日本統計学会誌
第 39 巻, 第 2 号, 2010 年 3 月
315 頁 ～ 332 頁

ノンパラメトリック回帰を用いた直線性の検定

竹澤　邦夫*, 辻谷　将明†

Test of Linearity by Nonparametric Regression

Kunio Takezawa* and Masaaki Tsujitani†

単回帰の妥当性を調べることを目的として, 単回帰の結果とノンパラメトリック回帰の結果をF検定を用いて比較すると, 単回帰が適切, という帰無仮説が誤って棄却される確率が, 想定している危険率とはかなり異なることが多い. そこで, F検定に代えてブートストラップ法を用いる方法を提案する. こちらは, 帰無仮説が誤って棄却される確率が適切で, 検出力も高い.

A comparison by F-test between the result of simple regression and that of nonparametric regression for examining the appropriateness of simple regression often gives an incorrect probability of rejecting the null hypothesis erroneously; the probability may be substantially different from the preset probability as the level of significance. A bootstrap method is proposed to replace F-test. This technique features an appropriate probability of rejecting the null hypothesis erroneously and a high power of test.

キーワード: F検定, 直線性, ノンパラメトリック回帰, ブートストラップ法, 平滑化スプライン

1. はじめに

1つの予測変数と1つの目的変数の間の関係を最小2乗法を用いて1次式に回帰するとき, そのデータに1次式をあてはめることが妥当かどうかが問題になることがある. そのとき, 特段の理由がなければ1次式をあてはめた結果を使いたいけれども, 1次式をあてはめることが妥当でないと考えるべき理由があれば, 代わりにノンパラメトリック回帰による回帰式を使う, という方針をとることがある. 具体的には, 当該データが1次式が生み出したものである, という帰無仮説を設定し, ノンパラメトリック回帰の結果と比較することによって, 帰無仮説が棄却されるか否かを調べる. そうした目的のための手段として, 統計量としてF値を用い, それに基づくF検定を行うことで帰無仮説の妥当性を検討する方法が多くの文献で紹介されている（例えば, Loader (1999, p.165), 丹後 (2000, p.109), Ruppert *et al.* (2003, p.148), Faraway (2005, p.236)).

* 中央農業総合研究センター データマイニング研究チーム, 〒 305-8666 茨城県 つくば市 観音台 3-1-1.
† 大阪電気通信大学 情報通信工学部 情報工学科, 〒 572-8530 大阪府 寝屋川市 初町 18-8.

ここでは，この方法による検定がかなりのバイアスを持つ結果をもたらすことをシミュレーションによって示す．そして，この方法に代わる手段として，ブートストラップ法を用いる方法を提案し，その特性を調べるためのシミュレーションを行う．

2. F検定を用いる方法

データとして$\{x_i, y_i\}$ $(1 \leq i \leq n)$（x_iがデータの予測変数の部分，y_iがデータの目的変数の部分，nがデータ数）という形で与えられたとき，以下に示す1次式に回帰することを単回帰と呼ぶ．

$$y = ax + b \tag{2.1}$$

aとbは回帰係数である．すなわち，以下の式がデータを生み出したと考える．

$$y_i = ax_i + b + \epsilon_i \tag{2.2}$$

$\{\epsilon_i\}$ $(1 \leq i \leq n)$は，それぞれが独立で，平均が0で，分散が一定の誤差である．aとbの値は，以下の式が与える値を最小にすることによって求める．

$$RSS_1 = \sum_{i=1}^{n}(y_i - ax_i - b)^2 \tag{2.3}$$

このようにして得られた回帰係数を$\hat{a}$と$\hat{b}$と書く．そのときのRSS_1を残差2乗和と呼ぶ．以下では，RSS_1をこの意味で用いる．

式(2.1)への回帰が妥当かどうかを知るための手段として，ノンパラメトリック回帰によって得られた回帰式と比較する方法がある．その際は，以下のようなノンパラメトリックな関数を用いて回帰式を作成する．

$$y = s(x) \tag{2.4}$$

$s(x)$はノンパラメトリックな回帰関数である．すなわち，以下の式がデータを生み出したと考える．

$$y_i = s(x) + \epsilon_i' \tag{2.5}$$

$\{\epsilon_i'\}$ $(1 \leq i \leq n)$は，それぞれが独立で，平均が0で，分散が一定の誤差である．データを使って推定した$s(x)$を$\hat{s}(x)$とする．

その際，以下のような帰無仮説（H_0）と対立仮説（H_1）を設定する．

H_0: $\{x_i, y_i\}$は式(2.2)が生み出した．
H_1: $\{x_i, y_i\}$は式(2.5)が生み出した．

そして，回帰の結果について検討を加え，帰無仮説が妥当でないことを示す根拠が発見された場合，対立仮説を採用する．

その際，以下のように定義される値（F 値）を用いる．

$$F = \frac{\dfrac{RSS_1 - RSS_2}{df_2 - df_1}}{\dfrac{RSS_2}{n - df_2}} \tag{2.6}$$

ここで，RSS_2 の定義は以下のものである．

$$RSS_2 = \sum_{i=1}^{n} (y_i - \hat{s}(x_i))^2 \tag{2.7}$$

ここで，$\hat{s}(\cdot)$ がノンパラメトリック回帰によって得られた回帰式である．df_1 は単回帰における有効自由度である．単回帰においては a と b の 2 つの回帰係数を用いるので，有効自由度は 2 である．df_2 はノンパラメトリック回帰における有効自由度である．ノンパラメトリック回帰に対応するハット行列の対角要素の和を用いることが多い（例えば，Loader (1999, p.28) の ν_1，Wood (2006, p.171)）．F 分布の定義における自由度は正の整数とされることが多い．しかし，自由度が正の実数の場合にも，F 分布の確率密度関数の定義をそのまま用いたものを自由度が正の実数の場合の F 分布として用いることができる（例えば，Loader (1999, p.166)）．

帰無仮説が正しいとき，式 (2.6) による F 値は，第 1 自由度が $(df_2 - df_1)$，第 2 自由度が df_2 の F 分布に近似的に従う．このことを以下のように書く．

$$F = \frac{\dfrac{RSS_1 - RSS_2}{df_2 - df_1}}{\dfrac{RSS_2}{n - df_2}} \sim F_{(df_2 - df_1), df_2} \tag{2.8}$$

手元のデータに対して，単回帰とノンパラメトリック回帰を行って，それらに基づく F 値を算出し，第 1 自由度が $(df_2 - df_1)$，第 2 自由度が df_2 の F 分布において，その F 値に対応する p 値を求める．その p 値が 0.05 より小さい値であれば，5 %の危険率で帰無仮説が棄却される，と結論づける．すなわち，単回帰の結果を放棄してノンパラメトリック回帰の結果を使うべき，と判断する．

