骨髄移植データに関する
イベントヒストリー解析

大阪電気通信大学　大学院　情報工学専攻

辻谷研究室

中井崇人

目次

1. はじめに

・白血病は骨髄のガン化が原因
・他人の骨髄を体内にいれるのでさまざまな問題がある
・患者が知りたいはどのくらい生きれるか
→1年後の生存確率

2. Multi-stateモデル

骨髄移植　完治しやすいパターン

血小板回復：手術が成功して骨髄が正常に機能

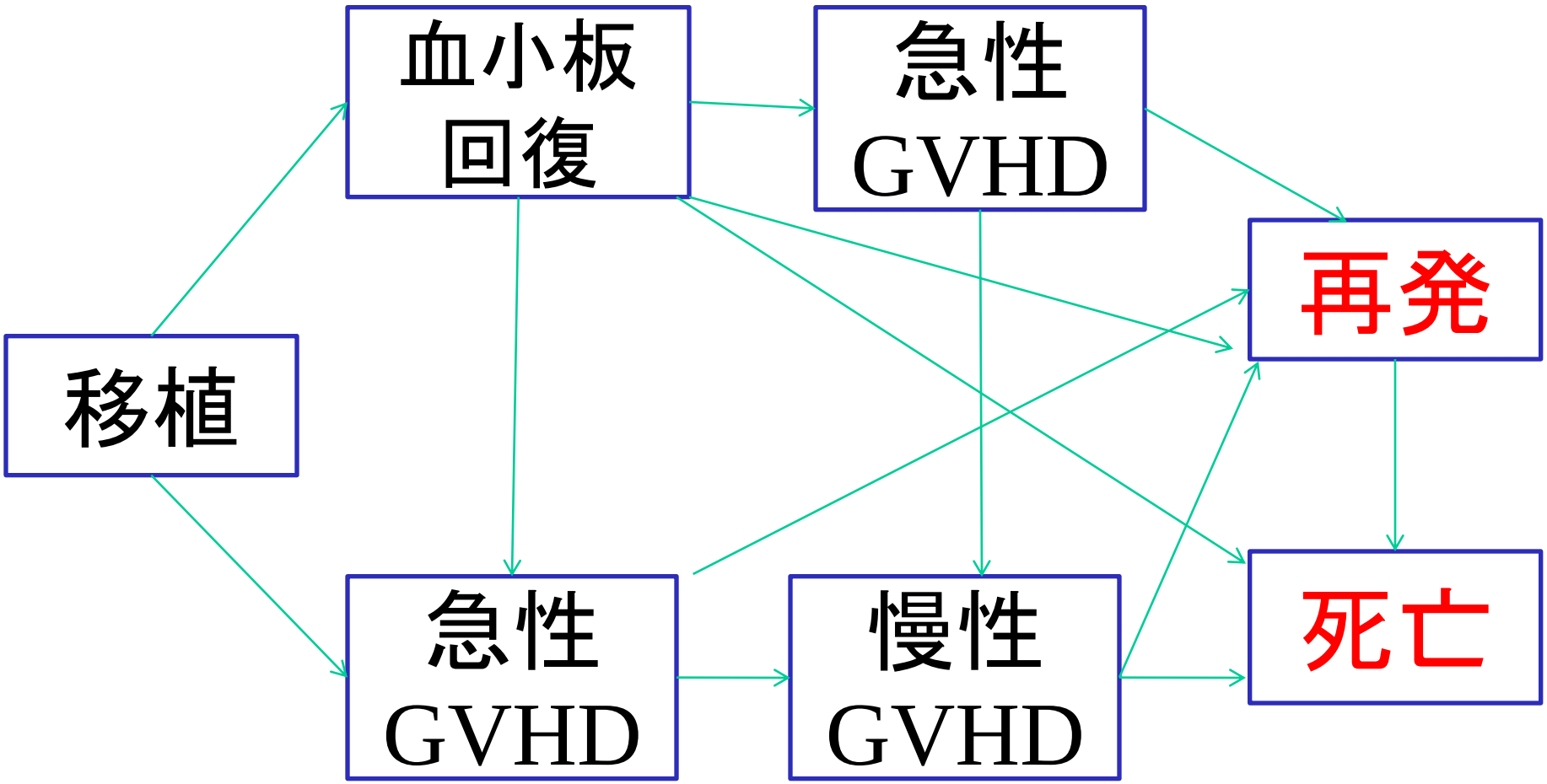
↓

移植片対宿主病（graft versus host disease通称GVHD）：
移植された臓器（血液）が周りの臓器を攻撃

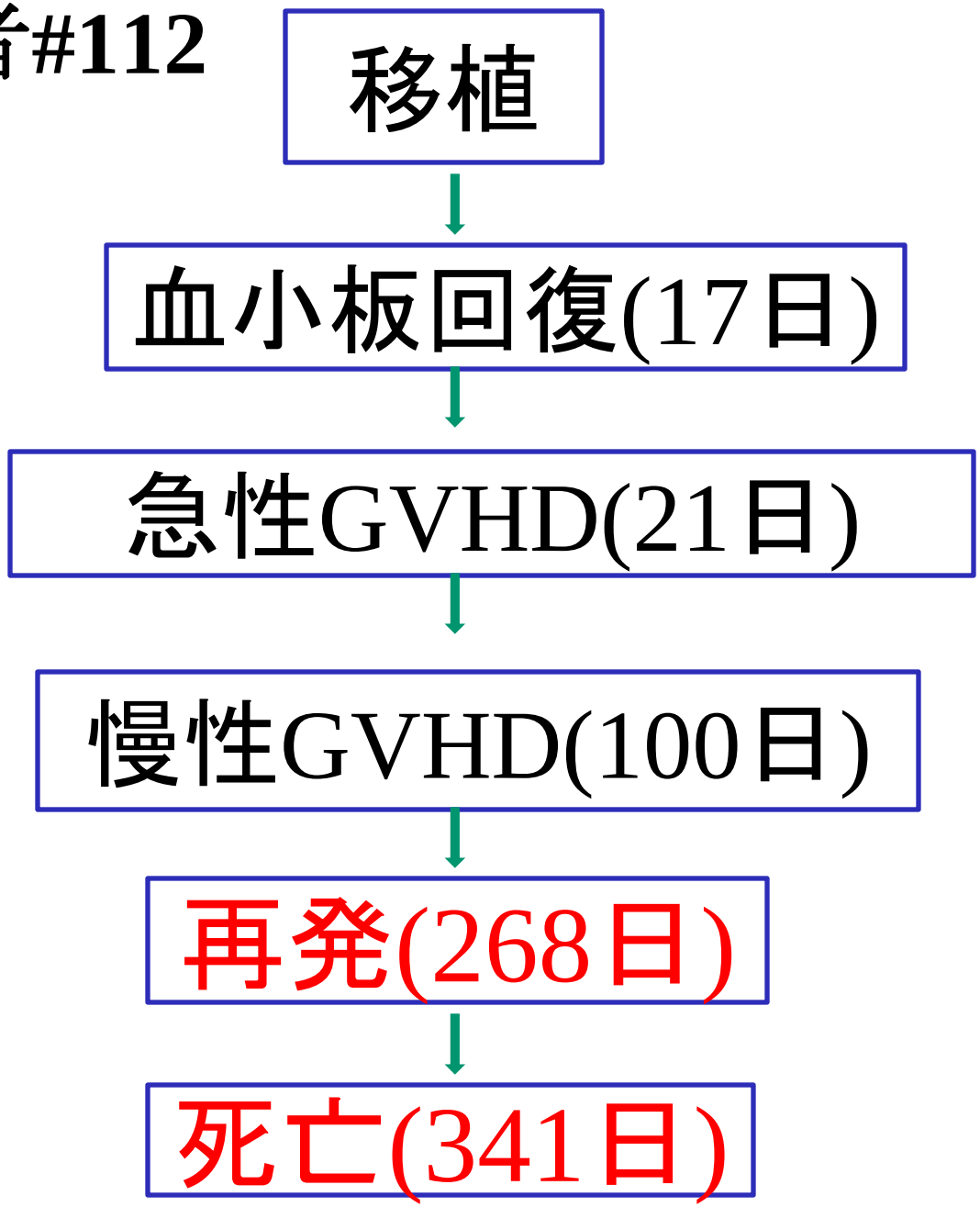
↓

完治

様々なパターン

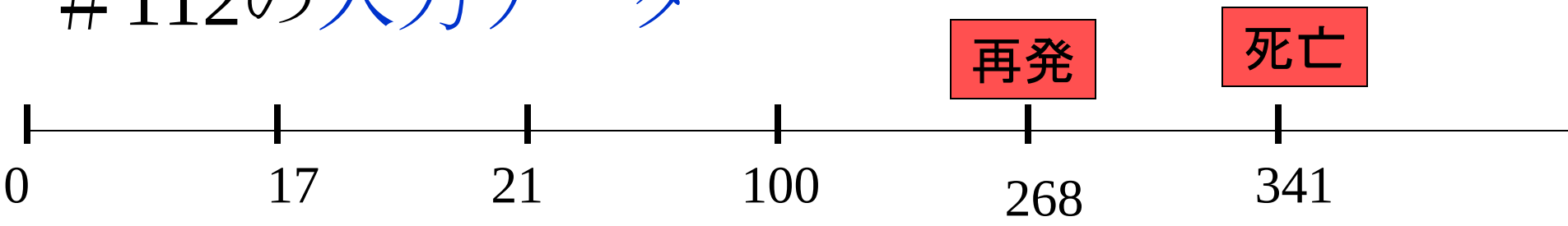


●患者#112



Cens:(1-死亡,0-打切り)
Time：発生日数(時間依存型)
Delta3=1:再発(時間依存型)
Za=1:急性GVHD発症(時間依存型)
Zc=1:慢性GVHD発症(時間依存型)
Zp=1:血小板回復(時間依存型)
Z1:患者年齢
Z2:ドナー年齢
Z3:患者の性別(1-男性,0-女性)
Z4:ドナーの性別(1-男性,0-女性)
Z5:患者のｻｲﾄﾒｶﾞﾛｳｲﾙｽ(CMV)の免疫状態(1-陽性,0-陰性)
Z6:ドナーのCMV (1-陽性,0-陰性)
Z7:移植までの待時間
Z8:French-American-Britih(FAB)の分類
(1-FAB分類で4か5でAML症状,0-それ以外の症状)
Z9:病院(1-OSU,2-AH,3-SVH,4-HU)
