岡本安晴　2001.21

回帰分析（単回帰モデル）

Ｎ個のデータの組、(y_i, x_i); $i = 1 \cdots N$、が与えられている場合について考えてみます。例えば、y_iが国語の点数、x_iが英語の点数とします。このとき、y_iをx_iの一次式

$$ax_i + b$$

で表わすことを考えます。

y_iをx_iの一次式で表わしたときの誤差をe_iとおけば、

$$y_i = ax_i + b + e_i$$

と書けます。

上の一次式は回帰式、ａは回帰係数と呼ばれています。グラフ上で回帰式を表わしたものは、回帰直線と呼ばれています。

一次式における定数aとbの値を、誤差の２乗e_i^2の和が最小になるように求めることを考えます。

２乗和をSSEとおけば、

$$SSE = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

となります。

SSEを最小にするaとbを求めるために、SSEをaとbで偏微分します。

$$\frac{\partial SSE}{\partial a} = \sum_{i=1}^{N} 2 \cdot (y_i - ax_i - b) \cdot (-x_i)$$

$$\frac{\partial SSE}{\partial b} = \sum_{i=1}^{N} 2 \cdot (y_i - ax_i - b) \cdot (-1)$$

$\partial SSE / \partial a = 0$とおいて、式を変形します。

$$\sum_{i=1}^{N} (y_i - ax_i - b) x_i = 0 \qquad (1)$$

$$a\sum_{i=1}^{N}x_i^2+b\sum_{i=1}^{N}x_i=\sum_{i=1}^{N}x_iy_i \tag{2}$$

$\P SSE/\P b=0$とおいて、同じように変形します。

$$\sum_{i=1}^{N}(y_i-ax_i-b)=0 \tag{3}$$

$$a\sum_{i=1}^{N}x_i+bN=\sum_{i=1}^{N}y_i \tag{4}$$

$(1.5.2)\times N-(1.5.4)\times\sum_{i=1}^{N}x_i$ より、

$$a\left\{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2\right\}=N\sum_{i=1}^{N}x_iy_i-\left(\sum_{i=1}^{N}x_i\right)\left(\sum_{i=1}^{N}y_i\right)$$

$$a=\frac{N\sum_{i=1}^{N}x_iy_i-\left(\sum_{i=1}^{N}x_i\right)\left(\sum_{i=1}^{N}y_i\right)}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2} \tag{5}$$

となります。

式$(2)\times\sum_{i=1}^{N}x_i-$式$(4)\times\sum_{i=1}^{N}x_i^2$ より、

$$b\left\{\left(\sum_{i=1}^{N}x_i\right)^2-N\sum_{i=1}^{N}x_i^2\right\}=\left(\sum_{i=1}^{N}x_iy_i\right)\left(\sum_{i=1}^{N}x_i\right)-\left(\sum_{i=1}^{N}y_i\right)\left(\sum_{i=1}^{N}x_i^2\right)$$

$$b=\frac{\left(\sum_{i=1}^{N}x_i^2\right)\left(\sum_{i=1}^{N}y_i\right)-\left(\sum_{i=1}^{N}x_i\right)\left(\sum_{i=1}^{N}x_iy_i\right)}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2} \tag{6}$$

となります。

次に、一次式ax_i+bの係数aおよびbを、式（5）および（6）で与えたときの誤差の２乗

和$SSE = \sum \{y_i - (ax_i + b)\}^2$の大きさについて考えます。

まず、$ax_i + b$の平均値を求めてみます。

式（4）の両辺をNで割って

$$a\frac{1}{N}\sum_{i=1}^{N} x_i + b = \frac{1}{N}\sum_{i=1}^{N} y_i$$

よって、

$$\frac{1}{N}\sum_{i=1}^{N}(ax_i + b) = a \cdot \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right) + b$$

$$= \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$= \overline{y}$$

となります。

すなわち

$$\frac{1}{N}\sum_{i=1}^{N}(ax_i + b) = a\overline{x} + b = \overline{y}$$

です。

但し、

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i 、 \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

