第14回　高度な分析

- これまで13回は自分でできるようになって欲しい内容
- 論文にはもっと高度な分析もある
- 今回はそういう高度な分析の紹介
  - モデルの当てはめ
    - 重回帰分析・変数選択・多重共線性・尤度比検定
    - 共分散分析
    - ロジスティック回帰分析
    - 因子分析
  - 生存時間解析
    - Kaplan-Meier法
    - ログランク検定
    - コックス回帰

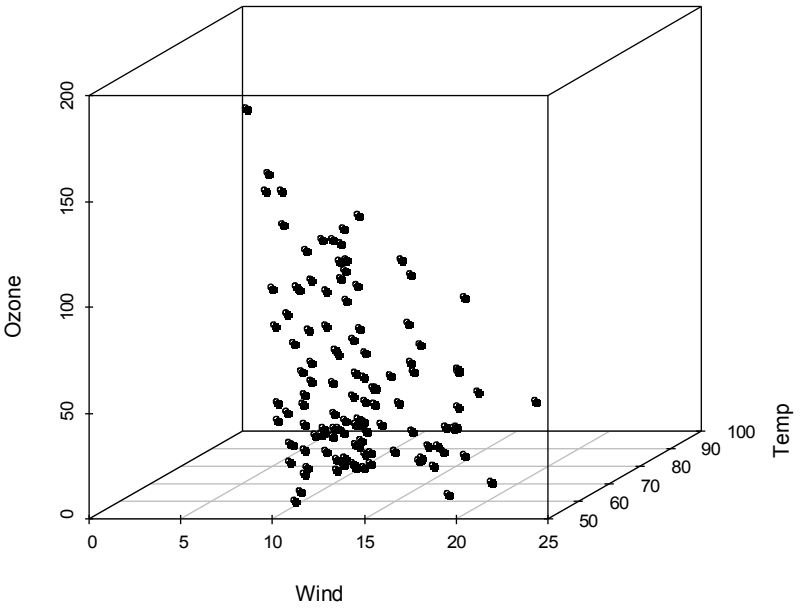
モデルの当てはめ

- 回帰モデル：従属変数のばらつきを独立変数のばらつきで説明。
- 独立変数は複数の変数の線形結合でもOK。
- 独立変数が1つのとき単回帰分析，2つ以上のとき重回帰分析。
- 例えば，収縮期血圧値SBPを，塩分摂取量SALTと年齢AGEで説明するモデルの場合，
  $SBP = \beta 0 + \beta 1\ SALT + \beta 2\ AGE + \varepsilon$
  という形になる（εは誤差項。β1，β2は偏回帰係数）
- 偏回帰係数の推定の際，多重共線性には注意（例えば，VIFを使ってチェックする）。
  (cf.) library(fmsb); ?VIF

(例)ニューヨーク気象データ:オゾン濃度の分析

- 単回帰分析では
```
attach(airquality)
plot(Ozone ~ Wind)
res1 <- lm(Ozone ~ Wind)
summary(res1)
detach(airquality)
```
としていた。

- 重回帰では風速(Wind)だけでなく，例えば気温(Temp)のオゾン濃度(Ozone)への影響も同時にみることができる




```
attach(airquality)
library(rgl) # 動かせる
plot3d(Wind,Temp,Ozone)
library(scatterplot3d) # きれいな出力
scatterplot3d(Wind,Temp,Ozone,pch=20,angle=30,scale.y=0.5)
res2 <- lm(Ozone ~ Wind + Temp)
summary(res2)
sdd <- c(0, sd(res2$model$Wind), sd(res2$model$Temp))
stb <- coef(res2)*sdd/sd(res2$model$Ozone)
stb
detach(airquality)
```