以下ではノンパラメトリック回帰を行うための手法として平滑化スプラインを用いる．平滑化スプラインによる推定値は自然スプラインになるので (Hastie and Tibshirani (1990))，上記の仮説検定の H_1 における $s(x)$ が自然スプラインであることを仮定したことになる．滑らかな関数として自然スプラインを用いることは，自然スプラインがいくつかの点を通る「もっとも滑らかな」補間によって得られるものである（桜井 (1981, 第 2 章)）ことから妥当であると考えられる．また，その自然スプラインを導くための手段として平滑化スプラインを用いることの正当性は，平滑化スプラインの様々な性質（例えば，Green and Silverman (1993, 第 3 章）が裏付けている．また，経験的にも，予測変数が 1 つの場合で

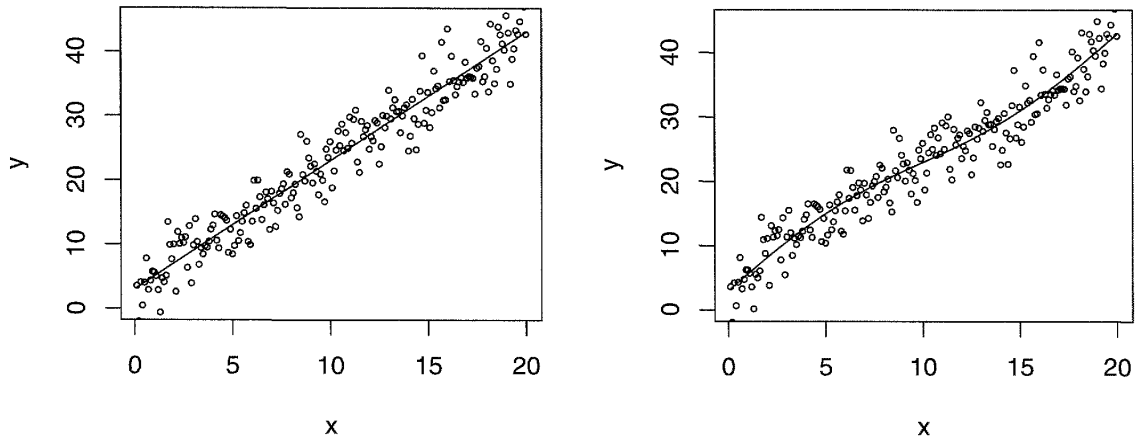

図1 シミュレーションにおいて用いたデータの例．$\alpha = 0$ のときのシミュレーション・データの1つ（○）と単回帰の結果（実線）（左）．$\alpha = 2$ のときのシミュレーション・データの1つ（○）と平滑化スプラインの結果（実線）（右）．

データが特に扱いにくいものでないときは，ノンパラメトリック回帰による平滑化の各手法はほぼ同一の結果をもたらすと言える．従って，ノンパラメトリック回帰による平滑化手法として平滑化スプライン以外の平滑化手法を用いた際にも結果にはほとんど影響しないと考えられる．

3. F検定を用いる方法とダービンワトソン比を用いる方法のシミュレーション

前節で紹介した方法の特性を調べるために，以下のような疑似データを用いたシミュレーションを行った．データは，$\{x_i, y_i\}$ $(1 \leq i \leq 200)$（x_i がデータの予測変数の部分，y_i がデータの目的変数の部分）で，以下の式を用いて得られる．

$$x_i = 0.1 \times i \tag{3.1}$$

$$y_i = 2x_i + 3 + \alpha \times \sin(0.1\pi x_i) + e_i \tag{3.2}$$

$\{e_i\}$ $(1 \leq i \leq 200)$ は $N(0.0, 3.0^2)$（平均が0.0，分散が 3.0^2 の正規分布）の実現値である．α の値として，$0, 0.5, 1, 1.5, 2, 2.5, 3$ の7通りを設定した．$\alpha = 0$ のとき，この200個のデータの目的変数の値は，1次式が与える結果に誤差を加えることで得られたものなので，このデータに対して前節で用いた検定を行えば，5％の確率で帰無仮説が棄却されるはずである．また，α の値が正のときは，高い確率で帰無仮説が棄却されることが好ましい．

そこで，乱数の初期値を代えることで1000組のシミュレーション・データを作成した．それらを使い，7つの α の値のそれぞれに対するシミュレーション・データを生み出した．$\alpha = 0$ と $\alpha = 2$ のときのデータの例を図1に示した．そして，単回帰と平滑化スプラインによる回帰も行った．その際に用いる平滑化パラメータの値は GCV(Generalized

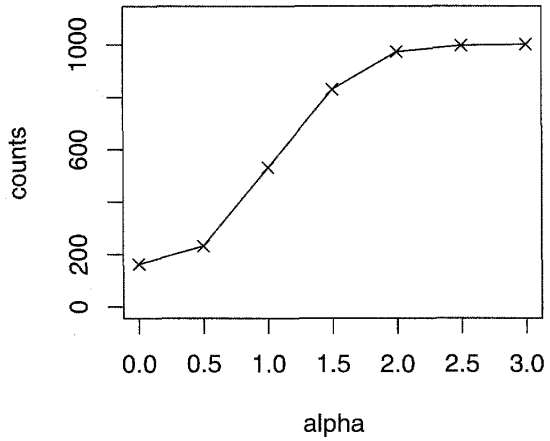

図 2 式 (3.1) と式 (3.2) による 1000 組のシミュレーション・データを用い，式 (2.8) による検定を行った結果.

Cross-Validation) を使って最適化した．以下が，GCV の定義である．

$$GCV = \frac{\sum_{i=1}^{n}(y_i - \hat{s}(\mathbf{x}_i))^2}{n \cdot \left(1 - \frac{\sum_{i=1}^{n}[\mathbf{H}]_{ii}}{n}\right)^2} \tag{3.3}$$

ここで，$[\mathbf{H}]_{ii}$ が線形回帰に伴うハット行列の対角要素である．ノンパラメトリック回帰において GCV を用いて平滑化パラメータの最適化を行うと，得られる推定値が凸凹が大きすぎるものになりやすいことや，得られる平滑化パラメータの値がデータの僅かな変化に大きく影響されるという意味で不安定であることが指摘されている（Hastie and Tibshirani (1990, p.50)，Simonoff (1998, p.175)，シモノフ (1999, p.207)）．また，ノンパラメトリック回帰における平滑化パラメータの最適化のための手法には他に多くのものがある．にもかかわらずここで GCV を用いるのは，ここで平滑化スプラインによる平滑化を行うために用いた `smooth.spline()`（R に標準で所収されている R プログラム）において，デフォルトでは GCV が用いられることに加えて，R を用いて平滑化スプラインとそれを発展させた内容の計算を行うためのパッケージである「assist」(Wang (2011))，「gss」(Gu (2002))，「mgcv」(Wood (2006)) においてもデフォルトで GCV が用いられるので，平滑化スプラインに関連した回帰における平滑化パラメータの最適化においては GCV を用いることが標準と考えられるからである．

このシミュレーションの結果が図 2 である．$\alpha = 0$ のとき，1000 組のデータのうち 162 組のデータにおいて帰無仮説が棄却された．5 %のはずが 16.2 %になった．この検定は，帰無仮説が棄却されやすい方向のバイアスを伴っていることが分かる．

