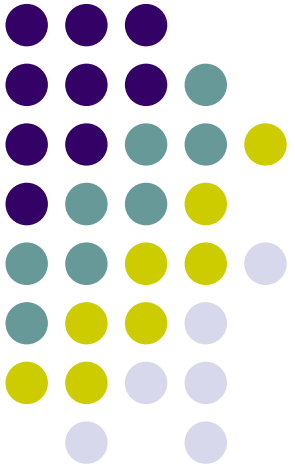
SAS Forumユーザー会 学術総会 2007

GAMとその周辺

○伊庭 克拓* 辻谷 将明**

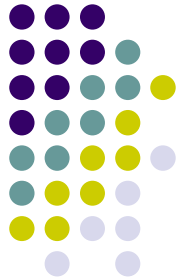
*東京CRO株式会社 DM統計本部 統計解析部

**大阪電気通信大学 情報通信工学部 情報工学科



2007年7月27日

本日の発表内容



- はじめに
- 平滑化スプライン
- 適用例
  - 脊柱後湾症データ
  - 糖尿病網膜症データ
- シミュレーション実験
- まとめ

はじめに

- 非線形性をもつ統計的多変量解析の発展
  - 薄板平滑化スプライン

    ⇒PROC TPSPLINE
  - 局所回帰

    ⇒PROC LOESS
  - 一般化加法モデル（Generalized Additive Models : GAM）

    ⇒PROC GAM
- 平滑化スプラインによるロジスティック判別の紹介と適用

２値データ

- ２値応答: Y_i

$$\Pr(Y_i = 1) = \pi_i$$

$$\Pr(Y_i = 0) = 1 - \pi_i$$

∴確率分布関数 $f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, y_i = 0 \text{ or } 1$

$$E[Y_i] = \pi_i, V[Y_i] = \pi_i(1-\pi_i)$$

ロジスティック判別

ロジット変換

$$\eta_i \triangleq \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

⇒PROC GENMOD
⇒PROC LOGISTIC etc.

- 判別分析

$$\text{応答 } y_i = \begin{cases} 1: \text{第1群} \\ 0: \text{第2群} \end{cases}$$

第i番目のデータが第1群に属す$(y_i = 1)$確率：π_i

第i番目のデータが第2群に属す$(y_i = 0)$確率：$1-\pi_i$

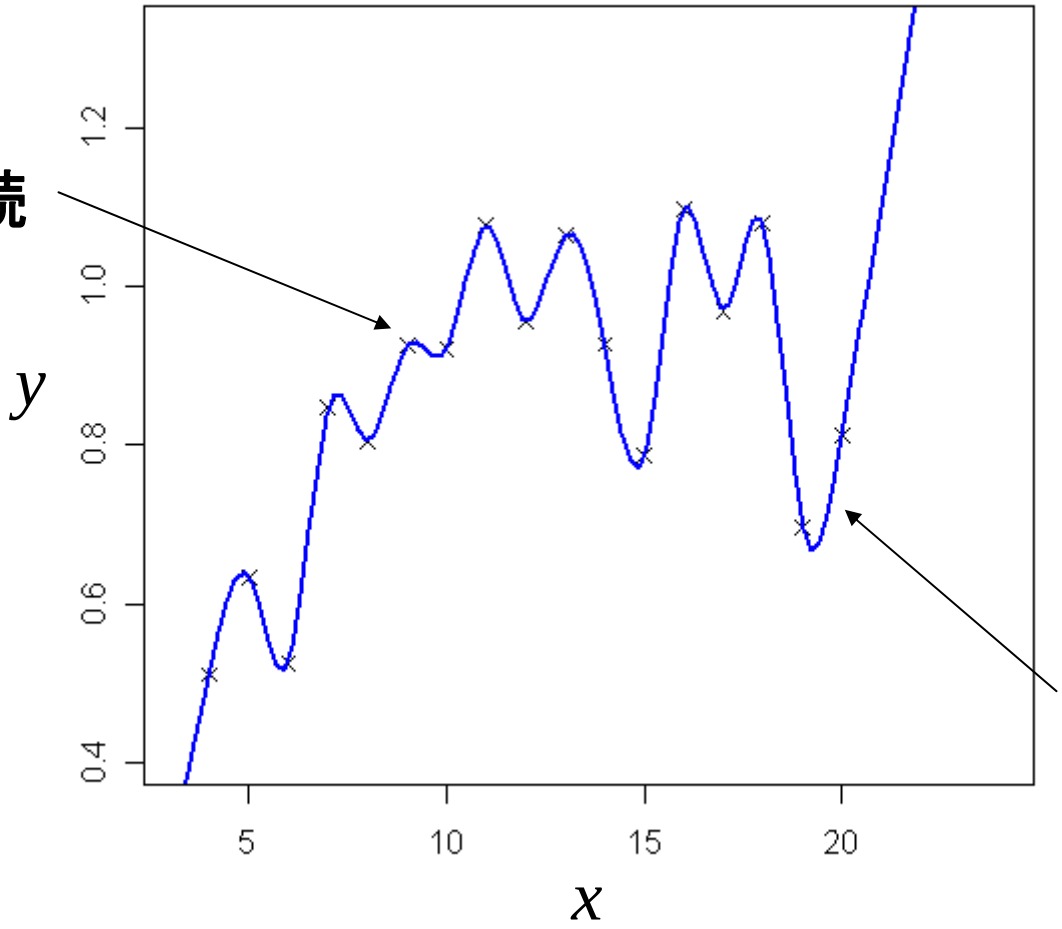
GAM（一般化加法モデル）

- 平滑関数 $s(x_{ij})$ の加法モデル
- 応答変数に指数分布族を仮定

$$\eta_i \triangleq \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + s(x_{i1}) + s(x_{i2}) + \cdots + s(x_{ip})$$

（3次の）自然スプライン

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



X：節点 x_d

区分的な3次式

（竹澤, 2007, 上巻）

ペナルティー付き残差平方和

$\lambda(\geq 0)$を大きくする⇔滑らかな曲線を求めることに重点をおく

$s(x)$ の曲率

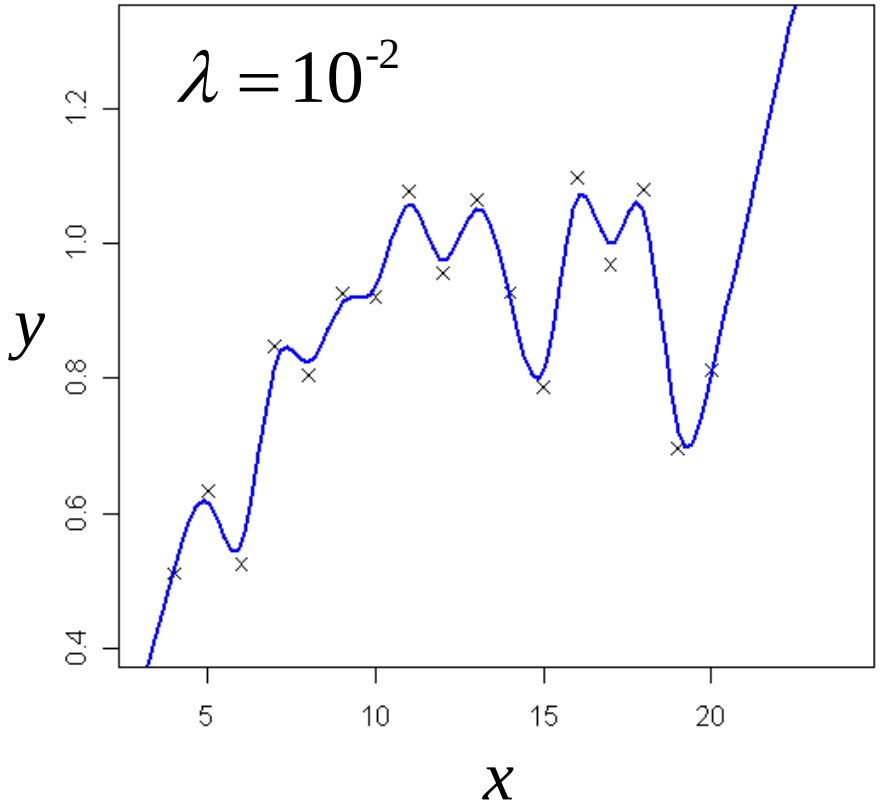
$$\sum_{i=1}^{n}\left[y_i - s(x_i)\right]^2 + \lambda\int\left\{s''(t)\right\}^2 dt$$

