SASによるノンパラメトリック回帰

SAS インスティチュートジャパン
小野 裕亮 小玉 奈津子 泉水 克之

The Power to Know.

The Power to Know.

始めに

- 発表の内容
  - ノンパラメトリック回帰について
  - SASでノンパラメトリック回帰を行うプロシジャの紹介
    - Loess回帰を行う LOESSプロシジャ
    - 薄板平滑化回帰を行う TPSLINEプロシジャ
    - 一般化加法モデルを実行する GAMプロシジャ

2

ノンパラメトリック回帰について

The Power to Know.

The Power to Know.

線形回帰

◆ **モデル式**

● **応答変数***Y*、**説明変数**を*X*とした場合の**線形回帰**の場合

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

誤差 **ε** が**N(0,σ²)**の正規分布に従う場合は、回帰パラメータ*β*₀
*β*₁を最小2乗法を利用して推定する。

4

The Power to Know.

ノンパラメトリック回帰（1）

- モデル式
  - 応答変数Yを説明変数をXとした場合

$$Y = g(X) + \varepsilon$$

g（x） → 1 滑らかな関数

2 関数の型は未知

5

The Power to Know.

ノンパラメトリック回帰（2）

- 既存の方法 （Version6でも可能）
  - 多項式回帰 （REGプロシジャ、GLMプロシジャ）
  - 区分多項式 （TRANSREGプロシジャ）
  - モデルを指定した非線形回帰 （NLINプロシジャ）
    （ニューラルネットワーク （Enterprise Miner ））

6

The Power to Know.

ノンパラメトリック回帰（3）

- 追加された新機能（Version8から）

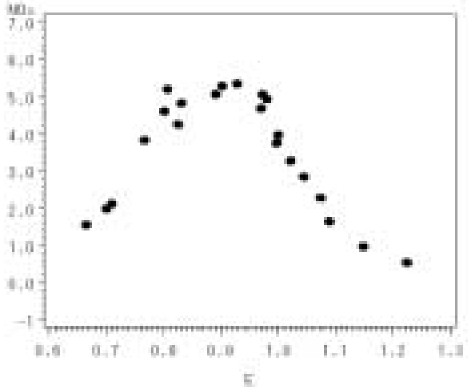
  - Loess（局所）回帰　（LOESSプロシジャ）
  - 薄板平滑化回帰　（TPSPLINEプロシジャ）
  - 一般化加法モデル（GAMプロシジャ）
    *** Version 8.1**では、評価版 Version8.2から製品版

7

The Power to Know.

ノンパラメトリック回帰の例（1）

- **Gas データ(Brinkman 1981)の例**



排気ガスに含まれる窒素酸化物の濃度（Nox）と、当量比（E）の関係に対して**多項式回帰**とLoess**回帰**を当てはめてみる。

・説明変数　当量比

当量比（エンジンで燃焼させるガスの空気とエタノールの混合比率）

・応答変数　窒素酸化物の濃度

8

ノンパラメトリック回帰の例（2）

- **多項式回帰**

- **モデル式**

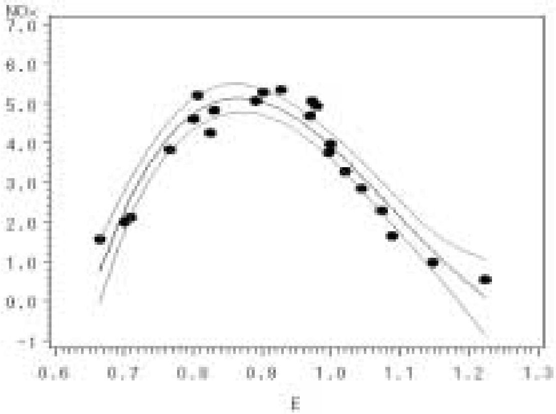
$$NOx = \beta_0 + \beta_1 E + \beta_2 E^2 + \beta_3 E^3 + \varepsilon$$

