生存時間分析

1226629C
弥島有希

生存時間分析とは？

- 故障や死亡などのイベントとそれが起きるまでにかかる時間との関係を調べる分析方法。

研究（観察）終了時までにイベントが起きなかったケース

→打ち切り

パッケージ：survival

●ﾉﾝﾊﾟﾗﾒﾄﾘｯｸﾓﾃﾞﾙ

●ｾﾐﾉﾝﾊﾟﾗﾒﾄﾘｯｸﾓﾃﾞﾙ

●ﾊﾟﾗﾒﾄﾘｯｸﾓﾃﾞﾙ　　の三種類

打ち切りとデータ

- gehanの構造

```
> library(survival);library(MASS)
 要求されたパッケージ splines をロード中です
> data(gehan);dim(gehan);
[1] 42  4
> gehan[1:6,]
  pair time cens   treat
1    1    1    1 control
2    1   10    1    6-MP
3    2   22    1 control
4    2    7    1    6-MP
5    3    3    1 control
6    3   32    0    6-MP
> Surv(gehan$time,gehan$cens)
 [1]  1  10  22   7   3  32+ 12  23   8  22  17   6   2  16  11  34+  8  32+
[19] 12  25+  2  11+  5  20+  4  19+ 15   6   8  17+ 23  35+  5   6  11  13
[37]  4   9+  1   6+  8  10+
>
```

行列・配列のサイズを返す

打ち切りか否か（打ち切り=0）

6-MP薬の投与の有無

+有り→打ち切りデータ

生存関数とハザード関数

- 生存時間・・・イベントが観測されるまでの時間T（確率変数）
- 生存関数・・・イベントがある時点tまで生起していない関数S(t)
- 確率密度関数・・・S（t）の微分f（t）
- 累積確率分布関数・・・F（t）

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t+\Delta t)}{\Delta t}$$

死亡時点

ものすごく小さい時間

- ハザード関数・・・t時まで生存した条件下で次の時刻に死亡する瞬間死亡率h（t）

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

時刻tのハザード（危険度）

- 累積ハザード関数をH（t）とするとS(t)とh（t）に下の関係が成り立つ

$$H(t) = \int_0^t h(t)dt = logS(t)$$

☆S(t)かh（t）のどちらかが求まれば、もう片方も求まる。
☆ハザード関数が瞬間時間での死亡の危険度を表現しているのに対してF(t)やS(t)はあるひとつの個体が時刻tまで時間が経過したときの死亡/生存確率を表す。

ノンパラメトリックモデル

- 確率分布の仮定をせず、時間以外の共変量を導入しない方法。2種類ある。

①経験分布による推定法

打ち切りなしの完全データのみ対象。カプランマイヤー法の特殊ケース。 $\hat{S}(t) = 1 - \mathrm{F(t)}$

S(t)の推定値

カプランマイヤー推定法・・・条件つき確立の考え方に基づく手法

$$\hat{\mathrm{S}}(\mathrm{t}) = \prod_{ti<t} \left(1 - \frac{di}{ri}\right)$$

ti時点での死亡数

リスクセットの数

i番目に大きな固有の生存時間

②ハザード関数による推定法

カンプラマイヤー法を用いて累積ハザード関数を推定。これを修正したものをネルソン（ネルソン-アーラン）推定量とよぶ。

関数survfit（１）

- survfit

  ノンパラメトリック法による生存時間を当てはめる関数

☆引数formulaをとる。formula→Surv(time , event)~groupという形式

打ち切りか否か

群分け
（説明変数）

☆引数typeを入れると推定法の指定可能（デフォルトはカプランマイヤー法）

・・・カプランマイヤー、フレミングハリントン、fh2　の三択

関数survfit(2)

```
> ge.sf<-survfit(Surv(time,cens)~treat,data=gehan)
> ge.sf
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)

                records n.max n.start events median 0.95LCL 0.95UCL
treat=6-MP           21    21       21      9     23      16      NA
treat=control        21    21       21     21      8       4      12
```

死亡者数

生存時間の中央値

両側の区間推定に関する情報

＜結果をsummaryする（一部省略）＞

```
> summary(ge.sf)
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
```

