# トレンドの変化点を推定可能な区分線形回帰の SAS による実装と応用事例のご紹介

○ビービ　彩加[1]、菊池　宏和[1]

([1]シミック（株）データサイエンス事業本部）

Techniques for Estimating Changes in Trends: Piecewise Linear Regression and Its Implementation in SAS

Ayaka Beebe, Hirokazu Kikuchi
Data Science Div., CMIC Co., Ltd.

## 要旨

途中で傾向が変わる線形回帰を求めることができる、区分線形回帰を紹介する。また、例を用いて SAS の PROC NLIN の使用方法を理解する。
キーワード：区分線形回帰モデル、複数の区分、変化点、PROC NLIN

## 1　はじめに

実データでは、データの中に区分ごとの傾きの違いが存在する場合、説明変数の範囲ごとに異なる関係性が現れることもある。このような状況では、従来の単一の線形回帰モデルを用いるだけでは変数間の関係の変化を十分に捉えることができず、推定結果に偏りが生じるリスクがある。区分線形回帰モデルはデータを複数の区間に分け、それぞれの区間で異なる線形関係を容認することによって、このような状態でも適切に対応できる臨機応変なモデル化手法である。

区分線形回帰モデルの推定では、各区間の回帰係数だけでなく、区切りとなる変化点の位置も同時に推定することが可能である。SAS では、PROC NLIN を用いることで独自のモデル式を柔軟に定義できるため、回帰係数と変化点を同時に推定する区分線形回帰モデルを非線形回帰として実装することが可能である。

本論文では、区分線形回帰モデルの基本的な理論と SAS での PROC NLIN による実装方法について、概要を説明する。SAS HELP のデータを用いた例を通じて、本手法の有用性や注意点について解説し、具体的な応用方法も示す。

## 2　区分線形回帰モデル

### 2.1　区分線形回帰

区分線形回帰（piecewise linear regression）は、説明変数の値域を複数区間に分割し、各区間で独立に線形回帰式を適用することにより、変化点（ブレークポイント）を境に関係性が変化するデータをモデルで表現する統計的手法である。特に、変化点の位置が未知の場合は、モデルが非線形となるため、パラメータ推定には非線形最小二乗法（nonlinear least squares method）が用いられることが多い。本手法を用いることで、従来の単一の線形回帰では対応できない傾向の変化を読み取ることが可能となり、区間ごとに異なる傾向を柔軟にモデリングできる点が特徴である。

### 2.2　非線形最小二乗法

区分線形回帰において変化点が未知の場合、回帰モデルは線形部分と変化点自体の両方をパラメータとして含むため、全体として非線形な推定問題となる。このとき適用される非線形最小二乗法は、観測値とモデルによる推定値の差（残差）の二乗和を最小化するパラメータ値を求める手法である。具体的には、初期推定値を設定した上で、パラメータ（変化点の位置および各区間の回帰係数）を段階的に調整し、残差平方和が最も小さくなるように最適化を行うアルゴリズムが用いられる。

区分線形回帰を使用し、変化点$(t)$が未知の場合、モデルは次のように表される。

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i, & x_i \le t \\ \beta_0 + \beta_1 t + \beta_2 (x_i - t), & x_i > t \end{cases} \qquad (1)$$

このとき、$(\beta_0, \beta_1, \beta_2)$は各区間の回帰係数、$t$は変化点の位置であり、説明変数を$x_i$、応答変数$y_i$と設定する。変化点$(t)$がモデル式の中に含まれ、区分ごとに線形性が異なるため、モデルはパラメータに対して非線形となり、通常の線形最小二乗法では推定ができない。

非線形最小二乗法は、次の目的関数を最小化するパラメータ$(\beta_0, \beta_1, \beta_2, t)$を探索する。

$$S(\beta_0, \beta_1, \beta_2, t) = \sum_{i=1}^{n} [y_i - f(x_i; \beta_0, \beta_1, \beta_2, t)]^2 \qquad (2)$$

