Proc GAMの紹介

東京大学大学院医学系研究科生物統計学
田中司朗*
松山裕

発表のアウトライン

- **イントロダクション**
  - Proc GAM**の特徴**
  - **散布図の平滑化**
- GAM（generalized additive model, **一般化加法モデル**）
  - **平滑化手法**（LOESS, **平滑化スプラインなど**）
  - **推定アルゴリズム**
  - **自由度の設定**
- Proc GAM**の文法・出力**
- **解析事例**: Kyphosis（**脊柱後弯症**）**データ**

Proc GAMで何が出来るか?

- 散布図の平滑化
- 柔軟な交絡調整
- 結果変数の予測
- 用量反応関係の把握
- …

Proc GAMの特徴

- **多機能な回帰分析を行うプロシジャ**
  - **結果変数: 連続データ, 2値データ, 計数データ...**
  - **平均構造のノンパラメトリックなモデル化**
- **結果のグラフィカルな提示に有用**
- **比較してみると...**
  - Proc LOESS
  - Proc GENMOD

Proc GAMと他のプロシジャとの比較

		結果変数の型	
		連続データ	カテゴリカルデータ
平均構造のモデル	パラメトリック	LOESS GENMOD GAM	GENMOD GAM
	ノンパラメトリック	LOESS GAM	GAM

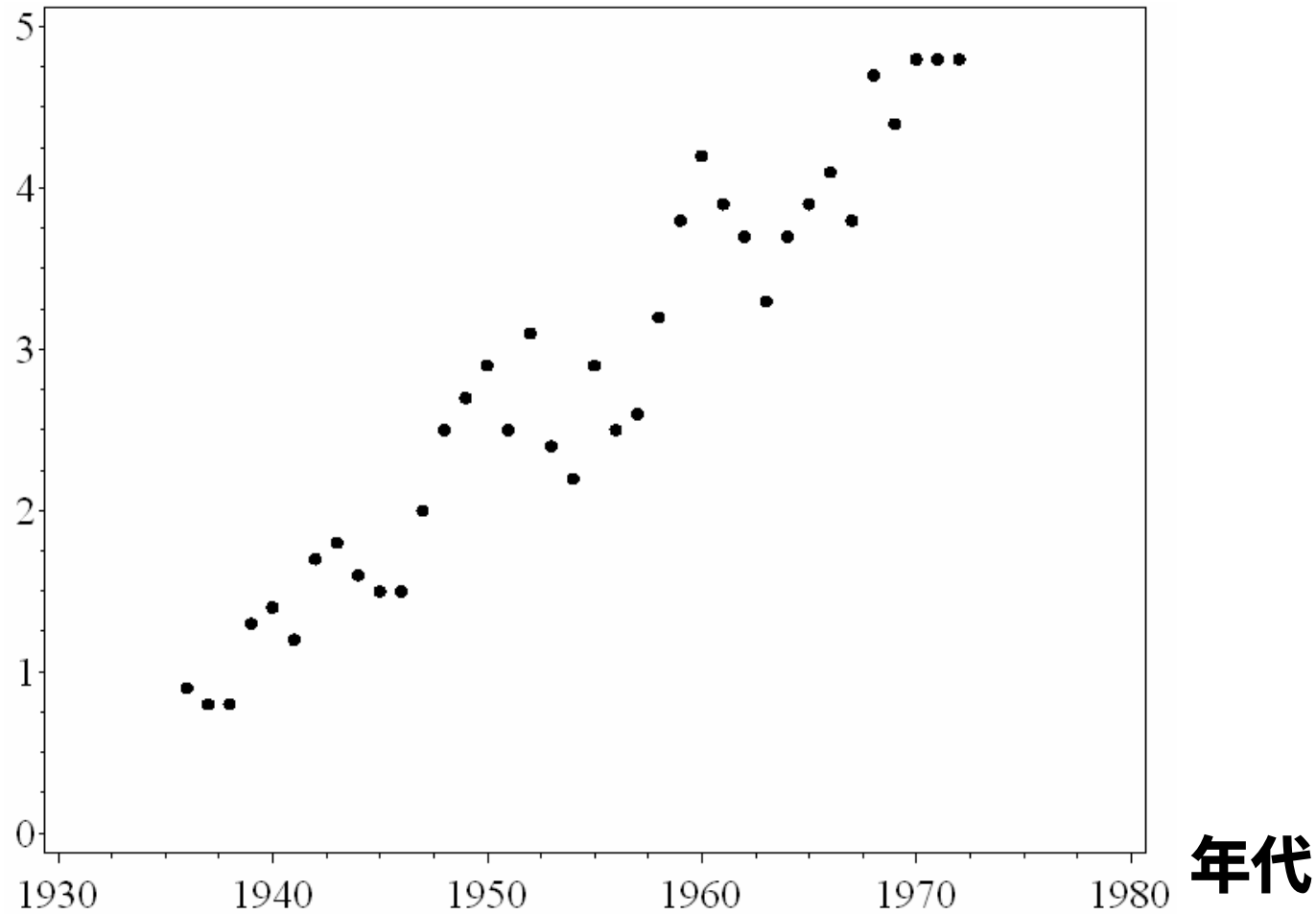
歴史的な流れ

- 70年代: 一般化線型モデル
  - カテゴリカルデータのモデルなどを統一的に表現
- 80年代: ノンパラメトリック回帰の発展
  - 探索的データ解析の一手法として
  - 散布図の平滑化: 連続量
- GAM
  - Hastie T, Tibshirani R. Generalized Additive Models, Chapman and Hall: London; 1990.
  - 古典的なノンパラメトリック回帰を
    カテゴリカルデータの場合に拡張

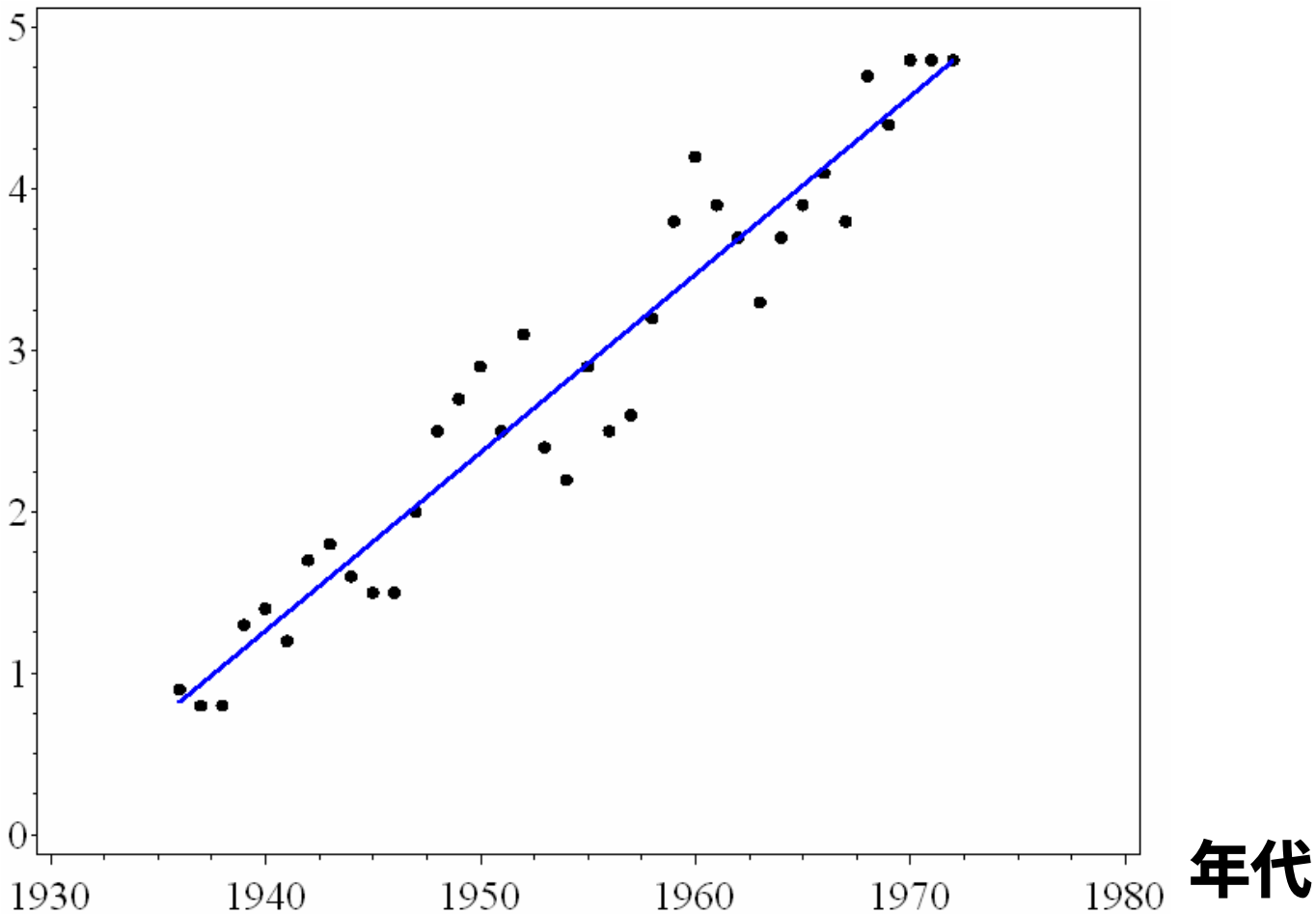
Melanoma (黒色腫) データ (Houghton D, et al)

- コネチカット州がん登録データベースによる記述疫学
  - 1936年から1972年までの悪性黒色腫発生の推移
  - 結果変数: 10万人あたりの年齢調整発生数
  - 説明変数: 年代

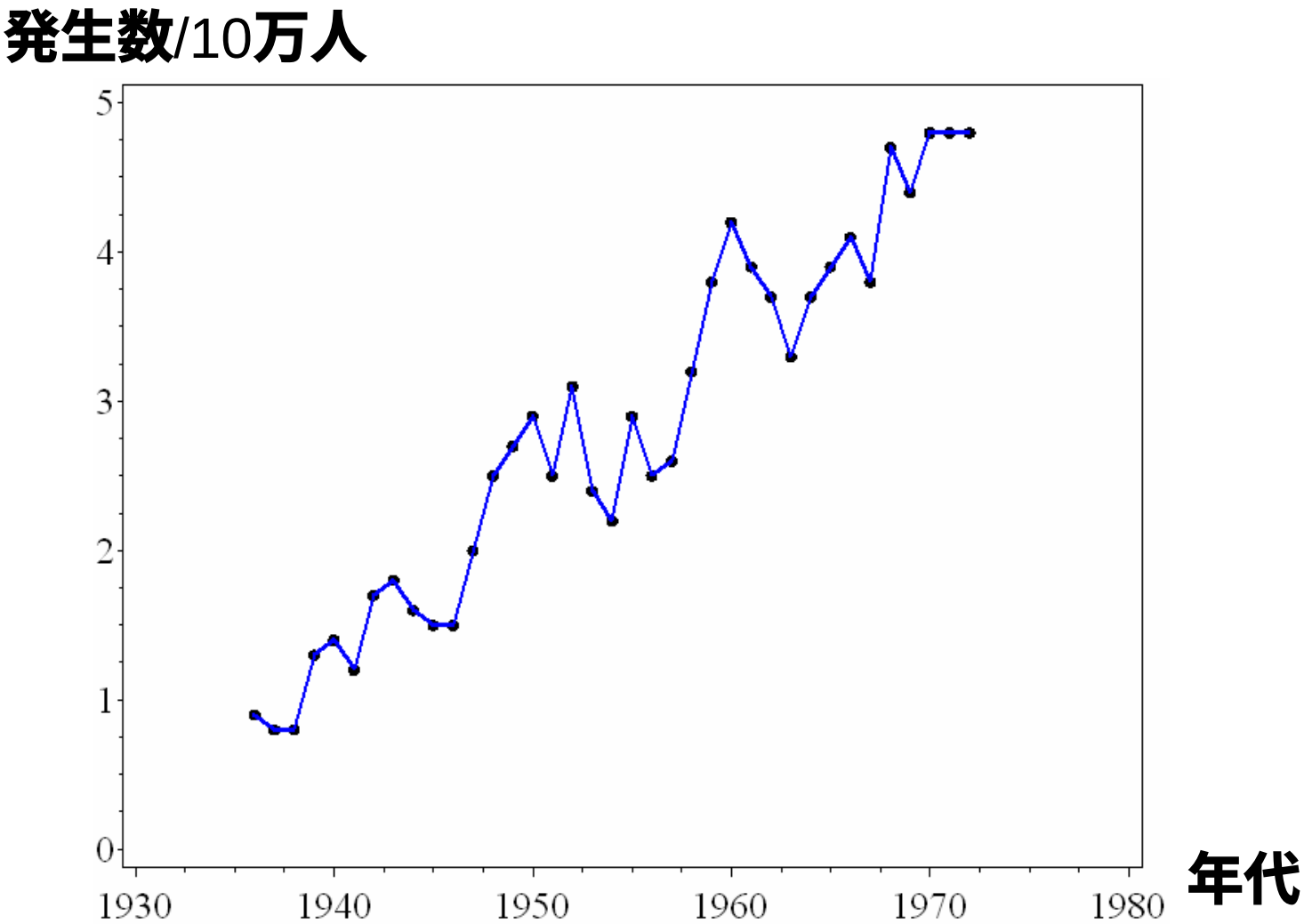
Melanomaデータの散布図



回帰分析の出力

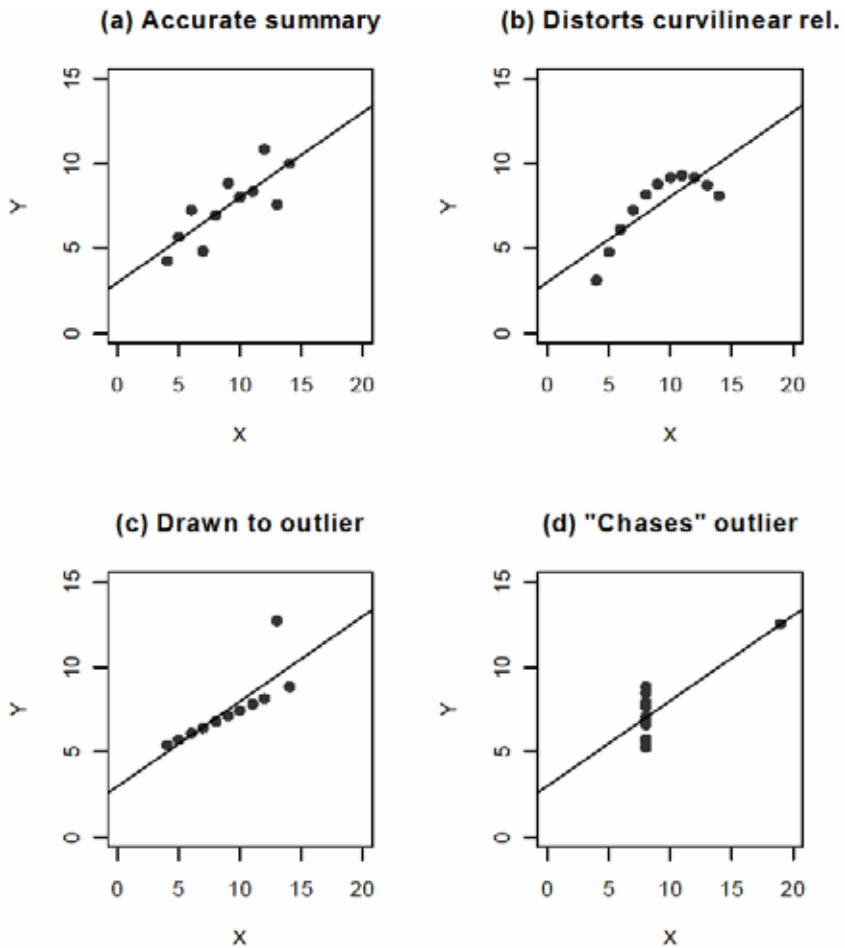


直線では捉えきれない曲線的な傾向



回帰分析: 線型性の仮定

- Anscombe 1973の例
  - 回帰分析の出力は同じ
    - 回帰係数
    - 残差平方和...
  - グラフを正しく要約しているのは(a)だけ
    - (b): **非線型**
    - (c), (d): 外れ値



ノンパラメトリック: 線型性の仮定を緩和

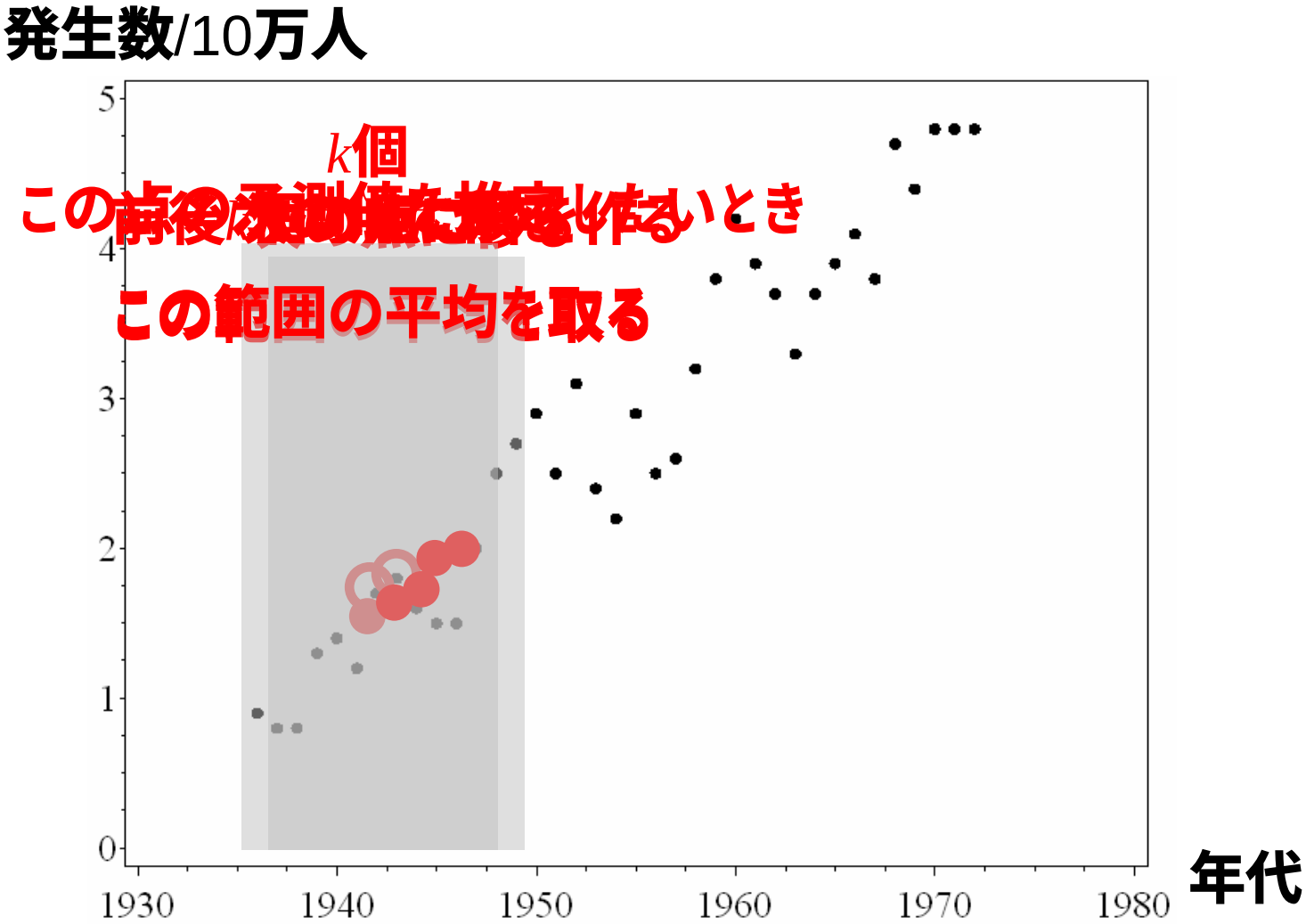
- **単回帰モデル**
  - yとxの関係を一次関数として特定
  - 推定する対象は**回帰係数** (β_0, β_1)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

