13
Phân tích sự kiện
(event history hay survival analysis)

Qua ba chương trước, chúng ta đã làm quen với các mô hình thống kê cho các biến phụ thuốc liên tục (như áp suất máu) và biến bậc thứ (như có/không, bệnh hay không bệnh). Trong nghiên cứu khoa học, và đặc biệt là y học và kĩ thuật, có khi nhà nghiên cứu muốn tìm hiểu ảnh hưởng đến các biến phụ thuộc mang tính thời gian. Nhà kinh tế học John Maynard Keynes từng nói một câu có liên quan đến chủ đề mà tôi sẽ mô tả trong chương này như sau: "Về lâu về dài tất cả chúng ta đều chết, cái khác nhau là chết sớm hay chết muộn mà thôi." Thành ra, ở đây việc theo dõi hay mô tả một biến bậc thứ như sống hay chết tuy quan trọng, nhưng … không chính xác. Cái biến số quan trọng hơn và chính xác hơn là thời gian dẫn đến việc sự kiện xảy ra.

Trong các nghiên cứu y học, kể cả nghiên cứu lâm sàng, các nhà nghiên cứu thường theo dõi bệnh nhân trong một thời gian, có khi lên đến vài mươi năm. Biến cố xảy ra trong thời gian đó như có bệnh hay không có bệnh, sống hay chết, v.v… là những biến cố có ý nghĩa lâm sàng nhất định, nhưng thời gian dẫn đến bệnh nhân mắc bệnh hay chết còn quan trọng hơn cho việc đánh giá ảnh hưởng của một thuật điều trị hay một yếu tố nguy cơ. Nhưng thời gian này khác nhau giữa các bệnh nhân. Chẳng hạn như thời điểm từ lúc điều trị ung thư đến thời điểm bệnh nhân chết rất khác nhau giữa các bệnh nhân, và độ khác biệt đó có thể tùy thuộc vào các yếu tố như độ tuổi, giới tính, tình trạng bệnh, và các yếu tố mà có khi chúng ta không/chưa đo lường được như tương tác giữa các gen.

Mô hình chính để thể hiện mối liên hệ giữa thời gian dẫn đến bệnh (hay không bệnh) và các yếu tố nguy cơ (risk factors) là mô hình có tên là "survival analysis" (có thể tạm dịch là *phân tích sống sót*). Cụm từ "survival analysis" xuất phát từ nghiên cứu trong bảo hiểm, và giới nghiên cứu y khoa từ đó dùng cụm từ cho bộ môn của mình. Nhưng như nói trên, sống/chết không phải là biến duy nhất, vì trong thực tế chúng ta cũng có những biến như có bệnh hay không bệnh, xảy ra hay không xảy ra, và do đó, trong giới tâm lí học, người ta dùng cụm từ "event history analysis" (phân tích biến cố) mà tôi thấy có vẻ thích hợp hơn là *phân tích sống sót*. Ngoài ra, trong các bộ môn kĩ thuật, người ta dùng một cụm từ khác, *reliability analysis* (phân tích độ tin cậy), để chỉ cho khái niệm *survival analysis*. Tuy nhiên, trong chương này tôi sẽ dùng cụm từ *phân tích biến cố.*

13.1 Mô hình phân tích số liệu mang tính thời gian

Ví dụ 1. **Thời gian dẫn đến ngưng sử dụng IUD**. Một nghiên cứu về hiệu quả của một y cụ ngừa thai trên 18 phụ nữ, tuổi từ 18 đến 35. Một số phụ nữ ngưng sử dụng y cụ vì bị chảy máu. Còn số khác thì tiếp tục sử dụng. Bảng số liệu sau đây là thời gian

(tính bằng tuần) kể từ lúc bắt đầu sử dụng y cụ đến khi chảy máu (tức ngưng sử dụng) hay đến khi kết thúc nghiên cứu (tức vẫn còn sử dụng đến khi chấm dứt nghiên cứu).

Bảng 13.1 Thời gian dẫn đến ngưng sử dụng hay tiếp tục sử dụng y cụ IUD

Mã số bệnh nhân	Thời gian (tuần)	Tình trạng (ngưng=1 hay tiếp tục=0)
1	18	0
2	10	1
3	13	0
4	30	1
5	19	1
6	23	0
7	38	0
8	54	0
9	36	1
10	107	1
11	104	0
12	97	1
13	107	0
14	56	0
15	59	1
16	107	0
17	75	1
18	93	1

Câu hỏi đặt ra là mô tả thời gian ngưng sử dụng y cụ. Thuật ngữ "mô tả" ở đây có nghĩa là ước tính số trung vị thời gian dẫn đến ngưng sử dụng, hay xác suất mà phụ nữ ngưng sử dụng vào một thời điểm nào đó. Tình trạng tiếp tục sử dụng có khi gọi là "survival" (tức "sống sót").

Để giải quyết vấn đề trên, đối những phụ nữ đã ngưng sử dụng vấn để ước tính thời gian không phải là khó. Nhưng vấn đề quan trọng trong dữ liệu mang tính thời gian này là một số phụ nữ vẫn còn tiếp tục sử dụng, bởi vì chúng ta không biết họ còn sử dụng bao lâu nữa, trong khi nghiên cứu phải "đóng sổ" theo một thời điểm định trước. Những trường hợp đó được gọi bằng một thuật ngữ khó hiểu là "censored" hay "survival" (tức còn sống, hay còn tiếp tục sử dụng, hay biến cố chưa xảy ra).

Gọi T là thời gian còn tiếp tục sử dụng (có khi gọi là *survival time*). T là một biến ngẫu nhiên, với hàm mật độ (probability density distribution) $f(t)$, và hàm phân phối tích lũy (cumulative distribution) là:

$$F(t) = \int_{-\infty}^{t} f(s)\,ds$$

Đây là xác suất mà một cá nhân ngưng sử dụng (hay kinh qua biến cố) tại thời điểm t. Hàm bổ sung $S(t) = 1 - F(t)$ thường được gọi là hàm "sống sót" (survival function).