ここで、$f(x_i; \beta_0, \beta_1, \beta_2, t)$は式(1)の区分線形回帰モデルを表している。この方法により、観測値$(y_i)$とモデルによる推定値との差（残差）の二乗和を最小にするパラメータセットが求められ、変化点と各区間の回帰係数を同時に推定することが可能となる。数式における $i$は、データの各観測点（各サンプル）を表すインデックス（番号）である。たとえば、データが $N$件あれば、$i$ は 1 から$N$ までの値を取り、上記の数式では、$i = 1$ から $N$ までの各データについて、「観測値 $y_i$」と「モデルから予測した値 $f(x_i; \beta_0, \beta_1, \beta_2, t)$」の差を二乗し、それらを全て足し合わせることで推定できる。

## 3　SAS の実装について

### 3.1　解析データセット

本分析で使用するデータセットは SAS Help ライブラリに記載されている LIDAR データセット[*1] である。本データセットは、平滑化手法や回帰モデルの例として、多くの書籍や論文で利用されている。LIDAR（Light Detection and Ranging）は、レーザーの反射を用いて大気中の化学物質を検出する技術であり、科学者によって広く用いられている。データセットの各変数の詳細を Table 1 に、データセットの値の中身を Table 2 に示す。

Table 1. LIDAR データセットの内容

| 変数名 | 内容 |
|---|---|
| RANGE | 光が反射して戻ってくるまでの距離 |
| LOGRATIO | レーザーから受け取る光の量の比を対数にしたもの |

Table 2. LIDAR データセット

| RANGE | LOGRATIO |
|---|---|
| 390 | -0.05035573 |
| 391 | -0.06009706 |
| … | … |
| 718 | -0.557754 |
| 720 | -0.8026684 |

## 3.2　RANGE と LOGRATIO の基本情報

本論文では RANGE を説明変数、LOGRATIO を応答変数として設定する。まず、LIDAR データセットの説明変数を均等に 3 つに分けたときの変化点（33 パーセンタイルと 66 パーセンタイル）を算出したものを Table 3 にまとめた。

Table 3. RANGE の 33 パーセンタイルと 66 パーセンタイル

| 変数 | 33rd Percentile | 66th Percentile |
|---|---|---|
| **RANGE** | 498 | 607 |

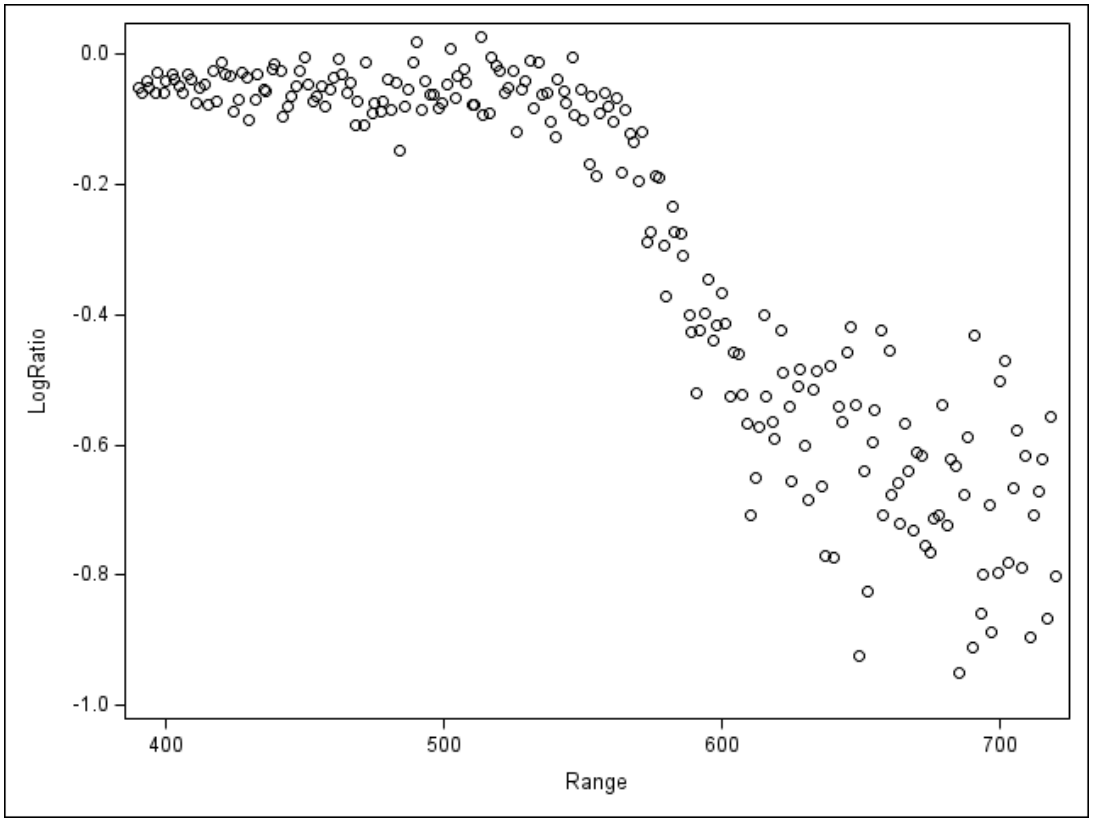
33 パーセンタイルと 66 パーセンタイルで RANGE のデータを分けた区分ごとに算出した LIDAR データセットの主要な統計量を Table 4 に記載した。

