Survival analysis 期末專題

吳翔宇

一、研究目的

本次分析希望藉由此資料集來分析各個因子影響乳癌的嚴重性與重要性，找出對於乳癌患者來說影響其生存時間的重要因素。

二、資料簡介

本次使用資料為德國乳腺癌研究組 (GBSG) 於 1984 年 7 月到 1989 年 12 月召募了 720 名原發性淋巴結陽性乳腺癌患者，以研究 3 和 6 週期的化療和使用 tamoxifen(泰莫西芬) 進行額外激素治療的有效性。

- Tamoxifen

  抗雌激素 Tamoxifen (泰莫西芬) 是早期的標準荷爾蒙治療選擇，可使用於不管停經前停經後，淋巴結轉移與否，只要荷爾蒙受體陽性 (ER+) 皆可適用。

變數介紹

- pid : 患者 ID
- age : 年齡 (年)
- meno : 更年期狀態 (0 = 停經前, 1 = 停經後)
- size : 腫瘤大小 (mm)
- grade : 腫瘤分級 (低到高依序為輕等、中等及嚴重)
- nodes : 陽性淋巴結數
- pgr : 孕激素受體 (fmol/l) 是一種會在乳房、卵巢、子宮和子宮頸中作用的蛋白質受體。
- er : 雌激素受體 (fmol/l)
- hormon : 是否接受激素療法 (0= 否, 1 = 是)
- rfstime : 復發、死亡或失去追蹤時間
- status : 0 = 存活, 1 = 死亡

資料檢視

```
library(survival)
data("cancer",package = "survival")
head(gbsg,3)
```

```
##    pid age meno size grade nodes pgr er hormon rfstime status
## 1  132  49    0   18     2     2   0  0      0    1838      0
## 2 1575  55    1   20     3    16   0  0      0     403      1
## 3 1140  56    1   40     3     3   0  0      0    1603      0
```

```
tail(gbsg,3)
```

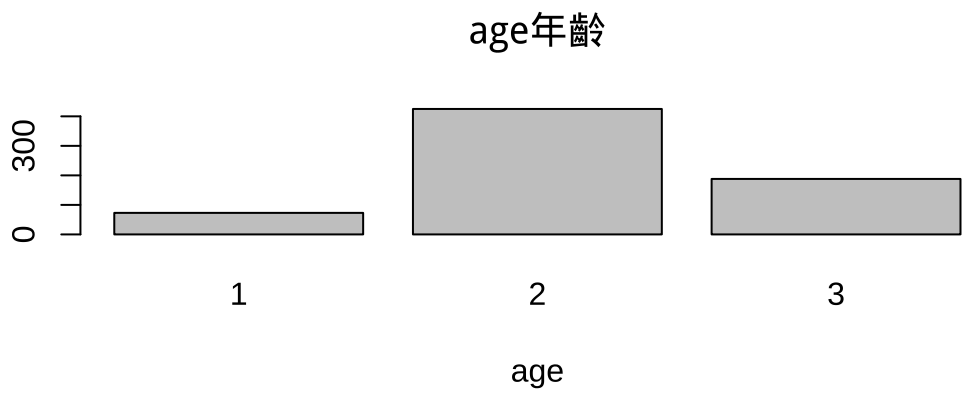
```
##      pid age meno size grade nodes  pgr  er hormon rfstime status
## 684 1525  57    1   35     3     1 1490 209      1    1342      0
## 685  736  44    0   21     2     3 1600  70      0     629      0
## 686  894  80    1    7     2     7 2380 972      1     758      0
```

資料處理 & 各變數介紹

1. age(年齡)

```
gbsg$age[gbsg$age <= 40] = 1
gbsg$age[gbsg$age > 40 & gbsg$age <= 60] = 2
gbsg$age[gbsg$age > 60 & gbsg$age <= 80] = 3
gbsg$age = factor(gbsg$age)
table(gbsg$age);plot(gbsg$age, main = "age 年齡", xlab = "age")
```

```
##
##   1   2   3
##  73 425 188
```

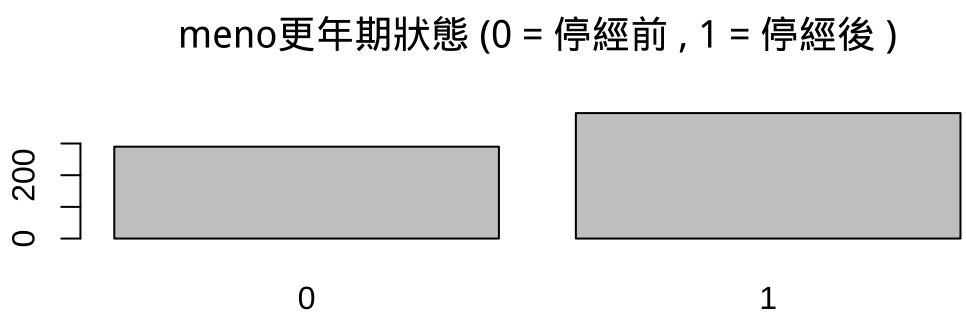


2. meno 更年期狀態 (0 = 停經前, 1 = 停經後)

```
gbsg$meno = factor(gbsg$meno)
table(gbsg$meno)
```

```
##
##   0   1
## 290 396
```

```
plot(gbsg$meno, main = "meno 更年期狀態 (0 = 停經前 , 1 = 停經後 )")
```

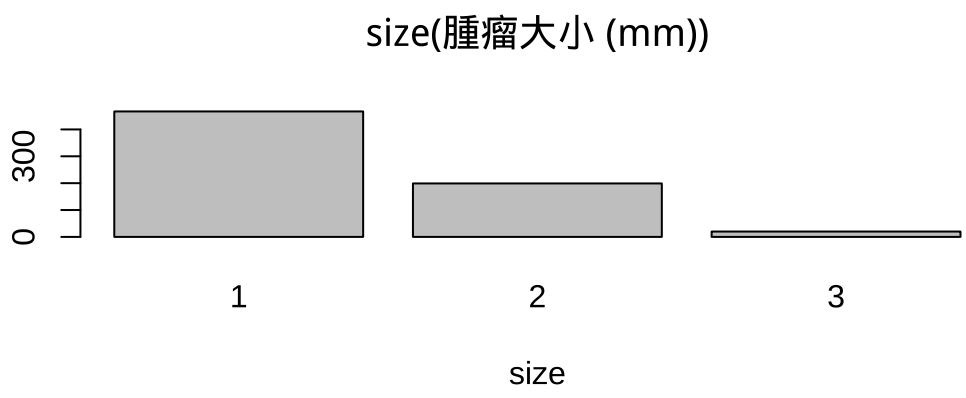


3. size(腫瘤大小 (mm))

將 size 分成三類，小於 30mm、介於 30-60mm 與大於 60mm

```
gbsg$size[gbsg$size <= 30] = 1
gbsg$size[gbsg$size > 30 & gbsg$size <= 60] = 2
gbsg$size[gbsg$size > 60] = 3
gbsg$size = factor(gbsg$size)
table(gbsg$size);plot(gbsg$size,main = "size(腫瘤大小 (mm))", xlab = "size")
```

```
##
##   1   2   3
## 467 199  20
```