重回帰モデルの結果の見方

```
Call:
lm(formula = Ozone ~ Wind + Temp)

Residuals:
    Min      1Q  Median      3Q     Max
-41.251 -13.695  -2.856  11.390 100.367

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -71.0332    23.5780  -3.013   0.0032 **
Wind         -3.0555     0.6633  -4.607 1.08e-05 ***
Temp          1.8402     0.2500   7.362 3.15e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.85 on 113 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared: 0.5687,     Adjusted R-squared: 0.5611
F-statistic:  74.5 on 2 and 113 DF,  p-value: < 2.2e-16
```

標準化偏回帰係数

```
> stb
(Intercept)        Wind        Temp
  0.0000000  -0.3311197   0.5291333
```

AICと尤度比検定

- モデルの当てはまりの悪さの指標**AIC**
  > AIC(res1) # 1093.187となる
  > AIC(res2) # 1049.741となる。
  重回帰モデルの方が当てはまりは良い→有意性は？
- **2**つのモデルを比べてみる
  Ozone = β0 + β1 Wind + ε (1)
  Ozone = β0 + β1 Wind + β2 Temp + ε (2)
  (1)は，**(2)**で**β2**が**0**という特殊例とみなせる
- 一般性がより低いモデル**(1)**の最大尤度の，一般性がより高いモデル**(2)**の最大尤度に対する比の対数をとってマイナス**2**倍した値が近似的にカイ二乗分布に従うことから，「**(1)**と**(2)**で当てはまりに差が無い」帰無仮説を尤度比検定できる

```
LL1 <- logLik(res1)
LL2 <- logLik(res2)
lambda <- -2*(as.numeric(LL1)-as.numeric(LL2))
dff <- attr(LL2,"df")-attr(LL1,"df")
1-pchisq(lambda, dff)
```

共分散分析

- 量的変数XとY，二値変数Bがあるとき
- 「YにBの2群間で差が無い」を検定したいけれども，XとYに相関があり，Xの大きさを調整した上でBの2群間での比較をしたいとき，
  $Y = \beta0 + \beta1\ X + \beta2\ B + \beta12\ X*B + \varepsilon$
  というモデルを考える。
- β12がゼロと有意差があるとき→BによってXとYの関係は異なる
- β12が有意でないとき→交互作用は有意でない
  $Y = \beta0 + \beta1\ X + \beta2\ B + \varepsilon$
  を考える。β2がゼロと有意差があれば最初に立てた帰無仮説は棄却できる。