このバイアスの原因として，F 値が本来より大きめに推定されることによる選択バイアスが考えられる (Wood (2006, p.195))．これは，手元のデータを使って平滑化パラメータの値を GCV を指標として最適化することは，有効自由度の割には残差の大きさが小さい回帰式を選択することだからである．このことによって，有効自由度と F 値の関係が，本

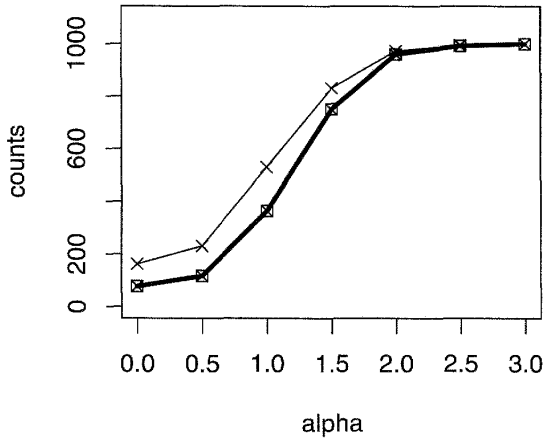

図 3 式 (3.1) と式 (3.2) による 1000 組のシミュレーション・データを用い，独立したデータを使って平滑化パラメータの値を求め，式 (2.8) による検定を行った結果（太線）．細線は，図 2 における曲線と同じもの．

来のものからずれたものになることが考えられる．すなわち，GCV を用いて平滑化パラメータを選ぶ，という選択行為によって，有効自由度と F 値の関係が歪むことによるバイアスなので，選択バイアスの一種と考えられる．このことが p 値にバイアスを与えるのであれば，平滑化パラメータの最適化においては検定を行うためのデータとは独立したデータを用いることでこの問題が解決することが期待できる．

そこで，平滑化パラメータの最適化のためのデータとして式 (3.1) と式 (3.2) を用いた 200 個のデータを使うものの，検定のためのデータを作成するときは乱数の初期値を代えたものを用いてシミュレーションを行った．そして，図 2 と同じ形式のグラフを描いたところ，図 3 が得られた．$\alpha = 0$ のとき，1000 組のデータのうち，77 組のデータにおいて帰無仮説が棄却された．先の，結果（162 組）に比べるとかなり減少しているので，選択バイアスの影響が明らかである．しかし，選択バイアスの影響を除いても 5 %よりやや大きい値になっていることは，選択バイアスだけではなく，p 値の算出方法そのものにも問題があることを示している．

一方，式 (2.8) を使う際の df_2（ノンパラメトリック回帰における有効自由度）としてノンパラメトリック回帰に対応するハット行列の対角要素の和とはやや異なるものを用いる流儀もある (Hastie and Tibshirani (1990, p.66), 丹後 (2000, p.109), Ruppert *et al.* (2003, p.90)）．以下の式を用いるものである．

$$df_2 = n - \mathrm{trace}(2\mathbf{H}) + \mathrm{trace}(\mathbf{H}\mathbf{H}^t) \tag{3.4}$$

$\mathbf{H}$ はノンパラメトリック回帰に対応するハット行列である．Ruppert *et al.* (2003, p.84) には，F 検定を行う場合には，自由度としてハット行列の対角要素の和よりも式 (3.4) の方が好ましいと書かれている．式 (3.4) の df_2 を用いると計算量が多くなるので，ここでは

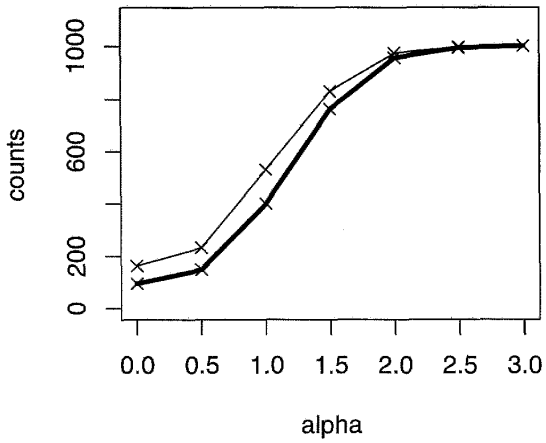

図 4 式 (3.1) と式 (3.2) による 1000 組のシミュレーション・データを用い，式 (2.8) と式 (3.4) による F 検定を行った結果（太線）．細線は，図 2 における曲線と同じもの．

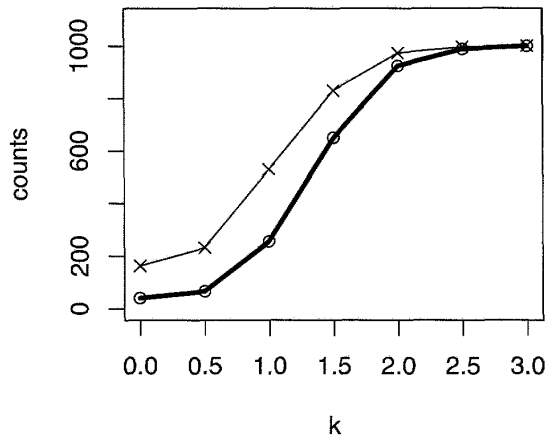

図 5 式 (3.1) と式 (3.2) による 1000 組のシミュレーション・データを用い，独立したデータを使って平滑化パラメータの値を求め，式 (2.8) と式 (3.4) による F 検定を行った結果（太線）．細線は，図 2 における曲線と同じもの．

以下の近似（Hastie and Tibshirani (1990, p.305)）を利用した．

$$df_2 \approx 1.25\mathrm{trace}[\mathbf{H}] - 0.5 \tag{3.5}$$

ただし，Hastie and Tibshirani (1990, p.66) では，式 (3.4) と式 (2.8) を用いた F 値が優れた近似になるのは，この $\mathbf{H}$ が与える推定値が不偏と見なされる場合であると解説されている．平滑化スプラインが与える推定値は滑らかな推定値を与える方向へのバイアスを含んでいるので不偏ではない．そのため，平滑化スプラインによる平滑化を行うときに，式 (3.4) を用いることの妥当性ははっきりしない．

式 (2.8) と式 (3.4) を用いて先のシミュレーションデータ（式 (3.2) を使って得られたもの）による検定を行ったところ，図 4 が得られた．$\alpha = 0$ のとき，1000 組のデータのうち，94 組のデータにおいて帰無仮説が棄却された．また，選択バイアスを回避する方法を用いた場合は，図 5 になった．$\alpha = 0$ のとき，1000 組のデータのうち，40 組のデータにおいて

