College Analysis 総合マニュアル

― 多変量解析２ ―

目次

１２．時系列分析

12.1　時系列分析とは

時系列分析は時間の経過とともに変化する変数の、過去のデータから、未来の値を予測する手法である。例えば企業の売上、在庫の受注、株価の変動など時系列的に変化するデータがこの分析の対象である。

分析方法には大きく分けて、古くから考えられてきた予測モデルという方法とデータの変動をいくつかの典型的な変動に分解する変動の分解モデルという方法がある。予測モデルには、予測値にこれまでの変動の差分を使う差の平均法、過去のデータにウェイトを付けて使う指数平滑法やブラウン法、過去の最も似た状況を探す最近隣法、重回帰分析を活用するARIMA などがあるが、これらはデータ数が少なく周期性を見抜くことが困難なデータに適用されることが多い。

一方変動が周期性を持っているようなデータに対しては変動の分解モデルが適用される。これは変動を「傾向変動」、「季節変動」、「循環変動」、「残差」などに分け、それぞれの特徴をとらえて予測値を求めるもので、長期的な予測もある程度可能な手法である。傾向変動はデータの平均的な変動を表し、予測には移動平均や回帰分析を基礎とした近似モデルが利用される。一般に季節変動は周期が一定の変動で、循環変動は周期が変化する変動を表す。

本来予測モデルと変動の分解モデルは別々に考えられたものであるが、後者の傾向変動に例えば ARIMA の結果を利用するなどということも可能であるため、我々のプログラムでは２つの手法を組み合わせて使うことができるようにした。元々の変動の分解モデルの傾向変動については、移動平均や線形近似、対数近似などの近似手法が利用されることが多いので、傾向変動を２つに分けて、「傾向変動１」としてこれらの近似手法を、「傾向変動２」として先に述べた予測モデルを用いることにした。もちろんどちらか１つを選んでもよい。これらの分解の後、必要があればデータの周期的な変動の分解を行う。

周期的な変動には周期が一定の季節変動と周期が変動する循環変動があるが、循環変動についてはまだプログラムに組み込んでいない。また、季節変動を「振幅変動」と振幅が一定の「周期変動」の積に分解し、これらをまとめて以下のモデルとした。

データ変動＝傾向変動１＋傾向変動２＋振幅振動×周期変動＋残差

プログラムでは振幅変動の平均が１に近くなるように設定し、周期変動の意味を理解し易くしている。

12.2　プログラムの利用法

メニュー［分析－多変量解析他－時系列分析］を選択すると、時系列分析の実行画面が図１のように示される。

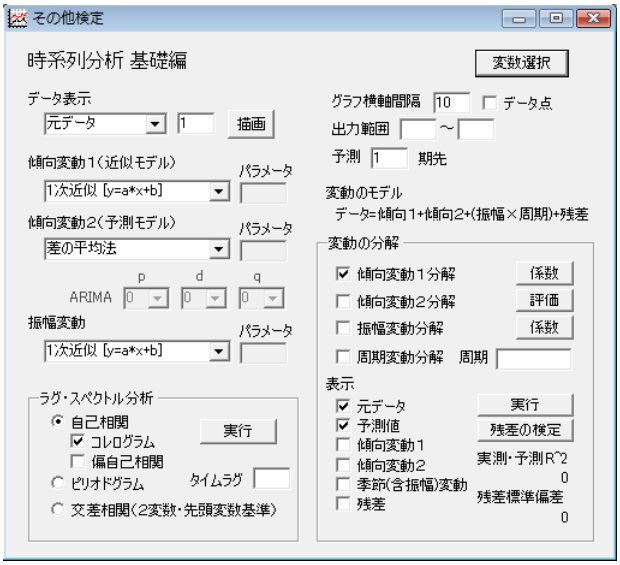


図１　時系列分析実行画面

以後は図２のデータ（時系列（decomp）.txt 2 頁目）を元にして、話を進める[4]。

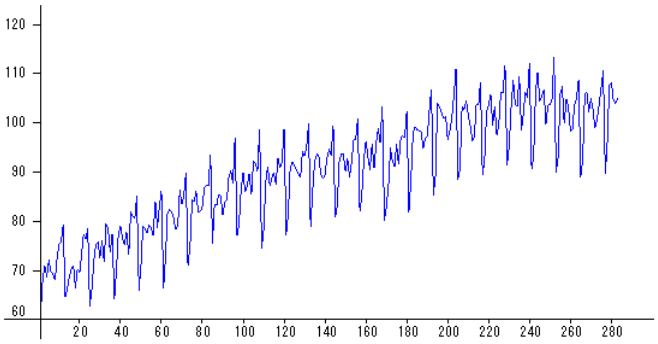


図２　時系列データ　decomp_food

このデータに対して、最初に述べたように、以下のモデルを考える。

データ変動＝傾向変動１＋周期変動＋残差

1）傾向変動の分解

傾向変動の抽出には、移動平均法による方法、最小２乗法を応用した関数の当てはめの方法（回帰分析はこれに含まれる）及び、局所回帰分析がある。移動平均法では 1 期先のデータをそれ以前の何期かのデータの平均として求める。最小２乗法を応用した近似手法の中には 1 次近似、対数近似、べき乗近似、指数近似、多項式近似等がある。このデータについては多項式近似の 2 次関数がよく当てはまる。具体的には以下の式で与えられる。

$$d_t = -0.0003t^2 + 0.2109t + 68.0017$$

この 2 次曲線を傾向変動とすると図 3 に示す結果となる。この段階での実測値と予測値の相関係数の 2 乗（決定係数）R^2 は 0.790 である。

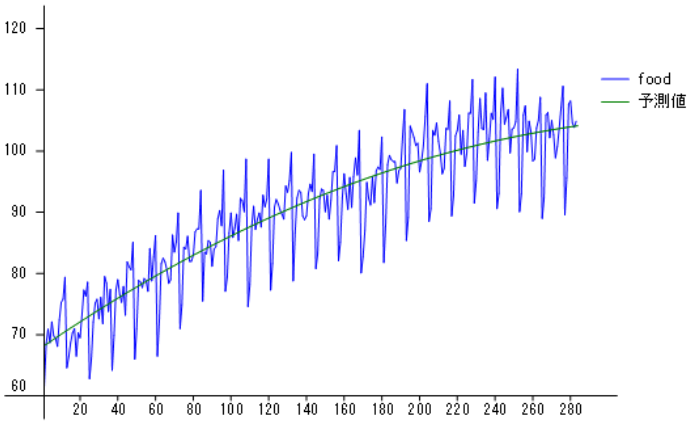


図３　２次式の傾向変動による予測

もう１つ傾向変動の分解に利用できる方法として、局所回帰分析がある。これは、予測したい位置の近くのデータにウェイトをかけた回帰分析である。ウェイトのかけ方はバンド幅という値によって調整されるが、通常利用されるのは、バンド幅が 0 から 1 の範囲が多い。バンド幅が小さい場合はウェイトの範囲は小さくなる。結果を図 4 に示す。

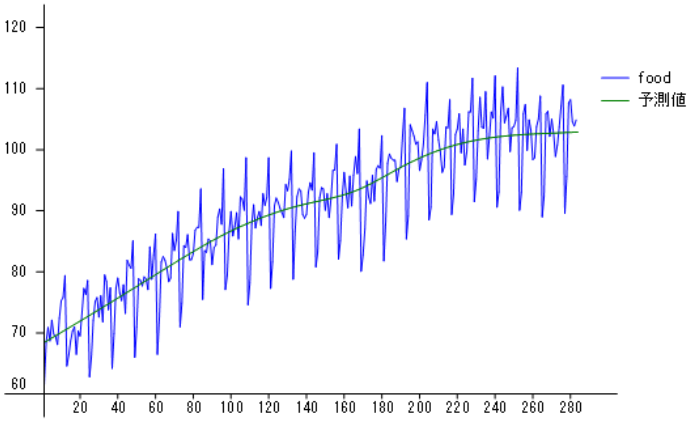


図４　局所回帰分析の傾向変動による予測

2）周期変動の分解

周期変動のラグ･スペクトル分析は傾向変動を除去したデータにどのような周期成分が含まれるかを知る重要な処理である。最初に時間的なラグ（遅れ）の影響を見るために自己相関係数を求め、ラグの値によってそれをプロットするコレログラムを作成する。結果を図 5 に示す。その際表示される表の結果によると 12 期ごとに相関の高い値が出ているので、変動の周期は 12 であることが分かる。

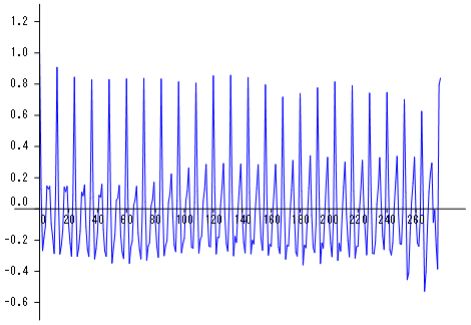


図５　コレログラム

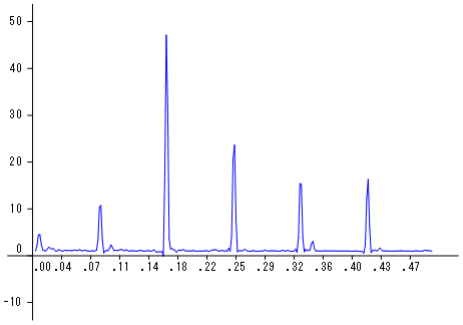


図６　ピリオドグラム

次に同じ処理をピリオドグラムというグラフで見ると周期性がより明確になる。結果を図6に示す。これを表で示される結果で詳細に見ると、まず周波数0.167（周期6）に大きなピークがあり、同様に周波数0.25（周期4）、周波数0.33（周期3）、周波数0.08（周期12）などにもピークがある。これらをまとめた周期はやはり12と考えられる。