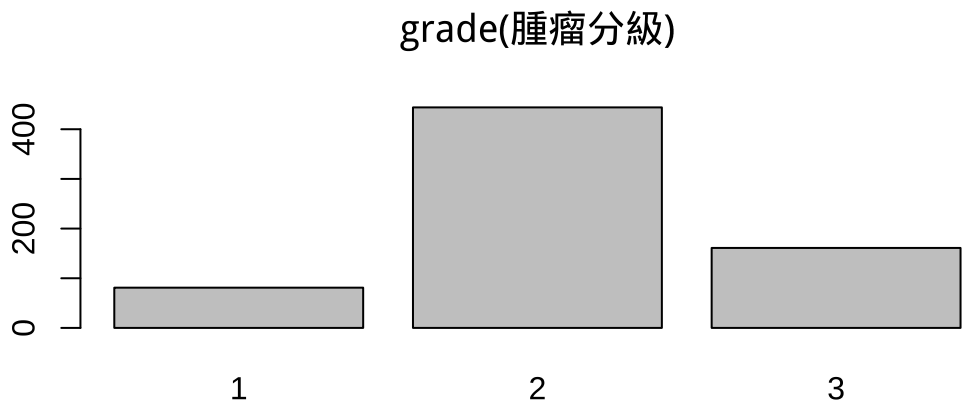


4. grade(腫瘤分級)

```
gbsg$grade = factor(gbsg$grade)
table(gbsg$grade)
```

```
##
##   1   2   3
##  81 444 161
```

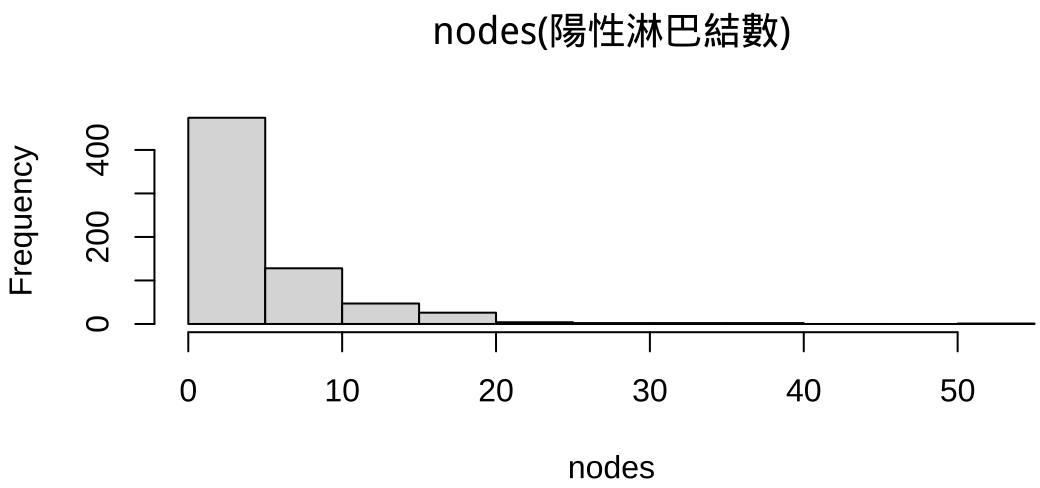
```
plot(factor(gbsg$grade), main = "grade(腫瘤分級)")
```



5. nodes(陽性淋巴結數)

```
summary(gbsg$nodes);hist(gbsg$nodes, main = "nodes(陽性淋巴結數)", xlab = "nodes")
```

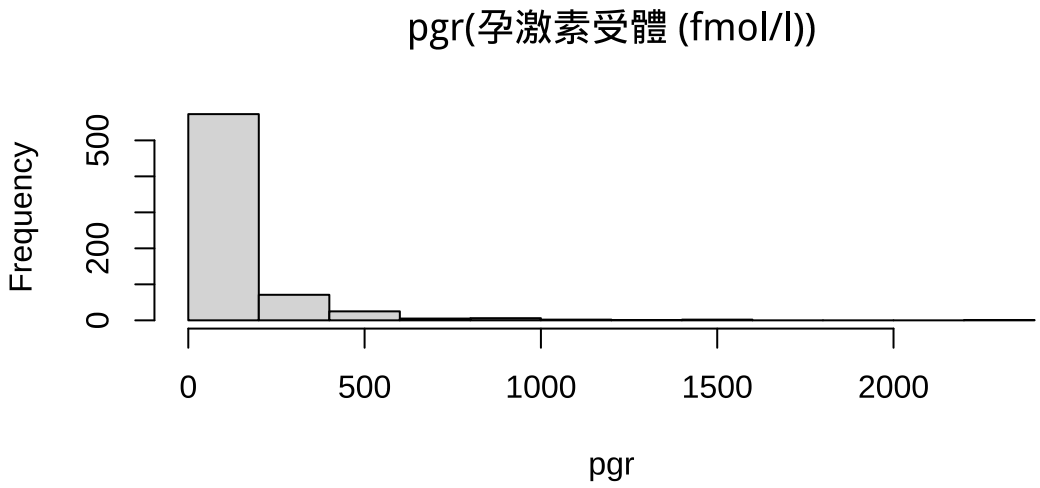
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    3.00    5.01    7.00   51.00
```



6. pgr(孕激素受體 (fmol/l))

```
summary(gbsg$pgr);hist(gbsg$pgr, main = "pgr(孕激素受體 (fmol/l))", xlab = "pgr")
```

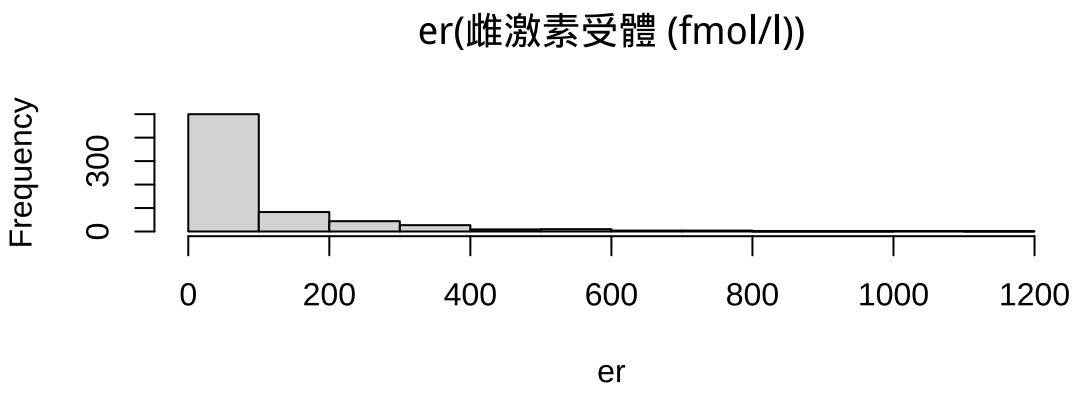
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     7.0    32.5   110.0   131.8  2380.0
```



7. er(雌激素受體 (fmol/l))

```
summary(gbsg$er);hist(gbsg$er, main = "er(雌激素受體 (fmol/l))", xlab = "er")
```

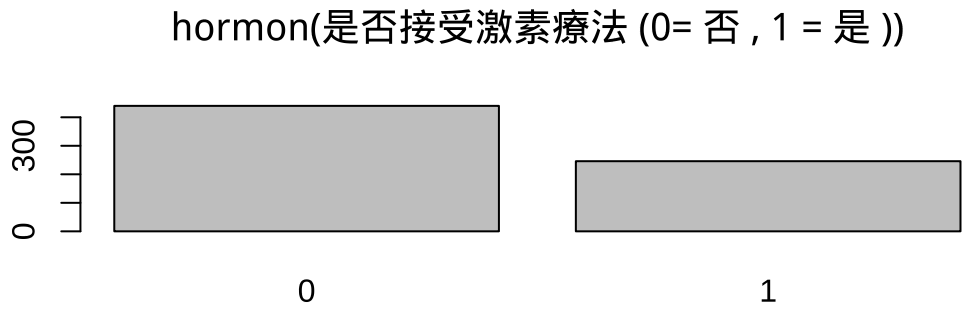
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   36.00   96.25  114.00 1144.00
```



8. hormon(是否接受激素療法 (0= 否, 1 = 是))