標準誤差

95%信頼区間の上下限値

プロットの作成

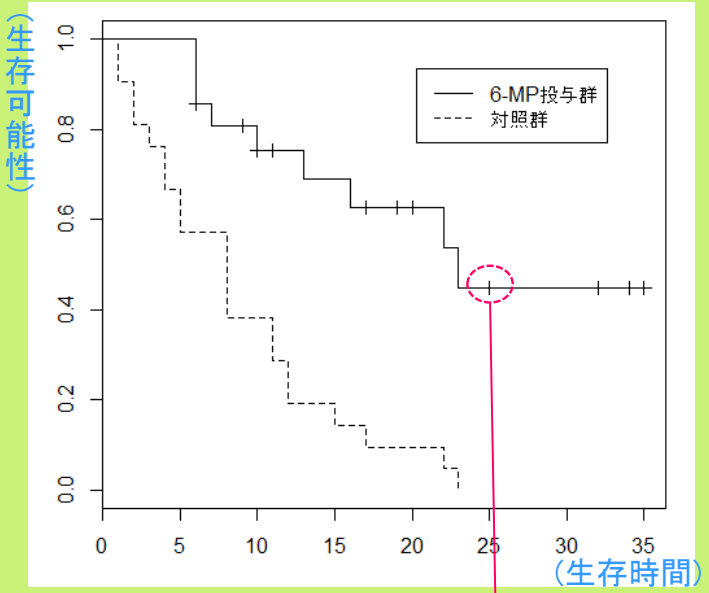
＜plot(散布図の作成)＞

```
> plot(ge.sf,lty=1:2)
> legend(locator(1),c("6-MP投与群","対照群"),lty=c(1,2))
```

