大学院・医学基礎技術演習・実験基本技術（医学統計学）補講テキスト

中澤　港（生態情報学　助教授）

2006 年 8 月 9 日

問い合わせ先：生態情報学　助教授　中澤　港（e-mail: nminato@med.gunma-u.ac.jp）

1　共分散分析とロジスティック回帰分析

共分散分析もロジスティック回帰分析も一般化線型モデルの枠組みで扱うことができる。共分散分析は通常の線型モデルでよいが，ロジスティック回帰分析は従属変数が二項分布に従うので，一般化線型モデルで扱う。Rcmdr でも実行可能である。

1.1　共分散分析

共分散分析は，典型的には，$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ というモデルになる。2 値変数 X_1 によって示される 2 群間で，量的変数 Y の平均値に差があるかどうかを比べるのだが，Y が量的変数 X_2 と相関がある場合に（このとき X_2 を共変量と呼ぶ），X_2 と Y の回帰直線の傾き (slope) が X_1 の示す 2 群間で差がないときに，X_2 による影響を調整した Y の修正平均（adjusted mean; 調整平均ともいう）に，X_1 の 2 群間で差があるかどうかを検定する。

R では，X_1 を示す変数名を `C`（注：`C` は `factor` である必要がある），X_2 を示す変数名を `X` とし，Y を示す変数名を `Y` とすると，`summary(lm(Y~C+X))` とすれば，`X` の影響を調整した上で，`C` 間で `Y` の修正平均（調整平均）が等しいという帰無仮説についての検定結果が得られる（`C` のカテゴリが `1` と `2` である場合，`C2` と表示される行の右端に出ているのがその有意確率である）。

ただし，この検定をする前に，2 本の回帰直線がともに有意にデータに適合していて，かつ 2 本の回帰直線の間で傾き (slope) が等しいかどうかを検定して，傾きが等しいことを確かめておかないと，修正平均の比較には意味がない。そこで，まず例えば，`summary(lm(Y[C==1]~X[C==1]); summary(lm(Y[C==2]~X[C==2])` として 2 つの回帰直線それぞれの適合を確かめ，`summary(lm(Y~C+X+C:X))`（または `summary(lm(Y~C*X))`）として傾きが等しいかどうかを確かめなければならない。傾きが有意に違っていることは，`C` と `X` の交互作用項が有意に `Y` に効いていることと同値なので，`Coefficients` の `C2:X` と書かれている行の右端を見れば，「傾きに差がない」という帰無仮説の検定の有意確率が得られる。そもそも回帰直線の適合が悪ければその独立変数は共変量として考慮する必要がないし，傾きが違っていれば群分け変数と独立変数の交互作用が従属変数に関して有意に影響しているということなので，2 群を層別して別々に解釈する方が良い。

例題

R の組み込みデータ `ToothGrowth` は，各群 10 匹ずつのモルモットに 3 段階の用量のビタミン C をアスコルビン酸としてあるいはオレンジジュースとして投与したときの象牙芽細胞（歯）の長さを比較するデータである。変数 `len` が長さ，`supp` が投与方法，`dose` が用量を示す。用量と長さの関係が投与方法によって異なるかどうかを共分散分析を使って調べよう。

データを使えるようにしてから，まずグラフを描いてみる。共分散分析をするような場面では，通常，下枠内のように，群によってマークを変えて散布図を重ね描きし，さらに線種を変えて群ごとの回帰直線を重ね描きするのだが，`coplot(len~dose | supp)` として横に 2 枚のグラフが並べて描かれるようにすることも可能である。

```
> data(ToothGrowth)
> attach(ToothGrowth)
> plot(dose,len,pch=as.integer(supp),ylim=c(0,35))
> legend(max(dose)-0.5,min(len)+1,levels(supp),pch=c(1,2))
> abline(lm1 <- lm(len[supp=='VC']~dose[supp=='VC']))
> abline(lm2 <- lm(len[supp=='OJ']~dose[supp=='OJ']),lty=2)
> summary(lm1)
> summary(lm2)
```