```
gbsg$hormon = factor(gbsg$hormon)
table(gbsg$hormon);plot(gbsg$hormon, main = "hormon(是否接受激素療法 (0= 否 , 1 = 是 ))")
```

```
##
##   0   1
## 440 246
```



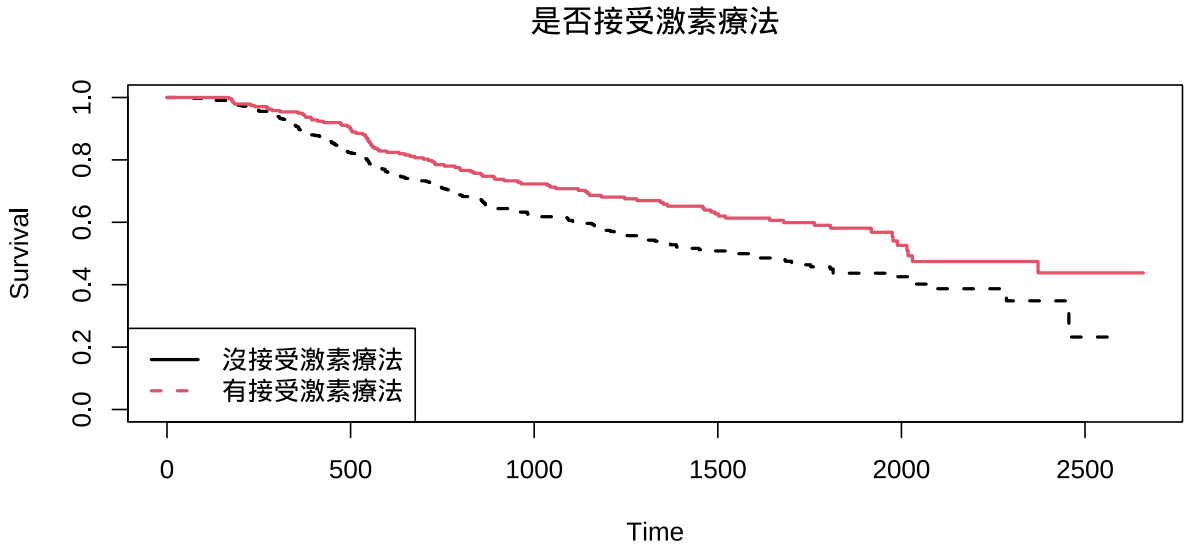
三、資料分析

1. Kaplan-Meier Survival estimate

首先使用 Kaplan-Meier Survival estimate 分別繪製出是否接受激素療法與停經前後存活曲線圖，接著使用 log rank test 檢定兩組之前是否真的有差別，最後因為此實驗為二因子實驗，我會將一因子固定的情況下照著上述的流程繪製出存活曲線圖，並做檢定。

是否接受激素療法

```
# 是否接受激素療法
fit_km_hormon = survfit(Surv(rfstime,status) ~ hormon, data = gbsg,
                        type="kaplan-meier",error="greenwood",conf.int=0.95)
plot(fit_km_hormon, lwd=2,lty=c(2,1),col = c(1,2),conf.int=F,mark.time=F,
     xlab="Time",ylab="Survival",main = " 是否接受激素療法")
legend("bottomleft",c(" 沒接受激素療法", " 有接受激素療法"),lwd=2,
       lty=c(1,2),col=c(1,2))
```




```
# log rank test
fit_diff_hormon = survdiff(Surv(rfstime,status) ~ hormon, data = gbsg)
fit_diff_hormon
```

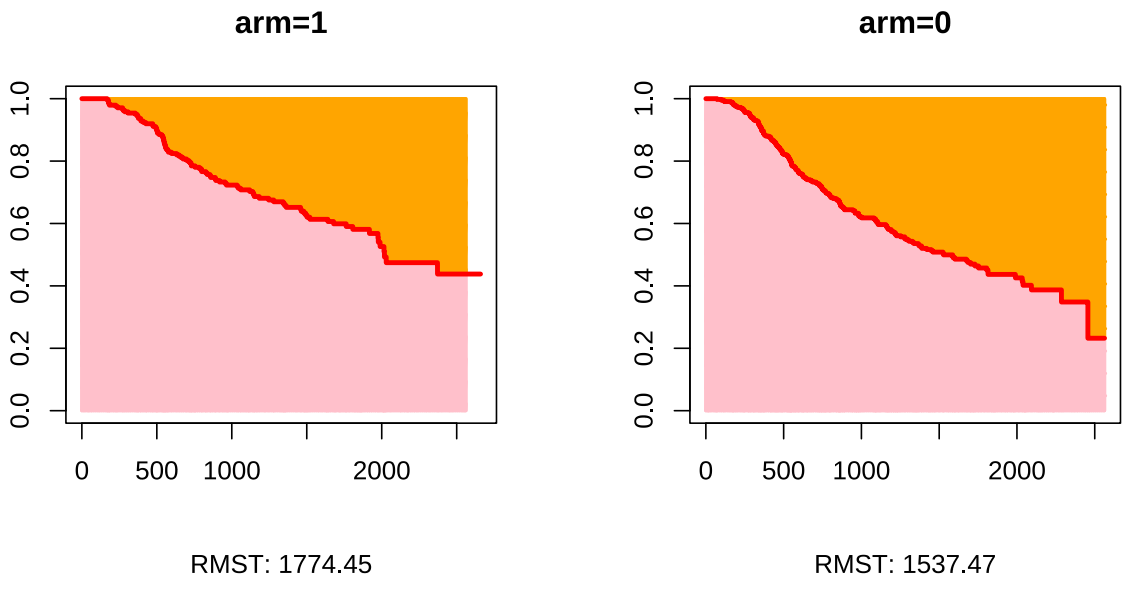
```
## Call:
## survdiff(formula = Surv(rfstime, status) ~ hormon, data = gbsg)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## hormon=0 440      205      180      3.37      8.56
## hormon=1 246       94      119      5.12      8.56
```

```
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

```
# mean survival time
library(survRM2)
(rmt = rmst2(gbsg$rfstime, gbsg$status, gbsg$hormon))
```

```
##
## The truncation time, tau, was not specified. Thus, the default tau  2563  is used.
##
## Restricted Mean Survival Time (RMST) by arm
##                   Est.     se lower .95 upper .95
## RMST (arm=1) 1774.447 61.636  1653.643  1895.251
## RMST (arm=0) 1537.470 49.253  1440.936  1634.004
##
##
## Restricted Mean Time Lost (RMTL) by arm
##                   Est.     se lower .95 upper .95
## RMTL (arm=1)  788.553 61.636   667.749   909.357
## RMTL (arm=0) 1025.530 49.253   928.996  1122.064
##
##
## Between-group contrast
##                           Est. lower .95 upper .95     p
## RMST (arm=1)-(arm=0) 236.977    82.340   391.613 0.003
## RMST (arm=1)/(arm=0)   1.154     1.052     1.266 0.002
## RMTL (arm=1)/(arm=0)   0.769     0.642     0.920 0.004
```