Table 4. LIDAR データセットの要約統計量

| 区分 | 変数 | N | Mean | SD | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|---|
| 全体 | RANGE | 221 | 554.8 | 95.9 | 390.0 | 472.0 | 555.0 | 637.0 | 720.0 |
| | LOGRATIO | 221 | -0.29 | 0.28 | -0.95 | -0.54 | -0.11 | -0.05 | 0.03 |
| $RANGE \leq 498$ | LOGRATIO | 73 | -0.05 | 0.03 | -0.15 | -0.07 | -0.05 | -0.03 | 0.02 |
| $498 < RANGE \leq 607$ | LOGRATIO | 73 | -0.17 | 0.16 | -0.53 | -0.28 | -0.09 | -0.06 | 0.03 |
| $RANGE > 607$ | LOGRATIO | 75 | -0.64 | 0.13 | -0.95 | -0.72 | -0.63 | -0.54 | -0.40 |

Figure 1 は、RANGE（x 軸）と LOGRATIO（y 軸）の関係を図に示している。

Figure 1. RANGE と LOGRATIO の散布図




### 3.3　区分線形回帰モデルを用いる背景

RANGE と LOGRATIO の関係性を考察する。Figure 1 から、両者の関係は明らかに非線形であり、通常の線形回帰モデルでは十分に捉えられないことが示唆される。そのため、より柔軟なモデリング手法が必要となる。また、図中には 2 つの変化点が認められ、RANGE と LOGRATIO の関係は 3 つの区間で異なる傾向を示している。本論文では、データをこれらの区間ごとに分割し、各区間に線形モデルを適用する区分線形回帰に着目する。

### 3.4　区分線形回帰モデルの数式

2 つの変化点を持った区分線形回帰モデルは以下の数式で表すことができる。パラメータの内容は Table 5 にまとめた。第一区間のモデル式は$b$を傾き、$a$を切片とした一般的な一次関数の式$y = a + bx$を設定し、第二区間では最初の変化点$(t_1)$の影響を考慮したモデル式を構築した。第三区間に対しても、第二区間のモデル式に 2 つ目の変化点$(t_2)$の影響を配慮したモデル式を設定している。

$$y_i = \begin{cases} a_1 + b_1 * x_i & x_i \le t_1 \\ a_1 + b_1 * t_1 + b_2 * (x_i - t_1) & t_1 < x_i \le t_2 \\ a_1 + b_1 * t_1 + b_2 * (t_2 - t_1) + b_3 * (x_i - t_2) & x_i > t_2 \end{cases} \quad (3)$$