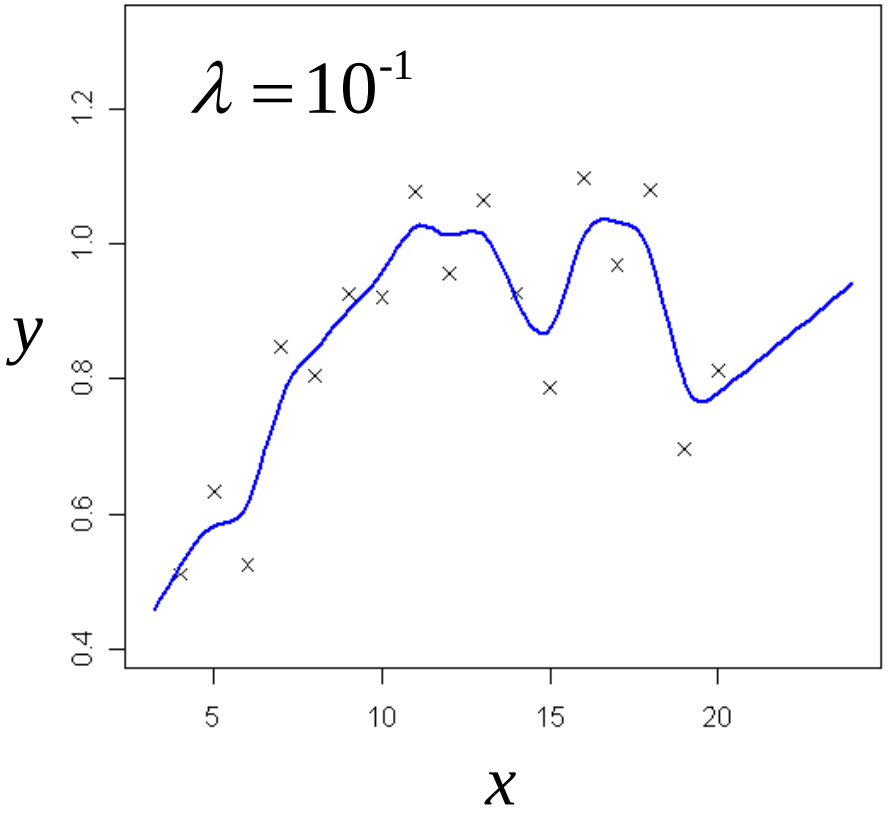
小さいほどモデルの
当てはまりは良い

小さいほど凸凹の少ない滑らかな
曲線（曲げ弾性エネルギー）

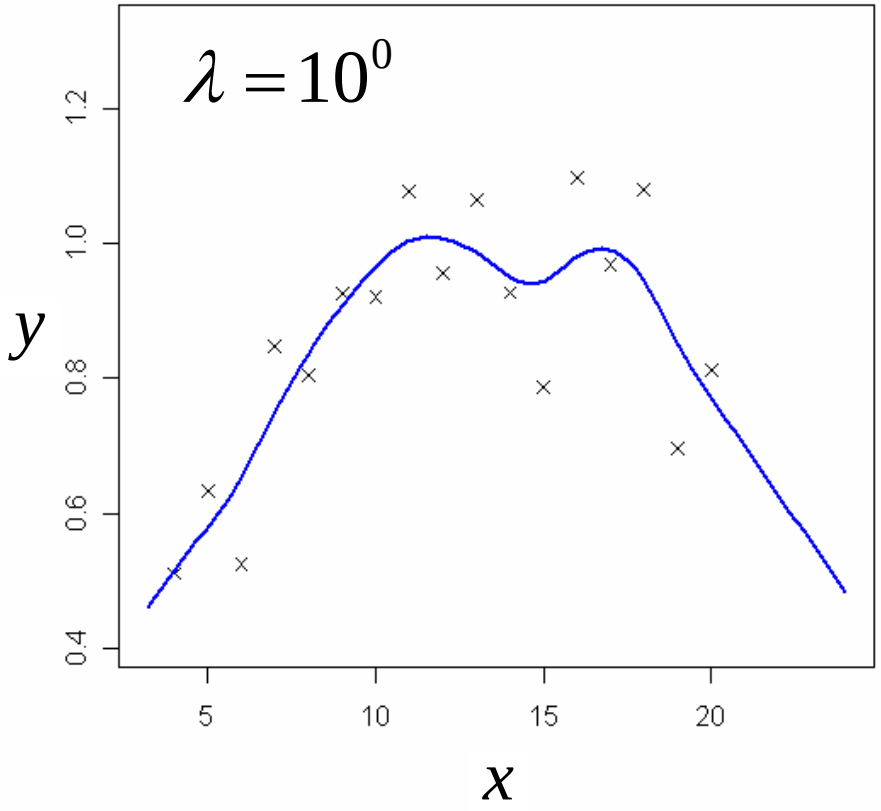
を最小にするような $y = s(x) = \hat{c}_0 + \hat{c}_1 x + \frac{1}{12}\sum_{d=1}^{n}\hat{\theta}_d \left|x - x_d\right|^3$

