

医学情報処理演習第13回「生存時間解析」^{*1}

2007年1月22日 中澤 港 (nminato@med.gunma-u.ac.jp)

演習サポート web ページ : <http://phi.med.gunma-u.ac.jp/medstat/>

生存時間解析概論

生存時間解析の特徴は、期間データを扱うことと、観察打ち切りケースが含まれるデータを扱えることである。

実験においては、化学物質などへの1回の曝露の影響を時間を追ってみていくことが良く行われる。時間ごとに何らかの量の変化を追うほかに、エンドポイント（観察期間の終点となるイベント）を死亡とした場合、死ぬまでの時間を分析することで毒性の強さを評価することができる。しかし、観察時間には限りがあるので観察中には死亡イベントが起こらないケースもある。そのような「観察打ち切り」ケースは相対的に長生きしている可能性が高いので、そのレコードを単純にデータから除去してしまうと全体の生存時間を過小評価してしまって不都合である。このような期間データを扱うには、一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれる分析法を用いる。

なかでもよく知られている指標が Kaplan-Meier の積・極限推定量である。現在では、普通、カプラン=マイヤ推定量と呼ばれている。イベントが起こった各時点での、イベントが起こる可能性がある人口（リスク集合^{*2}）あたりのイベント発生数を1から引いたものを掛け合わせて得られる関数値である。カプラン=マイヤ推定量が0.5を横切る時点は、イベントを経験せずにいる人数が全体の半分になるまでの時間を意味する。言い換えると、この点は、生存時間の中央値の、ノンパラメトリックな最尤推定量である。通常、推定と同時に階段状の生存曲線を描き、観察打ち切りレコードは曲線上に縦棒で表す。

コラム：生命表解析

データ数が多い場合は、個々の間隔データを集計して、生命表解析を行うこともある。生命表解析の代表的なものは、ヒトの平均寿命を計算するときに行われている（官庁統計としても、まさしく生命表という形で発表されている）。平均寿命とは0歳平均余命のことだが、これは、ある時点での年齢別死亡率に従って、ゼロ歳児10万人が死んでいったとすると、生まれてから平均してどれくらいの期間生存するのかという値である^a。

一般に x 歳平均余命は、 x 歳以降の延べ生存期間の総和 (T_x) を x 歳時点の個体数 (l_x) で割れば得られる。延べ生存期間の総和は、年齢別死亡率 q_x が変化しないとして、 $l_x(1 - q_x/2)$ によって x 歳から $x+1$ 歳まで生きた人口 L_x （開始時点の人口が決まっていれば死亡率も変化しないので x 歳の静止人口と呼ばれる）を求め、それを x 歳以降の全年齢について計算して和をとることで得られる。

ヒトの人口学では年齢別死亡率から q_x を求めて生命表を計算するのが普通だが、生物一般について考えるときは、同時に生まれた複数個体（コホート）を追跡して年齢別生存数として l_x を直接求めてしまう方法（コホート生命表）とか、たんに年齢別個体数を l_x と見なしてしまう方法（静態生命表、偶然変動で高齢の個体数の方が多い場合があるので平滑化するのが普通）がよく行われる。

^a 誤解されることが多いが、死亡年齢の平均ではないので注意されたい。

一方、複数の期間データ列の差の比較、すなわち生存時間の差の検定には、ログランク検定や一般化ウィルコクソン検定が使われる。

それらのノンパラメトリックな方法とは別に、イベントが起こるまでの時間が何らかのパラメトリックな分布に当てはまるかどうかを調べる方法もある。当てはめる分布としては指数分布やワイブル分布がある。イベ

^{*1} 本資料は <http://phi.med.gunma-u.ac.jp/medstat/it13-2006.pdf> としてダウンロード可能である。なお、講義時に配布したのから、かなり加筆修正してある。

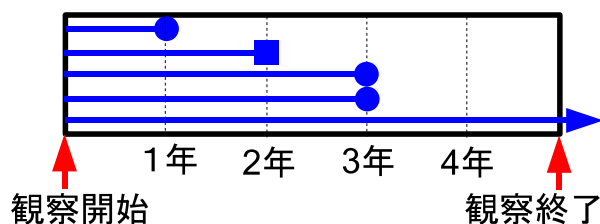
^{*2} population at risk の訳語として、大橋・浜田 (1995) に従っておく。

ントが起こるまでの期間に何らかの別の要因が与える効果を調べたいときはコックス回帰（それらが基準となる個体のハザードに対して $\exp(\sum \beta_i z_i)$ という比例定数の形で掛かるとする比例ハザード性を仮定する方法）と、パラメトリックなモデルに対数線形モデルの独立変数項として入れてしまう加速モデルがある。加速モデルは `survreg()` 関数で実行できるが、ここでは詳細には立ち入らない。

R では生存時間解析をするための関数は `survival` パッケージで提供されており、`library(survival)` または `require(survival)` とすれば使えるようになる^{*3}。

カプラン = マイヤ法：`survfit()` 関数

カプラン = マイヤ推定量について、まず一般論を示しておく。イベントが起こる可能性がある状態になってから、イベントが起こった時点をも t_1, t_2, \dots とし、 t_1 時点でのイベント発生数を d_1 、 t_2 時点でのイベント発生数を d_2 、以下同様であるとする。また、時点 t_1, t_2, \dots の直前でのリスク集合の大きさを n_1, n_2, \dots で示す。リスク集合の大きさは、その直前でまだイベントが起きていない個体数である。観察途中で転居などによって打ち切りが生じるために、リスク集合の大きさはイベント発生によってだけでなく、打ち切りによっても減少する。従って n_i は、時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる。例えば、下図のように 5 人のがん患者をフォローアップしていて、ちょうど 1 年で 1 人亡くなり、ちょうど 2 年で 1 人転出したためフォローアップ不能になり、ちょうど 3 年で 2 人が同時に死亡し、5 年目まで継続観察してまだ最後の 1 人は生き残っていたとすると、 t_1 が 1 年で t_2 が 3 年となり、 $d_1 = 1$ 、 $d_2 = 2$ 、 $n_1 = 5$ 、 $n_2 = 3$ となる。