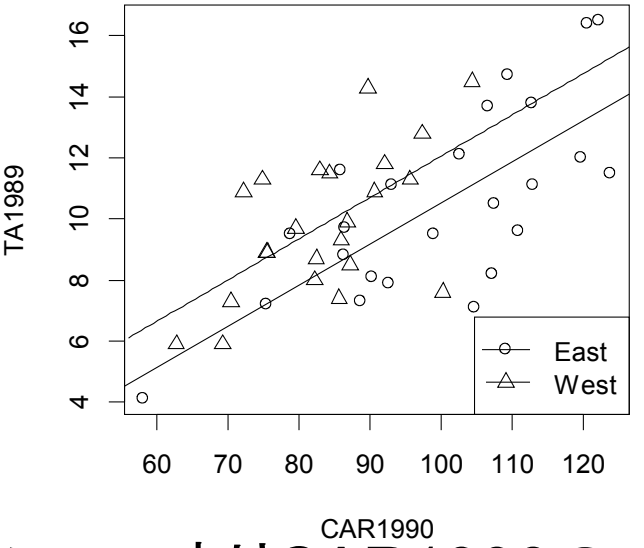
共分散分析の例

- 日本の各都道府県(PREF)の，
  1990年の100世帯当たり乗用車台数(CAR1990)，
  1989年の人口10万人当たり交通事故死者数(TA1989)，
  1985年の国勢調査による人口集中地区居住割合(DIDP1985)
  を含むhttp://phi.med.gunma-u.ac.jp/grad/sample3.dat
  を読み込み，東日本か西日本か(REGION)間で交通事故死者数を比較する。ただし乗用車台数は交通事故死者数と関連していると思われるので，その影響を共変量として調整する共分散分析実行
- コードは下記

```
x<-read.delim("http://phi.med.gunma-u.ac.jp/grad/sample3.dat")
plot(TA1989 ~ CAR1990, pch=as.integer(REGION), data=x)
east <- lm(TA1989 ~ CAR1990, data=subset(x, REGION=="East"))
summary(east); abline(east, lty=1)
west <- lm(TA1989 ~ CAR1990, data=subset(x, REGION=="West"))
summary(west); abline(west, lty=2)
legend("bottomright",pch=1:2,lty=1:2,legend=c("East","West"))
summary(lm(TA1989 ~ REGION*CAR1990, data=x))
ac <- lm(TA1989 ~ REGION+CAR1990, data=x)
summary(ac)
cfs <- dummy.coef(ac)
cfs[[1]] + cfs$CAR1990 * mean(ac$model$CAR1990) + cfs$REGION
```

共分散分析の結果の見方

- summary(east)で，CAR1990の係数は0.1346(p=5.7e-5)
  summary(west)で，CAR1990の係数は0.1352(p=0.0026) ともに5%水準で有意



- summary(lm(TA1989 ~ REGION*CAR1990, data=x))の結果から，REGIONWest:CAR1990の係数が0.00062(p=0.99)なので，REGIONとCAR1990の交互作用効果は5%水準で有意でない。つまりCAR1990のTA1989への影響はREGION間で差が無い→共分散分析へ

- summary(ac)の結果でAdjusted R-squaredが0.4488なので，共分散モデルは，データのばらつきの約45%を説明しているといえる。また

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.95787	2.21349	-1.336	0.1883	
REGIONWest	1.52190	0.68689	2.216	0.0319	*
CAR1990	0.13475	0.02177	6.189	1.78e-07	***

5%有意

＊100世帯当たり車保有が1台増えると10万人当たり交通事故が0.135件増え，車保有台数が同じなら西日本は東日本より1.52件事故が多い

- 最後に出てくるEast 9.4446とWest 10.9665が「修正平均」

ロジスティック回帰分析

- 目的変数（応答変数）が2値データ（イベントが起こる／起こらない等）で，正規分布ではなく二項分布に従う回帰モデル。glm()を用いる。
- （例）MASSライブラリのbacteriaデータ（オーストラリア）
  - http://www.menzies.edu.au/publications/anreps/MSHR00.pdf
  - Dr. A. LeachによるRCT。中耳炎の既往のある50人の子供に投薬し，定期的にH. Influenzaeの検出をチェックした結果
  - 変数は，y（菌の有無。nが無し，yが有り），ap（薬の種類。aが実薬，pがプラセボ），hilo（服薬遵守を促す程度，hiかlo），week（研究開始からの週数），ID（個人番号），trt（apとhiloを組み合わせた処理種類，"placebo"，"drug"，"drug+"）
- 投薬と服薬遵守を促す程度と週数が菌の有無に影響するかどうかを調べるデザインなので，目的変数が菌の有無という2値変数であり，ロジスティック回帰分析になる。

```
library(MASS)
res <- glm(y ~ ap+hilo+week, binomial, data=bacteria)
```

結果の解釈

対数オッズ比

```
Call:
glm(formula=y~ap+hilo+week, family = binomial, data = bacteria)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3763   0.3813   0.5212   0.6576   1.1194

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9278     0.3762   5.124 2.99e-07 ***
app           0.8343     0.3816   2.186  0.02879 *
hilolo       -0.5066     0.3546  -1.428  0.15317
week         -0.1167     0.0443  -2.633  0.00845 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 217.38  on 219  degrees of freedom
Residual deviance: 202.90  on 216  degrees of freedom
AIC: 210.9

Number of Fisher Scoring iterations: 4
```

