一般化加法モデル(Generalized Additive Models: GAM)による個人ローンデータの解析

大阪電気通信大学　大学院
修士課程　工学研究科　情報工学専攻
西　尾　政　人

①従来は線形モデルGLMによって解析されていたものを非線形モデルGAMを採用し、その非線形性の可視化

②影響観測値の検出には、逸脱度に基づくDIFDEVを提案し、モデルの当てはめに悪い影響を与えているとされる観測値の検出

③一例消去CV法を活用することで、平滑化パラメータの決定を行い、それと同時に影響観測値の検出が可能

④影響観測値を除去後、共変量の非線形性の変化を可視化→従来は、逸脱度の変化やCV値の変化のみ

実際の個人ローンデータ（2000人のデータ）

顧客No.	ローン破産の有無	共変量				
		年齢	ローン額(ドル)	観測期間(年)	保険料(ドル)	ローン目的
1	0	53	1000	18	104	2
•	•	•	•	•	•	•
876	1	31	2000	14	219	3
•	•	•	•	•	•	•
1503	0	35	1000	5	88	2
•	•	•	•	•	•	•
2000	0	22	500	12	0	2

応答値(ローン破産) $y=\begin{cases}0：破産なし\\1：破産あり\end{cases}$

ローン目的 $=\begin{cases}1：ローリスク\cdots家電製品の購入、旅行等\\2：ミディアムリスク\cdots車や家の購入等\\3：ハイリスク\cdots借金のための借金等\end{cases}$

1. 個人ローンデータについて

X1 :年齢
X2 :ローン額
X3 :観測期間
X4 :保険料
X5 :ローンの目的（何のために組んだか）
0.ローリスク :新婚旅行、電化製品の購入など
1.ミディアムリスク :家の改築、車の購入など
2.ハイリスク :家の購入、借金のための借金など

共変量

＜データ＞

2000人分のデータ:共変量はローンを借りたときの情報（年齢、ローン額、ローン目的など）

＜定義＞

解析研究を行なう上で、説明変数に上記のような共変量を用いる。それぞれをX1～X5と定義する。ただし、ローン目的は離散値であり、1ならローリスク、2ならミディアムリスク、3ならハイリスクとなっている。