なお、イベント発生と打ち切りが同時点で起きている場合は、打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。このとき、カプラン = マイヤ推定量 $\hat{S}(t)$ は、

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)\dots = \prod_{i < t} (1 - d_i/n_i)$$

として得られる。その標準誤差はグリーンウッドの公式により、

$$\text{var}(\hat{S}) = \hat{S}^2 \times \sum_{i < t} \frac{d_i}{n_i(n_i - d_i)}$$

で得られる。なお、カプラン = マイヤ推定では、階段状のプロットを同時に行うのが普通である。

R では、`library(survival)` または `require(survival)` としてパッケージを呼び出し、`Surv`(生存時間, 打ち切りフラグ) 関数で生存時間データを作る。打ち切りフラグは通常、1 でイベント発生、0 が打ち切り、つ

^{*3} `survival` は起動時にロードされていないが Recommended なライブラリなので、Windows 版では R 本体をインストールするだけで同時にインストールされる。ちなみに、R パイナリに built-in なライブラリは、`base`, `datasets`, `grDevices`, `graphics`, `grid`, `methods`, `splines`, `stats`, `stats4`, `tcltk`, `tools`, `utils` であり、Recommended なライブラリで、将来全パイナリに built-in される予定なのは、`survival` の他には、`KernSmooth`, `MASS`, `boot`, `class`, `cluster`, `foreign`, `lattice`, `mgcv`, `nlme`, `nnet`, `rpart`, `spatial` がある。`search()` でロード済みライブラリ一覧、`.packages(all.avail=T)` でインストール済み一覧が表示される。ロード済みのライブラリをアンロードするには `detach(package=survival)` などとする。

まり観察期間終了時に生存していることを示すが，TRUE がイベント発生，FALSE が生存としてもいいし，1 でイベント発生，2 で打ち切りとすることもできる。また，区間打ち切りレコード^{*4}を扱うときは，Surv(最終観察時点，観察不能になった時点，打ち切りフラグ)とする。打ち切りフラグは，0 が右側打ち切り，1 がちょうどイベント発生，2 が左側打ち切り，3 が区間打ち切りを示す。打ち切りレコードがないときは，打ち切りフラグそのものを省略することもできる。例えば，1 人目については 81 ヶ月より後で 92 ヶ月より前にイベントが発生した区間打ち切り，2 人目については 22 ヶ月でイベント発生，3 人目については 29 ヶ月まで観察して，まだイベントが発生していないという場合，次の枠内のように生存時間データ dat を定義する。

```
time <- c(81,22,29)
time2 <- c(92,22,NA)
event <- c(3,1,0)
dat <- Surv(time,time2,event,type="interval")
```

生存時間データを dat に付値したとすると，res <- survfit(dat) で Kaplan-Meier 法によるメディアン生存時間が得られ，plot(res) とすれば階段状の生存関数が描かれる。イベント発生時点ごとの値を見るには，summary(res) とすればよい。

コラム：日時を扱う関数

生データとして生存時間が与えられず，観察開始とイベント発生の日付を示している場合，それらの間隔として生存時間を計算するには，difftime() 関数や ISOdate() 関数を使うと便利である。例えば，下枠内では，まず x というデータフレームに変数 names (名前)，dob (誕生年月日) と dod (死亡年月日) が付値している。次に difftime() 関数で 4 人分の死亡年月日と誕生年月日の差 (= 生存日数) を計算し，[x\$names=="Robert"] で Robert (これは言うまでもなくロベルト・コッホのことである) についての生存日数が得られ，それが alivedays に付値される。その次の行のように 365.24 で割れば，生存年数に換算される。日数の与え方は，日本語ロケールの場合，ダブルクォーテーションマークで括って，年，月，日がハイフンまたはスラッシュでつながれた形で与えることもできるし (それが日付であることを明示するには as.Date() に渡せばよい)，その次の行のように ISOdate(年，月，日) という形で与えることもできる。ちなみにこの行は，この 4 人がもし現在生きていたら満何歳になるかという計算結果を与える。

```
it13-1-2006.R
x <- data.frame(
  names = c("Edward", "Shibasaburo", "Robert", "Hideyo"),
  dob = c("1749-5-17", "1853-1-29", "1843-12-11", "1876-11-9"),
  dod = c("1823-1-26", "1931-6-13", "1910-5-27", "1928-5-21"))
alivedays <- difftime(x$dod, x$dob)[x$names=="Robert"]
as.numeric(alivedays/365.24)
as.numeric(difftime(ISOdate(2007, 1, 22), x$dob)/365.24)
```

^{*4} ある間隔の中で観察打ち切りあるいはイベント発生したことはわかっているが，正確なその時点が不明のケースは，区間打ち切りレコードとして扱う。例えば，ある年に観察して生存していた人が，翌年行ったら転居してしまっていて追跡不能な場合，最後の観察時点と最初に観察不能になってしまった時点の間のどこかであることは既知だが，その 1 年のどこで打ち切りになったか不明なので，区間打ち切りレコードとなる。

