岡本安晴　2001.5

重回帰モデル

N組のデータ（y_i、x_{i1}、x_{i2}、…、x_{ip}）；i＝1、…、N、において、y_iを、次の1次式で表わすことを考えます。

$$y_i \approx b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

記号 ≈ は、近似式であることを表わします。

上式には定数項が明記されていませんが、これは、例えば、

$$x_{ip} = 1$$

とすることによって、定数項をもつ場合も含まれていることがわかります。

近似式の、データ全体についての精度を、誤差の2乗和Q_eで表わします。すなわち、

$$Q_e = \sum_{i=1}^{N} \{y_i - (b_1 x_{i1} + \cdots + b_p x_{ip})\}^2$$

Q_eを最小にするb_jを求めるために、Q_eのb_jに関する偏導関数を求めます。

$$\frac{\partial Q_e}{\partial b_j} = \sum_{i=1}^{N} 2\{y_i - (b_1 x_{i1} + \cdots + b_p x_{ip})\}(-x_{ij})$$

$$= -2\sum_{i=1}^{N} y_i x_{ij} + 2\sum_{i=1}^{N} (b_1 x_{i1} x_{ij} + \cdots + b_p x_{ip} x_{ij})$$

上の偏導関数を0とおくと、次式が得られます。

$$\sum_{i=1}^{N} y_i x_{ij} = \sum_{i=1}^{N} (b_1 x_{i1} x_{ij} + \cdots + b_p x_{ip} x_{ij})$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{p} x_{ij} x_{ik} b_k \qquad (1)$$

上式を行列を使った式で表わすために、

$$y = [y_1, \cdots, y_N]^t$$

$$X = \begin{bmatrix} x_{11} \cdots x_{1p} \\ x_{21} \cdots x_{2p} \\ \vdots \\ x_{N1} \cdots x_{Np} \end{bmatrix}$$

$$b = [b_1, \cdots, b_p]^t$$

とおきます。

なお、ベクトルあるいは行列の右肩にｔまたは’を付けたものは、転置行列を表わします。

このとき、式（1）は次のように表わせます。

$$X'y = X'Xb \qquad (2)$$

$X'X$ が正則である、すなわち逆行列$(X'X)^{-1}$が存在するとき、

$$b = (X'X)^{-1}X'y$$

となります。

なお、最初の１次式を、行列を使って表わすと、

$$y \approx Xb$$

となります。

いま、Ｘの第ｐ列目がすべて１であるとします（cf. 佐和, 1979）。

$$X = [X_1, \mathbf{1}_N]$$

とＸを分割して表わします。1_N は、１をＮ個並べた列ベクトルを表わします。

Ｘに合わせて、ｂも

$$b = \begin{bmatrix} b_0 \\ b_p \end{bmatrix}$$

と分割します。b_0は（ｐ－１）次元列ベクトル、b_pはｂのｐ番目の要素（スカラー）です。

Ｘが上の形のときは、yの成分y_iについてみると

$$y_i \approx b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1} + b_p$$

となっています。これは、１次式が定数項b_pをもつ形です。

このとき、式（4.2）は次のようになります。

$$[X_1, 1_N]' y = [X_1, 1_N]'[X_1, 1_N] \begin{bmatrix} b_0 \\ b_p \end{bmatrix}$$

$$= \begin{bmatrix} X_1'X_1 & X_1'\mathbf{1}_N \\ \mathbf{1}_N' X_1 & \mathbf{1}_N'\mathbf{1}_N \end{bmatrix} \begin{bmatrix} b_0 \\ b_p \end{bmatrix}$$

すなわち、

$$\begin{cases} X_1' y = X_1'X_1 b_0 + X_1' 1_N b_p \\ \mathbf{1}_N' y = \mathbf{1}_N' X_1 b_0 + \mathbf{1}_N'\mathbf{1}_N b_p \end{cases} \qquad (3)$$