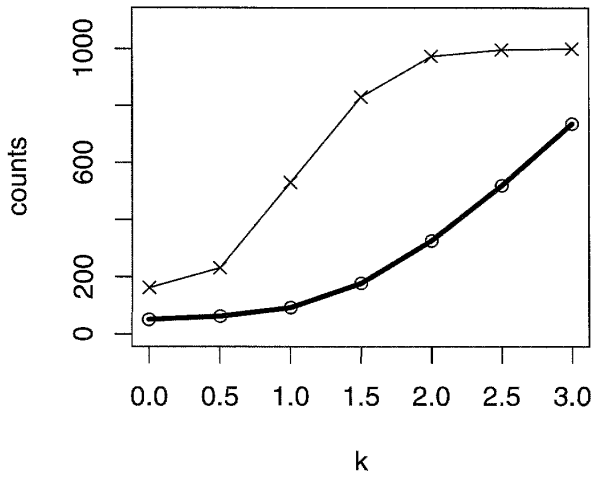

図 6 式 (3.1) と式 (3.2) による 1000 組のシミュレーション・データを用い，独立したデータを使って平滑化パラメータの値を求め，ダービンワトソン比による検定を行った結果（太線）．細線は，図 2 における曲線と同じもの．

帰無仮説が棄却された．つまり，選択バイアスを除くことができれば危険率が 5 パーセントよりも小さくなる．

以上のシミュレーションにより，かなり限定された条件の下でのものではあるけれども，自由度が整数ではないときに式 (2.8) を用いて得られる F 値はかなり粗い近似と見なさなければならず，誤差の方向さえはっきりしないことが明らかになった．丹後 (2000, p.108) では自由度が整数ではないときの F 値が「近似」であることが強調されている．しかし，どの方向にどのくらいの大きさの誤差が生じるかが分からなければ，実用上は厳密な F 値が得られていると考えても差し支えないのかどうかの判断はできない．上記のシミュレーションの結果は，自由度が整数ではないときの F 値の実用的な利用は控えるべきであることを明らかにした．従って，ノンパラメトリック回帰を利用した直線性の検定のための新たな手法が求められている．

他方，直線性を検定するための手法としてノンパラメトリック回帰を用いず，単回帰が与える残差に正の相関がないという帰無仮説の妥当性をダービンワトソン比（例えば，Myers (2000, p.287)）を用いて調べる方法を用いることも考えられる．単回帰が与える残差に正の相関がないという帰無仮説が棄却されれば，直線ではなく曲線をあてはめるべきと主張できると考えられるからである．そこで，先のシミュレーションデータ（式 (3.2) を使って得られたもの）を用いて，単回帰が与える残差に正の相関がないという仮説が棄却されるかどうかをダービンワトソン比を用いて検定した結果を，図 6 に示した．ダービンワトソン比を用いた検定を行うために，R のパッケージ「lmtest」に所収されている `dwtest()` を用いた片側検定（危険率が 5 パーセント）を行った．$\alpha = 0$ のとき，仮説が棄却されたデータセットの数が 51 組になっている．つまり，データが直線から生み出されものであるにもかかわらず，残差に正の相関があるという誤った結論を導いてしまう可能性がほぼ 5 パーセントなので，設定した危険率を正しく反映した結果が得られたと言える．しかし，α

の値が正のときに仮説が棄却されるデータセットの数が少ない．すなわち，検出力が弱い．これは，ダービンワトソン比を用いた検定が仮説を棄却しないのは，直線がデータが生み出したと考えられるときばかりではないことが原因と考えられる．つまり，単回帰が与える残差の相関が特定の部分では正で他の部分では負のときにはデータが直線が生み出したものと考えるべきではないにもかかわらず，仮説を棄却しないことがある．そのため，実際には曲線が生み出したデータセットであるにもかかわらず仮説を棄却しない場合が多くなる．このことは，帰無仮説が棄却されることはデータに曲線をあてはめるべきと主張する強い根拠になりうるけれども，帰無仮説が棄却されない場合でもデータに曲線をあてはめるべきであることが少なくないことを意味する．そのため，検出力が弱くなってしまう．このことから，データが直線が生み出したものであるか否かの検定を行うための手段としてダービンワトソン比を用いた検定は好ましくなく，単回帰の結果とノンパラメトリック回帰の結果を比較する方法の信頼性を高めるべきであることが分かる．

4. ブートストラップ法を用いた検定

前節の結果から，F 検定を用いる方法には，選択バイアスの影響による問題があることが分かった．また，得られる F 値が理論的な F 分布に正確には従わないことによる影響も考えられる．しかし，実際のデータにおいては，選択バイアスの影響を除く形で直線性の検定を行うことは難しい．また，ノンパラメトリック回帰が関連する F 値が厳密には理論的な F 分布に従わないかも知れない点については今後の統計数学の進歩を待たなければならない．そこで，F 検定を行う代わりに，ブートストラップ法（Efron and Tibshirani (1994)，Wu (1986)）の一種と考えられるいくつかの方法を提案する．これらの方法は，Härdle and Mammen (1993) の方法に比べて直感的に理解しやすく，簡単に実行できる．

最初に提案する方法では，単回帰による推定値と平滑化スプラインによる推定値の差（D_1 とする）が大き過ぎるとき，そのデータは 1 次式に回帰すべきではない，と考える．そのため，ブートストラップ法を用いて単回帰に相応しい疑似データを大量に作成し，それらを用い，単回帰による推定値と平滑化スプラインによる推定値の差（$\{D_2^b\}$ とする）の分布を求める．そして，その分布の中での D_1 の位置を調べることによって，元々のデータが単回帰に相応しいか否かを検定する．

以下がそのアルゴリズムである．

手順 (1) データ（$\{x_i, y_i\}$ $(1 \leq i \leq n)$）に最小 2 乗法を用いて 1 次式をあてはめる．得られた 1 次式を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \hat{y}_i\}$ $(1 \leq i \leq n)$ とする．

手順 (2) データ（$\{x_i, y_i\}$）を平滑化スプラインを用いて平滑化する．その際に用いる平滑化パラメータの値は GCV を用いて最適化する．平滑化スプラインによって得られたスプ

ライン関数を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \tilde{y}_i\}$ $(1 \leq i \leq n)$ とする．

手順 (3) 1次式が与える推定値と平滑化スプラインが与える推定値の間の距離を以下のように定義する．

$$D_1 = \sum_{i=1}^{n} (\hat{y}_i - \tilde{y}_i)^2 \tag{4.1}$$

手順 (4) 以下の式を用いて100組のブートストラップ・データ（$\{x_i, y_i^b\}$ $(1 \leq i \leq n, 1 \leq b \leq 100)$）を作成する．

$$y_i^b = \hat{y}_i + \epsilon_i^b \tag{4.2}$$

ここで，$\{\epsilon_i^b\}$ は $N(0.0, VAR)$（平均が0.0，分散が VAR の正規分布）の実現値である．この VAR の値は，以下の式を用いて求める．

$$VAR = \frac{\sum_{i=1}^{n} (y_i - \hat{s}(\mathbf{x}_i))^2}{n - \sum_{i=1}^{n} [\mathbf{H}]_{ii}} \tag{4.3}$$