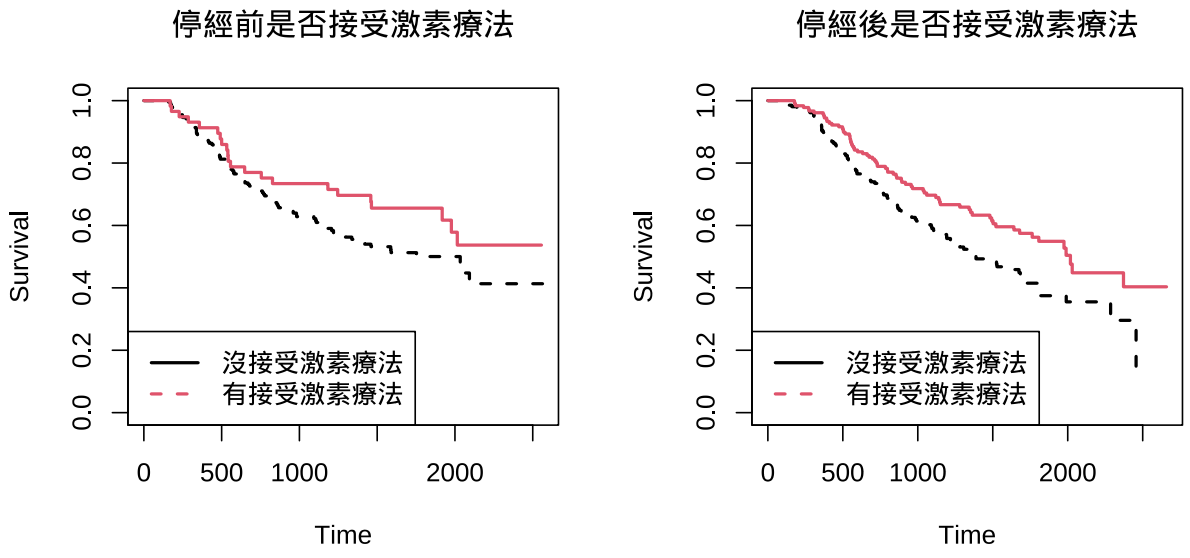
```
plot(rmt)
```



可以看到有接受激素療法的組別存活曲線一直在沒接受激素療法上，且在 log rank test 中 p-value 為 $0.003 < 0.05$，表示兩組之間存在差異，可以看到平均存活時間，在這個實驗下有接受激素療法的患者平均生存時間 95% 信賴區間為 (1654,1895) 天，沒接受激素療法的患者平均生存時間 95% 信賴區間為 (1441,1634) 天。

停經前後分別對於是否接受激素療法

```
par(mfrow = c(1,2))
# 停經前是否接受激素療法與否
fit_km_hormon0= survfit(Surv(rfstime,status) ~ hormon, data = gbsg,
                  subset = (meno == 0),
                  type="kaplan-meier",error="greenwood",conf.int=0.95)
plot(fit_km_hormon0, lwd=2,lty=c(2,1),col = c(1,2),conf.int=F,mark.time=F,
     xlab="Time",ylab="Survival",main = " 停經前是否接受激素療法")
legend("bottomleft",c(" 沒接受激素療法", " 有接受激素療法"),lwd=2,
       lty=c(1,2),col=c(1,2))

# 停經後是否接受激素療法與否
fit_km_hormon1= survfit(Surv(rfstime,status) ~ hormon, data = gbsg,
                  subset = (meno == 1),
                  type="kaplan-meier",error="greenwood",conf.int=0.95)
plot(fit_km_hormon1, lwd=2,lty=c(2,1),col = c(1,2),conf.int=F,mark.time=F,
     xlab="Time",ylab="Survival", main = " 停經後是否接受激素療法")
legend("bottomleft",c(" 沒接受激素療法", " 有接受激素療法"),lwd=2,
       lty=c(1,2),col=c(1,2))
```



```r
# log rank test with meno = 0
fit_diff_hormon0 = survdiff(Surv(rfstime,status) ~ hormon, data = gbsg,
                  subset = (meno == 0))
fit_diff_hormon0
```

```
## Call:
## survdiff(formula = Surv(rfstime, status) ~ hormon, data = gbsg,
##     subset = (meno == 0))
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## hormon=0 231       97     89.7     0.593      2.45
## hormon=1  59       22     29.3     1.817      2.45
##
##  Chisq= 2.4  on 1 degrees of freedom, p= 0.1
```

```r
# log rank test with meno = 1
fit_diff_hormon1 = survdiff(Surv(rfstime,status) ~ hormon, data = gbsg,
                      subset = (meno == 1))
fit_diff_hormon1
```

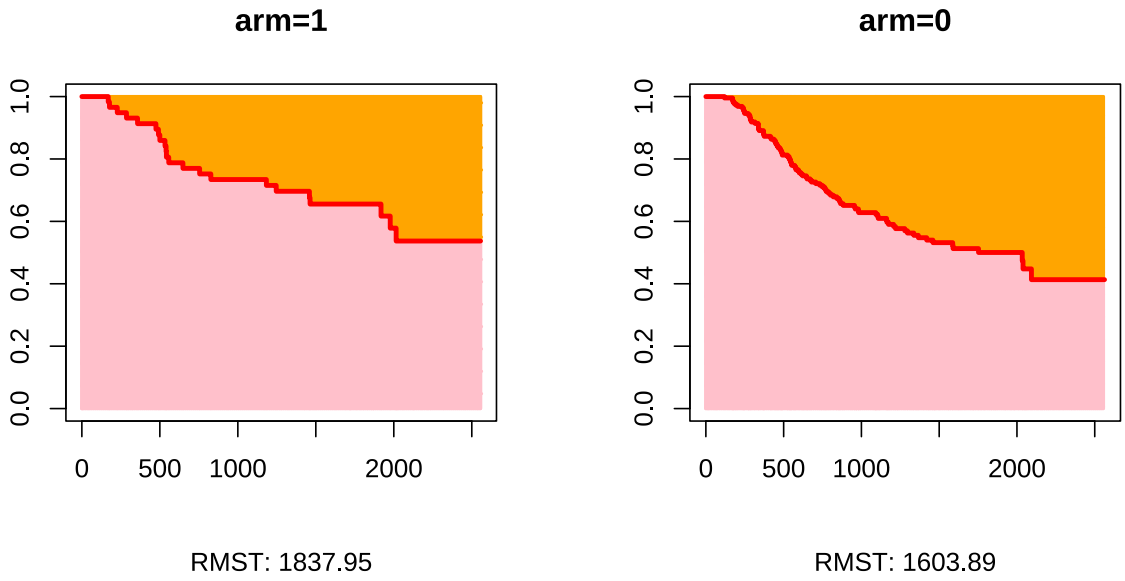
```
## Call:
## survdiff(formula = Surv(rfstime, status) ~ hormon, data = gbsg,
##     subset = (meno == 1))
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## hormon=0 209      108     90.1      3.54      7.12
## hormon=1 187       72     89.9      3.55      7.12
##
##  Chisq= 7.1  on 1 degrees of freedom, p= 0.008
```

```r
# mean survival time for meno = 0
par(mfrow = c(2,2))
(rmt1 = rmst2(gbsg$rfstime[gbsg$meno == 0], gbsg$status[gbsg$meno == 0], gbsg$hormon[gbs
```

```
##
## The truncation time, tau, was not specified. Thus, the default tau  2556  is used.
##
## Restricted Mean Survival Time (RMST) by arm
##                   Est.      se lower .95 upper .95
## RMST (arm=1) 1837.953 122.953  1596.970  2078.937
## RMST (arm=0) 1603.886  70.177  1466.341  1741.431
##
##
## Restricted Mean Time Lost (RMTL) by arm
##                  Est.      se lower .95 upper .95
## RMTL (arm=1) 718.047 122.953   477.063   959.030
```

```
## RMTL (arm=0) 952.114  70.177   814.569  1089.659
##
##
## Between-group contrast
##                          Est. lower .95 upper .95     p
## RMST (arm=1)-(arm=0) 234.067   -43.407   511.541 0.098
## RMST (arm=1)/(arm=0)   1.146     0.980     1.340 0.088
## RMTL (arm=1)/(arm=0)   0.754     0.523     1.087 0.130
```

```
plot(rmt1)
```