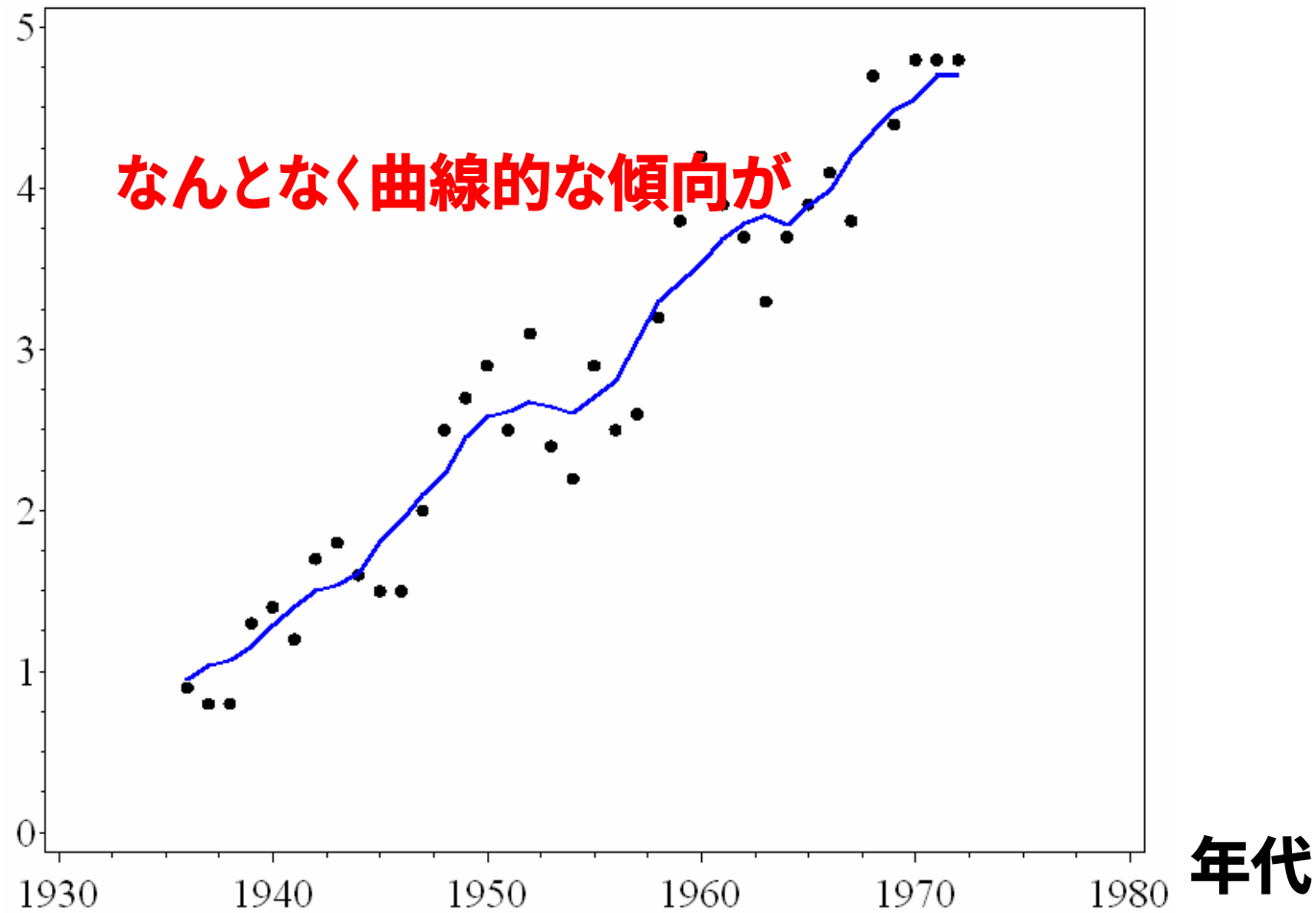
- **ノンパラメトリック回帰のモデル**
  - $s(x)$は特定しない (何らかの非線型の関数でよい)
  - **関数の形**をデータから推定する

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

単純なノンパラメトリック回帰 (移動平均)



移動平均の出力


発生数/10万人
なんとなく曲線的な傾向が
5
4
3
2
1
0
1930
1940
1950
1960
1970
1980
年代

移動平均

- **重み付き平均**の形で表される

$$\hat{s}(x_j) = \frac{\sum_{i=1}^{n} w_{ij} y_i}{k}$$

$$w_{ij} = \begin{cases} 1 & (x_i \text{が} x_j \text{の} k \text{近傍に含まれるとき}) \\ 0 & (\text{含まれないとき}) \end{cases}$$

Proc GAMでは

- **散布図の平滑化が可能**
- **より性能のよい平滑化手法を採用**
  - LOESS
  - **平滑化スプライン**

GAM（一般化加法モデル）

- 古典的な散布図平滑化の手法を
  カテゴリカルデータへ拡張
- まずは一般化線型モデルについて…

一般化線型モデル

- 線型モデルやカテゴリカルデータ解析のモデルを統一的に表現
  - ロジスティック回帰, Poisson回帰...
- 2つの構成要素から成る
  - サンプリングモデル
  - 平均構造のモデル (GAMではこの部分を拡張)

サンプリングモデル

- **結果変数の従う分布は指数型分布族**
  - **正規分布, 二項分布, Poisson分布などを含む**

平均構造のモデル

- **結果変数Y_iの平均と説明変数X_iとの関係**
  - **ただし、**$\mathrm{E}(y_i)=\mu_i$, $g(\mu_i)$**はリンク関数**

$$g(\mu_i)=X_i\beta$$

  - **線型回帰**: $g(\mu_i)=\mu_i$
  - **ロジスティック回帰**: $g(\mu_i)= \log[\mu_i/(1-\mu_i)]$

GAM: 平均構造に加法モデルを仮定

- 結果変数Y_iの平均と説明変数X_iとの関係
  - ただし、$\mathrm{E}(y_i)=\mu_i$, $g(\mu_i)$はリンク関数

$$g(\mu_i) = s_0 + \sum_{j=1}^{p} s_j(x_{ij})$$

  - $s_j(x)$は特定しない（何らかの非線型の関数でよい）
  - 関数の形をデータから推定する

発表のアウトライン

- **イントロダクション**
  - Proc GAM**の特徴**
  - **散布図の平滑化**
- GAM（generalized additive model, **一般化加法モデル**）
  - **いくつかの平滑化手法**（LOESS, **平滑化スプラインなど**）
  - **推定アルゴリズム**
  - **自由度の設定**
- Proc GAM**の文法・出力**
- **解析事例**: Kyphosis（**脊柱後弯症**）**データ**

LOESS
⟨Locally Weighted Scatterplot Smoother⟩

- **移動平均では**
  - **推定する点の近くには1, 遠くには0の重みを与えて平均を取ったらうまくいった**

- **LOESSの2つの工夫**
  - **最小二乗法による推定**
  - **よりよい重み**
    - **局所重み付き** ⟨Locally Weighted⟩ **最小二乗推定**

LOESSの最小二乗推定

- 以下の重み付き残差平方和を最小にする推定量
  - w_{ij}は重み（後述）

$$RSS = \sum_{i=1}^{n} w_{ij}[y_i - [\underbrace{\beta_{0j} + \beta_{1j}(x_i - x_j)}]]^2$$

単回帰のモデルを当てはめる

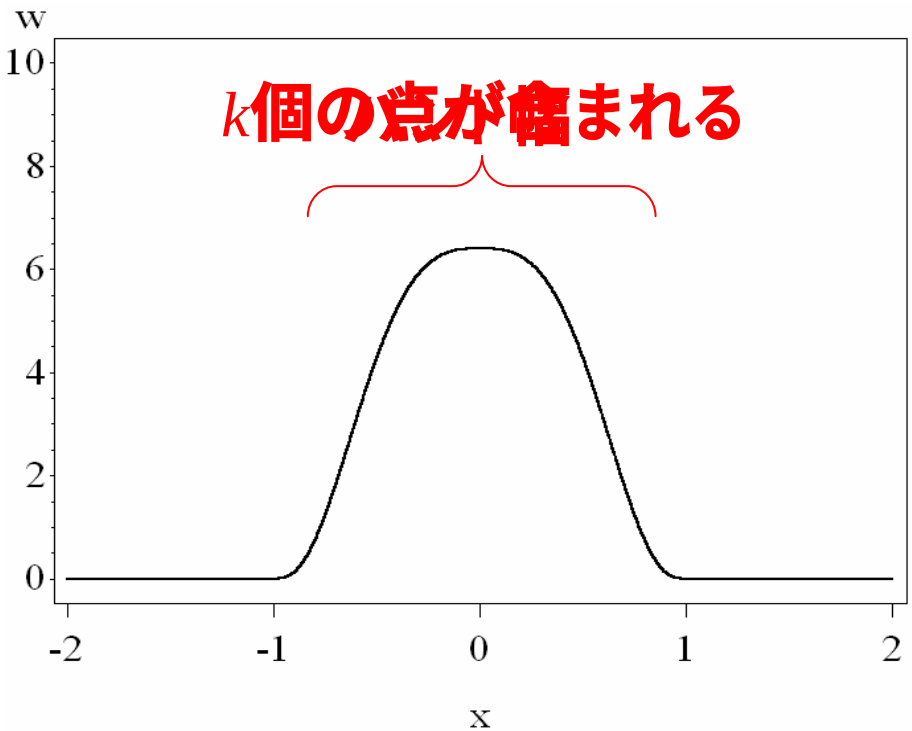
LOESSの重み

- **数式で表すと複雑だが…**

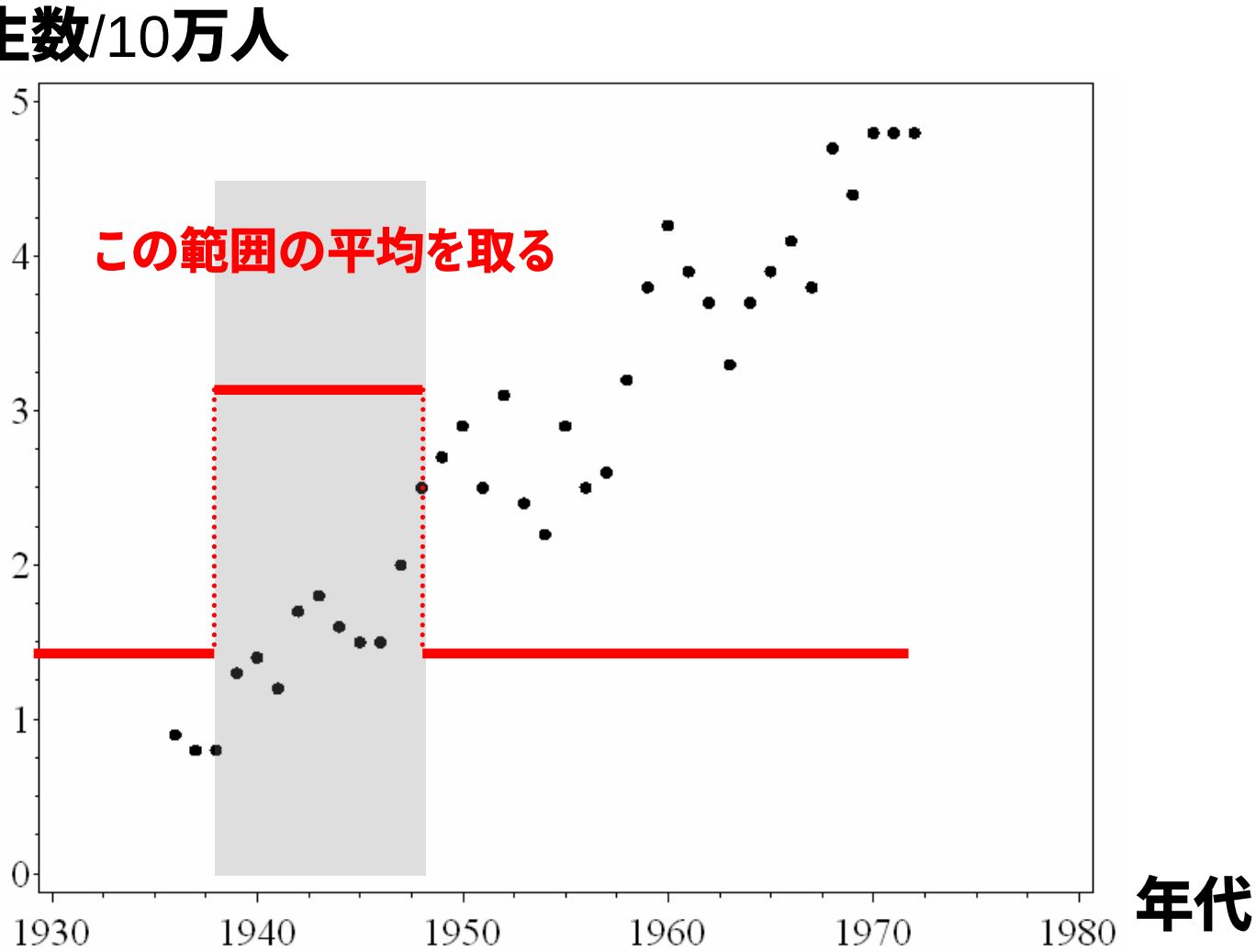
$$w_{ij} = \begin{cases} \dfrac{32}{5}\left[1-\left(\dfrac{|x_i - x_j|}{\max|x_i - x_j|}\right)^3\right]^3 & (x_i\text{が}x_j\text{の}k\text{近傍に含まれるとき}) \\ 0 & (\text{含まれないとき}) \end{cases}$$

LOESSの重み

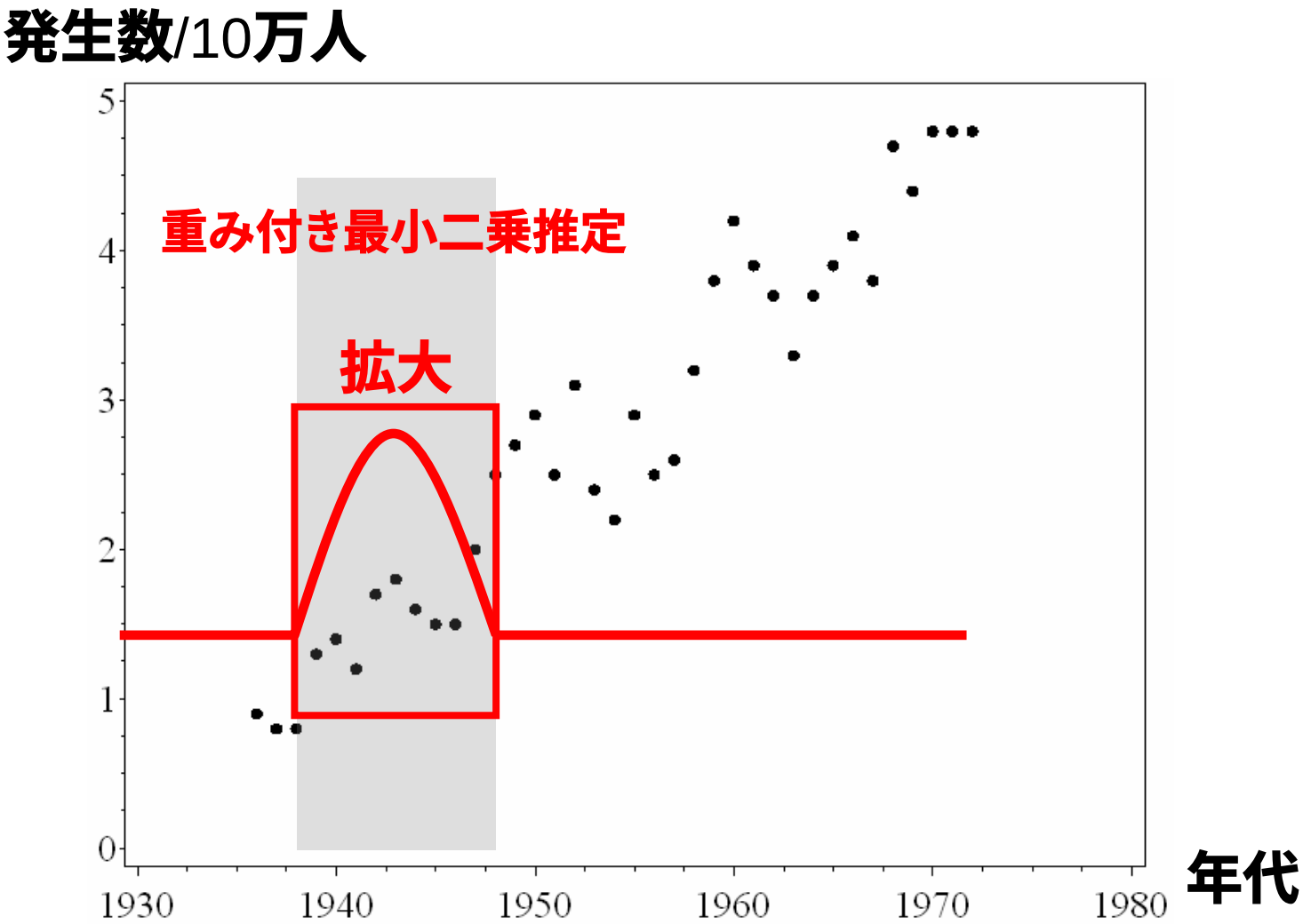
- 重みをグラフにしてみると
  - 近くには大きい重み、遠くには小さい重み



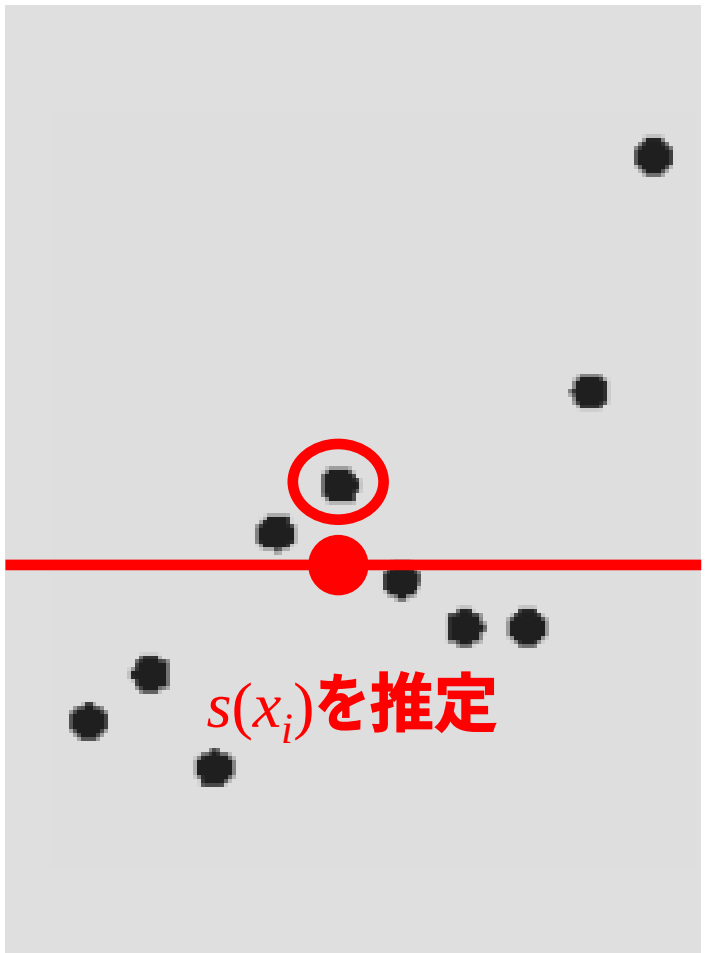
重みのイメージ (移動平均)