Số liệu thời gian T thường được mô phỏng bằng hai hàm xác suất: hàm sống sót và hàm nguy cơ (*hazard function*). Hàm sống sót như định nghĩa trên là xác suất một cá nhân còn "sống sót" (hay trong ví dụ trên, còn sử dụng y cụ) đến một thời điểm t. Hàm nguy cơ, thường được viết bằng kí hiệu $h(t)$ hay $\lambda(t)$ là xác suất mà cá nhân đó ngưng sử dụng (hay xảy ra biến cố) ngay tại thời điểm t.

$$h(t) = \lim_{\Delta \to 0} \frac{\Pr\left[\left(t \le T < t + \Delta t\right) | T \ge t\right]}{\Delta t} = \frac{f(t)}{S(t)}$$

sao cho $h(t)\ \delta t$ là xác suất một cá nhân ngưng sử dụng trong khoảng thời gian ngắn δt với điều kiện cá nhân đó sống đến thời điểm t. Từ mối liên hệ:

Pr(sống sót đến $t+\delta t$) = Pr(sống sót đến t) . Pr(sống sót đến δt | sống đến t)

chúng ta có:

$$1-F\left(t+\delta t\right)=\left(1-F\left(t\right)\right)\times\left(1-h\left(t\right)\delta t\right)$$

Từ đó, chúng ta có:

$$\delta tF'\left(t\right)=\left(1-F\left(t\right)\right)h\left(t\right)\delta t$$

Thành ra, hàm nguy cơ là:

$$h\left(t\right)=\frac{f\left(t\right)}{1-F\left(t\right)}$$

Và hàm nguy cơ tích lũy:

$$\Lambda\left(t\right)=\int_{-\infty}^{t}\lambda\left(u\right)du$$

Từ định nghĩa hàm nguy cơ $-h\left(t\right)=\frac{-f\left(t\right)}{1-F\left(t\right)}$, chúng ta có thể viết:

$$\Lambda\left(t\right)=-\log\left(1-F\left(t\right)\right)$$

Một số hàm nguy cơ có thể ứng dụng để mô tả thời gian này. Hàm đơn giản nhất là một hằng số, dẫn đến một mô hình Poisson (thuộc nhóm các luật phân phối mũ):

$$f\left(t\right)=\lambda e^{-\lambda t}\quad\left(t\geq 0\right)$$

Do đó:

$$F\left(t\right)=1-e^{-\lambda t}$$

Thành ra:

$$h(t)=\lambda$$

Những lí thuyết trên đây thoạt đầu mới xem qua có vẻ tương đối rắc rối, nhưng với số liệu thực tế thì sẽ dễ theo dõi hơn. Bây giờ chúng ta quay lại với số liệu từ **Ví dụ 1**. Để tiện việc theo dõi và tính toán, chúng ta cần phải sắp xếp lại dữ liệu trên theo thứ tự thời gian, bất kể đó là thời gian ngưng sử dụng hay còn tiếp tục sử dụng:

```
10  13*  18*  19  23*  30   36  38*  54*
56* 59   75   93  97  104* 107 107* 107*
```

Trong dãy số liệu trên dấu "*" là để đánh dấu thời gian censored (tức còn tiếp tục sử dụng IUD). Cách đơn giản nhất là chia thời gian từ 10 tuần (ngắn nhất) đến 107 tuần (lâu nhất) thành nhiều khoảng thời gian như trong bảng phân tích sau đây:

Bảng 13.2. Ước tính xác suất tích lũy cho mỗi khoảng thời gian

Mốc thời gian (t)	Khoảng thời gian (tuần)	Số phụ nữ lúc bắt đầu thời điểm (n_t)	Số phụ nữ ngưng sử dụng (d_t)	Xác suất ngưng sử dụng $h(t)$	Xác suất còn sử dụng p_t	Xác suất tích lủy S(t)
1	0 – 9	18	0	0.0000	1.0000	1.0000
2	10 – 18	18	1	0.0555	0.9445	0.9445
3	19 – 29	15	1	0.0667	0.9333	0.8815
4	30 – 35	13	1	0.0769	0.9231	0.8137
5	36 – 58	12	1	0.0833	0.9167	0.7459
6	59 – 74	8	1	0.1250	0.8750	0.6526
7	75 – 92	7	1	0.1428	0.8572	0.5594
8	93 – 96	6	1	0.1667	0.8333	0.4662
9	97 – 106	5	1	0.2000	0.8000	0.3729
10	107 –	3	1	0.3333	0.6667	0.2486

Trong bảng tính toán trên, chúng ta có:

- Cột thứ nhất là mốc thời gian (tạm kí hiệu là t). Cột này không có ý nghĩa gì, ngoại trừ sử dụng để làm chỉ số;

- Cột thứ 2 là khoảng thời gian (duration) tính bằng tuần. Như đề cập trên, chúng ta chia thời gian thành nhiều khoảng để tính toán, chẳng hạn như từ 0 đến 9 tuần, 10 đến 18 tuần, v.v… Chú ý rằng trong thực tế, chúng ta không có số liệu cho thời gian từ 0 đến 9 tuần, nhưng khoảng thời gian này đặt ra để làm cái mốc khởi đầu để tiện cho việc ước tính sau này. Đây chỉ là những phân chia tương đối tùy tiện và chỉ có tính cách minh họa; trong thực tế máy tính có thể làm việc đó cho chúng ta;

- Cột thứ 3 là số đối tượng nghiên cứu n_t (hay cụ thể hơn là số phụ nữ trong nghiên cứu này) *bắt đầu* một khoảng thời gian. Chẳng hạn như khoảng thời gian 0-9, tại thời điểm bắt đầu 0, có 18 phụ nữ (hay cũng có thể hiểu rằng số phụ nữ được theo dõi/quan sát ít nhất 0 tuần là 18 người).

  Trong khoảng thời gian 10–18, ngay tại thời điểm bắt đầu 10, chúng ta có 18 phụ nữ; nhưng trong khoảng thời gian 19–29, ngay tại thời điểm bắt đầu 19, chúng ta có 15 phụ nữ (cụ thể là: `19 23* 30 36 38* 54* 56* 59 75 93 97 104* 107 107* 107*`); vân vân.

Nói cách khác, cột này thể hiện số đối tượng với thời gian quan sát tối thiểu là *t*. Do đó, trong khoảng thời gian 97 – 106, chúng ta có 5 phụ nữ với thời gian theo dõi từ 97 tuần trở lên (`97 104* 107 107* 107*`).

- Cột thứ 4 trình bày số phụ nữ ngưng sử dụng y cụ d_t (hay biến cố xảy ra) trong một *khoảng thời gian*. Chẳng hạn như trong khoảng thời gian 10–18 tuần, có một phụ nữ ngưng sử dụng(tại `10` tuần); trong khoảng thời gian 19 – 29 tuần cũng có một trường hợp ngưng sử dụng (tại `19` tuần), v.v…

- Cột thứ 5 là xác suất nguy cơ $h(t)$ trong một khoảng thời gian. Một cách đơn giản, $h(t)$ được ước tính bằng cách lấy d_t chia cho n_t. Ví dụ trong khoảng thời gian 10-18 có 1 phụ nữ ngưng sử dụng (trong sô 18 phụ nữ), và xác suất nguy cơ là 1/18 = 0.0555. Xác suất này được ước tính cho từng khoảng thời gian.

- Cột thứ 6 là xác suất còn sử dụng cho một khoảng thời gian, tức lấy 1 trừ cho $h(t)$ trong cột thứ 5. Xác suất này không cung cấp nhiều thông tin, nhưng chỉ được trình bày để dễ theo dõi tính toán trong cột kế tiếp.

- Cột thứ 7 là xác suất tích lũy còn sử dụng y cụ $S(t)$ (hay cumulative survival probability). Đây là cột số liệu quan trọng nhất cho phân tích. Vì tính chất "tích lũy", cho nên cách ước tính được nhân từ hai hay nhiều xác suất.

  Trong khoảng thời gian 0-9, xác suất tích lũy chính là xác suất còn sử dụng trong cột 6, (vì không có ai ngưng sử dụng).