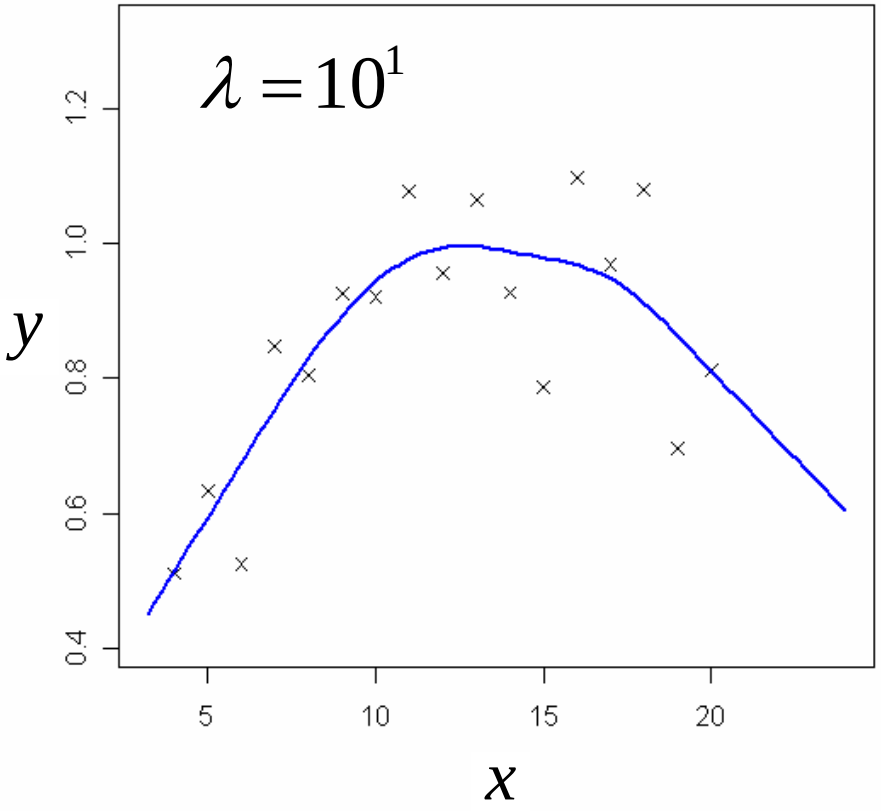
平滑化パラメータ λ



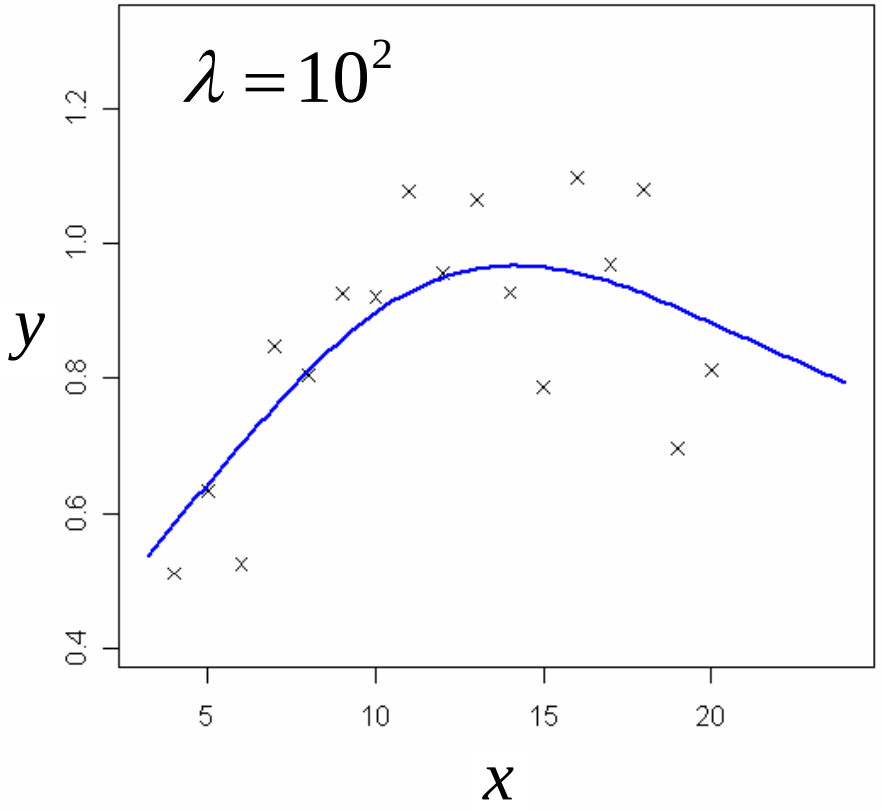


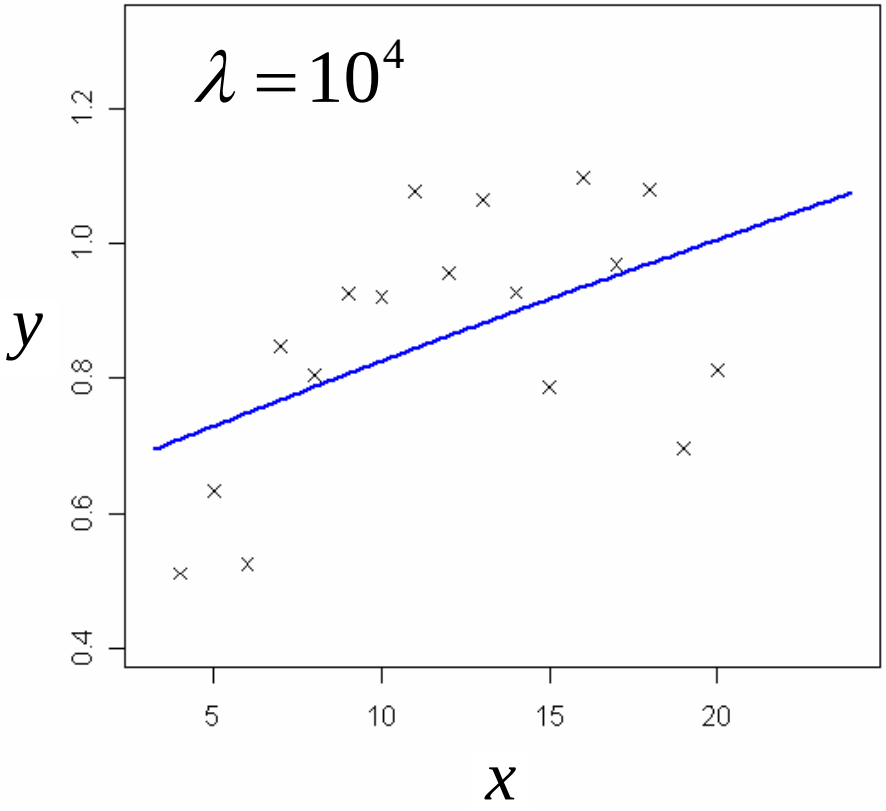
平滑化パラメータ λ





平滑化パラメータ λ





平滑化スプラインの特徴

- 非線形モデル
- 当てはまりの良さと滑らかさとのバランスをとる

平滑化パラメータ λ の決定が必要

$$\sum_{i=1}^{n}\left[y_i - s(x_i)\right]^2 + \lambda \int \left\{s''(t)\right\}^2 dt$$

小さいほどモデルの
当てはまりは良い

曲げ弾性エネルギー
（小さいほど滑らかな曲線）

一般化クロスバリデーション（GCV）

- ハット行列

$$\text{応答}\boldsymbol{y}\text{の予測値}: \hat{\boldsymbol{y}} = \boldsymbol{H}_{\lambda}\boldsymbol{y}$$

- モデルの自由度
  - 実行自由度（＝有効パラメータ数）

$$df = tr(\boldsymbol{H}_{\lambda})$$

  - 平滑化パラメータ λ の決定

$\longleftrightarrow$ 自由度 df の決定

一般化クロスバリデーション（*GCV*）

- クロスバリデーション（leaving-one-out Cross Validation：*CV*）の近似

$$GCV(\lambda) = \frac{n\left\|\boldsymbol{y} - \boldsymbol{H}_\lambda \boldsymbol{y}\right\|^2}{\left\{tr(\boldsymbol{I} - \boldsymbol{H}_\lambda)\right\}^2}$$

が最小になるように λ を決定する

$$GCV \cong \exp(AIC/n) \quad as \quad n \to \infty$$

（竹澤, 2007, 上巻）

適用例1

- 脊柱後湾症データ(n=83)
  - 脊柱後湾症の子供に対する椎弓切除術(背骨の矯正手術)の結果
  - 術後の症状の有無に影響する予後因子の探索

y : kyp　術後の症状の有無(2値データ)