Z10:GVHD防止薬としてMTXの使用(1-使用,2-使用しなかった)
g:病状グループ(1-ALL,2-AML低ﾘｽｸ,3-AML高ﾘｽｸ)

＃112の入力データ



時間区間	生存時間	再発の有無
1	17	0
2	21	0
3	100	0
4	268	1
5	341	1

時間依存型

＃112の入力データ

時間依存型

Cens	Time	Delta3	Za	Zc	Zp	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	g
0	17	0	0	0	1	20	23	0	1	1	1	180	1	1	0	3
0	21	0	1	0	1	20	23	0	1	1	1	180	1	1	0	3
0	100	0	1	1	1	20	23	0	1	1	1	180	1	1	0	3
0	268	1	1	1	1	20	23	0	1	1	1	180	1	1	0	3
1	341	1	1	1	1	20	23	0	1	1	1	180	1	1	0	3

3. 平滑化スプライン(一般化加法モデル:GAM)

再発

$$\ln\left(\frac{p}{1-p}\right) = c_0 + c_1 x_1 + c_2 x_2 + \cdots$$

$$+ \frac{1}{12}\sum_{d=1}^{n} \theta_d \left| x_{20} - x_{20,d} \right|^3$$

$s(Time)$

ペナルティ付き残差平方和

平滑化パラメータ

$$\sum_{i=1}^{n}\left[y_i - s(x_i)\right]^2 + \lambda \int \left\{ s''(x) \right\}^2 dx$$

