岡本安晴　2001．5

回　帰　分　析

重回帰分析のための確率モデルについて説明します。

Gauss-Markoff setup（Timm, 1975）と呼ばれているモデル

$$y = Xb + e$$

において、Xとbは定数部分で、それらの積Xbからのランダムな変動分eを加えたものとして、yが表わされています。

$$e = [e_1, \cdots, e_N]'$$

において、各e_iは、平均0、分散s^2の正規分布に従い、お互いに独立であるとします。

XがN行q列の行列で階数がrであるとします。このとき、残差平方和Q_eをs^2で割ったものは、自由度N－rのカイ２乗分布に従います。

$$\frac{Q_e}{s^2} \sim c^2_{N-p} \qquad (1)$$

ここで、Q_eは次式

$$Q_e = \min_b (y - Xb)'(y - Xb)$$

$$= (y - Xb)'(y - Xb)$$

で与えられ、bは次式

$$b = (X'X)^- X'y$$

で与えられるbの推定値です。ここで、$(X'X)^-$は$X'X$の一般逆行列を表しています。

Q_eを求めるときは、$(y - Xb)'(y - Xb)$を最小にするという最小２乗基準でbの推定を行っています。$b = [b_1, \cdots, b_p]'$に対するある制約、例えば$b_1 = 0$というような制約のもとでbの推定を行うときは、その制約条件を満たす範囲でbの値を求めることになります。

制約条件として、次の１次式で表わされるものについて考えます。

$$c_1 b_1 + \cdots + c_p b_p = 0$$

例えば、$b_1 = 0$は

$$1 \cdot b_1 + 0 \cdot b_2 + \cdots + 0 \cdot b_p = 0$$

という形で表わすことができます。

$$c = [c_1, \cdots, c_p]$$

とおき、cをコントラストと呼びます。コントラストはお互いに独立なものを行として複数個まとめて行列

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1p} \\ & \vdots & \\ c_{g1} & \cdots & c_{gp} \end{bmatrix}$$

の形で与えることができます。

制約条件

$$Cb = \mathbf{0}$$

もとでの$(y - Xb)'(y - Xb)$の最小値を、Q_tで表わします。

$$Q_t = \min_{Cb=0} (y - Xb)'(y - Xb)$$

このとき、次式が成り立ちます。

$$Q_h = Q_t - Q_e = (Cb)'[C(X'X)^{-}C']^{-1}(Cb)$$

Q_hは、$Cb = \mathbf{0}$という制約を付けたために増えた、残差平方和Q_eの増加分です。このQ_hをs^2で割ったものは、仮説「$Cb = \mathbf{0}$」が真であるとき、自由度gのカイ２乗分布に従い、自由度Q_e / s^2と独立です。

$$\frac{Q_h}{s^2} \sim c_g^2 \qquad (2)$$

（１）式と（２）式より、帰無仮説「$cb = 0$」のもとで、

$$F = \frac{Q_h / s^2}{(Q_e / s^2)/(N-p)}$$

$$= \frac{Q_h}{Q_e/(N-p)}$$

は、自由度（g、N－r）のF分布に従います。

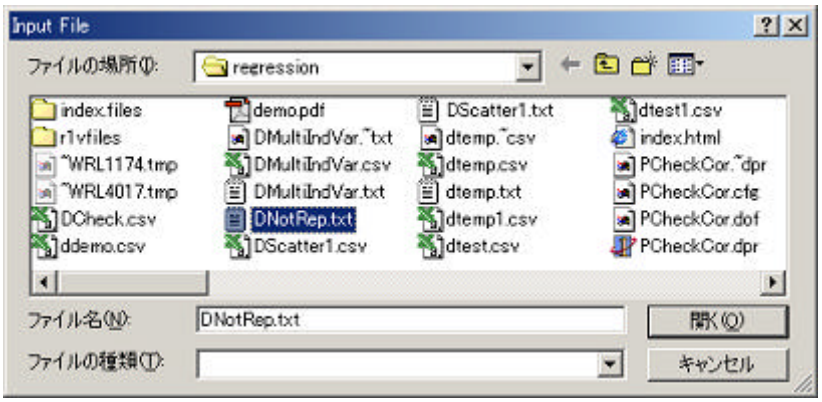
Q_hは、母数bの真の値に対して、Cbが0から離れた値であるほど大きい値になりやすいので、F値が有意水準aで設定された値F_aより大きいとき、帰無仮説「$Cb = 0$」は棄却されます。F_aは、自由度（g、N－r）のF分布に従う確率変数をZとするとき、