Table 5. 区分線形回帰モデルのパラメータ

| パラメータ | 内容 |
|---|---|
| $a_1$ | 第一区間の切片 |
| $b_1$ | 第一区間の傾き |

| | |
|---|---|
| $b_2$ | 第二区間の傾き |
| $b_3$ | 第三区間の傾き |
| $t_1$ | 第一変化点 |
| $t_2$ | 第二変化点 |

定義した区分線形回帰モデルを SAS で実装するには PROC NLIN が適応可能である。

## 3.5 PROC NLIN とは

SAS の PROC NLIN は、非線形最小二乗法によりパラメータを推定するためのプロシジャであり、モデル式の自由な設定やパラメータの初期値の指定が可能である。区分線形回帰モデルにおいても、変化点の位置や回帰係数など複数のパラメータを一括して推定ができる。

本論文では、PROC NLIN を用いて区分線形回帰モデルの構築およびパラメータ推定を行う。

最適なパラメータを決定するプロセスは反復的に行われる。まず初期パラメータ値を設定する必要がある。初期パラメータ値は、データの要約統計量および散布図を参考に設定する。PROC NLIN は、パラメータの値を順次調整しながら、モデルの当てはまりを向上させる。このパラメータの調整の 1 回分を反復（イテレーション）と呼ぶ。次の反復でも同様に、パラメータを修正することでさらに当てはまりの改善を試みる。各反復で当てはまりの改善が得られなくなった時点で、モデルは収束したと判断される。

## 3.6 PROC NLIN を使った区分線形回帰モデルの実装

初期パラメータは Table 4 の要約統計量や Figure 1 の散布図を元に判断し、決定する。Table 6 に今回設定した初期パラメータの値をまとめた。

Table 6. 区分線形回帰モデルの初期パラメータ

| パラメータ | 内容 | 値 |
|---|---|---|
| $a_1$ | 第一区間の切片 | -0.1 |
| $b_1$ | 第一区間の傾き | 0.0001 |
| $b_2$ | 第二区間の傾き | -0.001 |
| $b_3$ | 第三区間の傾き | -0.0005 |
| $t_1$ | 第一変化点 | 500 |
| $t_2$ | 第二変化点 | 600 |

以下のコードを用いて SAS で PROC NLIN を実装する。また、SAS での実装は、SAS Version 9.4 を使用した。

```
proc nlin data=lidar converge=1e-3;
   /* 初期パラメータ値の設定 */
   parms a1=-0.1        /* 第一区間の切片 */
         b1=0.0001      /* 第一区間の傾き */
         b2=-0.001      /* 第二区間の傾き */
         b3=-0.0005     /* 第三区間の傾き */
         t1=500         /* 第一変化点 */
         t2=600;        /* 第二変化点 */

   /* 区分線形回帰モデルの定義 */
   if (range <= t1) then pred = a1 + b1*range;
   else if (range <= t2) then pred = a1 + b1*t1 + b2*(range-t1);
   else pred = a1 + b1*t1 + b2*(t2-t1) + b3*(range-t2);

   /* モデル式を設定 */
   model logratio = pred;

   /* 変化点に制約を設定 */
   bounds t1 > 390, t2 > t1, t2 < 720;

   /* 結果の出力 */
   output out=results p=predicted r=residual;
run;
```

まず、parms ステートメントにより Table 6 に示した初期パラメータ値を指定し、model ステートメントでは logratio と pred の関係でモデル式を構築した。pred の数式については、区間ごとに if else ステートメントを用いて応答変数と予測式を定義した。さらに、bounds ステートメントによって変化点のパラメータ範囲を設定した。