データの曲線の関係を示す為に、3次の項までの多項式回帰を行なう。

9

ノンパラメトリック回帰の例（3）

- **多項式回帰の結果（1）**



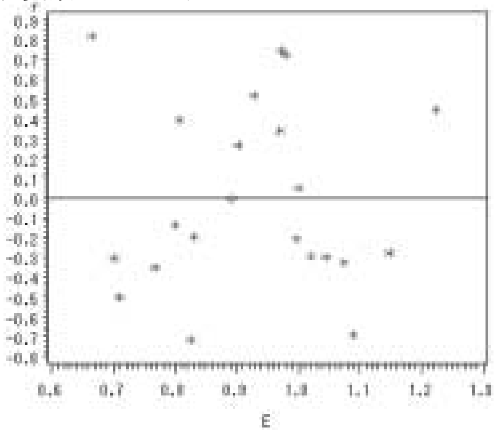
10

The Power to Know.

ノンパラメトリック回帰の例(3)

● **多項式回帰の結果（2）**

（残差と説明変数のプロット）



11

The Power to Know.

ノンパラメトリック回帰の例(4)

- **Loess**回帰

  ● **モデル式**

$$NOx = g(E) + \varepsilon$$

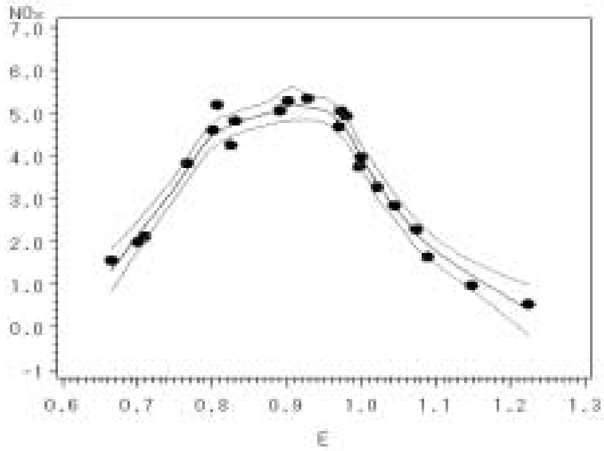
滑らかな関数 **g**( )を、**Loess回帰**を利用して推定する。

12

SAS e-Intelligence — The Power to Know.

ノンパラメトリック回帰の例（5）

- **Loess 回帰の結果（1）**



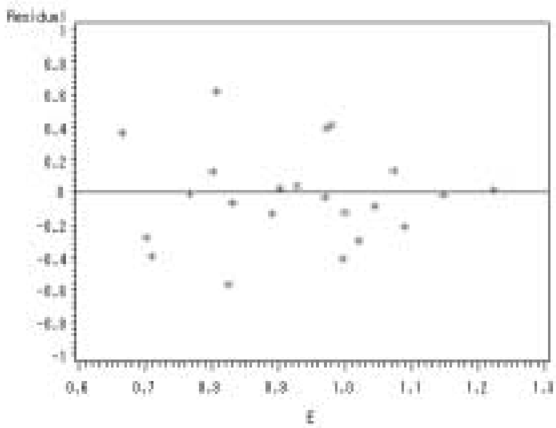
13

SAS e-Intelligence — The Power to Know.

ノンパラメトリック回帰の例（5）

- **Loess 回帰の結果（2）**

（残差と説明変数のプロット）



14

The Power to Know.

既存の方法

- 長所
  - 計算が簡単である
  - 結果の解釈が容易である
- 短所
  - モデル式を明示的に指定する必要がある
  - １つのデータ点がモデル全体に影響を与える

15

The Power to Know.

ノンパラメトリック回帰

- 長所
  - モデルを明示的に指定しなくてすむ。
（但し、滑らかさの度合いを決めなくてはいけない）
- 短所
  - モデルの解釈が複雑である
  - 大量なデータでは計算時間がかかる
  - 外挿ができない（補間のみ）

16

The Power to Know.

Loess（局所）回帰を行うLOESSプロシジャ

The Power to Know.