このようにして求まる周期変動を加えた予測結果が図 7 である。この段階での実測値と予測値の R^2 は0.965である。

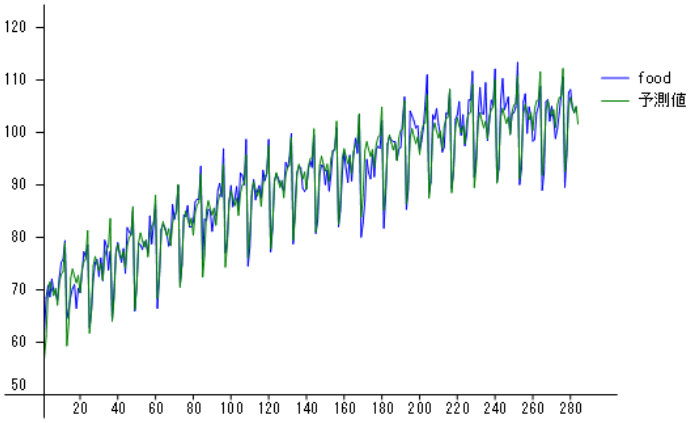


図７　季節変動の分解

さらに周期性を詳細にながめてみる。図 8 と図 9 にコレログラムとピリオドグラムを示す。コレログラムを見ると大きな周期の変動がありそうである。

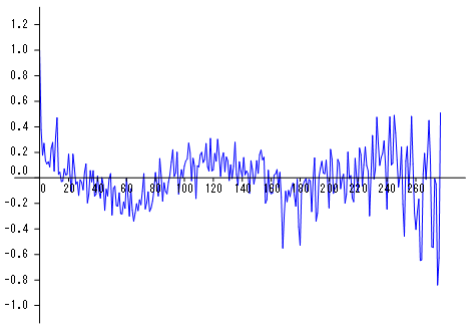


図８　残差のコレログラム

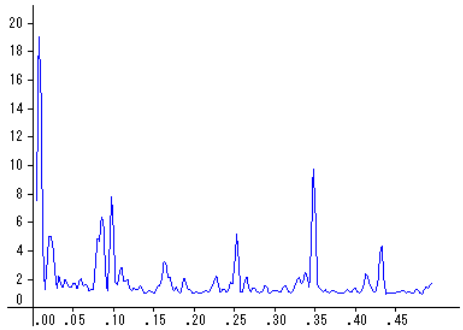


図９　残差のピリオドグラム

ピリオドグラムを見ると、0の近くにピークがあり、これは周期110近傍のピークである

ことが分かる。残差の標準偏差を最小にするように選んでやると、周期は 115 となる。そこでこの周期変動を加えて、最終的に図 10 のような予測になる。この場合の実測値と予測値の R^2 は 0.984 となる。

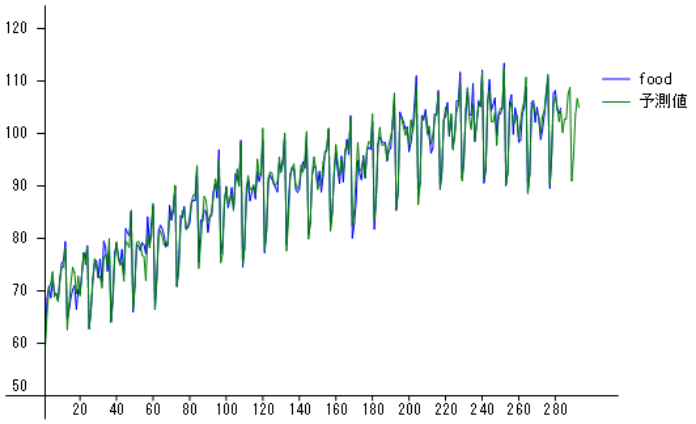


図 10　時系列データの予測

この手法によって、指定した先までの予測はできるが、これはデータの傾向のみに基づく分析であり、何らかの因果関係に基づく分析ではない。他の要素を考慮する分析にはパネル重回帰分析と呼ばれる手法がある。

問題１

時系列分析 1.txt の売上 2 について、以下の問いに答えよ。但し、ここでは周期変動２と振幅変動については考えないものとする。

１）このデータの傾向変動を１次式で推定するとどのような式になるか。
　売上＝［　　］×時間＋［　　］

２）上の傾向変動を除いた場合の残差標準偏差の値はいくらか。　［　　］

３）傾向変動を除いた残差から、ピリオドグラム等を用いて季節変動の周期を求めるといくらか。　［　　］

４）上の季節変動を除いた場合の残差標準偏差の値はいくらか。　［　　］

５）データを上の傾向変動と季節変動で予測するモデルの R^2 の値はいくらか。
　［　　］

６）このモデルでの１期先の予測値はいくらか。　［　　］

７）このモデルでの 5 期先の予測値はいくらか。　［　　］

問題２

時系列分析 1.txt の売上 4 について、以下の問いに答えよ。但し、ここでは周期変動２と振幅変動については考えないものとする。

１）このデータの傾向変動として、バンド幅 0.5 の局所重回帰分析を選んだときの残差標準偏差の値はいくらか。　残差標準偏差［　　］

２）傾向変動を除いたデータから、ピリオドグラム等を用いて季節変動の周期を求めるといくらか。　［　　］
３）傾向変動と季節変動を除いた残差標準偏差はいくらか。　［　　］
４）傾向変動と季節変動を除いたデータから、ピリオドグラム等を用いて再度季節変動（長周期の周期変動）の周期を求めるといくらか。　［　　］
５）上の変動をすべて除いた残差標準偏差はいくらか。　［　　］
６）データを上の傾向変動、季節変動、循環変動で予測するモデルの R^2 の値はいくらか。［　　］
７）このモデルでの１期先の予測値はいくらか。　［　　］
８）このモデルでの 10 期先の予測値はいくらか。　［　　］

問題１解答

１）このデータの傾向変動を１次式で推定するとどのような式になるか。
売上＝［ 0.511 ］×時間＋［ 51.19 ］
２）上の傾向変動を除いた場合の残差標準偏差の値はいくらか。　［ 0.75 ］
３）傾向変動を除いた残差から、ピリオドグラム等を用いて季節変動の周期を求めるといくらか。　［ 8 ］
４）上の季節変動を除いた場合の残差標準偏差の値はいくらか。　［ 0.58 ］
５）データを上の傾向変動と季節変動で予測するモデルの R^2 の値はいくらか。［ 0.983 ］
６）このモデルでの１期先の予測値はいくらか。　［ 65.98 ］
７）このモデルでの５期先の予測値はいくらか。　［ 69.00 ］

問題２解答

１）このデータの傾向変動として、バンド幅 0.5 の局所重回帰分析を選んだときの残差標準偏差の値はいくらか。　残差標準偏差［ 3.95 ］
２）傾向変動を除いたデータから、ピリオドグラム等を用いて季節変動の周期を求めるといくらか。　［ 4 ］
３）傾向変動と季節変動を除いた残差標準偏差はいくらか。　［ 1.78 ］
４）傾向変動と季節変動を除いたデータから、ピリオドグラム等を用いて再度季節変動（長周期の周期変動）の周期を求めるといくらか。　［ 50 ］
５）上の変動をすべて除いた残差標準偏差はいくらか。　［ 1.22 ］
６）データを上の傾向変動、季節変動、循環変動で予測するモデルの R^2 の値はいくらか。［ 0.984 ］
７）このモデルでの１期先の予測値はいくらか。　［ 42.08 ］
８）このモデルでの 10 期先の予測値はいくらか。　［ 47.74 ］

12.3　時系列分析のデータ

時間を過去から未来へ等間隔で区切ったとき、ある時点t（$t=1,\cdots,N$）でのある変数Xの値をx_tとする。時系列分析はこの変数の変化を分析し、モデルを作成して今後の予測を行うことを目的とする。以後このデータ書式を用いて予測モデルと変動の分解モデルの理論について説明する。

時系列分析では、データをそのままの形で使うより、何らかの変換を加えてから分析を進

める方がよりはっきりとした結果が得られることがある。ここではよく利用されるデータの変換について述べる。

変数が値の増大とともに変動の大きさも大きくなっていくような場合は、元の変数の対数をとって新しい変数とすると分析が容易になる場合がある。また、比率や確率のように $[0,1]$ 区間の値の場合は、以下のロジット変換によって値域が $(-\infty,\infty)$ の時系列に変換できる。

対数変換　　$z_t = \log_e x_t$

ロジット変換　　$z_t = \log_e\left(\dfrac{x_t}{1-x_t}\right)$

また、時系列データの差分を使って新しい変数を作り出すことも行われる。

差分（i期）　　$z_t = x_t - x_{t-i}$

差分比（i期）　　$z_t = x_t / x_{t-i}$

12.4　予測モデルの理論

時系列データの周期性が明らかでない場合やデータの数が周期性を見るのに十分でない場合、予測モデルと呼ばれる方法を用いて時系列データの予測が行われる。これからは図１のデータを用いて各種の予測モデルを紹介する[1]。

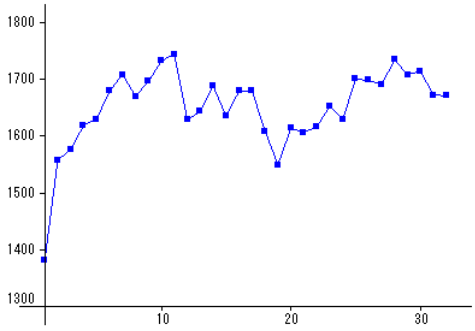


図１　時系列データ

これらのモデルは基本的にt時点までのデータを元に$t+1$時点での予測値を求めるもので、長期の予測には向かない。

1)　差の平均法（差分法）

これは$t+1$時点の予測値 y_{t+1} をt時点のデータ x_t とこれまでの２時点間の差分の平均で与えるものである。

$$y_{t+1} = x_t + A_t$$

ここに

$$A_t = \frac{(x_2 - x_1) + (x_3 - x_2) + \cdots + (x_t - x_{t-1})}{t-1} = \frac{x_t - x_1}{t-1}$$

差の平均法を用いた予測を図 2 に示す。これを見るとデータが上下している場合、残差の

平均は相殺され、予測値は 1 期前の値と余り変わらない様子が見える。この手法はデータに上昇傾向や下降傾向が見られる場合に適用できる。

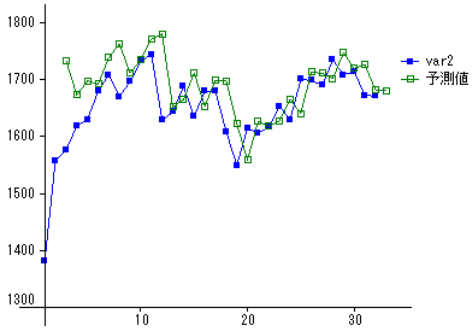


図 2　差の平均法を用いた予測

2 期以上の予測には実測値の代わりに予測値を使うことにすれば、予測は直線的に伸びて行く。

2)　指数平滑法

この方法は$t+1$期の予測値y_{t+1}をt期の実測値x_tと予測値y_tを使って以下のように与えるものである。

$$y_{t+1} = \alpha x_t + (1-\alpha) y_t \qquad \text{但し、} \ y_1 = x_1 \ （\text{または} \ y_2 = x_1）とする。$$