有意水準5%で有意にオッズ比が1と異なる

作表方法

- ロジスティック回帰分析の結果の表には，対数オッズ比を掲載しても仕方ないので，通常は指数関数を使ってオッズ比に戻す。
  exp(coef(res)) # オッズ比の点推定量を表示
  exp(confint(res)) # 95%信頼区間を表示
- 下のように作表する

		95%信頼区間		
説明変数	オッズ比	下限	上限	p値
プラセボ投与	2.30	1.11	5.03	0.029
遵守指導低	0.60	0.29	1.21	0.153
週数	0.89	0.81	0.98	0.008

- プラセボ投与群は治療薬投与群に比べ，菌が存在する確率が2.3倍あり，1週間経つごとに菌が存在する確率が0.89倍になると解釈できる

因子分析

- 目的：観察された変数の背後の潜在因子を推定
  - 結果的に，相互に関連のある多変数の次元の縮約ができる。その意味では主成分分析に似ている
- 基本：
  - 入力データ：通常300ケース以上，変数はケース数の1/2～1/10が普通。変数は正規分布すべきで，外れ値は極力排除しておくべき。他の変数とまったく関係が無い変数や他の変数に線形従属する変数はモデルに入れない。
  - 出力：(1)因子負荷量。潜在因子と各変数との相関を意味する。(2)因子得点。各潜在因子への各個人の寄与を意味する
  - 回転：因子同士が独立であることを維持した直交回転と，必ずしも独立でなくてもいい斜交回転がある。直交回転の代表的なものとして，バリマックス回転が有名
  - ツール：スクリープロット，バートレットの球面性検定，KMOのサンプリング適切性基準，クロンバックのαなど。

因子の評価と基本モデル

- 抽出された因子には適切な命名が必要。
- 3つ以上の変数が寄与しているのが基本
- 1つか2つしかないときは因子を抽出しすぎか多重共線性の可能性
- 基本モデル
  - 10変数，300人のデータ。2つの潜在因子を仮定すると，
    $X1 = \beta1.1\ F1 + \beta2.1\ F2 + \varepsilon1$
    $X2 = \beta1.2\ F1 + \beta2.2\ F2 + \varepsilon2$
    ....
    $X10 = \beta1.10\ F1 + \beta2.10\ F2 + \varepsilon10$
  - βは因子負荷量。εは誤差分散（uniquenessとも呼ばれる）。
    εがなるべく小さくなるように，多様な方法（主因子法や最尤法など）でβを推定。n番目の人について
    $\beta1.1\ X1[n] + \beta1.2\ X2[n] + ... + \beta1.10\ X10[n]$
    の値は第1因子得点となる。

Rで因子分析を実行するには

- 使える関数
  - factanal()：標準で入っている。factors=オプションで因子数指定が必要
  - paf(): relaパッケージ所収。最適な因子数が自動的に決まる
  - fa(): psychパッケージ所収。nfactors=オプションで因子数指定が必要。
  - alpha(): psychパッケージ所収。クロンバックのαを計算
  - cortest.bartlett(): psychパッケージ所収。バートレットの球面性検定を実行
  - fa.parallel(): psychパッケージ所収。パラレル分析により，適切な因子数を計算する

因子分析実行例

```
library(MASS) # データbiopsyがMASSパッケージに入っているので
# biopsyはウィスコンシン大学病院の乳房の腫瘍699例のバイオプシー
# IDは患者ID，classはbenign（良性）かmalignant（悪性）
# 他のV1～V9はバイオプシー所見10段階評価
# V1:clump thickness. V2:uniformity of cell size.
# V3:uniformity of cell shape. V4:marginal adhesion.
# V5:single epithelial cell size.
# V6:bare nuclei (16 values are missing).
# V7:bland chromatin. V8:normal nucleoli. V9:mitoses.
BP <- subset(biopsy,complete.cases(biopsy)) # 欠損値除去
(F3 <- factanal(BP[,2:10], factor=3))
(F4 <- factanal(BP[,2:10], factor=4))
(F5 <- factanal(BP[,2:10], factor=5))
# factor=6にすると因子数が多すぎるというエラーが出る
barplot(colSums(F5$loadings^2)) # スクリープロット
```