とおきます。

次に、

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 、$$

$$s_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 、$$

および

$$s_{xy}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})$$

とおきます。

s_x^2、s_y^2およびs_{xy}は、次のように変形できます。

$$s_x^2=\frac{1}{N}\left\{\sum_{i=1}^{N}x_i^2-2\overline{x}\left(\sum_{i=1}^{N}x_i\right)+N\overline{x}^2\right\}$$

$$=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)^2$$

同様にして

$$s_y^2=\frac{1}{N}\sum_{i=1}^{N}y_i^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}y_i\right)^2$$

$$s_{xy}=\frac{1}{N}\left\{\sum_{i=1}^{N}x_iy_i-\overline{x}\left(\sum_{i=1}^{N}y_i\right)-\overline{y}\left(\sum_{i=1}^{N}x_i\right)+N\overline{xy}\right\}$$

$$=\frac{1}{N}\left(\sum_{i=1}^{N}x_iy_i\right)-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)\left(\sum_{i=1}^{N}y_i\right)$$

従って、式（5）は、次のように書けます。

$$a=\frac{\frac{1}{N}\sum_{i=1}^{N}x_iy_i-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)\left(\sum_{i=1}^{N}y_i\right)}{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)^2}$$

$$=\frac{s_{xy}}{s_x^2} \qquad (7)$$

以上より、誤差の２乗和は、次のように計算できます。

$$\sum_{i=1}^{N}\{y_i-(ax_i+b)\}^2=\sum_{i=1}^{N}\{(y_i-\overline{y})-(ax_i+b-\overline{y})\}^2$$

$$=\sum_{i=1}^{N}(y_i-\overline{y})^2-2\sum_{i=1}^{N}(y_i-\overline{y})(ax_i+b-\overline{y})+\sum_{i=1}^{N}(ax_i+b-\overline{y})^2$$

$$= Ns_y^2 - 2\sum_{i=1}^{N}(y_i - \overline{y})(ax_i + b - a\overline{x} - b) + \sum_{i=1}^{N}(ax_i + b - a\overline{x} - b)^2$$

$$= Ns_y^2 - 2\sum_{i=1}^{N}(y_i - \overline{y})a(x_i - \overline{x}) + \sum_{i=1}^{N}a^2(x_i - \overline{x})^2$$

$$= Ns_y^2 - 2a\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) + a^2\sum_{i=1}^{N}(x_i - \overline{x})^2$$

$$= Ns_y^2 - 2\frac{s_{xy}}{s_x^2}Ns_{xy} + \left(\frac{s_{xy}}{s_x^2}\right)^2 Ns_x^2$$

$$= Ns_y^2 - N\frac{s_{xy}^2}{s_x^2}$$

従って、誤差の２乗和の平均値の、y_iの分散s_y^2に対する比は、次のようになります。

$$\frac{\frac{1}{N}\sum_{i=1}^{N}\{y_i - (ax_i + b)\}^2}{s_y^2} = \frac{s_y^2 - \frac{s_{xy}^2}{s_x^2}}{s_y^2}$$

$$= 1 - \left(\frac{s_{xy}}{s_x s_y}\right)^2$$

$$= 1 - r^2 \qquad (8)$$

ここで、r は相関係数で、次式で与えられます。

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

$$= \frac{s_{xy}}{s_x s_y}$$

式（8）より、r の絶対値が１に近いほど、誤差の小さいことがわかります。式（8）で与えられる$1 - r^2$は、非決定係数と呼ばれています。これに対して、r の２乗r^2は、直線回帰の決定係数と呼ばれています。