ここにαは$0 < \alpha < 1$のパラメータである。またこの式は以下のように書き換えると、指数平滑の意味が分かり易い。

$$y_{t+1} = \alpha x_t + (1-\alpha)[\alpha x_{t-1} + (1-\alpha) y_{t-1}]$$

$$\cdots$$

$$= \alpha x_t + \alpha(1-\alpha) x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots + \alpha(1-\alpha)^{t-2} x_2 + (1-\alpha)^{t-1} x_1$$

これを見るとαの値が小さいほど過去からの影響を受けやすくなっていることが分かる。これは今期以前の指数平滑値を次期の予測値とするものである。この方法を用いて時系列データの変動をα=0.74 として予測した結果を図 3 に示す。パラメータの値は図 4 のようにパラメータの値を変えて残差の平均を調べ、最小値をとることによって求めた。

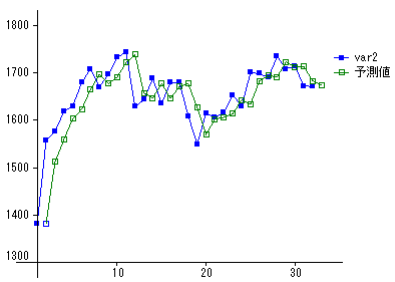


図 3　指数平滑法による予測（α=0.74）

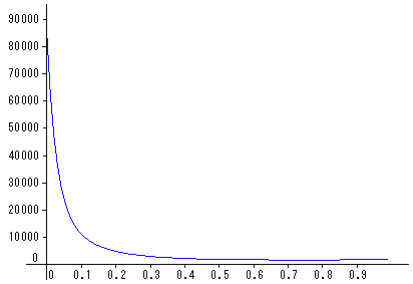


図 4　パラメータの推定

差の平均法と同様この場合も変動は平滑され、予測値は 1 期前の実測値に近い値になっている。また 2 期以上先の予測値は、実測データを予測データで置き換えると同じ値が続く。

この予測値を見ると 1 期前の実測値にかなり引きずられていることが分かる。指数平滑法も上がり下がりのあるデータには向かない。

3)　ブラウン法（ブラウンの 2 重指数平滑法）

指数平滑法は単純に今期までの指数平滑値を予測値としたものであって、予測値の精度については考慮されていない。この精度を考慮した方法がブラウン法である。

ここで比較のために指数平滑法の公式を少し書き換えておく。

$y_{t+1} = u_t$

$u_t = \alpha x_t + (1-\alpha)u_{t-1}$　　t時点の x の指数平滑値（$t+1$時点の x の予想値）

ブラウン法は、指数平滑法で予測される$t+1$期の予測値u_tに、この予測値と指数平滑法によるu_tの予測値v_{t-1}との差（のm'倍）を足して来期を予測するものである。指数平滑を 2 度行うので 2 重指数平滑法と呼ばれる。

$y_{t+1} = u_t + m'(u_t - v_{t-1})$

$u_t = \alpha x_t + (1-\alpha)u_{t-1}$　　t時点の u の値（$t+1$時点の x の予想値）

$v_{t-1} = \beta u_{t-1} + (1-\beta)v_{t-2}$　　$t-1$時点のvの値（t時点の u の予想値）

ここにm, α, βはパラメータである。

この式を分かり易く表現すると以下となる。

x の補正予測値＝$t+1$時点の x の予測値＋m'(t時点の u の値－t時点の u の予測値)

＝$t+1$時点の x の予測値＋$t+1$時点の予測補正項

実際の計算では、参考文献 1 に従い、$m'=1$，$\alpha=\beta$ としており、

$$y_{t+1} = a_t + b_t,\quad a_t = 2u_t - v_t,\quad b_t = \frac{\alpha}{1-\alpha}(u_t - v_t)$$

以下の初期値をおいている。

$u_1 = v_1 = x_1$，　$a_1 = x_1$，$b_1 = [(x_2 - x_1) + (x_4 - x_3)]/2$

このため予測値は、$t=5$から求める。

ブラウン法による最適なパラメータでの予測を図 5 に示す。ここでも明らかなように増加・減少のあるデータに対してブラウン法はあまり有効とは言えない。

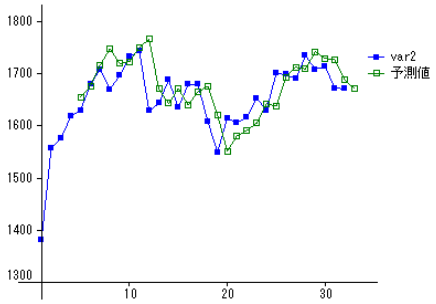


図 5　ブラウン法による予測（α=0.42）

4)　最近隣法

最近隣法は現在とその 1 期前のデータに似た過去のデータを探して、次期のデータの予測値を決めるものである。

最近隣法は以下の形で予測を行う。現在とその 1 期前のデータを x_t, x_{t-1} とし、過去のデータ x_{t-m}, x_{t-m-1} との距離 d_m を以下のように考える。

$$d_m = \sqrt{(x_t - x_{t-m})^2 + (x_{t-1} - x_{t-m-1})^2}$$

距離の最小値 $d_{\min}$ を求め、距離がその 1.62 倍未満のデータを集める。

$$S = \{d_m \mid d_m < 1.62 \times d_{\min}\}$$

この 1.62 は黄金分割比と呼ばれ、実用上多く使われる。その集めた距離の逆数を利用して重み w_m（$d_m \in S$）計算する。但し、距離が 0 の場合はある小さな値（このソフトの場合は 0.0001）としている。

$$w_m = \frac{1/d_m}{\sum_{d_k \in S} 1/d_k}$$

この重みを使って予測値 y_{t+1} を以下のように求める。

$$y_{t+1} = \sum_{d_m \in S} w_m x_{m+1}$$

実際に最近隣法を用いた予測は図 6 のようになる。

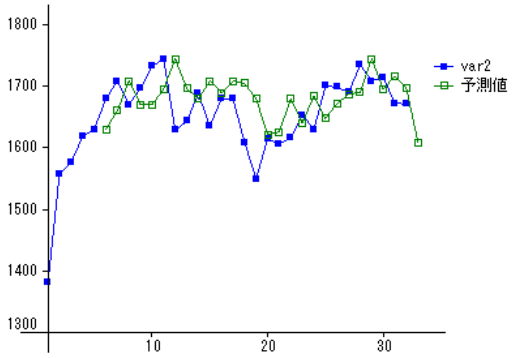


図 6　最近隣法による予測

この方法はデータの上がり下がりの変動が大きいほど有効で、上昇下降傾向があるデータには向かない。また過去の似た状況を探すことから、一般に過去のデータが多いほど予測の精度は上がる。

5)　自己相関モデル（ARIMA）

このモデルには 3 つのパラメータ p, d, q があり、これらのパラメータを用いて、ARIMA(p, d, q) と表される。以後各パラメータについて説明し、最後に全体を見渡す。

最初にパラメータ d について述べる。これはデータの差分の回数である。差分は傾向変動などを取り除く 1 つの手段である。$x_t^{(1)}$ を 1 回の差分、$x_t^{(2)}$ を 2 回の差分とするとそれぞれ元のデータを用いて以下のように表される。

$$x_t^{(1)} = x_t - x_{t-1}$$
$$x_t^{(2)} = x_t^{(1)} - x_{t-1}^{(1)} = x_t - 2x_{t-1} + x_{t-2}$$

d回の差分データに対して ARMA(p, q)モデルを適用する手法が ARIMA(p, d, q) モデルである。但し、d回の差分データでは利用できるデータが、d+1 期から t期までとなる。

MA モデル

次にパラメータ q について考える。このパラメータは MA(q)と呼ばれるモデルのパラメータである。このモデルは$t \geq t_0$に対して以下の仮定が基礎になっている。

$$x_t = b_1 u_{t-1} + b_2 u_{t-2} + \cdots + b_q u_{t-q} + b_0 + u_t$$

ここに$u_t, u_{t-1}, \cdots, u_{t-q}$は各時点のホワイトノイズである。特に$b_0 = 0$の場合が教科書などに載っている。

1 期先の予測値y_{t+1}を実測値x_{t+1}からホワイトノイズu_{t+1}を引いたものと定義すると以下のような関係が得られる。

$$\begin{aligned} y_{t+1} &= x_{t+1} - u_{t+1} \\ &= b_1 u_t + b_2 u_{t-1} + \cdots + b_q u_{t-q+1} + b_0 - u_{t+1} + u_{t+1} \\ &= b_1(x_t - y_t) + b_2(x_{t-1} - y_{t-1}) + \cdots + b_q(x_{t-q+1} - y_{t-q+1}) + b_0 \end{aligned}$$

計算手順はまず$t < t_0$の間のノイズ$x_t - y_t$の初期値を決める。我々はこれを $N(0,1)$ の正規乱数としている。次にこれらの初期値を用いて$t = t_0$の場合に上式から重回帰分析を用いて予測値y_{t_0+1}求める。但し、計算が可能なためには初項の時期をずらしたデータの組が q 個必要であり、少なくとも$t_0 > 2q$でなければならない。我々はこれを$t_0 = 2q + 2$にしている。ここで得た予測値y_{t_0+1}を使って、上式を用いて再度重回帰分析を行うことによって新しい予測値y_{t_0+2}を得る。これを繰り返して行くことで、最終的な予測値y_{t+1}を得る。

この処理では長期予測は不可能である。長期予測のためには実測値の代わりに予測値を用いるしかないが、そうすると説明変数が 0 になって行き、前の予測値が続くようになる。

MA(1) と MA(2) による予測グラフを図 7a と図 7b に示す。

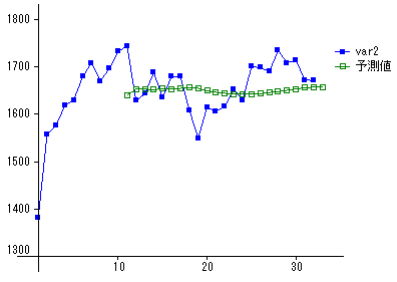


図 7a　MA(1) モデルによる予測

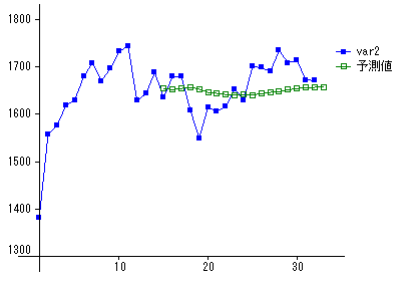


図 7b　MA(2) モデルによる予測

AR モデル

パラメータ p は AR(p) と呼ばれるモデルのパラメータである。このモデルは以下の仮定が基礎になっている。

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + a_0 + u_t$$

ここにu_tはt時点のホワイトノイズである。特に$a_0 = 0$の場合が教科書などによく載っている。

1 期先の予測値y_{t+1}を実測値x_{t+1}からホワイトノイズu_{t+1}を引いたものと定義すると$t \geq t_0$に対して以下のような関係が得られる。

$$y_{t+1} = a_1 x_t + a_2 x_{t-1} + \cdots + a_p x_{t-p+1} + a_0$$

計算は重回帰分析を用いるが、手順は過去の予測値を使う必要がないので MA モデルと比べると単純である。但し、計算が可能なためには初項の時期をずらしたデータの組が p 個必要であり、少なくとも$t_0 > 2p$でなければならない。我々はこれを$t_0 = 2p + 2$にしている。