$s(x)$の曲率

小さいほどモデルの
当てはまりは良い

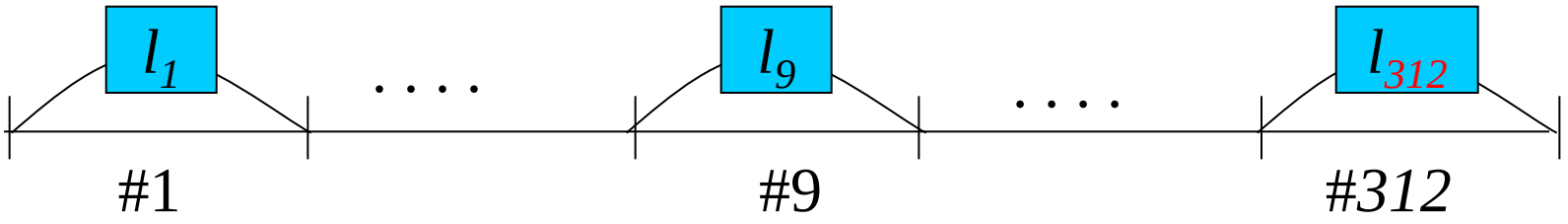
小さいほど滑らかな曲線
（曲げ弾性エネルギー）

⇒ 最小にするスプライン関数

平滑化スプライン

$$y = s(x) = c_0 + c_1 x + \frac{1}{12}\sum_{d=1}^{n} \theta_d \left| x - x_d \right|^3$$

平滑化スプラインλ_iの最適選択
■ 変形n-重CV法



初期標本 $\boldsymbol{X} = \left\{ \boldsymbol{X}^{\langle 1 \rangle}, \boldsymbol{X}^{\langle 2 \rangle}, \mathrm{K}, \boldsymbol{X}^{\langle n \rangle} \right\}, \boldsymbol{X}^{\langle d \rangle} = \left\{ \boldsymbol{x}^{\langle d \rangle}; d^{\langle d \rangle} \right\},$

$$\boldsymbol{x}^{\langle d \rangle} = \left\{ x_1^{\langle d \rangle}, \mathrm{L}, x_I^{\langle d \rangle} \right\}$$

訓練標本 $\boldsymbol{X}_{[d]} = \left\{ \boldsymbol{X}^{\langle 1 \rangle}, \boldsymbol{X}^{\langle 2 \rangle}, \mathrm{K}, \boldsymbol{X}^{\langle d-1 \rangle}, \boldsymbol{X}^{\langle d+1 \rangle}, \mathrm{K}, \boldsymbol{X}^{\langle n \rangle} \right\}$