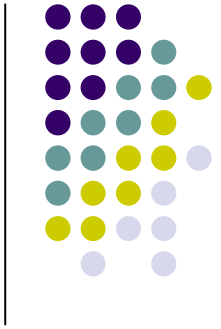
x_1 : age　手術時の年齢(月齢)

x_2 : start 上から何番目の脊椎から先を手術したか

x_3 : num 手術した脊椎の個数
(取り除いた脊椎の個数)

(小西 & 北川, 2004)

PROC GAMの文法



```
proc gam data=KYP ;
model kyp = param( start num ) spline( age , df=&df )
                                                    / link = logit dist = binomial ;
run;
```

① ② ③ ④

① model文の左辺は応答変数、右辺はモデル式

② 線形項 param(*variable* ・・・)

LOESS loess(*variable* ,df=*number*)

平滑化スプライン spline(*variable* ,df=*number*)

薄板平滑化スプライン spline2(*variable1,variable2* ,df=*number*)

③ df値を指定しないとき、デフォルトdf=4

④ ③の場合、/の後にmethod=gcvを指定するとGCVを使ってグリッド検索による平滑化パラメータの探索

最適自由度の選択

- PROC　GAMにおける最適自由度の決定
  - グリッド検索による平滑化パラメータの探索
  - しばしば数値計算がill-conditionになる

- 提案法
  - モデル全体の*GCV*で各項の自由度を探索

*GCV*逸脱度（Generalized Cross-validated deviance）

$$V_g = \frac{nDev(\boldsymbol{y}, \hat{\boldsymbol{\pi}})}{\left\{tr(\boldsymbol{I} - \boldsymbol{H}_\lambda)\right\}^2}$$

← 逸脱度（**deviance**）

（Hastie et al., 1990;竹澤, 2007, 下巻;Wood, 2006）

最適な自由度の決定

- 最適な自由度の選択結果

	GCV	*df*	逸脱度	誤判別率
提案法	0.763	3	54.50	0.133

参考　leaving-one-out CV : df=2　,逸脱度=56.39
　　Wood法（R）　　: df=2.5,逸脱度=54.97

- 他の判別手法との比較

誤判別率＝
- 0.133 : GAM
- 0.205 : ロジスティック判別
- 0.193 : 線形判別

（辻谷 & 外山, 2007; Wood, 2004;Wood, 2006）

適合度検定

- ブートストラップ法による棄却点の算出

$\boldsymbol{X} = (x_i, y_i)$: 初期標本 $\xrightarrow{(b = 1, \cdots, B)}$ $\boldsymbol{X}_b^* = (x_i, y_i^*)$: ブートストラップ標本

$Dev = -2 \ln L\left(\boldsymbol{X}; \hat{\boldsymbol{\beta}}(\boldsymbol{X})\right)$
: 初期標本による逸脱度

$Dev(b) = -2 \ln L\left(\boldsymbol{X}_b^*; \hat{\boldsymbol{\beta}}(\boldsymbol{X}_b^*)\right)$
: ブートストラップ標本による逸脱度

