林学のための教育用ＢＡＳＩＣ プログラム（その１）

―― 多項式のあてはめ ――

統 数 研　仁 木 直 人

私はここ数年東京農工大学においてプログラム言語とデータ解析の講義をしています。以前は，FORTRAN のみを教えていたのですが，昨年から言語をBASICに半分切り換えました。最近のマイクロ・コンピュータの普及ぶりからすると，これからの学生にはBASICの知識の方がむしろ不可欠となるような気がします。

この講義の中で例題として作ったプログラムのいくつかを，これから数回に亘って，私の講義ノートの形で連載していきたいと思います。例題は，学生が実際に実習などで経験する（した）ものや，これから卒論のデータ整理などに使えるものをなどを選ぶようにしていますから，皆さんのお役にも少しは立つのではないかと思います。

なお，ここに載せますプログラムは，ほとんどの計算機についてそのままで実行できるよう，なるべく共通語で書いたつもりです。しかし，いわゆる basic BASIC に従って書きますと，いたずらにプログラムが長くなりますから，最近のほとんどのBASICには備わっている３つの機能を追加してあります。

第１は，代入文におけるＬＥＴの省略。第２は，コロン（：）で区切ることによる複文。そして，ＩＦ～ＴＨＥＮのあとに複文が書けること。以上の３点です。

グラフック機能など，どうしても計算機ごとに異なる命令を使わなければならない部分はサブルーチンにしておきます。必要ならばお持ちの計算機用に作り直して下さい。それが大変でしたら，先頭の行をRETURN 文にして，残りの部分は消して下さい。

多項式のあてはめ

①　考え方と解法

変数ｘとｙについて，ｎ組のデータ x_1 ，x_2 ，……，x_n ；y_1 ，y_2 ，……y_n が与えられている。そしてｘとｙの間には，

$$y \fallingdotseq C_0 + C_1 x + C_2 x^2 + \cdots\cdots + C_m x^m$$

のような関係が成立つと考えられる。このような場合，データからどのように C_0 ，C_1 ，……，C_m を決めるか。

最小二乗法に従って考えてみる。最小二乗法では，測定値と推定値の差を二乗したものの和を考え，この和を最も小さくするように係数 C_0 ，C_1 ，……，C_m を決める。すなわち，

$$Q = \sum_{i=1}^{n} \{ y_i - (C_0 + C_1 x_i + C_2 x_i^2 + \cdots\cdots + C_m x_i^m) \}^2$$

を最小とする C_0 ，C_1 ，……，C_m を求める。この解は，

$$\frac{\partial Q}{\partial C_j} = 0 \qquad (j=1, 2, \cdots\cdots, m)$$

より（少し計算すれば解るように），

$$\begin{pmatrix} \Sigma 1 & \Sigma x & \Sigma x & \cdots\cdots\cdots \Sigma x^m \\ \Sigma x & \Sigma x^2 & \Sigma x^3 & \cdots\cdots\cdots \Sigma x^{m+1} \\ \cdots\cdots & & & \\ \Sigma x^m & \Sigma x^{m+1} & \Sigma x^{m+2} & \cdots\cdots \Sigma x^{2m} \end{pmatrix} \begin{pmatrix} C_0 \\ C_1 \\ \cdots\cdots \\ C_m \end{pmatrix} = \begin{pmatrix} \Sigma y \\ \Sigma xy \\ \cdots\cdots \\ \Sigma x^m y \end{pmatrix}$$

なる連立一次方程式を解けば得られる（x_i，y_iの添字は省略した。和はi＝1からnまでについてとる）。

次に最尤法に従って考えてみる。今度は変数x，y間にかなり具体的なモデルを導入する。

モデル：yは平均$\mu(x) = C_0 + C_1 x + \cdots\cdots C_m x^m$，分散$\sigma^2$の正規分布に従う。

これは，測定値と推定値の差は平均0分散σ^2の正規分布に従う，といっても同じ。観測値y_1，y_2，……，y_nが得られる確率，すなわち尤度は，

$$P_r\{y_1, y_2, \cdots, y_n\} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\{y_1 - \mu(x_i)\}^2}{2\sigma^2}\right]$$