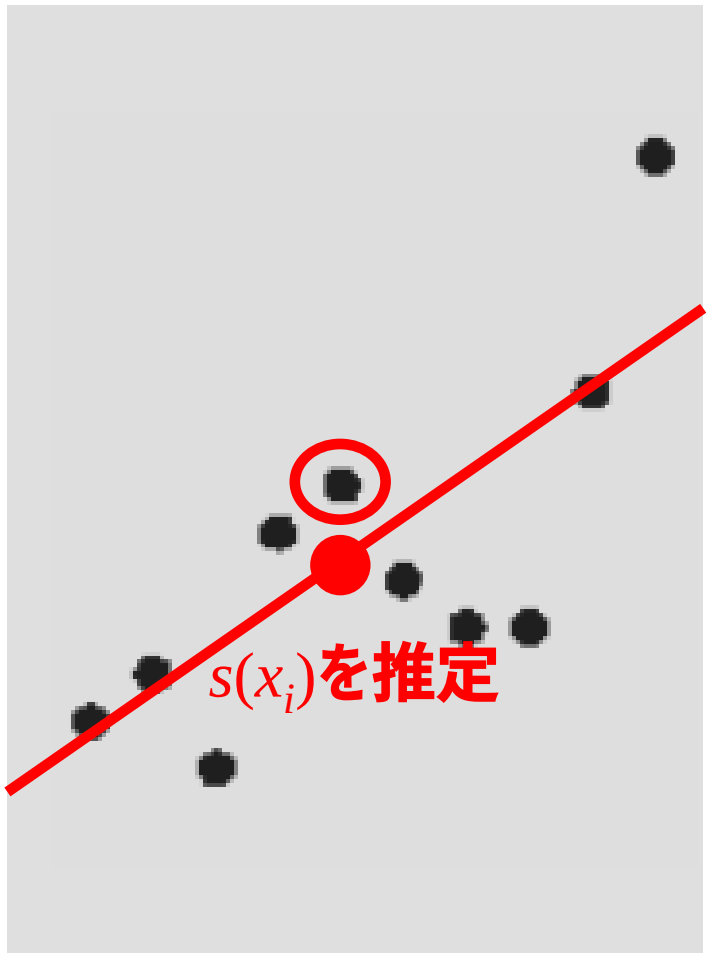
重みのイメージ (LOESS)



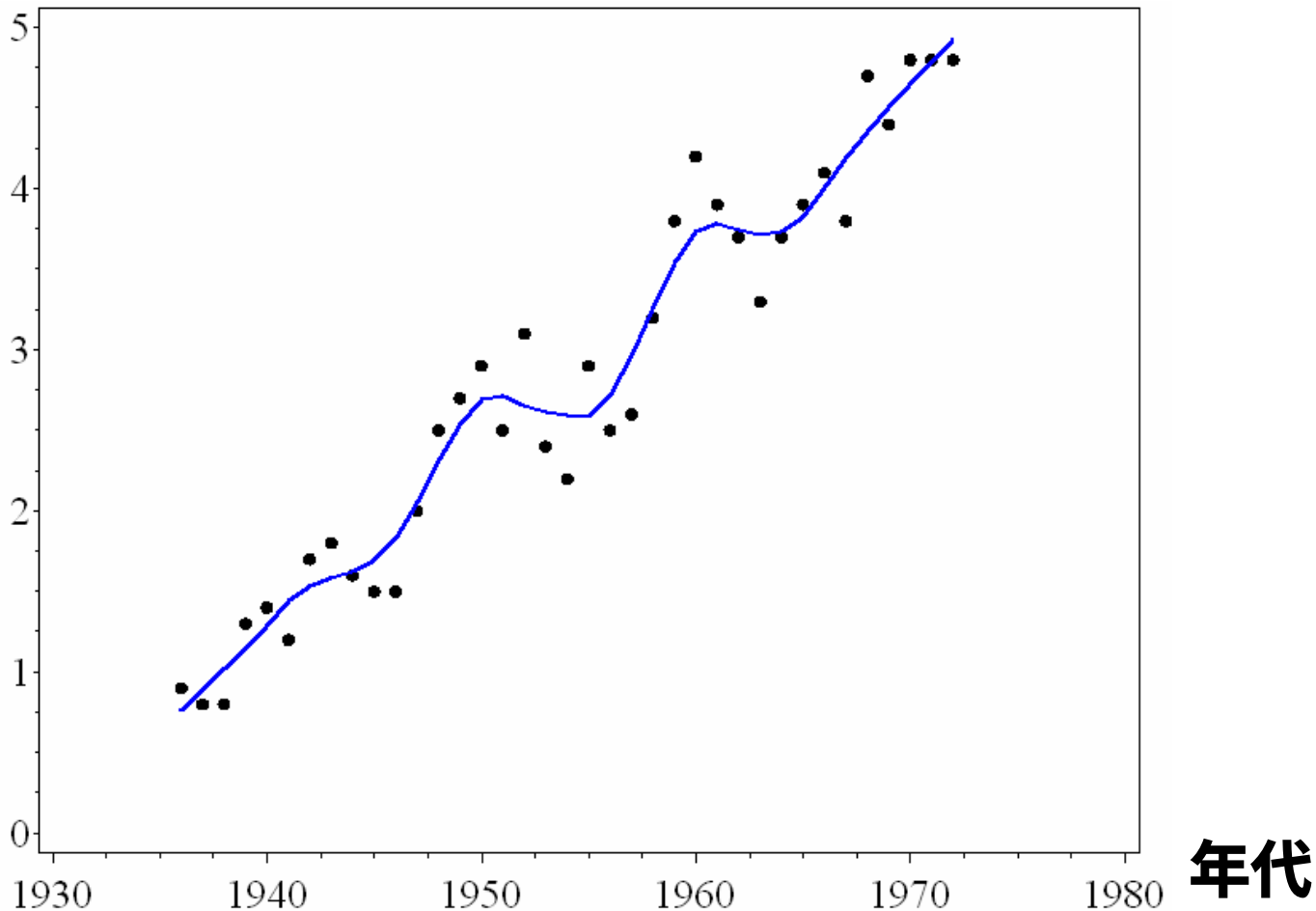
局所での推定のイメージ (移動平均)



局所での推定のイメージ（LOESS: 単回帰）



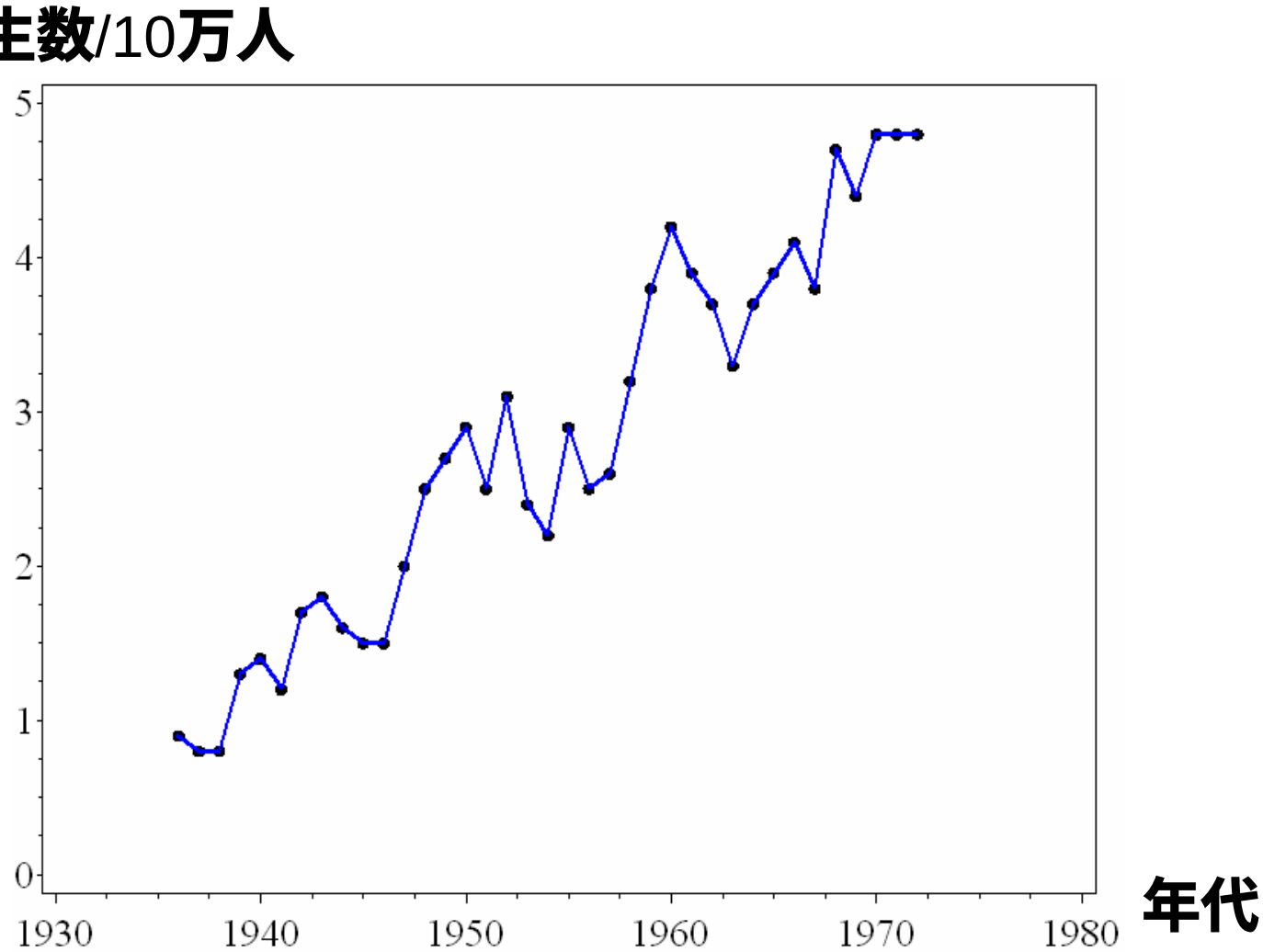
LOESSの出力



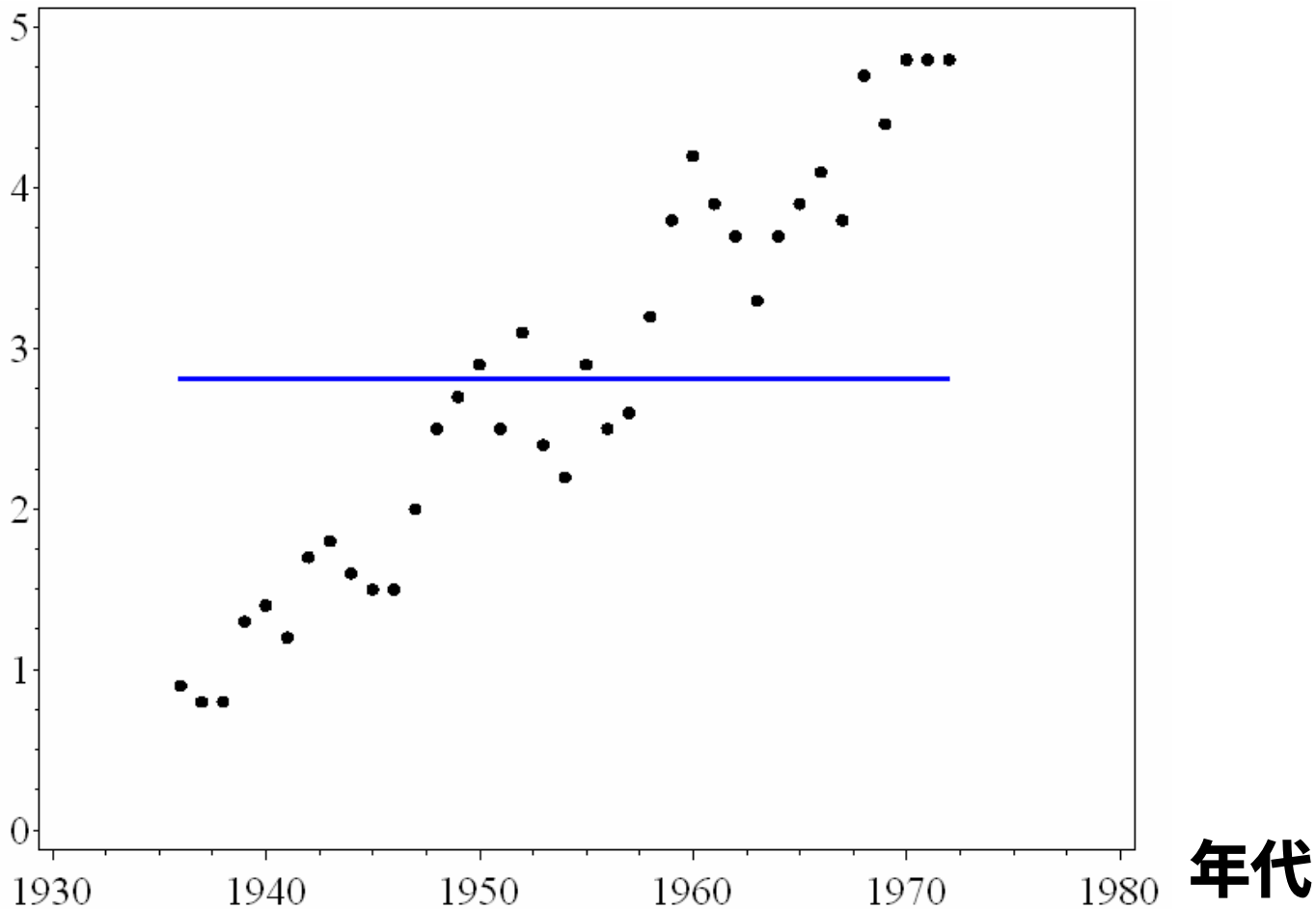
移動平均はk個の点の平均

- $k=1$のとき
  - $s(x_i)$の推定量は、その点自身（自由度はn）
  - バイアスはないが、推定精度が悪い
- $k=n$のとき
  - $s(x_i)$の推定量は、全体平均（自由度は1）
  - 推定精度はよいが、バイアスがある
- 精度とバイアスのトレードオフは**バンド幅**で決まる

k=1のとき（バイアスはないが精度が悪い）



$k=n$のとき (精度はよいがバイアスがある)



バンド幅の意味

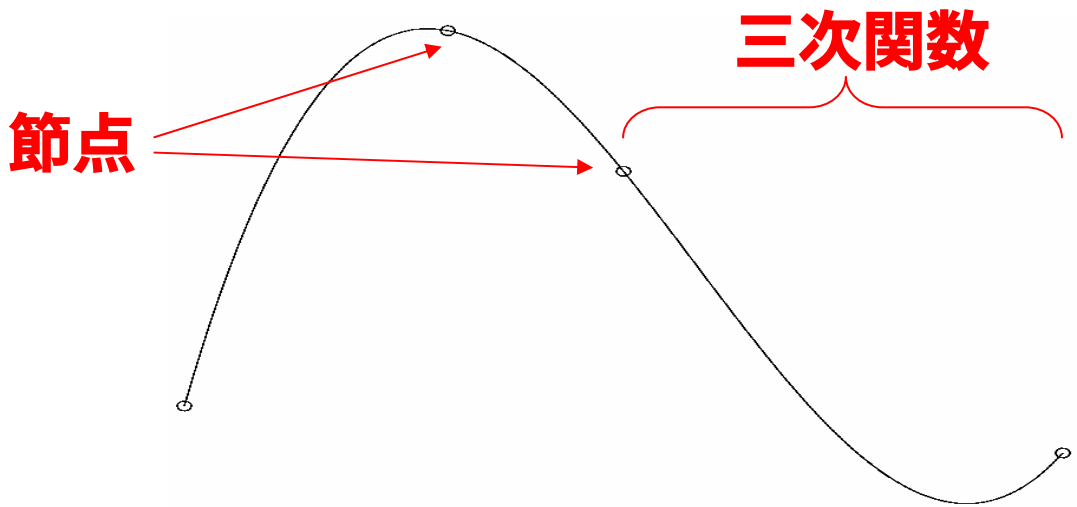
- バンド幅を決めるkは**平滑化パラメータ**
  - 平滑化の程度を規定する
  - 推定精度とバイアスのトレードオフ
  - 解析者が指定するチューニングパラメータ