凡例作成の関数

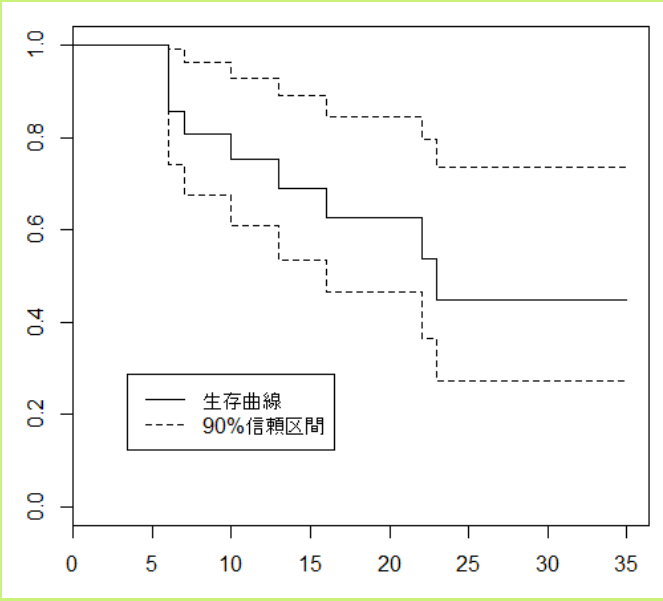
マウスで凡例の位置を指定する関数

線のタイプを指定（1実線、2点線）



打ち切りのある地点
（関数mark.t=F非表示）

☆引数conf.intを加えると信頼区間表示できる。



=の後ろ指定可
デフォルトは.95（95%信頼区間）

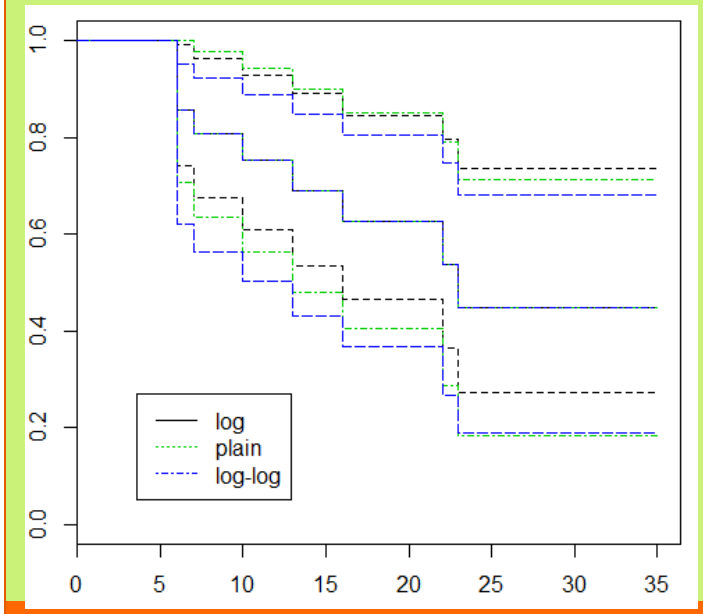
```
> ge2<-subset(gehan,treat=="6-MP")
> ge2.s<-survfit(Surv(time,cens)~treat,conf.int=.9,data=ge2)
> plot(ge2.s,mark.t=F)
> legend(locator(1),lty=c(1,2),legend=c("生存曲線","90%信頼区間"))
```

様々なプロット

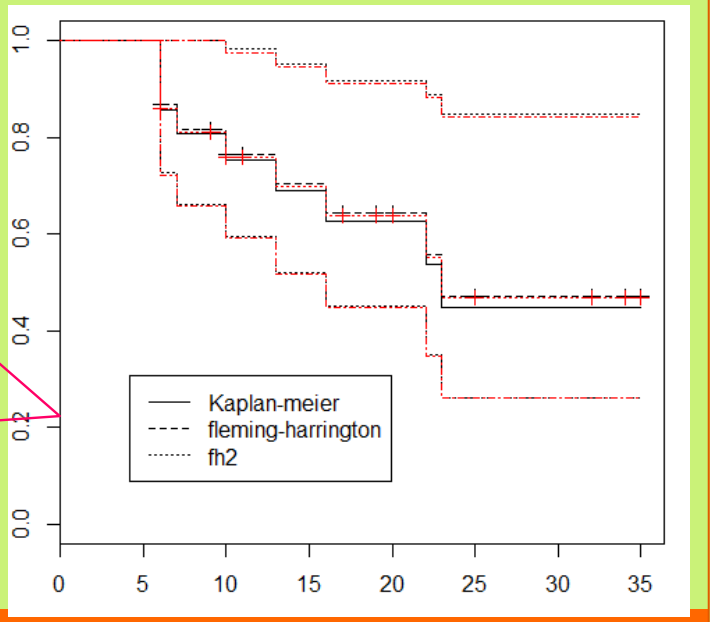
- 信頼区間の推定法を指定する関数conf.type
  →none, plain, log, log-log から一つ選択（デフォルトはlog）

```
> plot(ge2.s,conf.int=TRUE,mark.t=F)
> lines(survfit(Surv(time,cens)~treat,
+ data=ge2,conf.type="plain"),mark.t=F,conf.int=TRUE,lty=3,col=3)
> lines(survfit(Surv(time,cens)~treat,
+ data=ge2,conf.type="log-log"),mark.t=F,conf.int=TRUE,lty=4,col=4)
> legend(locator(1),c("log","plain","log-log"),lty=c(1,3,4),col=c(1,3,4))
```

線の色



Typeを指定すると、
あらゆる推定量で返したプロットが作成できる。



生存関数の検定

- ログ・ランク検定・・・郡ごとのイベントをありとなしに集計したクロス表のカイ二乗値の検定値を検定統計量とする。
- 関数survdiff ： G-rhoファミリの検定。引数rho=0でログランク、1でゲーハン―ウィルコクソン検定を行う。

```
> survdiff(Surv(time)~treat,data=gehan)
Call:
survdiff(formula = Surv(time) ~ treat, data = gehan)

                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP     21       21     29.2      2.31      8.97
treat=control  21       21     12.8      5.27      8.97

 Chisq= 9  on 1 degrees of freedom, p= 0.00275
```