手順 (5) それぞれのブートストラップ・データに最小2乗法を用いて1次式をあてはめる．得られた1次式を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \hat{y}_i^b\}$ $(1 \leq i \leq n)$ とする．

手順 (6) それぞれのブートストラップ・データを平滑化スプラインを用いて平滑化する．その際に用いる平滑化パラメータの値は GCV を用いて最適化する．平滑化スプラインによって得られたスプライン関数を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \tilde{y}_i^b\}$ $(1 \leq i \leq n)$ とする．

手順 (7) それぞれのブートストラップ・データにおいて，1次式が与える推定値と平滑化スプラインが与える推定値の間の距離を以下のように定義する．

$$D_2^b = \sum_{i=1}^{n} (\hat{y}_i^b - \tilde{y}_i^b)^2 \tag{4.4}$$

手順 (8) 手順 (7) が与える $\{D_2^b\}$ $(1 \leq b \leq 100)$ のうち，手順 (3) が与える D_1 よりも大きいものが5個以下しかなかったとき，帰無仮説（H_0: $\{x_i, y_i\}$ は式 (2.2) が生み出した）を棄却し，対立仮説（H_1: $\{x_i, y_i\}$ は式 (2.5) が生み出した）を採用する．

この方法は，元々のデータを使って1次式に回帰して得られる推定値と平滑化スプラインに回帰して得られる推定値の間の距離を調べ，次に，帰無仮説を仮定して得られるブートストラップ・データにおける同様の距離を調べ，この2種類の距離の関係が，帰無仮説の元では実現する可能性が低いとき，帰無仮説を棄却するものである．2種類の距離の関係だけを利用しているので，F 分布のような特定の分布を仮定していない．その意味でもノンパラメトリックである．しかし，手順 (4) において作成するブートストラップ・データは，$N(0.0, VAR)$ を用いて作成する．式 (2.2) の $\{\epsilon_i\}$ と式 (2.5) の $\{\epsilon_i'\}$ は正規分布に従

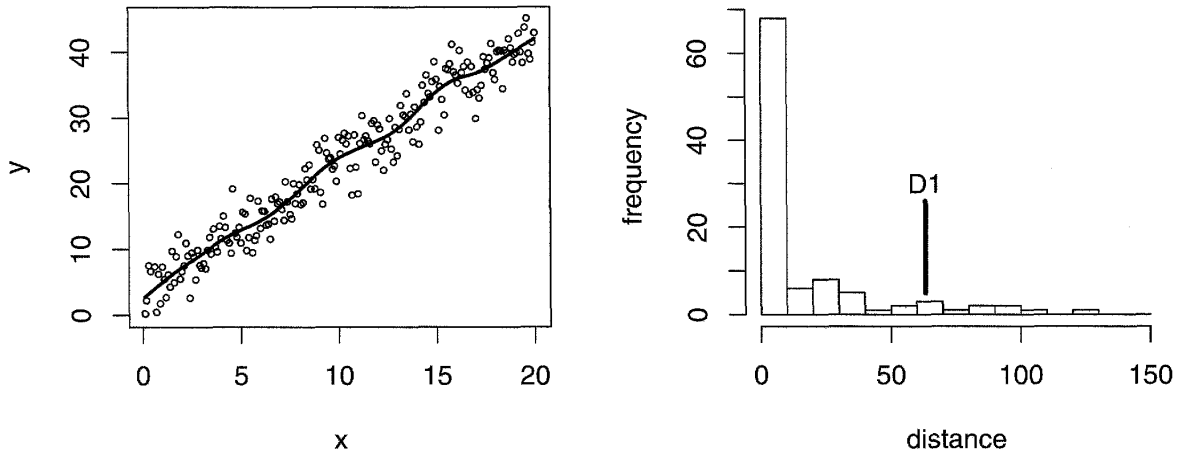

図 7 $\alpha = 0$ のときのシミュレーション・データのうちの 1 つを使った結果．データ（○）とノンパラメトリック回帰の結果（実線）（左）．ラグで示したのが 100 個のブートストラップ・データによる $\{D_2^b\}$ の値で，$D1$ と示した位置が D_1 の値である．10 個の値が D_1 より大きい（右）．

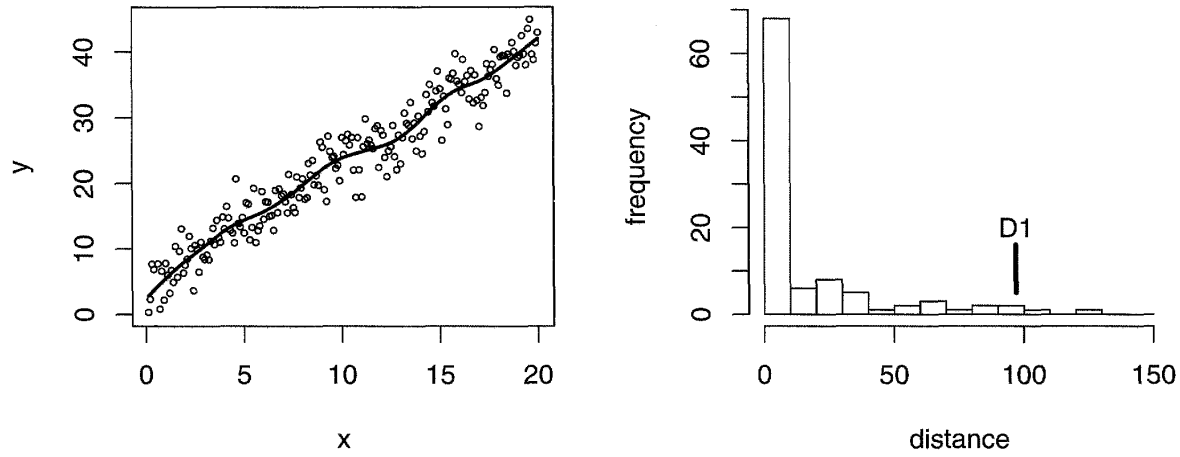

図 8 $\alpha = 1.5$ のときのシミュレーション・データのうちの 1 つを使った結果．データ（○）とノンパラメトリック回帰の結果（実線）（左）．ラグで示したのが 100 個のブートストラップ・データによる $\{D_2^b\}$ の値で，$D1$ と示した位置が D_1 の値である．3 個の値が D_1 より大きい．（右）．

うことを仮定していないので，正規分布を仮定してブートストラップ・データを作成することは，誤差が正規分布に従うという条件の下でのパラメトリック・ブートストラップ法を実行していることに相当する．

図 7 が，このアルゴリズムを使い，前節で示したシミュレーション・データのうち，$\alpha = 0$ のときのシミュレーション・データの内の 1 つを用いた結果である．図 7（左）は，このときのデータの値とノンパラメトリック回帰による推定値を示している．図 7（右）は，100 個のブートストラップ・データによる $\{D_2^b\}$ の値のヒストグラムである．100 個の $\{D_2^b\}$ の内，10 個の値が D_1 より大きい．従って，帰無仮説（H_0: $\{x_i, y_i\}$ は式 (2.2) が生み出した）は棄却されない．一方，図 8 は，$\alpha = 1.5$ のときのシミュレーション・データの内の 1 つを用いて図 7 と同様の作業を行った結果である．100 個の $\{D_2^b\}$ の内，3 個の値が D_1 より大きい．従って，帰無仮説（H_0: $\{x_i, y_i\}$ は式 (2.2) が生み出した）は棄却される．