例題 1

大橋・浜田 (1995) の p.60-61 に掲載されている Gehan の白血病治療データは、42 人の白血病患者を 6-MP とプラセボを投与するペアにランダムに割り付けて治療し、寛解が続いている週数である (イベントは白血病の再発)。R では MASS ライブラリに含まれている。これは有名なデータで、<http://data.princeton.edu/wws509/datasets/gehan.dat> としてインターネット上にも公開されている。このデータを用い、2 群別々にカプラン=マイヤ推定をせよ。

プログラムは次の枠内の通り。

it13-2-2006.R

```
require(MASS)
require(survival)
print(res<-survfit(Surv(time,cens)~treat,data=gehan))
par(family="sans",las=1)
plot(res,lty=c(1,2),main="Gehan のデータについてのカプラン=マイヤプロット")
legend(30,0.2,lty=c(1,2),legend=levels(gehan$treat))
summary(res)
```

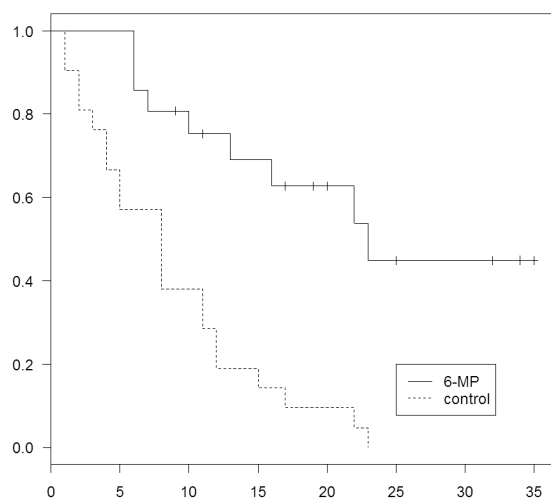
3 行目により以下が表示される。この結果だけで、2 群別々のメディアン生存時間 (カプラン=マイヤ推定量) とその 95%信頼区間はわかる。

```
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)
```

	n	events	median	0.95LCL	0.95UCL
treat=6-MP	21	9	23	16	Inf
treat=control	21	21	8	4	12

4-6 行目により以下のグラフが描かれる。カプラン=マイヤ法でメディアン生存時間の推定をした場合、このような生存曲線を描くのが普通である。縦棒は観察打ち切りの生じた時点を示す。

Gehanのデータについてのカプラン=マイヤプロット



7 行目の `summary(res)` によって、次の枠内の通り、すべてのイベント発生時点における生残確率と、その 95%信頼区間が表示される (生存曲線はこの結果を使って描かれる)。

```
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)
```

```
treat=6-MP
time n.risk n.event survival std.err lower 95% CI upper 95% CI
 6    21     3   0.857  0.0764    0.720    1.000
 7    17     1   0.807  0.0869    0.653    0.996
10    15     1   0.753  0.0963    0.586    0.968
13    12     1   0.690  0.1068    0.510    0.935
16    11     1   0.627  0.1141    0.439    0.896
22     7     1   0.538  0.1282    0.337    0.858
23     6     1   0.448  0.1346    0.249    0.807
```

```
treat=control
time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1    21     2   0.9048  0.0641    0.78754    1.000
 2    19     2   0.8095  0.0857    0.65785    0.996
 3    17     1   0.7619  0.0929    0.59988    0.968
 4    16     2   0.6667  0.1029    0.49268    0.902
 5    14     2   0.5714  0.1080    0.39455    0.828
 8    12     4   0.3810  0.1060    0.22085    0.657
11     8     2   0.2857  0.0986    0.14529    0.562
12     6     2   0.1905  0.0857    0.07887    0.460
15     4     1   0.1429  0.0764    0.05011    0.407
17     3     1   0.0952  0.0641    0.02549    0.356
22     2     1   0.0476  0.0465    0.00703    0.322
23     1     1   0.0000    NA          NA          NA
```

例題 2

survival ライブラリに含まれているデータ aml は、急性骨髄性白血病 (acute myelogenous leukemia) 患者が化学療法によって寛解した後、ランダムに 2 群に分けられ、1 群は維持化学療法を受け (維持群)、もう 1 群は維持化学療法を受けずに (非維持群)、経過観察を続けて、維持化学療法が生存時間を延ばすかどうかを調べたデータである^a。以下の 3 つの変数が含まれている。

time 生存時間あるいは観察打ち切りまでの時間 (週)

status 打ち切り情報 (0 が観察打ち切り, 1 がイベント発生)

x 維持化学療法が行われたかどうか (Maintained が維持群, Nonmaintained が非維持群)

薬物維持化学療法の維持群と非維持群で別々に、生存時間の中央値を Kaplan-Meier 推定し、生存曲線をプロットせよ。

^a 出典: Miller RG: Survival Analysis. John Wiley and Sons, 1981. 元々は, Embury SH, Elias L, Heller PH, Hood CE, Greenberg PL, Schrier SL: Remission maintenance therapy in acute myelogenous leukaemia. *Western Journal of Medicine*, 126, 267-272, 1977. のデータ。研究デザインは Gehan と似ている。

it13-3-2006.R

```
require(survival)
print(res <- survfit(Surv(time,status)~x, data=aml))
summary(res)
par(family="sans",las=1)
plot(res,lty=c(1,2),main="急性骨髄性白血病の維持化学療法の有無別 Kaplan-Meier 推定")
legend(100,0.8,lty=c(1,2),legend=c("維持群","非維持群"))
```