$$P(Z \geq F_a) = a$$

を満たす値として求まります。

プログラムPRegU.dprは、上で説明した重回帰分析を行うものです。

このプログラムを実行すると、下図の入力データ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。



入力データファイルはテキストファイルとしてリスト1のように用意します。

リスト１　入力データファイル例

```
      岡本安晴「Delphi で学ぶデータ分析法」ＣＱ出版社,1998
                    表 2.2.4.1.1 より
/
  3
      83    1 0 0        文学部
      84    1 0 0
      74    1 0 0
      58    1 0 0
      73    1 0 0
      74    1 0 0
      66    1 0 0
      85    1 0 0
      61    1 0 0
      77    1 0 0
      65    0 1 0        法学部
      80    0 1 0
      81    0 1 0
      81    0 1 0
      80    0 1 0
      72    0 1 0
      56    0 1 0
      64    0 1 0
      70    0 1 0
      66    0 1 0
      77    0 0 1        経済学部
      72    0 0 1
      70    0 0 1
      75    0 0 1
      65    0 0 1
      53    0 0 1
      67    0 0 1
      75    0 0 1
      60    0 0 1
      63    0 0 1
/

/
  2
     1 -1   0
     1   0 -1


/
  -1
```

リスト１のデータは次の形式に従っています。

行の先頭が/で始まる行を区切り行として用います。

最初の区切り行の次の行に独立変数x_1**、・・・、**x_p**の数ｐを書きます。リスト１では３が書かれています。独立変数の数を書いた行の次の行からデータ**y**、**x_1**、・・・、**x_p**を１行に１組ずつ書きます。データを各行の先頭は空白文字（半角）を１つ以上置きます。行の**

先頭が/で始まる区切り行が現れた前の行でデータは終わります。

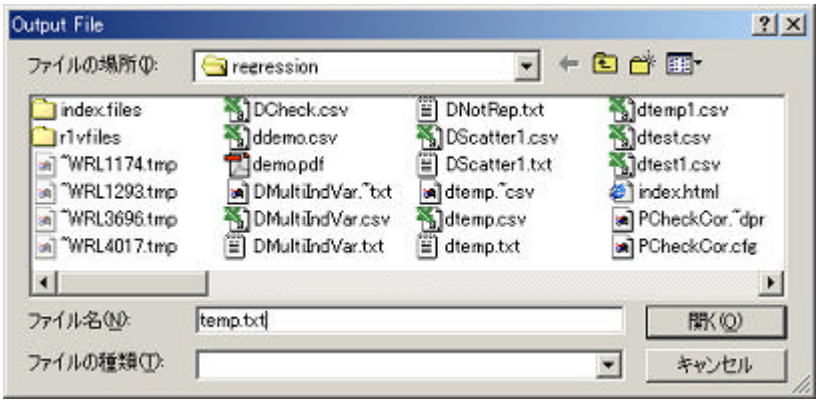
次の区切り行に続けてコントラスト行列を書きます。コントラスト行列は、先ず行数ｇを書いてから行単位で書きます。リスト１では行数２のコントラスト行列