LOESSプロシジャ（１）

- LOESSプロシジャについて
  - Loess回帰（局所回帰）
  - 平滑化パラメータ
  - 平滑化パラメータの自動選択
  - 外れ値の存在や裾の重い誤差分布に対する工夫
  - 大容量データに対する工夫
  - データの標準化

18

The Power to Know.

LOESSプロシジャ（2）

- **Loess回帰（局所回帰）**

$$y_i = g(x_i) + \varepsilon_i$$

説明変数　x_i　は行ベクトル

非説明変数　y_i

近くの点ほど**大きな重み**を与えて、**各点**ごとに回帰分析を行う。

重みつけの方法によって、曲面の滑らかさが変化する。

19

The Power to Know.

LOESSプロシジャ（3）

- **平滑化パラメータ**
  - **ある点xにおいて、回帰分析を行う時にx_iに対して与える重みw_i**

平滑化パラメータs≦1の時

$$w_i = \begin{cases} \frac{32}{5}\left(1-\left(\frac{d_i(x)}{d_{(ns)}(x)}\right)^3\right)^3 & d_i(x) \le d_{(ns)}(x) \\ 0 & d_i(x) > d_{(ns)}(x) \end{cases}$$

・$d_i(x)$ は、xとx_i　の距離

・$d_{(ns)}(x)$ は、xからns番目のデータまでの距離

この関数は、近くに位置するものほど、重みが大きな値になる。

Sを0に近づけていくと、より近くのデータしか利用しなくなる。

20

sas e-Intelligence — The Power to Know.

LOESSプロシジャ（4）

平滑化パラメータ　S >1の時

$$w_i = \frac{32}{5}\left(1-\left(\frac{d_i(x)}{d_{(n)}(x)s^{\frac{1}{n}}}\right)^3\right)^3$$

Sの値を大きくすると、すべての重みが一定になっていき、

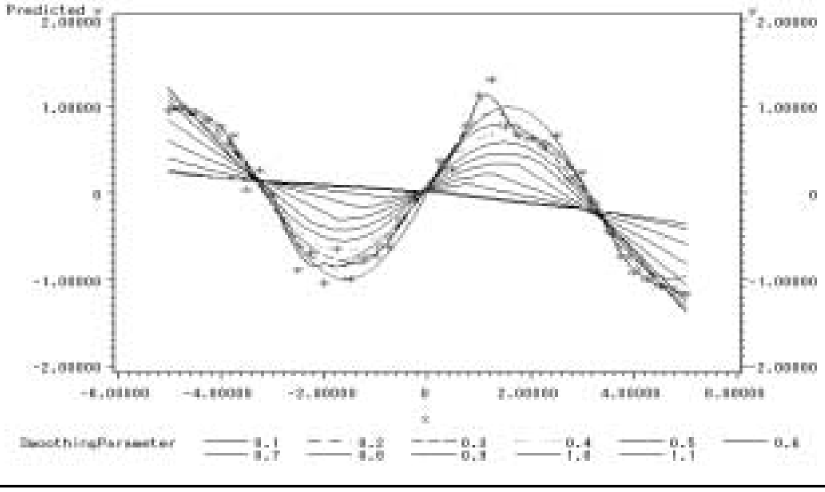
通常の回帰分析に近づいていく。

21

sas e-Intelligence — The Power to Know.

LOESSプロシジャ（5）

- 平滑化パラメータと滑らかさ



22

The Power to Know.

LOESSプロシジャ（6）

- **平滑化パラメータの自動選択（v8.2）**
  - **平滑化パラメータの選択基準**
    - **GCV（一般化クロスバリデーション）による選択**

      (データ数が少なく、誤差（ノイズ）が大きいようなデータに対しては、over fitting (= under smoothig)する傾向がある。）
    - **AIC$_{C1}$による選択**
    - **AIC$_{C}$による選択**

23

The Power to Know.