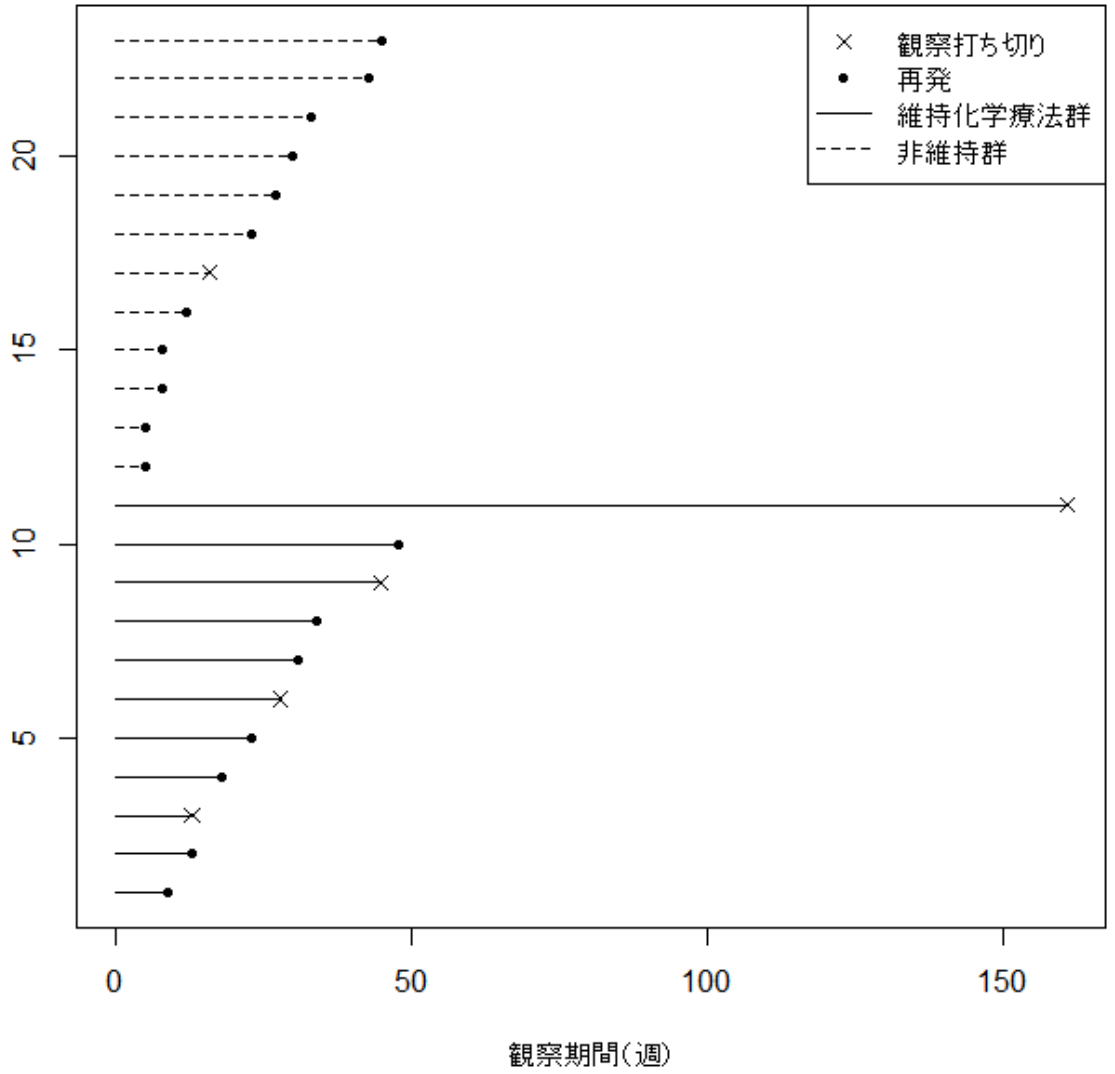
因子数5の結果

```
Uniquenesses:
   V1    V2    V3    V4    V5    V6    V7    V8    V9
0.445 0.032 0.127 0.293 0.382 0.236 0.316 0.005 0.005

Loadings:
   Factor1 Factor2 Factor3 Factor4 Factor5
V1 0.407   0.296   0.219   0.163   0.477
V2 0.795   0.405   0.290   0.205   0.214
V3 0.679   0.417   0.310   0.190   0.324
V4 0.375   0.669   0.248   0.216   0.102
V5 0.529   0.366   0.289   0.283   0.200
V6 0.318   0.642   0.207   0.118   0.439
V7 0.459   0.547   0.327   0.118   0.231
V8 0.340   0.306   0.844   0.197   0.185
V9 0.179   0.149   0.143   0.954

                 Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings        2.139   1.834   1.257   1.208   0.719
Proportion Var     0.238   0.204   0.140   0.134   0.080
Cumulative Var     0.238   0.441   0.581   0.715   0.795

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 0.52 on 1 degree of freedom.
The p-value is 0.471
```