発表のアウトライン

- **イントロダクション**
  - Proc GAM**の特徴**
  - **散布図の平滑化**
- GAM（generalized additive model, **一般化加法モデル**）
  - **いくつかの平滑化手法**（LOESS, **平滑化スプライン**など）
  - **推定アルゴリズム**
  - **自由度の設定**
- Proc GAM**の文法・出力**
- **解析事例**: Kyphosis（**脊柱後弯症**）**データ**

三次のスプライン関数

- 区分多項式の一種
  - 範囲を適当に区分する
  - 区間ごとに三次関数を当てはめる
  - ただし, 区間の繋ぎ目 (節点) では連続で滑らか



平滑化スプライン

- 以下のペナルティ付き残差平方和を最小とする推定量

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

- xの実現値を節点とした三次の自然スプラインが解

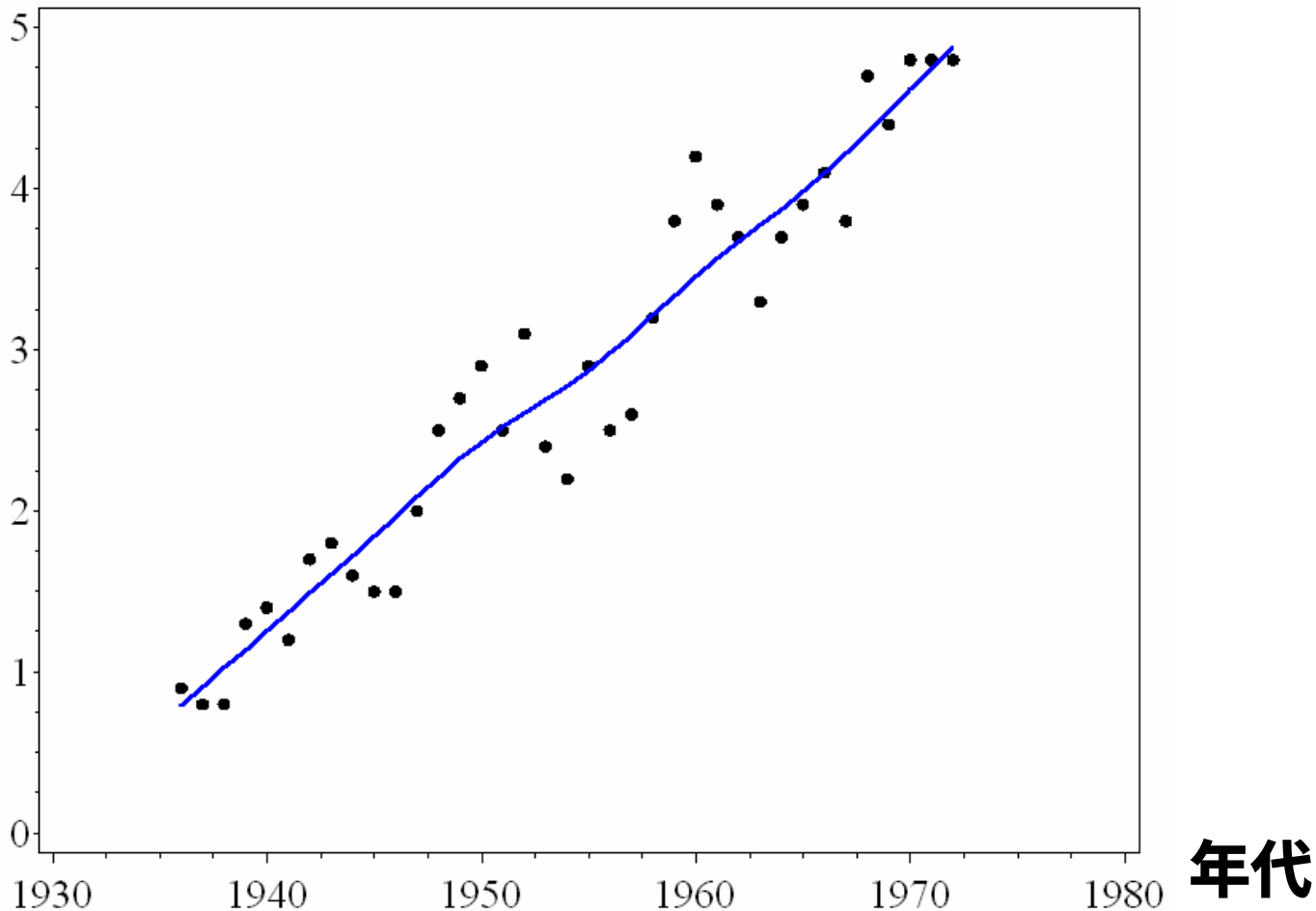
平滑化スプライン

- 以下のペナルティ付き残差平方和を最小とする推定量

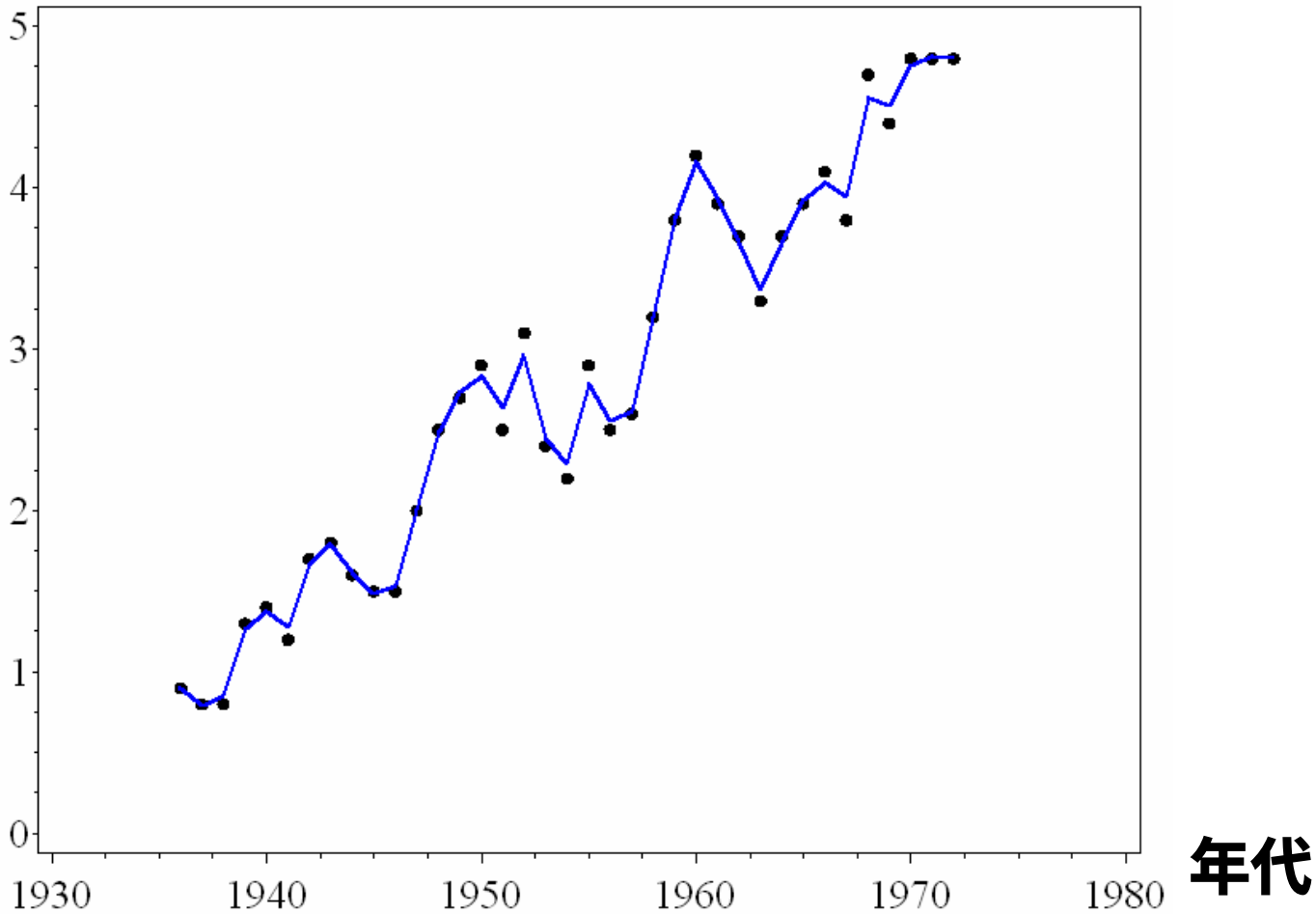
$$RSS = \sum_{i=1}^{n} [y_i - s(x_i)]^2 + \frac{\lambda}{2} \int [s''(x)]^2 \, dx$$

- xの実現値を節点とした三次の自然スプラインが解
- λは**平滑化パラメータ**

平滑化スプラインの出力 ($\lambda \to \infty$)



平滑化スプラインの出力 (λ　0)



平滑化スプライン

- 以下のペナルティ付き残差平方和を最小とする推定量

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

- 第一項はデータへの当てはまりをよくする基準
- 第二項は曲線の変動を小さくする基準
  - λは2つのバランスを規定する

自由度

- 平滑化パラメータは手法により異なる
  - LOESSではk
  - 平滑化スプラインではλ
- 自由度により統一的に表現
  - 実質的なパラメータの数に変換

自由度

- 回帰分析では, 最小二乗推定量は行列の形で表せる
  - Aはハット行列

$$\hat{y} = X(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}y$$
$$= Ay$$

- 自由度 (パラメータの数) は行列Aのランク
  - 単回帰における帰無仮説「$\beta_1=0$」のt検定
    - モデルの自由度は2 (切片と傾き)
    - 残差の自由度はn-2

自由度

- LOESS, 平滑化スプライン推定量もこの形に表せる
  - Aは平滑化パラメータと一対一に対応

$$\hat{s} = Ay$$

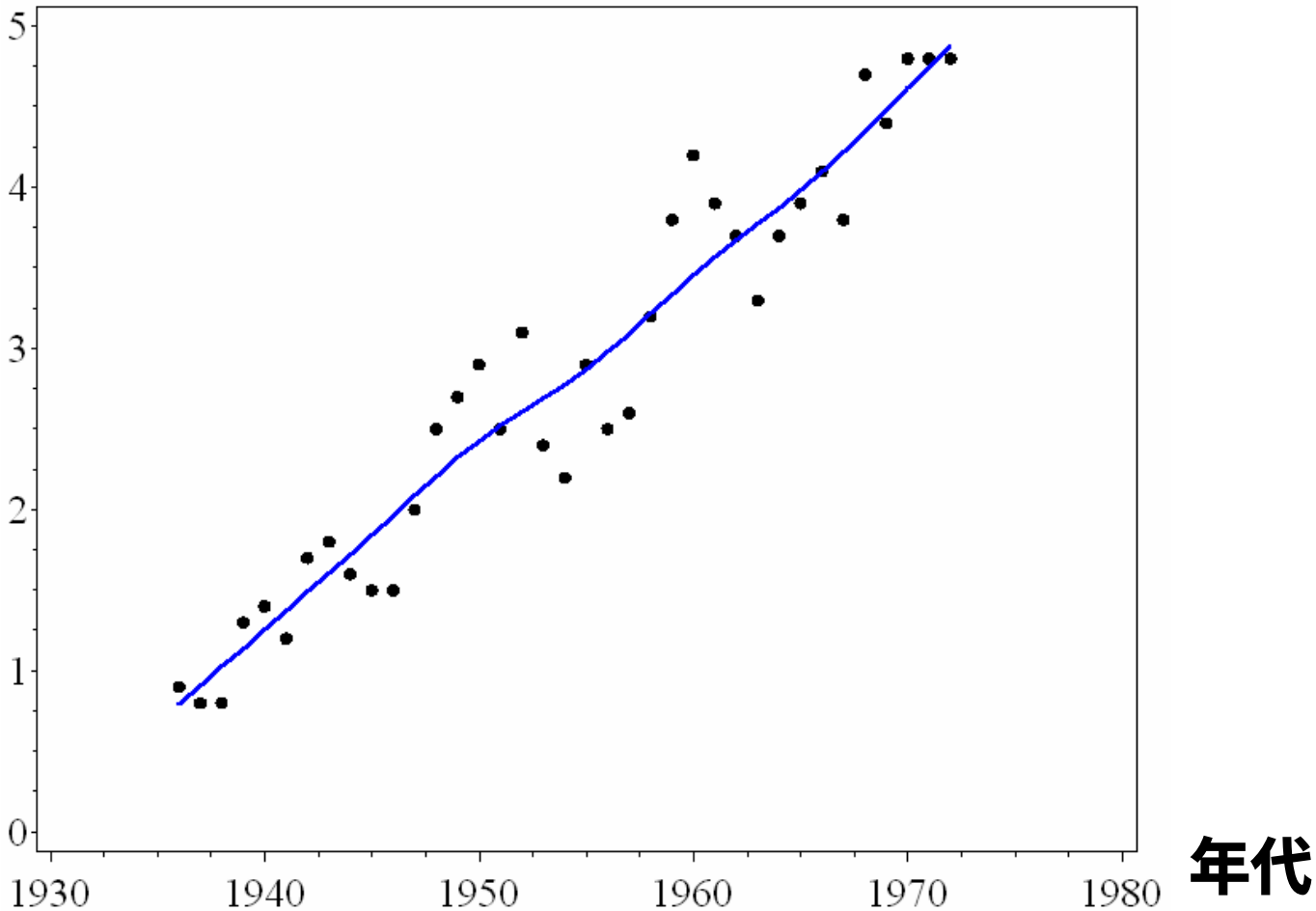
- 自由度
  - LOESSではA^tAのトレース (対角要素の和)
  - 平滑化スプラインではAのトレース
- こう決めておけば平滑化パラメータを統一的に扱える

自由度の設定

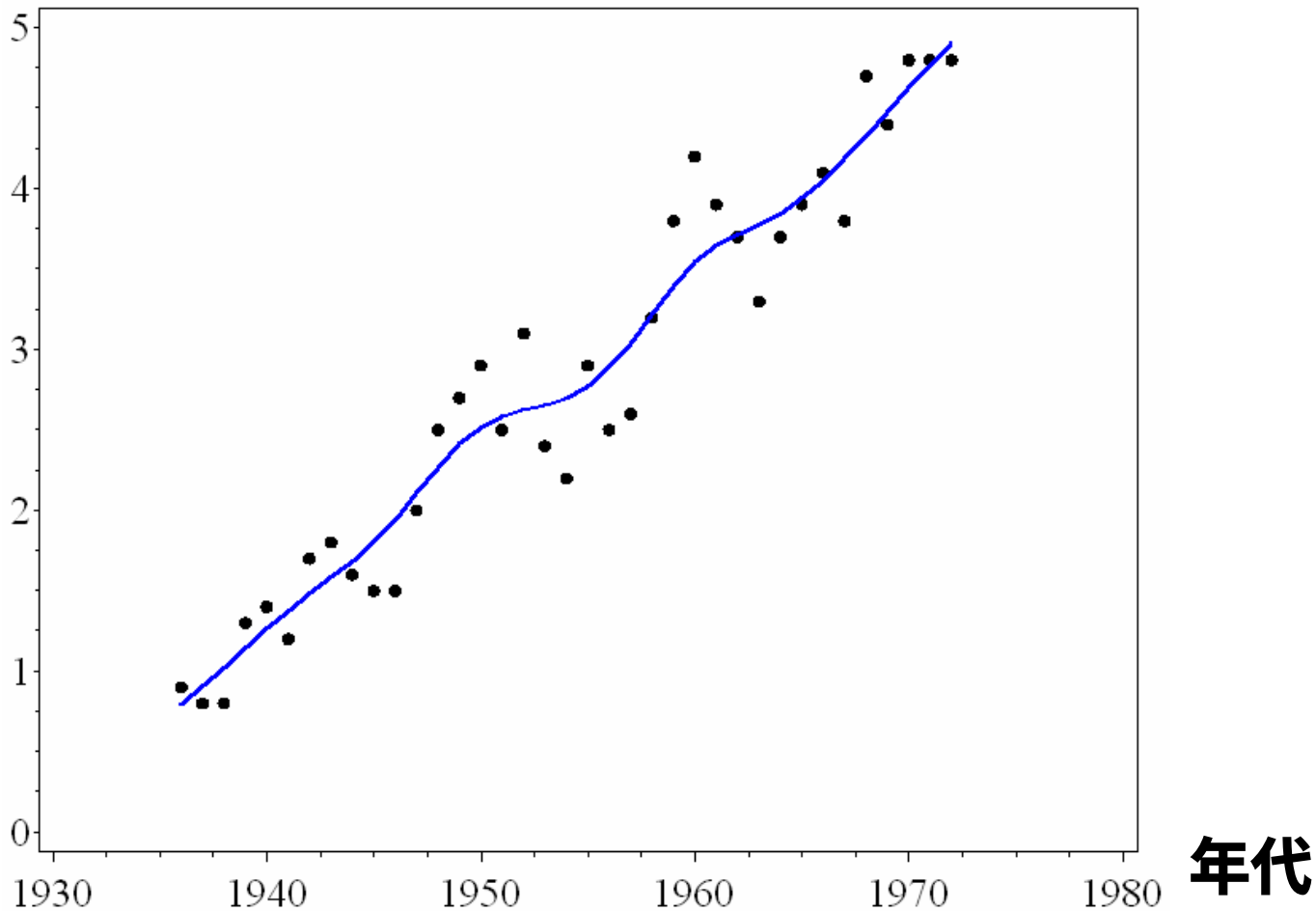
- 解析者が指定
  - 扱っているデータ解析の背景から
- データから推定
  - GCV (一般化クロスバリデーション) 法
  - 以下の基準を最小にする自由度を選ぶ

$$GCV = \frac{n\sum_{i=1}^{n}[y_i - \hat{s}(x_i)]^2}{[n - \mathrm{trace}(A)]^2}$$