上枠内のように入力すると、2行目でカプラン=マイヤ法での計算がなされ、下枠内が表示される。

	n	events	median	0.95LCL	0.95UCL
x=Maintained	11	7	31	18	Inf
x=Nonmaintained	12	11	23	8	Inf

この表は、維持群が11人、非維持群が12人、そのうち死亡まで観察された人がそれぞれ7人と11人いて、維持群の生存時間の中央値が31週、非維持群の生存時間の中央値が23週で、95%信頼区間の下限はそれぞれ18週と8週、上限はどちらも無限大であると読む。4行目を入力すると、死亡が観察された7人と11人について、それぞれの死亡時点までの生存確率が標準誤差と95%信頼区間とともに表示される。最後の3行で、例題1と同様に生存曲線が描かれる。

ログランク検定：survdif() 関数

次に、ログランク検定を簡単な例で説明する。

8匹のラットを4匹ずつ2群に分け、第1群には毒物Aを投与し、第2群には毒物Bを投与して、生存期間を追跡したときに、第1群のラットが4,6,8,9日目に死亡し、第2群のラットが5,7,12,14日目に死亡したとする。この場合、観察期間内にすべてのラットが死亡し、正確な生存時間がわかっているため、観察打ち切りがないデータとなっていて計算しやすい。

ログランク検定の思想は、大雑把に言えば、死亡イベントが起こったすべての時点で、群と生存/死亡個体数の2×2クロス集計表を作り、それをコクラン=マンテル=ヘンツェル流のやり方で併合するということである。

このラットの例では、死亡イベントが起こった時点1~8において各群の期待死亡数を計算し、各群の実際の死亡数との差をとって、それに時点の重みを掛けたものを、各時点における各群のスコアとして、群ごとのスコアの合計を求める。2群しかないため、各時点において群1と群2のスコアの絶対値は同じで符号が反対になる。2群の生存時間に差がないという帰無仮説を検定するためには、群1の合計スコアの2乗を分散で割った値をカイ二乗統計量とし、帰無仮説の下でこれが自由度1のカイ二乗分布に従うことを使って検定する。

なお、重みについては、ログランク検定ではすべて1である。一般化ウィルコクソン検定では、重みを、2群を合わせたリスク集合の大きさとする（そうした場合、もし打ち切りがなければ、検定結果は、ウィルコクソンの順位和検定の結果と一致する）。つまり、ログランク検定でも一般化ウィルコクソン検定でも、実は期間の情報はまったく使われず、死亡順位の情報だけが使われているのである。

記号で書けば次の通りである。第*i*時点の第*j*群の期待死亡数 e_{ij} は、時点*i*における死亡数の合計を d_i 、時点*i*における*j*群のリスク集合の大きさを n_{ij} 、時点*i*における全体のリスク集合の大きさを n_i とすると、

$$e_{ij} = d_i \cdot n_{ij} / n_i$$

と表される^{*5}。上の例では、 $e_{11} = 1 \cdot n_{11} / n_1 = 4/8 = 0.5$ となる。時点*i*における第*j*群の死亡数を d_{ij} 、時点の重みを w_i と表せば、時点*i*における群*j*のスコア u_{ij} は、

$$u_{ij} = w_i (d_{ij} - e_{ij})$$

となり、ログランク検定の場合（以下、重みは省略してログランク検定の場合のみ示す）の群1の合計スコ

^{*5} 打ち切りデータは、リスク集合の大きさが変わることを通してのみ計算に寄与する。打ち切り時点ではスコアは計算されないことに注意しよう。

アは

$$u_1 = \sum_i (d_{i1} - e_{i1})$$

となる。上の例では、

$$u_1 = (1 - 4/8) + (0 - 3/7) + (1 - 3/6) + (0 - 2/5) + (1 - 2/4) + (1 - 1/3) + (0 - 0/2) + (0 - 0/1)$$

である。これを計算すると約 1.338 となる。分散は、分散共分散行列の対角成分を考えればいいので、

$$V = V_{jj} = \sum_i \frac{(n_i - n_{ij})n_{ij}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

となる。この例の数値を当てはめると、

$$V = \frac{(8-4) \times 4}{8^2} + \frac{(7-3) \times 3}{7^2} + \frac{(6-3) \times 3}{6^2} + \frac{(5-2) \times 2}{5^2} + \frac{(4-2) \times 2}{4^2} + \frac{(3-1) \times 1}{3^2}$$

となり、 $4 \times 4/64 + 4 \times 3/49 + 3 \times 3/36 + 3 \times 2/25 + 2 \times 2/16 + 2 \times 1/9$ で計算すると、約 1.457 となる。したがって、 $\chi^2 = 1.338^2/1.457 = 1.23$ となり、この値は自由度 1 のカイ二乗分布の 95%点である 3.84 よりずっと小さいので、有意水準 5%で帰無仮説は棄却されない。つまりこれだけのデータでは、差があるとはいえないことになる（もちろん、サンプルサイズを大きくすれば違う結果になる可能性もある）。

R でログランク検定を実行するには、観察時間を示す変数を time、打ち切りフラグを event、グループを group とすれば、`survdif(Surv(time,event)~group)` とすればよい。この例の場合なら、下枠内の通り。なお、一般化ウィルコクソン検定をするには `survdif(Surv(time,event)~group, rho=1)` とすればよい。

it13-4-2006.R

```
require(survival)
time <- c(4,6,8,9,5,7,12,14)
event <- c(1,1,1,1,1,1,1,1)
group <- c(1,1,1,1,2,2,2,2)
survdif(Surv(time,event)~group)
```