LOESSプロシジャ（7）

- **自由度による指定（v8.2）**

予測値 $\hat{y}$ は、行列Lを利用して**実測値** y の る。

通常の線形回帰の場合

L=x(xx\`)$^{-1}$x\`

$$\hat{y} = Ly$$

LOESSプロシジャでは、3種類の自由度をこの行列Lから計算する。

・$df1 = trace(L)$

・$df2 = trace(L^{`}L)$

・$df3 = 2trace(L) - trace(L^{`}L)$

24

The Power to Know.

LOESSプロシジャ(8)

- **外れ値の存在や裾の重い誤**

MODELステートメントの

ITERATIONS= **オプション**で
最大反復回数を指定

最小2乗基準を最小にする推定の他に、**Tukeyの双加重関数**.
(biweight function)が最小となるように**反復推定**を行い、外れ値や、
誤差の裾が重い場合に頑健性のある回帰を行なうことができる。

25

The Power to Know.

LOESSプロシジャ(5)

- **大容量データに対する工夫**

データ**すべての点**に対して**局所回帰**を行なうのではなく、**kd Tree法**
にもとづいて選択した点に対して、局所回帰を行なう。

MODELステートメントの

BUCKET= オプションで
点の数を指定

26

The Power to Know.

LOESSプロシジャ(6)

- **データの標準化**

Loess回帰の重みW_iは、距離によって決められる。

説明変数が複数ある場合は、標準化するか、しないかによって結果が変化する。

LOESSプロシジャでは、元データのまま利用す 分析をするかが選択することが可能。

MODELステートメントの SCALE=SD( ) オプションでトリムする 割合を指定

標準化には、**トリム化標準偏差**が使われる。

27

The Power to Know.

LOESSプロシジャ(9)

- LOESSプロシジャで平滑化パラメータを自動選択する実行例

- **データ** → Sinカーブに誤差を加えて作成したデータ

```
data sample;
   do x=-5 to 5 by 0.25;
      y=sin(x)+rannor(12345)*0.4;
      output;
    end;
run;
```

28

The Power to Know.

LOESSプロシジャ(10)

●プログラム1

```
/***GCVによる平滑化パラメータの選択**/
  proc loess data=sample;
     model y=x /select=gcv ;
     ods output OutputStatistics=statsgcv;
  run;
```

29

The Power to Know.

LOESSプロシジャ(11)

●プログラム2

/***平滑化パラメータをAIC_C基準により選択**/

```
  proc loess data=sample;
     model y=x /select=aicc;
     ods output OutputStatistics=statsaicc;
  run;
```

30

sas e-Intelligence The Power to Know.

TPSPLINEプロシジャ(1)

キーワード

- Thin-Plate Spline（薄板スプライン）
- ペナルティ付き最小2乗法
- セミパラメトリック
- GCV functionを用いたパラメータ選択

31

sas e-Intelligence The Power to Know.

TPSPLINEプロシジャ(2)

- 最小2乗法

$$y_i = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon_i, \quad \varepsilon_i : i.i.d.\ N(0, \sigma^2)$$

残差平方和 $\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=1}^{p} \beta_i x_i)^2$ を最小とする

$\beta_0, \beta_1, \ldots, \beta_p$ を求める

32

The Power to Know.

TPSPLINEプロシジャ(3)

- **ペナルティ付き（罰則付き）最小2乗法(1)**

$$y_i = f(x_i) + z_i\beta + \varepsilon_i$$

x_i, z_i	説明変数（ベクトル）
y_i	被説明変数
f	滑らかな関数（未知）
β	パラメータ（ベクトル）

33

The Power to Know.

TPSPLINEプロシジャ(4)

- **ペナルティ付き（罰則付き）最小2乗法(2)**

“ペナルティ付き残差平方和”

無限の面積を持つ弾性薄板を凹ますときのエネルギーがこの形

$$D^{\alpha} f(x) = \sum \frac{\alpha}{\alpha_1! \cdots \alpha_d!}\left[\frac{\partial^{\alpha}}{\partial x_1^{\alpha_1} \cdots \partial x_1^{\alpha_d}} f(x)\right]$$

$$S_\lambda(\beta, f) = \frac{1}{n}\underbrace{\sum_{i=1}^{n}(y_i - z_i\beta - f(x_i))^2}_{\text{Goodness of fit}} + \lambda \underbrace{\int (D^m f(x))^2 dx}_{\text{Smoothness of fit}}$$

0 **rough** λ **smooth** ∝

34

sas e-Intelligence

The Power to Know.