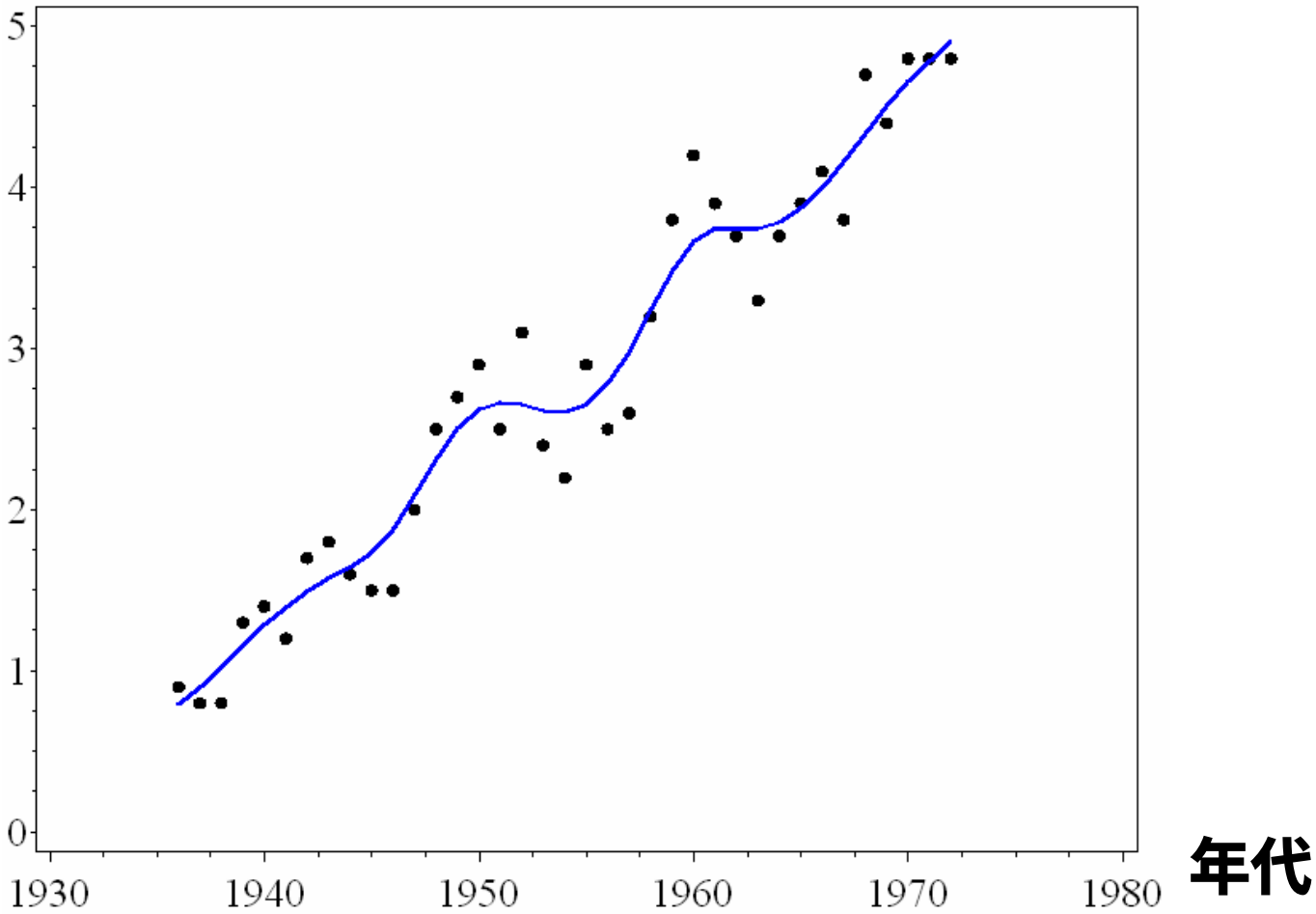
自由度5



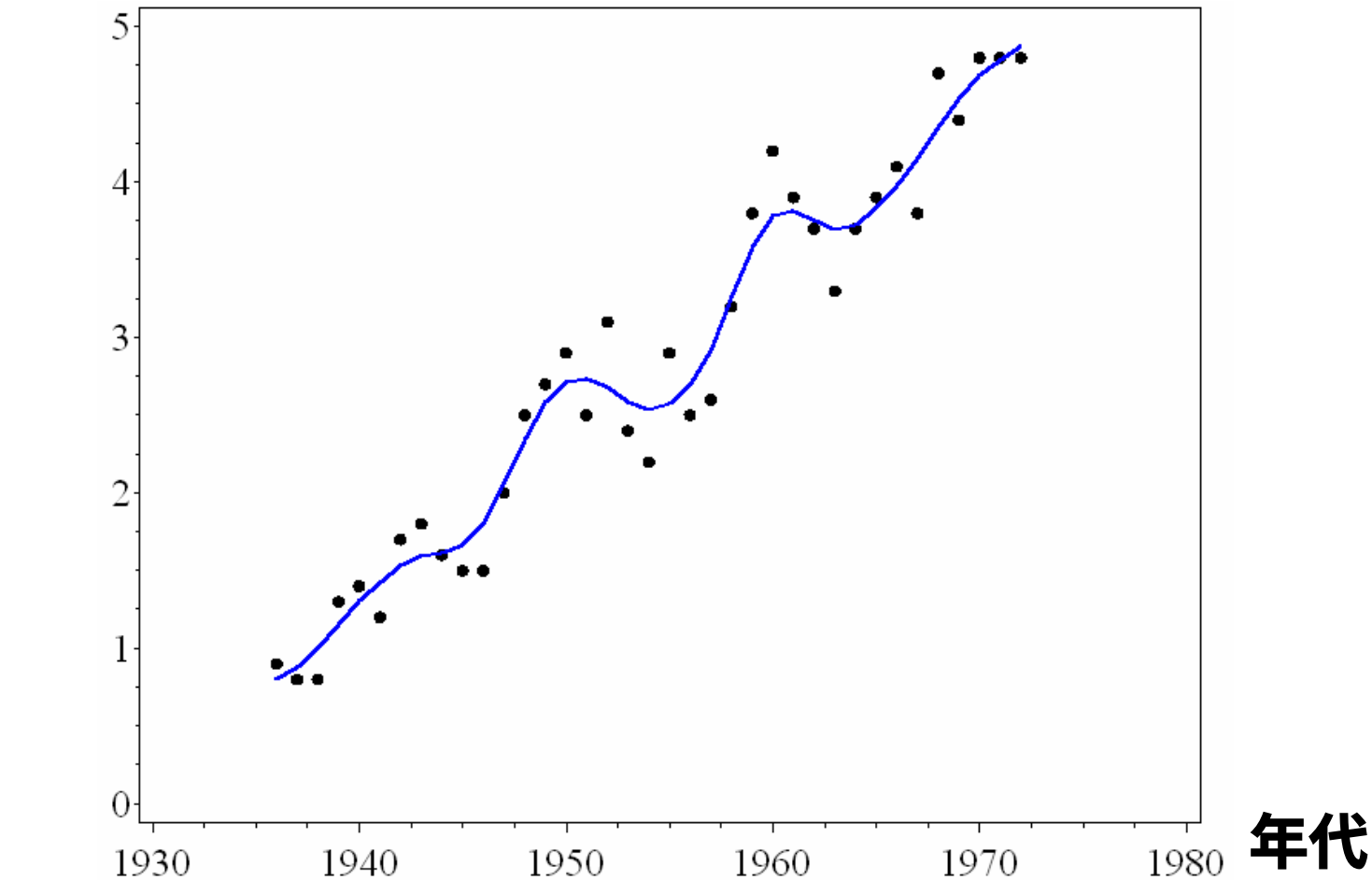
自由度7



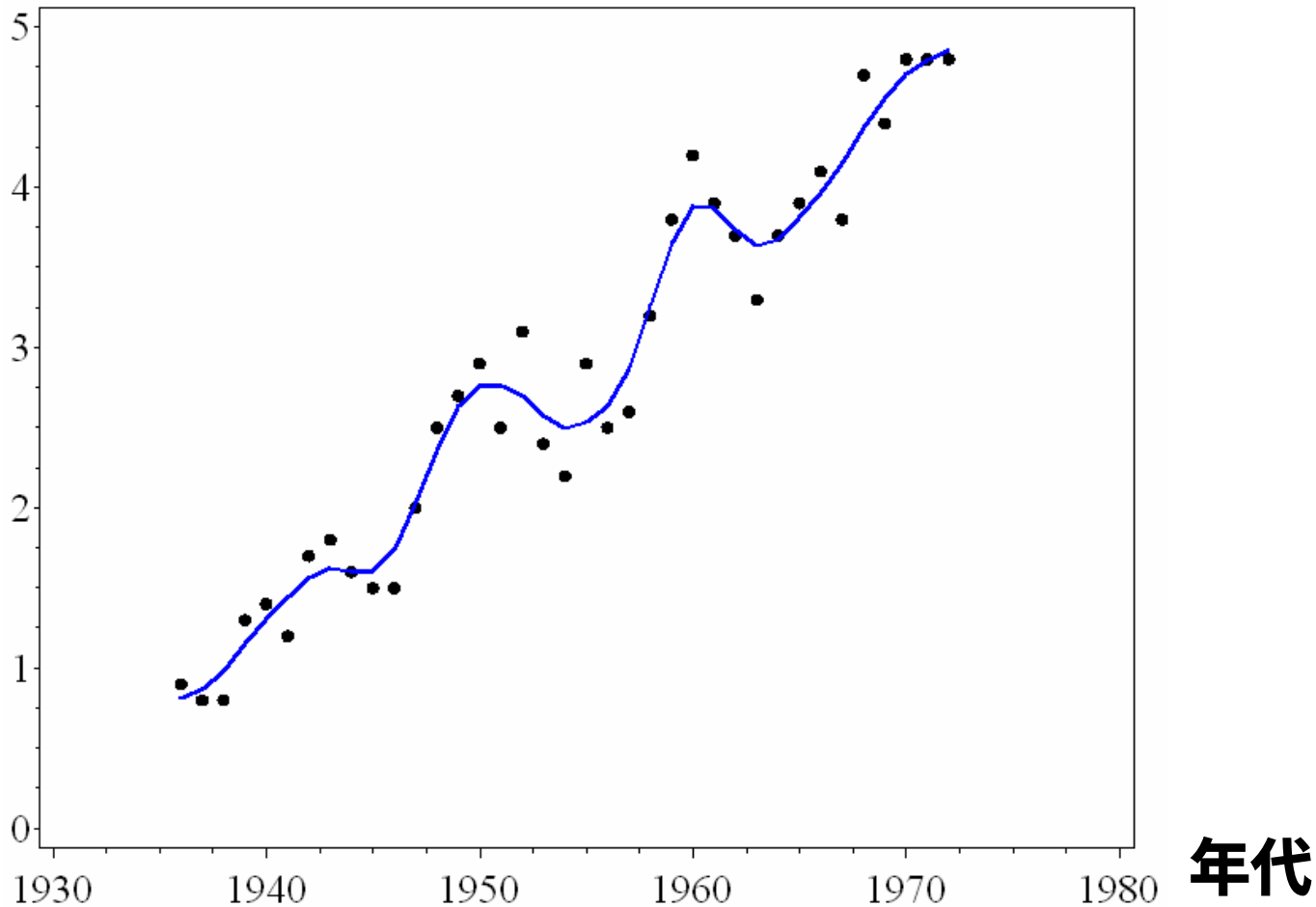
自由度9



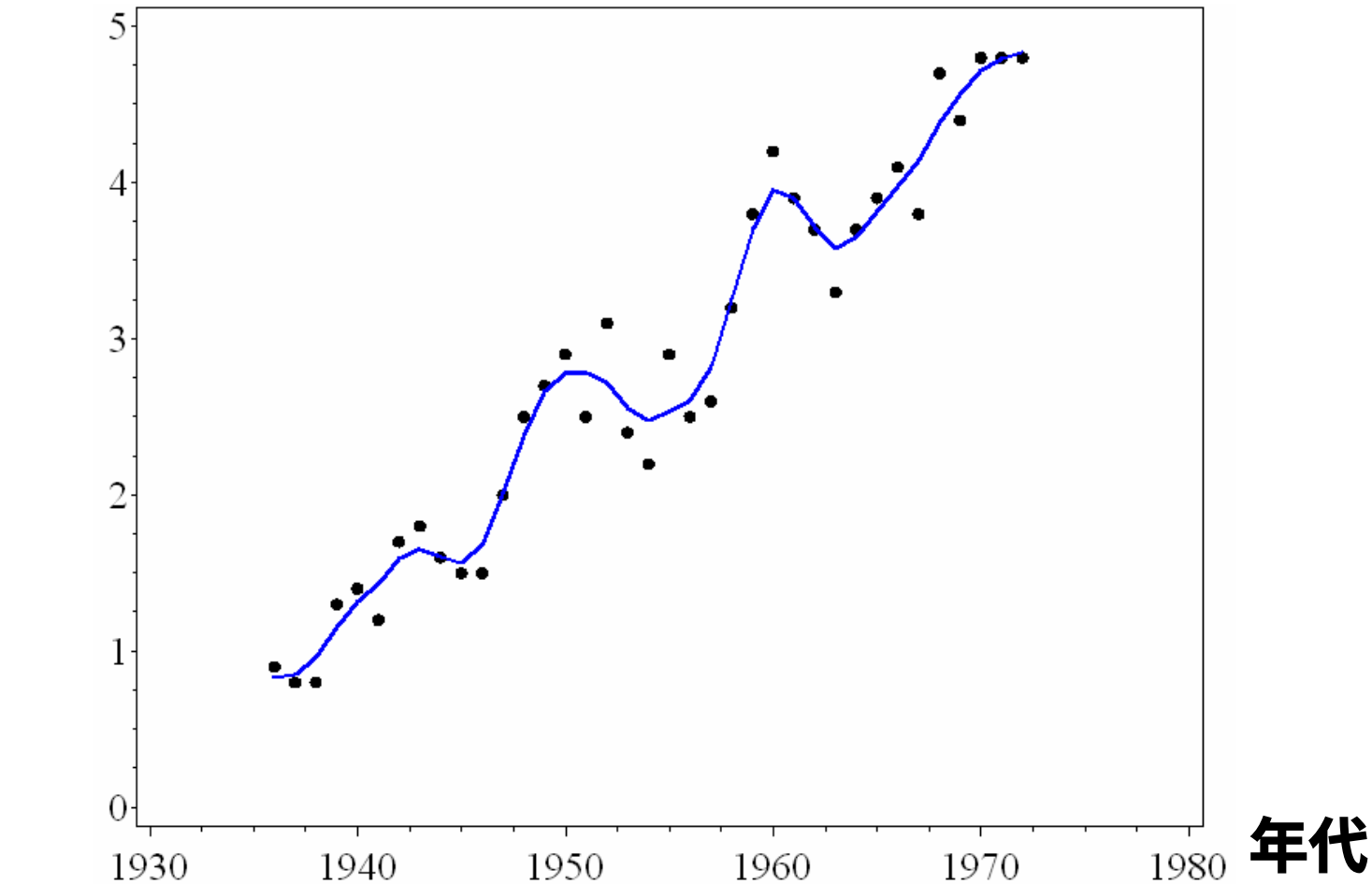
自由度11



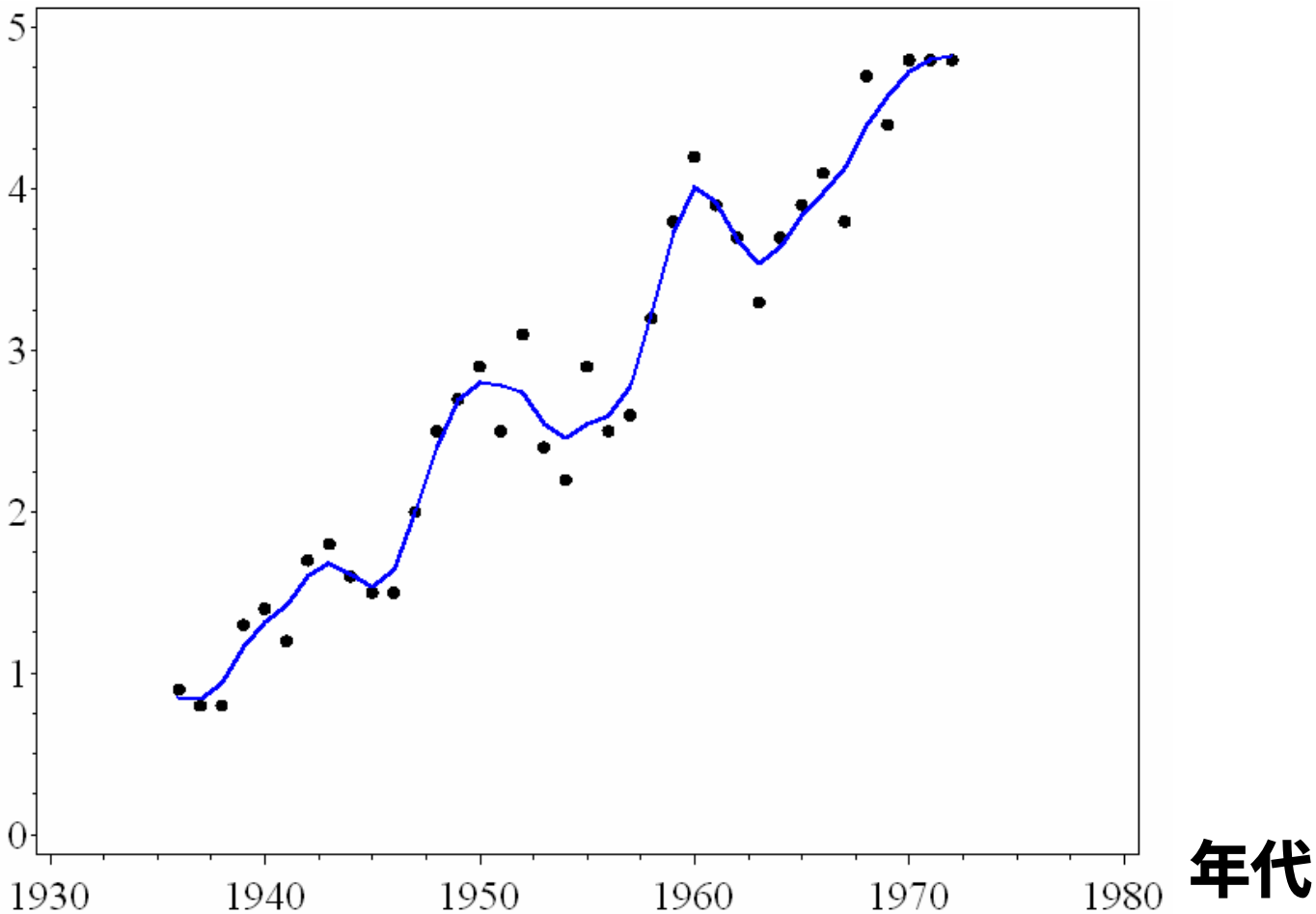
自由度13