いま、y_iの分散s_y^2を、次のように分解します。

$$Ns_y^2 = \sum_{i=1}^{N}(y_i - \overline{y})^2$$

$$= \sum_{i=1}^{N}\{y_i - (ax_i + b) + (ax_i + b) - \overline{y}\}^2$$

$$= \sum_{i=1}^{N}\{y_i - (ax_i + b)\}^2 + 2\sum_{i=1}^{N}\{y_i - (ax_i + b)\}\{(ax_i + b) - \overline{y}\} + \sum_{i=1}^{N}\{(ax_i + b) - \overline{y}\}^2$$

上式の第２項は、次のように変形できます。

$$\sum_{i=1}^{N}\{y_i - (ax_i + b)\}\{(ax_i + b) - \overline{y}\}$$

$$= \sum_{i=1}^{N}\{y_i - (ax_i + b)\}ax_i + \sum_{i=1}^{N}\{y_i - (ax_i + b)\}(b - \overline{y}) \qquad (9)$$

式（9）の第１項は、式（１）を用いると、

$$\sum_{i=1}^{N}\{y_i - (ax_i + b)\}ax_i = a\sum_{i=1}^{N}(y_i - ax_i - b)x_i$$

$$= 0$$

となります。

式（9）の第２項は、式（３）を用いると、

$$\sum_{i=1}^{N}\{y_i - (ax_i + b)\}(b - \overline{y}) = (b - \overline{y})\sum_{i=1}^{N}(y_i - ax_i - b)$$

$$= 0$$

となります。

したがって、

$$Ns_y^2 = \sum_{i=1}^{N}\{y_i - (ax_i + b)\}^2 + \sum_{i=1}^{N}\{(ax_i + b) - \overline{y}\}^2$$

$$= Ns_y^2(1 - r^2) + \sum_{i=1}^{N}\{(ax_i + b) - \overline{y}\}^2 \qquad （式（8）より）$$

ゆえに、

$$\frac{\frac{1}{N}\sum_{i=1}^{N}\{(ax_i+b)-\bar{y}\}^2}{s_y^2}=\frac{\frac{1}{N}\sum_{i=1}^{N}\{(ax_i+b)-(a\bar{x}+b)\}^2}{s_y^2}=r^2$$

となります。

上式より、決定係数r^2は、回帰直線による予測値の分散の、データ値の分散s_y^2に対する比になっていることがわかります。

回帰直線を求めるプログラム POneIndVar.dpr を用意しました。

このプログラムのフォームは、図１のように用意されています。

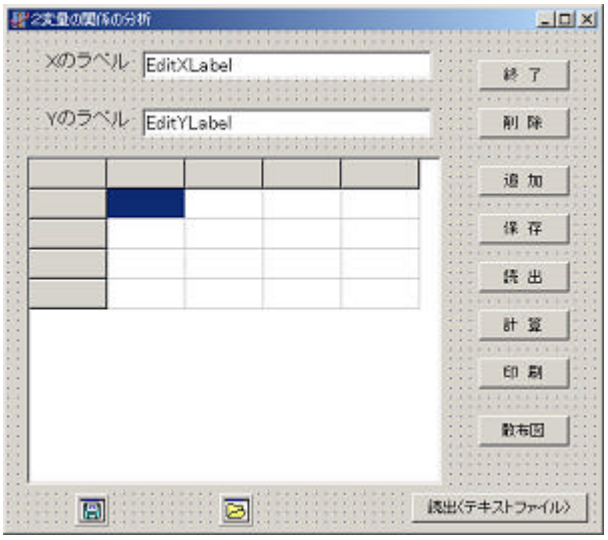


図１　フォームの準備

散布図や回帰直線は、実行時において図１の「散布図」ボタンをクリックしたときに生成・表示されるフォームの Image コンポーネントの Canvas に描きます。このフォームはプログラミング時には次のようになっています。