P値＝0．00275なので、有意水準5%を基準とすると、両群（投薬群と対照群）の生存曲線には有意な差が認められる

セミノンパラメトリックモデル

- 確率分布の仮定をせず、時間以外の共変量を導入する方法。コックス比例ハザードモデルが一般的。

コックス比例ハザードモデル

$$\text{定義:}\quad h(t|x) = h_0(t)\exp(\beta x) = h_0(t)\exp\left(\sum_{i=1}^{m}\beta_i x_i\right)$$

→生存時間を目的変数とした回帰モデル。
　モデルのパラメータ$\beta = (\beta_1, \beta_2, \ldots, \beta_m)$は条件付き確率の部分尤度を最大化する方法で推定できる。

パラメータの推定

- 推定方法は直接法、ブレスローの近似法、エフロンの近似法がある。関数はcoxph.(引数formula：共変量の指定、method：推定法の指定（デフォルトはエフロン近似法））

```
> data(kidney)
> kidney.cox<-coxph(Surv(time,status)~sex+disease,data=kidney)
> summary(kidney.cox)
```

打ち切りか否か

病気の種類（GN,AN,PKD,Other）

```
coxph(formula = Surv(time, status) ~ sex + disease, data = kidney)

  n= 76, number of events= 58

             coef exp(coef) se(coef)      z Pr(>|z|)
sex       -1.4774    0.2282   0.3569 -4.140 3.48e-05 ***
diseaseGN  0.1392    1.1494   0.3635  0.383   0.7017
diseaseAN  0.4132    1.5116   0.3360  1.230   0.2188
diseasePKD -1.3671    0.2549   0.5889 -2.321   0.0203 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
sex           0.2282     4.3815   0.11339    0.4594
diseaseGN     1.1494     0.8700   0.56368    2.3437
diseaseAN     1.5116     0.6616   0.78245    2.9202
diseasePKD    0.2549     3.9238   0.08035    0.8084

Concordance= 0.696  (se = 0.045 )
Rsquare= 0.206   (max possible= 0.993 )
Likelihood ratio test= 17.56  on 4 df,   p=0.001501
Wald test            = 19.77  on 4 df,   p=0.0005533
Score (logrank) test = 19.97  on 4 df,   p=0.0005069
```