目的変数 :ローン破産の有無を用いる。0であれば破産しておらず、逆に1であれば返済中に破産

ロジット変換

【例 :ローン破産の起きる確率とローン額との関係】

ローン破産が起きる確率

$$P = 0.2 + 0.7x$$

ローン額

$x \to +\infty$ のとき $P \to +\infty$

$x \to -\infty$ のとき $P \to -\infty$

Pの値は確率であるため

$$0 \leq P \leq 1$$

で、なければならない。

ロジット変換

$$\ln\left(\frac{P}{1-P}\right)=0.2+0.7x$$

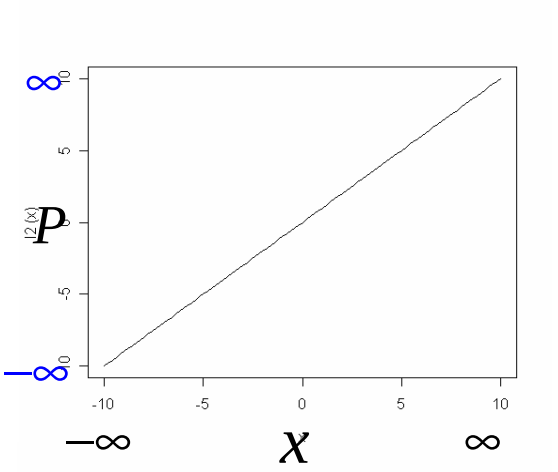
$x \to +\infty$のとき$P \to 1$

$x \to -\infty$のとき$P \to 0$

$\therefore 0 \le P \le 1$

ロジット変換しない場合

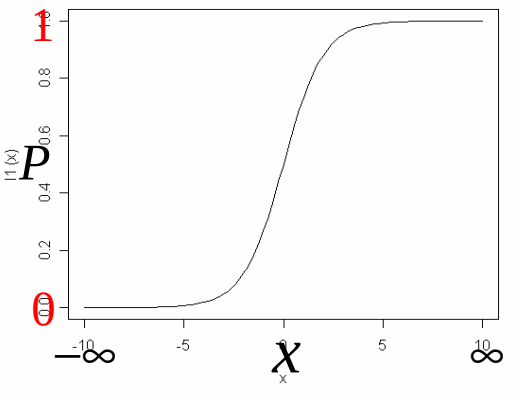
$$P = c_0 + c_1 x$$



ロジット変換した場合

$$\ln\left(\frac{P}{1-P}\right) = c_0 + c_1 x$$

$$P = \frac{1}{1+\exp\{-(c_0 + c_1 x)\}}$$

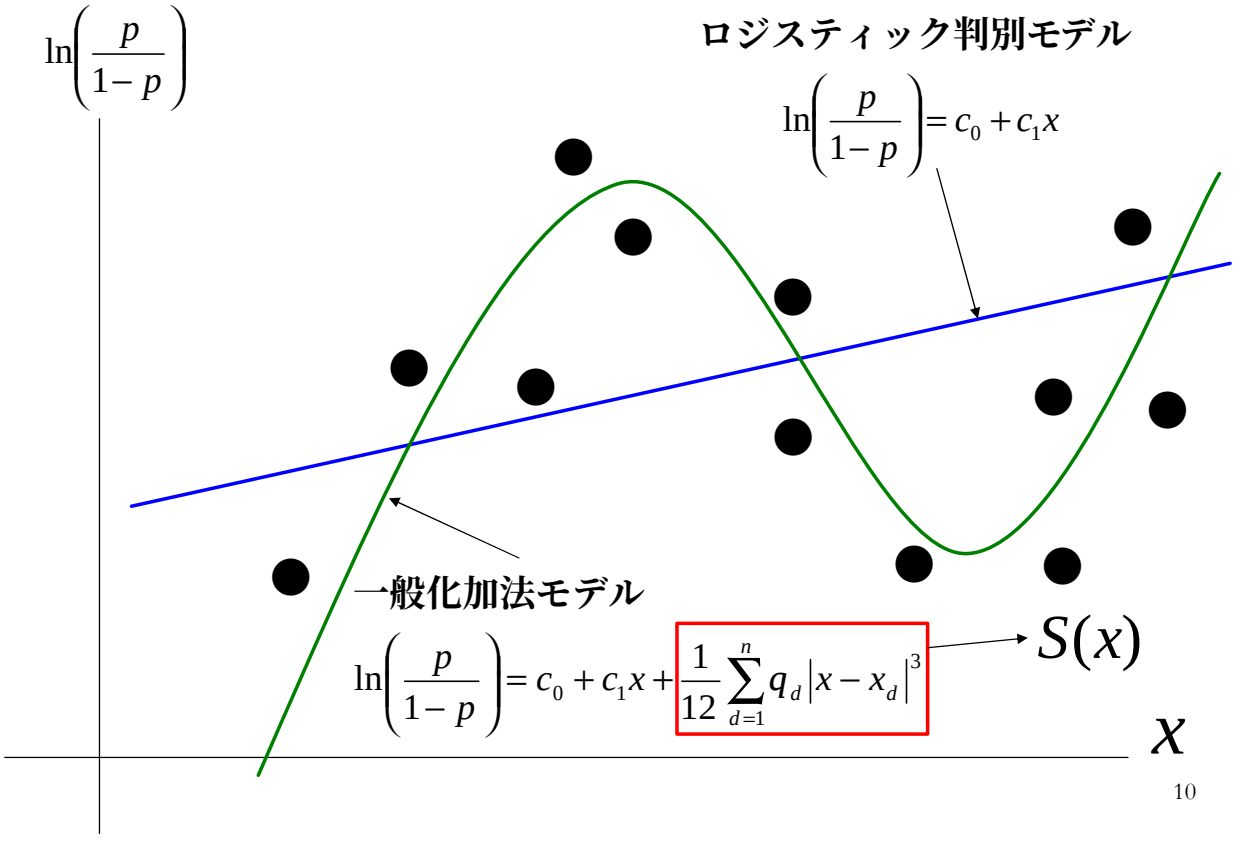


ロジット変換とその必要性

- ロジット変換とは回帰式などによって算出される値を0～1の間にある数値へと変換する方法である。

- その必要性については、次ページに示すが、回帰式によって算出した確率Pは、Xの増減によって-∞～∞の値を取ってしまう。ゆえに、ロジット変換を行なうことによりXの増減に関係なく、確率が0～ 1の確率を取る。

ロジスティック判別モデルと一般化加法モデル



線形モデルと一般化加法モデルとの違い
(共変量が一個の場合)

線形モデル

$$\ln\left(\frac{p}{1-p}\right)=c_0+c_1x$$ ⇒ 一次式

一般化加法モデル

$$\ln\left(\frac{p}{1-p}\right)=c_0+c_1x+\frac{1}{12}\sum_{d=1}^{n}q_d\left|x-x_d\right|^3$$ ⇒ 三次式

ペナルティ付き残差平方和

小さいほどモデルの当てはまりは良い

第1項と第2項のバランスを調節する

曲げ弾性エネルギー (小さいほど滑らかな曲線)