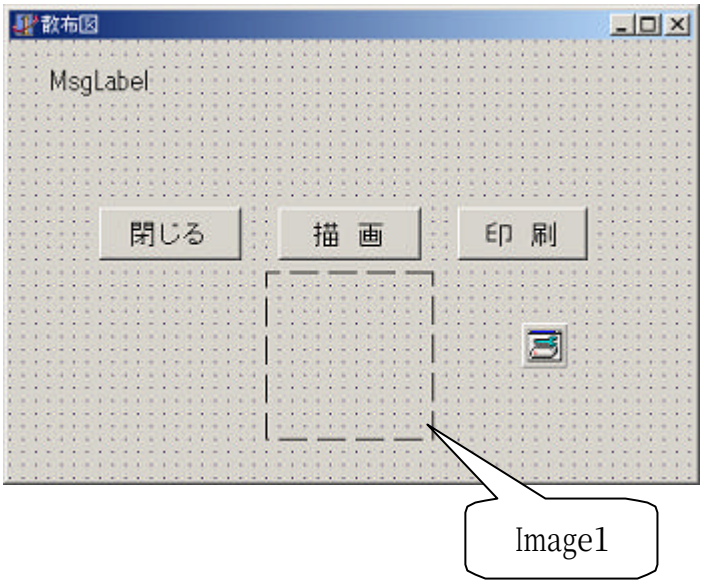


Image1 の Canvas において、第２変量を表わすＹ軸のラベルを縦方向に描くために（図３における「国語」）、Image コンポーネント Image2 を実行中に生成して、この Canvas に描いたラベル（文字列）を Image1 の Canvas に９０度回転してコピーすることにします。

プログラムの実行開始時のフォームは、図２のようになっています。

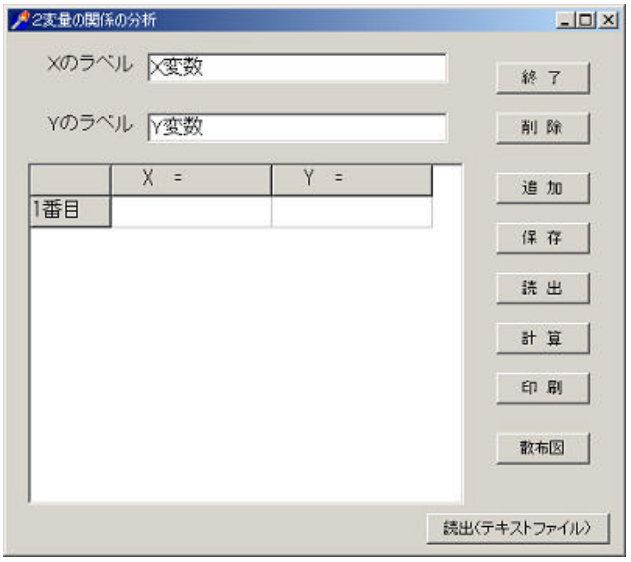


図２　実行開始時のフォーム

「追加」ボタンを押して行を必要なだけ追加しデータを設定します。「Xのラベル」「Yのラベル」の右側の Edit コンポーネントにはそれぞれの変数の適当なラベル（名前）を設