で与えられる。計算にはこの対数をとった対数尤度の方が便利で，

$$L = -n\ell n\sqrt{2\pi} - n\ell n\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \{y_i - \mu(x_i)\}^2$$

$$= -n\ell n\sqrt{2\pi} - n\ell n\sigma - Q/(2\sigma^2)$$

ゆえに，尤度（または対数尤度）を最大にするC_0，C_1，…，C_mは

$$\text{より} \quad \frac{\partial L}{\partial C_j} = 0 \qquad (j=1, 2, \cdots, m)$$

$$\frac{\partial Q}{\partial C_j} = 0 \qquad (j=1, 2, \cdots, m)$$

が得られるから，最小二乗法の解と一致する。逆に考えれば，最小二乗法の裏には「測定値と推定値の差が平均0の正規分布に従う」という暗黙の仮定があるとも言える。

なお，

$$\frac{\partial L}{\partial \sigma} = 0$$

より，Lを最小にするσは，最小となったQの値をminQで表わすと，

$$\sigma^2 = \min Q/n$$

で求められる。これは測定値と推定値の差の分散（残差分散）に一致する。

また，相関係数ρは元のyの分散

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_y) \qquad \left(\mu_y = \frac{1}{n} \sum_{i=1}^{n} y_i \ : \ \text{yの平均}\right)$$

と残差分散$\sigma_e^2 = \min Q/n$を用いて，

$$\rho^2 = 1 - \sigma_e^2 / \sigma_y^2$$

で定義されるから，

$$\rho = \sqrt{1 - \sigma_e^2 / \sigma_y^2}$$

で，符号が負のものを考える必要はない。

② 多項式の次数

あてはめる多項式の次数を上げれば，必ず相関係数が大きくなる。相関係数は何でもかでも大きい方が良いのだ，という考えは誤まっている。

極端な話が，$x_i \neq x_j$（$i \neq j$）の場合，n－1次の多項式を用いれば相関係数を常に1にできる。しかし，この多項式のグラフはメチャクチャな形をしているはずだ。

これほどでなくても，必要以上に次数を上げると，データの小さな特徴に引きずられて，返ってデータの本質的な構造をとらえることができなくなる。この辺の様子は後に例題で見てみる。

相関係数がさほど違わなければ，低い次数の式を使うのがよい。しかし，あまりに低い次数の式を採用するのでは，データの大きな構造をとらえることすらできなくなる恐れが強い（相関係数がかなり小さくなる）。

次数の決定に一番良いのは，（x_i，y_i）の点をグラフ用紙にプロットし，データの全体像を把握することだろう。異常データ（測定・転記などのミス，環境の急激な変化など）も見いだすこともある。このときＡＩＣの値を計算しておくと判断の大きな助けとなる。

ＡＩＣ（Akaike's Information Criterion）は，最尤推定した分布モデルが“真の分布”をどの程度近似しているかを測るための量である（詳しく言えば，モデルと真の分布の違いを，Kullbuck－Leibler情報量で測った値の，定数部分を除いた部分の推定値）。

ＡＩＣ＝－2×（モデルを仮定したときの最大対数尤度）
＋2×（モデルに含まれる自由なパラメータの数）

m次式をあてはめた場合，最尤法のところで見たように

$$\begin{aligned} AIC &= -2(\max L) + 2(m+2) \\ &= n(2\ell n\sqrt{2\pi} + 2\ell n\sigma_e + 1) + 2(m+2) \\ &= n\ell n\sigma_e^2 + 2(m+2) + (\text{定数}) \end{aligned}$$

となるから，残差分散σ_e^2を求めればＡＩＣは簡単に計算できる（定数項は比較のためには不要であるので除いてよい）。

ＡＩＣが小さい値を持つモデルほど近似の程度が良いと考えられる。いくつかの次数mについてＡＩＣを算出し，最もＡＩＣを小さくするmを最適な次数として採用する。

ＡＩＣ自身が推定値であるから，データ数nはある程度大きくなくてはならない（一応の目安としては10（m＋2）以上，できれば20（m＋2）以上）。データが少な目のときは，ＡＩＣの小さな差（小数点以下）にはこだわらず，低めの次数を選択するのが賢明である。

③ プログラム

通常のデータで4次式以上をあてはめる必要のあるものはごく稀である。そこで0次式（無相関）から4次式までのあてはめのできるプログラムとした。0次式（平均で推定）は，xとyの間に相関

関係があるかどうか，をＡＩＣにより判定する際に用いる。

データは，始めにデータ数ｎを置き，以下x_1，y_1，x_2，y_2，…，x_n，y_nという順に与える。

まず，ｘ，ｙの平均・標準偏差を計算し，表示する（１３０～２４０行）。次に，４次式のあてはめの計算に必要な連立一次方程式の係数行列および定数ベクトルを求めてしまう（２５０～４３０行）。このときｘの８乗までに関する加算が行なわれるので，計算精度を確保するため，ｘの値を平均０分散１と正規化して使っている（２８０行）。