また、

$$\mathbf{1}_N' Xb = \mathbf{1}_N'[X_1, \mathbf{1}_N] \begin{bmatrix} b_0 \\ b_p \end{bmatrix}$$

$$= \mathbf{1}_N' (X_1 b_0 + \mathbf{1}_N b_p)$$

$$= \mathbf{1}_N' y \qquad \text{（式（3）より）}$$

すなわち、ｙの予測式

$$\hat{y} = Xb$$

の平均$m_{\hat{y}}$が、ｙの平均m_yに等しい、という次式が成り立ちます。

$$m_{\hat{y}} = \frac{1}{N} 1'_N Xb = \frac{1}{N} 1'_N y = m_y \qquad (4)$$

いま、ｙの総変動分（平方和）

$$Q_y = (y - m_y 1_N)'(y - m_y 1_N)$$

を、次のように分解します。

$$Q_y = (y - m_y 1_N)'(y - m_y 1_N)$$

$$= (y - Xb + Xb - m_y 1_N)'(y - Xb + Xb - m_y 1_N)$$

$$= (y - Xb)'(y - Xb) + (y - Xb)'(Xb - m_y 1_N)$$

$$+ (Xb - m_y 1_N)'(y - Xb) + (Xb - m_y 1_N)'(Xb - m_y 1_N)$$

ここで、

$$(y - Xb)'(Xb - m_y 1_N) = y'Xb - y'm_y 1_N - b'X'Xb + b'X'm_y 1_N$$

上式に、

$$b = (X'X)^{-1}X'y$$

を代入すると、

$$\text{上式} = y'Xb - y'm_y 1_N - y'X(X'X)^{-1}X'Xb + b'X'm_y 1_N$$

$$= -y'm_y 1_N + b'X'm_y 1_N$$

$$= -y'm_y 1_N + m_y(1'_N\ Xb)'$$

$$= -y'm_y 1_N + m_y(1'_N\ y)' \qquad \text{（式（4）より）}$$

$$= 0$$

同様にして、

$$(Xb - m_y 1_N)'(y - Xb) = 0$$

となります。

したがって、Q_y は次のように分解できます。

$$Q_y = (y - Xb)'(y - Xb) + (Xb - m_y 1_N)'(Xb - m_y 1_N)$$

$$= (y - Xb)'(y - Xb) + (Xb - m_{\hat{y}} 1_N)'(Xb - m_{\hat{y}} 1_N)$$

すなわち、総変動平方和

$$Q_y = (y - m_y 1_N)'(y - m_y 1_N)$$

は、残差平方和

$$Q_e = (y - Xb)'(y - Xb) = (y - \hat{y})'(y - \hat{y})$$

と、1次式$\hat{y} = Xb$で説明される変動平方和

$$Q_{\hat{y}} = (Xb - m_{\hat{y}}1_N)'(Xb - m_{\hat{y}}1_N) = (\hat{y} - m_{\hat{y}}1_N)'(\hat{y} - m_{\hat{y}}1_N)$$