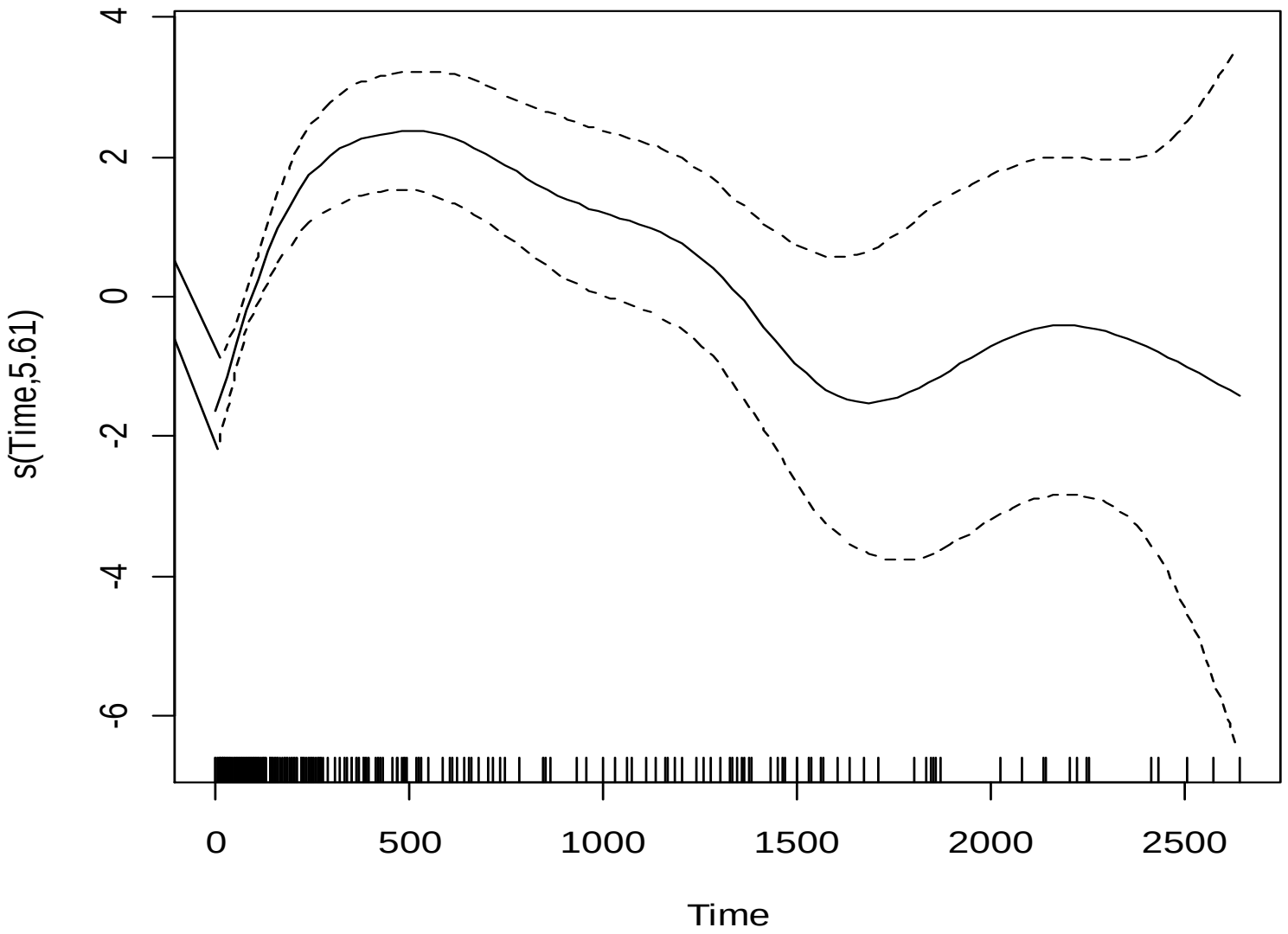
: $\boldsymbol{X}$の第d番目を除去

$$CV\text{規準}: -2\ln L = -2\sum_{d=1}^{n}\sum_{l=1}^{l_d} d_l^{[d]} \ln h_l^{[d]}\left(\boldsymbol{x}_l^{[d]}\right) + \left(1 - d_l^{[d]}\right)\ln\left\{1 - h_l^{[d]}\left(\boldsymbol{x}_l^{[d]}\right)\right\}$$

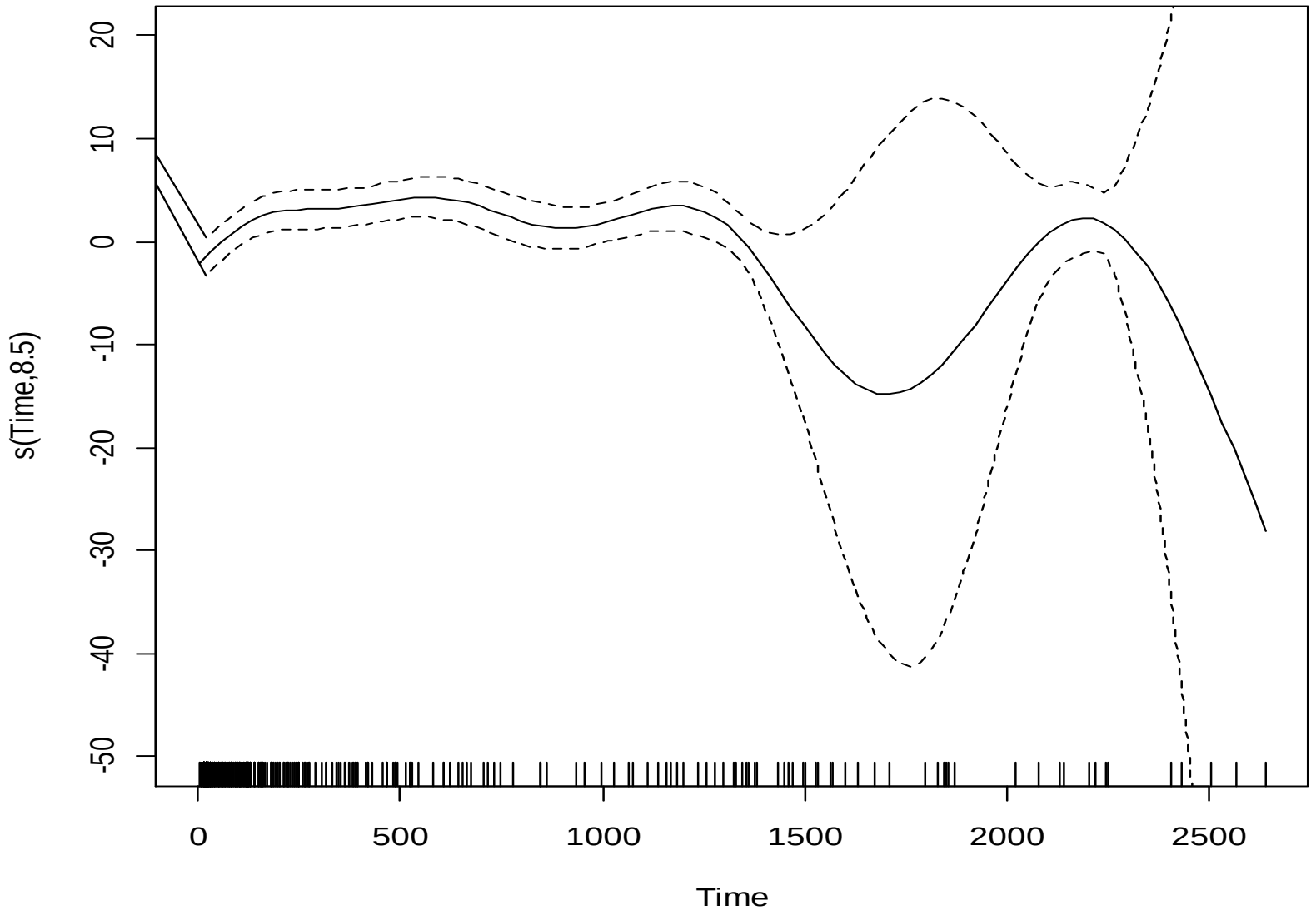
CV規準は、TICと漸近同等(Stone(1977),Shibata(1989))

共変量の有意性検定

共変量	回帰係数	p値
再発の有無	0.635	0.0794
患者年齢	-1.078	0.0065
ドナー年齢	0.1876	0.0199
血小板回復	-3.0516	0.0294
慢性GVHD	-1.0776	0.0065
血小板回復×患者年齢	0.2366	0.0237
血小板回復×ドナー年齢	-0.2003	0.0185