この処理でも長期予測は不可能である。長期予測のためには実測値の代わりに予測値を用いるしかないが、a_iが殆ど変わらない状況では例えば$p = 1$，$|a_1| < 1$の場合、

$$y_n = a_1 y_{n-1} + a_0, \quad \lim_{n\to\infty} y_n = a_0/(1 - a_1)$$

となり、前の予測値に近い値が続くようになる。AR(1)と AR(2) による予測グラフをそれぞれ図 8a と図 8b に示す。

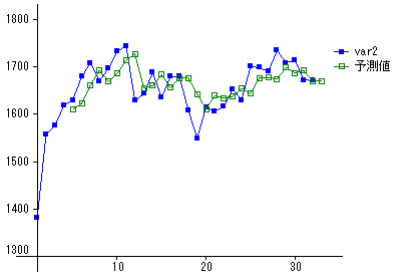


図 8a　AR(1) モデルによる予測

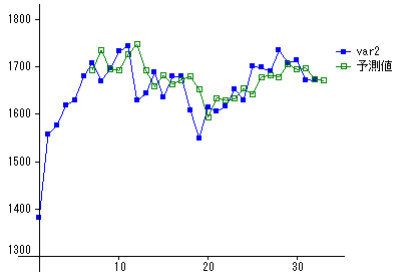


図 8b　AR(2) モデルによる予測

ARIMA モデル

ここではこれまで学んできたモデルを複合した場合を考える。今 d 回の差分データを$x_t^{(d)}$とすると、ARIMA(p, d, q) モデルは$t \geq t_0$で以下のように表される。

$$x_t^{(d)} = \sum_{i=1}^{p} a_i x_{t-i}^{(d)} + \sum_{i=1}^{q} b_i u_{t-i}^{(d)} + c + u_t$$

これを用いて予測値$y_{t+1}^{(d)}$は以下のようになる。

$$y_{t+1}^{(d)} = \sum_{i=1}^{p} a_i x_{t-i+1}^{(d)} + \sum_{i=1}^{q} b_i (x_{t-i+1}^{(d)} - y_{t-i+1}^{(d)}) + c$$

計算手順は、まず$t < t_0$以前のノイズ$x_t^{(d)} - y_t^{(d)}$を標準正規乱数で初期化する。後は MA モデルの場合と同様に、$t = t_0$の場合の予測値$y_{t_0+1}^{(d)}$を重回帰分析で求めて、これを利用してさらに次の予測値を求める方法をとる。但し計算が可能なためには、上式に必要なデータが$r = \max(p, q)$個、それを時期をずらして $p + q$ 期分必要であることから、少なくとも$t_0 > r + p + q + d$でなければならない。我々は少し大きくとって、以下としている。

$$t_0 = r + p + q + d + 2$$

計算が可能であることで上のような条件を付けたが、計算の正確さを考えると十分でない。MA モデルでは計算の初期値を乱数で与えているので、t_0 の近くの推定値は良い近似ではない。我々は値が安定するまで待つ必要がある。そのため、誤差の計算や表示に利用するのは実際には経験的に以下にしている。

$t_0 = 2p + d + 2$　　$q = 0$ の場合

$t_0 = (r + p + q + d + 2) + (2q + 5)$　　$q > 0$ の場合

これで $t_0 + 1$ 期からの予測値 $y_{t+1}^{(d)}$ が求められたが、これは差分を d 回取ったデータの予測値である。我々はこれを元のデータに戻す必要がある。データ間に

$$x_{t+1}^{(d-1)} = x_t^{(d-1)} + x_{t+1}^{(d)}$$

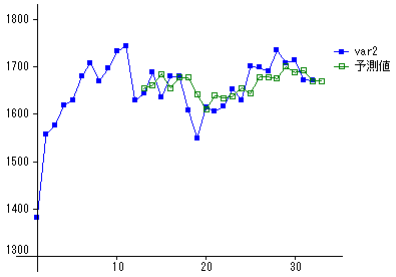
の関係があることから、これを以下のように拡張する。

$$y_{t+1}^{(d-1)} = x_t^{(d-1)} + y_{t+1}^{(d)}$$

即ち、以下のように求められる。

$$y_{t+1} = y_{t+1}^{(0)} = x_t^{(0)} + y_{t+1}^{(1)} = x_t^{(0)} + x_t^{(1)} + y_{t+1}^{(2)} = \cdots = \sum_{i=0}^{d-1} x_t^{(i)} + y_{t+1}^{(d)}$$

ARIMA(1,0,1) , ARIMA(1,1,1) による予測グラフを図 9a と図 9b に示す。



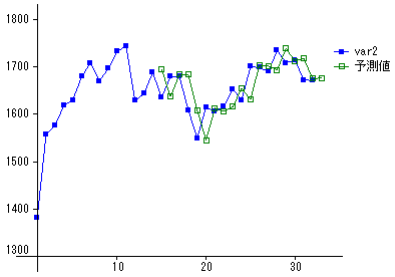


図 9a　ARIMA(1,0,1) モデルによる予測　図 9b　ARIMA(1,1,1) モデルによる予測

差分を入れると 1 期前の実測値に差分の予測値を足すことになり、やはり 1 期前の状態に引きずられるようである。

12.5　変動の分解モデルの理論

具体的なイメージを持ってもらうために、今後しばらく図 1 のデータを元にして話を進める[2],[4]。

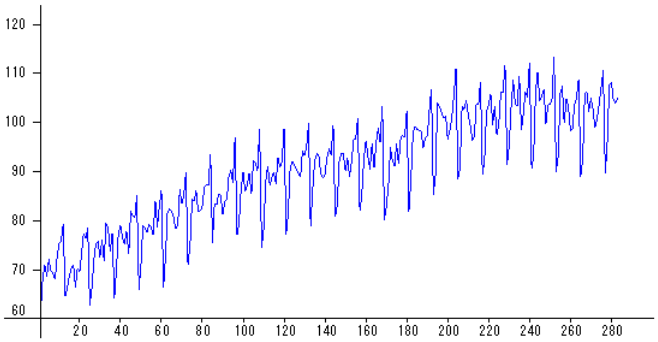


図 1　時系列データ decomp_food

データは様々な要因で変動するが、我々は大きくこれを、傾向変動T、季節変動S、循環変動C、残差変動Rに分ける。ここに傾向変動は長期にわたる継続的な変化で、季節変動は周期が一定の変化、循環変動は周期が一定でないものの周期性が認められる変化、残差変動は観測誤差などのゆらぎである。一般に変数Xはこれらの変動の関数として以下のように表される。

$$X = f(T, S, C, R)$$

この一般の関係の中で、実際の分析のためには様々な仮定を置くことが多い。我々のプログラムでは周期が変化する循環変動については考えず、それぞれの変動の合計で表される以下の加法モデルを採用している

$$X = T + S + R$$

但し、傾向変動には通常、移動平均や回帰近似が利用されるが（これを近似モデルと呼ぶ）、我々は傾向変動を２つに分け、近似モデルT_1と５節で述べた予測モデルT_2の和と考える。これによって予測モデルだけの処理も変動の分解モデルと合わせた処理も可能になる。また季節変動について、振幅の変化も考え、季節変動を振幅変動Aと振幅一定の季節変動S'（以後これを周期変動と呼ぶ）の積に分解する。ここで振幅変動には回帰近似を用い、大きさの平均を１に近くなるようにとる。これらを合わせて、我々のプログラムでは以下のようなモデルを扱う。

$$X = T_1 + T_2 + A \times S' + R$$

以後予測モデルT_2を除いて、それぞれの変動の分解について詳細に説明する。

1） 傾向変動の分解

傾向変動の抽出は主に移動平均法による方法と最小２乗法の手法を応用した方法（回帰分析はこれに含まれる）がある。n期の移動平均法では時点tのデータの値を以下のようにして、データの平滑化を行う。

$$d_t = \frac{1}{2m+1}\sum_{i=-m}^{m} x_{t+i} \qquad n = 2m+1 \text{の場合}$$

$$d_t = \frac{1}{2m+2}\left\{\sum_{i=-m}^{m} x_{t+i} + \frac{1}{2}(x_{t-m-1} + x_{t+m+1})\right\} \qquad n = 2m+2 \text{の場合}$$

これは中心法と呼ばれる方法であるが、移動平均を予測に用いる場合には、以下のような方法が使われる。我々はこの方法を用いる。

$$d_t = \frac{1}{n}\sum_{i=-n}^{-1} x_{t+i}$$

また、時間のずれに対して重み係数を掛ける場合もある。データに周期性がある場合、この方法では傾向変動に周期成分が残るが、移動平均を行ったデータに再度移動平均を行うとさらになめらかな傾向が得られる。但し、移動平均では時系列データの前後、または前が使えなくなるので、ある程度データ数も必要である。我々のプログラムでは複数回の移動平均は考えていない。

予め大雑把なデータの変化を近似的につかんでおくことは重要である。最小 2 乗法の手法を応用した近似手法の中で線形回帰分析を利用するものは計算が容易である。よく使われる線形回帰の方法には以下のようなものがある。

1 次近似　　$d_t = at + b$

対数近似　　$d_t = a\log t + b$

べき乗近似　　$d_t = bt^a$

指数近似　　$d_t = be^{at}$

多項式近似　　$d_t = a_p t^p + a_{p-1}t^{p-1} + \cdots + a_1 t + a_0$

ここにべき乗近似と指数近似については両辺の対数をとって線形回帰分析を行う。また、多項式近似は重回帰分析を用いてパラメータの推定を行う。例として 2 次式による近似結果を図 2 に示す。