定します。行を削除するときは、その行のセルをクリックしてアクティブにしてから「削除」ボタンをクリックします。

データはリスト１のように用意されているファイルから読み込むことも出来ます。

リスト１　入力データファイル例

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

第１行目に第１変数のラベル、第２行目に第２変数のラベルを置きます。

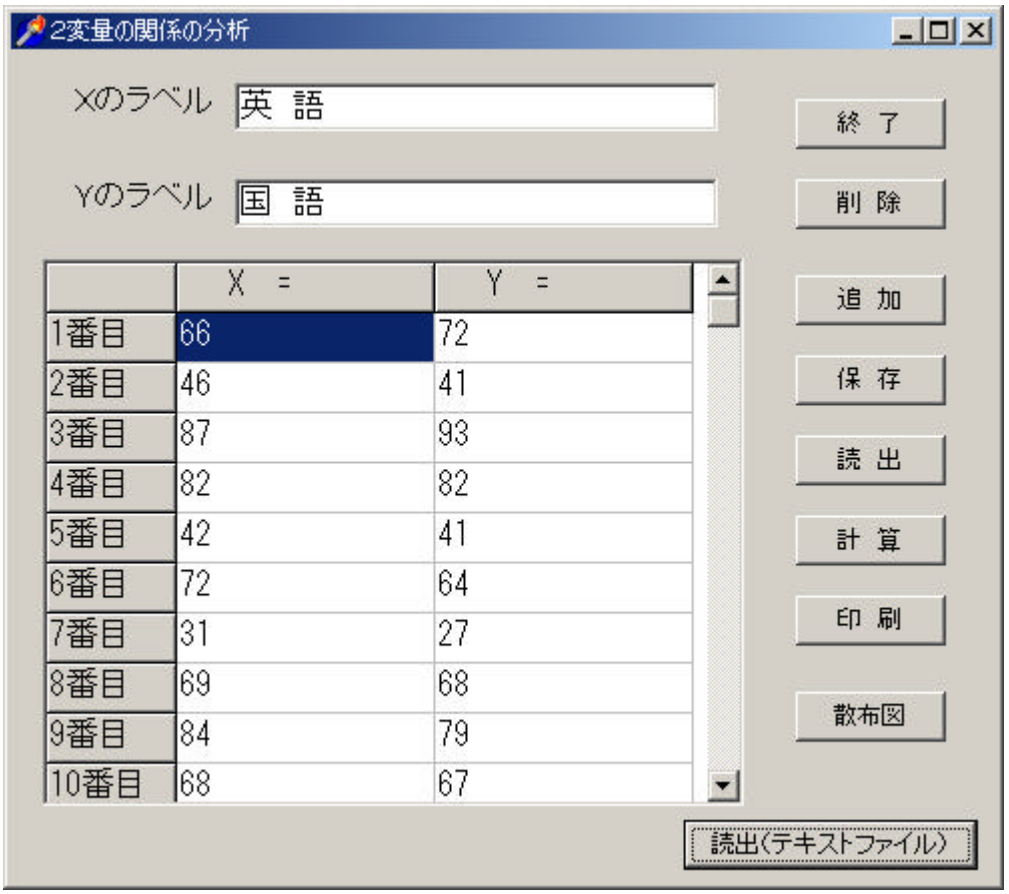
次に、第３行目に、どのデータ値よりも小さい適当な値を、基準値として置きます。リスト１では-100 がおかれています。数値の後ろは、データとしては無視されるので、１つ以上の空白を置いた後、「英語」、「国語」などの文字列がコメントとして置かれています。

４行目から、各行に１組ずつ、データがx_i、y_iの順に並べられています。

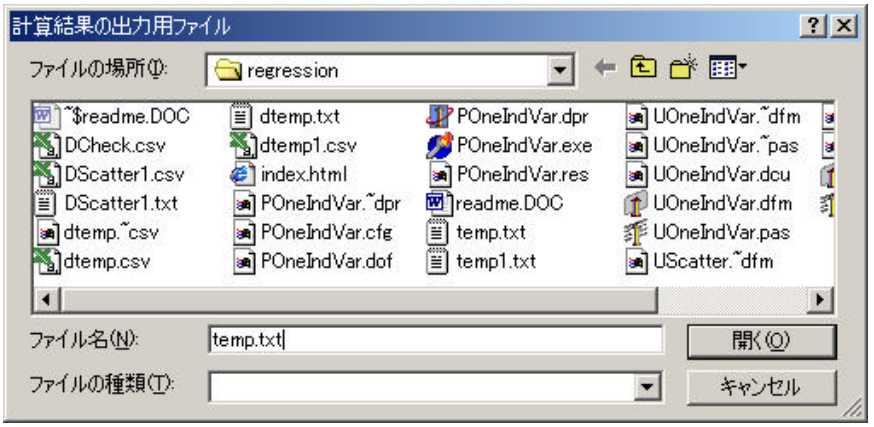
最後のデータが並べられている行の次の行は、３行目においた数値より小さい値を置きます。リスト１では-1000 が置かれています。