実行の過程で、パラメータの調整は計 11 回の反復を経て収束判定基準に達し、反復ごとの推定値を Table 7 に掲載した。

Table 7. Iteration Phase

| **Iteration** | **$a_1$** | **$b_1$** | **$b_2$** | **$b_3$** | **$t_1$** | **$t_2$** | **Sum of Squares** |
|---|---|---|---|---|---|---|---|
| 0 | -0.1000 | 0.000100 | -0.00100 | -0.00050 | 500.0 | 600.0 | 18.6247 |
| 1 | -0.0685 | 0.000030 | -0.00173 | -0.00086 | 526.0 | 685.0 | 11.7717 |
| 2 | -0.0428 | -0.00002 | -0.00400 | -0.00062 | 541.5 | 603.0 | 8.6361 |
| 3 | -0.0270 | -0.00006 | -0.00712 | -0.00180 | 555.2 | 624.7 | 1.5101 |
| 4 | 0.0134 | -0.00015 | -0.00794 | -0.00169 | 554.5 | 615.5 | 1.3612 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | -0.00415 | -0.00011 | -0.00873 | -0.00189 | 555.7 | 608.4 | 1.3385 |
| 6 | 0.0135 | -0.00015 | -0.00893 | -0.00175 | 557.3 | 611.1 | 1.3325 |
| 7 | 0.0151 | -0.00015 | -0.00929 | -0.00182 | 558.1 | 608.6 | 1.3302 |
| 8 | 0.0152 | -0.00015 | -0.00917 | -0.00179 | 557.9 | 609.5 | 1.3295 |
| 9 | 0.0163 | -0.00015 | -0.00916 | -0.00177 | 558.0 | 609.9 | 1.3294 |
| 10 | 0.0166 | -0.00015 | -0.00916 | -0.00176 | 558.0 | 610.0 | 1.3294 |
| 11 | 0.0163 | -0.00015 | -0.00915 | -0.00176 | 558.0 | 610.0 | 1.3294 |

反復回数が増加するにつれて、誤差項の平方和は減少し、モデルの適合度が高まる。モデルの適合性は Table 8 の分散分析表にて表すことができる。

Table 8. Analysis of Variance

| **Source** | **Degrees of Freedom** | **Sum of Squares** | **F value** | **p-value** |
|---|---|---|---|---|
| Model | 5 | 16.2248 | 524.80 | <0.0001 |
| Error | 215 | 1.3294 | | |
| Corrected Total | 220 | 17.5542 | | |