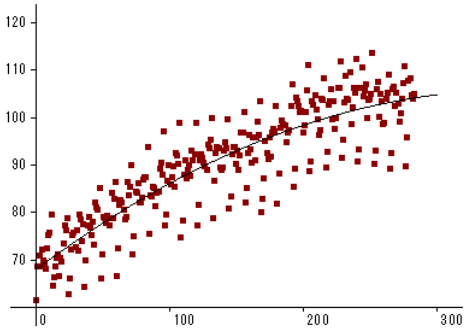


図 2　2 次曲線の当てはめ

このデータについては以下の 2 次曲線が最良である。

$$y = -0.00029t^2 + 0.211t + 68.002$$

これら以外の近似には非線形最小 2 乗法など他の方法を利用する。

この傾向変動の結果を元データから分離するには、我々のモデルでは引き算を用いる。

$$y_t = x_t - d_t$$

この 2 次曲線を傾向変動として取り除くと図 3 の結果となる。この段階での実測値と予測値の相関係数の 2 乗（決定係数）R^2 は 0.790 である。

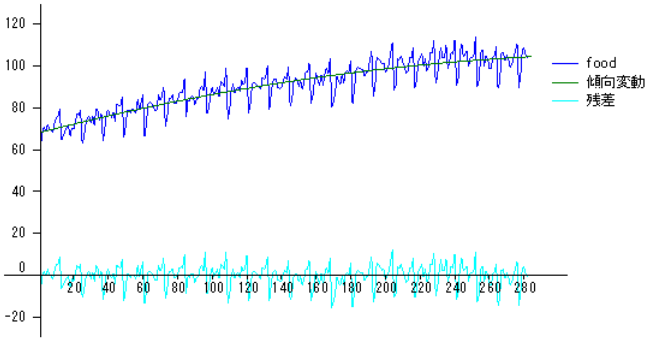


図 3　傾向変動の分離

もう 1 つ傾向変動の分解に利用できる方法として、局所回帰分析が考えられる。これは、

ウェイトをかけた回帰分析である。予測したい点を要求点として、その近傍に大きなウェイトをかけ、それから離れるに従ってウェイトを小さくする。これにより、関数形を定めることなく、非線形の予測を行うことができる。ウェイトの範囲はバンド幅と呼ばれる値によって決めることができるが、バンド幅が 100 以上の場合はほぼ完全に線形回帰分析となる。通常利用されるのは、バンド幅が 0 から 1 の範囲が多い。

予測モデルの分解については、前節で述べたので省略する。

2) 振幅変動の分解

振幅変動の推定は以下の振幅変動データに対して近似曲線を考えることによって与えることにする。

振幅変動データ＝傾向変動の残差の絶対値

÷傾向変動の残差の絶対値の平均値

これによって振幅変動の値はほぼ 1 に近い値となり、周期変動を平均的な振幅を持つ季節変動と意味付けることができるようになる。図 4 に近似直線を求める図を示す。振幅変動を分離した残差は傾向変動残差÷振幅変動推定値で与えられる。

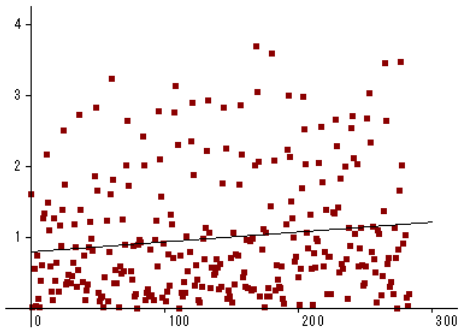


図 4　振幅変動の推定

3) 周期変動の分解

周期変動のスペクトル抽出は傾向変動と振幅変動を除去したデータ y_t にどのような周波数成分が含まれるかを知る重要な処理である。最初に時間的なラグの影響を見るために自己相関係数を求め、ラグの値によってそれをプロットするコレログラムを作成する。

自己相関係数 r_k （$k=1,2,\cdots,L<N-1$）は以下の式により求められる。

$$r_k=\frac{s_k^2}{s_0^2},\qquad \text{ここに}\ \ s_k^2=\frac{1}{N-k}\sum_{t=k+1}^{N}(x_t-\bar{x}_{k+1}^{N})(x_{t-k}-\bar{x}_1^{N-k}),\quad \bar{x}_a^b=\sum_{t=a}^{b}x_t$$

図 5 に最大周期を 70 にしたコレログラムを示す。これによると変動の周期は 12 であることが分かる。

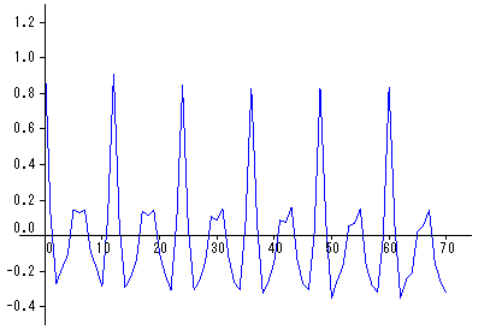


図５　コレログラム

次にこのコレログラムに対してその周波数成分を見ると周期性がより明確になる。このような問題には関数のフーリエ（Fourier）展開の手法が用いられるが、ここでは参考のために期間$2L$を周期に持つ関数$f(x)$のフーリエ展開の公式を与えておく。

$$f(x)=\frac{a_0}{2L}+\frac{1}{L}\sum_{k=1}^{\infty}\left(a_k\cos k\pi x/L+b_k\sin k\pi x/L\right)$$

$$a_k=\int_{-L}^{L}f(x)\cos k\pi x/L\,dx,\qquad b_k=\int_{-L}^{L}f(x)\sin k\pi x/L\,dx$$

この式は関数を周波数$f_k=k/2L$（$k=1,2,3,\cdots$）の正弦波成分の合計で表したもので、各成分の強さは係数a_kとb_kで与えられる。

我々の時系列データでは関数が離散的であるため、離散フーリエ変換という手法を利用する。nを時系列データx_tの周期として、離散フーリエ展開の公式を以下に与える。

$$x_t=\frac{1}{n}\sum_{k=0}^{n-1}(a_k\cos 2\pi kt/n+b_k\sin 2\pi kt/n)\qquad(1)$$

$$a_k=\sum_{t=1}^{n}x_t\cos 2\pi kt/n,\qquad b_k=\sum_{t=1}^{n}x_t\sin 2\pi kt/n$$

この公式を自己相関係数r_iに対して適用する。自己相関係数は$r_i=r_{-i}$であるため、$-m\le i<m$の範囲で偶関数である。その際には周期を$2m$として、以下の形で与えられる。

$$r_i=\frac{1}{2m}\sum_{k=-m}^{m-1}(a_k\cos 2\pi k/2m+b_k\sin 2\pi kt/2m)=\frac{1}{m}\sum_{k=0}^{m-1}a_k\cos\pi kt/m$$

$$a_k=\sum_{i=-m}^{m-1}r_i\cos 2\pi ki/2m=2\sum_{i=0}^{m-1}r_i\cos\pi ki/m$$

この量a_kを周波数$f_k=k/2m$の生スペクトルと呼び、これをラグごとに表したグラフをピリオドグラムという。実用上は生スペクトルより、平滑化という処理を行ったピリオドグラムがよく用いられる[2]。

実際のデータに対する平滑化したピリオドグラムを図６に示す。

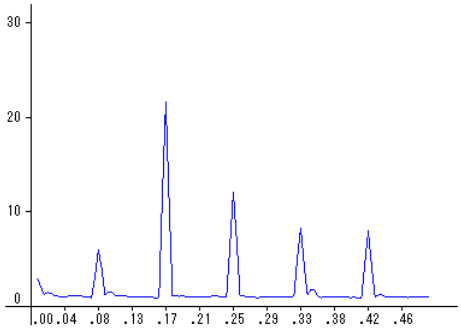


図 6　ピリオドグラム

これを詳細に見るとまず、周波数 0.167（周期 6：これらは別に表示されるデータから読み取れる）に大きなピークがあり、同様に周波数 0.25（周期 4)、周波数 0.33（周期 3)、周波数 0.08（周期 12）などにもピークがある。これらの全体的な周期は、ここに現れた周期の重ね合わせ（最小公倍数、但し時系列の長さの半分より小さいこと）と考えると周期 12 である。

この変動の分離には一般の離散フーリエ変換の式 (1) を利用するが、上で考えた周期を n として残差 y_t に適用し、周期変動 u_t を得る。

$$u_t = \frac{1}{n}\sum_{k=0}^{n-1}(a_k \cos 2\pi kt/n + b_k \sin 2\pi kt/n)$$

$$a_k = \sum_{t=1}^{n} y_t \cos 2\pi kt/n\,, \qquad b_k = \sum_{t=1}^{n} y_t \sin 2\pi kt/n$$

時系列のデータには周期性があると言っても、各周期間には揺らぎが見られる。しかし上の計算では時系列中どの１周期を考えればよいのか分からない。そこで実際の計算には特定の１周期を選ぶのではなく、各周期中の同一時点の残差の平均 $\bar{y}_t$ を用いて計算を行った。

このようにして季節変動を除去した結果が図 7 である。ここでは除去した季節変動と残差のみ示してある。この段階での実測値と予測値の R^2 は 0.9647 である。

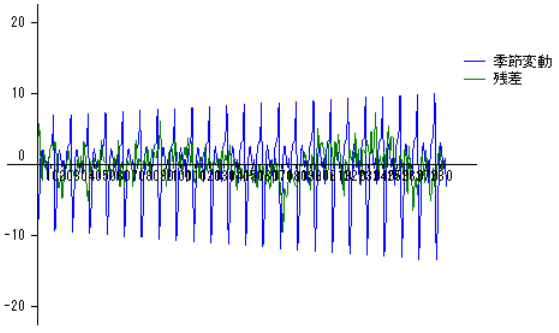


図 7　季節変動の分解

もう少し詳細に残差の周波数をながめて（タイムラグ 200 まで）図 8 でピリオドグラムを描いてみる。

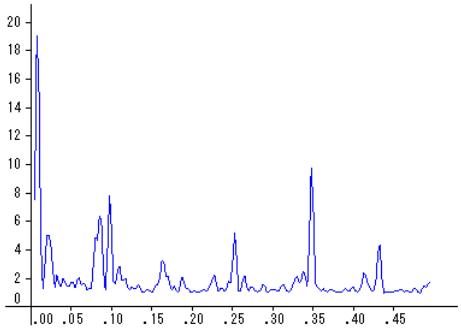


図８　残差のピリオドグラム

これを見ると、0 の近くにピークがあり、これは周期 130 近傍のピークであることが分かる。残差の標準偏差を最小にするように選んでやると、周期は 129 となる。そこでこの周期変動を差し引いて、最終的に図 9 の分解になる。最終的な実測値と予測値の R^2 は 0.9838 となる。振幅変動を分離しない場合の R^2 は 0.9830 であり、この場合振幅変動の分解の効果はわずかである。