キーボードから何次式をあてはめるかを入力する（５００～５５０行）。このとき与えるデータの整数部を次数ｍとして使い，小数部はオプション選択のために用いる。オプションは，「ｙとｙの推定値を表示する（０.５）」「点群（x_i，y_i）とあてはめた多項式をグラフ表示する（０.９）」のふたつを用意してある。

与えられたｍに従って係数行列および定数ベクトルの必要な部分を取出し（５７０～６００行），連立一次方程式を解く（６２０行）。解法は消去法による（１０００～１２３０行）。ここで求まったC_0，C_1，…，C_nは正規化したｘについてであるから，元のｘについての値に直してから表示する(６３０～７８０行)。

次にｙの推定値に関して，その平均（ｙの平均に一致するはずであるが念のため），標準偏差およびｙとの共分散を計算し，相関係数を求める（７９０～９３０行）。このとき，オプションの指定があれば，ｙとｙの推定値を次々に表示する。

残差分散を相関係数から逆に求め，ＡＩＣを計算し，相関係数とともに表示する(９４０～９６０行)。

グラフ表示オプションがあれば１５００行以下のサブルーチンを実行する。画面を切換える前に"PRESS ANY KEY"と表示する。任意のキーを押せば，画面を一度消去したのち，始めに（x_i，y_i）の点群を水色で表示し，再度押せば，今度は多項式を黄色で表示する。見終ったらもう一度任意のキーを押す。

与えられた次数ｍに関する処理が終わると，別の次数（とオプション）の入力要求を行い，以下同様の処理をくり返す。

テスト・データ（５０～７０行）について，計算を行なってみる。０～４次式のあてはめの結果，２次式をあてはめたときがＡＩＣ最少となった。注意すべきことは，２次式を用いたときの相関係数が，１次式を用いたときの相関係数と比べるとある程度の差があるのに対し，３次式以上をあてはめても相関係数がほとんど改善されない，という点である。同じ程度のあてはまりの良さであれば，次数の低い方が良い，という「ケチの原理」が働いている。実際，４次式のあてはめでの解を見れば，かなり絶対値の大きな値をもつ項を足し引きすることで帳尻を合わせている様子が読み取れる。このような場合の解は不安定なことが多い。

グラフィック機能を用いれば，別の面を観察できる。データの存在する範囲では，２～４次式のどれをとってもほとんど見分けがつかないほどであるが，範囲外の様子は著く違い，とくに４次式では不自然さが目につく。ｘの最大・最小値付近のデータのちょっとした変化で，各項の係数が大きく変化することを確かめるとよい。

```
        X             Y

MEAN    1.254         32.416
S.D.    .719155       9.64664

DEGREE OF POLYNOMIAL . OPTION ? 0

    CONST.

 0  32.416


R = 0          AIC = 230.661

DEGREE OF POLYNOMIAL . OPTION ? 1

    CONST.

 0  16.3105
 1  12.8433


R = .957463    AIC = 108.374

DEGREE OF POLYNOMIAL . OPTION ? 2

    CONST.

 0  20.1944
 1  3.76464
 2  3.58941


R = .972437    AIC = 89.061

DEGREE OF POLYNOMIAL . OPTION ? 3

    CONST.

 0  21.0324
 1  2.7504E-03
 2  7.26655
 3 -.968476


R = .972842    AIC = 90.3305

DEGREE OF POLYNOMIAL . OPTION ? 4

    CONST.

 0  22.1716
 1 -8.00631
 2  21.095
 3 -9.41569
 4  1.67329


R = .973302    AIC = 91.4874

DEGREE OF POLYNOMIAL . OPTION ? 2.5

    CONST.

 0  20.1944
 1  3.76464
 2  3.58941
```

	Y	ESTIMATED Y
1	21.8	20.6067
2	22.2	20.6067
3	19.5	20.6067
4	22.2	21.0909
5	19.7	21.3599
6	22.3	21.6468
7	23.1	21.9517
8	19.9	22.2745
9	25.2	22.6153
10	19.1	22.6153
11	25.9	22.974
12	26.2	23.3507
13	19.8	23.7453
14	23.6	24.5884
15	26	24.5884
16	24.1	25.5033
17	25.4	25.9877
18	23.9	25.9877
19	23.8	27.0102
20	28.3	27.0102
21	29.7	28.1046
22	27.1	28.1046
23	22.9	28.6787
24	31.7	29.8807
25	30.4	30.5086
26	34.5	30.5086
27	34.6	31.8183
28	29.6	31.8183
29	33.5	33.1998
30	33.7	33.1998
31	33.6	33.9175
32	40.4	34.6531
33	37.4	36.1782
34	33	36.1782
35	37.1	36.9676
36	39.7	37.775
37	39.5	38.6004
38	41.3	40.3049
39	42.8	41.1841
40	39.2	42.0813
41	42.8	42.9964
42	41.6	43.9294
43	41	44.8804
44	45.1	44.8804
45	46.6	46.8362
46	46.8	47.841
47	51	48.8638
48	51.3	48.8638
49	50.9	50.9631
50	50	50.9631
MEAN	32.416	32.416
S.D.	9.64664	9.38081