$$C=\begin{bmatrix}1 & -1 & 0\\ 1 & 0 & -1\end{bmatrix}$$

が書かれています。

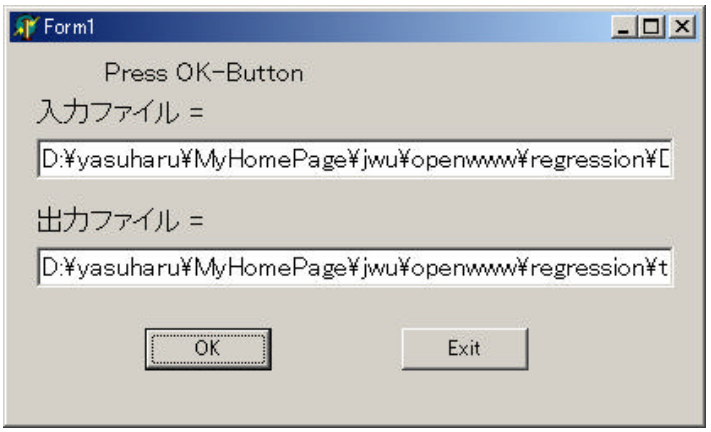
区切り行の次に負の整数が書かれたところでデータファイルの終了とみなされます。

入力データファイル名を設定した後「開く」ボタンをクリックすると、計算結果を書き出す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。

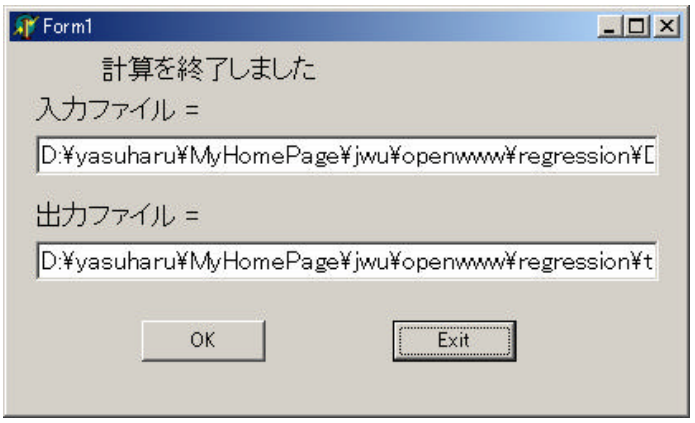


計算結果はテキストファイルとして出力されるので、プログラムの実行終了後適当なエディタで開いて見ることができます。

出力ファイル名の設定後、「開く」ボタンをクリックすると次のフォームが表示されます。



「OK」ボタンのクリックで計算が始まり、計算が終了すると次のフォームになります。



「Exit」ボタンのクリックでプログラムの実行終了となります。計算結果は、リスト１のデータの場合リスト２となっています。

リスト２　　計算結果の例

```
入力データファイル = D:\yasuharu\MyHomePage\jwu\openwww\regression\DNotRep.txt

                83.000    1 0 0
                84.000    1 0 0
                74.000    1 0 0
                58.000    1 0 0
                73.000    1 0 0
                74.000    1 0 0
                66.000    1 0 0
                85.000    1 0 0
                61.000    1 0 0
                77.000    1 0 0
                65.000    0 1 0
                80.000    0 1 0
                81.000    0 1 0
                81.000    0 1 0
                80.000    0 1 0
                72.000    0 1 0
```

```
                56.000    0 1 0
                64.000    0 1 0
                70.000    0 1 0
                66.000    0 1 0
                77.000    0 0 1
                72.000    0 0 1
                70.000    0 0 1
                75.000    0 0 1
                65.000    0 0 1
                53.000    0 0 1
                67.000    0 0 1
                75.000    0 0 1
                60.000    0 0 1
                63.000    0 0 1

B =
        73.5
        71.5
        67.7

Qe              = 2017.1
Qy              = 2190.7
1-Qe/Qy         = 0.0792440772
sqrt(1-Qe/Qy)   = 0.281503246

コントラスト
    1.0   -1.0    0.0
    1.0    0.0   -1.0

     F = 1.16186605      df1 = 2    df2 = 27
```

コントラストの次に、そのコントラストに対する F 値が書き出されています。リスト２の場合、「F = 1.16　df1 = 2　df2 = 27」は、リスト１の表す１要因３水準のデータに対する分散分析の結果と一致しています。

参考文献

（１）Tim, N. H. (1975) *Multivariate Analysis with Applications in Education and Psychology*. Brooks/Cole Publishing Company, Inc.

（２）岡本安晴（1998）Delphiで学ぶデータ分析法．CQ出版社.