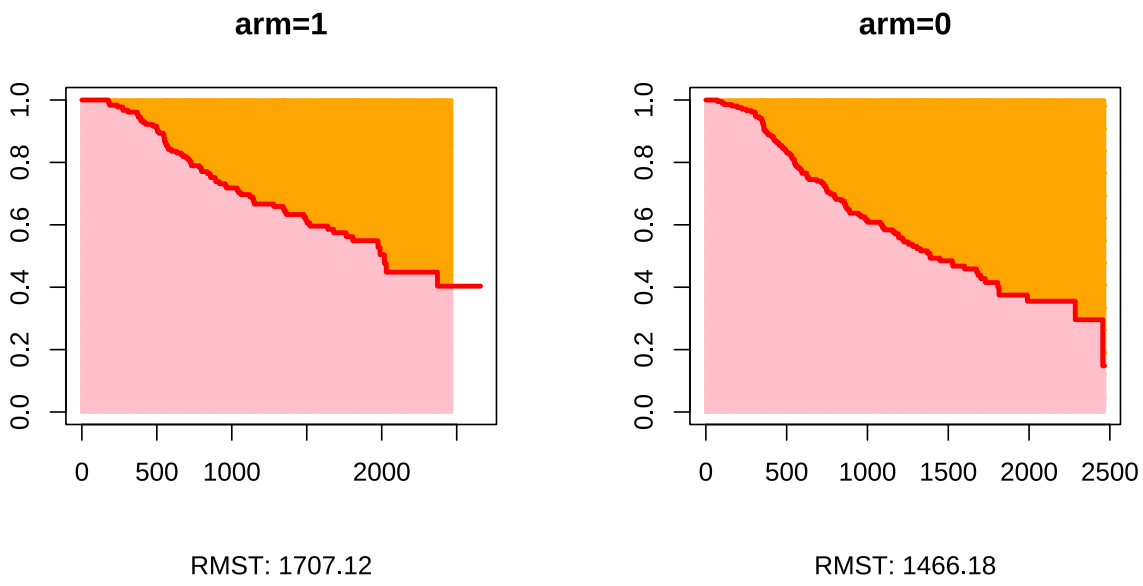



```
# mean survival time for meno = 1
(rmt2 = rmst2(gbsg$rfstime[gbsg$meno == 1], gbsg$status[gbsg$meno == 1], gbsg$hormon[gbs
```

```
##
## The truncation time, tau, was not specified. Thus, the default tau  2467  is used.
##
## Restricted Mean Survival Time (RMST) by arm
##                   Est.     se lower .95 upper .95
## RMST (arm=1) 1707.118 67.284  1575.244  1838.992
## RMST (arm=0) 1466.184 63.906  1340.930  1591.438
##
##
## Restricted Mean Time Lost (RMTL) by arm
##                   Est.     se lower .95 upper .95
## RMTL (arm=1)  759.882 67.284   628.008   891.756
## RMTL (arm=0) 1000.816 63.906   875.562  1126.070
##
##
```

```
## Between-group contrast
##                           Est. lower .95 upper .95     p
## RMST (arm=1)-(arm=0) 240.934    59.057   422.811 0.009
## RMST (arm=1)/(arm=0)   1.164     1.038     1.306 0.010
## RMTL (arm=1)/(arm=0)   0.759     0.613     0.940 0.012
```

```
plot(rmt2)
```



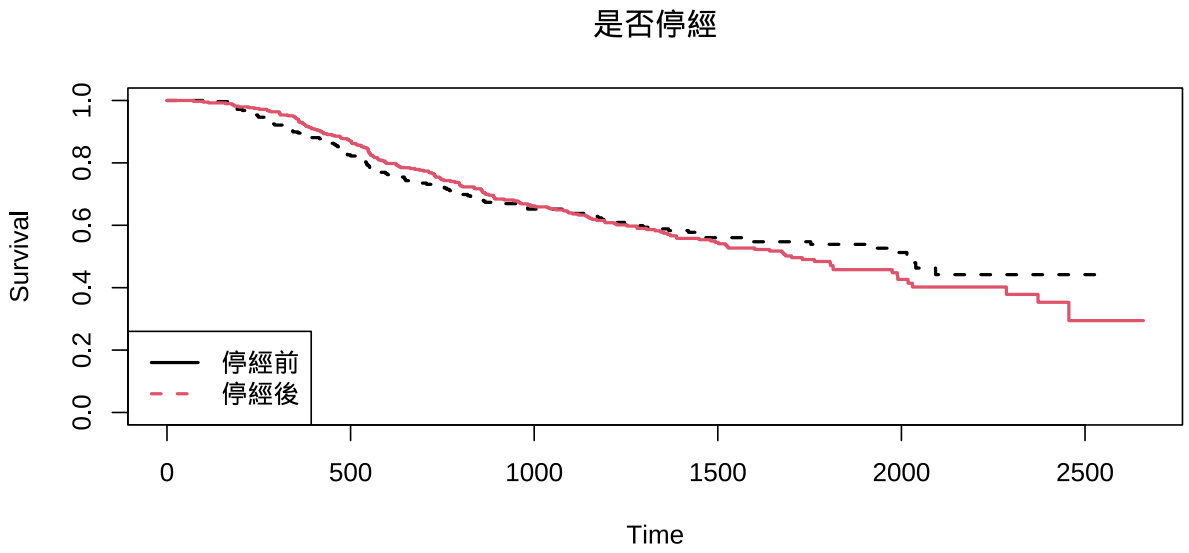
RMST: 1707.12

RMST: 1466.18

根據繪製出的存活曲線圖可以看到不管是在停經前或是停經後，有接受激素療法的存活曲線始終高於沒接受激素療法，但在 log rank test 中，停經前的 p-value 為 $0.1 > 0.05$ 表示停經前無論是否接受激素療法對於患者之存活率無顯著影響，停經後之 p-value 為 $0.008 < 0.05$ 表示停經後是否接受激素療法對於患者之存活率有顯著影響，可以看到平均存活時間，在這個實驗下停經前有接受激素療法的患者平均生存時間 95% 信賴區間為 $(1597,2079)$ 天，停經前沒接受激素療法的患者平均生存時間 95% 信賴區間為 $(1466,1741)$ 天，停經後有接受激素療法的患者平均生存時間 95% 信賴區間為 $(1575,1839)$ 天，停經後沒接受激素療法的患者平均生存時間 95% 信賴區間為 $(1341,1591)$ 天，可以發現停經前的平均存活時間是高於停經後的平均存活時間的。

是否停經

```
# 是否停經
fit_km_meno = survfit(Surv(rfstime,status) ~ meno, data = gbsg,
                      type="kaplan-meier",error="greenwood",conf.int=0.95)
plot(fit_km_meno, lwd=2,lty=c(2,1),col = c(1,2),conf.int=F,mark.time=F,
     xlab="Time",ylab="Survival",main = " 是否停經")
legend("bottomleft",c(" 停經前", " 停經後"),lwd=2,lty=c(1,2),col=c(1,2))
```