$Dev \geq Dev^* \Rightarrow$ モデルは妥当ではない

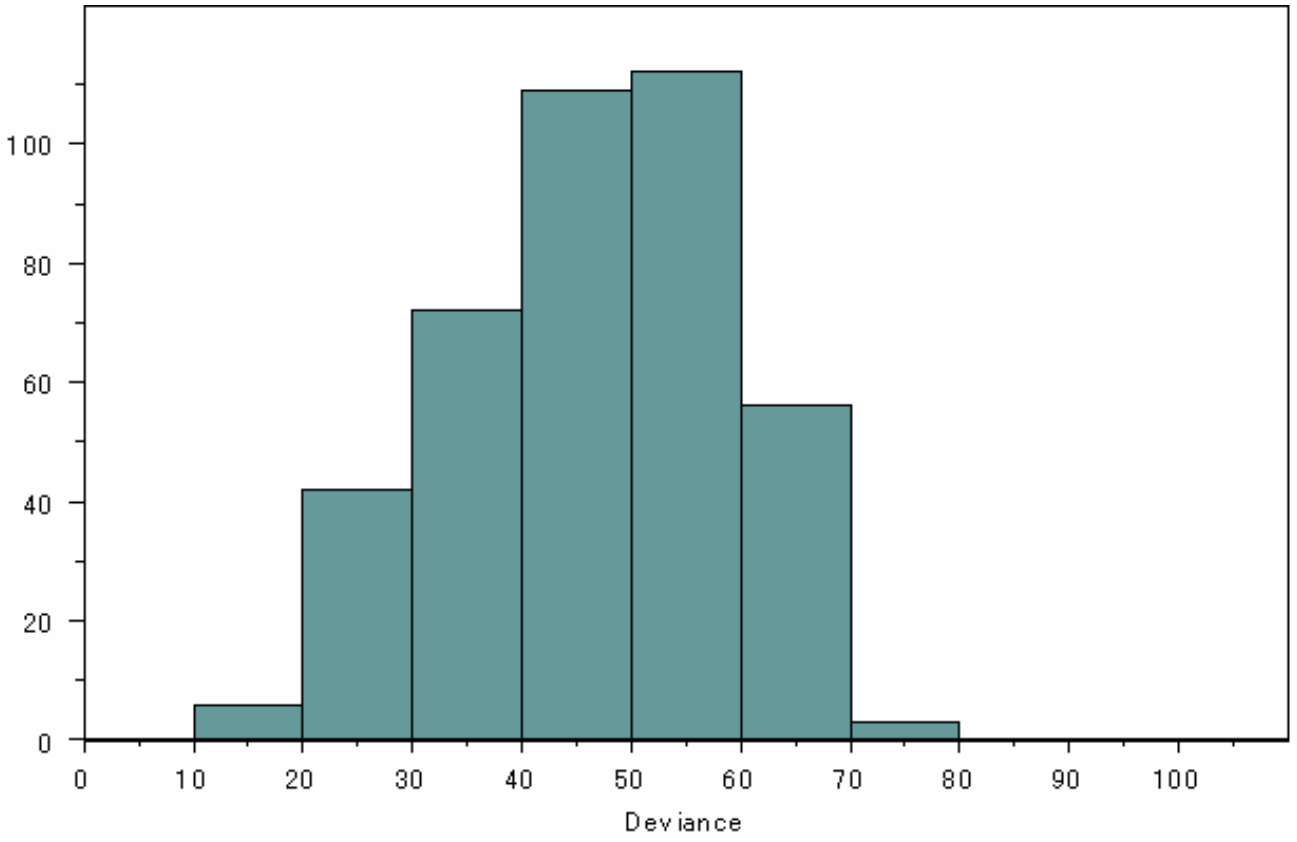
$Dev^* = Dev(b)$を小さい順に並べたときのj番目の値

有意水準 : $\alpha = 1 - j/(B+1)$

適合度検定

- ブートストラップされた逸脱度のヒストグラム

$$Dev = 54.50 < 66.06 = Dev^{*} \quad (\alpha = 0.05, B = 400)$$



共変量の有意性検定

- 尤度比検定

$$Dev(M_1) - Dev(M_0) \sim \text{自由度}(\nu_1 - \nu_0)\text{の}\chi^2\text{分布}$$

帰無仮説のモデルでの
逸脱度（自由度 ν_1）

もとのモデルでの
逸脱度（自由度 ν_0）

  - 共変量の効果

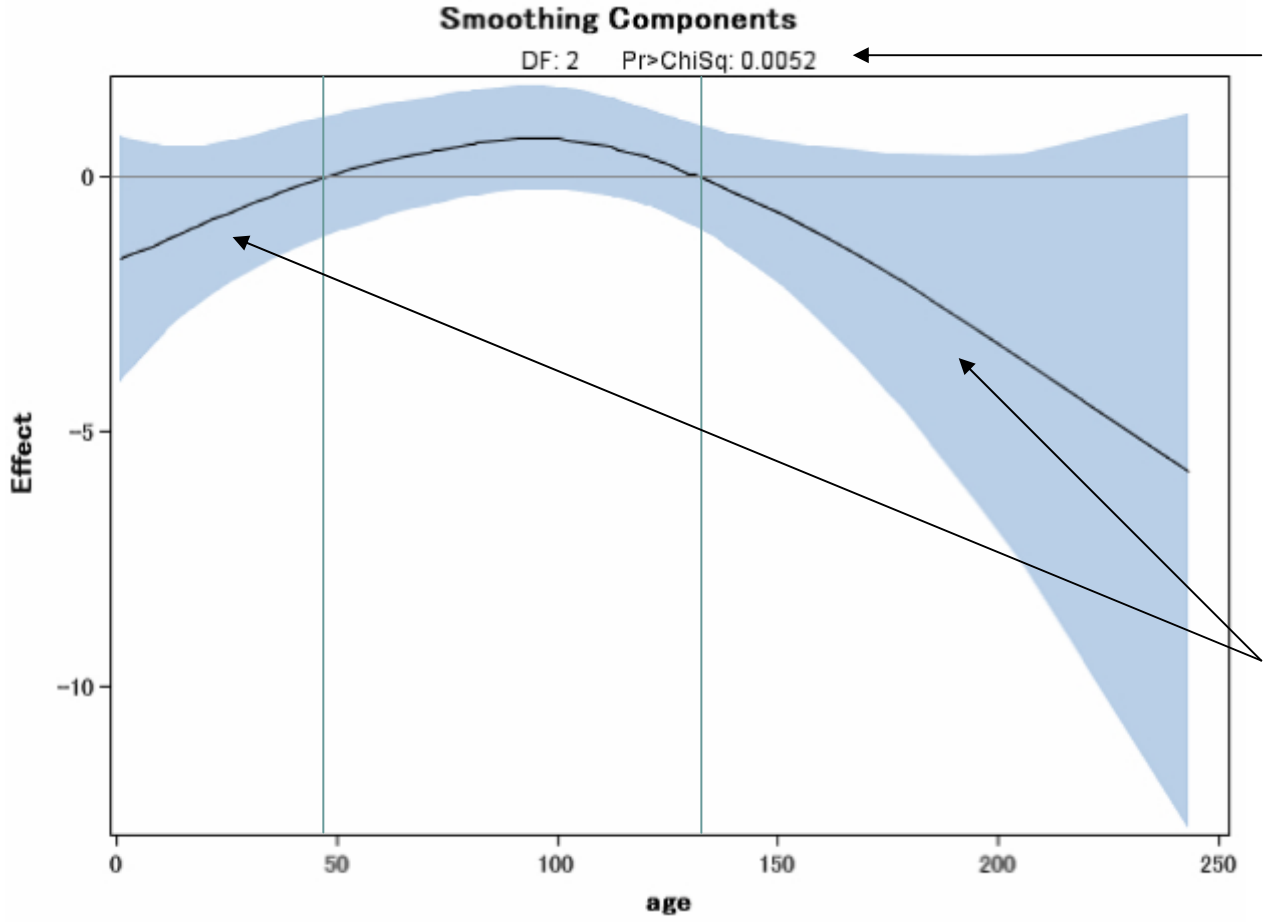
    M_0 : kyp = param(start num) spline(age)
    M_1 : kyp = param(start num)

  - 共変量の非線形効果

    M_0 : kyp = param(start num) spline(age)
    M_1 : kyp = param(start num age)

手術時の月齢での平滑スプライン

$$Dev(M_1) - Dev(M_0) = 11.74^{**}; df = 3$$



残りやすい

症状の残りやすさ

残りにくい

非線形効果の尤度比検定

手術の成功率が高い

適用例2

- 糖尿病網膜症データ（n=669）
  - 糖尿病患者における糖尿病網膜症の発症と進行のリスクファクターを探索した疫学研究の結果

y : ret　糖尿病網膜症の進行（2値データ）

x_1 : dur　糖尿病の罹病期間 ; year

x_2 : gly　糖化ヘモグロビン ; %

x_3 : bmi　肥満度 ; kg/m^2

（Wahba et al., 1995; Wang et al.,1997）

薄板平滑化スプライン

- 罹病期間とBMIには交互作用の可能性あり
- 薄板平滑化スプラインの利用

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = s(\boldsymbol{x}_i) \triangleq \sum_{\alpha<\beta} s_{\alpha\beta}(x_\alpha(i), x_\beta(i))$$

```
proc gam data=WESDR ;
model ret = param( gly ) spline2( dur , bmi , df=&df )
                                        / link = logit dist = binomial ;
run;
```