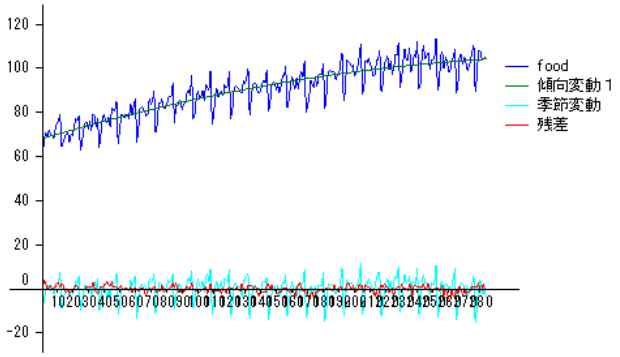


図９　時系列データの分解

実はこの残差にはまだ周期性が残っており、これに対して周期性の分離を行い、さらに残差を小さくできる。実際、例えば 91,90,41 と周期性を取り除いていくと実測値と予測値の R^2 は 0.9941 と大きくできる。これを見ると予測精度が上がっているように思われるが、すでに周期成分 129 を入れているのでこのデータの数 283 個から見れば、わずか 2 周期分を用いて予測を行っていることになる。3 周期目はそれ以前と少しずれることを考えると、いくら残差が小さくできたからといって予測が正しくなる保証はない。ある程度のところで止めておくべきであろう。

さて分解がうまくいき、これ以上分解が難しくなる場合もある。そのとき残差の自己相関係数は 0 に近い値となり、ピリオドグラムは平坦に近くなる。このような波をホワイトノイズと呼ぶ。ホワイトノイズの検定には、Ljung-Box 検定が用いられる。それには、利用するデータ数を t、ラグ i の母相関係数と標本相関係数をそれぞれ ρ_i，r_i として、以下の関係が利用される。

帰無仮説： $\rho_1 = \rho_2 = \cdots = \rho_m = 0$

$$Q = t(t+2)\left\{\frac{r_1^2}{t-1} + \frac{r_2^2}{t-2} + \cdots + \frac{r_m^2}{t-m}\right\} \quad \sim \quad \chi_m^2$$

4)　変動の分解モデルによる予測

時系列データの変動の分解は、データにある程度の周期性があること、その数が最低でも 2 周期分以上あることが条件で可能となる。また傾向変動２（予測手法）を使うと長期予測は難しい。これまで見てきたデータについて 100 期先までの長期予測をしてみよう。見易くするために $t = 200$ からのデータを図 10 に表示する。

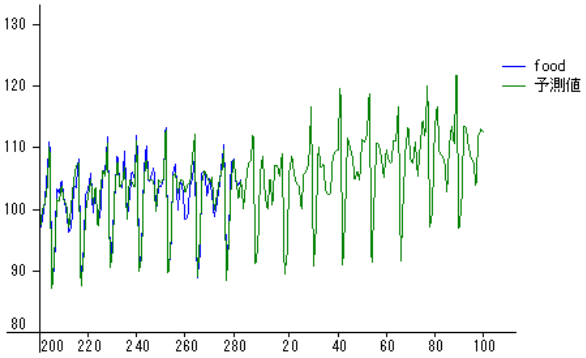


図 10　時系列データと長期予測

参考文献

[1] 高橋玲子他著, 上田太一郎監修, Excel で学ぶ時系列分析と予測, オーム社, 2006.

[2] 北川源四郎, 時系列解析入門, 岩波書店, 2005.

[3] 石村貞夫, SPSS による時系列分析の手順［第２版］, 東京図書, 2006.

[4] 統計数理研究所のホームページの中の Web Decomp のサンプルデータを使用させていただきました。（http://ssnt.ism.ac.jp/inets/inets.html）

１３．共分散構造分析

13.1　共分散構造分析とは

共分散構造分析は、重回帰分析と因子分析を組み合わせて、観測される変数と潜在的な変数の間の関係をネットワーク構造図で表現する統計モデルである。変数間の影響の強さを表すパラメータの値は構造方程式と呼ばれる方程式を通して、観測変数の共分散行列から推定する。その際パラメータ値を決定するために、パラメータ数は共分散行列の独立な成分数より少ない場合を考え、パラメータ数より方程式数が多い状態が生じる（同じ場合もある）。そのため、パラメータの推定にはある評価関数を用いて、これを最小化するような方法を考える。この評価関数の選び方によって、推定値の導出にはいくつかの方法がある。その中で最もよく利用されるのが最小２乗法や最尤法である。

ここではまず図１の構造モデルを例として共分散構造分析の概要を説明をする。

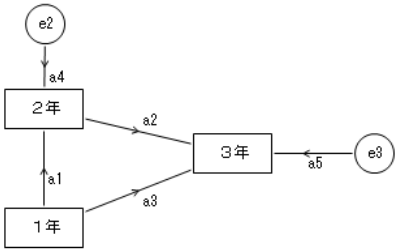


図１　構造モデル

四角や楕円（ここには含まれていない）や円で表される量はモデルに含まれる変数で、形によりその意味するところが異なり、それぞれラベルが付けられている。矢印は因果関係を表すパラメータで、これにもラベルが付けられている。また、双方向の矢印（ここには含まれていない）は相関を表すパラメータである。

変数は通常、いくつかの視点から以下のように分けられる。

観測変数と潜在変数

観測変数とは実測値の分っている変数であり、図 １ の構造モデルでは１年、２年、３年の変数がこれに相当し、構造図では四角形で表現される。潜在変数とは直接には観測されない変数で、因子分析の因子や誤差などがこれに当り、構造図では楕円や円で表現される。図１の例では e2, e3 である。ここでは楕円（含まれていない）が因子、円が誤差変数である。

外生変数と内生変数

外生変数は構造モデルで相関を除いてどこからも影響を受けない（片側矢印が入らない）変数で、図１の構造モデルでは１年, e2, e3 がこれに当る。内生変数はそれ以外の変数で２年，３年などである。

構造変数と誤差変数

構造変数とは後に述べるモデルの構成要素に使われる変数で、図 1 の構造モデルでは１年、２年、３年がこれに当る。潜在変数の中の因子も構造変数である。誤差変数とはモデルでは説明できないゆらぎの成分を表すもので e2, e3 がこれに当る。

このモデルをよく利用される影響行列の形で表現すると表 1 のようになる。左側の変数が始点、上側の変数が終点である。

表 1 構造モデルの影響行列

	１年	２年	３年	e2	e3
１年		a1	a3		
２年			a2		
３年					
e2		a4			
e3			a5		

これらの変数の関係は構造方程式と呼ばれる式で表現される。図 1 の構造モデルでは以下となる。

２年＝a1×１年＋a4×e2

３年＝a3×１年＋a2×２年＋a5×e3

この方程式の左辺をすべての構造変数に拡張し、以下のような式を考える

１年＝１年

２年＝a1×１年＋a4×e2

３年＝a3×１年＋a2×２年＋a5×e3

構造方程式の左辺には構造変数と呼ばれる変数を取るが､そのうちの内生変数は必ず誤差変数からの影響を受けるようにする。

これらの関係から､共分散（同じ変数同士だと分散）を求めると以下の関係を得る。但し、構造変数（因子も含む）と別の誤差、誤差と誤差の相関はなく、共分散は 0 であると仮定する。また因子や誤差の分散は 1 とする。

Cov(１年, １年)＝Cov(１年, １年)

Cov(２年, ２年)＝a1²×Cov(１年, １年)＋a4²× 1

Cov(３年, ３年)＝a3²×Cov(１年, １年)＋a2²×Cov(２年, ２年)

＋a3×a2×Cov(１年, ２年)＋a5²× 1

Cov(１年, ２年)＝a1×Cov(１年, １年)

Cov(１年, ３年)＝a3×Cov(１年, １年)＋a2×Cov(１年, ２年)

Cov(２年, ３年)＝a1×a3×Cov(１年, １年)＋a1×a2×Cov(１年, ２年)

+a4×a2×a4×Cov(e2×e2)　←この部分注意

Cov(,)の部分は実測データから求められるので、これはパラメータ a1,a2,a3,a4,a5 に対する連立方程式である。しかし、変数の数と方程式の数は一般には一致せず、厳密な解は求まらない。そこで、パラメータが良い値を取るように、例えば実測の共分散とモデルから得られ

た予測の共分散ができるだけ近い値を取るようにしたり、実測変数が正規分布するとして、予測した共分散をできるだけ正規分布に適合させるようにしたりして、パラメータの近似値を求める。前者が最小２乗法、後者が最尤法である。実際の計算では可能な限り最尤法が利用される。このように共分散を元にパラメータを推測する方法として共分散構造分析という名前が付けられている。

ここでこのようにして求めたモデルの良し悪しを評価するいくつかの指標とその性質についてまとめておく。

自由度

自由度は以下のように定義する（これまでの解釈を変更する必要がある）。ここでは、「外生構造変数の分散が、パラメータを決定する構造方程式を決めない」という分析の性質から、少しくどくなるがいろいろな書き方をしてみる。

自由度＝構造方程式を決める分散・共分散の値－全パラメータ数
＝観測変数の分散共分散数＋潜在構造変数の分散数－外生構造変数の分散数
－全パラメータ数
＝観測変数の分散共分散数＋潜在構造変数の数
－全パラメータ数－誤差を除く外生変数の数
＝観測変数の分散共分散数＋（外生潜在構造変数の数＋内生潜在構造変数の数）
－（誤差を除くパラメータ数＋誤差数）－（外生変数の数－誤差数）
＝観測変数の分散共分散数＋（外生潜在構造変数の数＋内生潜在構造変数の数）
－誤差を除くパラメータ数－外生変数の数
＝観測変数の分散共分散数＋（外生潜在構造変数の数＋内生潜在構造変数の数）
－誤差を除くパラメータ数
－（外生観測変数の数＋外生潜在構造変数の数＋誤差数）
＝観測変数の分散共分散数＋内生潜在構造変数の数
－誤差を除くパラメータ数－外生観測変数の数
－誤差数
＝観測変数の分散共分散数＋内生潜在構造変数の数
－誤差を除くパラメータ数－外生観測変数の数
－（内生観測変数への誤差数＋内生潜在構造変数への誤差数）
＝観測変数の分散共分散数＋内生潜在構造変数の数
－誤差を除くパラメータ数－外生観測変数の数
－（内生観測変数の数＋内生潜在構造変数の数）
＝観測変数の分散共分散数
－（誤差を除くパラメータ数＋外生観測変数の数＋観測変数への誤差数）

以上より、実用上は以下の式を使うと簡単に分かる。

自由度＝ｎ（ｎ＋１）/２－ｍ

ｎ：観測変数の数　→　ｎ（ｎ＋１）/２：観測変数の分散共分散数

ｍ：誤差を除くパラメータ数＋外生観測変数の数＋観測変数への誤差数

パラメータの検定

パラメータの値を 0 と比較した検定結果が表示されるが、この多くに有意差が表れることが基本である。

解の検定

構成されたモデルが正しいかどうかを判定する検定である。P<0.05 で正しくないと判定する。この検定はデータ数を増やして精度を上げるほど対立仮説である「モデルは正しくない」という結果が出やすくなるという矛盾を含んでいる。