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

最小にする関数

$s(x)$の曲率

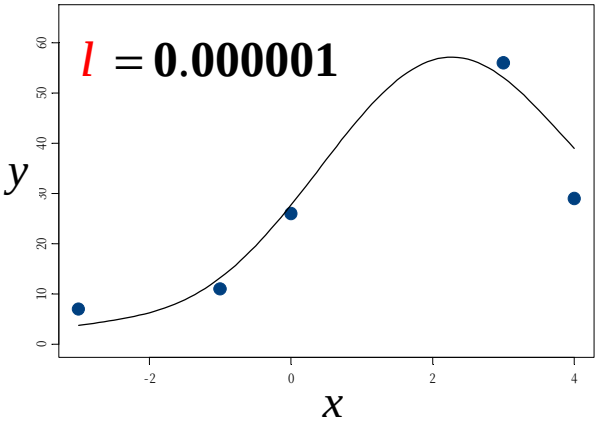
平滑化スプライン（3次自然スプライン関数）

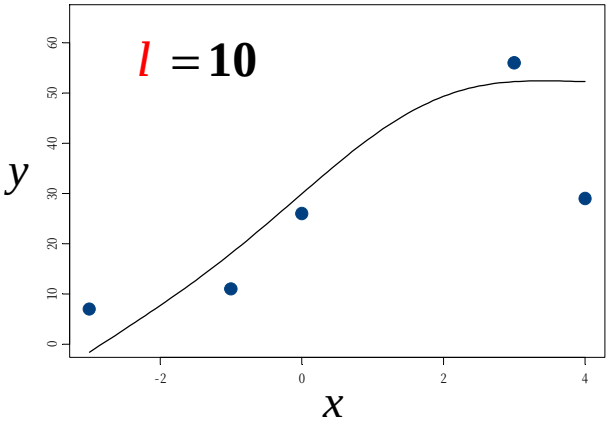
$$\ln\left(\frac{P}{1-P}\right)=s(x)=c_0+c_1x+\frac{1}{12}\sum_{d=1}^{n}q_d\left|x-x_d\right|^3$$

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

平滑化パラメータl　とは？

曲げ弾性エネルギーが大　曲げ弾性エネルギーが小





- **1例消去CV法（厳密解）** ⇐ 計算量が膨大

初期標本 $\boldsymbol{X} = \left\{ \boldsymbol{X}^{\langle 1 \rangle}, \boldsymbol{X}^{\langle 2 \rangle}, \ldots, \boldsymbol{X}^{\langle 2000 \rangle} \right\}$

⇓ d番目を除去 : $\boldsymbol{X}^{\langle d \rangle} = \left\{ x_1^{\langle d \rangle}, \ldots, x_5^{\langle d \rangle}; y^{\langle d \rangle} \right\}$

$$\boldsymbol{X}_{[d]} = \left\{ \boldsymbol{X}^{\langle 1 \rangle}, \boldsymbol{X}^{\langle 2 \rangle}, \ldots, \boldsymbol{X}^{\langle d-1 \rangle}, \boldsymbol{X}^{\langle d+1 \rangle}, \ldots, \boldsymbol{X}^{\langle n \rangle} \right\}$$

$$CV = -2 \sum_{d=1}^{2000} \left\{ y^{\langle d \rangle} \ln \hat{p}_{[d]}^{\langle d \rangle} + \left(1 - y^{\langle d \rangle} \right) \ln \left(1 - \hat{p}_{[d]}^{\langle d \rangle} \right) \right\}$$

$\boldsymbol{X}_{[d]}$で構築したモデルの，$\boldsymbol{y}^{\langle d \rangle}$ の予測値

- **Wood(**Generalized-Cross-Validation**)法**

  GCV :1**例消去CV(厳密解）の近似解**

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

AIC = **モデルの当てはまりの悪さ**+**モデルの複雑さ**

平滑化パラメータの最適選択

- 最適選択の結果

2^{-20}～2^{0}まで動かす	平滑化パラメータ			
	観測期間	*年齢*	*ローン額*	*保険料*
1例消去CV法	2^{-13}	2^{-10}	2^{-15}	2^{-10}
Wood法 (GCV)	5.03×10^{-5}	2.20×10^{-6}	7.04×10^{-8}	2.35×10^{-4}