回帰係数

標準誤差

指数関数

尤度比の検定

ワイルド検定

スコア検定

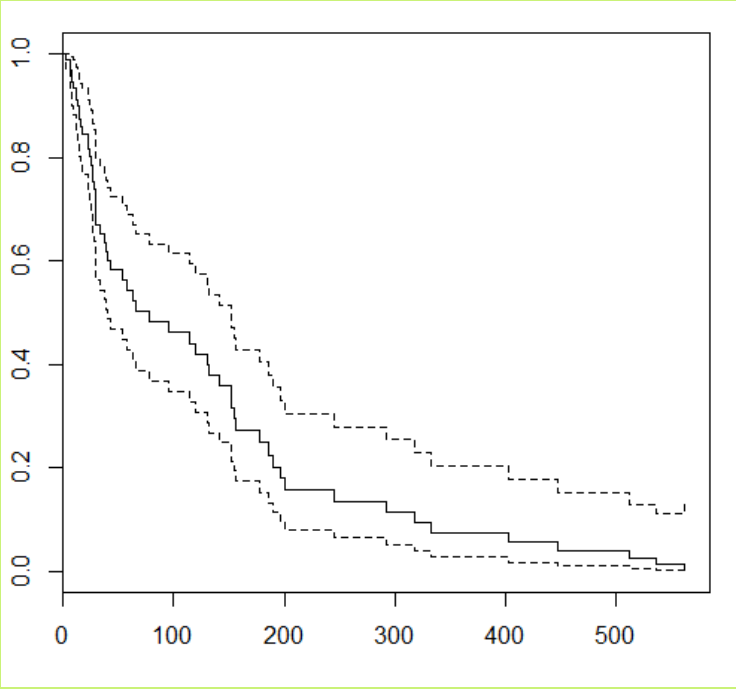
生存時間の推定

- 関数survfitを用いて生存時間を当てはめる

```
> kidney.fit<-survfit(kidney.cox)
> summary(kidney.fit)
Call: survfit(formula = kidney.cox)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     76       1  0.99018 0.00985      0.971059        1.000
    7     71       2  0.96856 0.01831      0.933335        1.000
    8     69       2  0.94641 0.02422      0.900102        0.995
    9     65       1  0.93499 0.02685      0.883820        0.989
   12     64       2  0.91106 0.03177      0.850869        0.976
   13     62       1  0.89844 0.03411      0.834007        0.968
   15     60       2  0.87273 0.03844      0.800550        0.951
   16     58       1  0.85959 0.04046      0.783848        0.943
   17     56       1  0.84591 0.04246      0.766654        0.933
   22     55       1  0.83144 0.04446      0.748706        0.923
   23     53       1  0.81625 0.04646      0.730084        0.913
   24     52       1  0.80113 0.04830      0.711843        0.902
   25     50       1  0.78592 0.05004      0.693710        0.890
   26     48       1  0.76976 0.05184      0.674573        0.878
   27     47       1  0.75369 0.05350      0.655802        0.866
```

〈推測されたkidney.fitの生存曲線〉



- 図示する

```
> plot(kidney.fit,mark.t=F)
```

残差分析

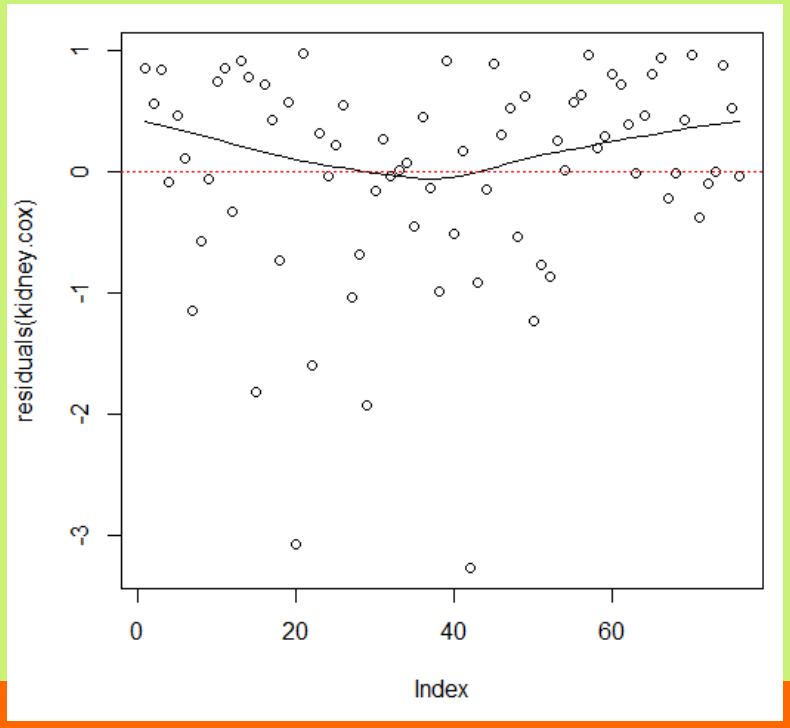
- マルチンゲール残差が一般的。関数residuals.coxph(residuals,resid)デフォルトはマルチンゲール残差。関数scatter.smoothで残差のプロットを作成。

```
> scatter.smooth(residuals(kidney.cox))
> abline(h=0,lty=3,col=2)
```