因子数5が十分であるという仮説は棄却されない（因子数3や4だと5%水準で棄却される）

生存時間解析

- イベント発生までの時間を分析する
- イベント発生の有無より，時間の分布を調べる方が情報量が多い
- 「イベントがまだ発生していない」状態の扱いが鍵
  - 観察終了時まではイベントが発生していないので，イベント発生までの時間が長めな人である可能性が高い
    - 単純に除去するとイベント発生までの時間が過小評価される
    - 右側打ち切りレコードとして扱う！

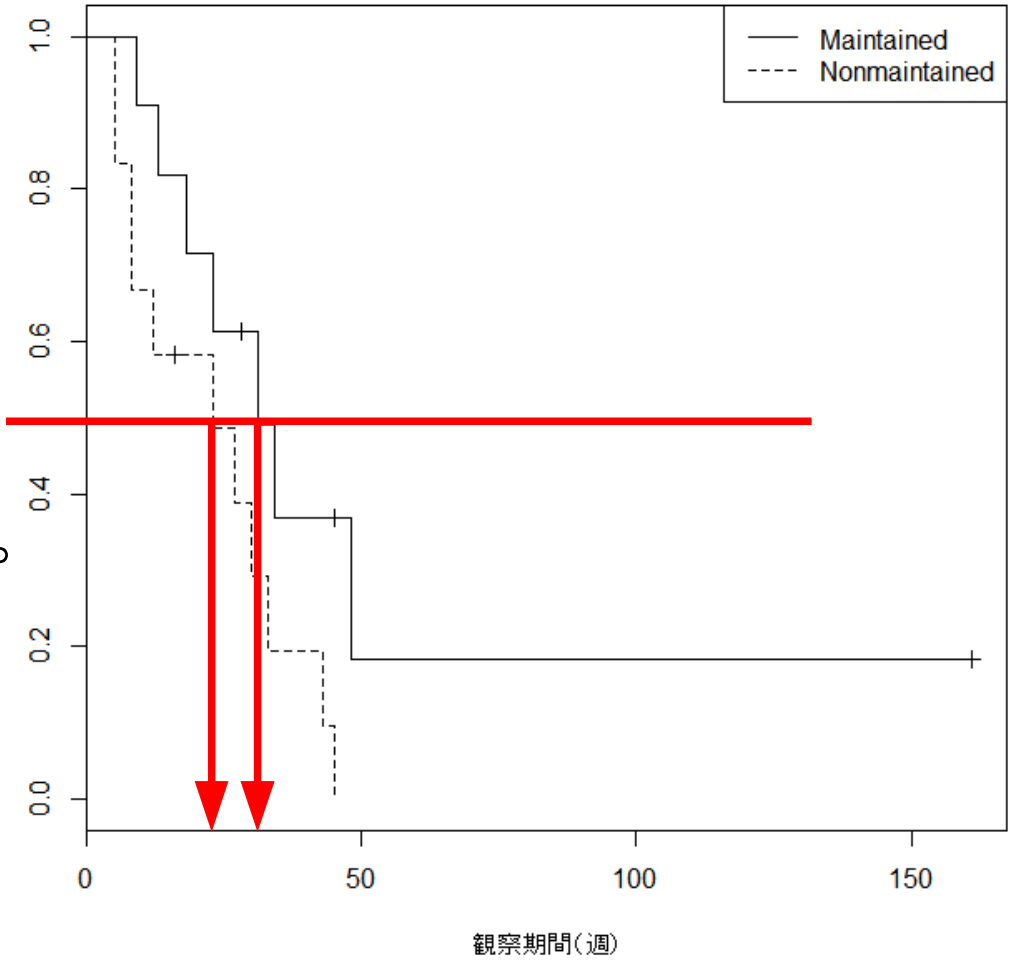


上図を描くコードは下記

```
library(survival)
plot(1:23~seq(0,max(time),len=23),type="n",data=aml,
  xlab="観察期間（週）", ylab="")
segments(rep(0,23),1:23,aml$time,1:23,
 lty=ifelse(aml$x=="Maintained",1,2))
points(aml$time,1:23,pch=ifelse(aml$status==0,4,20))
legend("topright",pch=c(4,20,NA,NA),lty=c(NA,NA,1,2),
 legend=c("観察打ち切り","再発","維持化学療法群","非維持群"))
```

Kaplan-Meier法

- 右側打ち切りを考慮した生存時間の中央値（カプラン＝マイヤ推定量）を求める方法
- 時点0では生存率が1。各死亡イベントが起こった時点ごとに，その時点で観察中の人数でその時点の死亡数を割った値を1から引いた値を掛けていくと，順次その時点までの生残率が得られる。
- 死亡イベントと死亡イベントの間で打ち切りがあると観察中人数が減る。
- 生残率が0.5を横切った時点が生存時間の中央値（維持群31週，非維持群23週）。



- 先のamlデータでのコードは下記

```
library(survival) # ライブラリ呼び出しが必須
(KM <- survfit(Surv(time,status)~x,data=aml))
