数理解析研究所講究録
第1951巻 2015年 98-101

FindFitをExcel感覚で活用

日本大学・生物資源科学部　五十嵐　正夫 (Masao Igarashi)[1]
日本大学・生物資源科学部　陳　静 璇 (Chen Jing Xuan)
College of Bioresource Sciences,
Nihon University

1　はじめに

研究室[2]に持ち込まれるほとんどのデータは非線形性の強いデータである．非線形性の弱いデータの場合は，データを適当に小分けにし，小分けにしたデータ毎に線形回帰を実行し，データ全体の性質を調べる方法が多いようである．小分けにしたデータにExcelを適用し回帰直線を求め，データと直線の当てはまり具合を調べるにはR^2の値が利用されることが多い．

ところが最近の非線形ソルバー，例えばMathematicaの`FindFit`はデータを小分けにしなくても，適当な非線形モデルを与えると，その係数を容易に求めてくれる場合が多い．データ全体に最小2乗法を適用し連立非線形方程式を生成し，ソルバーが適当に初期値決めてNewton法を利用して数値解を導き出していると思われる．

このような数学ソフトウエアは，汎用的に作られている場合が多い．授業で利用する場合には，利用方法に一寸した工夫を行うと更に実用性が高まり，教育的効果の向上がみられる場合がある．この論文では学生が持ち込んだデータを基に，この汎用的ソフトウエアの実践的工夫法を紹介する．

2　データの種類と解析方法

学生，あるいは教員から持ち込まれるデータには次のようなものがある．

1. 細胞 (Cell) の増殖データ
2. 地域の植生 (NDVI) の変化データ [2]
3. 陸地の気温変化 (LST) のデータ
4. 海面の温度変化 (SST) のデータ
5. CO_2 濃度の変化のデータ

[1]E-mail igarashi.masao@nihon-u.ac.jp
[2]一般教養　数理情報研究室

6. 硝酸態窒素の除去率のデータ

ここでは細胞の増殖データにLogisticモデルとGompertzモデルを当てはめ`FindFit`によりそのモデルの係数を決める手順を示す.

1. データをプロットし，その形状から非線形曲線のモデルを決定する.
2. `FindFit`を利用し，その非線形モデルの係数を決める.
3. モデルの係数が出力されれば，データと曲線の算術的相対誤差の平方和平均(CR)を計算し，その値をデータと曲線の当てはまり具合の尺度とする.
4. モデルの係数が出力されない場合は，非線形モデルの係数の符号を逆にし`FindFit`を適用する．この「符号を逆にする」操作をここでは「ひと工夫」と呼んでいる[3].
5. 「ひと工夫」の結果，モデルの係数が出力されれば前と同様にデータと推定値の平方和平均(CR)を計算し，当てはまり具合を調べる.
6. 「ひと工夫」しても解が出力されない場合は，データの非線形性を弱める工夫をする．例えばNDVIデータは12か月の周期を持つと想定されるので，その周辺値をモデルの既知パラメータとして設定し，各パラメータ毎に当てはまり具合を計算し，最適なパラメータを決定する [3].

上の手法は，Mathematicaの組み込み関数`FindFit`本体には何の手も加えていないので，全くの初心者でも利用できる.

3 数式実験

細胞の増殖実験dataをLogistic曲線に当てはめるプログラムを次に示す.

```
data={10, 8, 24, 77, 87, 144, 178, 209}
f[t_]=K/(1+C*E^(r*t))
f[t]/. FindFit[data,f[t],{K,C,r},{t}]
```

変数はt，モデルは`f[t_]=K/(1+CE^(rt))`，見つけるべきパラメタ - はK, C, rである.

Mathematicを実行させると

$$f(t) = \frac{92.125}{1 + 102.236e^{-54.5473t}} \tag{1}$$

の解が出力される．モデル式からKは200程度の数値が期待されるが出力は92.125，eのべき係数rが-54.5473であることより，出力された解はデータに当てはまっていないことが分かる[4].

[3]数学的は符号を逆にしても式は等価である.