出力結果を見ると、 $\chi^2 = 1.2$ 、自由度 1、 $p = 0.268$ となっているので、有意水準 5%で、2 群には差がないことがわかる。なお、ログランク検定だけでなく、カプラン=マイヤ法により生存時間の中央値と生存曲線の図示もするのが普通である。

例題 3

例題 2 のデータで、維持群と非維持群の間に生存時間の有意差はあるか、有意水準 5%でログランク検定せよ。

`survdif(Surv(time,status)~x,data=aml)` と打つだけである。カイ二乗値が 3.4 で、有意確率が 0.0653 と計算される。したがって、有意水準 5%で帰無仮説は棄却されず、ログランク検定では、維持群と非維持群の間の生存時間の差は有意ではないといえる。なお、このデータは打ち切りを含んでいるが、R で計算する限りにおいては、そのことを意識する必要はない（適切に処理してくれる）。

コックス回帰—比例ハザードモデル：coxph() 関数

カプラン=マイヤ推定やログランク検定は、まったく母数の分布を仮定しない方法だった。コックス回帰は、「比例ハザード性」を仮定する。そのため、比例ハザードモデルとも呼ばれる。

コックス回帰の基本的な考え方は、イベント発生に影響する共変量ベクトル $z_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ をもつ個体 i の、時点 t における瞬間イベント発生率 $h(z_i, t)$ （これをハザード関数と呼ぶ）として、

$$h(z_i, t) = h_0(t) \cdot \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \dots + \beta_p z_{ip})$$

を想定するものである。 $h_0(t)$ は基準ハザード関数と呼ばれ、すべての共変量のイベント発生への影響がゼロである「基準人」の、時点 t における瞬間死亡率を意味する。 $\beta_1, \beta_2, \dots, \beta_p$ が推定すべき未知パラメータであり、共変量が $\exp(\beta_x z_{ix})$ という比例定数の形でイベント発生に影響するので、このことを「比例ハザード性」と呼ぶ。なお、Cox が立てたオリジナルのモデルでは、 z_i が時間とともに変わる、時間依存性共変量の場合も考慮されていたが、現在、通常行われるコックス回帰では、共変量の影響は時間に依存しないもの（時間が経過しても増えたり減ったりせず一定）として扱う。

そのため、個体間のハザード比は時点によらず一定になるという特徴をもつ。つまり、個体 1 と個体 2 で時点 t のハザードの比をとると基準ハザード関数 $h_0(t)$ が分母分子からキャンセルされるので、ハザード比は常に、

$$\frac{\exp(\beta_1 z_{11} + \beta_2 z_{12} + \dots + \beta_p z_{1p})}{\exp(\beta_1 z_{21} + \beta_2 z_{22} + \dots + \beta_p z_{2p})}$$

となる。このため、比例ハザード性を仮定できれば、基準ハザード関数の形について（つまり、生存時間分布について）特定のパラメトリックモデルを仮定する必要がなくなる。この意味で、比例ハザードモデルはセミパラメトリックであるといわれる。

二重対数プロット

ここで生存関数とハザード関数の関係について整理しておこう。まず、 T をイベント発生までの時間を表す非負の確率変数とする。生存関数 $S(t)$ は、 $T \geq t$ となる確率である。 $S(0) = 1$ となることは定義より自明である。ハザード関数 $h(t)$ は、ある瞬間 t にイベントが発生する確率なので、

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} = -\frac{dS(t)}{dt} \frac{1}{S(t)} = -\frac{d(\log(S(t)))}{dt}$$

である。累積ハザード関数は、 $H(t) = \int_0^t h(u) du = -\log S(t)$ となる。これを式変形すると、 $S(t) = \exp(-H(t))$ とも書ける。

そこで、共変量ベクトルが z である個体の生存関数を $S(z, t)$ 、累積ハザード関数を $H(z, t)$ とすれば、

$$H(z, t) = \int_0^t h(z, u) du = \int_0^t h_0(u) \exp(\beta z) du = \exp(\beta z) H_0(t)$$

$$S(z, t) = \exp(-H(z, t)) = \exp\{-\exp(\beta z) H_0(t)\}$$

となる。したがって、比例ハザード性が成立していれば、

$$\log(-\log S(z, t)) = \beta z + \log H_0(t)$$

が成り立つことになるので、共変量で層別して、横軸に生存期間をとり、縦軸に生存関数の対数の符号を逆にしてもう一度対数をとった値をとって散布図を描くと、層間で βz だけ平行移動したグラフが描かれることになる。これを二重対数プロットと呼ぶ。

逆に考えれば、二重対数プロットを描いてみて、層ごとの散布図が平行になっていなければ、「比例ハザード性」の仮定が満たされないので、コックス回帰をするのは不適切といえる。

コックス回帰のパラメータ推定

パラメータ β の推定には、部分尤度という考え方が用いられる。時点 t において個体 i にイベントが発生する確率を、時点 t においてイベントが 1 件起こる確率と、時点 t でイベントが起きたという条件付きでそれが個体 i である確率の積に分解すると、前者は生存時間分布についてパラメトリックなモデルを仮定しないと不明だが、後者はその時点でのリスク集合内の個体のハザードの総和を分母として、個体 i のハザードを分子として推定できる。すべてのイベント発生について、後者の確率だけをかけあわせた結果を L とおくと、 L は、全体の尤度から時点に関する尤度を除いたものになり、その意味で部分尤度とか偏尤度と呼ばれる。

