サバイバルツリー法の改良と SAS/PHREGプロシジャによる実行

浜田 泉[1)], 川口 淳[2)]
株式会社イベリカCRD[1)], 久留米大学バイオ統計センター[2)]

Extended survival tree method and SAS/PHREG procedure

Izumi Hamada[1)], Atsushi Kawaguchi[2)]
IBERICA CRD Co.,Ltd[1)], Biostatistics Center, Kurume University[2)]

SAS ユーザー総会 アカデミア/テクノロジー&ソリューション セッション 2010

要旨:

生存時間と予後因子の関係を明らかにするための手法の一つであるサバイバルツリー法を改良し, その手法を適用するにあたりSAS/PHREGプロシジャを有効に用いた実際のプログラムを紹介する.

キーワード：サバイバルツリー, Cox比例ハザードモデル, PHREGプロシジャ, SBC, スコア検定統計量, 予後予測

Outline

- はじめに
- サバイバルツリーとは
- 改良方法
- 実データ適用例
- SAS/PHREGプロシジャを用いた処理
- まとめ
- 付録

はじめに

生存時間とその共変量
(性別, 年齢, 治療前の検査データなど)
の関係を明らかにしたい.

■ 生存時間に関連しているのはどの因子か
■ どれくらいの程度で効いているのか
⇒YとXの回帰問題
回帰関係を明らかにすることで新しくきた患者さんのリスクを予測したりするために用いる

それは生存時間の多変量モデル解析です
Cox比例ハザードモデルがあるじゃないか！！

はじめに

多変量解析，重回帰分析・・・・

$$\log \lambda(t \mid \mathbf{z}) = \lambda_0(t) + \beta_a AGE + \beta_s SEX + \beta_d DOSE + \beta_{sd} SEXLD$$

表 3.5 ステージ (STG)，ボールマン分類 (BORR)，手術時年齢 (AGE) およびそれらの交互作用を共変量とするステップワイズ Cox 回帰解析の結果

変数名	変数追加の χ^2 値	変数除去の χ^2 値	p 値	推定係数 (標準誤差)
STG	219.32		0.000	0.830 (0.068)
BORR	110.38		0.000	0.710 (0.068)
AGE	54.77		0.000	0.922 (0.126)
STG × AGE	18.12		0.000	−0.190 (0.045)
BORR × AGE	4.00		0.045	−0.114 (0.057)
STG × BORR	11.64		0.000	−0.098 (0.027)
STG × BORR × AGE		0.08	0.878	

0.83STG+0.71BORR+0.922AGE-0.19STG×AGE
-0.114BORR×AGE-0.098×STG×BORR ？？

例えば，生存時間のリスクグループを共変量の条件で分けて予後が良好，普通，悪いなどざっくり3グループに分けて解釈できないだろうか

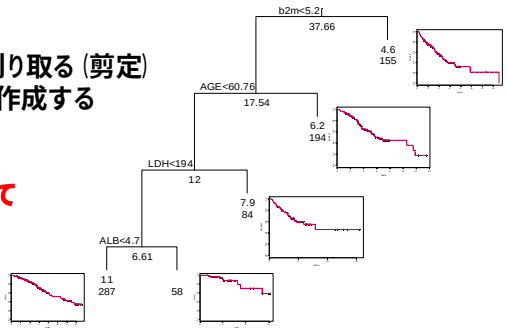
そういう場合サバイバルツリーは有用です

サバイバルツリーとは

目的変数と説明変数間の関係を樹形図であらわすツリーモデルに生存時間データを目的変数とする方法

Step1.木を育てる（分岐）
Step2.育ち過ぎた木を刈り取る（剪定）
Step3.リスクグループを作成する

ルートノードから各患者のプロファイルをたどっていくと予後予測ができる



サバイバルツリーとは

サバイバルツリーの利点と問題点

■利点

①共変量の測定尺度についての制約がない

②結果を樹形図として表すことにより得られた
リスクグループの説明が比較的容易
(予後因子の二値化もできる)

LeBlanc and Crowley (1992)

