６章　相関係数の検定と回帰分析

この章では２つの量的なデータの関係を調べる検定手法を学びます。２つの量的なデータを表示するには散布図がよく用いられ、描画された点の散らばり方によって、相関係数が計算されました。この相関係数はピアソン（Pearson）の相関係数と呼ばれ、２つのデータの間の線形の（散布図では直線的な）関係を調べるものでした。しかし、一般に関係は線形なものとは限らず、非線形な（散布図では曲線的な）関係も多く見られます。我々はこのような関係も考える必要がありますが、このように非線形な、但し大小関係だけは考えた、相関係数にスピアマン（Spearman）の順位相関係数があります。これは２つの変数の大きさの順位を用いた相関係数です。我々はこれらの２つの相関係数を使って、相関の有無を調べることにします。

6.1 （Pearson の）相関係数

ここでは Pearson の相関係数が、統計的に 0 と異なるかという検定を紹介します。但し、2 つのデータが 2 変量正規分布に従う場合にこの検定は有効です。2 変量正規分布の検定については College Analysis にありませんので、散布図でデータがラグビーボール状に描画されており、各変数に正規性がある（正確にはないことはない）ことを調べればよいでしょう。形が歪んでいたり、どちらかまたは両方の変数に正規性がなかった場合には次節で述べるスピアマンの順位相関係数を用います。分析メニューは［分析－基本統計－相関と回帰分析］を選択して、図 6.1.1 のメニュー画面を表示します。

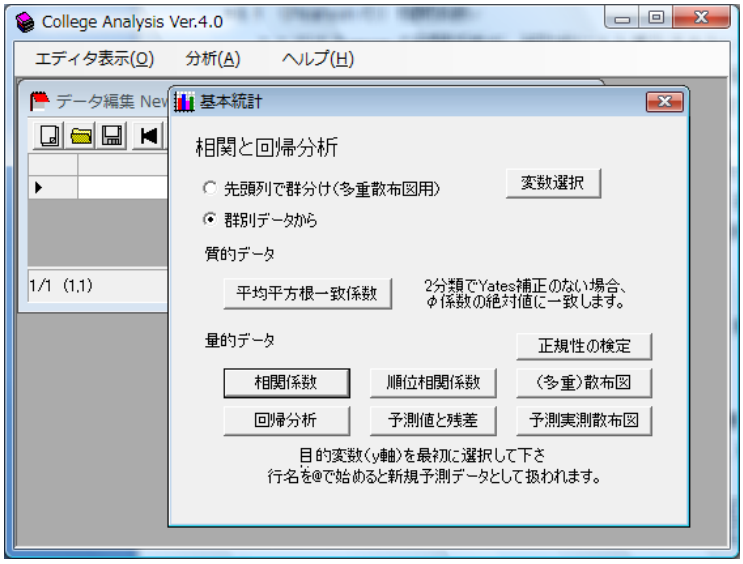


図 6.1.1 相関と回帰分析メニュー画面

最初にピアソンの相関係数の検定を行います。以下の例を見て下さい。

例

２つの商品 A, B の地域別使用率 (%) のデータは以下の通りである。それぞれの商品の使用率に線形の相関が認められるか。正規性を仮定して、有意水準 5%で検定せよ。

A(%)	33	24	30	50	42	15	15	56	13	45	44	21	18	31	27	40
B(%)	20	34	50	20	58	23	12	34	26	56	42	5	25	51	19	27

ファイル Samples¥テキスト 6.txt を開いて、最初のページですべての変数を選択し、「相関係数」コマンドボタンをクリックして、図 6.1.2 のような結果を得ます。