サンプルサイズを大きくすると真の値に収束し、分布が正規分布で近似でき、分散もその推定量としては最小になるという意味での、「良い」推定量として、パラメータ β を推定するには、この部分尤度 L を最大にするようなパラメータを得ればよいことを Cox が予想したので（後にマルチンゲール理論によって証明された）、比例ハザードモデルをコックス回帰という^{*6}。R でのコックス回帰の基本は、`coxph(Surv(time, cens) ~ grp+covar, data=dat)` という形になる。

例題 4

例題 2 のデータで維持の有無が生存時間に与える影響をコックス回帰せよ。

it13-5-2006.R

```
require(survival)
summary(res <- coxph(Surv(time,status)~x,data=aml))
loglogplot <- function(X) {
  S <- X$surv
  T <- X$time
  G <- X$ntimes.strata
  GG <- names(X$strata)
  GX <- rep(GG,G)
  xr <- c(0,max(T)*1.5)
  mas <- ifelse(max(S)==1,0.99,max(S))
  mis <- ifelse(min(S)==0,0.01,min(S))
  yr <- c(log(-log(mas)),log(-log(mis)))
  plot(T[GX==GG[1]],log(-log(S[GX==GG[1]])),type="l",lty=1,xlim=xr,ylim=yr,
       xlab="time",ylab="log(-log(S))",main="二重対数プロット")
  for (i in 2:length(GG)) {
    lines(T[GX==GG[i]],log(-log(S[GX==GG[i]])),lty=i)
  }
  legend(max(T),-2,legend=GG,lty=1:length(GG))
}
KM <- survfit(Surv(time,status)~x,data=aml)
par(family="sans",las=1,mfrow=c(1,2))
plot(KM,main="aml データのカプラン = マイヤプロット")
loglogplot(KM)
```

2 行目で得られる結果は、下枠内の通りである。

^{*6} なお、同時に発生したイベントが 2 つ以上ある場合は、その扱い方によって、Exact 法とか、Breslow の方法、Efron の方法、離散法などがあるが、可能な場合は Exact 法を常に使うべきである。また、離散法は、離散ロジスティックモデルに対応する推定法となっていて、生存時間が連続量でなく、離散的にしか得られていない場合に適切である。Breslow 法を使うパッケージが多いが、R の `coxph()` 関数のデフォルトは Efron 法である。Breslow 法よりも Efron 法の方が Exact 法に近い結果となる。

```

Call:
coxph(formula = Surv(time, status) ~ x, data = aml)

n= 23

      coef exp(coef) se(coef)      z      p
xNonmaintained 0.916      2.5    0.512 1.79 0.074

      exp(coef) exp(-coef) lower .95 upper .95
xNonmaintained      2.5      0.4    0.916    6.81

Rsquare= 0.137 (max possible= 0.976 )
Likelihood ratio test= 3.38 on 1 df,  p=0.0658
Wald test              = 3.2 on 1 df,  p=0.0737
Score (logrank) test = 3.42 on 1 df,  p=0.0645

```

有意水準 5%で「維持化学療法の有無が生存時間に与えた効果がない」という帰無仮説は棄却されない*7。従って差はないと解釈される。exp(coef) の値 2.5 が、2 群間のハザード比の推定値になるので、維持群に比べて非維持群では 2.5 倍死亡ハザードが高いと考えられるが、95%信頼区間が 1 を挟んでいるので、有意水準 5%では有意でない。

3 行目以降により、左に 2 群別々に推定した Kaplan-Meier プロットが描かれ*8、右に二重対数プロットがなされる。

例題 5

例題 1 にあげた Gehan の白血病治療データで、対照群に対する 6-MP 処置群のハザード比を推定せよ。

プログラムは次の枠内の通りである。

```

it13-6-2006.R
require(MASS)
require(survival)
res <- coxph(Surv(time,cens)~treat,data=gehan)
summary(res)
plot(survfit(res))

```

summary(res) により、次の枠内の結果が出力される。plot(survfit(res)) では、コックス回帰で推定されるベースラインの生存曲線が 95%信頼区間つきで描かれる。

*7 最終行に Score (logrank) test とあるが、これは Rao の Score 検定の結果である。survdif() により実行されるログランク検定の結果とは微妙に異なる。

*8 コックス回帰の場合は、通常、群の違いは比例ハザード性を前提として 1 つのパラメータに集約させ、生存関数の推定には 2 つの群の情報を両方使い、共変量の影響も調整して推定したベースラインの生存曲線を 95%信頼区間つきで描かせるため、plot(survfit(coxph(Surv(time,cens)~treat+pair,data=gehan))) のようにするのだが、推定値は共変量の影響を無視して plot(survfit(Surv(time,cens),data=gehan)) とすると、あまり大きくは変わらないことが多い。共変量でコックス回帰したベースラインの生存曲線を 2 群別々に描きたい場合は、coxph() 関数の中で、subset=(treat=="6-MP") のように指定することによって、群ごとにパラメータ推定をさせることができるが、この場合は独立変数に群別変数を入れてはいけない。2 つ目のグラフをプロットするときは par(new=T) をしてから色を変えてプロットすればいいのだが、信頼区間まで重ね描きされると見にくいので、あまりお勧めしない。他の共変量はとりあえず無視して Kaplan-Meier プロットするので通常は十分であろう。

```

Call:
coxph(formula = Surv(time, cens) ~ treat, data = gehan)

n= 42

      coef exp(coef) se(coef)      z      p
treatcontrol 1.57      4.82    0.412 3.81 0.00014

      exp(coef) exp(-coef) lower .95 upper .95
treatcontrol    4.82      0.208    2.15    10.8

Rsquare= 0.322 (max possible= 0.988 )
Likelihood ratio test= 16.4 on 1 df,  p=5.26e-05
Wald test              = 14.5 on 1 df,  p=0.000138
Score (logrank) test = 17.3 on 1 df,  p=3.28e-05