TPSPLINEプロシジャ(5)

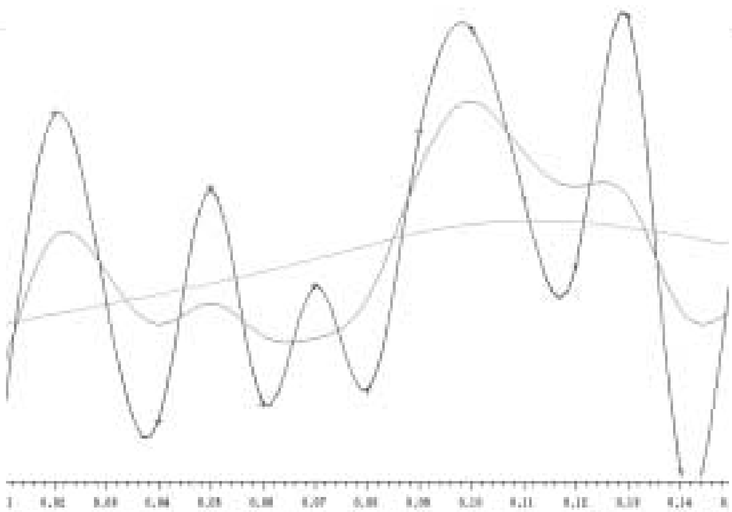
- **ペナルティ付き（罰則付き）最小2乗法(3)**

m=2, d=1, λ ↓ 0のとき
結果はcubic spline!
(SAS/IMLのSPLINE関数,
SAS/INSIGHT,
PROC GPLOT)

35

sas e-Intelligence

The Power to Know.



36

TPSPLINEプロシジャ(6)

- ペナルティ付き（罰則付き）最小2乗法(4)

λ はgiven

$S_\lambda(\beta, f)$ を最小化する！

37

TPSPLINEプロシジャ(7)

- β と f の決め方(roughly)

1. $f(x_i) = \theta_0 + \theta \cdot x_i + \sum_{j=1}^{n} \delta_j E(x_i - x_j)$

$E(y) = \mathrm{const} \cdot \| y \|^2 \log(\| y \|)$

と書ける

2. $S_\lambda(\beta, f)$は行列の形に落ちる

38

TPSPLINEプロシジャ(8)

問題

λをどうやって決めればいいの？？？

PROC TPSPLINEにおける回答

⇒⇒⇒GCV function
（一般化クロスバリデーション関数）

39

TPSPLINEプロシジャ(9)

- **GCV**(Generalized Cross Validation)**function**

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}(y_i - f_\lambda(x_i))^2}{(n - \mathrm{Trace}(A(\lambda)))^2}$$

$A(\lambda)$は$\hat{y} = A(\lambda)y$ **を満たす行列**

GCVの値を最小にするλを選択する

$$GCV \approx CV$$

$$CV(\lambda) = \sum_{i=1}^{n}(y_i - f_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\frac{(y_i - f_\lambda(x_i))^2}{(1 - a_{ii})^2}$$

40

The Power to Know.

余話

**GCVで選択しても、余りよろしくない結果が
出てしまう、との報告もある(**Hurvich, Simonoff, and Tsai, 1998)

AICC??

AICC$_1$???

41

The Power to Know.