summary(lm1) と summary(lm2) をみると，投与方法別の回帰係数がゼロと有意差があることがわかる。そこで次に，これらの回帰係数間に有意差がないという帰無仮説を検定する。モデルの右辺に独立変数間の交互作用項を含めればいい。

```
> lm3 <- lm(len ~ supp*dose)
> summary(lm3)
Call:
lm(formula = len ~ supp * dose)

Residuals:
     Min       1Q   Median       3Q      Max
-8.22643 -2.84625  0.05036  2.28929  7.93857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.550      1.581   7.304 1.09e-09 ***
suppVC        -8.255      2.236  -3.691 0.000507 ***
dose           7.811      1.195   6.534 2.03e-08 ***
suppVC:dose    3.904      1.691   2.309 0.024631 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-Squared: 0.7296,     Adjusted R-squared: 0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

この結果から，suppVC:dose の従属変数 len への効果（交互作用効果）がゼロという帰無仮説の検定の有意確率が 0.024631 となるので，有意水準 5%で帰無仮説は棄却される。従って，この場合は，**投与経路によって投与量と長さの関係の傾きが有意に異なるので，と提示した上で，**先に計算済みの，投与経路別の回帰分析の結果を解釈すればよい（修正平均の差の検定はしても意味がない）。

例題 2

組み込みデータ swiss（1888 年頃のスイスのフランス語を話す 47 州についての，標準化された出生力水準 Fertility，農業就業割合 Agriculture，陸軍の試験で最高ランクを記録した人の割合 Examination，初等教育を超える教育を受けた人の割合 Education，カソリック信者割合 Catholic，乳児死亡割合 Infant.Mortality からなるデータ）を使って，教育水準が高いほど出生力が低いけれども，それがカソリック信者割合に影響を受ける（カソリック信者の方がプロテスタント信者よりも一般に出生力が高い）という仮説を検討してみよう。

Rcmdr を使ってみる。まず，library(Rcmdr) として R コマンダーを呼び出す。次に，「データ」→「パッケージ内のデータ」→「アタッチされたパッケージからデータセットを読み込む」として，パッケージとして datasets，データとして swiss を選択する。次に，「データ」→「アクティブデータセット内の変数の管理」→「数値変数を区間で区分」として，Catholic が 50%を超えるかどうかを割振る変数 MoreCatholic を作る。MoreCatholic で層別して散布

図を描かせてから，「統計量」→「モデルへの適合」→「線型モデル」で，モデルとして左辺に `Fertility` を，右辺に `MoreCatholic*Education` を指定すれば交互作用項により傾きの差が検討できる。この場合，傾きの差は有意ではないので，もう一度「線型モデル」を呼び出して，右辺を `MoreCatholic+Education` とすれば，教育水準を調整してもカソリックが多いかどうかによって標準化された出生力の調整平均に差があるかどうかがわかる。

1.2 ロジスティック回帰分析

ロジスティック回帰分析は，従属変数（ロジスティック回帰分析では反応変数と呼ぶこともある）が 2 値変数であり，二項分布に従うので `lm()` ではなく，`glm()` を使う一般化線型モデルとなる。ロジスティック曲線とは関係ない。従属変数がポアソン分布に従う場合も `glm()` で扱えるが，それはポアソン回帰と呼ばれる。

ロジスティック回帰分析の思想としては，例えば疾病の有無を，複数のカテゴリ変数によって表される要因の有無で説明する（量的な変数によって表される交絡を調整しながらオッズ比を計算できるのが利点であり，医学統計ではもっともよく使われる手法の一つである）。

この問題は，疾病の有病割合を P とすると，$\ln(P/(1-P)) = b_0 + b_1X_1 + ...b_kX_k$ と定式化できる。X_1 が要因の有無を示す 2 値変数で，$X_2, ...X_k$ が交絡であるとき，$X_1 = 0$ の場合を $X_1 = 1$ の場合から引けば，

$$b1 = \ln(P_1/(1-P_1)) - \ln(P_0/(1-P_0)) = \ln(P_1 * (1-P_0)/(P_0 * (1-P_1)))$$

となるので，b_1 が他の変数の影響を調整したオッズ比の対数になる。対数オッズ比が正規分布するとすれば，オッズ比の 95%信頼区間が

$$\exp(b_1 \pm 1.96 \times \mathrm{SE}(b_1))$$

として得られる。

例題 3 として，`library(MASS)` にある `data(birthwt)` を使った実行例を示す。

Springfield の Baystate 医療センターの 189 の出生について，低体重出生とそのリスク因子の関連を調べるためのデータである。`str(birthwt)` とすると変数がわかる。

low	低体重出生の有無を示す 2 値変数（児の出生時体重 2.5 kg 未満が 1）
age	年齢
lwt	最終月経時体重（ポンド[a]）
race	人種（1＝白人，2＝黒人，3＝その他）
smoke	喫煙の有無（1＝あり）
ptl	非熟練労働経験数
ht	高血圧の既往（1＝あり）
ui	子宮神経過敏の有無（1＝あり）
ftv	妊娠の最初の 3 ヶ月の受診回数
bwt	児の出生時体重 (g)

[a] 略号 lb. で，1 lb. は 0.454 kg に当たる。

