２章 重回帰分析

複数の変数で、１つの変数を予測するような手法を「重回帰分析」といいます。前の巻でところで述べた回帰分析は、１つの説明変数で目的変数を予測（説明）する手法でしたが、この説明変数が複数個になったと考えればよいでしょう。重回帰分析はこの予測式を与える分析手法です。以下の例を見て下さい。

例

以下のデータ（Samples¥重回帰分析 1.txt）をもとに体重を身長と胸囲の１次関数で予測せよ。

体重	身長	胸囲	体重	身長	胸囲
61.0	167.0	84.0	49.5	164.7	78.0
55.5	167.5	87.0	61.0	171.0	90.0
57.0	168.4	86.0	59.5	162.6	88.0
57.0	172.0	85.0	58.4	164.8	87.0
50.0	155.3	82.0	53.5	163.3	82.0
50.0	151.4	87.0	54.0	167.6	84.0
66.5	163.0	92.0	60.0	169.2	86.0
65.0	174.0	94.0	58.8	168.0	83.0
60.5	168.0	88.0	54.0	167.4	85.2
49.5	160.4	84.9	56.0	172.0	82.0

体重を身長と胸囲の１次式で予測（説明）するのですから、体重を目的変数、身長と脅威が説明変数となります。説明変数を独立変数、目的変数を従属変数と呼ぶ場合もあります。予測式は以下の形になります。

$$体重 = b_1 \times 身長 + b_2 \times 胸囲 + b_0$$

この式は重回帰式と呼ばれ、係数 b_1, b_2, b_0 は偏回帰係数と呼ばれます。それでは実際に重回帰分析を実行してみましょう。

データ Samples¥重回帰分析 1.txt を読み込んで、メニュー［分析－多変量解析他－重回帰分析］を選択すると図 2.1 の分析メニューが表示されます。

図 2.1 重回帰分析メニュー

分析メニュー中に目的変数を最初に選択するよう書いてありますので、「変数選択」ボタンで体重を最初に選択し、他の変数を後から選択します（All の選択でそのようになります)。

College Analysis では基本的に分析名の書いてあるボタンをクリックすると最も大事な結果が表示されるようになっていますので、この場合はまず「重回帰分析」ボタンをクリックします。すると図 2.2 の結果が表示されます。

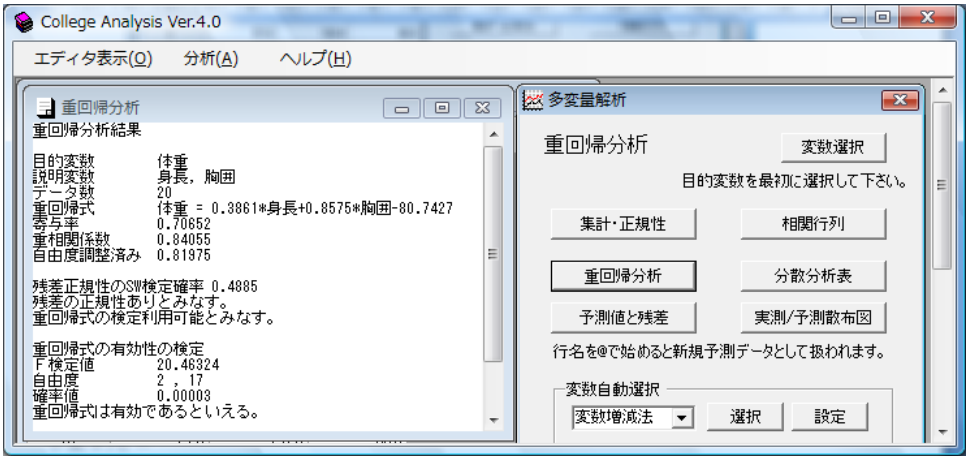

図 2.2 重回帰分析結果

ここで重回帰式と偏回帰係数は数式の形で表示されています。重相関係数は体重の実測値と予測値の相関係数で、寄与率はこの重回帰式がどの程度体重の変動を説明できているかを表しており、重相関係数の２乗で与えられます。自由度調整済みとなっているのは自由度調整済み重相関係数のことで、説明変数をたくさん選ぶことで重相関係数が高くなっていくことを調整した指標です。

その下には残差の正規性の検定を行っている部分があります。回帰分析の回帰式の検定のときにも行ったものと同じ検定を実施するためには、回帰分析の体重の実測値と予測値の差である残差について正規性が成り立つことが必要ですが、ここではその検定を行っています。