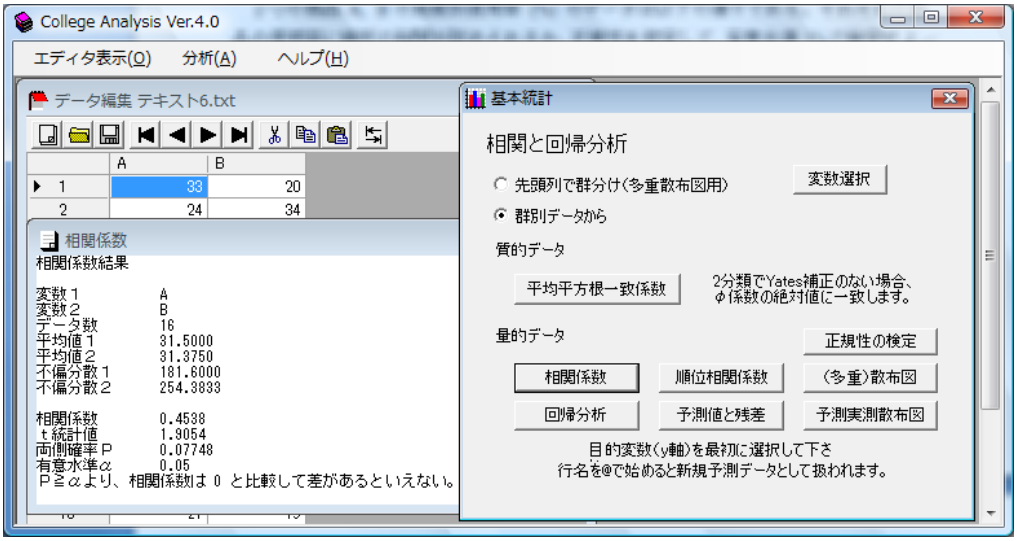


図 6.1.2 ピアソンの相関係数の検定画面

相関係数は 0.45 を超えていますが、ここではデータ数が少ないために相関係数は 0 と比べて有意差が見られません。もう少しデータを集める必要があるようです。

2 つの変数間に関係があると考えられるのは、ある程度相関係数の絶対値が大きく（必然性はありませんが、前に述べたように 0.5 程度から）、なおかつ検定結果に有意差が見られるような場合とする必要があります。ここで使った検定手法は以下の通りです。