```
> require(MASS)
> data(birthwt)
> attach(birthwt)
> low <- factor(low)
> race <- factor(race, labels=c("white","black","other"))
> ptd <- factor(ptl>0)
> smoke <- (smoke>0)
> ht <- (ht>0)
> ui <- (ui>0)
> ftv <- factor(ftv)
> levels(ftv)[-(1:2)] <- "2+"
> bw <- data.frame(low,age,lwt,race,smoke,ptd,ht,ui,ftv)
> detach(birthwt)
> summary(res <- glm(low ~ ., family=binomial, data=bw))
> summary(res2 <- step(res))
```

変数選択後の結果をみると，smokeTRUE の係数（対数オッズ比）は 0.866582 で，その SE が 0.404469 である。したがって，最終的なモデルに含まれる他の変数（最終月経時体重，黒人，他の有色人種，非熟練労働経験あり，高血圧既往あり，子宮神経過敏あり）の影響を調整した喫煙の低体重出生への効果（オッズ比とその 95%信頼区間）は，下枠内によって得られる。なお，人種は 3 つのカテゴリがあるので，自動的にダミー変数化されて処理される。

```
> exp(0.866582)
[1] 2.378766
> exp(0.866582 - qnorm(0.975)*0.404469)
[1] 1.076616
> exp(0.866582 + qnorm(0.975)*0.404469)
[1] 5.255847
```

この結果から，喫煙者は非喫煙者に比べて約 2.38 倍（95%信頼区間は [1.08, 5.26]），低体重出生児をもちやすいということを示している（95%信頼区間の下限が 1 より大きいので，有意水準 5%で有意な影響があったといえる）。

2 生存時間解析

生存時間解析は，観察打ち切りデータを扱うために必須である。他の多くの解析手法が一時点での属性や測定値を扱うのに対して，生存時間解析で扱われるデータは時間的な間隔である。いまのところ Rcmdr では扱えないので，コンソールからコマンドを打たねばならない。生存時間解析を行う関数は survival ライブラリに含まれているが，これは Recommended ライブラリなので，Windows 版の R では標準でインストールされているが，ロードはされていない。そのため，まず，library(survival) として survival に含まれている関数が使えるようにしなくてはならない。

大橋・浜田 (1995) で説明に用いられている Gehan の白血病治療データは，R では MASS ライブラリに含まれているので，これら 2 つのライブラリを呼び出す必要がある。既にインストールされているライブラリを呼び出すには，library() または require() を用いる。

2.1 カプラン＝マイヤ推定

大橋・浜田 (1995) の p.60-61 にあるような，Gehan の白血病治療データでの 6-MP 投与群と対照群別々のカプラン＝マイヤ推定をするには，R では以下のようにする。

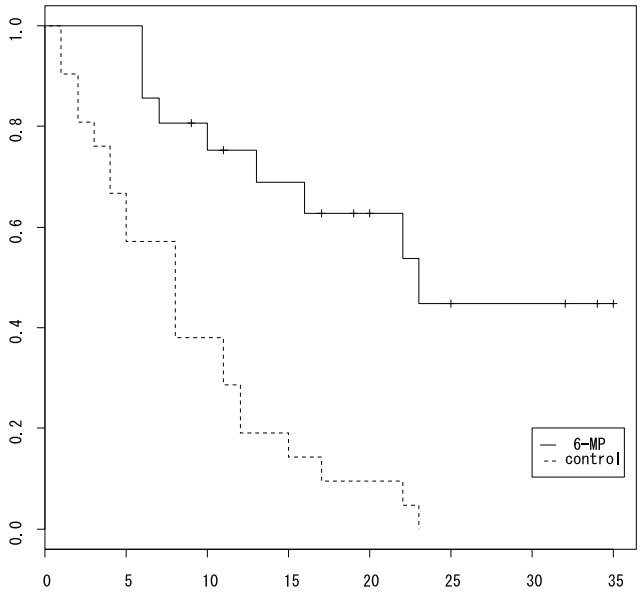
Surv() は期間データと打ち切りフラグから生存時間型のデータを構成する関数であり，生存時間解析の関数は，この型のデータを扱うことができる。summary() は詳しい出力が欲しいときにつける[*1]。

[*1] 生存時間解析の結果に限らず，多くの結果オブジェクトに使える。ただし，オブジェクトによっては summary メソッドを持っていない場合