サバイバルツリーの方法に基づいたデータ解析を実施．骨髄腫の試験SWOG8229に適用

SAS ユーザー総会 アカデミア/テクノロジー&ソリューション セッション 2010

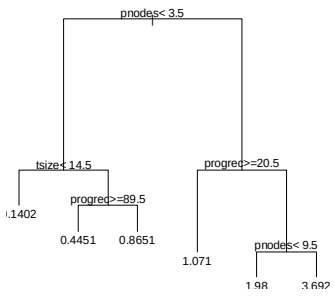
サバイバルツリーとは

■問題点

①SASでサバイバルツリー解析ができない・・・

②分岐に使用する共変量の出現頻度が偏る (場合もある)

変数名	説明	単変量cox p値
horTh	ホルモン治療の有無 (yesまたはno)	0.0036
age	年齢 (歳)	0.446
menostat	閉経状態 (pre=閉経前, post=閉経後)	0.5988
tsize	腫瘍サイズ (mm)	<.0001
tgrade	腫瘍グレード (I < II < III)	0.0001
pnodes	陽性結節数 (個)	<.0001
progrec	プロゲステロン受容体 (fmol)	<.0001
estrec	エストロゲン受容体 (fmol)	0.0411
time	生存時間 (日)	
cens	打ち切り変数 (0=打ち切り,1=イベント)	

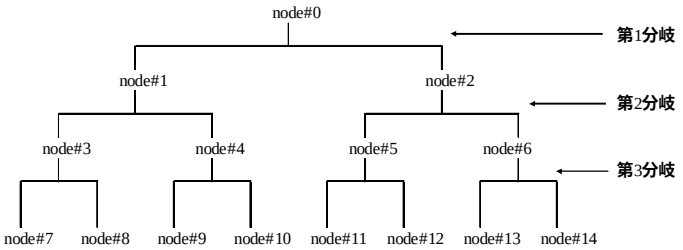


そこで、木の本数を増やしてみる ⇒ 改良方法「複数木作成」

改良方法（準備）

基本ツリー

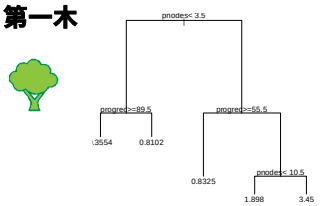
1. 第3分岐まで行う
2. ノードのnに含まれるイベント数が全体の5%未満になった時点で分岐を終了する



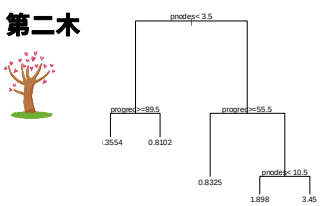
改良方法のアルゴリズム（例：二本まで）

1) 集団全員を対象に第一木を育てる.
2) 第一木を表すダミー変数を作成する（第一木モデル）.
3) ダミー変数を調整因子としてモデルに含ませ，集団全員を対象に第二木を育てる.
4) 第二木を表すダミー変数を作成する（第二木モデル）.
5) 得られた第一木モデルのSBC，第一木モデル＋第二木モデルのSBCをそれぞれ計算し，SBCが小さい方を最終モデルとして採用する.

第一木



第二木



$$\lambda(t \mid x) = \lambda_0(t)\exp(\beta x)$$

$$\lambda(t \mid x, \omega) = \lambda_0(t)\exp(\beta x + \gamma'\boldsymbol{\omega})$$

第一木を表したもの＝調整因子

改良方法 分岐 (比例ハザードモデル)

第一木

$$\lambda(t \mid x) = \lambda_0(t)\exp(\beta x)$$

二値化後変数

H0=β=0
のスコア検定統計量が一番大きい
ものから順に取り出し分岐

第二木

$$\lambda(t \mid x, \omega) = \lambda_0(t)\exp(\beta x + \gamma'\omega)$$

二値化後変数

第一木をダミー変数表示 (調整因子)

H0=β=0
のスコア検定統計量
＝調整済みスコア統計量
が一番大きいもの

第一木モデル

$$\lambda(t \mid \omega) = \lambda_0(t)\exp(\gamma_1\omega_1 + \gamma_2\omega_2 + \cdots + \gamma_k\omega_k)$$

第二木モデル

$$\lambda(t \mid \omega, \psi) = \lambda_0(t)\exp(\gamma_1\omega_1 + \gamma_2\omega_2 + \cdots + \gamma_k\omega_k + \eta_1\psi_1 + \eta_2\psi_2 + \cdots + \eta_m\psi_m)$$

実データの適用例

Rのipredパッケージに
含まれるGBSG2データ
(Schumacher et al, 1994)
を使用する.
(さっきと同じです)
686例の乳がん患者データ
を用いて得られた
グループ化とその条件を
示す.