適合度指標

GFI（Goodness of Fit Index）

これは実測値による共分散行列とパラメータで表された共分散行列の類似の程度を見る指標である。この指標の値は 0.9 以上が良いとされるが、モデルの自由度が大きくなると値を大きくすることが難しくなる。

AGFI（Adjusted Goodness of Fit Index）

これは GFI の自由度の問題を改善した指標で、相関を加えて自由度を見かけ上小さくしても値が改善されるとは限らない指標である。一般に AGFI≦GFI の関係がある。

情報量基準

AIC（Akaike's Information Criterion）

これは一般の統計モデルの評価指標として有名である。この値が小さいほど良いモデルとされる。この指標には、標本数が多い場合、自由度が小さい（パラメータ数が多い）モデルが良いモデルと判断される傾向がある。

CAIC（Consistent Akaike's Information Criterion）

これは AIC の標本数の影響を抑えた指標である。

自由度当りのモデルの分布と最尤モデルとの乖離を表す指標

RMSEA（Root Mean Square Error of Approximation）

≦0.05 で当てはまりが良い、≧0.1 で当てはまりが悪いと言われている。

13.2　プログラムの利用法（基本）

ここでは前節で述べた例を用いてプログラムの動作を説明する。プログラムを起動すると図 1 のような実行画面が表示される。これはできるだけ簡易化したメニューである。この中で「拡張メニュー」というボタンがあるが、この利用法については理論の詳細を述べた後に示す。

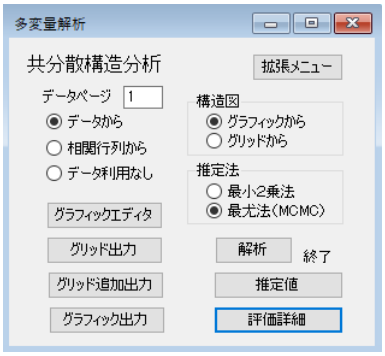


図 1　共分散構造分析実行画面

データは図 2 のような形式である。

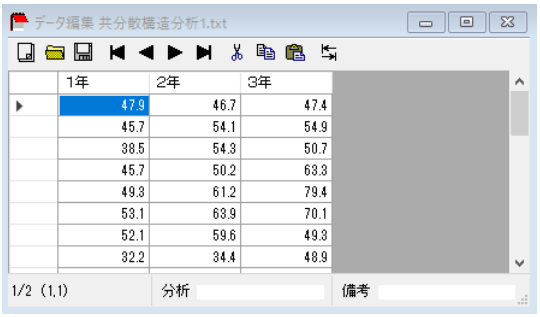


図 2　共分散構造分析のデータ

データのあるページの番号を「データページ」テキストボックスに記入し（この場合は 1 頁)、「グラフィックエディタ」ボタンをクリックすると図 3 のようなグラフィックエディタの画面が表示される。

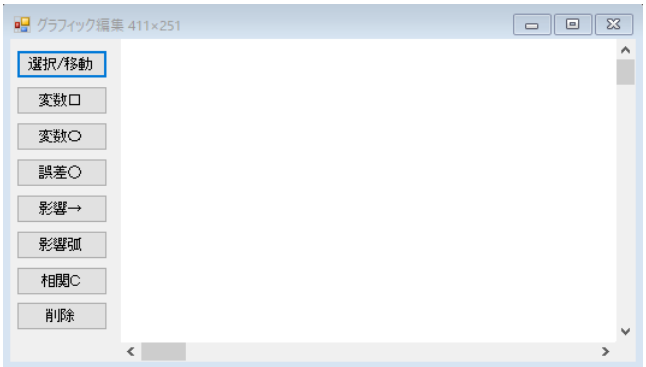


図 3　グラフィックエディタ

左側のボタンをクリックして動作を選び、マウスを使って構造図を図 4 のように描く。図形の名前を変更するには、図形や線の上でマウスをダブルクリックし、出てきたテキストボックスに文字列を入力して、テキストボックスの外をクリックする。

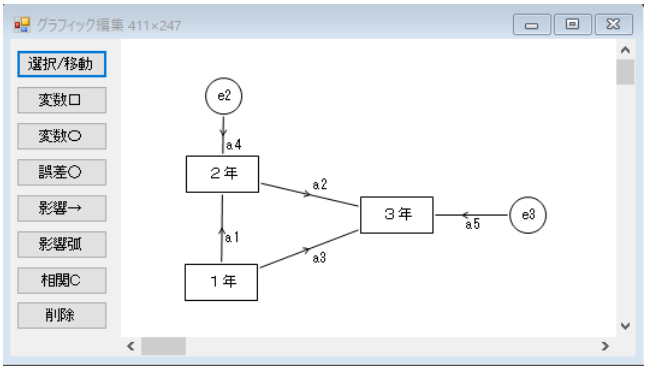
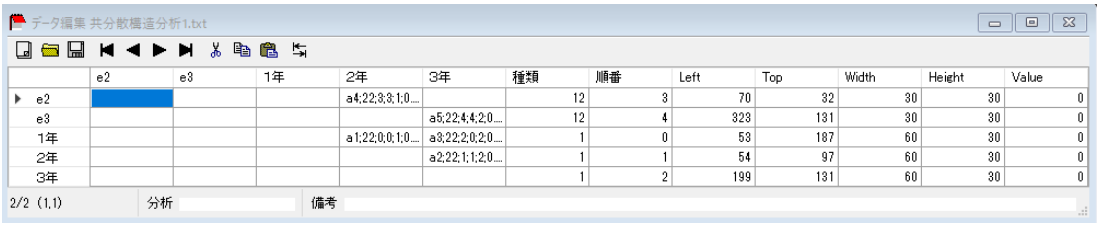


図 4　構造図

この描き方については、総合マニュアルのツールの中にもある。このデータはグラフィックエディタのメニュー［編集－画面コピー］で画像としてコピーされる。また、メニュー［編集－グリッド出力］で現在のグリッドエディタのページに、メニュー［編集－グリッド追加出力］で、グリッドエディタの最後のページに追加されてコピーされる。以後図形はグラフィックエディタのデータとして扱われる。その形式を図 5 に示す。



	e2	e3	1年	2年	3年	種類	順番	Left	Top	Width	Height	Value
e2				a4;22;3;3;1;0...		12	3	70	32	30	30	0
e3					a5;22;4;4;2;0...	12	4	323	131	30	30	0
1年				a1;22;0;0;1;0...	a3;22;2;0;2;0...	1	0	53	187	60	30	0
2年					a2;22;1;1;2;0...	1	1	54	97	60	30	0
3年						1	2	199	131	60	30	0

図 5　グラフィックエディタのデータ形式

図 5 のデータにしておくと「グラフィック出力」ボタンで図 4 のような構造図に変えることができる。

計算はグラフィックエディタのデータを元にするか、グリッドエディタに保存されたデータを元にするか、「グラフィックから」、「グリッドから」ラジオボタンで指定することができる。通常は「グラフィックから」にしておいて、良い構造図ができたらグリッドに保存するようにすればよい。推定法も「最尤法」使う方が良いが、構造が複雑だと結果が出ないようなこともあるので使い分ける。

分析を実行するにはまず「解析」ボタンをクリックする。終了の表示が出たら計算の終了である。「推定値」のボタンをクリックして図 6a と図 6b の図と表での結果を得る。

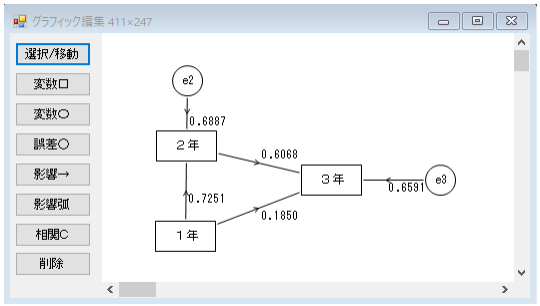


図 6a　推定値（図）

	1年	2年	3年	e2	e3
1年		0.725	0.185		
2年			0.607		
3年					
e2		0.689			
e3			0.659		
分散	1.000	1.000	1.000	1.000	1.000
評価関数値	0.000				

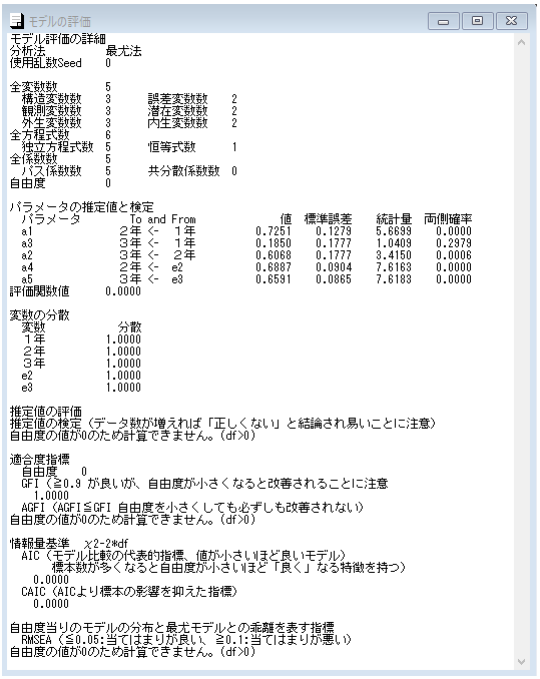
図 6b　推定値（表）

評価の詳細については、「評価詳細」ボタンをクリックすると図 7 と図 8 の結果を得る。


パラメータ推定値

	変数		変数	推定値	標準誤差	検定値	両側確率
a1	2年	<-	1年	0.725	0.016	5.670	0.0000
a3	3年	<-	1年	0.185	0.032	1.041	0.2979
a2	3年	<-	2年	0.607	0.032	3.415	0.0006
a4	2年	<-	e2	0.689	0.008	7.616	0.0000
a5	3年	<-	e3	0.659	0.007	7.618	0.0000
評価関数値	0.000						