  Trong khoảng thời gian 10-18, xác suất tích lũy được ước tính bằng cách lấy xác suất còn sử dụng trong thời gian 0-9 nhân cho xác suất còn sử dụng trong thời gian 10-18, tức là: 1.000 x 0.9445 = 0.9445. Ý nghĩa của ước tính này là: xác suất còn sử dụng cho đến thời gian 9 tuần là 94.45%.

  Tương tự, trong khoảng thời gian 19-29 tuần, xác suất tích lũy còn sử dụng được tính bằng cách lấy xác suất tích lũy còn sử dụng đến tuần 10-18 nhân cho xác suất còn sử dụng trong khoảng thời gian 19-29: 0.9445 x 0.9333 = 0.8815. Tức là, xác suất còn sử dụng đến tuần 29 là 88.15%.

  Nói chung, công thức ước tính $S(t)$ là $\hat{S}(t) = \prod_{t=1}^{k} \left(\frac{n_t - d_t}{n_t} \right)$. Chú ý dấu mũ "^" trên $S(t)$ là để nhắc nhở rằng đó là ước số. Nếu gọi xác suất còn sử dụng trong khoảng thời gian t là p_t (tức cột 6), thì $S(t)$ cũng có thể tính bằng công thức: $\hat{S}(t) = \prod_{t=1}^{k} p_t$.

Phép ước tính được mô tả trên thường được gọi là *ước tính Kaplan-Meier* (Kaplan-Meier estimates), hay thỉnh thoảng cũng được gọi là *product-limit estimate*.

13.2 Ước tính Kaplan-Meier bằng R

Tất cả các tính toán trên, tất nhiên, có thể được tiến hành bằng R. Trong R có một package tên là `survival` (do Terry Therneau và Thomas Lumley phát triển) có thể ứng dụng để phân tích biến cố. Trong phần sau đây tôi sẽ hướng dẫn cách sử dụng package này.

Quay lại với Ví dụ 1, việc đầu tiên mà chúng ta cần làm là nhập dữ liệu vào R. Nhưng trước hết, chúng ta phải nhập package `survival` vào môi trường làm việc:

```
> library(survival)
```

Kế đến, chúng ta tạo ra hai biến số: biến thứ nhất gồm thời gian (hãy gọi là `weeks` cho trường hợp này), và biến thứ hai là chỉ số cho biết đối tượng ngưng sữ dụng y cụ (cho giá trị 1) hay còn tiếp tục sử dụng (cho giá trị 0) và đặt tên biến này là `status`. Sau đó nhập hai biến vào một dataframe (và gọi là `data`) để tiện việc phân tích.

```
> weeks  <- c(10, 13, 18, 19, 23, 30, 36, 38, 54,
             56, 59, 75, 93, 97, 104, 107, 107, 107)
> status <- c(1, 0, 0, 1, 0, 1, 1,0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0)
> data <- data.frame(duration, status)
```

Bây giờ, chúng ta đã sẵn sàng phân tích. Để ước tính Kaplan-Meier, chúng ta sẽ sử dụng hai hàm `Surv` và `survfit` trong package `survival`. Hàm `Surv` dùng để tạo ra một biến số hợp (combined variable) với thời gian và tình trạng. Ví dụ, trong lệnh sau đây:

```
> survtime <- Surv(weeks, status==1)
> survtime
[1]  10   13+  18+  19   23+  30   36   38+  54+  56+  59   75   93   97
[15] 104+ 107  107+ 107+
```

chúng ta sẽ có `survtime` là một biến với thời gian và dấu "+" (chỉ còn sống sót, hay censored observation, hay trong trường hợp này là còn sử dụng y cụ). Biến số này chỉ có giá trị và ý nghĩa cho phân tích của R, chứ trong thực tế, có lẽ chúng ta không cần nó.

Còn hàm `survfit` cũng khá đơn giản, chúng ta chỉ cần cung cấp hai thông số: thời gian và chỉ số như ví dụ sau đây:

```
> survfit(Surv(weeks, status==1))
```

Hay nếu đã có object `survtime` thì chúng ta chỉ đơn giản "gọi":

```
> survfit(survtime)
Call: survfit(formula = survtime)

      n  events  median 0.95LCL 0.95UCL
```

```
      18       9      93      59     Inf
```

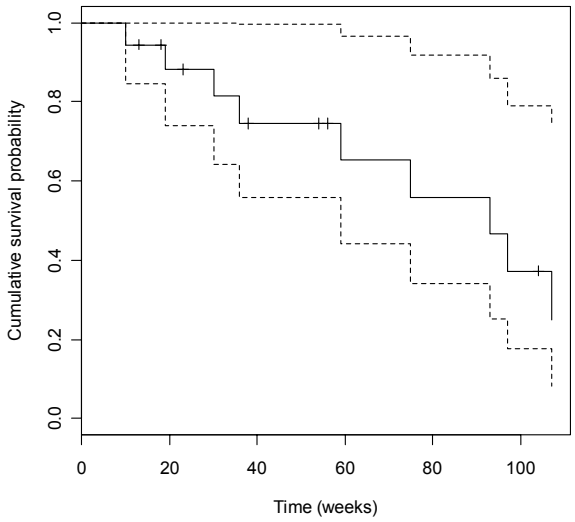
Kết quả trên đây chẳng có gì hấp dẫn, vì nó cung cấp những thông tin mà chún ta đã biết: có 9 biến cố (ngưng sử dụng y cụ) trong số 18 đối tượng. Thời gian (median - trung vị) ngưng sử dụng là 93 tuần, với khoảng tin cậy 95% từ 59 tuần đến vô cực (`Inf = infinity`). Để có thêm kết quả chúng ta cần phải đưa kết quả phân tích vào một object chẳng hạn như `kp` và dùng hàm `summary` để biết thêm chi tiết:

```
> kp <- survfit(Surv(weeks, status==1))
> summary(kp)
Call: survfit(formula = Surv(weeks, status == 1))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10     18       1    0.944  0.0540        0.844        1.000
   19     15       1    0.881  0.0790        0.739        1.000
   30     13       1    0.814  0.0978        0.643        1.000
   36     12       1    0.746  0.1107        0.558        0.998
   59      8       1    0.653  0.1303        0.441        0.965
   75      7       1    0.559  0.1412        0.341        0.917
   93      6       1    0.466  0.1452        0.253        0.858
   97      5       1    0.373  0.1430        0.176        0.791
  107      3       1    0.249  0.1392        0.083        0.745
```

Một phần của kết quả này (cột `time, n.risk, n.event, survival`) chúng ta đã tính toán "thủ công" trong bảng trên. Tuy nhiên R còn cung cấp cho chúng ta sai số chuẩn (standard error) của $S(t)$ và khoảng tin cậy 95%.