```

どの検定結果をみても有意水準 5%で「6-MP 処置が死亡ハザードに与えた効果がない」という帰無仮説は棄却される。exp(coef) の値 4.82 が、2 群間のハザード比の推定値になるので、6-MP 処置群に比べて対照群では 4.82 倍 (95%信頼区間が [2.15, 10.8]) 死亡ハザードが高いと考えられ、6-MP 処置は有意な延命効果をもつと解釈できる。

コックス回帰における共変量の扱い

コックス回帰で、共変量の影響をコントロールできることの意味をもう少し説明しておく。例えば、がんの生存時間を分析するとき、進行度のステージ別の影響は無視できないけれども、これを調整するには、大別して3つの戦略がありうる。

1. ステージごとに別々に分析する。
2. 他の共変量の影響はステージを通じて共通として、ステージを層別因子として分析する
3. ステージも共変量としてモデルに取り込む

3 番目の仮定ができれば、ステージも共変量としてイベント発生への影響を定量的に評価できるメリットがあるが、そのためには、ステージが違っててもベースラインハザード関数が同じでなければならず、やや非現実的である。また、ステージをどのように共変量としてコード化するかによって結果が変わってくる (通常はダミー変数化することが多い)。2 番目の仮定は、ステージによってベースラインハザード関数が異なることを意味する。R の coxph() 関数で、層によって異なるベースラインハザードを想定したい場合は、strata() を使ってモデルを指定する。例えば、この場合のように、がんの生存時間データで、生存時間の変数が time、打ち切りフラグが event、治療方法を示す群分け変数が treat、がんの進行度を表す変数が stage であるとき、進行度によってベースラインハザード関数が異なることを想定して、治療方法によって生存時間に差が出るかどうかコックス回帰で調べたければ、coxph(Surv(time, event) ~ treat + strata(stage)) とすればよい。

なお、コックス回帰はモデルの当てはめなので、一般化線型モデルで説明したのと同様、残差分析や尤度比検定、重相関係数の 2 乗などを用いて、よりよいモデル選択をすることができる。ただし、基準ハザード関数の型に特定の仮定を置かないと AIC は計算できない。

例題 6

survival ライブラリに含まれているデータ colon は結腸癌の補助的な化学療法の臨床試験の最初の成功例の 1 つである。Levamisole はそれまで動物の寄生虫処理に使われてきた毒性の低い化合物であり、5-FU は中等度の毒性をもつ化学療法の薬剤である。1 人あたり再発と死亡の 2 種類のレコードが含まれている。変数は以下の通りである。

```
id      個人の id 番号
study   全員が 1
rx      処理。3 水準をもつ要因型変数。Obs は経過観察のみ，Lev は Levamisole のみ投与，Lev+5FU は両方投与。
sex     性別を示す数値型変数。1 が男性，0 が女性。
age     満年齢を示す数値型変数。
obstruct 腫瘍による結腸の閉塞を示す数値型変数。1 が閉塞あり，0 が閉塞なし。
perfor 穿孔の有無を示す数値型変数。1 が穿孔あり，0 が穿孔なし。
adhere 隣接臓器への癒着の有無を示す数値型変数。1 が癒着あり，0 が癒着なし。
nodes  リンパ節転移数を示す数値型変数。
status  打ち切りかどうかを示す数値型変数。1 がイベント発生，0 が打ち切り。
differ  分化の程度を示す数値型変数。1 が高分化，2 が中等度，3 が低分化型。
extent  局所の広がり程度を示す数値型変数。1 が粘膜下，2 が筋，3 が漿膜，4 が隣接組織。
surg    手術後の時間の長さを示す。0 が短時間，1 が長時間。
node4   4 つ以上の癌細胞陽性のリンパ節があるかどうかを示す。1 があり，0 がなし。
time    イベント発生までの時間（日）を示す数値型変数。
etype   イベントの種類を示す数値型変数で，1 が再発，2 が死亡を意味する。
```

Levamisole 単独投与あるいは 5-FU と共に投与した場合に，経過観察に比べ，死亡のハザード比がいくつになるか，年齢と性別を共変量として調整してコックス回帰せよ。

まず死亡をエンドポイントとするサブセット colon2 を作る。colon2 <- subset(colon, etype==2) とすればよい。年齢と性別を共変量として調整したコックス回帰の実行は次の枠内の通り (loglogplot() は it13-5-2006.R で定義済みのもの)。

```
colon2$sex <- factor(colon2$sex)
KM <- survfit(Surv(time,status)~rx,data=colon2)
layout(1:2)
plot(KM)
loglogplot(KM)
res <- coxph(Surv(time,status)~rx+age+sex,data=colon2)
summary(res)
```

Levamisole と 5-FU を両方投与すると，経過観察だけのときに比べて，年齢と性別を調整したコックス回帰でハザード比が 0.688 倍 (95%信頼区間が [0.545,0.869]) で，ハザード比が 1 という帰無仮説の検定の有意確率は 0.0017 となるので，両方投与は結腸癌治療に有効であることが示された。このコックス回帰モデル自体のデータへの適合は， $R^2 = 0.013$ と低いが，尤度比検定で $\chi^2 = 12.5$, $d.f. = 4$, $p = 0.014$ と有意であり，意味はある。モデルの適合度を上げるには分化の程度など共変量を増やせばよい。ちなみに年齢と性別を調整したコックス回帰のベースラインの生存曲線を 95%信頼区間付きで処理別に描かせるには，次の枠内のよう
にすればよい^{*9}。