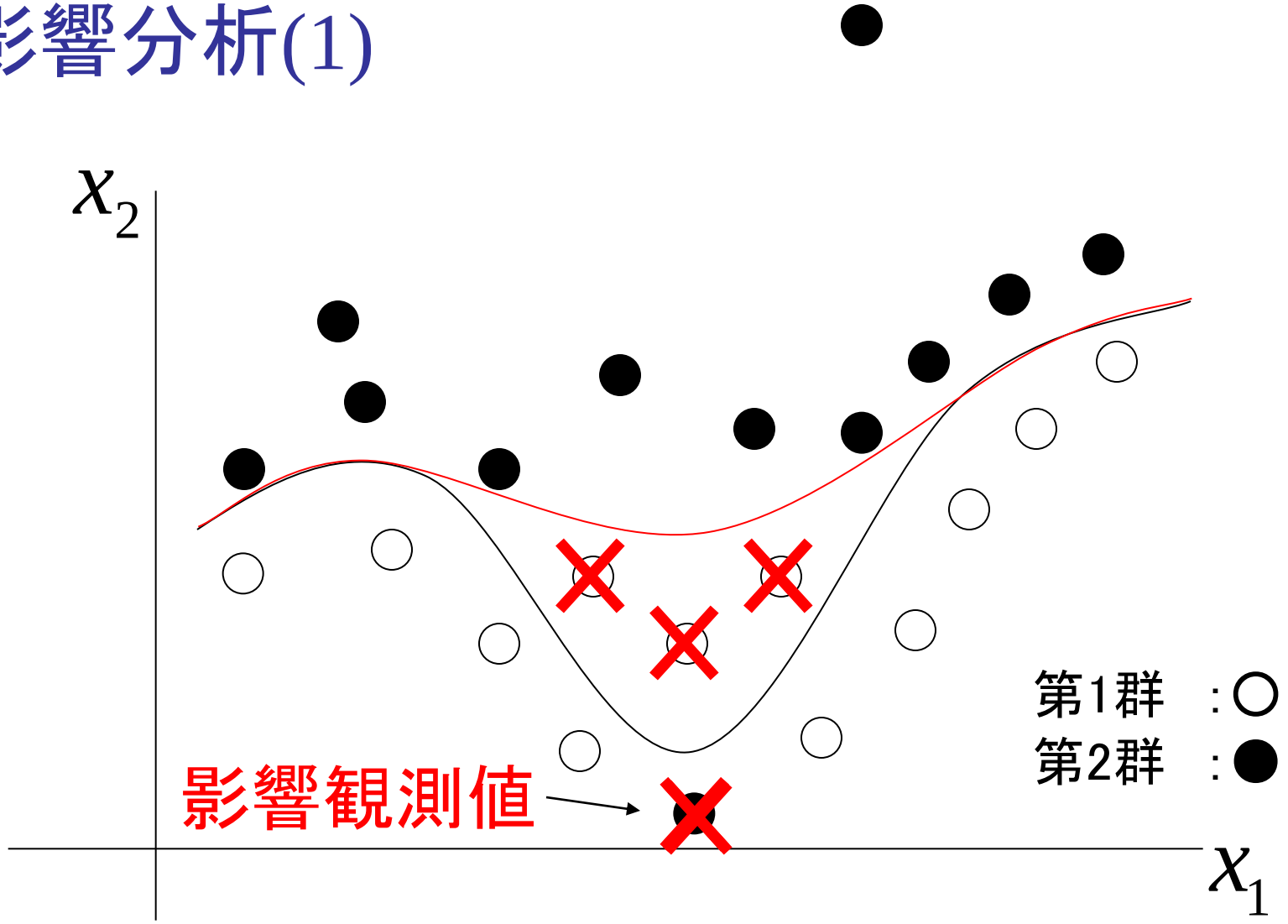


全データを用いた場合のTime(発生日数)のスプラインの図



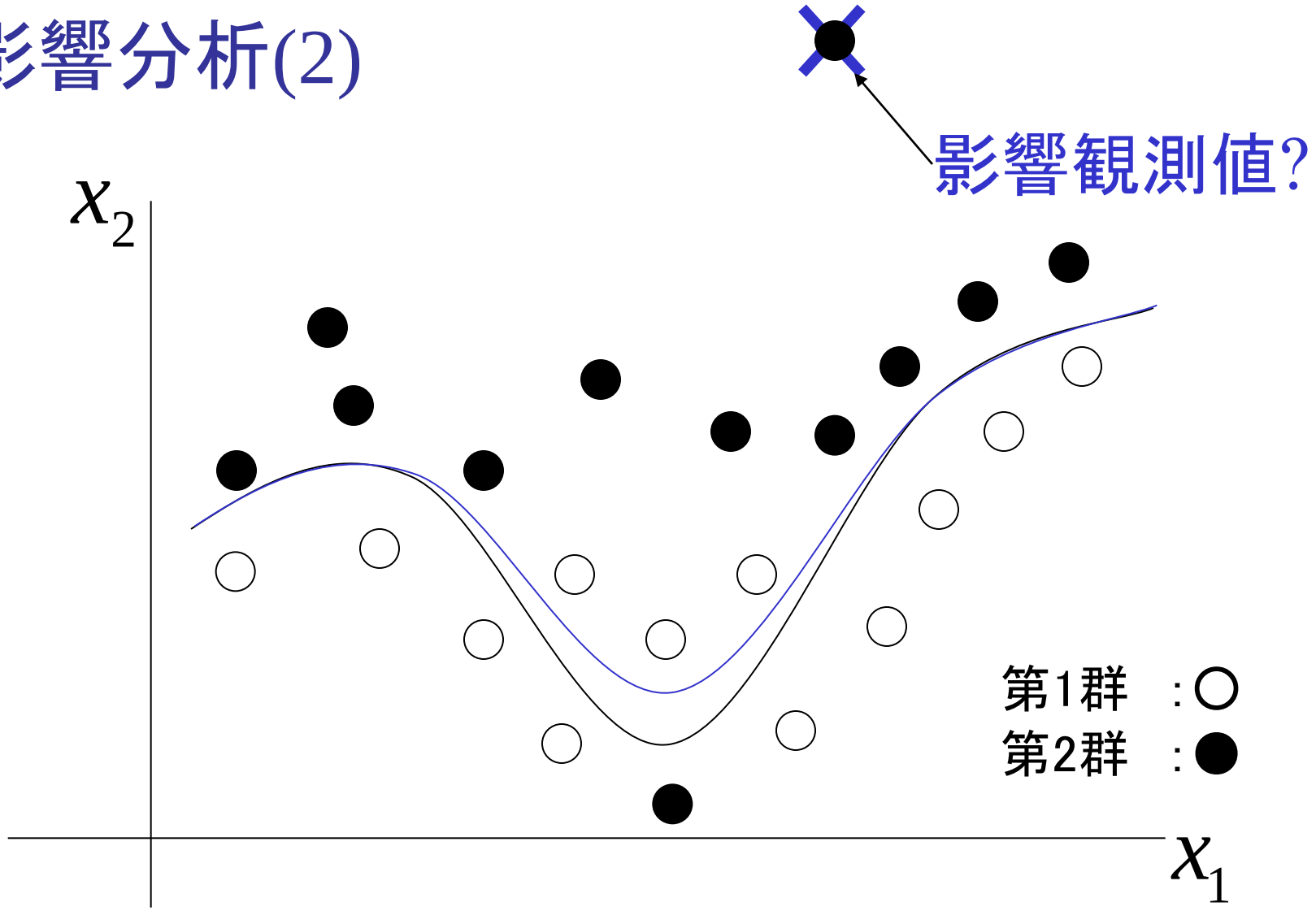
GCVを採用した場合のTimeのスプラインの図

影響分析(1)



曲線は簡単になり、誤判別は少なくなる

影響分析(2)



曲線はあまり変わらず、誤判別も変わらない

DIFDEV