Khoảng tin cậy 95% được ước tính từ công thức $\hat{S}(t) \pm 1.96 \times se\left[\hat{S}(t)\right]$, mà trong đó, $se\left[\hat{S}(t)\right] = \hat{S}(t) \times \left\{ \sum_{t=1}^{k} \frac{d_t}{n_t (n_t - d_t)} \right\}$. Công thức sai số chuẩn này còn được gọi là *công thức Greenwood* (hay *Greenwood's formula*). Chúng ta có thể thể hiện kết quả trên bằng một biểu đồ bằng hàm `plot` như sau:

```
> plot(kp,
       xlab="Time (weeks)",
       ylab="Cumulative survival probability")
```



Trong biểu đồ trên, trục hoành là thời gian (tính bằng tuần) và trục tung là xác suất tích lũy còn sử dụng y cụ. Đường chính giữa chính là xác suất tích lũy $\hat{S}(t)$, hai đường chấm là khoảng tin cậy 95% của $\hat{S}(t)$. Qua kết quả phân tích này, chúng ta có thể phát biểu rằng xác suất sử dụng y cụ đến tuần 107 là khoảng 25% và khoảng tin cậy từ 8% đến 74.5%. Khoảng tin cậy khá rộng cho biết ước số có độ dao động cao, đơn giản vì số lượng đối tượng nghiên cứu còn tương đối thấp.

13.3 So sánh hai hàm xác suất tích lũy: kiểm định log-rank (log-rank test)

Phân tích trên chỉ áp dụng cho một nhóm đối tượng, và mục đích chính là ước tính $S(t)$ cho từng khoảng thời gian. Trong thực tế, nhiều nghiên cứu có mục đích so sánh $S(t)$ giữa hai hay nhiều nhóm khác nhau. Chẳng hạn như trong các nghiên cứu lâm sàng, nhất là nghiên cứu chữa trị ung thư, các nhà nghiên cứu thường so sánh thời gian sống sót giữa hai nhóm bệnh nhân để đánh giá mức độ hiệu nghiệm của một thuật điều trị.

Ví dụ 2. Một nghiên cứu trên 48 bệnh nhân với bệnh mụn giộp (herpes) ở bộ phận sinh dục nhằm xét nghiệm hiệu quả của một loại vắc-xin mới (tạm gọi bằng mã danh `gd2`). Bệnh nhân được chia thành 2 nhóm một cách ngẫu nhiên: nhóm 1 được điều trị bằng gd2 (gồm 25 người), và 23 người còn lại trong nhóm hai nhận giả dược (placebo). Tình trạng bệnh được theo dõi trong vòng 12 tháng. Bảng số liệu sau đây trình bày thời gian (tính bằng tuần và gọi tắt là `time`) đến khi bệnh tái phát. Ngoài ra, mỗi bệnh nhân còn cung cấp số liệu về số lần bị nhiễm trong vòng 12 tháng trước khi tham gia công trình nghiên cứu (`episodes`). Theo kinh nghiệm lâm sàng, `episodes` có liên hệ mật thiết đến xác suất bị nhiễm (và chúng ta sẽ quay lại với cách phân tích biến số này một một phần sau). Câu hỏi đặt ra là `gd2` có hiệu nghiệm làm giảm nguy cơ bệnh tái phát hay không.

Bảng 13.1. Thời gian đến nhiễm trùng ở bệnh nhân với bệnh mụn giộp cho nhóm gd2 và giả dược

id	episodes	time	infected
1	12	8	1
3	10	12	0
6	7	52	0
7	10	28	1
8	6	44	1
10	8	14	1
12	8	3	1
14	9	52	1
15	11	35	1
18	13	6	1
20	7	12	1
23	13	7	0
24	9	52	0
26	12	52	0
28	13	36	1
31	8	52	0
33	10	9	1
34	16	11	0
36	6	52	0
39	14	15	1
40	13	13	1
42	13	21	1
44	16	24	0
46	13	52	0
48	9	28	1

id	episodes	time	infected
2	9	15	1
4	10	44	0
5	12	2	0
9	7	8	1
11	7	12	1
13	7	52	0
16	7	21	1
17	11	19	1
19	16	6	1
21	16	10	1
22	6	15	0
25	15	4	1
27	9	9	0
29	10	27	1
30	17	1	1
32	8	12	1
35	8	20	1
37	8	32	0
38	8	15	1
41	14	5	1
43	13	35	1
45	9	28	1
47	15	6	1

Chú thích: trong biến infected (nhiễm), 1 có nghĩa là bị nhiễm, và 0 là không bị nhiễm.

Trong trường hợp trên chúng ta có hai nhóm để so sánh. Một cách phân tích đơn giản là ước tính $S(t)$ cho từng nhóm và từng khoảng thời gian, rồi so sánh hai nhóm bằng một kiểm định thống kê thích hợp. Song, phương pháp phân tích này có nhược điểm là nó không cung cấp cho chúng ta một "bức tranh" chung của tất cả các khoảng thời gian. Ngoài ra, vấn đề so sánh giữa hai nhóm trong nhiều khoảng thời gian khác nhau làm cho kết quả rất khó diễn dịch.

Để khắc phục hai nhược điểm so sánh trên, một phương pháp phân tích được phát triển có tên là log-rank test (kiểm định log-rank). Đây là một phương pháp phân tích phi thông số để kiểm định giả thiết rằng hai nhóm có cùng $S(t)$. Phương pháp này cũng chia thời gian ra thành k khoảng thời gian, $t_1, t_2, t_3, \ldots, t_k$, mà khoảng thời gian t_j (j = 1, 2, 3…, k) phản ánh thời điểm j khi một hay nhiều đối tượng của hai nhóm cộng lại. Gọi d_{ij} là số bệnh nhân trong nhóm i (i=1, 2) bị bệnh trong khoảng thời gian t_j. Gọi $d_j = d_{1j} + d_{2i}$ là tổng số bệnh nhân mắc bệnh và đặt $n_j = n_{1j} + n_{2j}$ là tổng số bệnh nhân của hai nhóm trong khoảng thời gian t_j. Với j = 1, 2, 3…, k, chúng ta có thể ước tính:

$$e_{1j} = \frac{n_{1j} d_j}{n_j} \qquad \text{và} \qquad e_{2j} = \frac{n_{2j} d_j}{n_j}$$

$$v_j = \frac{n_{1j} n_{2j} d_j \left(n_j - d_j\right)}{n_j^2 \left(n_j - 1\right)}$$

(ở đây, e_{1j}, e_{2j} là số bệnh nhân trong nhóm 1 và 2 mà chúng ta tiên đoán là sẽ mắc bệnh nếu có cùng xác suất mắc bệnh trong cả hai nhóm (tức xác suất trung bình), v_j là phương sai). Ngoài ra, chúng ta có thể ước tính tổng số bệnh nhân mắc bệnh cho nhóm 1 và 2:

$$O_1 = \sum_{j=1}^{k} d_{1j} \qquad \text{và} \qquad O_2 = \sum_{j=1}^{k} d_{2j}$$

Và tổng số bệnh nhân mắc bệnh nếu có cùng chung xác suất mắc bệnh cho cả hai nhóm:

$$E_1 = \sum_{j=1}^{k} v_j \qquad \text{và} \qquad V = \sum_{j=1}^{k} v_j$$

Gọi T_i là một biến ngẫu nhiên phản ánh thời gian từ khi được điều trị đến khi mắc bệnh cho nhóm *i*, và gọi $S_i\left(t\right) = \Pr\left(T_i \geq t\right)$, kiểm định log-rank được định nghĩa như sau:

$$\chi^2 = \frac{\left(O_1 - E_1\right)^2}{V}$$

Nếu $\chi^2 > \chi^2_{1,\alpha}$ (trong đó, $\chi^2_{1,\alpha}$ là trị số Chi bình phương với độ ý nghĩa thống kê α=0.95), chúng ta có bằng chứng để kết luận rằng độ khác biệt về *S*(t) giữa hai nhóm có ý nghĩa thống kê.