最適な自由度の決定

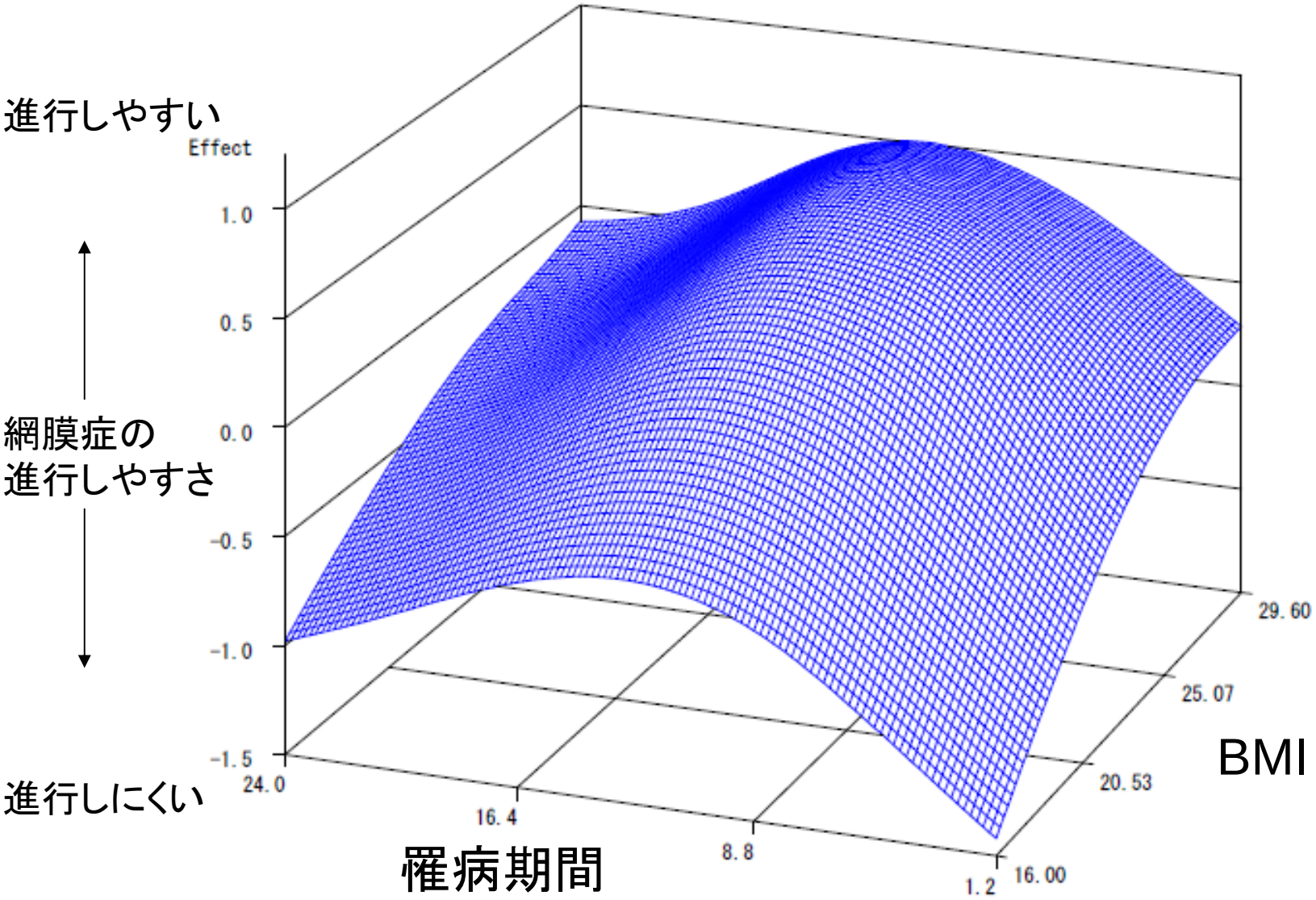
- 最適な自由度の選択結果

	GCV	*df*	逸脱度	誤判別率
SAS自動選択	1.148	8.3	744.81	0.274
提案法	1.146	8	745.48	0.280

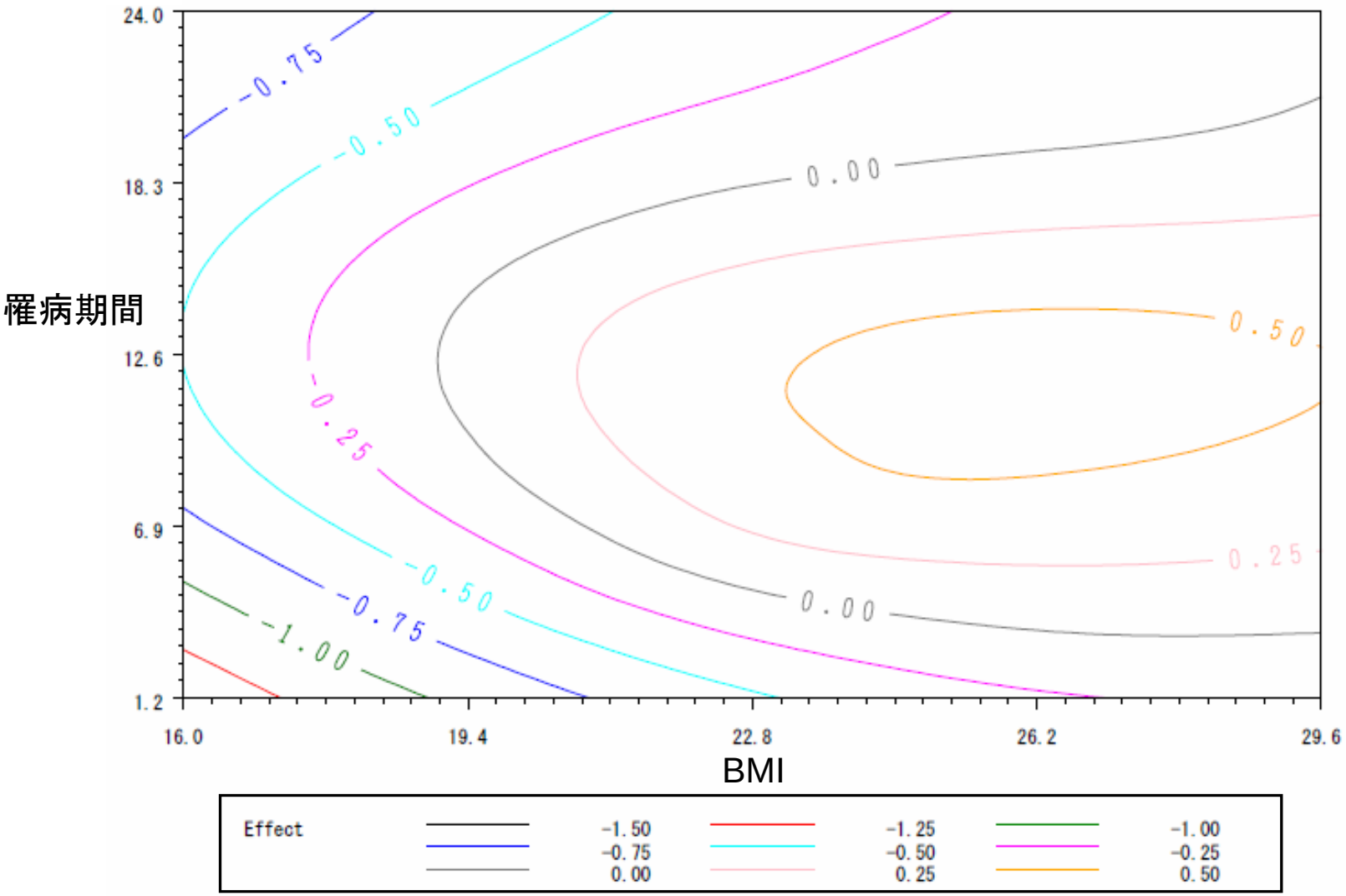
- 他の判別手法との比較

誤判別率＝
- 0.274 : GAM
- 0.306 : ニューラルネット
- 0.293 : ロジスティック判別
- 0.318 : 線形判別

罹病期間とBMIの薄板平滑化スプライン



罹病期間とBMIの薄板平滑化スプライン



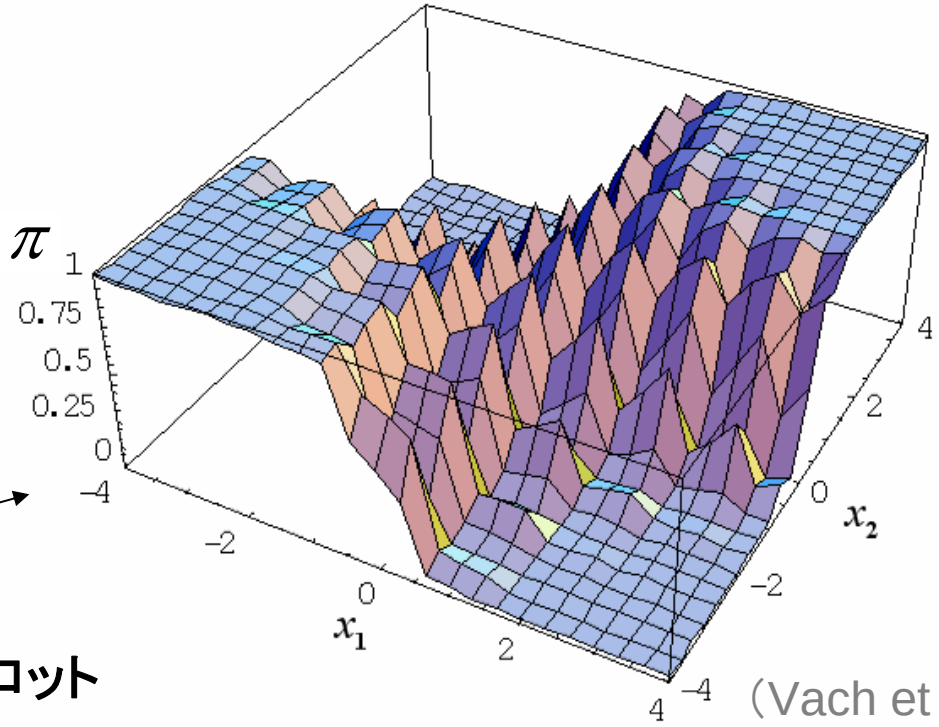
シミュレーション実験

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in N\left(\mathbf{0}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right)$$: 乱数