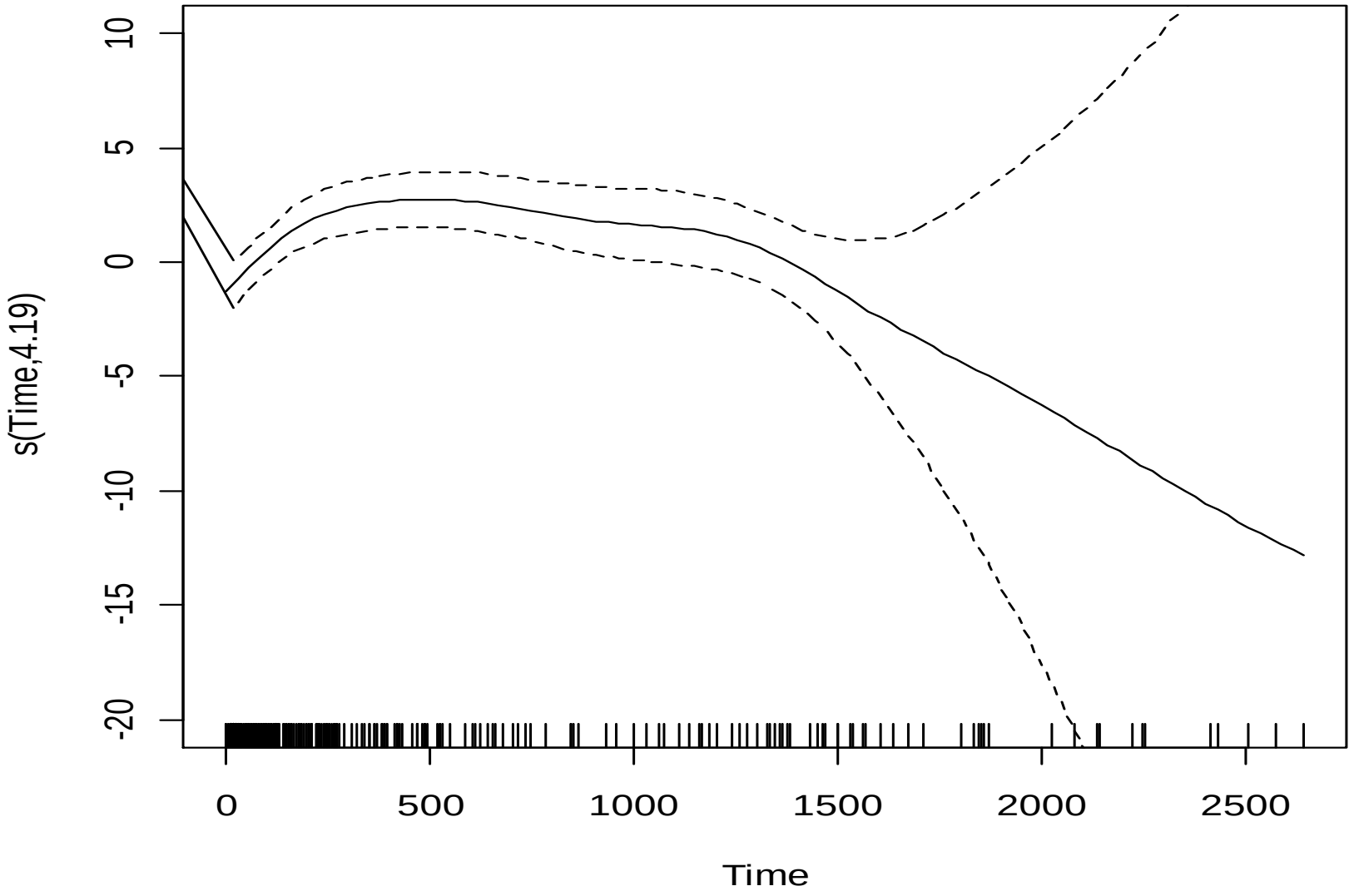
$$\Delta Dev_{[d]} = Dev - Dev_{[d]} \geq 0$$
$$\Delta df_{[d]} = df - df_{[d]}$$

$Dev(df)$: すべての個体を用いたときの逸脱度(残差自由度)

$Dev_{[d]}(df_{[d]})$: d番目の個体を取り除いたときの逸脱度(残差自由度)