plot(KM, lty=1:2, xlab="観察期間（週）")
legend("topright",lty=1:2,legend=names(table(aml$x)))
```

ログランク検定

- 「2群の生存時間に差が無い」帰無仮説の検定。
- 生存時間の順序の情報を用いる
- **aml**データで，維持群と非維持群で白血病再発までの時間に差が無いという帰無仮説をログランク検定するコードは，

```
library(survival)
survdiff(Surv(time, status) ~ x, data=aml)
```

- 結果は

```
                 N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained    11        7    10.69      1.27       3.4
x=Nonmaintained 12       11     7.31      1.86       3.4

 Chisq= 3.4  on 1 degrees of freedom, p= 0.0653
```

- **p**値が**0.05**より大きいので，有意水準**5%**で有意な差があるとはいえない。

コックス回帰

- 比例ハザードモデルともいう
- 仮想的な「基準人」の生存曲線を計算し，個々の人のイベント発生ハザードは基準人のハザードに比例するという「比例ハザード性」を仮定する
- 生存曲線の関数形は仮定しないのでセミパラメトリックなモデルと言われる
- 詳細は説明しないが，コードは下記

```
library(survival)
res<-coxph(Surv(time,status)~x, data=aml)
summary(res)
plot(survfit(res)) # 基準生存曲線と95%信頼区間
# 比例ハザード性の確認は二重対数プロットが平行ならOK
KM <- survfit(Surv(time,status)~x, data=aml)
plot(KM,fun=function(y){log(-log(y))},lty=1:2)
```