[4]eはMathematicaではE，乗算は$*$である．また，このような一見もっともらしい解は幻影解と呼ばれている

そこでモデル式の r の符号マイナスにして `f[t_]=K/(1+C*E^(-r*t))` とすると

$$f(t) = \frac{240.173}{1 + 61.2078e^{-0.745342t}} \tag{2}$$

の解が得られる．この曲線とデータを同一平面上にプロットすると，曲線とデータとよく一致することが分かる．式 (1) と (2) の基になる式は数学的には全く等価であるにもかかわらず，一方は幻影解，一方は合理的な解となっている．

このような現象を詳しく調べるために次の 4 つの例に関して，符号を変えながら数値実験を行った結果を次に示す．

モデル

$$f[t\_] = \frac{K}{1 + Ce^{rt}}, \quad f[t\_] = \frac{K}{1 - Ce^{rt}}, \quad f[t\_] = \frac{K}{1 + Ce^{-t}}, \quad f[t\_] = \frac{K}{1 - Ce^{-rt}} \tag{3}$$

数値例

例 1 : data={10, 8, 24, 77, 87, 144, 178, 209}

例 2 : data={20, 21, 56, 119, 127, 195, 234, 187}

例 3 : data={25, 23, 65, 143, 165, 197, 206, 187}

例 4 : data={30, 22, 76, 124, 124, 194, 250, 215}

実験結果

○は合理的な解が得られた場合，×は解が非合理な解がの場合を示す．この例では

表 1: Logistic 曲線：合理的な解○，合理的でない解×

	K=Max[data]				K は未知数			
係数の符号	例 1	例 2	例 3	例 4	例 1	例 2	例 3	例 4
$C,\ r$	×	×	×	×	×	×	×	×
$-C,\ r$	○	○	○	○	×	×	×	×
$C, -r$	○	○	○	○	×	○	○	○
$-C, -r$	×	×	×	×	○	○	○	○

Logistic 曲線にモデルを設定したが，同じ例題に対して Gompertz 曲線

$$L[t\_] = Ke^{Ce^{rt}}, \quad L[t\_] = Ke^{-Ce^{rt}}, \quad L[t\_] = Ke^{Ce^{-rt}}, L[t\_] = Ke^{-Ce^{-rt}} \tag{4}$$

を設定すると次のような結果が得られる．

表 2: Gompertz 曲線:合理的な解○，合理的でない解×

	K=Max[data]				K は未知数			
係数の符号	例 1	例 2	例 3	例 4	例 1	例 2	例 3	例 4
$C,\ r$	○	○	○	○	×	×	×	×
$-C,\ r$	×	○	○	○	×	×	×	×
$C, -r$	×	○	○	○	○	○	○	○
$-C, -r$	×	○	○	○	○	○	○	○

4 まとめ

Newton 法は初期値や反復停止則 [2] をうまく選ばないと数値解が得られない場合が多い [1]．計算機的にはオーバーフローやアンダーフローを起こしやすいからである．この理由を学生に FindFit を使う前に理解させることはなかなか困難である．

そこで本論文ではモデルの係数，例えば C を $-C$ に置き換え，また r を $-r$ に置き換えることにより FindFit が行っているであろう初期値の反転を行う数値実験を行ってみた．この操作はソルバー FindFit の本体に何の手も加えていないという意味で Excel 的な操作と考えられる．「できあがりのお総菜にひと手間かけたレシピ」に過ぎないが，このような手法を知っているか，いないかにより数学ソフトウエアに対する親密度や理解度は大きく変わり「解が得られない徒労感」から，少しは解放されると期待される．

あるデータに Logistic 曲線と Gompertz 曲線の両方を適用し，どちらの曲線がよりデータによく当てはまっているか，を知りたい場合がある．通常は線形の場合と同じように相関係数 R^2 が使われる場合が多いが，ここでは次の相対誤差の平方和平均 CR を利用した．この値が小さければ，当てはまり具合はよいことになる [3]．

$$\mathrm{CR} = \frac{\sum_{k=1}^{m} \dfrac{|\widehat{y_k} - y_k|}{\max\left(\widehat{|y_k|}, |y_k|\right)}}{m} \tag{5}$$

ここで m はデータの大きさ，$\{\widehat{y_k}\}$ 観測値，$\{y_k\}$ は回帰曲線による推定値である，各相対誤差を取るとき分母で max を取るのはゼロ割りを避けるためである．

参考文献

[1] 山本哲朗, 数値解析入門, サイエンス社, 1976.

[2] Masao Igarashi, A termination Criterion for iterative methods used to find the zeros of polynomials, Math Comp. 1984.

[3] Masao Igarashi et al., A fitness criterion on finding nonlinear curve for the NDVI Data, Information, 2010.