変数名	説明	単変量cox p値
horTh	ホルモン治療の有無 (yesまたはno)	0.0036
age	年齢 (歳)	0.446
menostat	閉経状態 (pre=閉経前, post=閉経後)	0.5988
tsize	腫瘍サイズ (mm)	<.0001
tgrade	腫瘍グレード (I < II < III)	0.0001
pnodes	陽性結節数 (個)	<.0001
progrec	プロゲステロン受容体 (fmol)	<.0001
estrec	エストロゲン受容体 (fmol)	0.0411
time	生存時間 (日)	
cens	打ち切り変数 (0=打ち切り,1=イベント)	

データセットの一部⇒

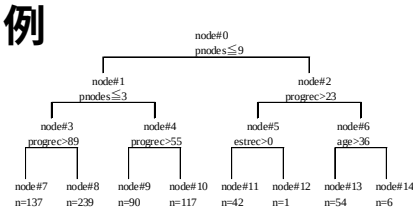
horTh	age	menostat	tsize	tgrade	pnodes	progrec	estrec	time	cens
no	70	Post	21	II	3	48	66	1814	1
yes	56	Post	12	II	7	61	77	2018	1
yes	58	Post	35	II	9	52	271	712	1
yes	59	Post	17	II	4	60	29	1807	1
no	73	Post	35	II	1	26	65	772	1
no	32	Pre	57	III	24	0	13	448	1
yes	59	Post	8	II	2	181	0	2172	0
no	65	Post	16	II	1	192	25	2161	0
no	80	Post	39	II	30	0	59	471	1
⋮									

686レコード

実データへの適用例

第一木

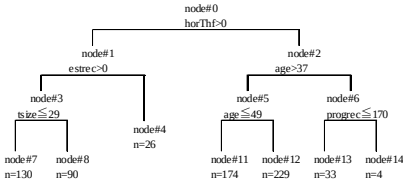
SBC=3468.799



第一+二木

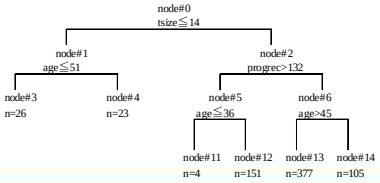
SBC=3453.013

二本木を採用



第一+二+三木

SBC=3468.451



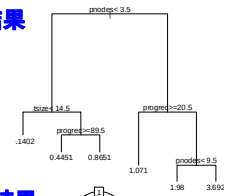
実データへの適用例 ー 1本と複数の比較

646 (日) をカットオフ値と設定し
(生存時間の中央値)
予後良い・悪い群に振り分けた場合

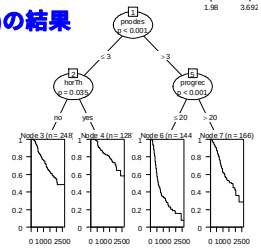
予後が悪いリスクグループの因子条件
pnode＞9のとき
1. progrec≦23
2. progrec＞23 かつ horTh≦0
pnode≦9のとき
1. ptogrec≦55 かつ horTh≦0
2. progrec≦89 かつ estrec≦0
3. progrec＞170 かつ age≦37 かつhorTh≦0

このデータ例での結果では一本木では出現しなかった
estrec, horThが予測条件に
組み込まれたことが確認できた.
これらは単変量解析においても有意だった因子
(estrecのp値= 0.0411, horThのp値= 0.0036) .

R(rpart)の結果



R(party)の結果

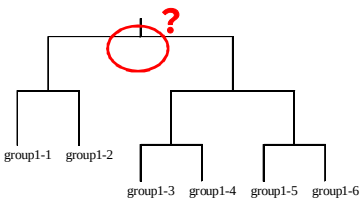


従来型サバイバルツリーでは・・・
死亡リスクの高いリスク因子の条件
はpnode>3 かつprogrec<=20

SAS/PHREGを用いた処理

何をしなければいけないか

まず一番最初の第一分岐はどの変数の
どのカットオフ値かを特定しなければならない



例えば・・・・・・・・	カットオフパターン数
性別：男，女	1
年齢：20～60歳	39
腫瘍のサイズ：0.1～5mm	100
腫瘍グレード：Ⅰ，Ⅱ，Ⅲ	2
血圧：	30
LED値：	50

=計200回くらいの2群検定をやって統計量が
一番大きいのが第一分岐のカットオフ

SAS/PHREGを用いた処理

分岐のための統計量

1群なら0,2群なら1という値をとるダミー変数を作成し
回帰モデルで2群検定を行うことができることを利用

分岐に使う最大のスコア検定統計量が一つのステップで求められる

```
PROC PHREG DATA = indata;
  MODEL time * cens(0) = すべてのダミー変数 /
  SLE = 0.99 SELECTION=FORWARD STOP =1 ;
RUN;
```

SAS/PHREGを用いた処理

分岐のための統計量　SASアウトプット

Note:　The model has reached STOP=1.