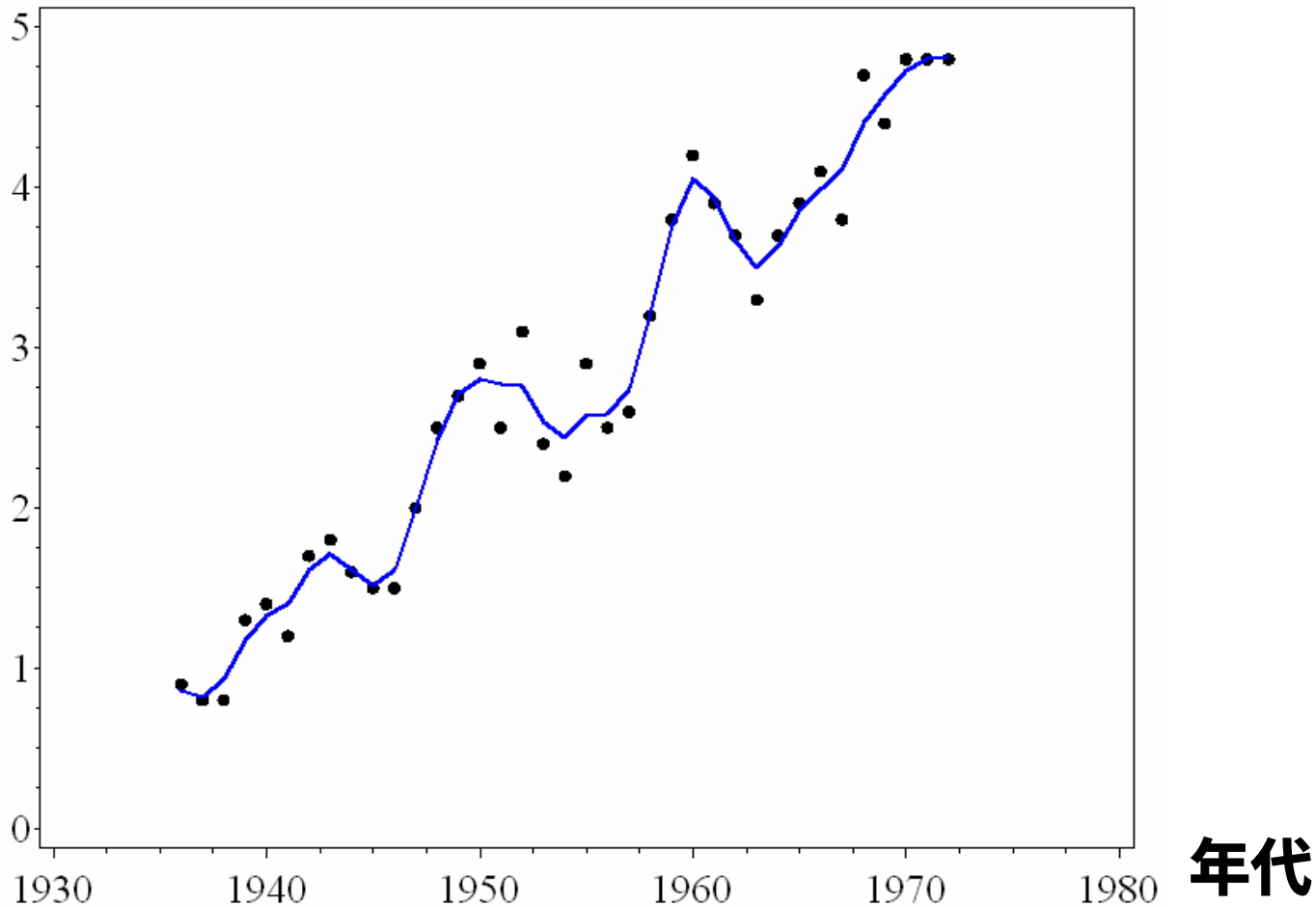
自由度15



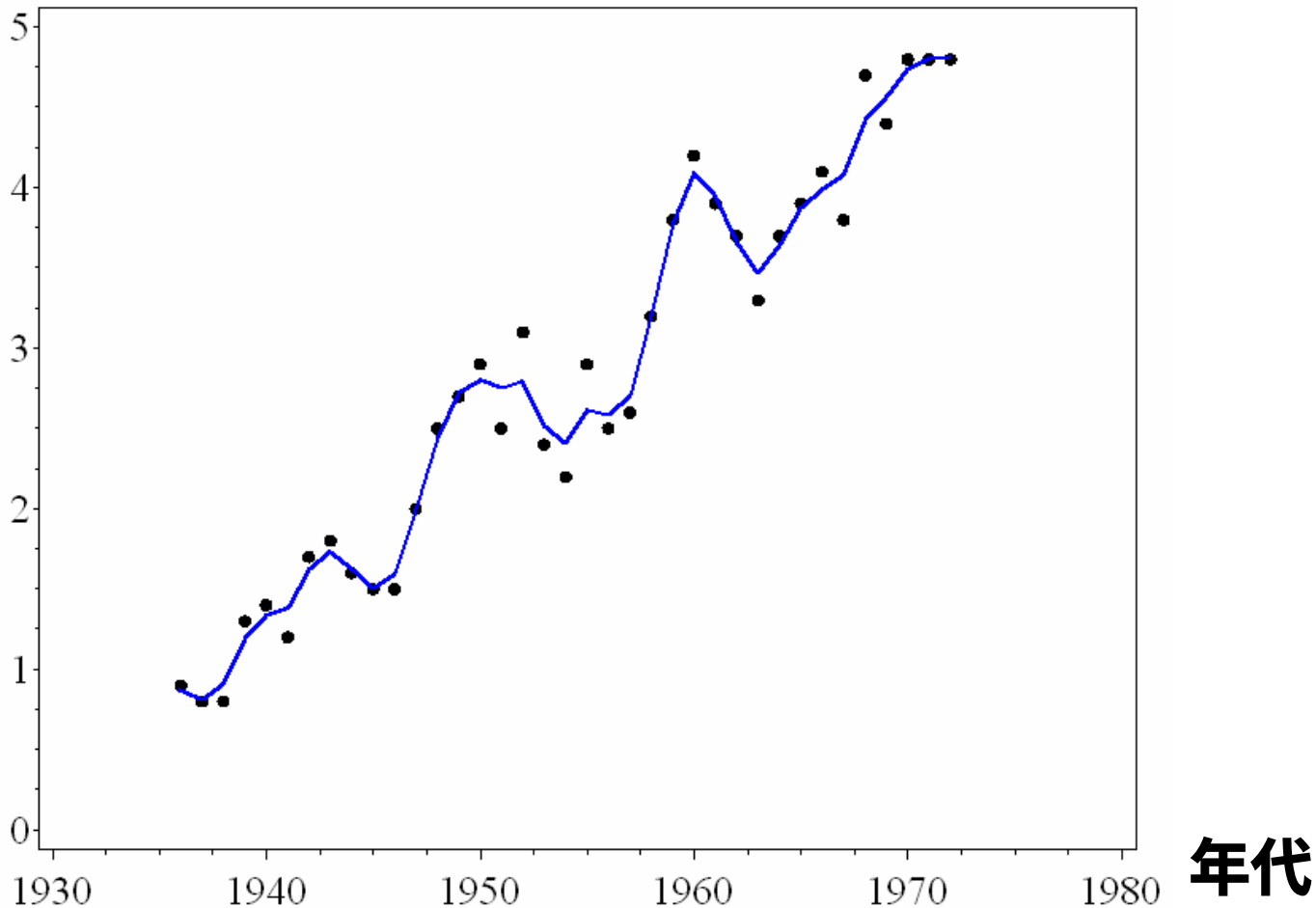
自由度17



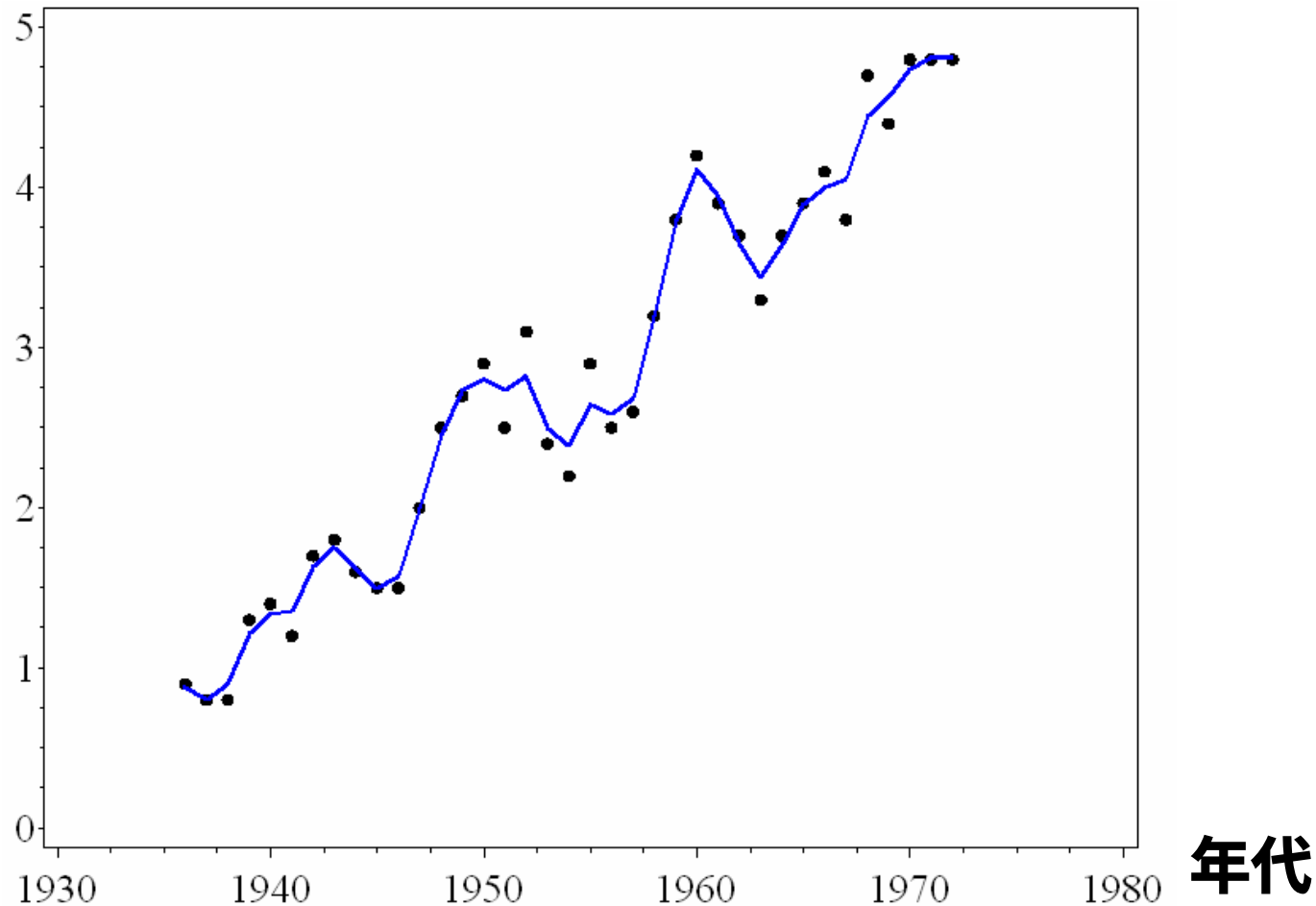
自由度19



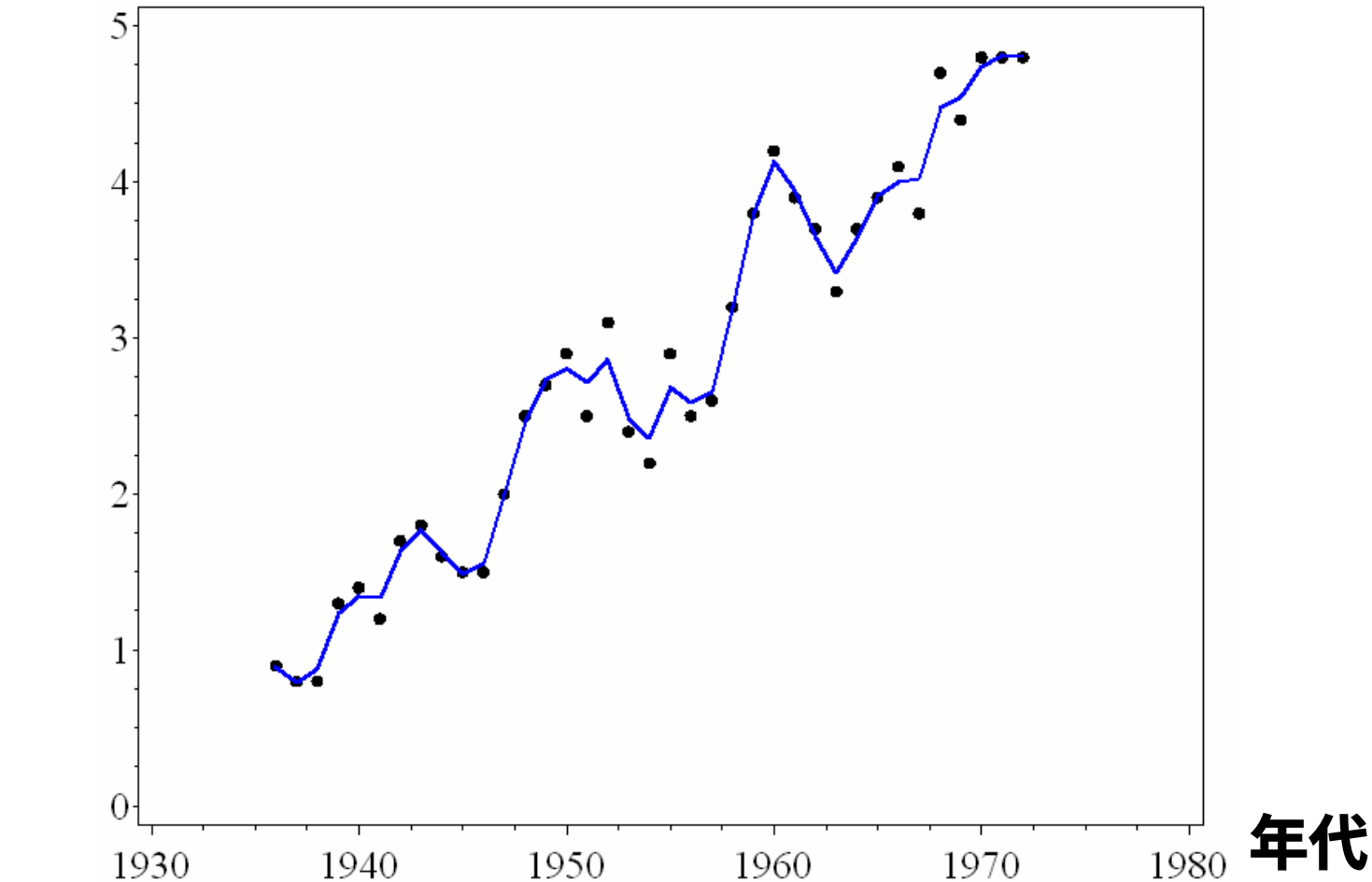
自由度21



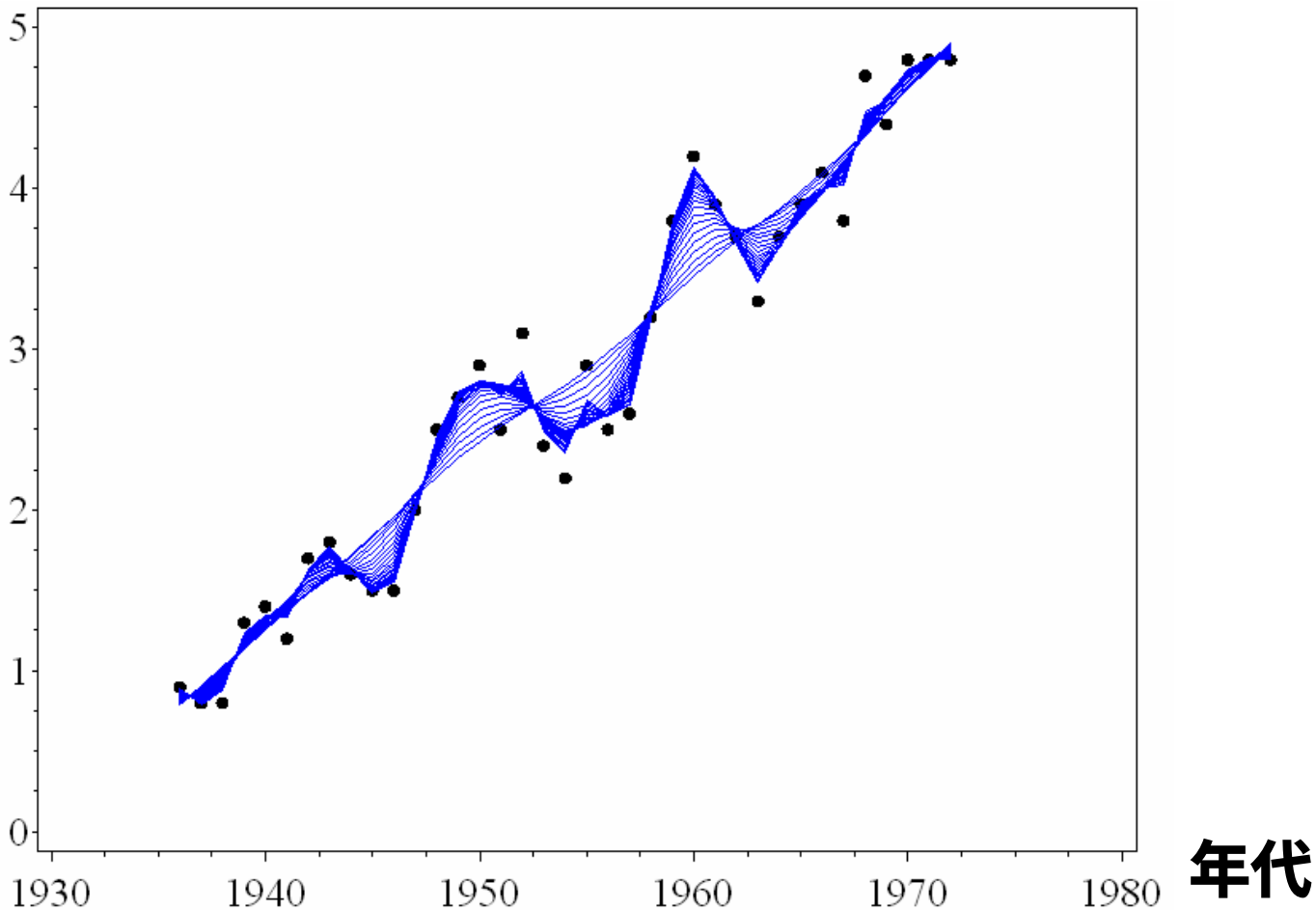
自由度23



自由度25



重ねてみると



説明変数が複数ある場合は?