$$s(x_1, x_2) = \sin(2 \times 3.14 \times x_1) + x_1 x_2 + \sin(2 \times 3.14 \times x_2)$$: モデル

$$\pi = \frac{1}{1 + \exp\{-s(x_1, x_2)\}}$$

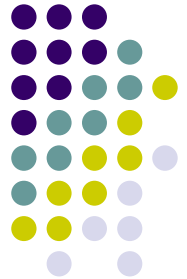
$$y = \begin{cases} 1(\text{第1群}) : \pi \geq 0.5 \\ 0(\text{第2群}) : \pi < 0.5 \end{cases}$$



群を決定する関数
(π, x_1, x_2) の3Dプロット

（Vach et al., 1996）

シミュレーション実験

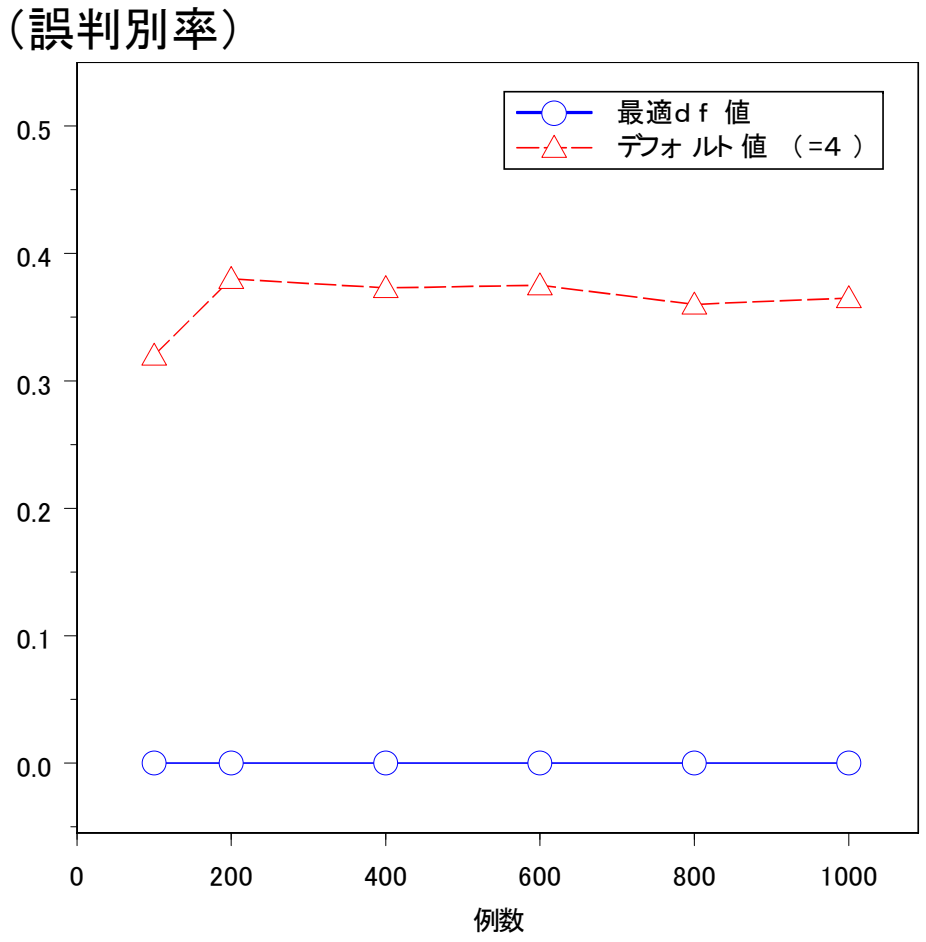


- 訓練標本：モデルの推定用
- 検証標本：モデルの評価用
  （将来のデータに対する精度を評価）

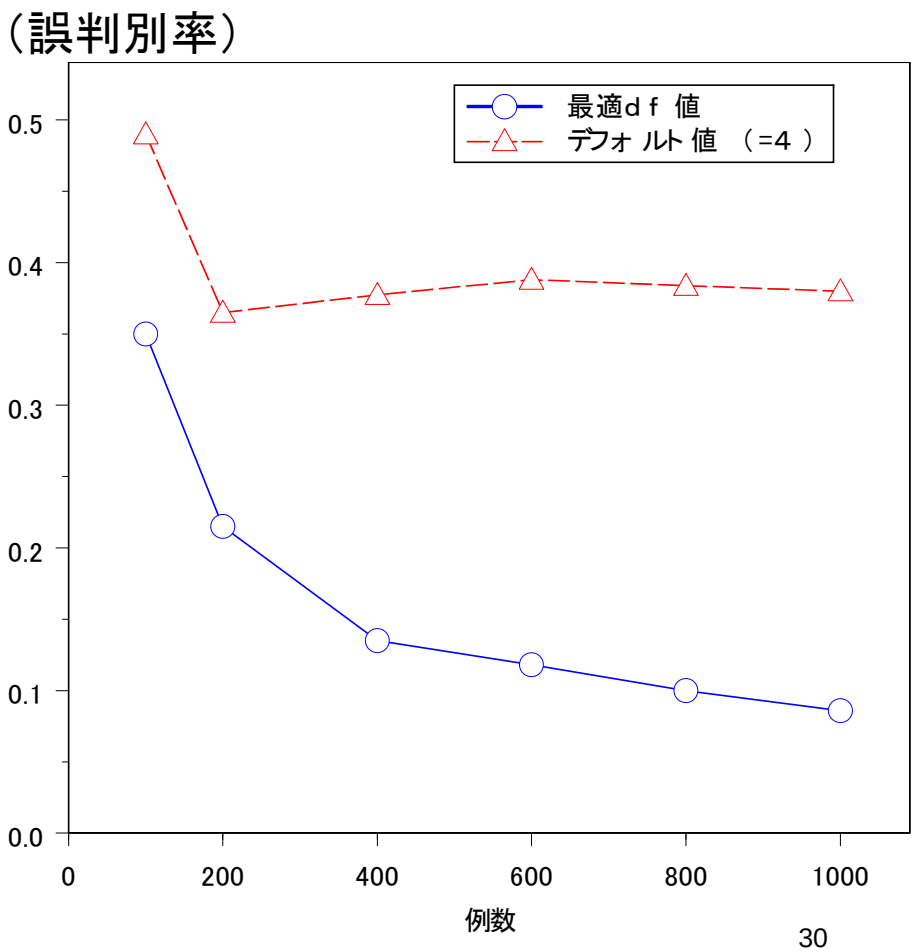
1. 自由度の選択　なし v.s. あり
2. GAM　v.s.　他の判別モデル

