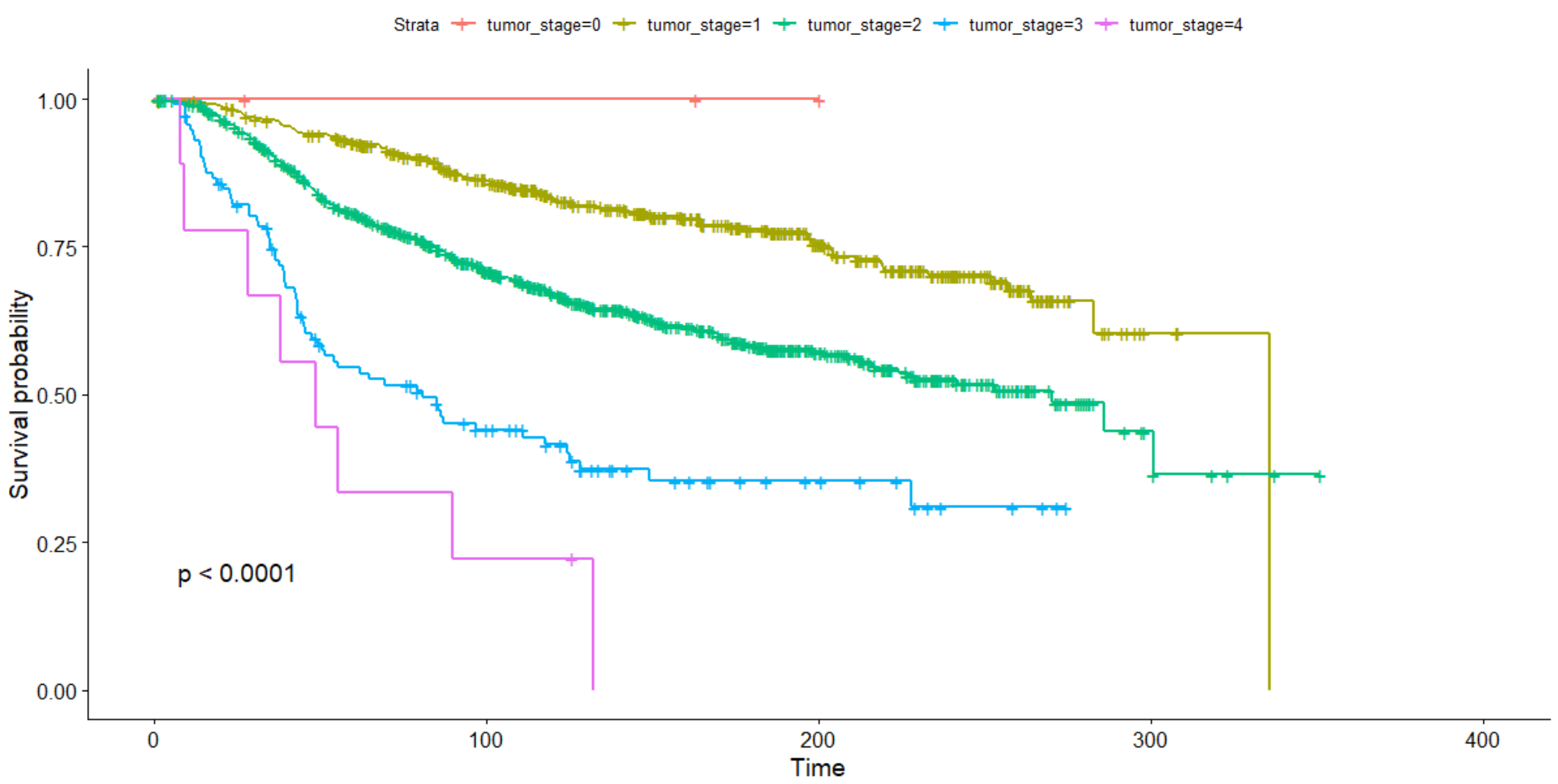
Log-rank test及Gehan’s test都說明沒有證據證明有/無接受放射治療的患者生存曲線存在差異，由圖也可以看出兩條生存曲線幾乎疊在一起。



就算以tumor stage分組後再分析，仍然都沒有顯著。

### tumor\_stage



Log-rank test及Gehan’s test都說明有證據證明不同期的患者生存曲線存在差異，由圖也可以明顯看出五期病人的生存曲線幾乎完全分開，而越後期的病人生存機率越低也非常符合常理。