- 1. 多次元の関数を考える
  - 実質的なパラメータ数が加速的に増加

$$y_i = s(x_{i1}, x_{i2}, \ldots, x_{ip}) + \varepsilon_i$$

- 2. **加法モデル**の制約を入れる
  - 1つ1つの説明変数の効果は足し算で効く
  - 交互作用はない

$$y_i = s_0 + \sum_{j=1}^{p} s_j(x_{ij}) + \varepsilon_i$$

薄板スプライン

- 二変量を扱える平滑化手法
- 以下のペナルティ付き残差平方和を最小とする推定量

$$RSS = \sum_{i=1}^{n} [y_i - s(x_{1i}, x_{2i})]^2 + \frac{\lambda}{2} \int \left[\left(\frac{\partial^2 s}{\partial {x_1}^2} \right)^2 + 2 \left(\frac{\partial^2 s}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 s}{\partial {x_2}^2} \right)^2 \right] \mathrm{d}x_1 \mathrm{d}x_2$$

- Proc GAMで指定可能

推定アルゴリズム

- 二つとも推定は最小二乗法
  - LOESSだと局所重み付き…
  - 平滑化スプラインだとペナルティ付き…
- 最小二乗解を求めるには?
  - 一度に全ての説明変数について推定するのは大変
  - ここで加法モデルが効いてくる

バックフィッティングアルゴリズム

- 特徴
  - 説明変数ごとに解を更新する
  - 順番に残差に対して当てはめを行う
- 例えば, k番目の関数s_kの更新では
  - s_k以外に関する残差を求める
  - R_{ik}を結果変数, x_{ik}のみを説明変数として解を求める

$$R_{ik} = y_i - \hat{s}_0 - \sum_{j=1}^{k-1} \hat{s}_j(x_{ij}) - \sum_{j=k+1}^{p} \hat{s}_j(x_{ij})$$

バックフィッティングアルゴリズム

- **ステップ1 (初期値を与える)**

  $\hat{s}_0^{(1)} = \bar{Y}, \hat{s}_1^{(1)} = \hat{s}_2^{(1)} = \ldots = \hat{s}_p^{(1)} = 0$

- **ステップ2 (反復計算)**

  R_1にx_1を当てはめて, $\hat{s}_1^{(1)}$を$\hat{s}_1^{(2)}$に更新

  R_2にx_2を当てはめて, $\hat{s}_2^{(1)}$を$\hat{s}_2^{(2)}$に更新

  ⋮

  全ての説明変数が終わったら次のループへ

- **ステップ3 (収束の判定)**

  残差平方和がこれ以上減少しなくなったらストップ

Proc GAMの文法

```
PROC GAM < option > ;
   CLASS variables ;
   MODEL response = covariates < /options > ;
   SCORE data=dataset out=dataset ;
   OUTPUT <out=dataset> <keyword> ;
   BY variables ;
   ID variables ;
   FREQ variable ;
```

Proc GAMの文法

```
PROC GAM < option > ;
   CLASS variables ;
   MODEL response = covariates < /options > ;
   SCORE（予測用のデータセットを指定）
   OUTPUT <OUT=dataset> <keyword> ;
   BY variables ;
   ID（ID用の変数を指定）
   FREQ variable ;
```

MODELステートメント

MODEL response = covariates < /options > ;

- 当てはめるモデルを指定する
  - 平滑化手法
  - 自由度
  - 結果変数の分布

MODELステートメント

```
MODEL response = SPLINE (covariate1, DF=3)
                 PARAM (covariate2, covariate3)
                / DIST=BINOMIAL;
```

- 平滑化手法
- 自由度
- 結果変数の分布

MODELステートメント

- 平滑化手法
  - PARAM, LOESS, SPLINE, SPLINE2 (薄板スプライン)
- 自由度
  - 直接指定, オプション (METHOD=GCV)
- 結果変数の分布
  - GAUSSIAN, BINOMIAL, BINARY, GAMMA, IGAUSSIAN, POISSON
  - リンク関数はカノニカルリンク

OUTPUTステートメント

OUTPUT <OUT=dataset> <keyword> ;

- 予測値･信頼区間･回帰診断統計量などを出力
- キーワード: 変数名
  - PRED　: 予測値
  - UCLM　: 上側信頼区間, LCLMだと下側
  - ADIAG : テコ比統計量
  - RESID　: 残差統計量
  - STD　　: 標準偏差
  - ALL　　: 上記の変数全てを出力

OUTPUTステートメント

OUTPUT OUT=dataset ALL;

- 予測値･信頼区間･回帰診断統計量などを出力
- キーワード: 変数名
  - PRED　: P_response, P_covariate1
  - UCLM　: UCLM_covariate1
  - ADIAG : ADIAG_covariate1
  - RESID　: R_covariate1
  - STD　　: STD_covariate1（スプライン, LOESSではSTDP_）
  - ALL　　: 上記の変数全てを出力

平滑化する説明変数の扱いについて

- 一次の項と二次以上に分けて出力
- 例えば, 年齢の効果は
  - 回帰係数×Age + P_Age
  - 一次の効果　　: 回帰係数×Age
  - 二次以上の効果: P_Age
- 上側信頼区間
  - 回帰係数の上側信頼区間×Age + UCLM_Age
- GCV基準により選択された自由度
  - 1を引いて出力

Kyphosis (脊柱後弯症) データ (Bell D, et al)

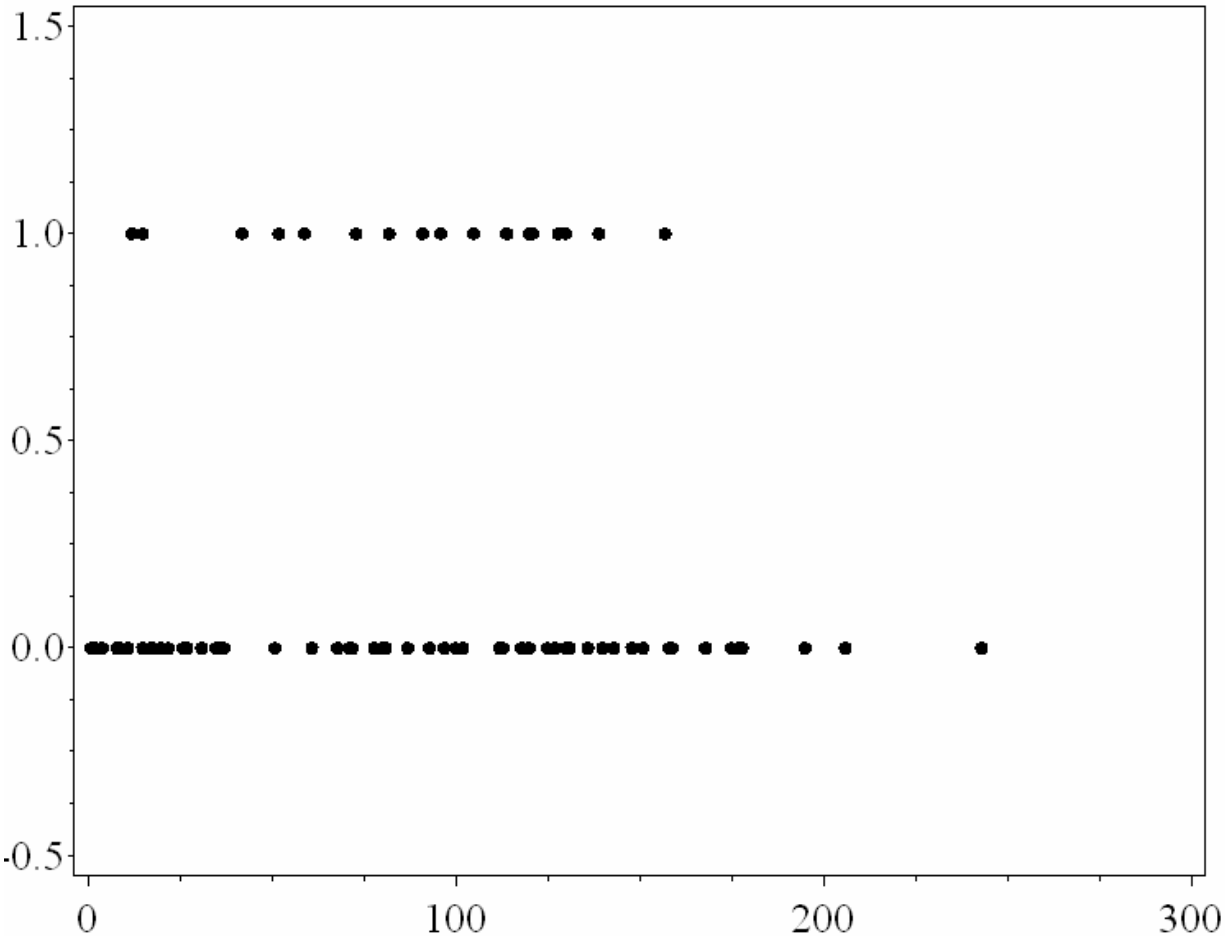
- **目的**: Kyphosisの予後因子の探索
- **対象**：椎弓切除術を受けた子供83人
- **エンドポイント**: Kyphosisの有無
- **予後因子**
  - Age: **年齢**
  - **その他の変数**
    - StartVert: (**手術を脊椎のどの部位から行ったか**)
    - NumVert: (**手術を行った部位の数**)

Kyphosisデータの入力

```
data Kyphosis;
    input Age StartVert NumVert Kyphosis @@;
    datalines;
    71  5  3  0     158 14 3 0    128 5   4 1
    2   1  5  0     1   15 4 0    1   16  2 0
    61  17 2  0     37  16 3 0    113 16  2 0
    59  12 6  1     82  14 5 1    148 16  3 0
    18  2  5  0     1   12 4 0    243 8   8 0
   ...
;
```

2値データだと…

Kyphosisの有無



SASのプログラム

- 結果変数
  - Kyphosis
- 説明変数
  - 平滑化（自由度3）: Age
  - 線型　　:StartVert, NumVert
- 入出力データセット
  - 解析用データセット: Kyphosis
  - 出力データセット: Kyphosis_out2
  - 予測用入力データセット: Kyphosis_in
  - 予測用出力データセット: Kyphosis_out1

SASのプログラム

```
PROC GAM DATA=Kyphosis;
MODEL Kyphosis = PARAM(StartVert NumVert)
                 SPLINE(Age, DF=3)
                 / DIST=BINOMIAL;
SCORE DATA=Kyphosis_in out=Kyphosis_out1;
OUTPUT OUT=Kyphosis_out2 ALL;
RUN;
```

SASの出力

大切なのはこの3つ

The GAM Procedure
Dependent Variable: Kyphosis
Regression Model Component(s): StartVert NumVert
Smoothing Model Component(s): spline(Age)

Summary of Input Data Set	
Number of Observations	83
Number of Missing Observations	0
Distribution	Binomial
Link Function	Logit

Iteration Summary and Fit Statistics	
Number of local score iterations	9
Local score convergence criterion	7.95016E-10
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	1.5075369E-9
The Deviance of the Final Estimate	54.394484949

The local score algorithm converged.

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-2.19516	1.54821	-1.42	0.1603
StartVert	-0.20190	0.06936	-2.91	0.0047
NumVert	0.45591	0.21021	2.17	0.0332
Linear(Age)	0.01156	0.00825	1.40	0.1650

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(Age)	0.999965	2.000000	55.171372	66

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(Age)	2.00000	10.478479	10.4785	0.0053

Parameter Estimates

- パラメトリックで指定した変数の回帰係数の出力
- 平滑化した変数 (Linear(Age)) は一次の部分だけ

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	−2.19516	1.54821	−1.42	0.1603
StartVert	−0.20190	0.06936	−2.91	0.0047
NumVert	0.45591	0.21021	2.17	0.0332
Linear(Age)	0.01156	0.00825	1.40	0.1650

Fit Summary of Smoothing Components

- 平滑化した変数の自由度やGCV基準の当てはまり
- 自由度は2
  - 一次の項は除いて考える

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(Age)	0.999965	2.000000	55.171372	66

Analysis of Deviance
(デビアンスを用いた分散分析)

- 平滑化した変数の検定
- Ageは有意

Smoothing Model Analysis Analysis of Deviance				
Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(Age)	2.00000	10.478479	10.4785	0.0053

OUTPUTステートメントの出力

	Kyphosis	Age	StartVert	NumVert	P_Kyphosis	P_Age	Adiag_Age	Std_Age	UCLM_Age	LCLM_Age	R_Kyphosis
1	0	71	5	3	-0.53	0.483	0.29	0.539	1.539	-0.57	-0.77
2	0	158	14	3	-2.86	-1.03	0.714	0.845	0.623	-2.69	-0.24
3	1	128	5	4	0.234	0.136	0.249	0.499	1.114	-0.84	0.889
4	0	2	1	5	-1.68	-1.58	1.45	1.204	0.778	-3.94	-0.43
5	0	1	15	4	-5.01	-1.62	1.533	1.238	0.807	-4.05	-0.08
6	0	1	16	2	-6.12	-1.62	1.533	1.238	0.807	-4.05	-0.05
7	0	61	17	2	-3.71	0.301	0.315	0.561	1.4	-0.8	-0.16
8	0	37	16	3	-3.95	-0.32	0.412	0.642	0.942	-1.57	-0.14
9	0	113	16	2	-2.65	0.554	0.246	0.496	1.526	-0.42	-0.27
10	1	59	12	6	-0.94	0.261	0.321	0.567	1.371	-0.85	1.6
11	1	82	14	5	-1.14	0.653	0.278	0.527	1.686	-0.38	1.77
12	0	148	16	3	-2.96	-0.61	0.469	0.685	0.73	-1.95	-0.23
13	0	18	2	5	-1.09	-0.98	0.65	0.806	0.603	-2.56	-0.58
14	0	1	12	4	-4.4	-1.62	1.533	1.238	0.807	-4.05	-0.11
15	0	243	8	8	-3.1	-5.75	12.26	3.502	1.115	-12.6	-0.21
16	0	168	18	3	-4.02	-1.5	1.076	1.037	0.531	-3.54	-0.13
17	0	1	16	3	-5.67	-1.62	1.533	1.238	0.807	-4.05	-0.06
18	0	78	15	6	-0.99	0.597	0.281	0.53	1.635	-0.44	-0.61
19	0	175	13	5	-2.37	-1.86	1.415	1.19	0.475	-4.19	-0.31
20	0	80	16	5	-1.6	0.626	0.279	0.528	1.661	-0.41	-0.45
21	0	27	9	4	-2.53	-0.65	0.491	0.701	0.72	-2.03	-0.28
22	0	22	16	2	-5.09	-0.83	0.563	0.751	0.639	-2.3	-0.08
23	1	105	5	6	1.43	0.686	0.26	0.51	1.686	-0.31	0.489
24	1	96	12	3	-1.39	0.747	0.271	0.52	1.767	-0.27	2.008
25	0	131	3	2	-0.34	0.032	0.262	0.512	1.036	-0.97	-0.84
26	1	15	2	7	-0.32	-1.09	0.737	0.859	0.595	-2.77	1.174
27	0	9	13	5	-3.75	-1.31	0.989	0.995	0.636	-3.26	-0.15
28	1	12	2	14	2.723	-1.2	0.849	0.921	0.606	-3.01	0.256
29	0	8	6	3	-3.3	-1.35	1.043	1.021	0.65	-3.35	-0.19
30	0	100	14	3	-1.77	0.732	0.267	0.517	1.744	-0.28	-0.41

SCOREステートメントの出力

	Age	StartVert	NumVert	Predicted Value of Kyphosis	Predicted Value of Age
1	71	3	5	0.782	0.483
2	158	3	5	0.272	-1.03
3	128	3	5	1.094	0.136
4	2	3	5	-2.08	-1.58
5	1	3	5	-2.13	-1.62
6	1	3	5	-2.13	-1.62
7	61	3	5	0.484	0.301
8	37	3	5	-0.41	-0.32
9	113	3	5	1.339	0.554
10	59	3	5	0.421	0.261
11	82	3	5	1.079	0.653
12	148	3	5	0.577	-0.61
13	18	3	5	-1.29	-0.98
14	1	3	5	-2.13	-1.62
15	243	3	5	-3.46	-5.75
16	168	3	5	-0.08	-1.5
17	1	3	5	-2.13	-1.62
18	78	3	5	0.977	0.597
19	175	3	5	-0.36	-1.86
20	80	3	5	1.029	0.626
21	27	3	5	-0.86	-0.65
22	22	3	5	-1.1	-0.83
23	105	3	5	1.378	0.686
24	96	3	5	1.335	0.747
25	131	3	5	1.025	0.032
26	15	3	5	-1.44	-1.09
27	9	3	5	-1.73	-1.31
28	12	3	5	-1.58	-1.2
29	8	3	5	-1.78	-1.35
30	100	3	5	1.366	0.732