- シュミレーションデータ n=100,200,400,600,1000

自由度の選択　なし v.s. あり

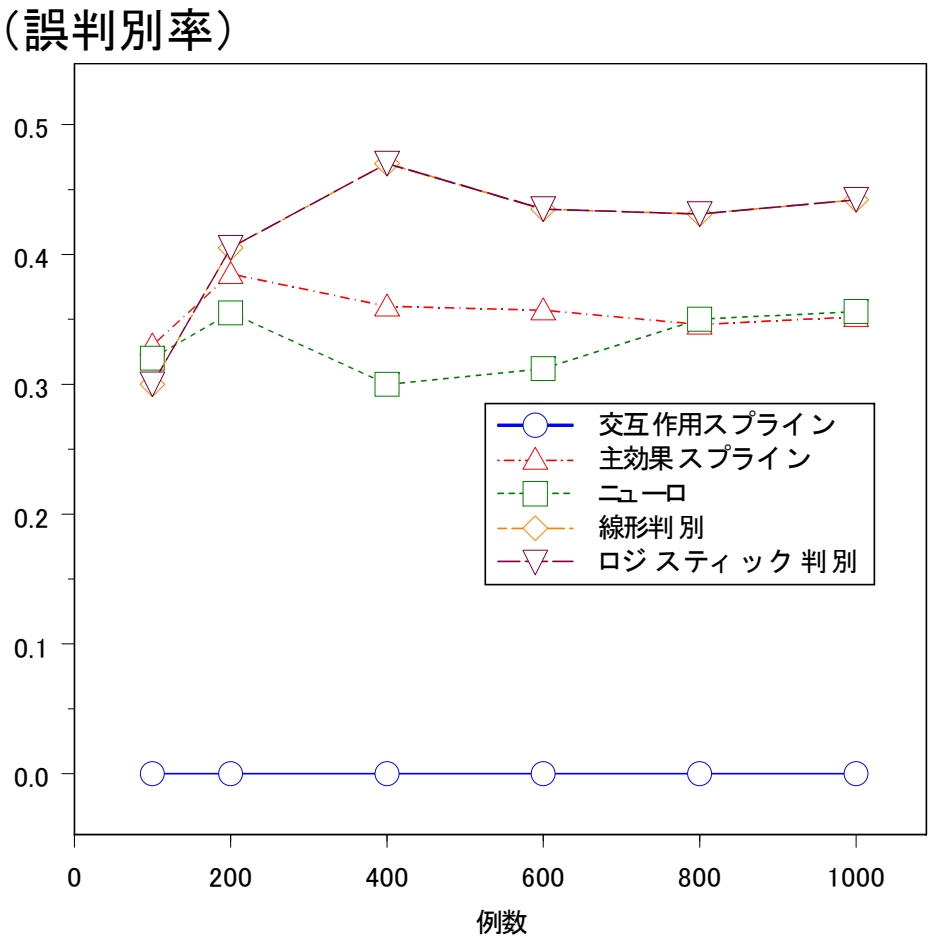


訓練標本（推定に用いたデータ）

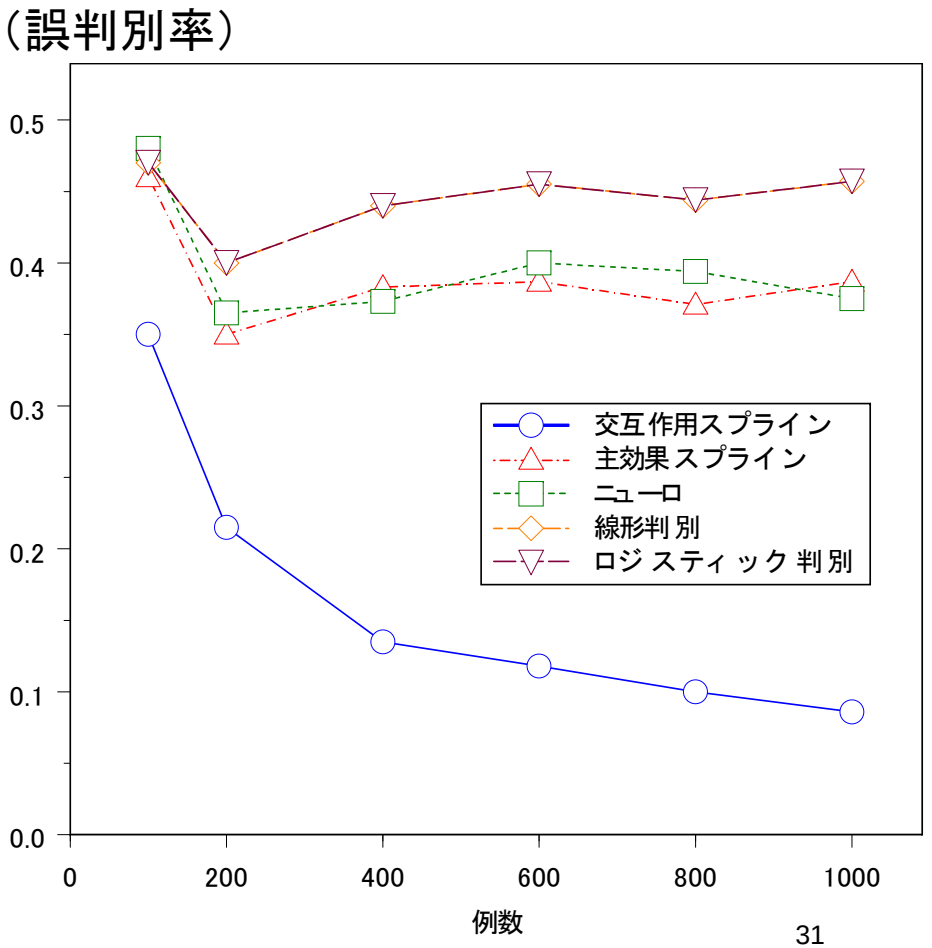


検証標本（推定に用いていないデータ）

GAM v.s. 他の判別モデル



訓練標本(推定に用いたデータ)



検証標本(推定に用いていないデータ)

まとめ

- 自由度の決定が必要
- 共変量の非線形性をグラフ表現
- 交互作用を考慮
- ill-conditionに落ち込む可能性あり

参考文献

1. Hastie, T.J. and Tibshirani, R.J.(1990): Generalized Additive Models, Chapman and Hall, London.
2. 小西貞則, 北川源四郎(2004): 情報量規準, 朝倉書店.
3. 竹澤邦夫(2007): みんなのためのノンパラメトリック回帰(上,下) 第3 版, 吉岡書店
4. 辻谷将明, 外山信夫(2007): R によるGAM 入門, 行動計量学, 34, 111-131.
5. Vach, W. et al.1996): Neural betworks and logistic regression:Part II, Comput. Statist. & Data Analy., 21, 683-701.
6. Wahba, G. et al., (1995): Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiolpgical study of diabetic retinopathy, Ann. Statist., 23, 1865-1895.
7. Wang, Y. et al., (1997): Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy, Statist. Medicine, 16, 1357-1376.
8. Wood, S.N.(2004): Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Amer. Stat. Assoc. 99, 673-686.
9. Wood, S.N.(2006): Generalized Additive Models, An Introduction with R, New York, Chapman & Hall.