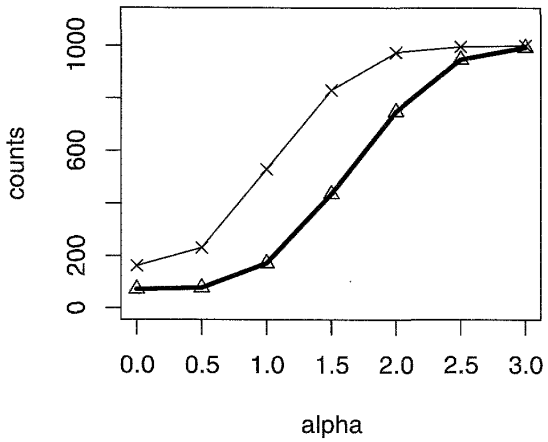

図 9 前節と同じシミュレーション・データを用い，ブートストラップ法を用いた検定を行った結果（太線）．細線は，図 2 における曲線と同じもの．

図 9 は，このシミュレーションの結果の全体を示している．$\alpha = 0$ のとき，1000 組のデータのうち，72 組のデータにおいて帰無仮説が棄却された．5 %にかなり近い．しかし，72 組のデータにおいて帰無仮説が棄却されたことは，図 3 において，$\alpha = 0$ のとき，1000 組のデータのうち，77 組のデータにおいて帰無仮説が棄却されたことと比べて，本質的な改善と言えるかどうかははっきりしない．また，α として正の値を与えたときの結果は，この方法の検出力がやや低いことを示している．従って，ブートストラップ法を用いた更に優れた方法を模索する必要があると考えられる．

5. ブートストラップ法を用いて棄却域を調整

前節の方法は F 分布を用いずに検定を行うものであった．しかし，F 分布はそのまま用いて，F 検定における棄却領域を決定するためにブートストラップ法を用いる方法も考えられる．多重比較におけるボンフェローニ（Bonferroni）の方法にブートストラップ法を組み合わせたような方法である．

以下がそのアルゴリズムである．

手順 (1) データ（$\{x_i, y_i\}$ $(1 \leq i \leq n)$）に最小 2 乗法を用いて 1 次式をあてはめる．得られた 1 次式を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \hat{y}_i\}$ $(1 \leq i \leq n)$ とする．そのときの残差 2 乗和を RSS_1 とする．有効自由度を df_1 $(= 2)$ とする．

手順 (2) データを平滑化スプラインを用いて平滑化する．その際に用いる平滑化パラメータの値は GCV を用いて最適化する．平滑化スプラインによって得られたスプライン関数を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \tilde{y}_i\}$ $(1 \leq i \leq n)$ とする．そのときの残差 2 乗和を RSS_2 とする．有効自由度（ハット行列の対角要素の和）を df_2 とする．

手順 (3) 手順 (1) と手順 (2) の結果を用いて，データの直線性を検定するためのF値（F_1）を求める．以下のものである．

$$F_1 = \frac{\frac{RSS_1 - RSS_2}{df_2 - df_1}}{\frac{RSS_2}{n - df_2}} \tag{5.1}$$

この F_1 に対応するp値を p_1 とする

手順 (4) 以下の式を用いて100組のブートストラップ・データ（$\{x_i, y_i^b\}$ $(1 \leq i \leq n, 1 \leq b \leq 100)$）を作成する．

$$y_i^b = \hat{y}_i + \epsilon_i^b \tag{5.2}$$

ここで，$\{\epsilon_i^b\}$ は $N(0.0, VAR)$（平均が0.0，分散が VAR の正規分布）の実現値である．

手順 (5) それぞれのブートストラップ・データに最小2乗法を用いて1次式をあてはめる．得られた1次式を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \hat{y}_i^b\}$ $(1 \leq i \leq n)$ とする．そのときの残差2乗和を RSS_1^b とする．有効自由度を df_1^b $(= 2)$ とする．

手順 (6) それぞれのブートストラップ・データを平滑化スプラインを用いて平滑化する．その際に用いる平滑化パラメータの値は GCV を用いて最適化する．平滑化スプラインによって得られたスプライン関数を用いて，それぞれのデータに対応する推定値を求める．その推定値を $\{x_i, \tilde{y}_i^b\}$ $(1 \leq i \leq n)$ とする．そのときの残差2乗和を RSS_2^b とする．有効自由度（ハット行列の対角要素の和）を df_2^b とする．

手順 (7) それぞれのブートストラップ・データにおいて，データの直線性を検定するためのF値（F^b）を求める．以下のものである．

$$F^b = \frac{\frac{RSS_1^b - RSS_2^b}{df_2^b - df_1^b}}{\frac{RSS_2^b}{n - df_2^b}} \tag{5.3}$$

この F^b に対応するp値を p^b とする

手順 (8) 手順 (7) が与える $\{p^b\}$ $(1 \leq b \leq 100)$ を小さい順に並べ，最初から5番目の値を p^* とする．

手順 (9) $p_1 < p^*$ のとき，帰無仮説（H_0: $\{x_i, y_i\}$ は式 (2.2) が生み出した）を棄却し，対立仮説（H_1: $\{x_i, y_i\}$ は式 (2.5) が生み出した）を採用する．

図10が，前々節と同じデータを用い，図9と同じシミュレーションを行った結果のうち，$\alpha = 0$ のときの p^* の値の分布のヒストグラムである．0.01くらいを最頻値として左右に広がる分布である．式 (2.6) を率直に用いた場合は0.05になることと比較すると，従来法が大きなバイアスを伴う方法であることが分かる．