リスト１の形式で用意されたテキストファイルを読み込むときは「読出（テキストファイ

ル）」ボタンをクリックします。下図はリスト１のデータを読み込んだ状態です。

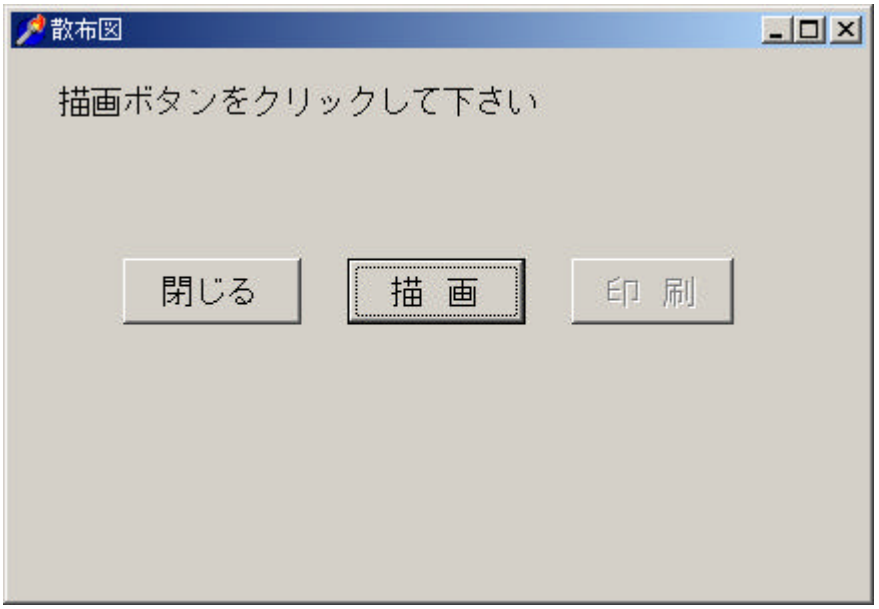


データの設定後、「散布図」あるいは「計算」ボタンをクリックすると計算が始まりまります。「計算」ボタンのクリックでは、各変数の平均値、標準偏差、相関係数などが計算されます。このボタンをクリックすると、まず計算結果出力用のファイル名を求めるダイアログボックスが表示されます。

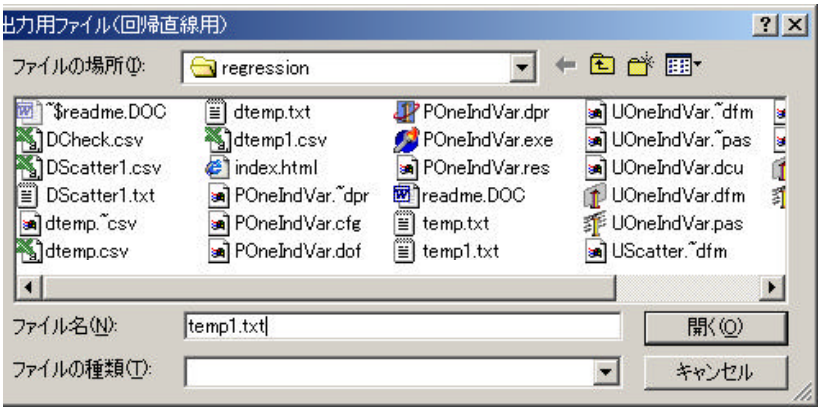


名前の設定後、「開く」ボタンのクリックで計算が始まり、計算結果が設定したファイル名のファイルに書き出されます。書き出されたファイルはテキストファイルなので、プログラムの実行終了後、適当なエディタで開いて見ることが出来ます。

「散布図」ボタンをクリックすると、まず下図のフォームが表示されます。



「描画」ボタンのクリックで回帰曲線の描画が行われますが、その前に計算結果を出力するテキスト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。