- 1例消去CVとGCVとの比較
  - 1例消去CV: 影響分析が可能 :計算量が膨大
  - GCV: 過剰当てはめの傾向 :計算は1回で済む

連続値グラフ 箱ひげ図）の見方

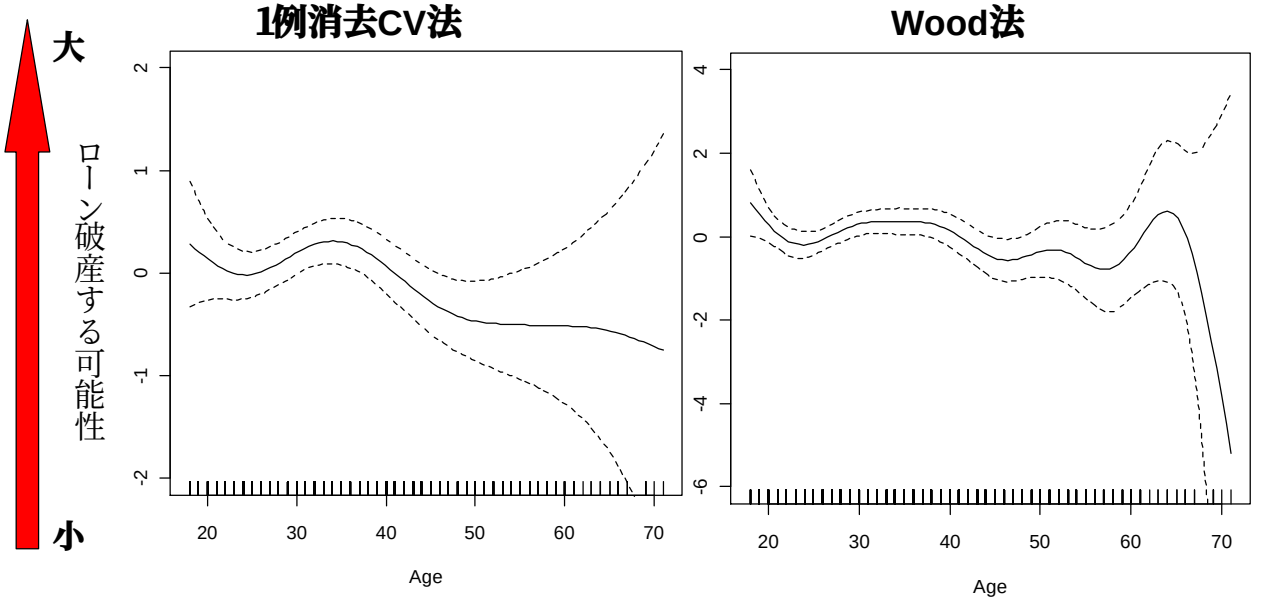
- グラフは各々の説明変数にスプライン関数を追加したグラフである。また、x軸は各々対象となった説明変数となっている。y 軸は上にいくほどローン破産する確率が高くなっていることを表わしている。見方としては、x軸の増加に伴ってy軸の動きがどのように変化していくかを観察する。

- 箱ひげ図について、y 軸は上記と同様である。x軸についてはカテゴリで区別する。見方としては、最大値、最小値はもちろんのこと、中央値の増減関係、箱の形の変化などカテゴリ別でどのような違いがあるのかを観察する。

非線形関数の可視化

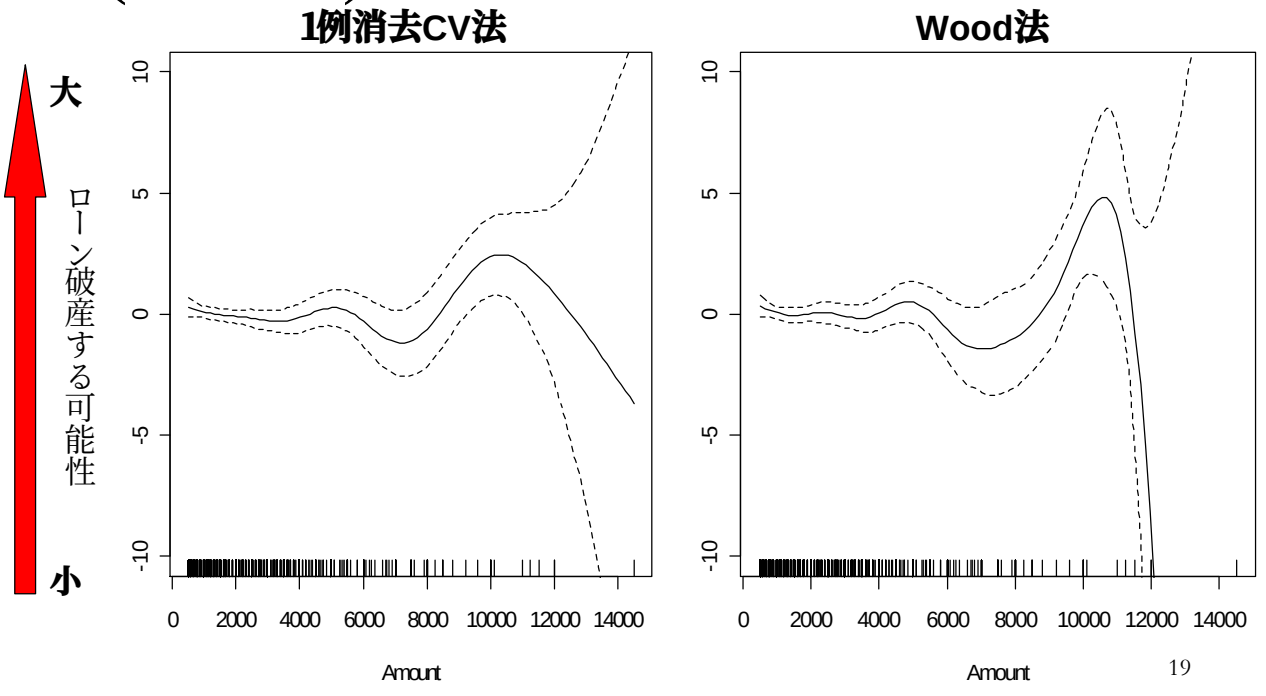
S(年齢)

年齢



ローン額

S(ローン額)


1例消去CV法
Wood法
大
ローン破産する可能性
小
10
5
0
-5
-10
0
2000
4000
6000
8000
10000
12000
14000
Amount