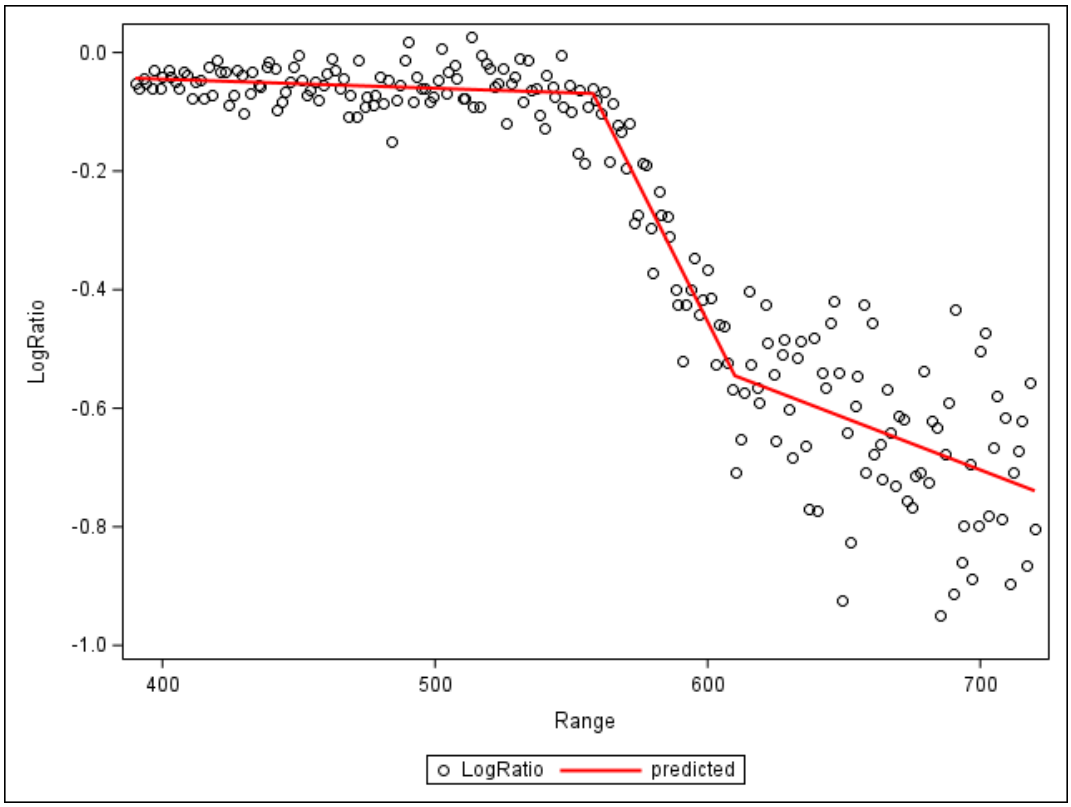
最終的に算出されたモデルのパラメータ推定値を Table 9 に示した。

Table 9. Parameter Estimates

| **Parameter Name** | **Parameter** | **Estimate** | **Standard Error** | **95% Confidence Limits** | |
|---|---|---|---|---|---|
| **第一区間の切片** | $a_1$ | 0.0163 | 0.0729 | -0.1273 | 0.1599 |
| **第一区間の傾き** | $b_1$ | -0.00015 | 0.000153 | -0.00046 | 0.000149 |
| **第二区間の傾き** | $b_2$ | -0.00915 | 0.000877 | -0.0109 | -0.00742 |
| **第三区間の傾き** | $b_3$ | -0.00176 | 0.000285 | -0.00233 | -0.00120 |
| **第一変化点** | $t_1$ | 558.0 | 3.3211 | 551.4 | 564.5 |
| **第二変化点** | $t_2$ | 610.0 | 4.3981 | 601.3 | 618.6 |

PROC NLIN で推定した区分線形回帰モデルを Figure 1 の散布図に加えたグラフは以下のとおりである。

Figure 2. モデルを加えた散布図




### 3.7　PROC NLIN を使った区分線形回帰モデルの解説

PROC NLIN による算出結果の回帰係数と変化点の値を式(3)に代入と以下になる。

$$LOGRATIO = \begin{cases} 0.0163 + (-0.00015) * RANGE & 390 < RANGE \leq 558.0 \\ 0.0163 + (-0.00015) * 558.0 + (-0.00915) * (RANGE - 558.0) & 558.0 < RANGE \leq 610.0 \\ 0.0163 + (-0.00015) * 558.0 + (-0.00015) * (610.0 - 558.0) + (-0.00176) * (RANGE - 610.0) & 610.0 < RANGE < 720 \end{cases} \quad (4)$$

最終的な結果は以下のモデルとなった。

$$LOGRATIO = \begin{cases} 0.0163 - 0.00015 * RANGE & 390 < RANGE \leq 558.0 \\ 5.0363 - 0.00915 * RANGE & 558.0 < RANGE \leq 610.0 \\ 0.9984 - 0.00176 * RANGE & 610.0 < RANGE < 720 \end{cases} \quad (5)$$

本分析では、LOGRATIO と RANGE の関係を区分線形モデルにより表現した。PROC NLIN を用いることで、非線形最小二乗法によるパラメータ推定が行われ、二つの変化点に対しても推定値および信頼区間が算出された。Table 9 の結果では、最初の変化点が 558.0 (95%CI: [551.4, 564.5])、2 つ目が 610.0 (95%CI: [601.3, 618.6])に存在することを示すものであり、それぞれの区間における回帰係数も同時に推定される。具体的には、RANGE が 390〜558.0 の区間（第一変化点まで）では LOGRATIO は RANGE の増加とともに緩やかに減少(傾き: -0.00015)し、558.0〜610.0 の区間（第一変化点から第二変化点まで）ではより急激に減少(傾き: -0.00915)する傾向が認められた。610.0〜720 の区間（第二変化点以降）では再び減少の傾きが小さくなっている(傾き: -0.00176)。以上より、RANGE の区分ごとに LOGRATIO の減少傾向が異なり、558.0 および 610.0 付近で関係性が変化することが明らかとなった。