ファイル名の設定後、「開く」ボタンをクリックすると計算が始まります。x[i]、y[i]と、直線回帰による値a*x[i]+b、誤差y[i]-(a*x[i]+b)が出力ファイルに書き出され、続いてaとbの値も書き出されます。

次に、相関係数rが求められて、決定係数r^2とともに書き出されてから、出力用ファイルが閉じられます。

出力用ファイルへの書き出しが終わると、散布図と回帰直線が図３のように描画されます。

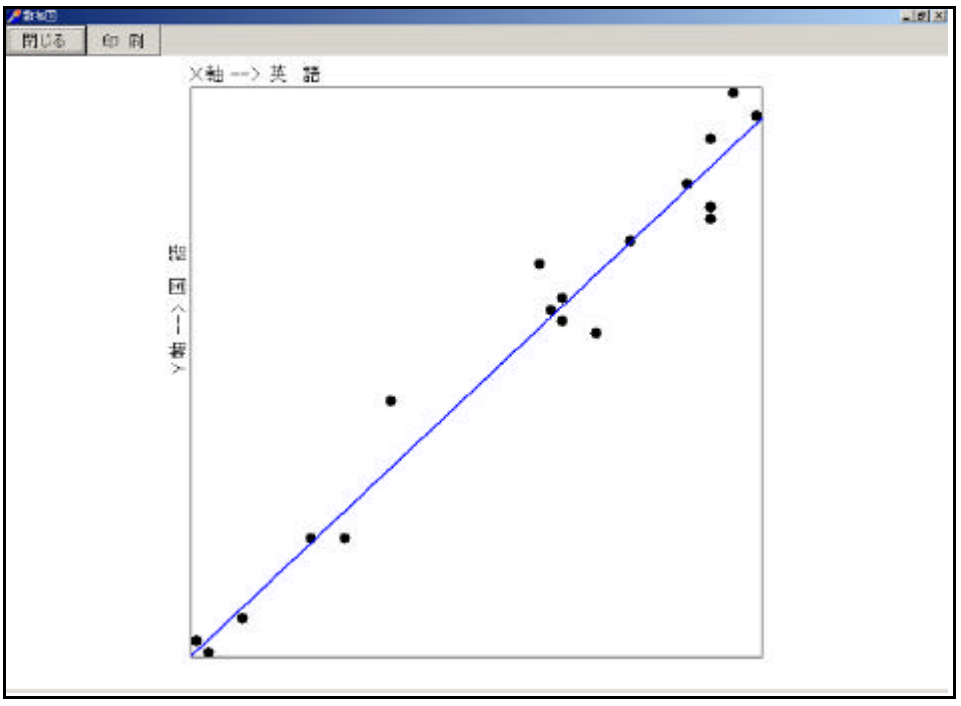


図３　回帰直線と散布図の描画画面

図３において、「印刷」ボタンをクリックすると回帰直線と散布図がプリンタに出力され

ます。

出力ファイルは、リスト２のようになっています。

リスト２　出力ファイル

```
変量Ｘのラベル --> 英　語
変量Ｙのラベル --> 国　語

    X =       Y =       a*X+b =    誤差
    66.000    72.000    65.026     6.974
    46.000    41.000    44.618    -3.618
    87.000    93.000    86.455     6.545
    82.000    82.000    81.353     0.647
    42.000    41.000    40.536     0.464
    72.000    64.000    71.149    -7.149
    31.000    27.000    29.311    -2.311
    69.000    68.000    68.087    -0.087
    84.000    79.000    83.394    -4.394
    68.000    67.000    67.067    -0.067
    50.000    57.000    48.699     8.301
    90.000    90.000    89.516     0.484
    85.000    87.000    84.414     2.586
    29.000    29.000    27.270     1.730
    76.000    75.000    75.230    -0.230
    69.000    66.000    68.087    -2.087
    84.000    78.000    83.394    -5.394
    35.000    31.000    33.393    -2.393

a =          1.02042
b =         -2.32185

r =          0.98063
決定係数 =   0.96164
```