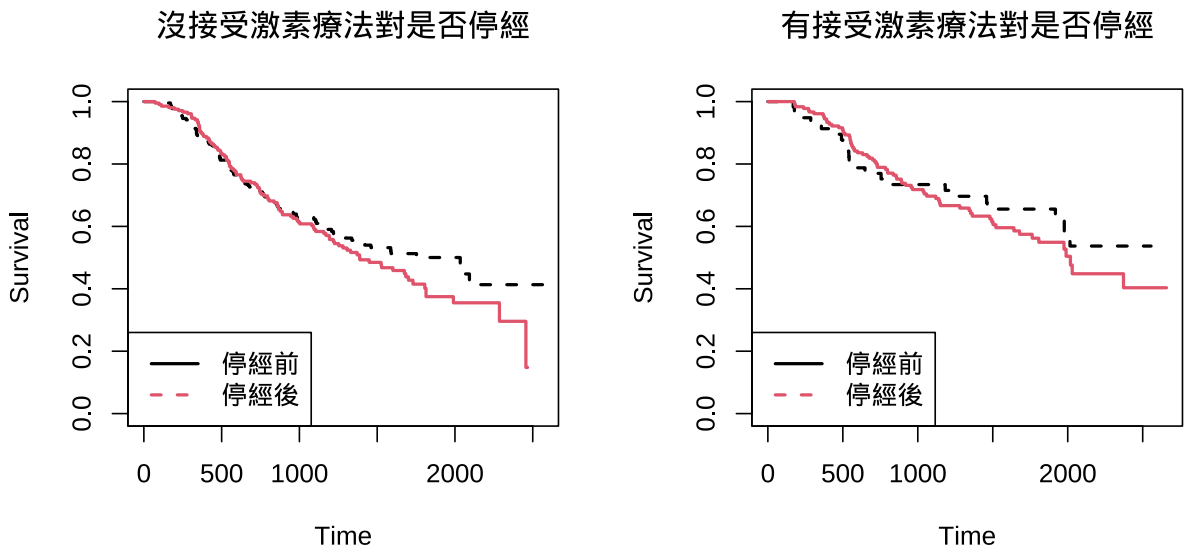



```
# log rank test
fit_diff_meno = survdiff(Surv(rfstime,status) ~ meno, data = gbsg)
fit_diff_meno
```

```
## Call:
## survdiff(formula = Surv(rfstime, status) ~ meno, data = gbsg)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## meno=0 290      119      124     0.164      0.28
## meno=1 396      180      175     0.115      0.28
##
##  Chisq= 0.3  on 1 degrees of freedom, p= 0.6
```

是否接受激素療法分別對是否停經

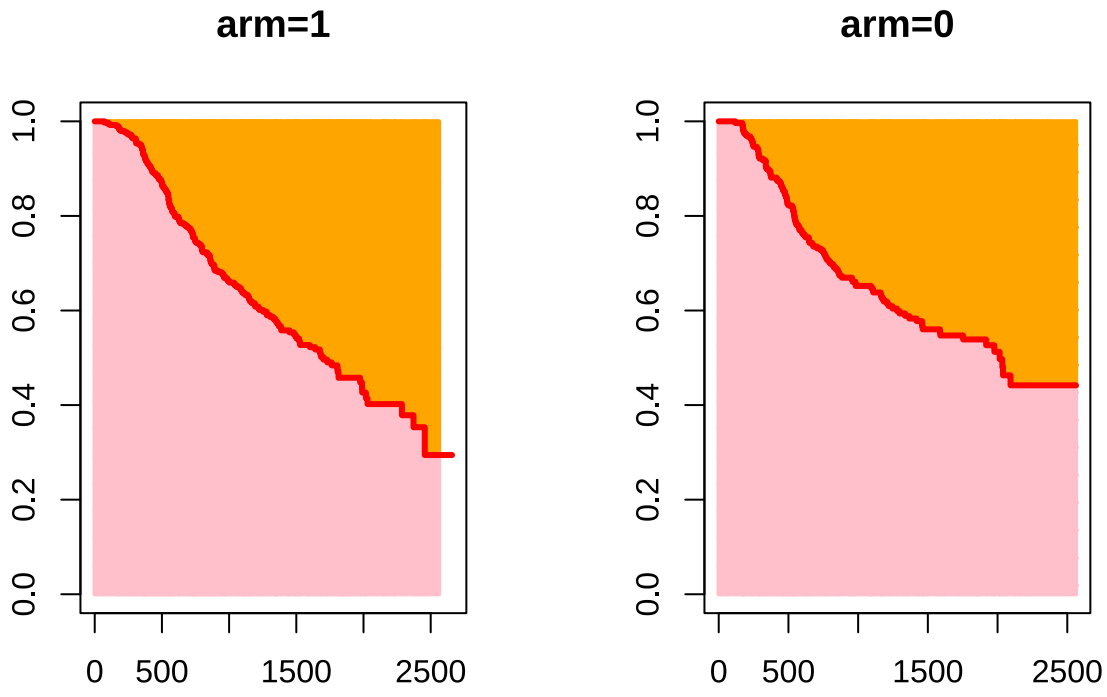
```
par(mfrow = c(1,2))
# 沒接受激素療法對是否停經
fit_km_meno0= survfit(Surv(rfstime,status) ~ meno, data = gbsg,
                  subset = (hormon == 0),
                  type="kaplan-meier",error="greenwood",conf.int=0.95)
plot(fit_km_meno0, lwd=2,lty=c(2,1),col = c(1,2),conf.int=F,mark.time=F,
     xlab="Time",ylab="Survival", main = " 沒接受激素療法對是否停經")
legend("bottomleft",c(" 停經前", " 停經後"),lwd=2,
       lty=c(1,2),col=c(1,2))

# 有接受激素療法對是否停經
fit_km_meno1= survfit(Surv(rfstime,status) ~ meno, data = gbsg,
                  subset = (hormon == 1),
                  type="kaplan-meier",error="greenwood",conf.int=0.95)
plot(fit_km_meno1, lwd=2,lty=c(2,1),col = c(1,2),conf.int=F,mark.time=F,
     xlab="Time",ylab="Survival", main = " 有接受激素療法對是否停經")
legend("bottomleft",c(" 停經前", " 停經後"),lwd=2,
       lty=c(1,2),col=c(1,2))
```



```r
# mean survival time
(rmt = rmst2(gbsg$rfstime, gbsg$status, gbsg$meno))
```

```
##
## The truncation time, tau, was not specified. Thus, the default tau  2563  is used.
##
## Restricted Mean Survival Time (RMST) by arm
##                   Est.     se lower .95 upper .95
## RMST (arm=1) 1609.853 49.396  1513.037  1706.668
## RMST (arm=0) 1660.333 61.137  1540.506  1780.160
##
##
## Restricted Mean Time Lost (RMTL) by arm
##                  Est.     se lower .95 upper .95
## RMTL (arm=1) 953.147 49.396   856.332  1049.963
## RMTL (arm=0) 902.667 61.137   782.840  1022.494
##
##
## Between-group contrast
##                           Est. lower .95 upper .95     p
## RMST (arm=1)-(arm=0) -50.480  -204.531   103.571 0.521
## RMST (arm=1)/(arm=0)   0.970     0.883     1.065 0.519
## RMTL (arm=1)/(arm=0)   1.056     0.893     1.248 0.523
```

```r
plot(rmt)
```