その下に重回帰式の有効性の検定の結果が表示されていますが、これは残差の変動と重回帰式の変動の大きさを比べるもので、残差の変動が大きすぎると重回帰式の有効性が疑われることになります。ここでは、重回帰式が有効であることが示されています。

図 2.2 の結果表示と同時に図 2.3 で与えられるグリッド（表）も出力されます。

	偏回帰係数	標準化係数	t 検定値	自由度	確率値	相関係数	偏相関係数
▶ 身長	0.3861	0.4333	3.2335	17	0.0049	0.5591	0.6171
胸囲	0.8575	0.6401	4.7768	17	0.0002	0.7253	0.7570
切片	-80.7427	0.0000	-3.5761	17	0.0023		

図 2.3 重回帰分析の結果のグリッド出力

この表では、重回帰式の係数である偏回帰係数の他に、データを平均 0、不偏分散 1 に標準化した場合の偏回帰係数である標準化偏回帰係数（標準化係数となっています）も表示されています。標準化偏回帰係数は重回帰式における各変数の重要性を表す指標です。通常の偏回帰係数では変数の大きさの影響でその値だけで重要性を判断することはできません。次の t 検定値から確率値までは各偏回帰係数（切片も含めて）が統計的に 0 でないことを調べる検定結果です。確率値は偏回帰係数が 0 となる確率で有意水準以下で偏回帰係数が 0 でないと判断します。

相関係数は目的変数と各変数のピアソンの相関係数で、偏相関係数は他の説明変数

からの影響を取り除いた目的変数と説明変数の相関係数です。目的変数は説明変数から影響を受けますが、直接的な影響と間接的な影響が考えられ、この間接的な影響を取り除いたものです。

図 2.1 のメニューで「分散分析表」ボタンをクリックすると図 2.4 の結果が表示されます。

図 2.4 分散分析表出力結果

これは分散分析表と呼ばれ、全変動とその中の回帰変動、残差変動を表示したものです。また図 2.2 で表示された重回帰式の有効性の検定結果も表形式で表示しています。

図 2.1 のメニューで「予測値と残差」ボタンをクリックすると図 2.5 の画面が表示されます。

図 2.5 予測値と残差出力結果

ここでは目的変数の実測値と重回帰式による予測値、及びそれらの差である残差を表示しています。実測値と予測値の関係を図で見たいなら、「実測/予測散布図」ボタンをクリックします。図 2.6 のような散布図が得られます。

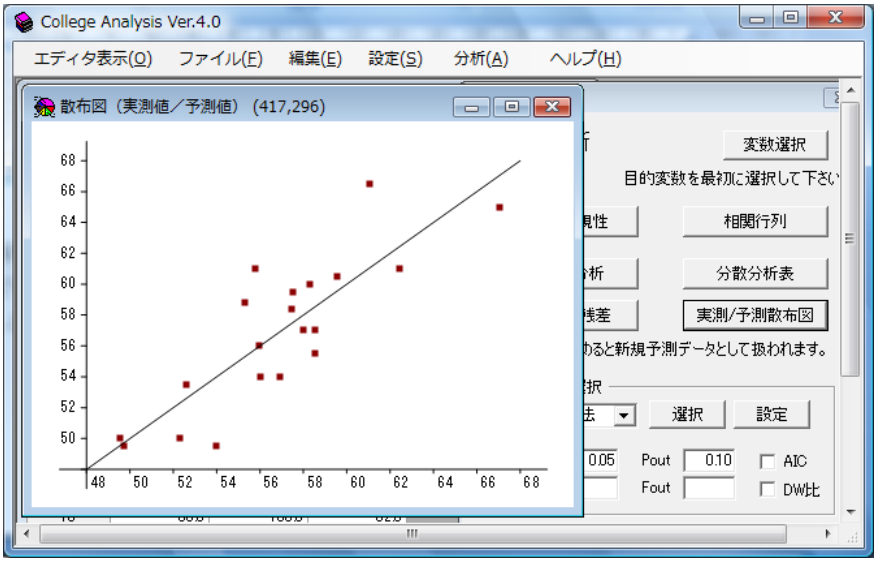

図 2.6 実測値と予測値の散布図

タイトルバーに（実測値／予測値）とありますが、これは実測値が縦軸、予測値が横軸であることを示しています。また斜めの線はこの散布図の回帰直線で、実測値＝予測値を表す直線になります。