TPSPLINEプロシジャ(10)

- **TPSPLINEプロシジャの実行例**
- **データ～「米国における硫酸塩の堆積量」**
- **説明変数～観測地点の緯度(latitude)と
経度(longitude)**
- **非説明変数～硫酸イオンの濃度(** g/m^2 **)**

42

TPSPLINEプロシジャ(11)

```
/*ノンパラメトリックモデル*/
proc tpspline data=so4;
 model so4 = (latitude longitude);
 score data=pred out=tp_out;
run;
```

43

TPSPLINEプロシジャ(12)

```
/*セミパラメトリックモデル*/
proc tpspline data=so4;
 model so4 = latitude (longitude);
 score data=pred out=tp_out;
run;
```

線形項
はこちら

ノンパラメトリック
な効果は（ ）の中

44

GAMプロシジャ(1)

- **<u>G</u>eneralized <u>A</u>dditive <u>M</u>odels （一般化加法モデル）**
- **Local Scoring Algorithms**
- **GCV**

45

GAMプロシジャ(2)

- **Generalized Additive Models**

ーイメージー

一般化線形モデル(GENMOD)

+

加法モデル

+

LOESS,TPSPLINE

46

GAMプロシジャ(3)

- **加法モデル**

$$Y = \beta_0 + s_1(X_1) + s_2(X_2) + \cdots + s_p(X_p) + \varepsilon$$

$s_i(x), i = 1,2,\ldots, p$は任意の関数

$$\updownarrow$$

- **（線形モデル）**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p + \varepsilon$$

47

GAMプロシジャ(4)

- **一般化線形モデル**

$$\eta = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p$$

$\mu = E[Y] \quad Y \sim$ 指数分布族

$\eta = g(\mu) \quad g(x)$はlink関数 $\mathrm{g(x) = x, logit(x), log(x)}, \ldots$

48

The Power to Know.

GAMプロシジャ(5)

SPLINE,LOESSを指定できる

- **一般化加法モデル(GAM)**

$$\eta = \beta_0 + s_1(X_1) + \cdots + s_p(X_p)$$

$$s_i(x) \text{は} smoothing\ \ function$$

$$E[s_i(X_i)] = 0$$

$$\mu = E[Y] \quad Y\text{〜指数分布族}$$

$$\eta = g(\mu) \quad g(x)\text{は`canonical' link関数}$$

49

The Power to Know.

GAMプロシジャ(6)

- **推定法**

Local Scoring Algorithms

動機~ Backfitting Algorithms
+重み付き反復最小2乗法

$$"B.A"\text{- - -} idea$$

$$R_j \overset{def}{=} Y - s_0 - \sum_{k \neq j} s_k(X_k) \text{と定めると、} E[R_j \mid X_j] = s(X_j)$$

50

The Power to Know.

GAMプロシジャ(7)

- λ の決め方

⇒⇒GCV

51

The Power to Know.

GAMプロシジャ(8)

```
/*GAMプロシジャ*/
PROC GAM data=so4;
model so4=spline(longitude)
      spline(latitude)/method=gcv;
score data=pred out=gam_out;
run;
```

関数として
SPLINE,SPLINE2,
LOESSを指定できる

52

The Power to Know.

GAMプロシジャ(9)

～まとめ～
LOESS,TPSPLINEと比較して

- 結果の解釈が比較的行いやすい
  ～その「加法性」より
- 計算時間が相対的に短い
- 得られる結果が芳しくない時があるので注意!!

53

The Power to Know.

終わりに

- ノンパラメトリック回帰を行うことにより、次のことが可能でしょう。
  - A) 特定のモデルを指定せずに、予測や補間を行う。
  - B) より精度の高い予測式の構築
- 課題
  - 平滑化パラメータの選択方法
  - 大規模データへの適用

54

参考文献(1)

- LOESS、TPSPLINE、GAMプロシジャの開発者による紹介
http://www.sas.com/rnd/app/papers/papers_da.html
- Generalized Additive Models
Hastie&Tibshirani,　Chapman&Hall
- Nonparametric Regression and Generalized Linear Models
Green&Silverman,　Chapman&Hall

55

参考文献(2)

- 平滑化とノンパラメトリック回帰への招待
S･シモノフ著　竹澤＆大森訳
農林統計協会

56

sas®

e-Intelligence

The Power to Know™