- **検定の方法**

$$\Delta Dev_{[d]} \sim \chi^2_{\Delta df_{[d]}}$$



#69(12日目に血小板回復;2204日目に死亡)を除去

#69除去後

共変量	回帰係数 (除去前)	p値 (除去前)
再発の有無	0.6458(0.6353)	0.0759(0.0794)
患者年齢	-0.2189(-0.2174)	0.0303(0.0312)
ドナー年齢	0.1890(0.1876)	0.0193(0.0199)
血小板回復	-3.0516(-3.052)	0.0294(0.0294)
慢性GVHD	-1.047(-1.0776)	0.0091(0.0065)
血小板回復×患者年齢	0.2374(0.2366)	0.0235(0.0237)
血小板回復×ドナー年齢	-0.1992(-0.2003)	0.0195(0.0185)

スプライン効果の有意性検定

平滑化スプライン（3次の自然スプライン）

$$y = s(x) = c_0 + c_1 x + \frac{1}{12}\sum_{d=1}^{n} \theta_d |x - x_d|^3$$

$\Rightarrow$ 尤度比検定

$$\Delta = Dev(0) - Dev(1) \sim \text{自由度}(\nu_0 - \nu_1)\text{の}\chi^2\text{分布}$$

$Dev(0)$：もとのモデルの逸脱度(残差自由度ν_0)

$Dev(1)$：帰無仮説のもとでの逸脱度(残差自由度ν_1)

時点のｽﾌﾟﾗｲﾝ効果の有意性検定

共変量	尤度比検定統計量	*d.f.*	p値
時点のｽﾌﾟﾗｲﾝ効果	34.153†	3.191◇	<<0.0001

†$34.153 = 317.822 - 262.022$

◇$3.191 = 372 - 368.809$

適合度検定

逸脱度(*Dev*iance):モデル適合度の評価

$$Dev = 2\left[\ln L(\max) - \ln L\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}\right)\right]$$
$$= -2\ln L\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}\right) \sim \chi^2 \text{ ?}$$

ブートストラップ法による棄却点の算出

手順0 初期標本 : $\boldsymbol{X} = \left\{ \boldsymbol{X}^{<1>}, \boldsymbol{X}^{<2>}, \ldots, \boldsymbol{X}^{<d>}, \ldots, \boldsymbol{X}^{<D>} \right\}$

手順1 ブートストラップ標本

$$\boldsymbol{X}_b^* = \left\{ \boldsymbol{X}_b^{<1>*}, \boldsymbol{X}_b^{<2>*}, \ldots, \boldsymbol{X}_b^{<d>*}, \ldots, \boldsymbol{X}_b^{<D>*} \right\}$$

の生成$(b = 1, \ldots, 400)$

手順2 逸脱度の計算

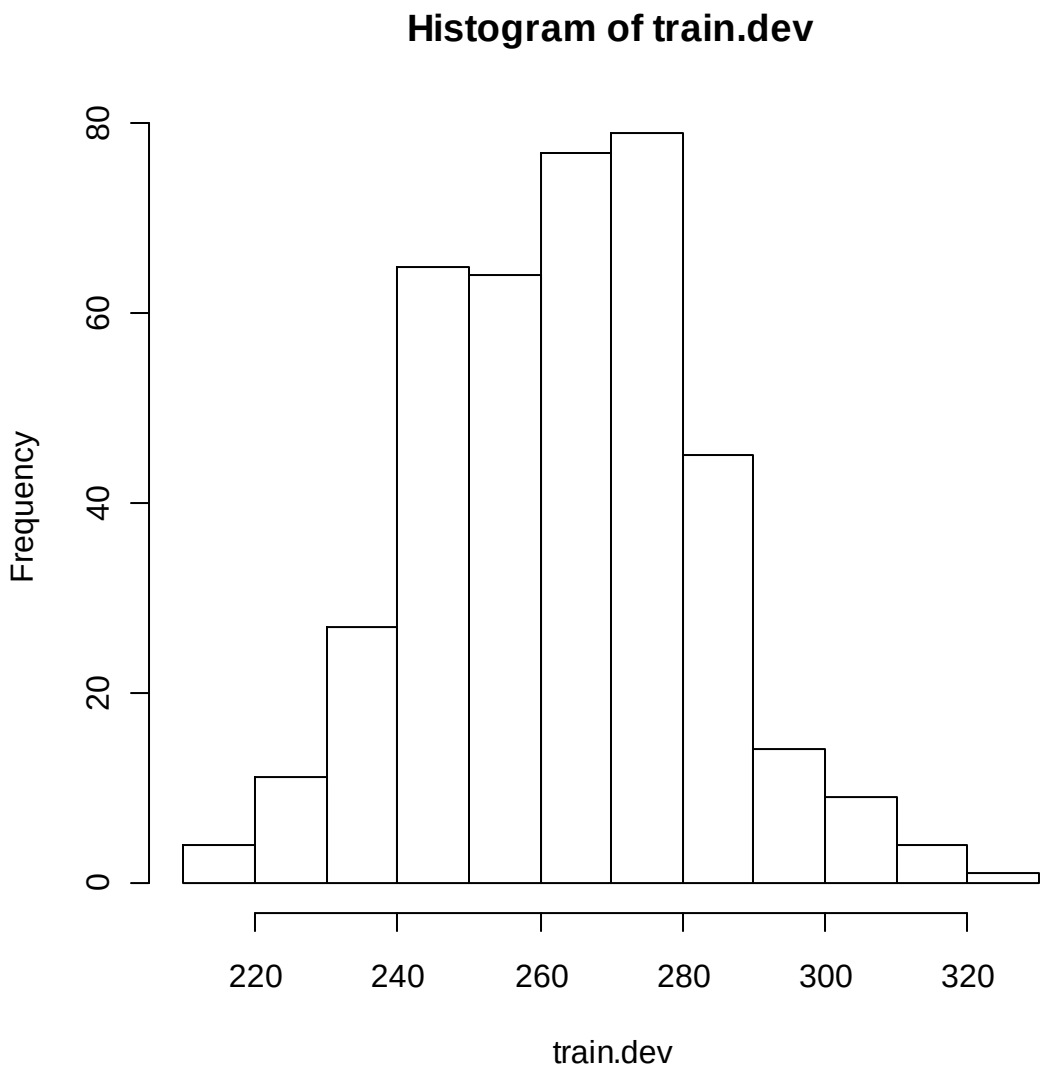
$$Dev(b) = 2\left[\ln L\left(\boldsymbol{X}; \max \right) - \ln L\left(\boldsymbol{X}^*(b); \hat{\theta} \right) \right], b = 1, 2, \ldots, B$$