リスト２から、リスト１のデータの場合、

a　1.0　　b　−2.3

であることがわかります。

a　1.0ということは、英語の点数１点分の変化量に対応する国語の点数の変化量が、約１点ということになります。

さらに、b　−2.3ですから、国語の点は、英語の点から約2.3点減じた値が直線回帰による値となります。

すなわち、

国語　　1　×　英語　ー　2.3

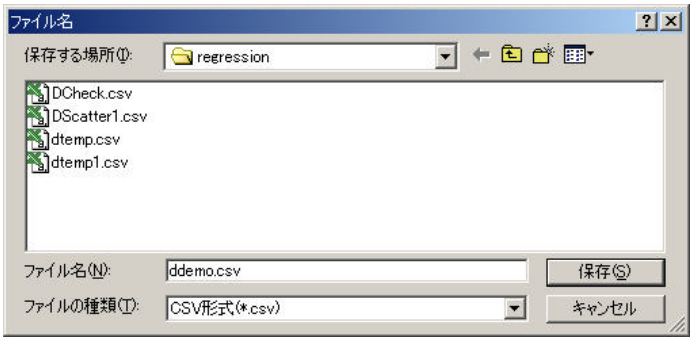
です。

決定係数は約 0.96 になっています。このことから、英語の点数によって、国語の点数の変動量の約 96%が予測できていると解釈できます。

プログラム POneIndVar.dpr では、図４のフォームにおいて

図４　データの設定されたフォーム

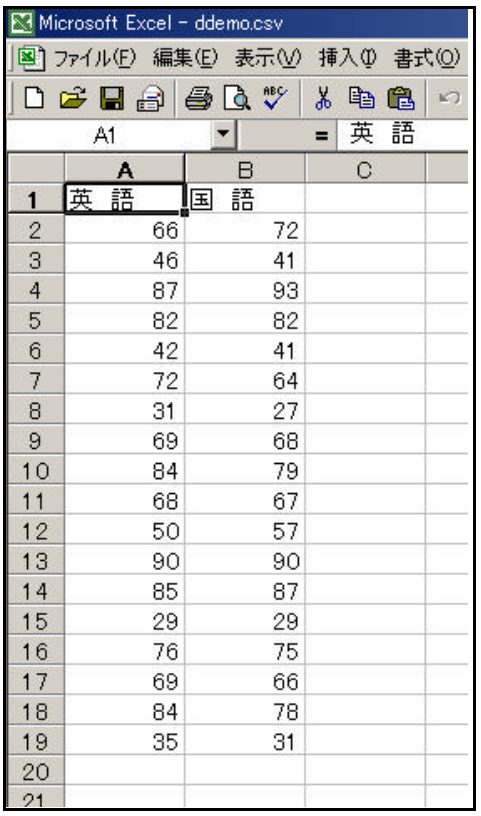
「保存」ボタンを押すと、CSV の形式でデータが保存されます。このときは次のダイアログボックス



が表示されるので、拡張子が.csv であるファイル名を設定します。この形式で保存したファ

イルは図２のフォームにおいて「読出」ボタンをクリックすると読み込むことが出来ます。

また、Excel でも読み込めます。Excel で読み込むと次図のようになります。



	A	B	C
1	英 語	国 語	
2	66	72	
3	46	41	
4	87	93	
5	82	82	
6	42	41	
7	72	64	
8	31	27	
9	69	68	
10	84	79	
11	68	67	
12	50	57	
13	90	90	
14	85	87	
15	29	29	
16	76	75	
17	69	66	
18	84	78	
19	35	31	
20			
21			

上図のように設定された Excel データを拡張子が.csv である CSV 形式で保存すると、プログラム POneIndVar.dpr から「読出」ボタンのクリックで読み込むことが出来ます。