理論

母相関係数を 0 と仮定して以下の性質を利用する。

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ 分布}$$

6.2 （Spearman の）順位相関係数

今度は前節のデータをもう一度利用して、正規性を仮定せずに相関を見てみましょう。

例

前節の問題で、それぞれの商品の使用率に相関（非線形のものも含む）が認められるか。正規性を仮定せずに、有意水準 5%で検定せよ。

「順位相関係数」ボタンをクリックすると図 6.1.3 の検定結果が得られます。

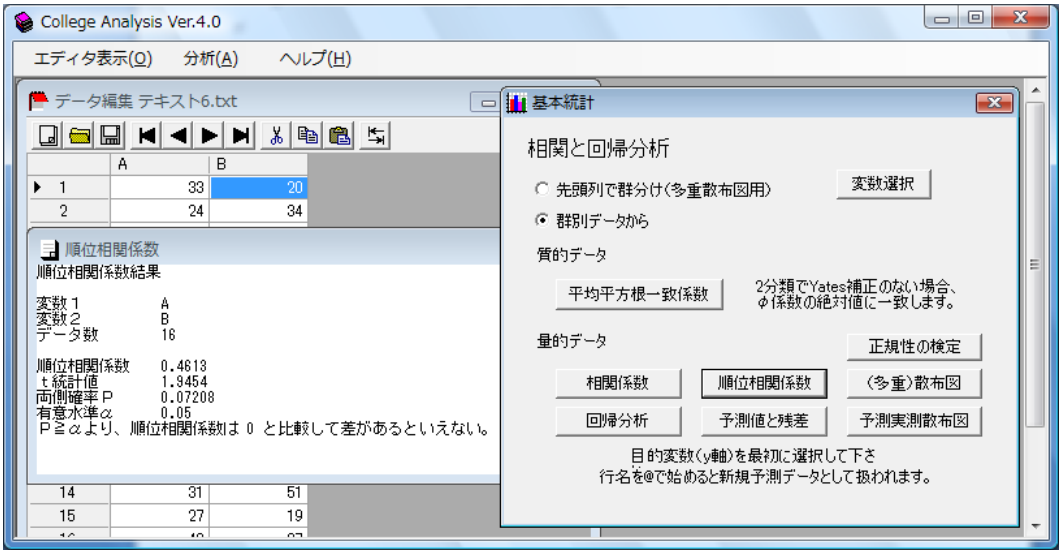


図 6.2.1 スピアマンの順位相関係数の検定画面

ここで使った検定手法は以下の通りです。

理論

順位相関係数 r_s を求め、母相関係数を 0 と仮定して以下の性質を用いる。

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2} \text{ 分布}$$

ところで本当はどちらの検定を考えるべきでしょうか。図 6.1.4 のように散布図を描き、図 6.1.5 のように正規性の検定を行います。

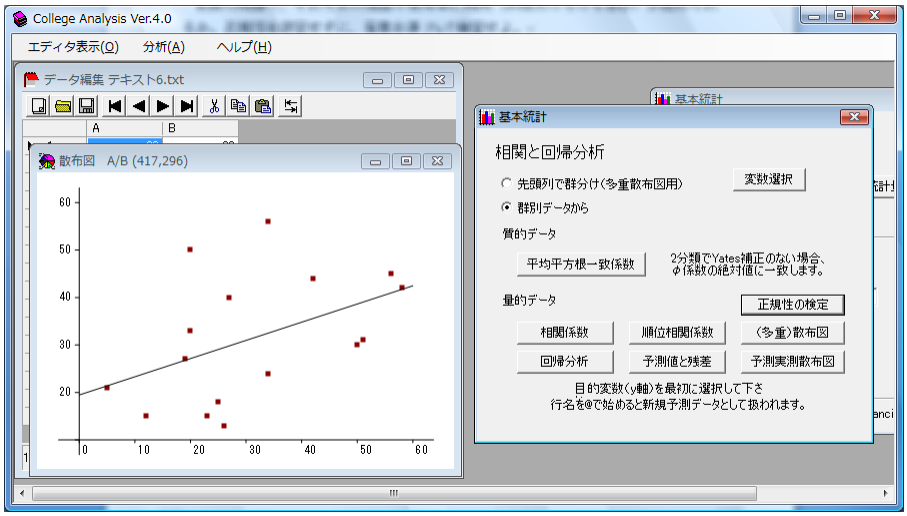


図 6.2.2 散布図の描画

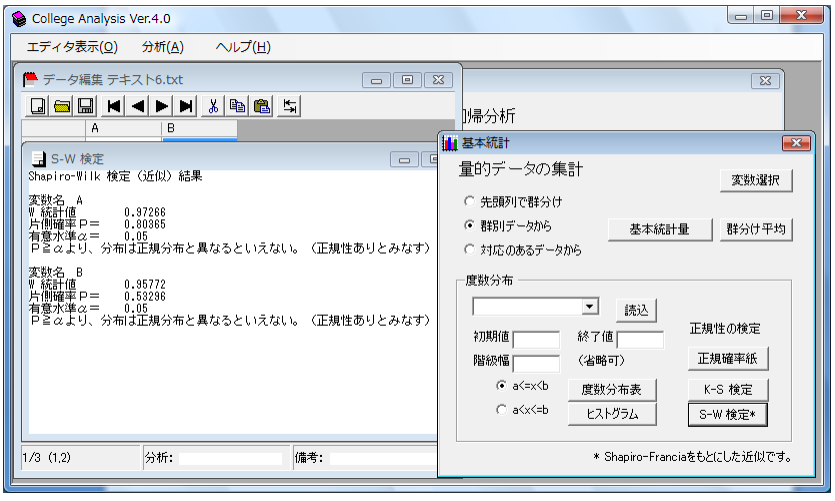


図 6.2.3 正規性の検定

これからピアソンの相関係数で調べることが分ります。

6.3 回帰分析

回帰分析の簡単な意味は、1.2 節で説明しましたが、ここでは回帰式の検定について話をします。回帰式には、回帰係数が 0 かどうかを調べる検定と回帰式の残差の変動と回帰式の変動（回帰変動）とを比べる回帰式の有効性の検定があります。回帰式の有効性の検定では、回帰変動が残差変動より有意に大きい場合に回帰式は有効であるいうことになります。これらの検定は一般には別物ですが、説明変数が 1 つの回帰式の場合、回帰直線の傾きを表す回帰係数の検定は回帰式の有効性の検定と一致します。またこれらの検定を利用するためには、残差の分布に正規性が必要なことは忘れられがちな性質です。それでは例を見てみましょう。