13.4 Kiểm định log-rank bằng R

Ví dụ 2 (tiếp tục). Chúng ta quay lại với ví dụ 2 và sẽ sử dụng R để tính toán kiểm định log-rank. Trước hết, chúng ta phải nhập các dữ liệu cần thiết bằng các lệnh thông thường như sau:

```
> group <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
           1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
           2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
           2, 2, 2, 2, 2, 2, 2, 2)

> episode <- c(12, 10, 7, 10, 6, 8, 8, 9, 11, 13, 7, 13, 9,
              12, 13, 8, 10, 16, 6, 14, 13, 13, 16, 13, 9,
              9, 10, 12,  7,  7,  7,  7, 11, 16, 16,  6, 15,
              9, 10, 17,  8,  8,  8,  8, 14, 13,  9, 15)

> time <- c(8, 12, 52, 28, 44, 14, 3, 52, 35, 6, 12, 7, 52,
          52, 36, 52,  9, 11, 52,15, 13, 21,24, 52,28,
          15,44, 2, 8,12,52,21,19, 6,10,15, 4, 9,27, 1,
```

```
         12,20,32,15, 5,35,28, 6)

> infected <- c(1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1,
              0, 1, 0, 0, 1, 1, 1, 0, 0, 1,
              1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1,
              1, 1, 0, 1, 1, 1, 1, 1)

> data <- data.frame(group, episode, time, infected)
```

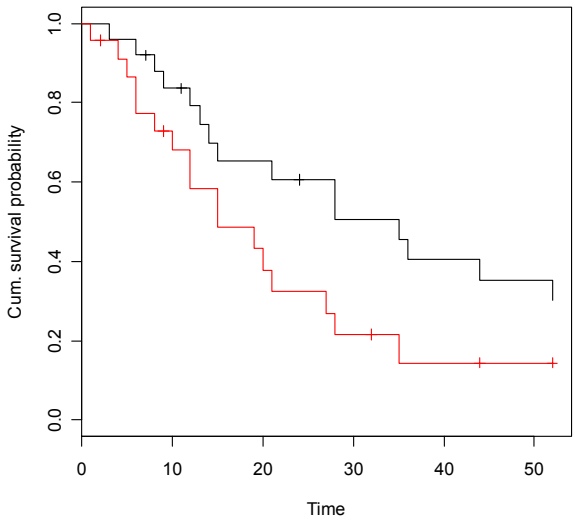
(a) Chúng ta ứng dụng hàm `survfit` để ước tính xác suất tích lũy $S(t)$ cho từng nhóm bệnh nhân và cho kết quả vào đối tượng `kp.by.group` như sau (chú ý cách cung cấp thông số ~ `group`):

```
> library(survival)
> kp.by.group <- survfit(Surv(time, infected==1) ~ group)
> summary(kp.by.group)
Call: survfit(formula = Surv(time, infected == 1) ~ group)

                group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3     25       1    0.960  0.0392        0.886        1.000
    6     24       1    0.920  0.0543        0.820        1.000
    8     22       1    0.878  0.0660        0.758        1.000
    9     21       1    0.836  0.0749        0.702        0.997
   12     19       1    0.792  0.0829        0.645        0.973
   13     17       1    0.746  0.0902        0.588        0.945
   14     16       1    0.699  0.0958        0.534        0.915
   15     15       1    0.653  0.1001        0.483        0.882
   21     14       1    0.606  0.1033        0.434        0.846
   28     12       2    0.505  0.1080        0.332        0.768
   35     10       1    0.454  0.1083        0.285        0.725
   36      9       1    0.404  0.1074        0.240        0.680
   44      8       1    0.353  0.1052        0.197        0.633
   52      7       1    0.303  0.1016        0.157        0.584

                group=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     23       1    0.957  0.0425       0.8767        1.000
    4     21       1    0.911  0.0601       0.8004        1.000
    5     20       1    0.865  0.0723       0.7346        1.000
    6     19       2    0.774  0.0889       0.6183        0.970
    8     17       1    0.729  0.0946       0.5650        0.940
   10     15       1    0.680  0.1000       0.5099        0.907
   12     14       2    0.583  0.1067       0.4072        0.835
   15     12       2    0.486  0.1088       0.3132        0.754
   19      9       1    0.432  0.1093       0.2630        0.709
   20      8       1    0.378  0.1082       0.2156        0.662
   21      7       1    0.324  0.1053       0.1712        0.613
   27      6       1    0.270  0.1007       0.1300        0.561
   28      5       1    0.216  0.0939       0.0921        0.506
   35      3       1    0.144  0.0859       0.0447        0.463
```

Và vẽ biểu đồ Kaplan-Meier cho từng nhóm như sau:

```
> plot(kp.by.group,
```

```
        xlab="Time",
        ylab="Cum. survival probability",
        col=c("black", "red"))
```



Qua biểu đồ trên, chúng ta có thể thấy khá rõ là nhóm được điều trị bằng `gd2` (đường màu đen phía trên) có xác suất nhiễm (hay bệnh tái phát) thấp hơn nhóm giả dược (đường màu đỏ, phía dưới). Nhưng phân tích trên không cung cấp trị số p để chúng ta phát biểu kết luận.

(b) Để có trị số p, chúng ta cần phải sử dụng hàm `survdiff` như sau:

```
> survdiff(Surv(time, infected==1) ~ group)
Call:
survdiff(formula = Surv(time, infected == 1) ~ group)

          N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 25       15     20.0      1.26      3.65
group=2 23       17     12.0      2.11      3.65

 Chisq= 3.7  on 1 degrees of freedom, p= 0.056
```

Kết quả phân tích log-rank cho trị số p=0.056. Vì p > 0.05, chúng ta vẫn chưa có bằng chứng thuyết phục để kết luận rằng `gd2` quả thật có hiệu nghiệm giảm nguy cơ tái phát bệnh.