の和に分解できます。

$$Q_y = Q_e + Q_{\hat{y}}$$

$Q_{\hat{y}}$とQ_yの比

$$R^2 = \frac{Q_{\hat{y}}}{Q_y} = 1 - \frac{Q_e}{Q_y}$$

は決定係数と呼ばれているもので、その正の平方根Rは重相関係数と呼ばれています（佐和、1979）。

R^2が1に近いほど、$Q_{\hat{y}}$はQ_yに近い、すなわち、誤差Q_eがQ_yに比べて相対的に小さい値であることを表わします。

プログラムPMultiIndVar.dprは、上の重回帰モデルによる分析を行うものです。

このプログラムのフォームは、プログラミング時には図1のように用意されています。

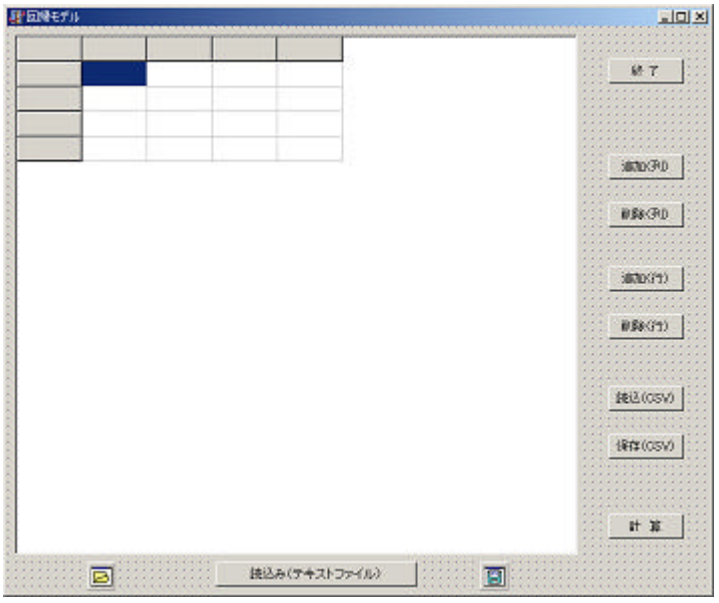


図1　プログラミング時のフォーム

プログラムを実行すると、図2のようなフォームになります。

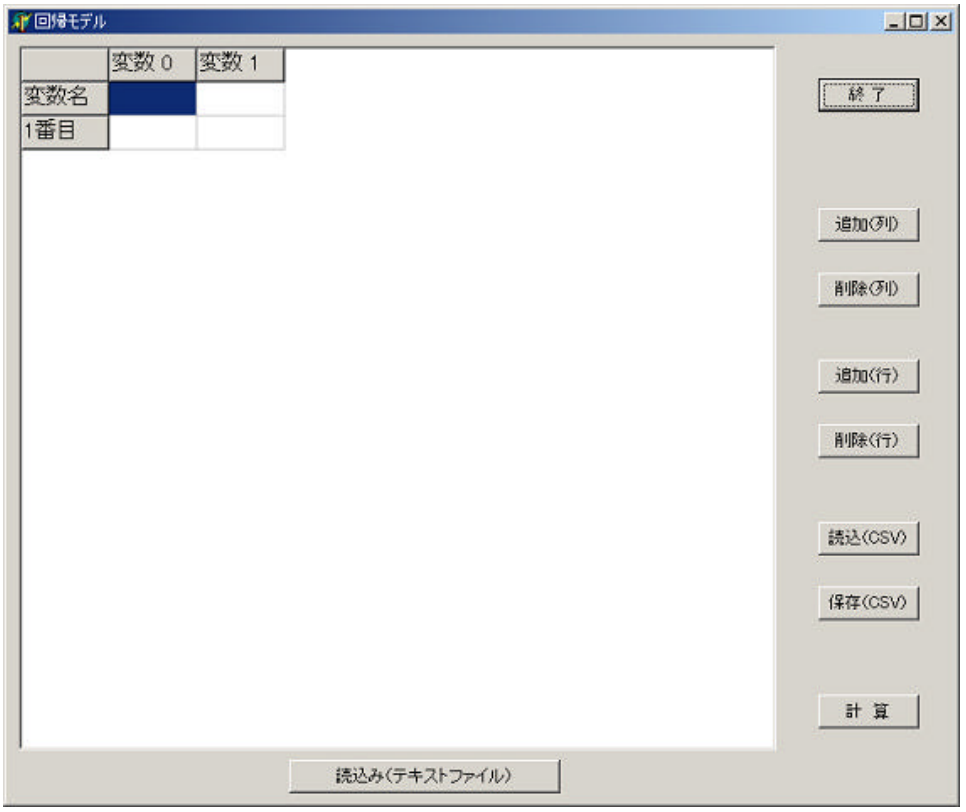


図２　実行開始時のフォーム

「追加」「削除」ボタンをクリックして、行および列を必要な数に調整します。変数０の欄に y_i、変数１の欄に x_{i1}、・・・、変数 p の欄に x_{ip} を書きます。i 番目のデータ、y_i、x_{i1}、・・・、x_{ip} は「 i 番目」の行に書きます。「変数名」の行には変数のラベルを書きますが、変数名は空白でもかまいません。

「読込み（テキストファイル）」ボタンをクリックすると、リスト１のような形式で用意されたテキストファイルデータを読み込むことができます。

リスト１　入力データ例（テキストファイル）

```
  -100
    2
国語
英語
切片
          72        66        1
          41        46        1
          93        87        1
          82        82        1
          41        42        1
          64        72        1
          27        31        1
          68        69        1
          79        84        1
          67        68        1
          57        50        1
          90        90        1
          87        85        1
          29        29        1
          75        76        1
          66        69        1
          78        84        1
          31        35        1
-1000
```

上のデータは、次式

$$\text{国語} = a \times \text{英語} + b + \text{誤差項}$$

を、

$$y_i = \text{国語}、\quad x_{i1} = \text{英語}、\quad x_{i2} = 1$$

とおいて、

$$y_i = \text{国語} = [\text{英語}, 1]\begin{bmatrix} a \\ b \end{bmatrix} + \text{誤差項} = [x_{i1}, x_{i2}]\begin{bmatrix} a \\ b \end{bmatrix} + \text{誤差項}$$