例

下の表のデータを用いて、身長により体重を推定する式を考える。ただし、式は 1 次式（体重 $= a \times$ 身長 $+ b$）と仮定し、その有効性を検討せよ。

体重	71	68	67	72	69	80	75	65	74	71
身長	169	175	170	179	176	174	173	181	179	178
体重	62	75	70	70	62	58	60	58	59	73
身長	170	180	177	175	172	166	168	173	169	170

Samples¥テキスト 6.txt のデータを開いて、目的変数、説明変数の順番に、体重、身長と変数選択し、図 6.1.1 のメニューで「回帰分析」ボタンをクリックします。結果は図 6.3.1 のように表示されます。

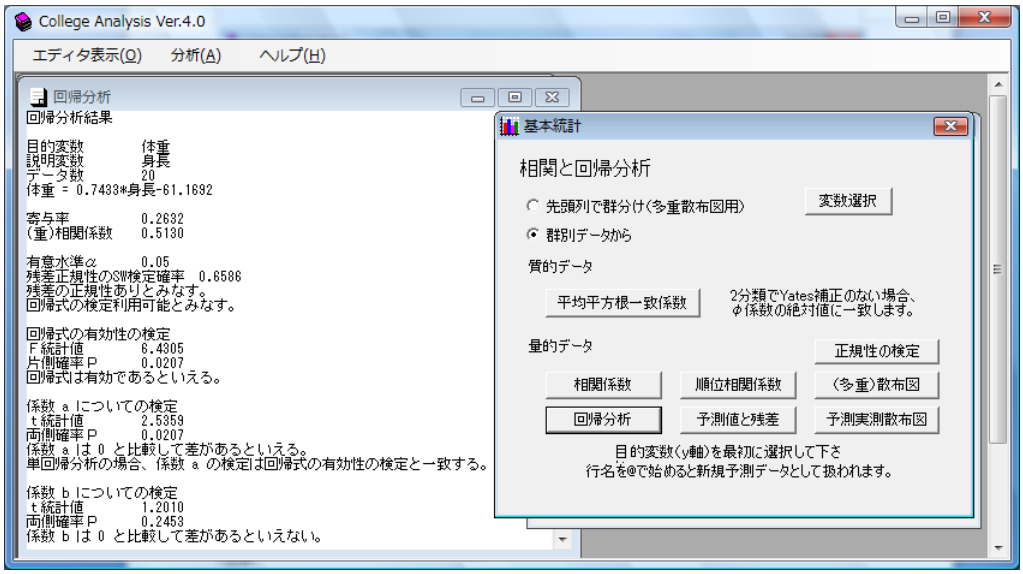


図 6.3.1 回帰分析結果

分析結果は、最初に回帰式が表示されます。次に重相関係数とその 2 乗である寄与率です。寄与率は決定係数とも呼ばれ、データの全変動のうちの何%が回帰変動で表されるかを示しています。そのためしばしば%で示されることがあります。また、重相関係数は目的変数の予測値と実測値の相関係数で、1 説明変数の回帰分析の場合は通常の相関係数に一致します。

次は、回帰式の有効性や回帰係数の検定ができるための、残差の正規性の検定結果が表示されています。ここでは数値的な S-W 検定（大まかな近似）の結果が参考のために示されています。残差の正規性が否定されなければ、回帰式の有効性や回帰係数の検定が利用できます。ここでは回帰式は有効であるという結論になっています。