最尤推定量の分析						
変数	自由度	パラメータ推定	標準誤差	カイ 2 乗	Pr > ChiSq	ハザード比
_pnodes_9	1	1.08744	0.13653	63.4378	<.0001	2.967

変数増加法の要約				
ステップ	変数の追加	取り込んだ数	スコア カイ 2 乗	Pr > ChiSq
1	_pnodes_9	1	69.7148	<.0001

17

SAS/PHREGを用いた処理

2本目はどうしますか

プロシジャのオプションをうまく利用する

①第一木グループのk個のダミー変数を用意する

$\boldsymbol{\omega}=(\omega_1,\omega_2,\ldots\ldots,\omega_k)'$

②

```
PROC PHREG DATA = indata;
  MODEL time * cens(0) = ω + すべてのダミー変数 /
  SLE = 0.99 SELECTION=FORWARD STOP = k+1
  INCLUDE= k;
RUN;
```

18

SAS/PHREGを用いた処理

2本目 SASアウトプット

ステップ 1. 変数 _horTh1_1 を追加します。モデルは次の説明変数を含みます:
_tree1fn_1 _tree1fn_2 _tree1fn_3 _tree1fn_4 _tree1fn_5 _tree1fn_6 _tree1fn_7 _horTh1_1

Note: The model has reached STOP=8.

最尤推定量の分析

変数	自由度	パラメータ推定	標準誤差	カイ 2 乗	Pr > ChiSq	ハザード比
_tree1fn_1	1	-3.60650	0.46523	60.0955	<.0001	0.027
_tree1fn_2	1	-2.84701	0.43619	42.6021	<.0001	0.058
_tree1fn_3	1	-2.77909	0.45706	36.9708	<.0001	0.062
_tree1fn_4	1	-1.95955	0.43788	20.0265	<.0001	0.141
_tree1fn_5	1	-2.38794	0.47522	25.2500	<.0001	0.092
_tree1fn_6	1	0.51808	1.08645	0.2274	0.6335	1.679
_tree1fn_7	1	-1.23340	0.44424	7.7087	0.0055	0.291
_horTh1_1	1	-0.36213	0.12604	8.2552	0.0041	0.696

変数増加法の要約

ステップ	変数の追加	取り込んだ数	スコア カイ 2 乗	Pr > ChiSq
1	_horTh1_1	8	8.3409	0.0039

SAS/PHREGを用いた処理

本数を決める

第一木モデル, 第二木モデル及び第三木から
それぞれの情報量基準SBC
(Schwarz 1978; Judge et al. 1980)**より**
小さいほうを最良モデルとして木の本数を決定する.

MODEL time * cens(0) = **第一木グループダミー変数**;
MODEL time * cens(0) = **第一+二木グループダミー変数**;
MODEL time * cens(0) = **第一+二+三木グループダミー変数**;

モデルの適合度統計量

基準	共変量なし	共変量あり
-2 LOG L	3576.346	3428.896
AIC	3576.346	3442.896
SBC	3576.346	3468.799

まとめ

- サバイバルツリーは見た目に解釈しやすい有用な方法
- しかし因子の抽出が十分ではない⇒改良方法を提案
- SASのPHREGプロシジャをうまく用いることで分岐特定や本数決定が容易にできる

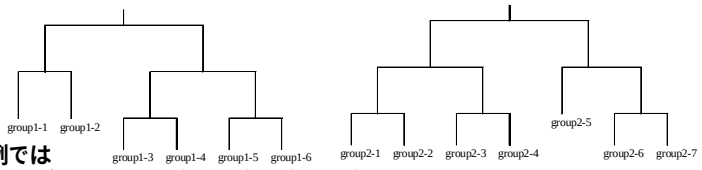
今後の課題など

- 改良方法に関して
  シミュレーションによる予測精度の評価（目下検討中）
- データ解析に関して
  複数木作成による結果のグループ化はやや面倒な作業でもあるのでツール化できれば実用性も向上する