箱ひげ図を用いる場合の例

例 :性別
- 男 :0
- 女 :1

→ カテゴリカルデータ 離散値）

連続値でないため散布図では表わせない）

↓

S-PLUSを用いた箱ひげ図の作成

箱ひげ図とは？

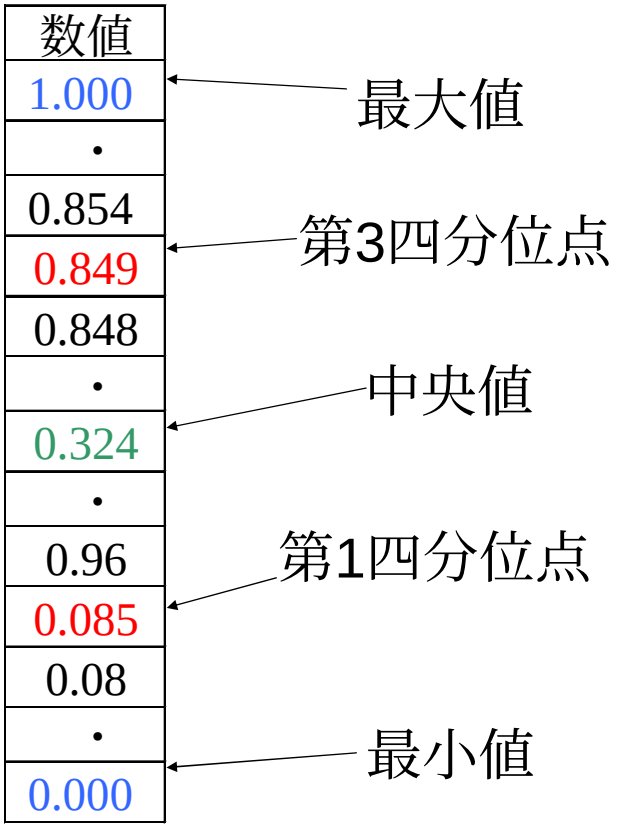
・右のような解析結果が得られたとする。

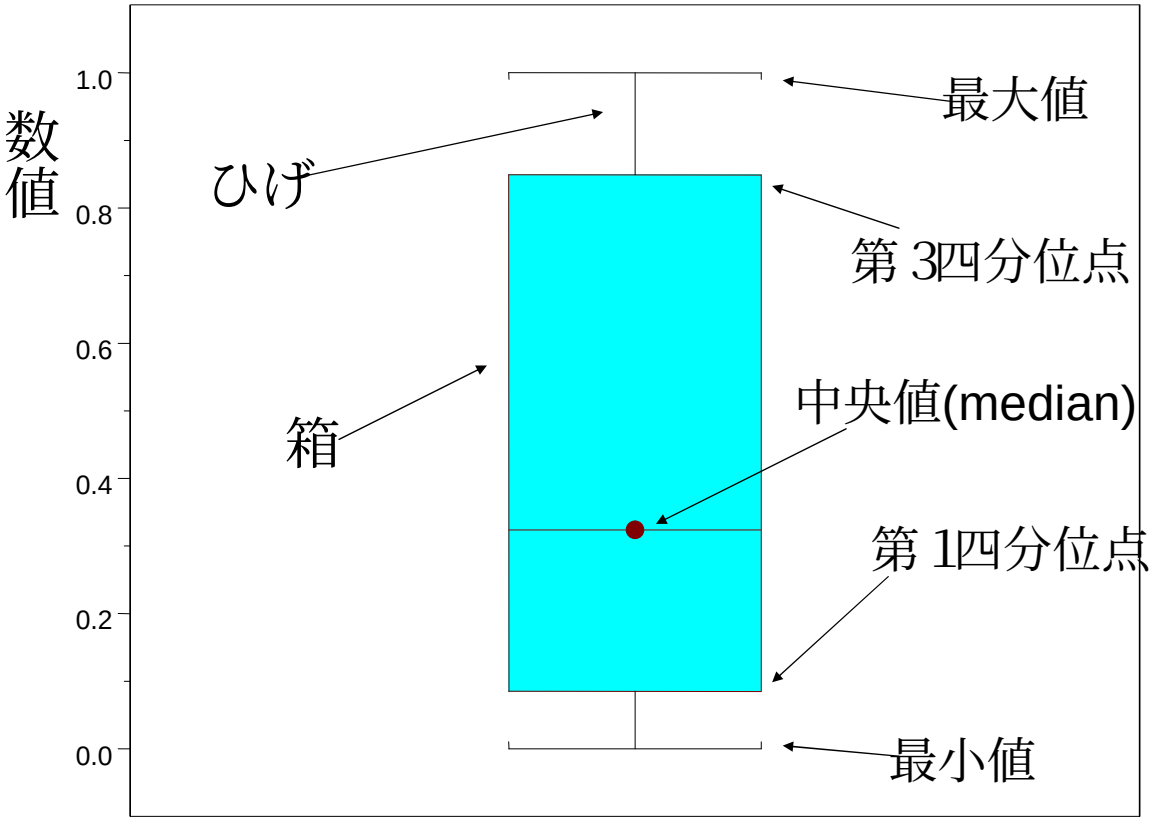
・これを箱ひげ図を描く場合、最大値1.0、最小値0となる。

・中央値とはデータの中のちょうど真ん中の数値のことを表わす。

・第1, 第3四分位点とは、最小値から数えて1/4番目にある数値を第1四分位点、3/4番目を第3四分位と呼ぶ。

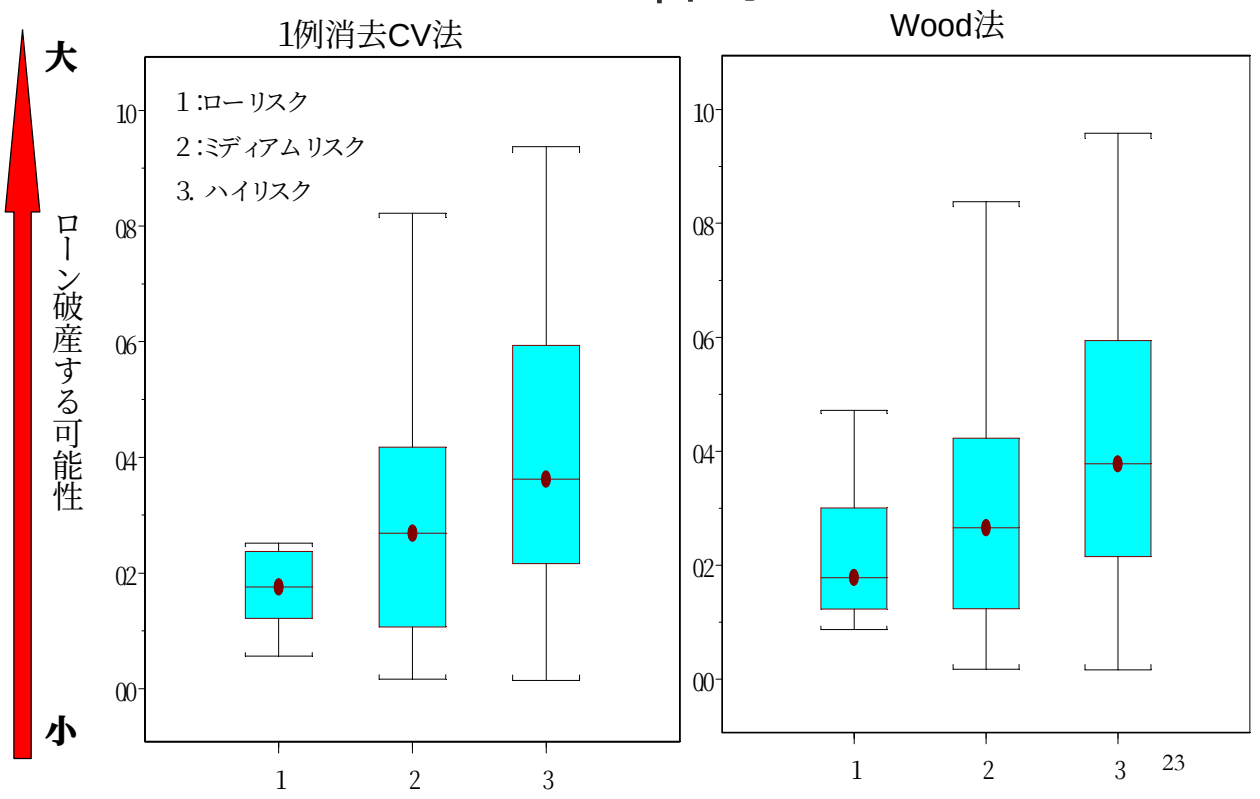
箱ひげ図で描いた場合を次ページに示す。





箱ひげ図 :見本

ローン目的


大
ローン破産する可能性
小
1例消去CV法
1 :ローリスク
2 :ミディアムリスク
3. ハイリスク
10
08
06
04
02
00
1
2
3
Wood法
10
08
06
04
02
00
1
2
3
23

影響分析

- 影響分析とはモデルの当てはめに悪影響を与えている影響観測値を検出する方法である。これを取り除くことでモデルの当てはまりは良くなり、より精密な予測が可能となる。

- 影響分析にはDIFDEVと呼ばれる情報量を算出し、その数値が検定の数値（6.63）よりも高くなれば、その検体はモデルに悪影響を与えているとみなし、除去対象となる。

$DIFDEV$ (***DIF****ference of* ***DEV****iance*)

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

$$Dev = -2\sum_{d=1}^{2000}\left\{ y^{\langle d \rangle} \ln \hat{p}^{<d>} + \left(1 - y^{\langle d \rangle}\right) \ln\left(1 - \hat{p}^{<d>}\right)\right\}$$