重回帰分析は説明変数をたくさん選ぶほど寄与率が高くなりますが、多ければ良いというものではありません。意味のある説明変数でシンプルに式を作ることこそモデルとして重要です。そこで、図 2.3 のところで見た偏回帰係数の検定を行い、有意なものだけを残すことを考える必要があります。これは一つ一つの変数を吟味しながら利用者が行うことをお勧めしますが、自動的に行うこともできます。それが図 2.1 のメニューの下の部分の変数自動選択です。その方法には、変数増減法、変数減少法、変数増加法が用意されていますが、良く利用されるのが変数増減法です。意味のある変数を追加し、重回帰分析を行い、その中で不要となった変数を除去するということを繰り返しますが、そのときの追加と削除の基準が Pin, Pout の確率値です。これは偏回帰係数の検定と同じなので、t 検定を用いてもよいのですが、2 乗して F 検定を利用するのが一般的です。Fin, Fout はそのときの F 値を使いますが、確率で考える方が意味がはっきりするように思います。

選択法を左のコンボボックスで選び、「選択」ボタンをクリックすると選択過程で得られた図 2.3 と同じ表が出力されます。ここでは例が説明変数 2 つなので図は省略します。得られた結果で良ければ、「設定」ボタンで選択変数を設定し、分析を実行することができるようになります。

最後にこれまでのことを簡単にまとめておきましょう。

重回帰分析とは以下の形で目的変数を予測する。

目的変数 ＝ b_1×説明変数１＋b_2×説明変数２＋・・・＋b_0

係数の値は？　→　偏回帰係数

説明変数の重要性は？　→　標準化偏回帰係数

どの程度予測できるか？　→　重相関係数，寄与率（決定係数）

このモデルは有効か？　→　Ｆ検定値と確率（要残差正規性）

それぞれの係数は有効か？　→　ｔ検定値と確率（要残差正規性）

他の変数の影響を除いた目的変数と各説明変数の相関は？　→　偏相関係数

どの程度予測できているのか図的に見たい　→　散布図

どの程度予測できているのかデータ毎に見たい　→　予測値と残差

まとめ

目的変数を体重に、説明変数を身長と胸囲にして、重回帰分析を行ったところ、以下の回帰式を得た。

体重 = 0.3861*身長+0.8575*胸囲-80.7427

予測体重と実測体重の相関である重相関係数は 0.84055 で、回帰式の寄与率は 0.70652 となった。これから体重変動の約 71%が説明できることが分かる。　各変数の予測における重要性を示す標準化偏回帰係数は、身長が 0.4333、胸囲が 0.6401 と胸囲が少し上回っている。

回帰式の妥当性の検定を行ったところ p=0.00003 となり、妥当性が有意に示された。また、各偏回帰係数が 0 と異なることを示す検定では、身長が p=0.00488、胸囲が p=0.00018、切片は p=0.00233 となり、各係数とも有意に 0 と異なっている。

以上のことからこの回帰式は予測モデルとして、かなり良いモデルになっている。

ここで利用した理論の公式は以下の通りです。

理論

標本番号	目的変数	説明変数 1	…	説明変数 p
1	y_1	x_{11}	…	x_{k1}
2	y_2	x_{12}	…	x_{k2}
:		:	…	:

n	y_n	x_{1n}	…	x_{kn}

目的

目的変数を最もよく説明する説明変数の線形モデルを与える。

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

偏回帰係数

目的変数のゆらぎ D を最も良く説明する偏回帰係数 b_0, b_i を求める。

$$Y_\lambda = b_0 + b_1 x_{1\lambda} + b_2 x_{2\lambda} + \cdots + b_k x_{k\lambda}$$

$$D = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \text{最小化}$$

標準化偏回帰係数

$y_\lambda^* = \dfrac{y_\lambda - \bar{y}}{u_y}$, $x_{i\lambda}^* = \dfrac{x_{i\lambda} - \bar{x}_i}{u_i}$ として、 y^* を説明する回帰式を求める。

$$Y_\lambda^* = b_1^* x_{1\lambda}^* + b_2^* x_{2\lambda}^* + \cdots + b_k^* x_{k\lambda}^* \qquad b_i^* = b_i \frac{u_i}{u_y}$$