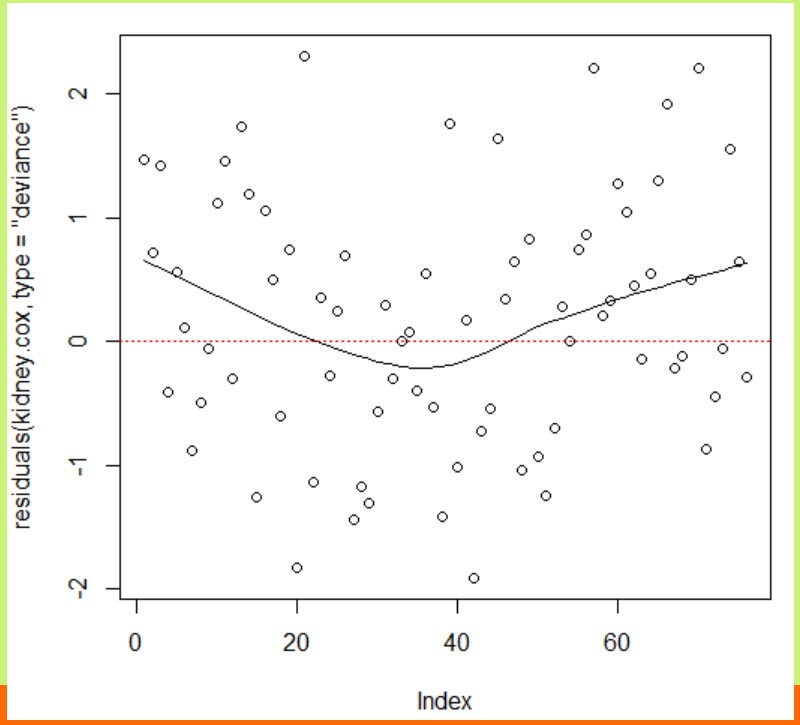
```
> scatter.smooth(residuals(kidney.cox,type="deviance"))
> abline(h=0,lty=3,col=2)
```

残差を指定

〈マルチンゲール残差〉



〈デヴァイアンス残差〉



ハザードの比例性の分析

- ハザード比が時間によらず一定である仮説を吟味する。関数cox.zph（シェーンフィールド残差を用いて仮説[$\beta(t) = \beta + \theta_g(t)$における$H_0: \theta = 0$]の検定に必要な統計量と標準化されたシェーンフィールド残差との相関関係を返す。）

- cox.zphの引数transformではg(t)を指定。（デフォルト：カプランマイヤー推定量）

```
> kidney.zph<-cox.zph(kidney.cox)
> kidney.zph
                rho    chisq     p
sex         0.18839  2.60676 0.106
diseaseGN  -0.01392  0.01123 0.916
diseaseAN   0.06162  0.20036 0.654
diseasePKD  0.00701  0.00438 0.947
GLOBAL           NA  4.20781 0.379
```