RMST: 1609.85 RMST: 1660.33

```
# log rank test with hormon = 0
fit_diff_meno0 = survdiff(Surv(rfstime,status) ~ meno, data = gbsg,
                  subset = (hormon == 0))
fit_diff_meno0
```

```
## Call:
## survdiff(formula = Surv(rfstime, status) ~ meno, data = gbsg,
##     subset = (hormon == 0))
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## meno=0 231       97      104     0.506      1.03
## meno=1 209      108      101     0.524      1.03
##
##  Chisq= 1  on 1 degrees of freedom, p= 0.3
```

```
# log rank test with hormon = 1
fit_diff_meno1 = survdiff(Surv(rfstime,status) ~ meno, data = gbsg,
                  subset = (hormon == 1))
fit_diff_meno1
```

```
## Call:
## survdiff(formula = Surv(rfstime, status) ~ meno, data = gbsg,
##     subset = (hormon == 1))
```

```
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## meno=0   59       22     25.3     0.424     0.588
## meno=1 187       72     68.7     0.156     0.588
##
##  Chisq= 0.6  on 1 degrees of freedom, p= 0.4
```

根據繪出的存活曲線圖可以看到兩組之間的存活曲線並沒有顯著差異，且在 logrank test 中 p-value 為 0.6 $>$ 0.05 表示停經前後對於存活率並沒有顯著影響，接著看到是否接受激素療法分別對是否停經的部分，在不管是否接受激素療法，p-value 分別為 0.3 及 0.4 大於 0.05 表示無論是否接受激素療法，停經前後對於存活率的影響皆不顯著，從平均存活時間來看停經後的平均存活時間 95% 信賴區間為 (1513,1707) 天，停經前的平均存活時間 95% 信賴區間為 (1541,1780) 天，兩者信賴區間互相重疊。

2. Cox PH model

(1) 考慮主效應

AIC function

```
AIC = function(fit){
  aic = -2*fit$loglik[2] + 2*length(fit$coefficients)
  return(aic)
}
```

step 1. 考慮 meno

```
fit_Cox1 = coxph(Surv(rfstime,status) ~ meno, data = gbsg)
summary(fit_Cox1)
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ meno, data = gbsg)
##
##   n= 686, number of events= 299
##
##          coef exp(coef) se(coef)    z Pr(>|z|)
## meno1 0.06265   1.06466  0.11824 0.53    0.596
##
##       exp(coef) exp(-coef) lower .95 upper .95
## meno1     1.065     0.9393    0.8444     1.342
##
## Concordance= 0.495  (se = 0.015 )
## Likelihood ratio test= 0.28  on 1 df,   p=0.6
## Wald test            = 0.28  on 1 df,   p=0.6
## Score (logrank) test = 0.28  on 1 df,   p=0.6
```

```
AIC(fit_Cox1)
```

```
## [1] 3577.928
```

可以看到 Cox PH model 下只加入 meno 的模型 Likelihood ratio test、Wald test 與 Score test 三者 p-value 皆大於 0.05，與先前 log rank test 檢定出的結果符合，表示停經前後對於存活率/風險比並沒有顯著影響，因此後面的模型不會將 meno 放入。

step 2. 考慮 hormon

```
fit_Cox2 = coxph(Surv(rfstime,status) ~ hormon, ties=c("breslow"), data = gbsg)
summary(fit_Cox2)
```

```
## Call:
```

```
## coxph(formula = Surv(rfstime, status) ~ hormon, data = gbsg,
##     ties = c("breslow"))
##
##   n= 686, number of events= 299
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## hormon1 -0.3639    0.6950   0.1250 -2.91  0.00361 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## hormon1     0.695      1.439    0.5439     0.888
##
## Concordance= 0.543  (se = 0.014 )
## Likelihood ratio test= 8.82  on 1 df,   p=0.003
## Wald test            = 8.47  on 1 df,   p=0.004
## Score (logrank) test = 8.56  on 1 df,   p=0.003
```

```
AIC(fit_Cox2)
```

```
## [1] 3569.53
```

將 hormon 放入模型，hormon 為顯著的且 AIC 為 3569.53

step 3. 加入第二個解釋變數

```
variable = names(gbsg)[c(2,4:8)]
aic = c(1:length(variable))

for(i in 1:length(variable)){
  dat = data.frame(gbsg[,colnames(gbsg) == variable[i]],gbsg[,c(9:11)])
  fit = coxph(Surv(rfstime,status) ~ hormon + .,
              ties=c("breslow"), data = dat)
  aic[i] = AIC(fit)
}
aic[which.min(aic)]
```

```
## [1] 3521.881
```

```
variable[which.min(aic)]
```

```
## [1] "nodes"
```

加入第二個解釋變數時，將 nodes 加入模型中使模型擁有最小的 AIC 且 AIC 低於只有 hormon 在模型內時，因此將 nodes 加入模型。

step 4. 加入第三個解釋變數

```
variable = names(gbsg)[c(2,4,5,7,8)]
aic = c(1:length(variable))

for(i in 1:length(variable)){
  dat = data.frame(gbsg[,colnames(gbsg) == variable[i]],gbsg[,c(6,9:11)])
  fit = coxph(Surv(rfstime,status) ~ hormon + nodes + .,
              ties=c("breslow"), data = dat)
  aic[i] = AIC(fit)
}
aic[which.min(aic)]
```

```
## [1] 3493.882
```

```
variable[which.min(aic)]
```

```
## [1] "pgr"
```

加入第三個解釋變數時，將 pgr 加入模型中使模型擁有最小的 AIC 且 AIC 低於模型內有 hormon 與 nodes 時，因此將 pgr 加入模型。

step 5. 加入第四個解釋變數

```
variable = names(gbsg)[c(2,4,5,8)]
aic = c(1:length(variable))

for(i in 1:length(variable)){
  dat = data.frame(gbsg[,colnames(gbsg) == variable[i]],gbsg[,c(6:7,9:11)])
  fit = coxph(Surv(rfstime,status) ~ hormon + nodes + .,
              ties=c("breslow"), data = dat)
  aic[i] = AIC(fit)
}
aic[which.min(aic)]
```

```
## [1] 3487.216
```

```
variable[which.min(aic)]
```

```
## [1] "grade"
```

加入第四個解釋變數時，將 grade 加入模型中使模型擁有最小的 AIC 且 AIC 低於模型內有 hormon、nodes 與 pgr 時，因此將 grade 加入模型。

step 6. 加入第五個解釋變數

```
variable = names(gbsg)[c(2,4,8)]
aic = c(1:length(variable))
```

```
for(i in 1:length(variable)){
  dat = data.frame(gbsg[,colnames(gbsg) == variable[i]],gbsg[,c(5:7,9:11)])
  fit = coxph(Surv(rfstime,status) ~ hormon + nodes + .,
              ties=c("breslow"), data = dat)
  aic[i] = AIC(fit)
}
aic[which.min(aic)]
```

```
## [1] 3487.37
```

```
variable[which.min(aic)]
```

```
## [1] "size"
```

加入第五個解釋變數時，將 size 加入模型中使模型擁有最小的 AIC，但 AIC 並沒有低於模型內有 hormon、nodes、pgr 與 grade 時太多，因此不將 size 加入模型，最終的模型為 hormon、nodes、pgr 與 grade。