Dev : すべての個体を用いたときの逸脱度

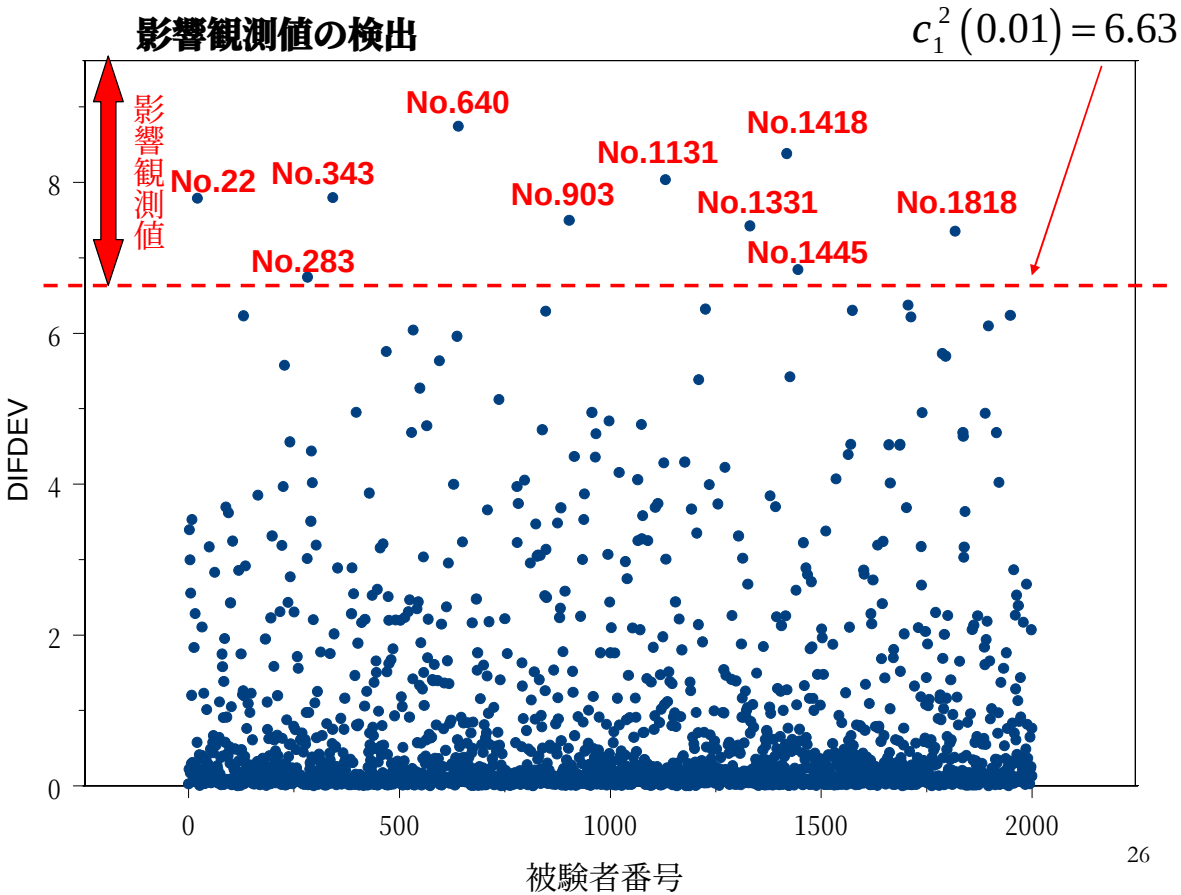
$Dev_{[d]}$: d番目の個体を取り除いたときの逸脱度

- **検定の方法**

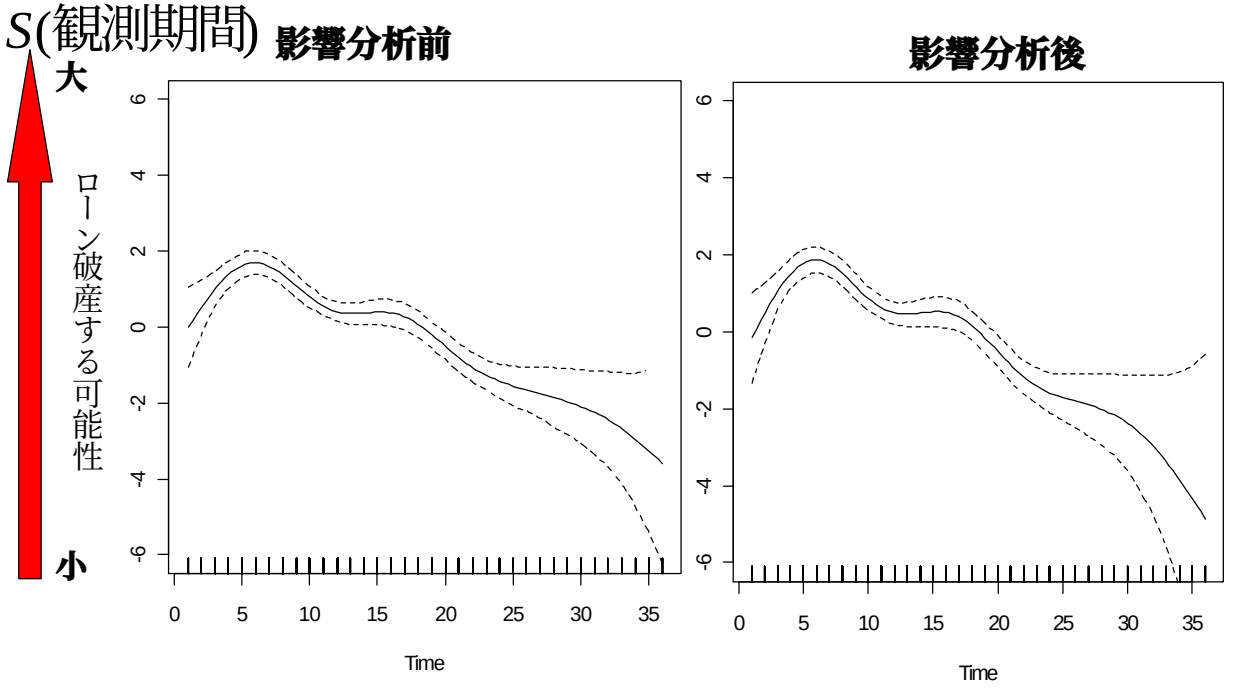
  自由度1のc^2分布に基づき、有意水準1%で検定

  $$\Delta Dev_{[d]} \geq c^2\left(0.01\right) = 6.63$$

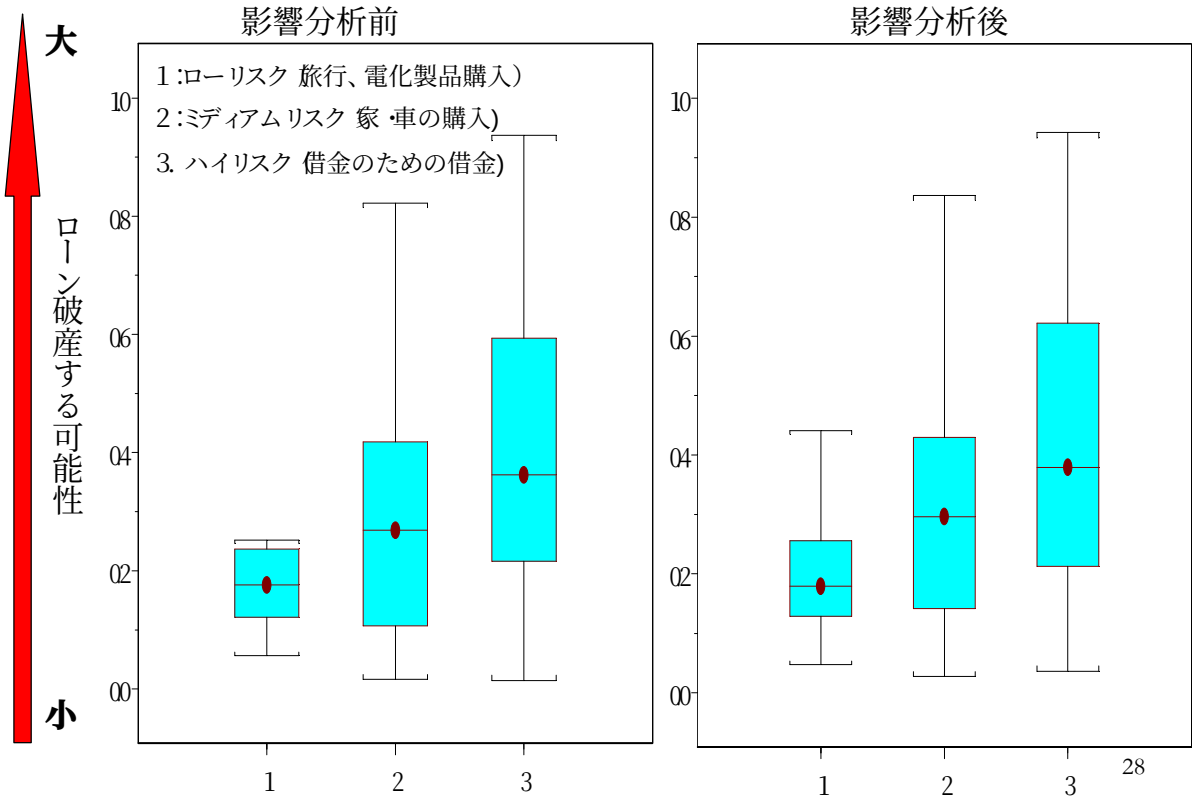
  ならば、d番目の個体を除去


影響観測値の検出
$c_1^2(0.01) = 6.63$
影響観測値
No.640
No.1418
No.1131
No.22
No.343
No.903
No.1331
No.1818
No.283
No.1445
DIFDEV
0
2
4
6
8
0
500
1000
1500
2000
被験者番号

観測期間における除去前後の比較



ローン目的における除去前後の比較(箱ひげ図)



研究発表は以上です。

ありがとうございました。