```
> require(MASS)
> require(survival)
> print(res<-survfit(Surv(time,cens)~treat,data=gehan))
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
> summary(res)
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
> plot(res,lty=c(1,2))
> legend(30,0.2,lty=c(1,2),legend=levels(gehan$treat))
```

グラフィック出力ウィンドウに下図が得られる（本資料は LaTeX で作成しているが，LaTeX に画像を取り込むために，graphicx パッケージを使って eps ファイルを読み込んだ。美しい eps ファイルを作成するため，グラフィック出力ウィンドウから OpenOffice.org の Draw に画像をコピー&ペーストしてサイズを決め，加工してからエクスポート機能で eps 出力を行っている）。

もあり，その場合は詳しい出力とはならない。



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。

```
> 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)
```

2.2 ログランク検定

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[*2]。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。

```
> 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。

[*2] 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。

2.3 コックス回帰—比例ハザードモデル—の考え方

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 における瞬間イベント発生率を意味する。$\beta_1, \beta_2, ..., \beta_p$ 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。

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 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。

2.4 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。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。

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するには以下のようにする。

```
> require(MASS)
> require(survival)
> res <- coxph(Surv(time,cens)~treat,data=gehan)
> summary(res)
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
> plot(survfit(res))
```

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[*3]。`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。

最後の行の `plot()` 関数により，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]。

[*3] 言うまでもなく，`Score (logrank) test` の結果は，`survdiff(Surv(time,cens)~treat,data=gehan)` の結果と同じである。

[*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=="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。

2.5 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。また，コックス回帰のパラメータ推定で，同時に発生したイベントが２つ以上ある場合の扱い方は，Breslow 法を採用しているパッケージが多いが，R のデフォルトは，より Exact 法に近いと言われている Efron 法である。

2.6 その他の技法

加速モデルは `survreg()` 関数で実行可能である。この他にも R には数多くの関数やライブラリが存在するので，前述の RjpWiki からリンクを辿って探せば，大抵のデータ解析はできるだろう。

3 参考文献

1. 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）
2. 中澤　港 (2003)『R による統計解析の基礎』（ピアソン・エデュケーション）
3. 間瀬 茂・神保 雅一・鎌倉 稔成・金藤 浩司 (2004)『工学のための数学３　工学のための データサイエンス入門 ― フリーな統計環境 R を用いたデータ解析 ― 』（数理工学社）
4. 岡田昌史（編） (2004)『The R Book －データ解析環境 R の活用事例集－』（九天社）
5. 舟尾暢男 (2005)『The R Tips －データ解析環境 R の基本技・グラフィック活用集』（九天社）
6. 鈴木義一郎 (1995) 『情報量基準による統計解析』（講談社サイエンティフィク）
7. Armitage P, Berry G, Matthews JNS (2002)『Statistical Methods in Medical Research, 4th ed.』(Blackwell Publishing)
8. Maindonald J, Braun J (2003) 『Data analysis and graphics using R』（Cambridge Univ. Press）