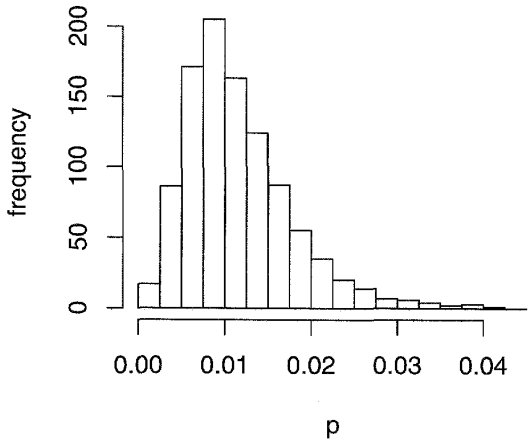

図 10 $\alpha = 0$ のときのシミュレーション・データ（1000 個）における，採択域と棄却域の境界の点の値（$\{p^*\}$）の分布

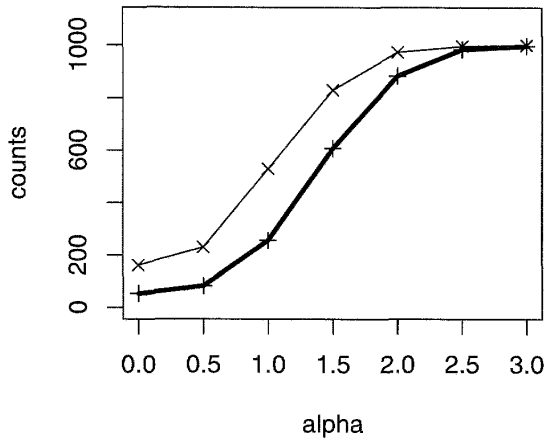

図 11 前々節と同じシミュレーション・データを用い，ブートストラップ法を用いた棄却域の調整を行った結果（太線）．細線は，図 2 における曲線と同じもの．

図 11 が，前々節と同じデータを用い，図 9 と同じシミュレーションを行った結果である．$\alpha = 0$ のとき，1000 組のデータのうち，53 組のデータにおいて帰無仮説が棄却された．5 %に近い．しかも，α として正の値を与えたときの結果は，検出力がかなり高いことを示している．

しかし，この方法は手順 (4) においてブートストラップ・データを作成する際に誤差が正規分布に従っていることを仮定している．この点の仮定を弱めるために，式 (5.2) の $\{\epsilon_i^b\}(1 \leq i \leq n)$ として，以下のものから重複を許してランダムにサンプルしたものを用いる方法も試みた．

$$\left\{ \frac{y_i - \hat{s}(\mathbf{x}_i)}{\sqrt{1 - \sum_{i=1}^{n} [\mathbf{H}]_{ii}/n}} \right\} \tag{5.4}$$

図 12 が，前々節と同じデータを用い，図 9 と同じシミュレーションを行った結果である．$\alpha = 0$ のとき，1000 組のデータのうち，46 組のデータにおいて帰無仮説が棄却された．図 12 を図 11 と比べると，誤差が正規分布に従っていることは直接的には仮定しなかった

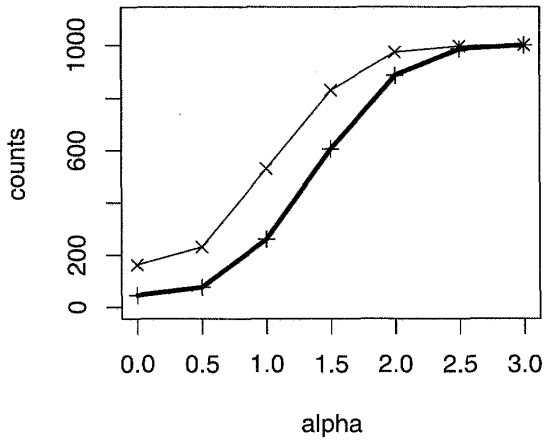

図 12 前々節と同じシミュレーション・データを用い，正規分布を仮定しないブートストラップ法を用いた棄却域の調整を行った結果（太線）．細線は，図 2 における曲線と同じもの．

ことと，ブートストラップ・データの誤差を得るために滑らかな確率密度関数を用いるのではなく残差に定数を掛けたものをリサンプリングしたものを用いたことによる検出力の低下は見られない．

6. 定数か否かの検定

データとして $\{x_i, y_i\}$ $(1 \leq i \leq n)$（x_i がデータの予測変数の部分，y_i がデータの目的変数の部分，n がデータ数）という形で与えられていたとき，以下に示すように，定数に回帰したくなることがある．

$$y = b \tag{6.1}$$

b が回帰係数である．すなわち，以下の式がデータを生み出したと考える．

$$y_i = b + \epsilon_i \tag{6.2}$$

$\{\epsilon_i\}$ $(1 \leq i \leq n)$ は，それぞれが独立で，平均が 0 で，分散が一定の誤差である．予測変数の値にかかわらず目的変数が定数にランダムな誤差を加えたものと見なされるときの回帰式である．予測変数が時間であれば，目的変数が時間に伴う特段の傾向を持たないことを意味する．b の推定値（$\hat{b}$）は，$\{y_i\}$ の平均値である．そのときの残差 2 乗和を RSS_1 とする．

式 (6.1) への回帰が妥当かどうかを知るために，以下のようなノンパラメトリックな関数を用いた回帰も行う．

$$y = s(x) \tag{6.3}$$

$s(x)$ はノンパラメトリックな回帰関数である．すなわち，以下の式がデータを生み出したと考える．

$$y_i = s(x) + \epsilon_i' \tag{6.4}$$

$\{\epsilon_i'\}$ $(1 \leq i \leq n)$ は，それぞれが独立で，平均が 0 で，分散が一定の誤差である．データを使って推定した $s(x)$ を $\hat{s}(x)$ とする．

その際，以下のような帰無仮説（H_0）と対立仮説（H_1）を設定する．

H_0: $\{x_i, y_i\}$ は式 (6.2) が生み出した．

H_1: $\{x_i, y_i\}$ は式 (6.4) が生み出した．

そして，回帰の結果について検討を加え，帰無仮説が妥当でないことを示す拠が発見されたとき，対立仮説を採用する．

その際に，以下のように定義される値を用いる．

$$F = \frac{\frac{RSS_1 - RSS_2}{df_2 - df_1}}{\frac{RSS_2}{n - df_2}} \tag{6.5}$$

ここで，RSS_2 の定義は以下のものである．

$$RSS_2 = \sum_{i=1}^{n} (y_i - \hat{s}(x_i))^2 \tag{6.6}$$

df_1 は定数への回帰における有効自由度（$= 1$）である．df_2 はノンパラメトリック回帰における有効自由度である．それらを用いて，式 (6.5) が与える F を用いて検定を行うのが従来法である．一方，前節で示した方法においても，$df_1 = 1$，$df_1^b = 1$ $(1 \leq b \leq 100)$ とすることによって，式 (6.1) に回帰すべきかどうかの検定を行うことができる．

これら 2 つの方法の特性を調べるために，以下のようなシミュレーションを行った．データは，$\{x_i, y_i\}$ $(1 \leq i \leq 200)$（x_i がデータの予測変数の部分，y_i がデータの目的変数の部分）で，以下の式を用いて得られる．

$$x_i = 0.1 \times i \tag{6.7}$$

$$y_i = 3 + \alpha \times \sin(0.1\pi x_i) + e_i \tag{6.8}$$

$\{e_i\}$ $(1 \leq i \leq 200)$ は $N(0.0, 3.0^2)$（平均が 0.0，分散が 3.0^2 の正規分布）の実現値である．α の値として，$0, 0.5, 1, 1.5, 2, 2.5, 3$ の 7 通りを設定した．このようにして作成したシミュレーション・データの例を図 13 に示した．