ここで利用された理論は以下の通りです。

理論

回帰式の決定

2 変数の関係を、$y = ax + b$ の直線で表わすとすると、x を説明変数、y を目的変数と呼ぶ。データ点からこの直線へ垂直におろした線の長さの２乗が最小となるように係数 a, b を決める。

平均 $\bar{x}, \bar{y}$，標準偏差 u_x, u_y，相関係数 r とすると

$$a = r\frac{u_y}{u_x}, \quad b = \bar{y} - r\frac{u_y}{u_x}\bar{x}$$

回帰式の検討

重相関係数 R　　目的変数の実測値と回帰式による予測値の相関係数

（説明変数が１つの場合　$R = r$）

寄与率（重決定係数）R^2　　目的変数の変動のうち回帰式が説明する割合

回帰式の有効性の検定（残差が正規分布する場合のみ利用可能）

予測式を $Y = ax + b$ であり、残差は $\varepsilon \sim N(0, \sigma^2)$ 分布とする。

回帰係数の有効性の検定は、データ数 n，残差変動 $EV = \sum_{i=1}^{n} (y_i - Y_i)^2$，説明変数の不偏分散 u_x^2 として以下の関係を用いる。

$$t_a = \frac{a}{\sqrt{\frac{EV}{n-2} \Big/ (n-1)u_x^2}} \sim t_{n-2}\text{分布} \qquad t_b = \frac{b}{\sqrt{\frac{EV}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)u_x^2}\right)}} \sim t_{n-2}\text{分布}$$

問題１

以下の 2 変数のデータを用いて問いに答えよ。

変数1	65	86	78	83	85	89	83	80	85	93	75	85	79	80
変数2	162	210	224	179	217	230	223	204	224	197	186	189	172	185

１）２変数の正規性が判定困難として、Pearson の相関係数と Spearman の順位相関係数を両方を求めよ。

相関係数	順位相関係数

２）それぞれ相関係数が 0 と異なるかどうか有意水準 5%で判定する。

	確率	判定
相関係数		相関があると［いえる・いえない］
順位相関係数		相関があると［いえる・いえない］

３）変数２を目的変数、変数１を説明変数として回帰分析を行う。

回帰式　変数２＝［　　］×変数１＋［　　］

重相関係数　［　　］

寄与率　　［　　］

４）回帰分析の有効性の検定は［行える・行えない］。

検定確率［　　］

回帰式は有効であると［いえる・いえない］

問題２

Samples¥テキスト 9.txt のデータを用いて以下の問題に答えよ。

１）年収と支出についての相関係数と順位相関係数を求める。

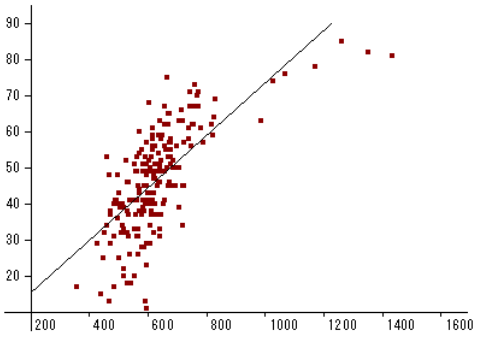
相関係数［　　］　順位相関係数［　　］

２）年収と支出に相関があるといえるか、相関係数を選んで有意水準 5%で判定する。

［相関係数・順位相関係数］で見る。

判定　確率［　　］　相関があると［いえる・いえない］

３）年収（横軸）と支出（縦軸）について以下のような散布図を描く。



４）支出を目的変数、年収を説明変数として回帰分析を行う。

回帰式　支出＝［　　］×年収＋［　　］

重相関係数　［　　］

寄与率　　［　　］

５）回帰分析の有効性の検定は［行える・行えない］。

検定確率［　　］

回帰式は有効であると［いえる・いえない］