R = .972437 AIC = 89.061

```
10 REM << POLYNOMIAL CURVE FITTING >>
20 REM   DEGREE OF POLYNOMIAL
22 REM     IS 0, 1, 2, 3 OR 4
24 REM   OPTION
26 REM    .5: PRINT Y AND ESTMTD. Y
28 REM    .9: PLOT DATA AND POLYNOMIAL
30 REM   CODED BY N.NIKI
50 DATA 50
52 DATA  .1  , 21.8 ,  .1  , 22.2 ,  .1  , 19.5 ,  .2  , 22.2 ,  .25 , 19.7
54 DATA  .3  , 22.3 ,  .35 , 23.1 ,  .4  , 19.9 ,  .45 , 25.2 ,  .45 , 19.1
56 DATA  .5  , 25.9 ,  .55 , 26.2 ,  .6  , 19.8 ,  .7  , 23.6 ,  .7  , 26
58 DATA  .8  , 24.1 ,  .85 , 25.4 ,  .85 , 23.9 ,  .95 , 23.8 ,  .95 , 28.3
60 DATA 1.05 , 29.7 , 1.05 , 27.1 , 1.1  , 22.9 , 1.2  , 31.7 , 1.25 , 30.4
62 DATA 1.25 , 34.5 , 1.35 , 34.6 , 1.35 , 29.6 , 1.45 , 33.5 , 1.45 , 33.7
64 DATA 1.5  , 33.6 , 1.55 , 40.4 , 1.65 , 37.4 , 1.65 , 33   , 1.7  , 37.1
66 DATA 1.75 , 39.7 , 1.8  , 39.5 , 1.9  , 41.3 , 1.95 , 42.8 , 2    , 39.2
68 DATA 2.05 , 42.8 , 2.1  , 41.6 , 2.15 , 41   , 2.15 , 45.1 , 2.25 , 46.6
70 DATA 2.3  , 46.8 , 2.35 , 51   , 2.35 , 51.3 , 2.45 , 50.9 , 2.45 , 50
100 DIM A(4,5),B(4,5),C(5)
110 DEF FNY(X)=C(0)+(C(1)+(C(2)+(C(3)+C(4)*X)*X)*X)*X
120 READ N
130 X1=0: X2=0: Y1=0: Y2=0
140 FOR I=1 TO N
150  READ X,Y
160  X1=X1+X: Y1=Y1+Y
170  X2=X2+X*X: Y2=Y2+Y*Y
180 NEXT I
190 X1=X1/N: Y1=Y1/N
200 X2=SQR(X2/N-X1*X1): Y4=Y2/N-Y1*Y1: Y2=SQR(Y4)
210 PRINT TAB(8);"X";TAB(22);"Y"
220 PRINT
230 PRINT "MEAN";TAB(6);X1;TAB(20);Y1
240 PRINT "S.D.";TAB(6);X2;TAB(20);Y2
250 M=4: M3=3: M5=5
260 RESTORE: READ N
270 FOR I=1 TO N
280  READ X,Y: X=(X-X1)/X2
290  X0=1: Y0=Y
300  FOR K=0 TO M
310   B(0,K)=B(0,K)+X0
320   B(K,M5)=B(K,M5)+Y0
330   X0=X0*X: Y0=Y0*X
340  NEXT K
350  FOR J=1 TO M
360   B(J,M)=B(J,M)+X0: X0=X0*X
370  NEXT J
380 NEXT I
390 FOR J=0 TO M3
400  FOR K=0 TO M3
410   B(J+1,K)=B(J,K+1)
420  NEXT K
430 NEXT J
500 REM < DEGREE OF POLYNOMIAL >
510 PRINT: PRINT "DEGREE OF POLYNOMIAL . OPTION";
520 INPUT M0
530 IF M0>=0 AND M0<5 THEN 550
540 PRINT "?? DEGREE MUST BE 0,1,2,3 OR 4.": GOTO 510
550 M=INT(M0): D=M0-M: M1=M+1
560 IF M=0 THEN C(0)=Y1: GOTO 670
570 FOR J=0 TO M
580  FOR K=0 TO M: A(J,K)=B(J,K): NEXT K
590  A(J,M1)=B(J,M5)
600 NEXT J
610 REM < SOLVE LINEAR EQUATION >
620 GOSUB 1000
630 REM < INVERSE TRANSFORMATION >