順位相関係数

カイ二乗検定の結果

☆仮説検定の結果はg(t)に依存

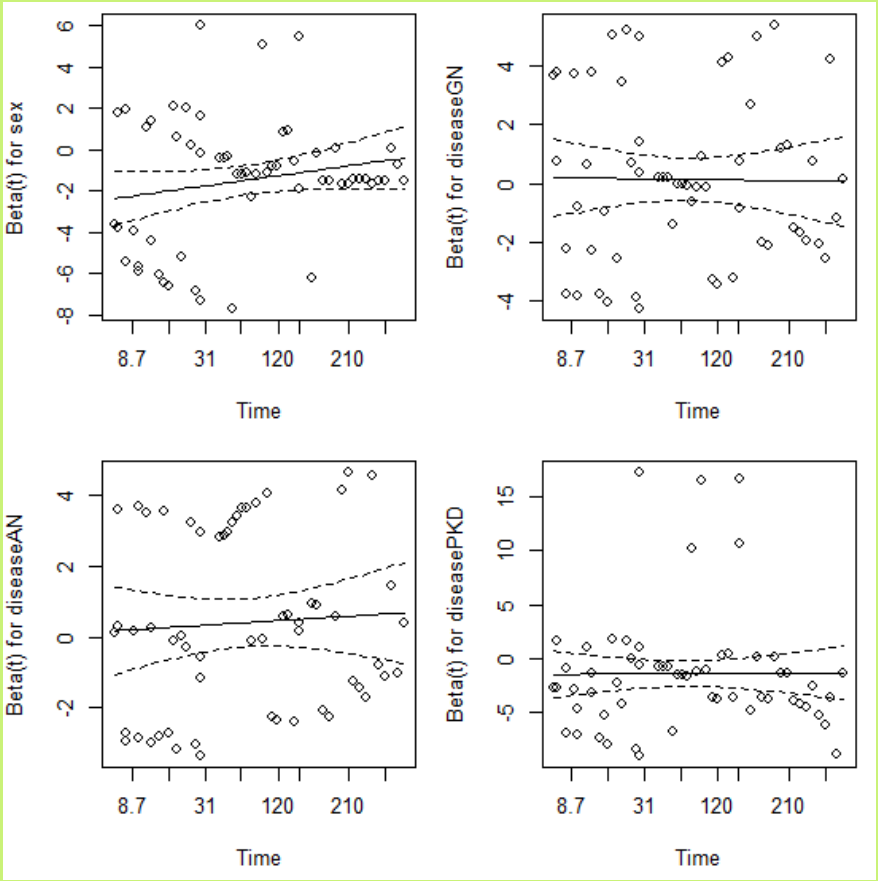
プロット化

- 変数ごとのシェーンフィールド残差プロットにスプライン平滑化（関数smooth.spline,自由度はdf=nで指定）された曲線と標準誤差の2倍の信頼区間が示されるグラフを作成する。

plotの実行　グラフの数を指定　グラフ間の余白を調整

```
> op<-par(mfrow=c(2,2),mar=c(4.5,4,1,1))
> plot(kidney.zph,df=2);par(op)
```

自由度



☆スプライン平滑化曲線の傾向に、時間に伴う明らかな変化パターンが見られなければ、比例ハザードの仮定に問題ない。

交互作用と変数の選択

- 説明変数の交互作用を取り入れたモデルの構築

```
> kidney.cox2<-coxph(Surv(time,status)~(sex+age+disease)^2,data=kidney)
> kidney.cox2
```

- あまり役に立たない（p値が高い）変数を取り除くAIC情報量を用いる関数stepAICを使用する

```
> library(MASS)
> stepAIC(kidney.cox2)
```

```
Start:  AIC=366.95
Surv(time, status) ~ (sex + age + disease)^2
```

〈・・・中略・・・〉

```
Step:  AIC=359.54
Surv(time, status) ~ sex + disease + sex:disease

              Df    AIC
<none>           359.54
- sex:disease  3 366.24
Call:
coxph(formula = Surv(time, status) ~ sex + disease + sex:disease,
    data = kidney)


                 coef exp(coef) se(coef)      z       p
sex            -2.525  8.00e-02    0.566 -4.462 8.1e-06
diseaseGN      -1.342  2.61e-01    1.304 -1.029 3.0e-01
diseaseAN      -1.418  2.42e-01    1.484 -0.956 3.4e-01
diseasePKD     -7.382  6.22e-04    1.949 -3.787 1.5e-04
sex:diseaseGN   0.831  2.30e+00    0.760  1.093 2.7e-01
sex:diseaseAN   1.075  2.93e+00    0.816  1.318 1.9e-01
sex:diseasePKD  4.214  6.77e+01    1.150  3.664 2.5e-04

Likelihood ratio test=30.3  on 7 df, p=8.5e-05  n= 76, number of events= 58
```

☆最適と判断された情報のみ返される

パラメトリックモデル

- 生存時間が確率分布に従うという仮定のもとで構築したモデル。この制約条件があるため応用範囲が限定。指数分布、ワイブル分布、対数正規分布、対数ロジスティック分布がよく使われる。

- 関数surverg（引数distで確率分布を指定：デフォルトはワイブル分布）

```
> survreg(Surv(time,status)~sex+disease,kidney,dist="lognormal")
Call:
survreg(formula = Surv(time, status) ~ sex + disease, data = kidney,
    dist = "lognormal")

Coefficients:
(Intercept)         sex   diseaseGN   diseaseAN  diseasePKD
  1.7923643   1.5062960  -0.3334601  -0.5321264   0.6810495

Scale= 1.129394

Loglik(model)= -329.1   Loglik(intercept only)= -340
        Chisq= 21.8 on 4 degrees of freedom, p= 0.00022
n= 76
```

対数尤度を返す→これを利用してAICを容易に求められる

指数分布；exponential

正規分布；gaussian

対数正規分布；log-normal

対数ロジスティック分布
；log-logistic

☆どの確率分布を用いるかは、AICを用いて評価可能。