付録

得られた樹形図に基づくグループ化

第一木グループ及び第二木グループの全ての組み合わせにより最終グループとする．



この例では
第一木モデルでの最終グループ：6グループ
第二木モデルでの最終グループ：6×7=42グループ

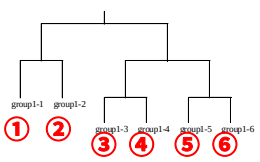
予測方法

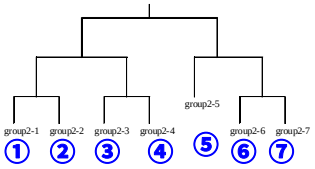
ある対象患者Lさんが属するグループの生存時間（カプランマイヤー法）に基づいて予測を行う．
最終グループiの$S_i(t)$を求め
$S_i(t_0) < 0.5$ →悪い
$S_i(t_0) \geqq 0.5$ →良い　※t_0は研究の目的によって自由に与えられる

付録

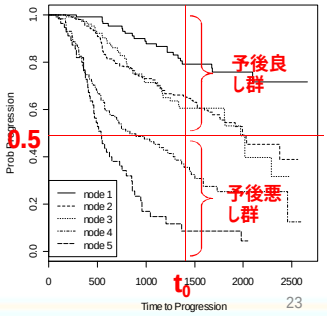
※第一木と第二木のターミナルノードからグループ化を行うイメージ

第一木　　第二木





group1=① and ①
group2=① and ②
group3=① and ③
：
group6×7=42



付録 二値化変数作成サンプルプログラム（ひとつの変数例）

```
/*値のバリエーションを抽出*/
proc freq data = data1 noprint;
  by keyall;
  tables &var. / out = freq data2;
run;
/*横にする*/
proc transpose data = data2
                    out = data3 prefix = x&var.x;
  by keyall;
  var &var.;
run;
/*バリエーション数をマクロ変数に格納*/
proc contents data = data2(keep =  x&var.:)
              out  = _temp noprint;
run;
data _null_;
  set _temp end = eof;
  if eof then
  call symput("OBS",compress(put(_N_,best.)));
run;
```

```
/*2群フラグをカットオフあるだけ作成*/
data temp3;
  merge data1 data3;
  by keyall;

  array col{*} x&var.:;
  array tempf{*} _&var._1-_&var._&obs.;

  do i=1 to dim(tempf);
    if       . < &var. <= col{i} then tempf{i}=0;
    if col{i} < &var.            then tempf{i}=1;
  end;
  drop x&var.:;
run;
```

付録 PROC PHREGのモデルステートメントの右辺に指定する変数文字列の作り方のサンプルプログラム

```
/*全変数情報を取得*/
proc contents data = master
              out  = data2(keep = varnum name) noprint;
run;

proc sort data = data2;
  by varnum;
run;

/*&varlistマクロ変数にブランク区切りで変数の文字列を作る*/
proc sql noprint;
  select name
  into: varlist separated by ' '
  from data2
run;
quit;

%put &varlist.;
```


参考文献

[1] LeBlanc, M., and Crowley, J., Relative risk trees for censored survival data, Biometrics (1992).

[2] M. Mizumoto, H. Harada, H. Asakura, T. Hashimoto, K. Furutani, H. Hashii, T. Takagi, H. Katagiri, M. Takahashi, T. Nishimura, Prognostic Factors and a Scoring System for Survival After Radiotherapy for Metastases to the Spinal Column, Wiley (2008).

[3] M. Radespiel-Troger, T. Rabenstein, H.T. Schneider, B. Lausen, Comparison of tree-based methods for prognostic stratification of survival data, Artificial Intelligence in Medicine (2003).

[4] XIAOGANG SU, CHIH-LING TSAI, Tree-augmented Cox proportional hazards models, Biostatistics (2005).

[5] Clinical Trials in Oncology, Second Edition. Green, S., Benedetti, J., Crowley, J. CRC Press, LLC (2003).

[6] 生存時間解析－SASによる生物統計．大橋靖雄，浜田知久馬．東京大学出版会．

[7] Cox比例ハザードモデル．中村剛．朝倉書店．