^{*9} なお，共変量を調整したコックス回帰で推定される生存関数について二重対数プロットを描かせるのはなかなか面倒であるが，余裕があれば試してみると面白いかもしれない。

```

attach(colon2)
xls <- c(0,max(time))
plot(survfit(coxph(Surv(time,status)~age+sex,subset=(rx=="Obs"))),col=1,xlim=xls)
par(new=T)
plot(survfit(coxph(Surv(time,status)~age+sex,subset=(rx=="Lev"))),col=2,xlim=xls)
par(new=T)
plot(survfit(coxph(Surv(time,status)~age+sex,subset=(rx=="Lev+5FU"))),col=3,xlim=xls)
detach(colon2)

```

例題 7

大橋、浜田 (1995) の付録 A に掲載されている膵臓癌データを入力してタブ区切りテキストファイルとして公開してあり、それを R で読み込んで変数の型などを適切に設定するプログラムも作成し公開してあるので、同書 p.138-150 に掲載されている SAS の解析結果と R の `coxph()` 関数による解析結果を比較せよ。

pcancer.R

```

dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/pcancer.txt")
dat$CENSOR <- 1-dat$CENSOR
dat$SEX <- factor(dat$SEX, labels=c("男性","女性"))
dat$TREAT <- factor(dat$TREAT, labels=c("照射なし","照射あり"))
dat$CH <- ordered(dat$CH,labels=c("CH0","CH1","CH2","CH3"))
dat$STAGE <- ordered(dat$STAGE,labels=c("III","IV"))
dat$PS <- ordered(dat$PS,labels=c("0,1","2","3","4"))
# data from 大橋・浜田 (1995) 「生存時間解析 SAS による生物統計」東京大学出版会付録 A.3
# 膵臓がんデータ (Nishimura et al. 1988. の一部)
# 手術中放射線照射の延命効果をみる目的のデータだがランダム割付ではない。
# 入力 Minato Nakazawa 20/1/2007 * R では死亡が 1 なので CENSOR のカテゴリに注意。
# *** 変数リスト ***
# CASENO 患者番号
# TIME 生存時間 (月) 連続量
# CENSOR 打ち切りの有無を示すフラグ (元データでは打ち切りが 1, 死亡が 0 で R と逆なので 1 から引く)
# AGE 手術時の年齢 連続量
# SEX 性別 0 が男性, 1 が女性
# TREAT 処置法 0 が術中照射無し, 1 が術中照射あり
# BUI 占拠部位 0 が頭部, 1 が頭部以外
# CH 膵内胆管への浸潤 順序尺度 1:CH0, 2:CH1, 3:CH2, 4:CH3
# P 腹膜播種性転移 0 がなし, 1 があり
# STAGE TNM 分類のステージ 順序尺度 3:III 期, 4:IV 期
# PS Performance Status (活動度) 順序尺度 1:0,1, 2:2, 3:3, 4:4

```

pcancer.R を実行後^{*10}、コックス回帰を行うコードは以下の通りである。同時発生イベントに対する当時の SAS のデフォルトは Breslow 法だったので、同書と結果を比較するために敢えて `method="breslow"` で実行したが、もちろん Efron 法や Exact 法の方が望ましい。

^{*10} `source("http://phi.med.gunma-u.ac.jp/medstat/pcancer.R")` とすればよい。

it13-7-2006.R

```
require(survival)
summary(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT, data=dat, method="breslow"))
summary(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT, data=dat, method="efron"))
summary(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT, data=dat, method="exact"))
res <- step(coxph(Surv(TIME,CENSOR)~AGE+SEX+TREAT+BUI+CH+P+STAGE+PS, data=dat,
                 method="breslow"))
summary(res)
```

結果の出力順序はやや異なるが、Breslow 法の結果は SAS とほぼ一致していた（わずかな丸め誤差と思われる違いは存在するが）。変数選択の結果も SAS と一致し、最終的に TREAT, BUI, STAGE が選ばれた。モデルの検定結果や有意確率もほぼ一致する。ただし、なぜか STAGE のリスク比だけが一致しない。

文献

- 大橋靖雄, 浜田知久馬 (1995) 「生存時間解析 SAS による生物統計」東京大学出版会.

課題

survival ライブラリに入っているデータ ovarian は、卵巣がんに対する 2 種類の治療法を比較する無作為化臨床試験の結果である。Eastern Cooperative Oncology Group の研究であり、含まれている変数は以下の通りである。

```
futime 生存時間または観察打ち切りまでの時間
fustat 打ち切りフラグ
age    年齢
resid.ds 残留疾病の有無 (1 がなし, 2 があり)
rx     治療種類 (処理群別を示す変数)
ecog.ps ECOG 能力状況 (0 が病気がないとまったく同じく何の制限もなく活動できる, 1 が強い運動はできないが軽い家事労働やオフィスワークならできる, 2 が起きている時間の半分くらいは活動できる, 3 が半分以上ベッドか椅子にいる, 4 がセルフケア不能, 5 が死亡を意味する)
```

このデータから、治療種類の違いによって卵巣がんの生存時間に差が出たか、年齢と残留疾病の有無を共変量として調整して分析せよ。

結果は A4 の紙に学籍番号と共に印刷し、氏名を自筆して提出すること。結果の提出をもって出席確認とする。なお、回答例は web サイトからダウンロードできるようにするので、各自参照されたい。