# 医学情報処理演習第 13 回「生存時間解析」[*1]

2006 年 1 月 30 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/

## 前回の訂正

前回課題で取り上げたサンプルデータ birthwt の説明で，母親の最終月経時の体重 lwt の単位がありませんでしたが，ポンド（略号 lb.，1 lb.=0.454 kg）でした。児の出生体重 bwt の単位が kg となっていましたが，g の間違いでした（現在 web で公開中のものでは訂正済み）。
また，データを加工した後，クリアせずに課題を実行したために，量的変数であるべき ftv がカテゴリ変数として扱われてしまい，正しいモデル選択ができなかった人がいました。データを全部クリアするには，rm(list=ls()) という命令を打つのが簡単です。

## 前回の課題の回答例

```
require(MASS)
data(birthwt)
attach(birthwt)
res <- lm(bwt~age+lwt+ptl+ftv)
summary(res2 <- step(res))
AIC(res2)
detach(birthwt)
```

上枠内のように従属変数を bwt，独立変数を age，lwt，ptl，ftv とする重回帰モデルをデータに当てはめ（VIF を計算すると，どれも 10 よりずっと小さいので多重共線性はない），step() 関数で変数減少法の変数選択を実行すると，最終的に以下のモデルが選択される。

出生時体重 (bwt) = 2464 + 4.0*母親の最終月経時体重 (lwt) - 194.0*非熟練労働経験数 (ptl)

係数がゼロという帰無仮説の検定の有意確率は，lwt について 0.0214，ptl について 0.0705 である。後者は有意水準 5%では有意でないが，重回帰モデル全体としては，F 値が 5.037（第一自由度 2，第二自由度 186）で，有意確率が 0.0074 なので，有意水準 5%で有意に当てはまっているといえる。符号も考えて係数を解釈すると，他の条件が同じなら，母親の最終月経時体重が 1 ポンド増えると児の出生体重は 4 グラム重くなり，非熟練労働経験数が 1 つ多いと児の出生体重は 194 グラム軽くなるといえる。しかし，自由度調整済み重相関係数の 2 乗が 0.041 しかないので，このモデルでは，出生時体重のばらつきの 4%強しか説明されないし，AIC も 3025 という大きな値になるので，このモデルで扱わなかった要因がもっと強く影響していると考えるべきである。
なお，上のコードに続けて下枠内を打てば，

```
sdd <- c(0,sd(lwt),sd(ptl))
coef(res2)*sdd/sd(bwt)
```

lwt と ptl の標準化偏回帰係数 ($\beta$) がそれぞれ，0.167，$-0.131$ となることがわかる。絶対値で比べると lwt の方が大きいので，母親の最終月経時体重の方が非熟練労働経験数よりも相対的に強く児の出生体重に影響しているといえる。
なお，http://phi.med.gunma-u.ac.jp/medstat/it12-kadai.R として以上のコード全体をダウンロード可能である。

[*1] 本資料は http://phi.med.gunma-u.ac.jp/medstat/it13-2005.pdf としてダウンロード可能である。

## 生存時間解析について

生存時間解析の特徴は，期間データを扱うことと，打ち切りデータを扱うことである。

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。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ものが 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）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。イベントを経験せずにいる人数が全体の半分になる時間を意味する。言い換えると，**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。

複数の期間データ列の差の比較には，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。中でも，今回は Mantel-Haenzel 流のログランク検定だけ紹介する。

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。

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。カプラン＝マイヤ法は `survfit()` 関数，ログランク検定は `survdiff()` 関数，コックス回帰は `coxph()` 関数，加速モデルは `survreg()` 関数で実行できる。

以下，カプランマイヤ法，ログランク検定，コックス回帰について簡単に説明し，演習するが，より詳しくは，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，などを参照されたい。

### 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て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。

## カプラン＝マイヤ法による生存曲線の推定

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 $t_1, t_2, ...$ 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 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは，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はなく，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，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，

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

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，

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

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。

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，下枠内のように打てば，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）が付値される。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）が計算され，`[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）についてだけの生存日数が得られ，それが `alivedays` に付値される。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，ダブルクォーテーションマークで括って，年，月，日がハイフンでつながれた形で与えることもできるし，最終行のように `ISOdate(`**年**`,`**月**`,`**日**`)` という形で与えることもできる。

it13-1.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$names=="Robert"]
alivedays/365.24
difftime(ISOdate(2005,1,31),x$dob)
```

R では, `library(survival)` または `require(survival)` としてパッケージを呼び出し, `dat <- 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，区間打ち切りデータを扱うときは，`Surv(`**観察開始時点，観察終了時点，打ち切りフラグ**`)` として，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ようになる。さらにいえば，まったく打ち切りデータがないときは，打ち切りフラグそのものを省略することもできる。生存時間データができたら，`res <- survfit(dat)` でカプラン＝マイヤ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。

例題 1

`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カプラン＝マイヤ推定し，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。

it13-2.R