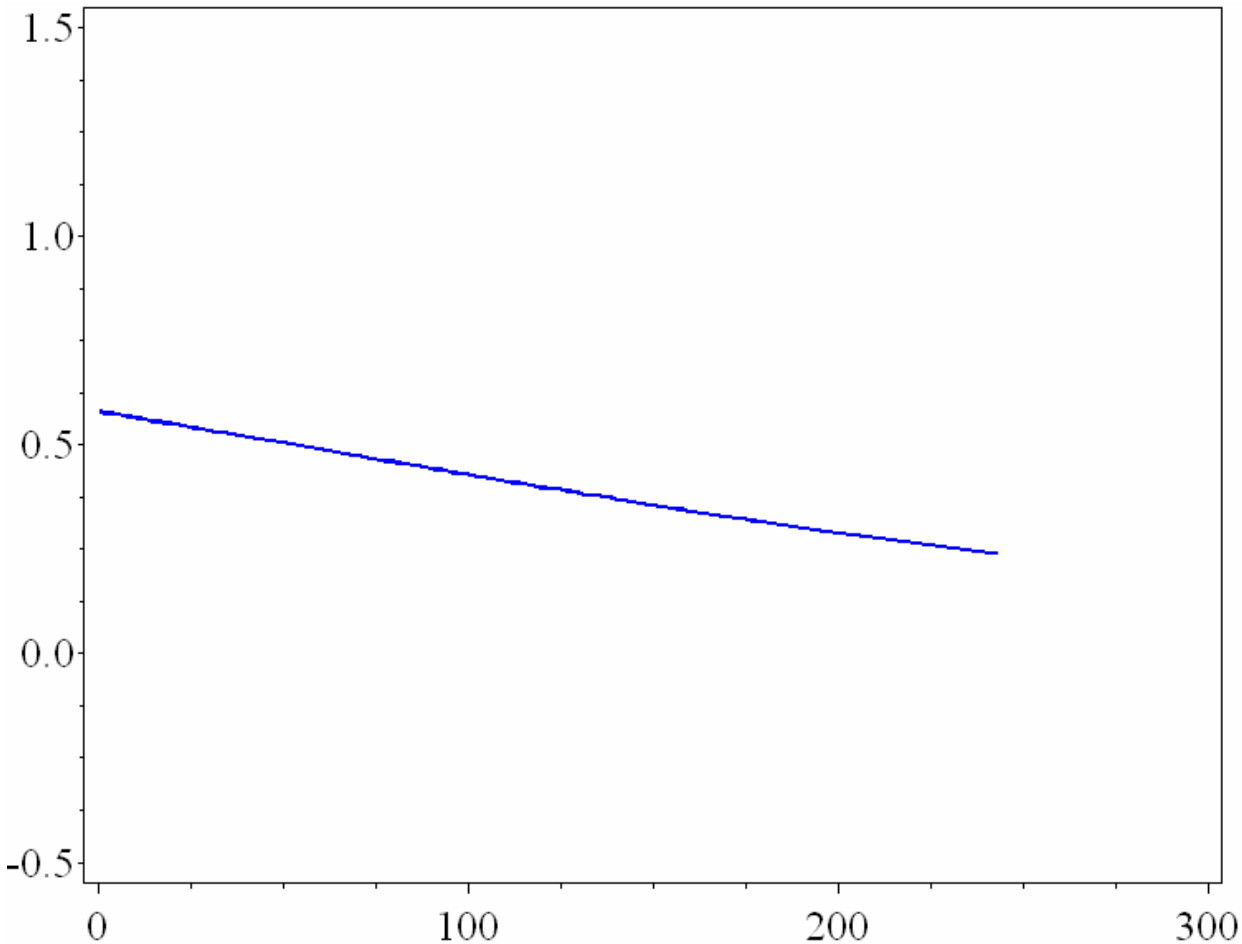
Proc GENMODで解析した結果

- StartVert, NumVertの結果は同様の結果
- Ageは有意ではない

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.2497	1.2424	-1.1853	3.6848	1.01	0.3145
Age	1	-0.0061	0.0055	-0.0170	0.0048	1.21	0.2713
StartVert	1	0.1972	0.0657	0.0684	0.3260	9.01	0.0027
NumVert	1	-0.3031	0.1790	-0.6540	0.0477	2.87	0.0904
Scale	0	1.0000	0.0000	1.0000	1.0000		

GENMODの結果をグラフにすると

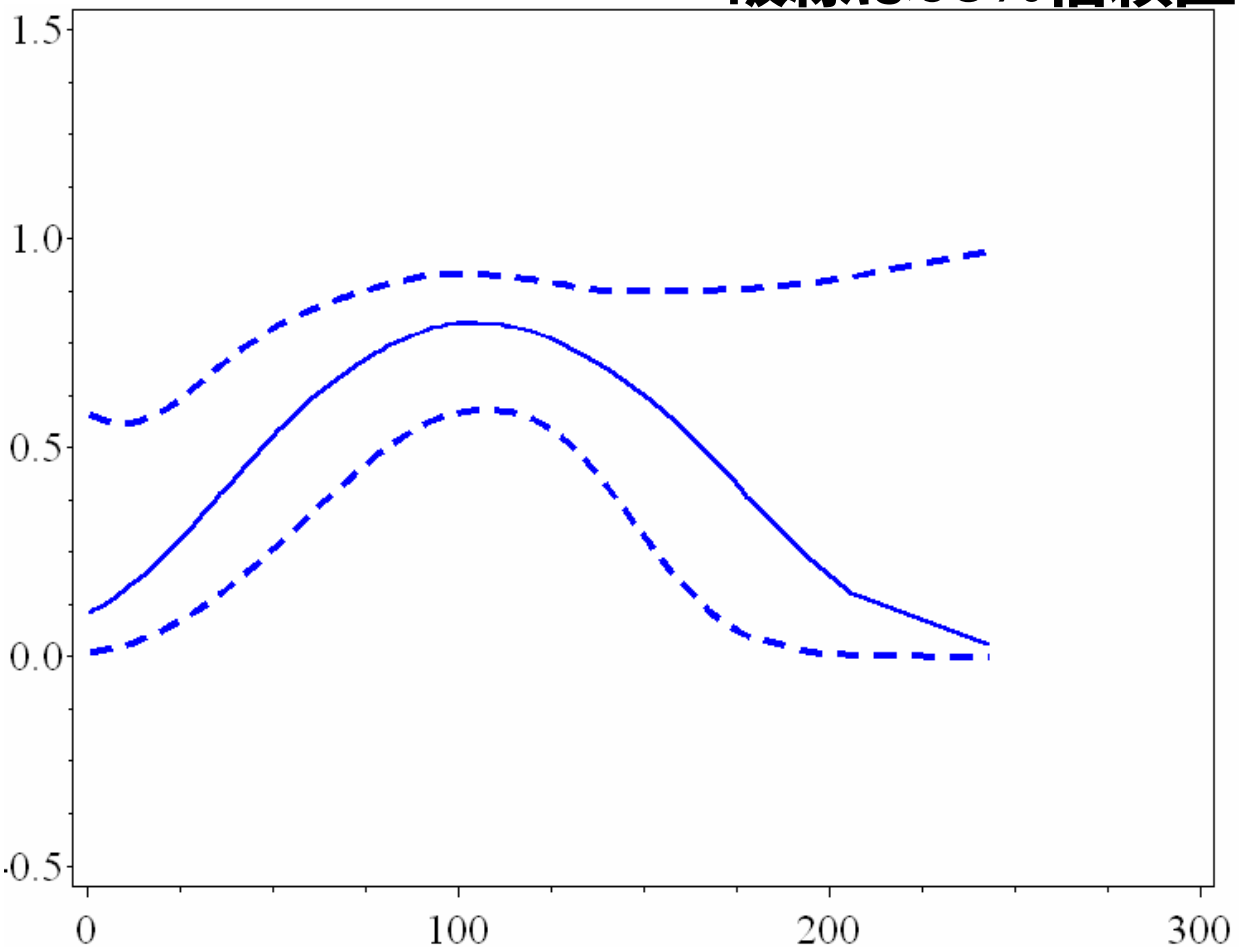
Kyphosisの有無



GAMの結果をグラフにすると

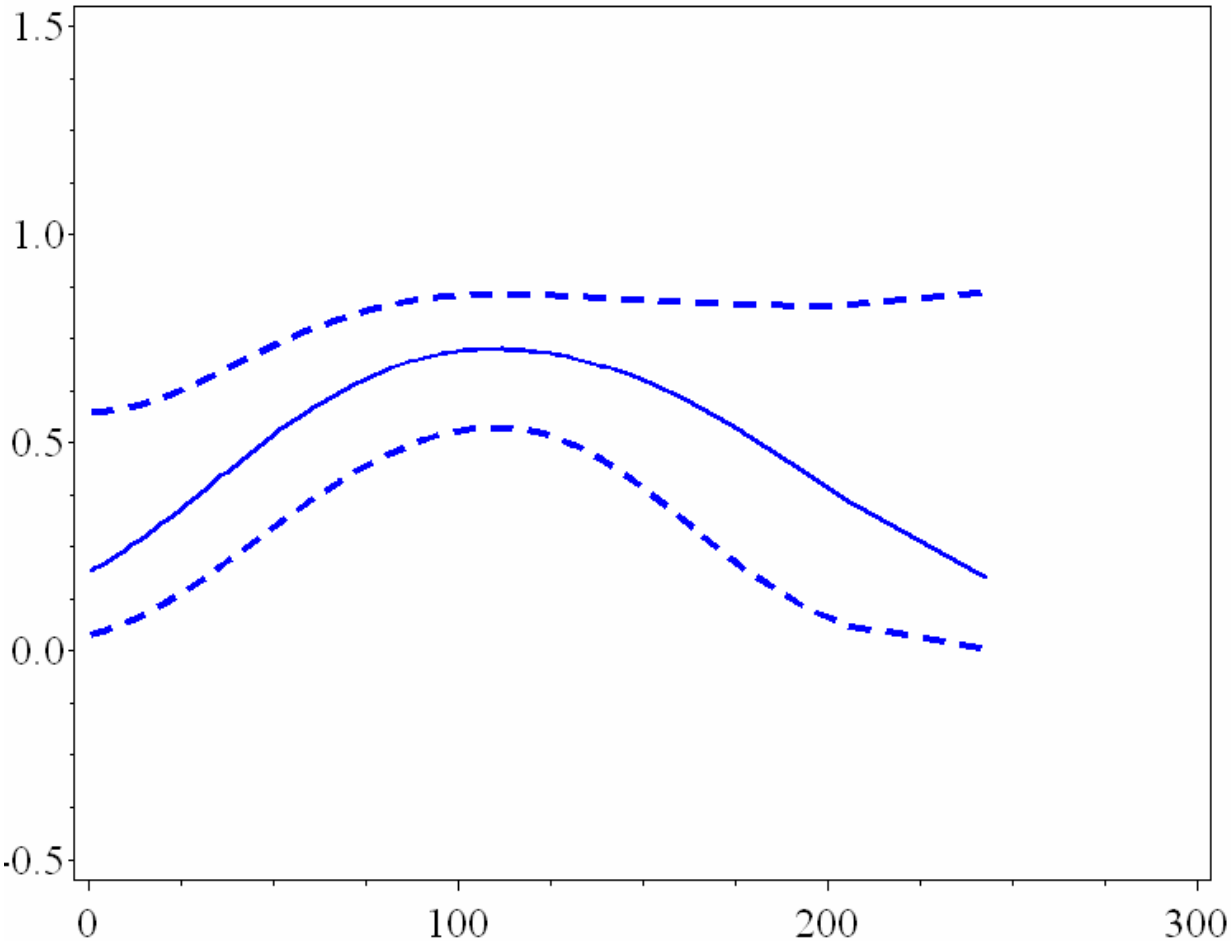
Kyphosisの有無

*破線は95%信頼区間



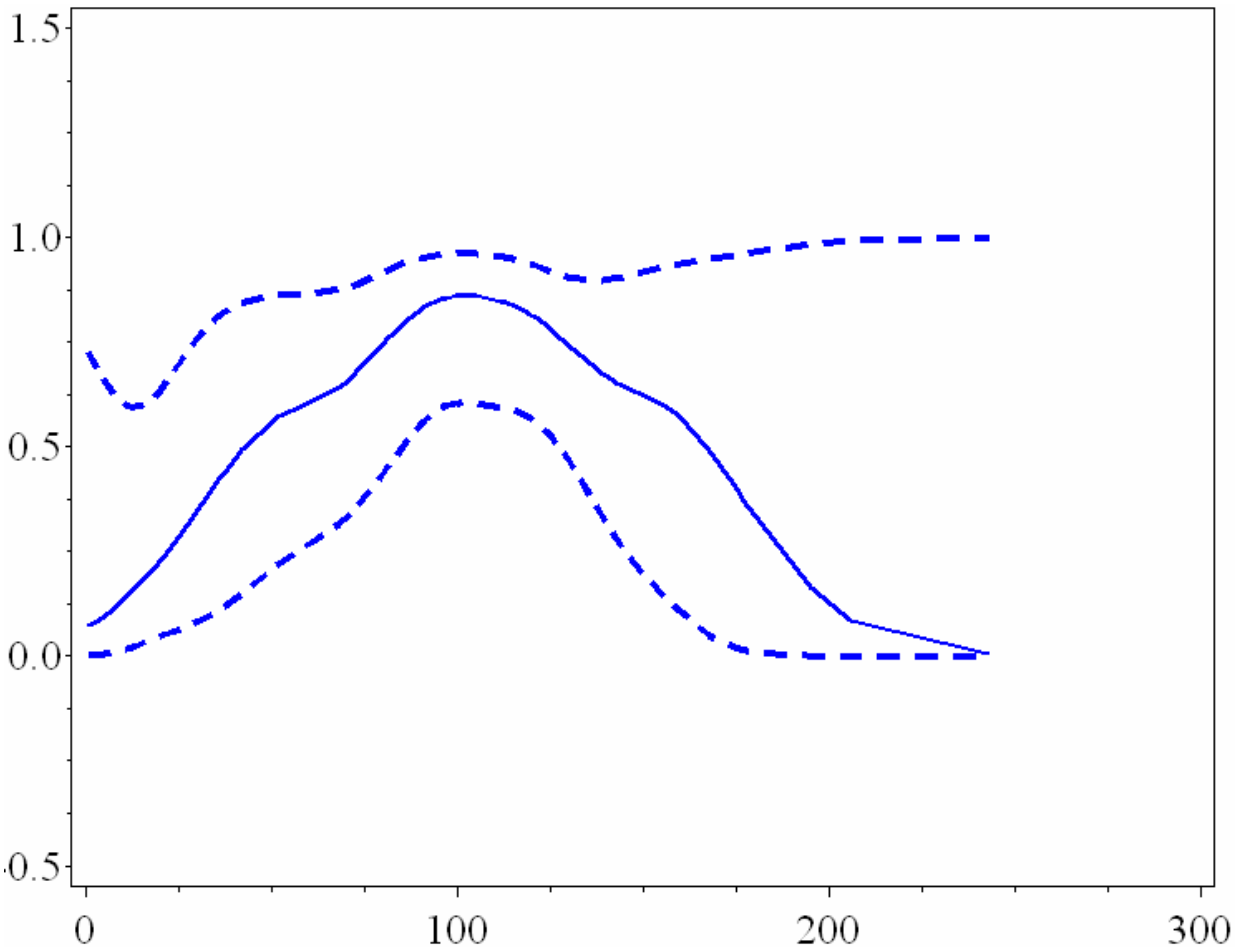
自由度2

Kyphosisの有無



自由度5

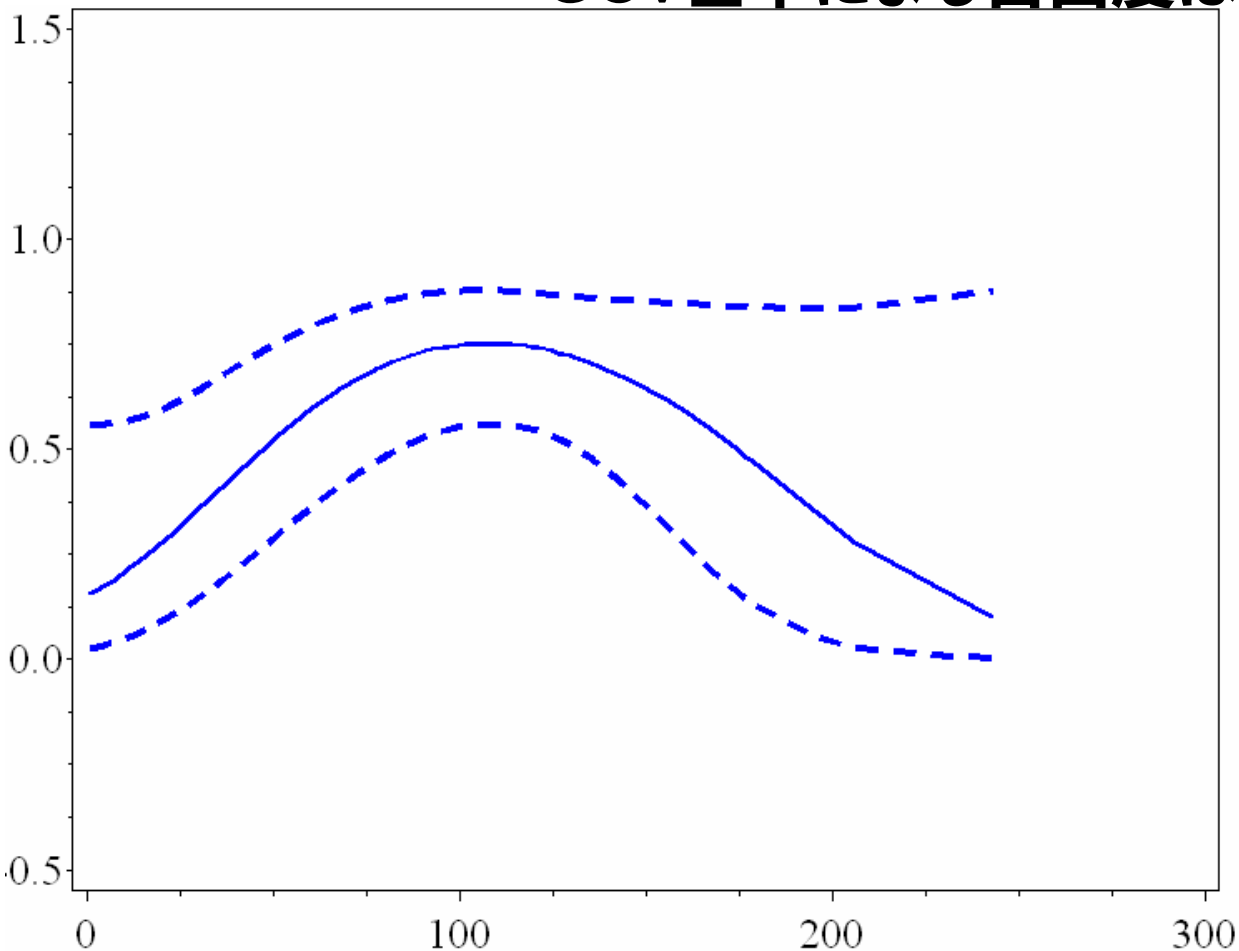
Kyphosisの有無



GCV基準を用いてみると

Kyphosisの有無

*GCV基準による自由度は2.27



まとめ: 発表内容

- ノンパラメトリック回帰
  - 散布図の平滑化
- GAMの紹介
  - 平滑化手法 (LOESS, 平滑化スプラインなど)
  - 推定アルゴリズム
  - 自由度の設定
- Proc GAMの文法・出力
- Kyphosisデータ解析の紹介

まとめ: 注意点

- 信頼区間・自由度などをグラフに示す
- 自由度の設定
  - 平滑化の程度によって印象がかなり異なる
  - いくつかの値を試してみる
- 平滑化した変数の出力
  - 一次の項と二次以上に分けて出力

参考文献

- **マニュアル･教科書**
  - SAS Institute. SAS/STAT® User's Guide Version9.1, SAS Institute, Cary, NC, U.S.A.; 2004.
  - Hastie T, Tibshirani R. Generalized Additive Models, Chapman and Hall: London; 1990.
  - Ruppert D, Wand MP, Carroll RJ. Semiparametric regression, Cambridge University Press: Cambridge; 2003.
- **原著**
  - Houghton AN, Flannery J, Viola MV. Malignant Melanoma in Connecticut and Denmark. *International Journal of Cancer* 1980; 25: 95-104.
  - Bell D, et al. Spinal Deformation Following Multi-Level Thoracic and Lumbar Laminectomy in Children. 1989; Submitted for publication.
  - Anscombe. Graphs in Statistical Analysis. *American Statistician* 1973; 27: 17-21