13.5 Mô hình Cox (hay Cox's proportional hazards model)

Kiểm định log-rank là phương pháp cho phép chúng ta so sánh $S(t)$ giữa hai hay nhiều nhóm. Nhưng trong thực tế, $S(t)$ hay hàm nguy cơ $h(t)$ có thể không chỉ khác nhau giữa các nhóm, mà còn chịu sự chi phối của các yếu tố khác. Vấn đề đặt ra là làm sao ước tính mức độ ảnh hưởng của các yếu tố nguy cơ (risk factors) đến $h(t)$. Chẳng hạn

như trong nghiên cứu trên, số lần bệnh nhân từng bị nhiễm (biến `episode`) được xem là có ảnh hưởng đến nguy cơ bệnh tái phát. Do đó, vấn đề đặt ra là nếu chúng ta xem xét và điều chỉnh cho ảnh hưởng của `episode` thì mức độ khác biệt về $S(t)$ giữa hai nhóm có thật sự tồn tại hay không?

Vào khoảng giữa thập niên 1970s, David R. Cox, giáo sư thống kê học thuộc Đại học Imperial College (London, Anh) phát triển một phương pháp phân tích dựa vào mô hình hồi qui (regression) để trả lời câu hỏi trên (D.R. Cox, Regression models and life tables (with discussion), Journal of the Royal Statistical Society series B, 1972; 74:187-220). Phương pháp phân tích đó, sau này được gọi là *Mô hình Cox*. Mô hình Cox được đánh giá là một trong những phát triển quan trọng nhất của khoa học nói chung (không chỉ khoa học thống kê) trong thế kỉ 20! Không thể kể hết bao nhiêu số lần trích dẫn bài báo của David Cox, vì bài báo gây ảnh hưởng cho toàn bộ hoạt động nghiên cứu khoa học.

Vì mô tả chi tiết mô hình Cox nằm ngoài phạm vi của chương sách này, nên tôi chỉ phát hoạ vài nét chính để bạn đọc có thể nắm vấn đề. Gọi x_1, x_2, x_3, … x_p là p yếu tố nguy cơ. x có thể là các biến liên tục hay không liên tục. Mô hình Cox phát biểu rằng:

$$h(t) = \lambda(t) e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p}$$

$h(t)$ được định nghĩa như phần trên (tức hàm nguy cơ), β_j (j = 1, 2, 3, …, p) là hệ số ảnh hưởng liên quan đến x_j, và $\lambda(t)$ là hàm số nguy cơ nếu các yếu tố nguy cơ x không tồn tại (còn gọi là baseline hazard function). Vì mức độ ảnh hưởng của một yếu tố nguy cơ x_j thường được thể hiện bằng tỉ số nguy cơ (*hazard ratio*, HR, cũng tương tự như odds ratio trong phân tích hồi qui logistic), hệ số $\exp(\beta_j)$ chính là HR cho khi x_j tăng một đơn vị.

Hàm `coxph` trong package R có thể được ứng dụng để ước tính hệ số β_j. Trong lệnh sau đây:

```
> analysis <- coxph(Surv(time, infected==1) ~ group)
```

Trong lệnh trên, chúng ta muốn kiểm định ảnh hưởng của hai nhóm điều trị đến hàm nguy cơ $h(t)$ và kết quả được chứa trong đối tượng `analysis`. Để tóm lược `analysis`, chúng ta sử dụng hàm `summary`:

```
> summary(analysis)
Call:
coxph(formula = Surv(time, infected == 1) ~ group)

  n= 48
       coef exp(coef) se(coef)     z    p
group 0.684      1.98    0.363 1.88 0.06

      exp(coef) exp(-coef) lower .95 upper .95
group      1.98      0.505     0.973      4.04

Rsquare= 0.071   (max possible= 0.986 )
```

```
Likelihood ratio test= 3.55  on 1 df,   p=0.0597
Wald test            = 3.55  on 1 df,   p=0.0596
Score (logrank) test = 3.67  on 1 df,   p=0.0553
```

Nên nhớ nhóm điều trị được cho mã số 1, và nhóm giả dược có mã số 2. Do đó, kết quả phân tích trên cho biết khi `group` tăng 1 đơn vị thì $h(t)$ tăng 1.98 lần (với khoảng tin cậy 95% dao động từ 0.97 đến 4.04). Nói cách khác, nguy cơ bệnh tái phát trong nhóm giả dược cao hơn nhóm điều trị `gd2` gần 2 lần. Tuy nhiên vì khoảng tin cậy 95% bao gồm cả 1 và trị số p = 0.06, cho nên chúng ta vẫn không thể kết luận rằng mức độ ảnh hưởng này có ý nghĩa thống kê.

Nhưng chúng ta cần phải xem xét (và điều chỉnh) cho ảnh hưởng của quá trình bệnh trong quá khứ được đo lường bằng biến số `episode`. Để tiến hành phân tích này, chúng ta cho thêm `episode` vào hàm coxph như sau:

```
> analysis <- coxph(Surv(time, infected==1) ~ group + episode)
> summary(analysis)
Call:
coxph(formula = Surv(time, infected == 1) ~ group + episode)

  n= 48
         coef exp(coef) se(coef)    z      p
group   0.874      2.40   0.3712 2.35 0.0190
episode 0.172      1.19   0.0648 2.66 0.0079

        exp(coef) exp(-coef) lower .95 upper .95
group        2.40      0.417      1.16      4.96
episode      1.19      0.842      1.05      1.35

Rsquare= 0.196   (max possible= 0.986 )
Likelihood ratio test= 10.5  on 2 df,   p=0.00537
Wald test            = 10.4  on 2 df,   p=0.00555
Score (logrank) test = 10.6  on 2 df,   p=0.00489
```

Kết quả phân tích trên cho chúng ta một diễn dịch khác và có lẽ chính xác hơn. Mô hình $h(t)$ bây giờ là:

$$h\left(t \mid group, episode\right) = \lambda\left(t\right) e^{0.874(group)+0.172(episode)}$$

Nếu `episode` tạm thời giữ cố định, tỉ số $h(t)$ giữa hai nhóm là:

$$\frac{h\left(t \mid group = 2\right)}{h\left(t \mid group = 1\right)} = e^{0.874(2-1)} = 2.40$$

Tương tự, nếu `group` tạm thời giữ cố định, khi `episode` tăng một đơn vị, tỉ số nguy cơ sẽ tăng 1.14 lần.

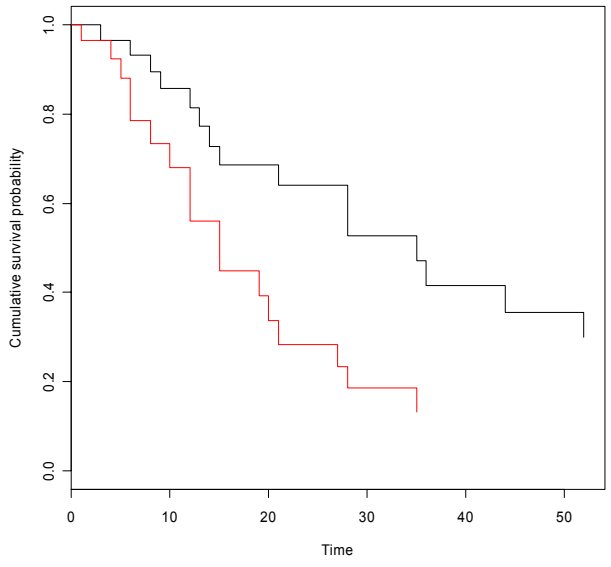
Nói cách khác, mỗi lần mắc bệnh trong quá khứ (tức `episode` tăng 1 đơn vị) làm tăng nguy cơ tái phát bệnh 19% (với khoảng tin cậy 95% dao động từ 5% đến 35%). Nhóm giả dược có nguy cơ bệnh tái phát tăng gấp 2.4 lần so với nhóm điều trị bằng `gd2` (và khoảng tin cậy 95% có thể từ 1.2 đến gần 5 lần). Cả hai yếu tố (nhóm điều trị) và `episode` đều có ý nghĩa thống kê, vì trị số $p<0.05$.