PROC NLIN による Table 8 の分散分析表では、p 値が 0.0001 未満となり、区分線形回帰モデルが有意に LIDAR データに適合していることが示された。

# 4　PROC NLIN の限界

本論文で用いた SAS の PROC NLIN は、非線形回帰モデルの推定において有用なプロシジャである。一方で、その使用にはいくつかの技術的な制約が存在する。特に、初期パラメータ値の設定が推定結果や収束性に大きく影響する点、線形モデル用プロシジャと比べて回帰診断に利用できる情報が限られている点、そしてランダム効果を含む階層構造や集団間変動を明示的にモデル化することができない点が挙げられる。以下では、これらの主な限界について詳述する。

## 4.1　初期パラメータ設定

非線形回帰に使用される PROC NLIN では、各パラメータの初期値を指定する必要がある。推定は初期値から反復的に進められるため、初期値が不適切であるとモデルが収束しない場合や、最適なパラメータ値に到達できず、局所解で推定が終了する可能性がある。その結果、得られる推定結果の信頼性が損なわれる場合がある。したがって、初期値の選定は推定結果の妥当性や収束性に大きく影響するため、十分な検討が必要となる。

## 4.2　回帰診断情報が少ない

PROC NLIN には診断指標の出力が非常に限定的であるという重要な技術的制約が存在する。例えば、PROC REG や PROC GLM などの線形モデル用プロシジャでは、標準化残差・スチューデント化残差、レバレッジ値、Cook の距離、DFFITS、分散比などの詳細な回帰診断指標が自動的に算出されるが、PROC NLIN ではこれらの指標が出力されない。そのため、データ点ごとの影響度や外れ値・高レバレッジ観測値の定量的評価、本格的なモデル適合度検証が困難となる。

## 4.3　ランダム効果導入不可

PROC NLIN は固定効果のみを持つ非線形モデルの推定しか行うことができず、ランダム効果や階層構造を持つモデルには対応していない。そのため、被験者内の繰り返し測定やグループ化されたデータなど、データ間の相関や構造的なばらつきを考慮する必要がある場合には使用できない。ランダム効果を含む非線形混合効果モデルについては、PROC NLMIXED など、専用の解析手法を用いる必要がある。

# 5　最後に

本論文では、データ内に区分ごとの変化点が存在する場合のモデル化手法として、区分線形回帰（piecewise linear regression）を取り上げ、その実装方法および解析上の有用性について検討を行った。SAS の PROC NLIN を活用し、変化点の位置および各区間の回帰係数を非線形最小二乗法で同時に推定する手順を具体的な LIDAR データセット（2 つの変化点を持つモデル）を用いて示した。その結果、区分線形回帰モデルは区間ごとに異なる傾向を明確に捉えることができ、従来の単純な線形回帰では十分に表現できないデータの構造的な変化を把握するうえで有効であることが示された。

一方で、本研究で用いた PROC NLIN にはいくつかの限界も認められる。特に、推定の際に設定する初期パラメータ値が結果に大きく影響するため、適切な初期値の選択が解の収束や妥当な推定結果の取得に不可欠となる。また、回帰モデルの診断情報が限定的であり、外れ値やモデル適合度の詳細な評価が困難であること、さらにランダム効果や階層構造を持つモデルへの対応ができないなど、解析の柔軟性に制約がある。これらの課題を十分に認識した上で、データや解析目的に応じて適切な手法を選択することが、今後の研究において重要となる。

# 参考文献

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