```
require(survival)
data(aml)
res <- survfit(Surv(time,status)~x, data=aml)
print(res)
summary(res)
plot(res,lty=c(1,2))
```

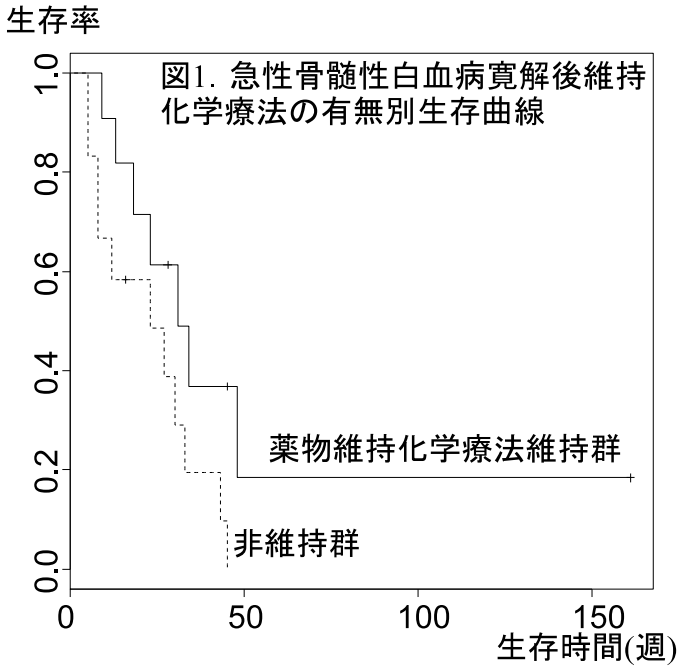
上枠内のように入力すると，3 行目の`~x` とすることで，群別に計算させることができる。4 行目を入力すると，下枠内が表示される。

```
                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。5 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。

最終行の `plot()` コマンドを打ってで表示される生存曲線を，メタファイル形式でクリップボードにコピー

し，OpenOffice.org の Draw というソフトに貼り付けて，フォントを変えたり注釈をつけたりして見やすくした図を下に示す[*2]。




## ログランク検定

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ので，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群しかないので，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

[*2] 通常，プレゼンテーションやポスターでは，このように図の中に細かく注釈を書き込むが，投稿論文の場合は，軸名以外の注釈や表題は，図の下に別枠で Figure legends として表示することが多い。

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[*3]。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*4/64+4*3/49+3*3/36+3*2/25+2*2/16+2*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` として，`survdiff(Surv(time,event)~group)` とすればよい。この例の場合なら，下枠内の通り。

it13-3.R

```
require(survival)
time2 <- 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)
survdiff(Surv(time2,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するのではなく，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。

[*3] 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。

例題 2

例題 1 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。

データを呼び出した後であれば，`survdiff(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

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 $z_i = (z_{i1}, z_{i2}, ..., z_{ip})$ 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，

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

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。$\beta_1, \beta_2, ..., \beta_p$ 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，

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

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，

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

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，横軸に生存期間の対数をとり，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 $\beta z$ 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。

## コックス回帰のパラメータ推定

パラメータ $\beta$ 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 $\beta$ 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。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（なお，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。

群分け変数も共変量となりうるので，生存時間を表す変数を `time`，打ち切りフラグを `event`，グループを `group` として，`coxph(Surv(time,event)~group)` とすれば，群間のハザード比が推定でき，それがゼロと差がないという帰無仮説が検定できる。イベント発生時間が同じ個体が 2 つ以上あるときの扱い方として Exact 法を用いるには，`coxph(Surv(time,event)~group, method="exact")` とすればよい。

例題 3

例題 1 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。

it13-4.R

```
require(survival)
data(aml)
res <- coxph(Surv(time,status)~x,data=aml)
summary(res)
plot(survfit(res))
```

4 行目で得られる結果は，下枠内の通りである。

```
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
```

どの検定結果をみても有意水準 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。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。

5 行目により，2 群を併せてコックス回帰を当てはめた生存曲線が，95%信頼区間付きでプロットされる[*4]。

## 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

[*4] 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用いる。2 群の生存曲線を別々に描きたい場合は，coxph() 関数の中で，subset=(x=="Maintained") 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必要がある。ただし信頼区間まで重ね描きされると見にくい。

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。

## 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 が病気がないの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。