Nhưng `episode` là một biến liên tục. Vấn đề đặt ra là sau khi điều chỉnh cho episode thì hàm $S(t)$ cho từng nhóm sẽ ra sao? Cách khác quan nhất là giả định cả hai nhóm `gd2` và giả dược có cùng số lần `episode` (như số trung bình chẳng hạn), và hàm $S(t)$ cho từng nhóm có thể ước tính bằng:

```
> Cox.model <- survfit(coxph(Surv(time, infected==1)~episode+strata(group)))
> plot(Cox.model,
       xlab="Time",
       ylab="Cumulative survival probability",
       col=c("black", "red"))
```

hay đơn giản hơn:

```
> plot(survfit(coxph(Surv(time, infected==1)~episode+strata(group))),
  xlab="Time",
  ylab="Cumulative survival probability",
  col=c("black", "red"))
```



13.6 Xây dựng mô hình Cox bằng Bayesian Model Average (BMA)

Cũng như trường hợp của phân tích hồi qui tuyến tính đa biến và phân tích hồi qui logistic đa biến, vấn đề tìm một mô hình "tối ưu" để tiên đoán biến cố trong trong điều kiện có nhiều biến độc lập là một vấn đề nan giải. Phần lớn sách giáo khoa thống kê học

trình bày ba phương án chính để tìm một mô hình tối ưu: forward algorithm, backward algorithm, và tiêu chuẩn AIC.

Với phương án forward algorithm, chúng ta khởi đầu tìm biến độc lập *x* có ảnh hướng lớn đến biến phụ thuộc *y*, rồi từng bước thêm các biến độc lập khác *x* cho đến khi mô hình không còn cải tiến thêm nữa.

Với phương án backward algorithm, chúng ta khởi đầu bằng cách xem xét tất cả biến độc lập *x* trong dữ liệu có thể có ảnh hướng lớn đến biến phụ thuộc *y*, rồi từng bước loại bỏ từng biến độc lập *x* cho đến khi mô hình chỉ còn lại những biến có ý nghĩa thống kê.

Hai phương án trên (forward và backward algorithm) dựa vào phần dư (residual) và trị số P để xét một mô hình tối ưu. Một phương án thứ ba là dựa vào tiêu chuẩn Aikaike Information Criterion (AIC) mà tôi đã trình bày trong chương trước. Để hiểu phương pháp xây dựng mô hình dựa vào AIC tôi sẽ lấy một ví dụ thực tế như sau. Giả dụ chúng ta muốn đi từ tỉnh A đến tỉnh B qua huyện C, và mỗi tuyến đường chúng ta có 3 lựa chọn: bằng xe hơi, bằng đường thủy, và bằng xe gắn máy. Tất nhiên, đi xe hơi đắt tiền hơn đi xe gắn máy, Mặt khác, đi đường thủy tuy ít tốn kém nhưng chậm hơn đi bằng xe hơi hay xe gắn máy. Nếu có tất cả 6 phương án đi, vấn đề đặt ra là chúng ta muốn tìm một phương án đi sao cho ít tốn kém nhất, nhưng tiêu ra một thời gian ngắn nhất! Tương tự, phương pháp xây dựng mô hình dựa vào tiêu chuẩn AIC là đi tìm một mô hình sao cho ít thông số nhất nhưng có khả năng tiên đoán biến phụ thuộc đầy đủ nhất.

Nhưng cả ba phương án trên có vấn đề là mô hình "tối ưu" nhất được xem là mô hình sau cùng, và tất cả suy luận khoa học đều dựa vào ước số của mô hình đó. Trong thực tế, bất cứ mô hình nào (kể cả mô hình "tối ưu") cũng có độ bất định của nó, và khi chúng ta có thêm số liệu, mô hình tối ưu chưa chắc là mô hình sau cùng, và do đó suy luận có thể sai lầm. Một cách tốt hơn và có triển vọng hơn để xem xét đến yếu tố bất định này là Bayesian Model Average (BMA).

Với phân tích BMA, thay vì chúng ta hỏi yếu tố độc lập *x* ảnh hưởng đến biến phụ thuộc có ý nghĩa thống kê hay không, chúng ta hỏi: xác suất mà biến độc lập *x* có ảnh hưởng đến *y* là bao nhiêu. Để trả lời câu hỏi đó BMA xem xét tất cả các mô hình có khả năng giải thích *y*, và xem trong các mô hình đó, biến *x* xuất hiện bao nhiêu lần.

Ví dụ 3. Trong ví dụ sau đây, chúng ta sẽ mô phỏng một nghiên cứu với 5 biến độc lập x_1, x_2, x_3, x_4, và x_5. Ngoại trừ x_1, 4 biến kia được mô phỏng theo luật phân phối chuẩn. Biến *y* là thời gian và kèm theo biến tử vong (`death`). Trong 5 biến *x* này, chỉ có biến x_1 có liên hệ với xác suất tử vong bằng mối liên hệ `exp(3*x1 + 1)`, còn các biến x_2, x_3, x_4, và x_5 được mô phỏng toàn độc lập với nguy cơ tử vong. Chúng ta sẽ sử dụng phương pháp xây dựng mô hình theo tiêu chuẩn AIC và BMA để so sánh.

```
# Nhập package survival và BMA để phân tích
> library(survival)
> library(BMA)
```

```
# Tạo ra 5 biến số độc lập
> x1 <- (1:50)/2 - 3
> x2 <- rnorm(50)
> x3 <- rnorm(50)
> x4 <- rnorm(50)
> x5 <- rnorm(50)

# Mô phỏng mối liên hệ risk=exp(beta*x1 + 1)
> model <- exp(3*x1 + 1)

# Tạo ra biến số phụ thuộc y
> y <- rexp(50, rate = model)

# Tạo ra biến sự kiện theo luật phân phối mũ, tỉ lệ 0.3
> censored <- rexp(50, rate=0.3)
> ycencored <- pmin(y, censored)
> death <- as.numeric(y <= censored)

# Cho tất cả biến số vào data frame tên simdata
> simdata <- data.frame(y, death, x1,x2,x3,x4,x5)

# Phân tích bằng mô hình Cox
> cox <- coxph(Surv(y, death) ~ ., data=simdata)
> summary(cox)
Call:
coxph(formula = Surv(y, death) ~ ., data = simdata)

  n= 50
      coef exp(coef) se(coef)       z       p
x1  3.2325    25.344    0.568  5.6908 1.3e-08
x2 -0.0319     0.969    0.331 -0.0963 9.2e-01
x3  0.3112     1.365    0.327  0.9518 3.4e-01
x4  0.1364     1.146    0.297  0.4600 6.5e-01
x5  0.4898     1.632    0.313  1.5643 1.2e-01

   exp(coef) exp(-coef) lower .95 upper .95
x1    25.344     0.0395     8.325     77.16
x2     0.969     1.0324     0.506      1.85
x3     1.365     0.7326     0.719      2.59
x4     1.146     0.8725     0.641      2.05
x5     1.632     0.6127     0.883      3.01

Rsquare= 0.992   (max possible= 0.997 )
Likelihood ratio test= 241  on 5 df,   p=0
Wald test            = 33.3  on 5 df,   p=3.36e-06
Score (logrank) test = 107  on 5 df,   p=0
```