```

```
640 FOR J=0 TO M
650  C(J)=A(J,M1)/X2^J
660 NEXT J
670 FOR J=M1 TO M5: C(J)=0: NEXT J
680 C(0)=C(0)-(C(1)-(C(2)-(C(3)-C(4)*X1)*X1)*X1)*X1
690 C(1)=C(1)-(2*C(2)-(3*C(3)-4*C(4)*X1)*X1)*X1
700 C(2)=C(2)-(3*C(3)-6*C(4)*X1)*X1
710 C(3)=C(3)-4*C(4)*X1
720 REM < PRINT CONSTANTS >
730 PRINT
740 PRINT "     CONST.": PRINT
750 FOR J=0 TO M
760  PRINT J;C(J)
770 NEXT J
780 PRINT
790 REM < CORRELATION >
800 IF M=0 THEN R=0: S9=Y4: GOTO 950
810 RESTORE: READ N
820 S1=0: S2=0: S3=0
830 IF D=.5 THEN PRINT TAB(8);"Y";TAB(20);"ESTIMATED Y": PRINT
840 FOR I=1 TO N
850  READ X,Y: Y0=FNY(X)
860  IF D=.5 THEN PRINT I;TAB(6);Y;TAB(20);Y0
870  S1=S1+Y0: S2=S2+Y0*Y0
880  S3=S3+Y*Y0
890 NEXT I
900 S1=S1/N: S2=SQR(S2/N-S1*S1)
910 R=(S3/N-S1*Y1)/Y2/S2
920 IF D=.5 THEN PRINT "MEAN";TAB(6);Y1;TAB(20);S1
930 IF D=.5 THEN PRINT "S.D.";TAB(6);Y2;TAB(20);S2
940 S9=(1-R*R)*Y4
950 A1=N*LOG(S9)+2*(M+2)
960 PRINT: PRINT "R ="; R, "AIC ="; A1
970 IF D>.89 THEN GOSUB 1500
980 GOTO 500
1000 REM < GAUSS-JORDAN METHOD >
1010 M1=M+1
1020 FOR I=0 TO M
1030  P=0: I1=I+1
1040  FOR J=I TO M
1050   P0=ABS(A(J,I))
1060   IF P<P0 THEN P=P0: J0=J
1070  NEXT J
1080  IF P<1E-06 THEN 1220
1090  P=A(J0,I)
1100  FOR K=I TO M1
1110   W=A(J0,K): A(J0,K)=A(I,K)
1120   A(I,K)=W/P
1130  NEXT K
1140  FOR J=0 TO M
1150   IF J=I THEN 1200
1160   P=A(J,I)
1170   FOR K=I1 TO M1
1180    A(J,K)=A(J,K)-P*A(I,K)
1190   NEXT K
1200  NEXT J
1210 NEXT I: RETURN
1220 PRINT "MATRIX IS NOT NORMAL"
1230 STOP
1500 REM < PLOT OPTION >
1510 REM   YOU CAN APPLY THIS SUBROUTINE IF YOUR MACHINE IS <PC-8001>,
1520 REM   IF NOT PLEASE PREPARE THE SBSTITE COMPATIBLE WITH YOURS.
1530 PRINT: PRINT "  PRESS ANY KEY."
1540 A$=INKEY$: IF A$="" THEN 1540
1550 CONSOLE 0,25,0,1: WIDTH 80,25
1560 COLOR 5: PRINT CHR$(12)
```

```
1570 RESTORE: READ N
1580 FOR I=1 TO N
1590  READ X,Y
1600  X0=(X-X1)/X2*25+80: Y=50-(Y-Y1)/Y2*15
1610  PSET(X0,Y)
1620 NEXT I
1630 COLOR 6
1640 A$=INKEY$:IF A$="" THEN 1640
1650 FOR X=X1-3*X2 TO X1+3*X2 STEP X2/26
1660 X0=(X-X1)/X2*25+80: Y0=50-(FNY(X)-Y1)/Y2*15
1670 IF Y0<0 OR Y0>99 THEN 1690
1680 PSET(X0,Y0)
1690 NEXT X
1700 A$=INKEY$: IF A$="" THEN 1700
1710 COLOR 7: WIDTH 40,20
1720 RETURN
```