の形として、データ値

国語の点数、　英語の点数、　定数１

を並べたものです。

すなわち、リスト１の入力データは次の形式で用意します。

まず、第 1 行目にデータ読込み終了の判定に用いる基準値を書きます。行の先頭の数値がこの１行目の値より小さいところで読込みを終了します。リスト１では、１行目に－100 が書かれていて、最後の－1000 がデータの終わりを示しています。２行目には、独立変数

x_{i1}、・・・、x_{ip}の数pを書きます。リスト１では２が書かれています。３行目から（３＋p）行目までに、各変数y_i、x_{i1}、・・・、x_{ip}のラベルを１行に１つずつ書きます。３行目にy_iのラベル、４行目にx_{i1}、・・・、（３＋p）行目にx_{ip}のラベルを書きます。リスト１では、「国語」、「英語」、「切片」が順番に書かれています。

（４＋p）行目からデータの組｛y_i、x_{i1}、・・・、x_{ip}｝次の形式

$$y_i \quad x_{i1} \quad \cdots \quad x_{ip}$$

で、1 行に 1 組ずつ書きます。

最後の組

$$y_N \quad x_{N1} \quad \cdots \quad x_{Np}$$

の次の行に、第 1 行に書いた数値より小さい数を書いてデータの終わりとします。リスト１では－1000 が書かれています。

リスト１のファイルを読み込むと図３のようになります。

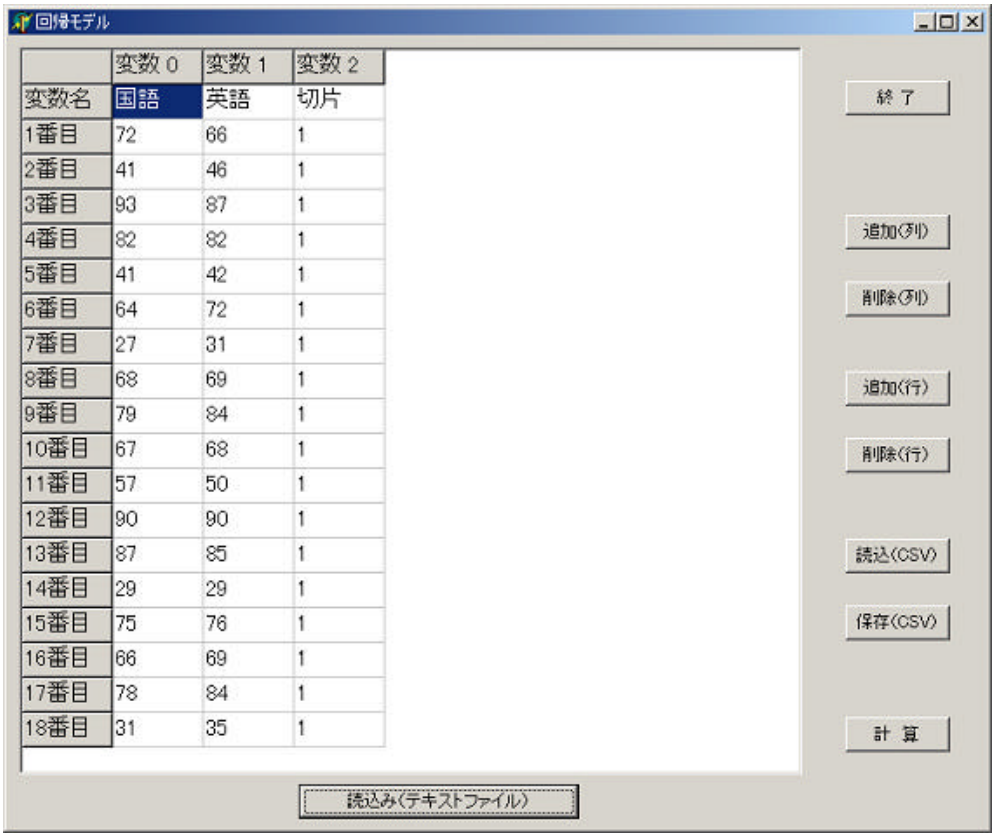
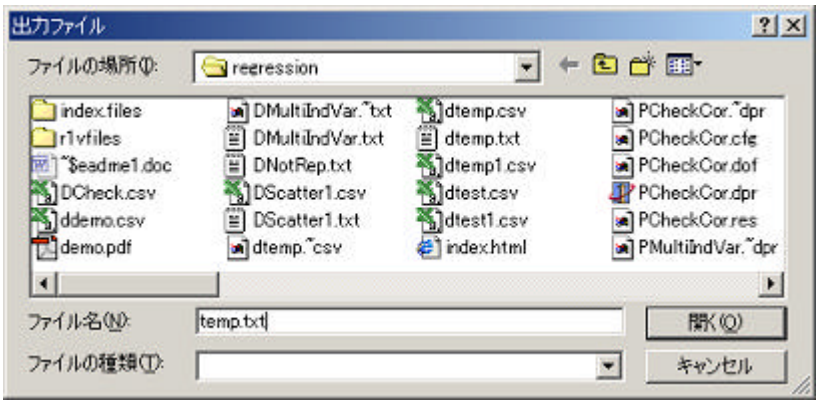