Kết quả trên cho thấy biến x1,x3 và x5 có ảnh hưởng có ý nghĩa thống kê đến biến y. Tất nhiên, đây làm một kết quả sai vì chúng ta biết rằng chỉ có x1 là có ý nghĩa thống kê mà thôi. Bây giờ chúng ta thử áp dụng cách xây dựng mô hình dựa vào tiêu chuẩn AIC:

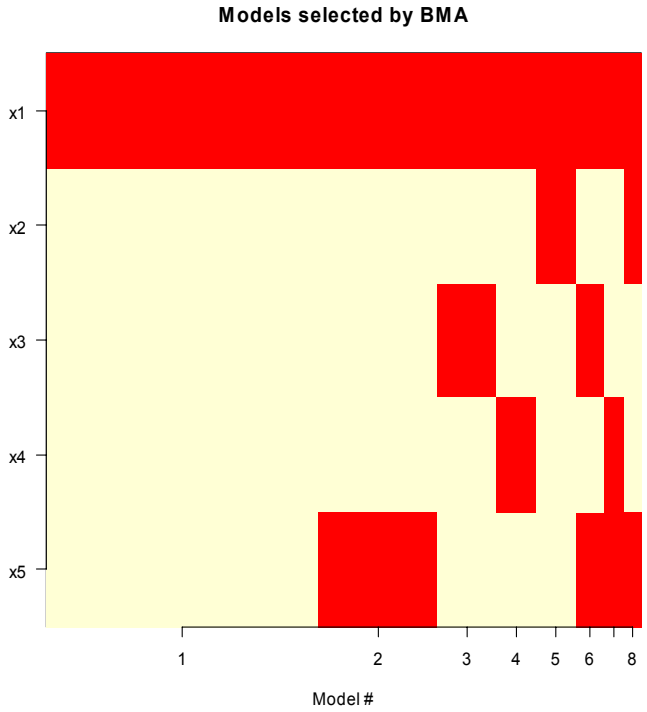
```
# Tìm mô hình dựa vào tiêu chuẩn AIC
> searchAIC <- step(cox, direction="both")
> summary(searchAIC)
Call:
coxph(formula = Surv(y, death) ~ x1 + x5, data = simdata)
```

```
  n= 50
    coef exp(coef) se(coef)    z       p
x1 3.126     22.79    0.529 5.91 3.4e-09
x5 0.429      1.54    0.297 1.45 1.5e-01

   exp(coef) exp(-coef) lower .95 upper .95
x1     22.79     0.0439     8.080     64.27
x5      1.54     0.6510     0.858      2.75

Rsquare= 0.992   (max possible= 0.997 )
Likelihood ratio test= 240  on 2 df,   p=0
Wald test            = 35.3  on 2 df,   p=2.18e-08
Score (logrank) test = 104  on 2 df,   p=0
```

Kết quả này cho thấy `x1` và `x5` là hai yếu tố độc lập có ảnh hưởng có ý nghĩa thống kê đến biến `y`. Một lần nữa, kết quả này sai! Bây giờ chúng ta sẽ áp dụng phép tính BMA:

```
#tìm mô hình bằng phép tính BMA
> time <- simdata$y
> death <- simdata$death
> xvars <- simdata[,c(3,4,5,6,7)]
> bma <- bic.surv(xvars, time, death)
> summary(bma)
> imageplot.bma(bma)

Call:
bic.surv.data.frame(x = xvars, surv.t = time, cens = death)


  8  models were selected
 Best  5  models (cumulative posterior probability =  0.8911 ):

      p!=0    EV    SD     model 1   model 2   model 3   model 4    model 5
x1    100.0  3.0360 0.509  2.98048   3.12625   3.03900   2.98288    2.98098
x2      9.6  0.0008 0.096   .         .         .         .         0.02136
x3     14.6  0.0410 0.155   .         .        0.27046    .          .
x4     10.0  0.0063 0.092   .         .         .        0.02497     .
x5     31.0  0.1349 0.261   .       0.42920     .         .          .

nVar                        1         2         2         2          2
BIC                  -233.774  -232.126  -230.713  -229.933  -229.930
post prob               0.458     0.201     0.099     0.067     0.067
```

Kết quả phân tích BMA cho thấy mô hình tối ưu là mô hình 1 chỉ có một biến có ý nghĩa thống kê: đó là biến `x1`. Xác suất mà yếu tố này có ảnh hưởng đến nguy cơ tử vong là 100%. Đây chính là kết quả mà chúng ta kì vọng, bởi vì chúng ta đã mô phỏng chỉ có `x1` có ảnh hưởng đến `y` mà thôi. Mô hình 2 có hai biến `x1` và `x5` (tức cũng chính là mô hình mà tiêu chuẩn AIC xác định), nhưng mô hình này chỉ có xác suất 0.201 mà thôi. Các mô hình 3(`x1` và `x3`), mô hình 4 (`x1` và `x4`) và mô hình 5 (`x1` và `x2`) cũng có khả năng nhưng xác suất quá thấp (dưới 0.1) cho nên chúng ta không thể chấp nhận được. Biểu đồ sau đây thể hiện các kết quả trên:



Biểu đồ trên trình bày 8 mô hình, và trong tất cả 8 mô hình, biến `x1` xuất hiện một các nhất quán (xác suất 100%). Còn các biến khác có ảnh hưởng nhưng không nhất quán. Qua so sánh giữa hai phương pháp xây dựng mô hình rõ ràng cho thấy cách phân tích BMA cung cấp cho chúng ta mô hình phù hợp đáng tin cậy nhất và có vẻ phù hợp với thực tế nhất.

Trên đây là những phương pháp phân tích biến cố thông dụng nhất trong khoa học thực nghiệm với mô hình Cox và kiểm định log-rank. Mô hình Cox có thể khai triển thành những mô hình phức tạp và tinh vi hơn cho các nghiên cứu phức tạp khác với nhiều biến và tương tác giữa các yếu tố nguy cơ. Tài liệu hướng dẫn cách sử dụng package `survival` có thể giúp bạn đọc tìm hiểu sâu hơn. Tài liệu này có tại trang web www.cran.R-project.org.