得られた結果が図 14 である．$\alpha = 0$ のときの結果は，従来法が 180 であるのに対して，前節で示した方法は 48 である．つまり，従来法は，帰無仮説を棄却する方向への強いバイアスを伴う結果をもたらすけれども，前節で示した方法は，妥当な結果を与える．α の値が正のときの結果は，前節で示した方法もかなりの検出力を持つことを示している．

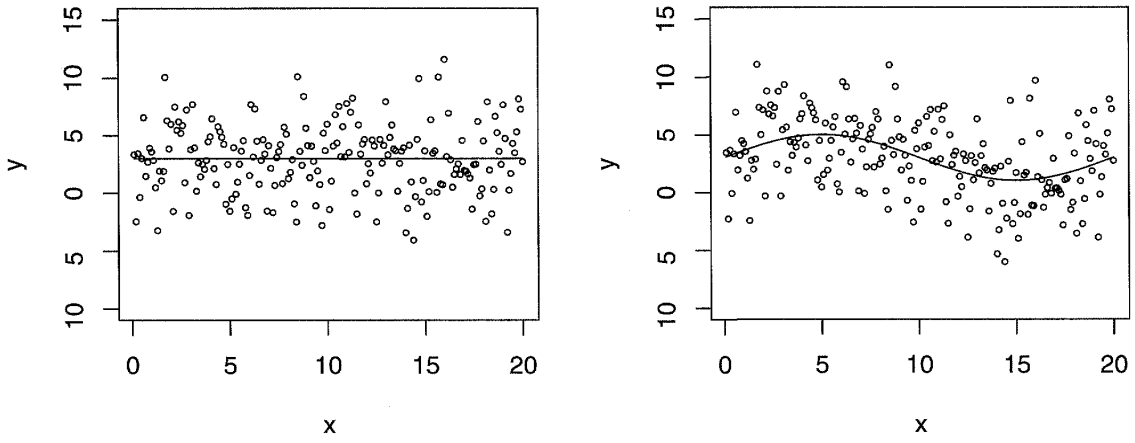

図 13 シミュレーションにおいて用いたデータの例．$\alpha = 0$ のときのシミュレーション・データの 1 つ（左）．$\alpha = 2$ のときのシミュレーション・データの 1 つ（右）．

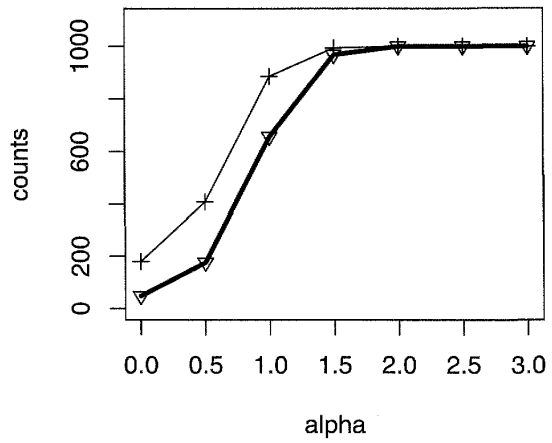

図 14 式 (6.7) と式 (6.8) による 1000 組のシミュレーション・データを用い，式 (6.5) による検定を行った結果（細線）．前節で示した方法による検定の結果（太線）．

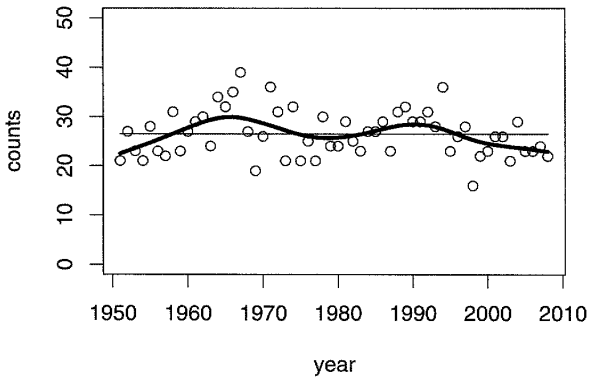

図 15 1951 年から 2008 年までの各年の台風の発生数．細線が定数をあてはめた結果．太線が平滑化スプラインによる平滑化の結果．

この方法を実際のデータに適用した．用いたデータは，1951 年から 2008 年までの各年の台風の発生数である（気象庁（2008））．データ数は 58 個になる．このデータと，このデータに定数をあてはめた結果，更に，平滑化スプラインによる平滑化（平滑化パラメータは GCV を用いて最適化）を行った結果を図 15 に描いた．式 (6.5) が与える F による p 値

は0.00649なので，従来法に従えば，危険率を1％に設定した場合でも，「H_0: $\{x_i, y_i\}$ は定数が生み出した」という帰無仮説は棄却される．しかし，前節で示した方法を用い，精度を高めるため，ブートストラップ・データの数を10000個とし，$\{p^b\}$ $(1 \leq b \leq 10000)$ を小さい順に並べ，最初から500番目の値を p^* と定義した．すなわち，危険率を5％に設定した．すると，p^* の値が0.00357になった．先に求めたp値（0.00649）より小さい値なので，危険率を5％にした場合でも帰無仮説は棄却されないことが結論づけられる．

7. おわりに

以上のシミュレーションの結果から，従来の方法よりもブートストラップ法を用いた方法の方が優れた結果を与えることが分かった．特に，ブートストラップ法を用いて棄却域を調整する方法は，危険率の点でも検出力の点でも望ましい特性を持つ．アルゴリズムが単純でプログラミングが容易である点でも従来法に取って代わるべき内容を備えている．

謝辞

査読者の方からの様々な御指摘によって本論文の内容が大きく向上した．厚く御礼申し上げる．

参 考 文 献

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, Chapman & Hall/CRC.

Faraway, J. J. (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, Chapman & Hall/CRC.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*, Chapman & Hall/CRC.

Gu, C. (2002). *Smoothing spline ANOVA models*, Springer.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits, *Ann. Stat.*, **21**(4), 1926–1947.

Hastie, T. and Tibshirani, R. J. (1990). *Generalized additive models*, Chapman & Hall/CRC.

気象庁 (2008). 台風の発生数（2007年までの確定値と2008年の速報値）http://www.data.jma.go.jp/fcd/yoho/typhoon/statistics/generation/generation.html.

Loader, C. (1999). *Local regression and likelihood*, Springer.

Myers, R. H. (2000). *Classical and modern regression with applications 2 edition*, Duxbury Press.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression*, Cambridge University Press.

桜井明 (1981).『スプライン関数入門』東京電機大学出版局.

Simonoff, J. S. (1998). *Smoothing methods in statistics, 2nd printing*, Springer.

シモノフ, J. S.，竹沢邦夫 (翻訳), 大森宏 (翻訳) (1999). 平滑化とノンパラメトリック回帰への招待（*Smoothing methods in statistics, 2nd printing* の日本語版），農林統計協会

丹後俊郎 (2000).『統計モデル入門』朝倉書店.

Wang, Y. (2011). *Spline smoothing: methods and applications with R*, Chapman & Hall/CRC.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*, Chapman & Hall/CRC.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *Ann. Stat.*, **14**(4), 1261–1295.