図３　データの設定例

上図のようにデータを設定後、「計算」ボタンをクリックすると計算が始まります。「計算」ボタンをクリックすると、まず計算結果を出力するためのファイル名の設定を求めるダイアログボックスが表示されます。



計算結果はテキストファイルとして書き出されるので、プログラムの実行終了後エディタなどで開いて見ることができます。

ファイル名の設定後「開く」ボタンをクリックすると、計算は一瞬で終わります。「終了」ボタンをクリックして実行を終了します。リスト１のデータの場合、出力ファイルの内容はリスト２のようになっています。

リスト２　リスト１のデータの分析結果

国語	英語	切片
72.000	66.000	1.000
41.000	46.000	1.000
93.000	87.000	1.000
82.000	82.000	1.000
41.000	42.000	1.000
64.000	72.000	1.000
27.000	31.000	1.000
68.000	69.000	1.000
79.000	84.000	1.000
67.000	68.000	1.000
57.000	50.000	1.000
90.000	90.000	1.000
87.000	85.000	1.000
29.000	29.000	1.000
75.000	76.000	1.000
66.000	69.000	1.000
78.000	84.000	1.000
31.000	35.000	1.000

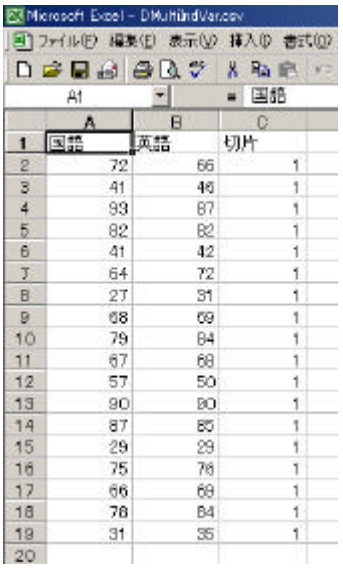
```
B =
        1.02042344   <英語>
        -2.32185061   <切片>

Qe          = 298.997754
Qy          = 7793.61111
決定係数    = 0.961635531
重相関係数  = 0.98063017
```

「B＝」の次行から、係数ベクトル$\mathbf{b}=[b_1,\cdots,b_p]$の各要素の値が１行に一つずつ書かれています。リスト２に示されている値は、次の回帰式を示しています。

国語＝1.02×英語－2.32＋誤差項

図３のように設定されたデータは、「保存（CSV）」ボタンをクリックするとCSV形式で保存することができます。このCSV形式のデータをExcelで読み込むと図４のようになります。



	A	B	C
1	国語	英語	切片
2	72	66	1
3	41	46	1
4	93	87	1
5	82	82	1
6	41	42	1
7	64	72	1
8	27	31	1
9	68	69	1
10	79	84	1
11	67	68	1
12	57	50	1
13	90	80	1
14	87	85	1
15	29	29	1
16	75	76	1
17	66	69	1
18	78	84	1
19	31	35	1
20			

図４　Excel で読み込んだデータ

Excel で図４の形式で作成したデータは、CSV 形式で保存すれば「読込（CSV）」ボタンをクリックすることによりプログラム PMultiIndVar.dpr に読み込むことができます。

参考文献

（１）佐和隆光 (1979)　回帰分析.　　朝倉書店.

（２）岡本安晴（1998）Delphi で学ぶデータ分析法. CQ 出版株式会社.