寄与率と重相関係数

$$SV = \sum_{\lambda=1}^{n} (y_\lambda - \bar{y})^2 = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 + \sum_{\lambda=1}^{n} (Y_\lambda - \bar{Y})^2 = EV + RV$$

全変動 SV , 回帰変動 RV , 残差変動 EV

寄与率 $R^2 = RV/SV$

重相関係数 $R = \sqrt{RV/SV}$ 観測値と予測値の相関係数でもある。

自由度調整済み重相関係数 $\bar{R} = \sqrt{1 - \dfrac{EV/(n-k-1)}{SV/(n-1)}}$

回帰式の有効性の検定

$$F = \frac{RV/k}{EV/(n-k-1)} \sim F_{p,n-p-1} \text{分布}$$

偏回帰係数の検定

$b_i = 0$ の検定 自由度 $n-k-1$ の t 検定

$b_0 = 0$ の検定 自由度 $n-k-1$ の t 検定

偏相関係数 $r_{iy\cdot 12\cdots i-1 i+1\cdots k}$

X_i ：他の説明変数で作った x_i の予測回帰式

Y_i ：他の説明変数で作った y の予測回帰式

$x_i' = x_i - X_i$, $y' = y - Y_i$ とした場合の、

x_i' と y' の相関係数（他の変数の影響を除いた相関係数）

残差

$$z_\lambda = y_\lambda - Y_\lambda$$

問題１

Samples¥重回帰分析 2.txt はある大学の学生について調べた、卒業試験の成績、入試点数、内申点数、ある５日間の勉強時間、授業への出席率のデータである。卒業試験の成績を他の変数で予測する重回帰分析を行い、結果をまとめにならって記述せよ。

問題２

Samples¥重回帰分析 2.txt について、重回帰分析を行い、以下の問いに答えよ。

１）回帰式を求めよ。

卒業試験＝　［　　］入試点数　＋［　　］内申点数
＋［　　］勉強時間　＋［　　］出席率
＋［　　］

２）この回帰式の寄与率を求めよ。［　　］

３）この場合残差の分布は正規分布といえるか。［正規分布・正規分布でない］

４）回帰式の係数の t 検定（偏回帰係数が 0 と異なるかどうかの検定）の確率値が 0.05 を超えるものの中で最大となる変数（最も不要な変数）を順次削除していくと、最終的に残るものは何か。各段階の検定確率値を記入せよ。但し、削除した変数のところは以後空欄にし、すべての確率が 0.05 未満になった場合は確定とする。

	４変数	３変数	２変数	１変数
入試点数				
内申点数				
勉強時間				
出席率				

５）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

卒業試験＝　［　　］入試点数　＋［　　］内申点数
＋［　　］勉強時間　＋［　　］出席率
＋［　　］

６）上の回帰式の寄与率を求めよ。［　　］

７）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。
［大きく下がっている・あまり下がっていない］

8）この式を新しい予測モデルとして採用するか。
［採用する・採用しない］

9）新しい予測モデルで、データ中の最初（1 番）の学生について卒業試験の実測値，その予測値，残差（実測値と予測値の差）はいくらか。
実測値［　　］　予測値［　　］　残差［　　］

10）上と同様のモデルで、質問項目の値が入試点数 70、内申点数 3.5、勉強時間 5、出席率 70%の学生の卒業試験はいくらに予測されるか。
［　　］

問題３

Samples¥重回帰分析 3.txt について、重回帰分析を行い、以下の問いに答えよ。

1）売上を従業員と資産で推測する回帰式を求めよ。
売上＝　［　　］従業員　＋［　　］資産
　　＋［　　］

2）上の回帰式の寄与率を求めよ。［　　］

3）log 売上を log 従業員と log 資産で推測する回帰式を求めよ。但し、この対数は底が 10 の常用対数である。
log 売上＝　［　　］log 従業員　＋［　　］log 資産
　　＋［　　］

4）上の回帰式の寄与率を求めよ。［　　］

5）$z = cx^a y^b$ の常用対数をとると以下のようになる。

$$\log_{10} z = a\log_{10} x + b\log_{10} y + \log_{10} c$$

ここに、$d = \log_{10} c$ とすると、$c = 10^d$（Excel で計算可能）

これを用いて３）の回帰式を以下の形に書き換えよ。
売上＝［　　］×従業員［　　］×資産［　　］

6）１）の回帰式と３）の回帰式はどちらがより優れていると思われるか。
どちらも良いモデルであるが、どちらかといえば［１・３］が優れている。