# Method and testing

## Survival function

設T為病人存活時間的分配，定義生存函數為

## Hazard function

定義風險函數為

## Cox model

用來建模的變數為

其中

比例風險模型假設為

為我們有感興趣且想要估計的的係數，其意義為

所以可以得知，當增加時

## AFT model

用來建模的變數為

其中

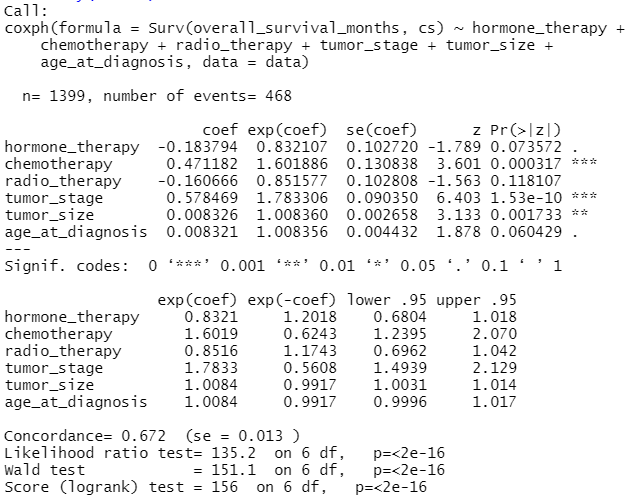
AFT model為

在此假設下可以推得

# Data analysis

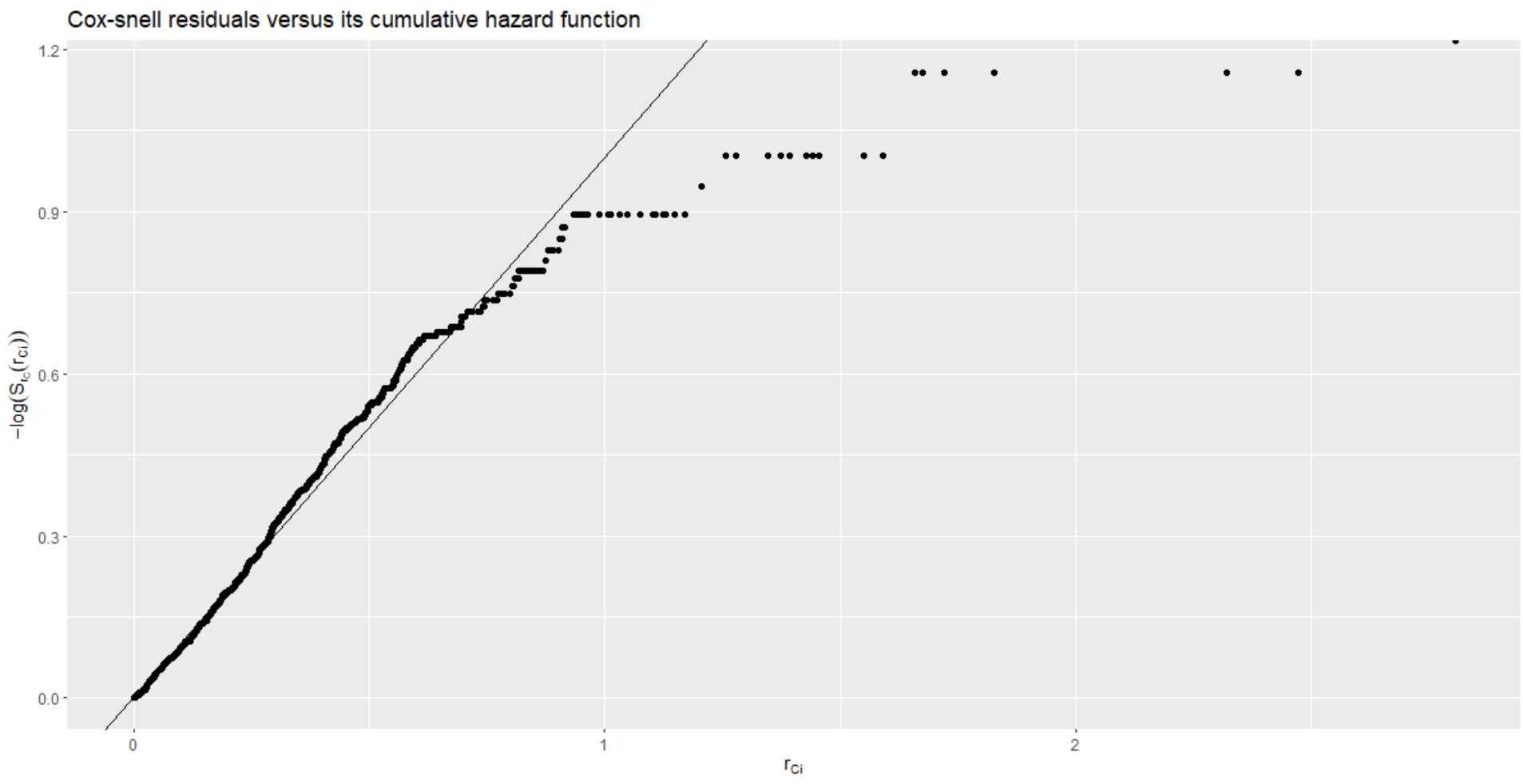
## Cox model

以比例風險模型套用於此資料，結果如下:



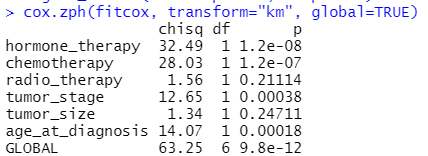
由報表可以看出較顯著的變數有chemotherapy、tumor\_stage及tumor\_size。三個治療方法中hormone thetapy和radio therapy都有負的係數，即有接受治療的病人生存函數較佳；chemotherapy卻反而有正的係數，原因可能是前面提到過的接受化療的病人本身病情就較嚴重，造成生存函數較差。

接著繪製Cox-Snell resuduals對其cumulative hazard function的圖以檢驗模型的配適度。



雖然尾端的點沒有貼近45度線，但資料點大多落在前端，即貼近45度線的地方，所以我認為若只由此圖判斷，此模型為可行的模型。

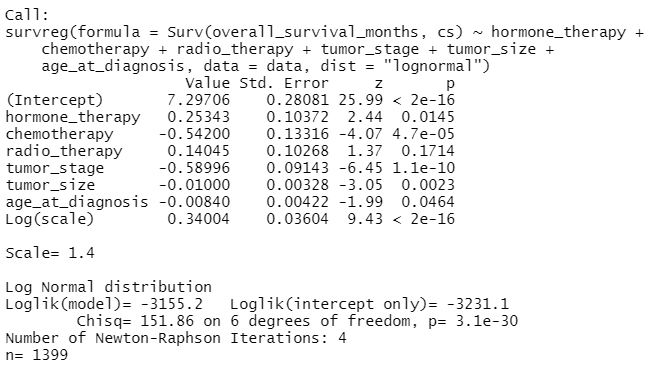
以survival套件裡的cox.zph函式，檢驗此模型是否符合比例風險的假設:

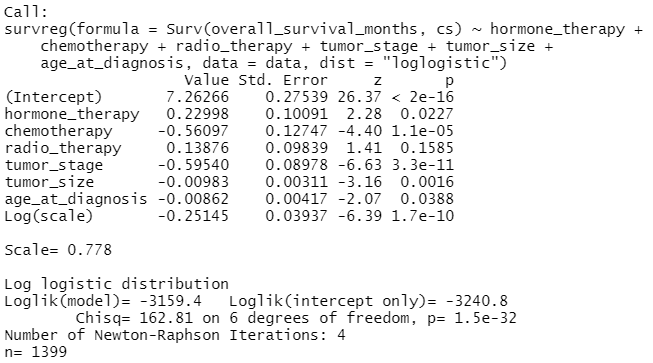


只有tumor\_size及radio\_therapy有符合比例風險假設。我認為其他變數不符合比例風險假設的原因可能有二:其一，如前面提到過若以是否接受hormone therapy分組，兩組病人的生存曲線有交叉的情況，這樣的變數自然不會符合比例風險假設。第二，像tumor\_stage是time dependent變數，但資料中沒有記錄隨著時間變數的變化，若直接將其視為time independent可能也是造成違背模型假設的原因。

總結上述結果，我認為Cox model對此資料並非非常好的模型，如果要得到更佳的結果，可能要試用其他更複雜的模型。

## AFT model

簡單的以套件嘗試了lognormal以及loglogistic分布的AFT模型，結果如下: 



兩個模型的結果皆是除了radio\_therapy以外其他變數皆顯著。其中hormone\_therapy對生存時間有正向影響，而chemothrrapy對生存時間有反向影響。

# Discussion

Log-rank test和Cox model的配適皆並不理想，所以我開始思考這個結果給了我什麼資訊，以及如何解決這些問題。

首先是化療對生存時間有反向影響，從Log-rank test、Cox model及AFT model的係數都得到這樣的結果。我認為此資料並非以隨機對照試驗(RCT)方式收集，即接受某種治療法的病人並非隨機分配的，所以生存函數的高低無法推得治療法的效用好壞。若是真的想探討治療法的效用，case control study可能是較可行方式。我的想法為先定義「治癒」，再從治癒者回推可能的成因(治療法)，並與配對病例(無治癒者)做對照，計算hazard ratio或odds ratio等指標來推論治療法的效用。

另外，生存曲線交叉的問題可以以two stage的Log-rank test執行，這也是可以再做進一步分析的部分。

模型配適的部分，既然最簡單通用的Cox model並不適用，AFT可能是可行的方法，但由於計算較困難，還需要再做進一步更精確的分析與配適度的檢驗。

總體而言，可以得到以下幾點結論:

1. Log\_rank test和Gehan’s test皆指出tumor\_stage顯著的影響生存機率；放射治療及賀爾蒙治療並沒有顯著的效果，但此結果有Log-rank test適用性的疑慮；化療對生存機率有顯著影響，但有接受化療者反而生存機率較低。
2. Cox model沒有很好的配適度，其中較顯著的變數有chemotherapy、tumor\_stage及tumor\_size。有接受化療者存活時間較短。
3. AFT model顯著的變數有age\_at\_diagnosis、chemotherapy、hormone\_therapy、tumor\_stage及tumor\_size。有接受賀爾蒙治療者存活時間較長，有接受化療者反而存活時間較短。