手順3 適合度検定

$Dev \geq Dev^* \Rightarrow$ ﾓﾃﾞﾙは妥当でない

$$\begin{cases} Dev = 2\left[\ln L\left(\boldsymbol{X}; \max \right) - \ln L\left(\boldsymbol{X}; \hat{\theta} \right) \right] \\ Dev^* = Dev(b) \text{を小さい順に並べたとき、の第} j \text{番目の値} \\ \alpha = 1 - j/(B+1) \end{cases}$$

$\therefore Dev = 262.0221 < 295.1631 = Dev^* \Rightarrow$ ﾓﾃﾞﾙは妥当


Histogram of train.dev
Frequency
0
20
40
60
80
220
240
260
280
300
320
train.dev

4. 生存率と予測

＃1の生存率

GAMの予測値（発生日:state）	生存率	計算式
0.0346 (13:血小板回復)	0.965	1-0.0346
0.0872 (67:急性GVHD)	0.881	0.965×(1-0.0872)
0.0747 (121:慢性GVHD)	0.815	0.881×(1-0.0747)
0.00003 (2081:打切り)	0.815	0.815×(1-0.00003)

患者#1(打切り例)

state (発生日)	生存率	
	GAM	線形
血小板回復(13)	0.9654	0.9185
急性GVHD(67)	0.8813	0.8438
慢性GVHD(121)	0.8154	0.7492
打切り(2081)	0.8154	0.6671

条件付き(1年後)生存率:患者#76(再発例)

state (発生日)	生存率
	GAM
血小板回復(20)	0.9582
急性GVHD(25)	0.9145
慢性GVHD(140)	0.8218
再発(421)	0.4026
1年後(2446)	0.2506

条件付き(1年後)生存率:0.6224=0.2506/0.4026

患者#101(打切り例;患者50歳,ﾄﾞﾅｰ36歳)

「患者が高齢」で”打ち切り”の典型的なパターン

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(32)	0.3262	0.8901
急性GVHD(32)	0.3196	0.8902
慢性GVHD(360)	0.6457	0.8509
打切り(1345)	0.9984	0.8526

患者#1(打切り例;患者26歳,ﾄﾞﾅｰ33歳)

「打切り例」の典型的なパターン

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(13)	0.4231	0.9189
急性GVHD(67)	0.4044	0.9190
慢性GVHD(121)	0.6509	0.8884
打切り(2081)	1.0000	0.8910

患者#18(再発例;患者20歳,ﾄﾞﾅｰ33歳)

「再発→死亡」の典型的なパターン

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(20)	0.4474	0.9255
急性GVHD(28)	0.4441	0.9255
再発(104) ⇒ 156日目に死亡	0.2785	0.6400

患者#90(死亡例;患者23歳,ﾄﾞﾅｰ16歳)

「慢性GVHDなし」の典型的なパターン

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(23)	0.3907	0.9094
再発(211) ⟹ 653日目に死亡	0.2407	0.5895

患者#78(再発例;患者14歳,ﾄﾞﾅｰ19歳)

「再発」までの生存時間が長い場合

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(18)	0.4406	0.9222
慢性GVHD(180)	0.6709	0.8928
再発(748) ⇒ 1156日目に死亡	0.7868	0.5451

患者#82(死亡例;患者30歳,ﾄﾞﾅｰ32歳)

「慢性GVHD」から「死亡」までの生存時間が長い場合

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(19)	0.4001	0.9135
慢性GVHD(120)	0.6316	0.8812
死亡(1074)	0.9800	0.8825

患者#86(死亡例;患者30歳,ﾄﾞﾅｰ35歳)

「血小板回復」せず、「急性GVHD」発症したが、「死亡」までの生存時間が短い場合

state (発生日)	1年後の条件付き生存率	
	GAM	線形
急性VHD(10)	0.0322	0.4475
死亡(80)	0.0478	0.4478

患者#136(死亡例;患者52歳,ﾄﾞﾅｰ48歳)

「血小板回復」したが「患者が高齢」のため「死亡」

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(19)	0.3431	0.8984
死亡(363)	0.4127	0.8989

患者#90(再発例;患者23歳,ﾄﾞﾅｰ16歳)

「血小板回復」したが、短期間に「再発」

state (発生日)	1年後の条件付き生存率	
	GAM	線形
血小板回復(23)	0.3907	0.9094
再発(211) ⇒ 653日目に死亡	0.2407	0.5900

参考文献

[1]大橋靖雄,浜田知久馬(1995):生存時間解析、東大出版
[2]柴田里程(1994):Sと統計モデル,共立出版
[3]辻谷将明,竹澤邦夫(2009):マシンラーニング, 共立出版.
[4]中村剛(2001):Cox比例ハザードモデル、朝倉書店
[5]Tsujitani, M. and Sakon, M. (2009). Analysis of survival data having time-dependent covariates. *IEEE Trans. Neural Networks*, **20**, 389-394.
[6] 辻谷将明、左近賢人(2005):時間依存型共変量を伴う生存データの解析, 応用統計学, **34**, 15-29.
[7]打波守(訳)(2009):生存時間解析、シュプリンガー
[8] Wood,S.N.(2006):Generalized Additive Models: An Introduction with R,Chapman&Hall.