Final Cox model

```
fit_Cox = coxph(Surv(rfstime,status) ~ hormon + nodes + pgr + grade,
                ties=c("breslow"), data = gbsg)
summary(fit_Cox)
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ hormon + nodes + pgr +
##     grade, data = gbsg, ties = c("breslow"))
##
##   n= 686, number of events= 299
##
##               coef  exp(coef)   se(coef)      z Pr(>|z|)
## hormon1 -0.3202656  0.7259562  0.1257009 -2.548  0.01084 *
## nodes    0.0542422  1.0557403  0.0067771  8.004 1.21e-15 ***
## pgr     -0.0022085  0.9977940  0.0005565 -3.968 7.23e-05 ***
## grade2   0.6522168  1.9197918  0.2487596  2.622  0.00874 **
## grade3   0.8080348  2.2434948  0.2677789  3.018  0.00255 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## hormon1    0.7260     1.3775    0.5674    0.9288
## nodes      1.0557     0.9472    1.0418    1.0699
## pgr        0.9978     1.0022    0.9967    0.9989
## grade2     1.9198     0.5209    1.1790    3.1261
## grade3     2.2435     0.4457    1.3274    3.7919
##
```

```
## Concordance= 0.687  (se = 0.015 )
## Likelihood ratio test= 99.13  on 5 df,   p=<2e-16
## Wald test            = 109.3  on 5 df,   p=<2e-16
## Score (logrank) test = 115.9  on 5 df,   p=<2e-16
```

最終模型為:

$$h(t|x) = h_0(t)e^{-0.324Z_{hormon}+0.049Z_{nodes}-0.002Z_{pgr}+0.644Z_{grade2}+0.788Z_{grade3}+0.007Z_{size}}$$

我們在此模型下有以下結論:

- hormon： 有接受激素療法的患者風險比是沒接受激素療法的患者 0.726 倍
- nodes： 陽性淋巴結數每多一單位風險比是 1.056 倍
- pgr： 孕激素受體每多一單位風險比是 0.998 倍
- grade： 腫瘤分級第二級的患者風險比是第一級患者的 1.920 倍，第三級患者是第一級患者的 2.243 倍

3. Accelerated failure time model (加速失效模型)

在建立完 Cox PH model 之後，我們確定了幾個對於風險比有顯著影響的變數包括 hormon、nodes、pgr 與 grade 四個變數，接著將這四個有顯著影響的變數對 AFT 模型做建模。

(1) Weibull distribution

```
fit_wei = survreg(Surv(rfstime,status) ~ hormon + nodes + pgr + grade,
                  data = gbsg, dist = "weibull")
(summary_fit_wei = summary(fit_wei))
```

```
##
## Call:
## survreg(formula = Surv(rfstime, status) ~ hormon + nodes + pgr +
##     grade, data = gbsg, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept)  8.120656   0.185966 43.67 < 2e-16
## hormon1      0.248519   0.090822  2.74  0.0062
## nodes       -0.042225   0.004938 -8.55 < 2e-16
## pgr          0.001638   0.000407  4.02 5.7e-05
## grade2      -0.485091   0.180655 -2.69  0.0072
## grade3      -0.605583   0.194072 -3.12  0.0018
## Log(scale)  -0.325854   0.048929 -6.66 2.7e-11
##
## Scale= 0.722
##
## Weibull distribution
## Loglik(model)= -2582.8   Loglik(intercept only)= -2637.3
##  Chisq= 108.93 on 5 degrees of freedom, p= 6.9e-22
## Number of Newton-Raphson Iterations: 6
## n= 686
```

```
# 加速因子
exp(summary_fit_wei$coefficients[-1])
```

```
##   hormon1     nodes       pgr    grade2    grade3
## 1.2821246 0.9586539 1.0016397 0.6156412 0.5457563
```

```
# \beta
exp(-summary_fit_wei$coefficients[-1]/summary_fit_wei$scale)
```

```
##   hormon1     nodes       pgr    grade2    grade3
## 0.7087510 1.0602354 0.9977332 1.9580603 2.3137303
```

我們在 Weibull AFT 模型下有以下結論:

- hormon： 有接受激素療法的患者壽命與風險分別是沒接受激素療法的患者 1.285 倍與 0.7088 倍

- nodes：陽性淋巴結數每多一單位壽命與風險分別為 0.959 倍與 1.06 倍
- pgr：孕激素受體每多一單位壽命與風險分別為 1.002 倍與 0.998 倍
- grade2：腫瘤分級第二級的患者壽命與風險分別是第一級患者的 0.616 倍與 1.958 倍
- grade3：腫瘤分級第三級的患者壽命與風險分別是第一級患者的 0.546 倍與 2.314 倍

(2) Loglogistic distribution

```
fit_logit = survreg(Surv(rfstime,status) ~ hormon + nodes + pgr + grade,
                    data = gbsg, dist = "loglogistic")
(summary_fit_logit = summary(fit_logit))
```

```
##
## Call:
## survreg(formula = Surv(rfstime, status) ~ hormon + nodes + pgr +
##     grade, data = gbsg, dist = "loglogistic")
##                 Value Std. Error      z       p
## (Intercept)  7.863956   0.176574  44.54 < 2e-16
## hormon1      0.311865   0.095048   3.28  0.0010
## nodes       -0.056167   0.008054  -6.97 3.1e-12
## pgr          0.001613   0.000385   4.19 2.8e-05
## grade2      -0.486416   0.170169  -2.86  0.0043
## grade3      -0.616247   0.188478  -3.27  0.0011
## Log(scale)  -0.552695   0.048467 -11.40 < 2e-16
##
## Scale= 0.575
##
## Log logistic distribution
## Loglik(model)= -2569.8   Loglik(intercept only)= -2627.9
##  Chisq= 116.39 on 5 degrees of freedom, p= 1.8e-23
## Number of Newton-Raphson Iterations: 4
## n= 686
```

```
# 加速因子
exp(summary_fit_logit$coefficients[-1])
```

```
##   hormon1     nodes       pgr    grade2    grade3
## 1.3659701 0.9453814 1.0016138 0.6148260 0.5399673
```

```
# \beta
exp(-summary_fit_logit$coefficients[-1]/summary_fit_logit$scale)
```

```
##   hormon1     nodes       pgr    grade2    grade3
## 0.5815844 1.1025372 0.9972014 2.3288084 2.9182770
```

我們在 Loglogistic AFT 模型下有以下結論:

- hormon：有接受激素療法的患者壽命與風險分別是沒接受激素療法的患者 1.366 倍與 0.582 倍
- nodes：陽性淋巴結數每多一單位壽命與風險分別為 0.945 倍與 1.103 倍
- pgr：孕激素受體每多一單位壽命與風險分別為 1.002 倍與 0.997 倍
- grade2：腫瘤分級第二級的患者壽命與風險分別是第一級患者的 0.615 倍與 2.329 倍
- grade3：腫瘤分級第三級的患者壽命與風險分別是第一級患者的 0.540 倍與 2.918 倍

四、結論

本次分析希望藉由此資料集來分析各個因子影響乳癌的嚴重性與重要性，我們藉由 Kaplan-Meier Survival estimate 方法繪出存活曲線，並使用 logrank test 檢定兩組的存活率是否有顯著差異，本研究的兩個主要因子為 meno(停經前與停經後) 和 hormon(是否接受激素療法)，在此方法下停經前後被檢定為不影響乳癌患者存活率的因子，但有接受激素療法的患者存活天數確實遠高於沒有接受激素療法的患者，後面我們也使用 Cox PH model、AFT model 來檢驗除了主要因子以外的其他因素哪些是嚴重影響乳癌患者風險與存活率，經由 Cox PH model 我們找出了幾個顯著影響乳癌患者風險的因子，像是 nodes(陽性淋巴結數)、pgr(孕激素受體)、grade(腫瘤分級) 與 size(腫瘤大小)，也算出了固定其他因子的情況下個別對乳癌患者壽命與風險的影響，未來可以將時間因素考慮進去或許可以更精進模型的分析能力。