図７　結果の詳細表示１


モデルの評価

```
モデル評価の詳細
分析法          最尤法
使用乱数Seed    0

全変数数        5
  構造変数数    3       誤差変数数    2
  観測変数数    3       潜在変数数    2
  外生変数数    3       内生変数数    2
全方程式数      6
  独立方程式数  5       恒等式数      1
全係数数        5
  パス係数数    5       共分散係数数  0
自由度          0

パラメータの推定値と検定
  パラメータ          To and From             値  標準誤差    統計量  両側確率
  a1                2年 <-  1年          0.7251    0.1279    5.6699    0.0000
  a3                3年 <-  1年          0.1850    0.1777    1.0409    0.2979
  a2                3年 <-  2年          0.6068    0.1777    3.4150    0.0006
  a4                2年 <-  e2           0.6887    0.0904    7.6183    0.0000
  a5                3年 <-  e3           0.6591    0.0865    7.6183    0.0000
評価関数値      0.0000

変数の分散
  変数              分散
  1年           1.0000
  2年           1.0000
  3年           1.0000
  e2            1.0000
  e3            1.0000

推定値の評価
推定値の検定（データ数が増えれば「正しくない」と結論され易いことに注意）
自由度の値が0のため計算できません。(df>0)

適合度指標
  自由度    0
  GFI（≧0.9 が良いが、自由度が小さくなると改善されることに注意
    1.0000
  AGFI（AGFI≦GFI 自由度を小さくしても必ずしも改善されない）
自由度の値が0のため計算できません。(df>0)

情報量基準　χ2-2*df
  AIC（モデル比較の代表的指標。値が小さいほど良いモデル）
       標本数が多くなると自由度が小さいほど「良く」なる特徴を持つ）
    0.0000
  CAIC（AICより標本の影響を抑えた指標）
    0.0000

自由度当りのモデルの分布と最尤モデルとの乖離を表す指標
  RMSEA（≦0.05:当てはまりが良い　≧0.1:当てはまりが悪い）
自由度の値が0のため計算できません。(df>0)
```

図８　結果の詳細表示２

最後にこれまで述べてきたことをまとめて、利用の手引きとしよう。

変数の分類

観測変数と潜在変数（実際に観測されるかどうかの分類）
構造変数と誤差変数（データ構造に関与するかどうかの分類，重要性）
外生変数と内生変数（影響を受けるかどうかの分類、影響するだけは外生変数）

図形表示

四角形　観測変数，構造変数，外生も内生もあり
楕円形　潜在変数，構造変数，外生も内生もあり
円形　　潜在変数，誤差変数，外生変数
矢印　　影響の方向を表す。
円弧　　相関を表す。

共分散構造分析の目的　　観測データからデータ間の因果関係を予想する。

構造図のルール　　内生変数には必ず誤差変数を付ける（外生変数には不要）。
　　プログラムの「評価詳細」ボタンで見たとき、自由度が１以上

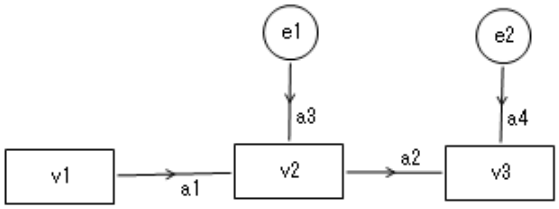
プログラムの実行　　「解析」ボタンをクリックした後、推定値や評価詳細を見る。

係数を見る　→　「推定値」ボタン

構造図の評価　→　「評価詳細」ボタン（表示画面で解説）

問題１

共分散構造分析 4.txt のデータから以下の構造を考え、最尤法を用いて共分散構造分析を実行し、以下の問いに答えよ。



１）変数の数はいくらか。[　　]

２）以下の変数名を書け。

構造変数［　　］　誤差変数［　　］

観測変数［　　］　潜在変数［　　］

外生変数［　　］　内生変数［　　］

３）パラメータ（係数）の数はいくらか。[　　]

４）パラメータの推定値を上の図に書き込め。

５）評価関数値はいくらか。[　　]

６）自由度の数はいくらか。[　　]

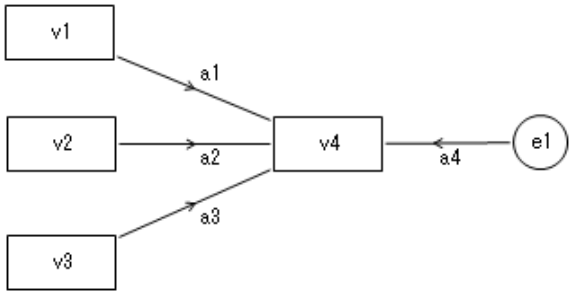
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。[　　]

８）適合度指標 GFI の値はいくらか。[　　]

９）良いモデルか。[良い・悪い]

問題２（重回帰分析）

共分散構造分析 2.txt のデータから考えた、以下のモデルのパラメータの推定値は重回帰分析の結果と等しいか調べる。



１）変数の数はいくらか。[　　]

２）以下の変数名を書け。

構造変数［　　］　誤差変数［　　］

観測変数［　　］　潜在変数［　　］

外生変数［　　］　内生変数［　　］

３）パラメータ（係数）の数はいくらか。［　　］

４）パラメータの推定値を上の図に書き込め。

５）評価関数値はいくらか。［　　］

６）自由度の数はいくらか。［　　］

７）有意に 0 と異なる誤差から以外のパラメータはいくつか。［　　］

８）適合度指標 GFI の値はいくらか。［　　］

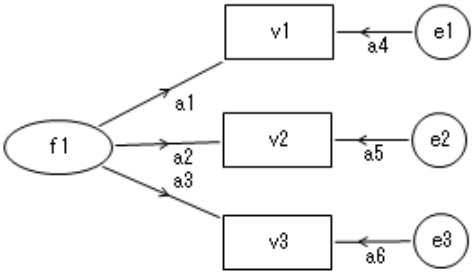
９）良いモデルか。［良い・悪い］

10）パラメータの推定値は最小２乗法と最尤法で等しいか。［等しい・異なる］

11）パラメータの推定値は重回帰分析のどの変数と等しいか。［偏回帰係数・標準化係数］

問題３（因子分析）

共分散構造分析 3.txt のデータから以下の構造を考え、最尤法を用いて共分散構造分析を実行し、以下の問いに答えよ。



１）変数の数はいくらか。［　　］

２）以下の変数名を書け。

構造変数［　　］　誤差変数［　　］

観測変数［　　］　潜在変数［　　］

外生変数［　　］　内生変数［　　］

３）パラメータ（係数）の数はいくらか。［　　］

４）パラメータの推定値を上の図に書き込め。

５）評価関数値はいくらか。［　　］

６）自由度の数はいくらか。［　　］

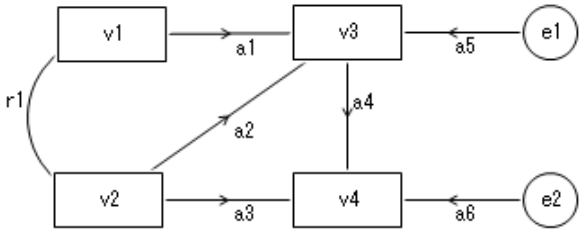
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。［　　］

８）適合度指標 GFI の値はいくらか。［　　］

９）良いモデルか。［良い・悪い］

問題４

共分散構造分析 5.txt のデータから以下の構造を考え、最尤法を用いて共分散構造分析を実行し、以下の問いに答えよ。



１）変数の数はいくらか。[　　]

２）以下の変数名を書け。

構造変数［　　］　誤差変数［　　］

観測変数［　　］　潜在変数［　　］

外生変数［　　］　内生変数［　　］

３）パラメータ（係数）の数はいくらか。[　　]

４）パラメータの推定値を上の図に書き込め。

５）評価関数値はいくらか。[　　]

６）自由度の数はいくらか。[　　]

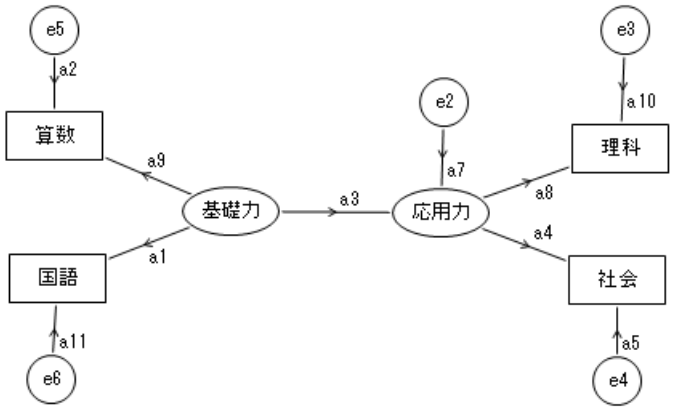
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。[　　]

８）適合度指標 GFI の値はいくらか。[　　]

９）良いモデルか。[良い・悪い]

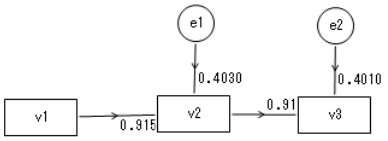
問題５

共分散構造分析 11. txt のデータから以下の構造を考え、最尤法を用いて共分散構造分析を実行し、以下の問いに答えよ。



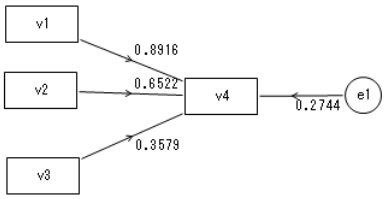
１）変数の数はいくらか。[　　]
２）以下の変数名を書け。
　構造変数［　　］　誤差変数［　　］
　観測変数［　　］　潜在変数［　　］
　外生変数［　　］　内生変数［　　］
３）パラメータ（係数）の数はいくらか。[　　]
４）パラメータの推定値を上の図に書き込め。
５）評価関数値はいくらか。[　　]
６）自由度の数はいくらか。[　　]
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。[　　]
８）適合度指標 GFI の値はいくらか。[　　]
９）良いモデルか。[良い・悪い]

問題１解答
１）変数の数はいくらか。[5]
２）以下の変数名を書け。
　構造変数［ v1,v2,v3 ］　誤差変数［ e1,e2 ］
　観測変数［ v1,v2,v3 ］　潜在変数［ e1,e2 ］
　外生変数［ v1,e1,e2 ］　内生変数［ v2,v3 ］
３）パラメータ（係数）の数はいくらか。[4]
４）パラメータの推定値を上の図に書き込め。



５）評価関数値はいくらか。[0.022]
６）自由度の数はいくらか。[1]
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。[2]
８）適合度指標 GFI の値はいくらか。[0.9742]
９）良いモデルか。[良い・悪い]

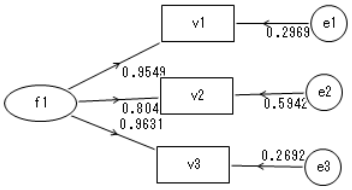
問題２解答（重回帰分析）
１）変数の数はいくらか。[5]
２）以下の変数名を書け。
　構造変数［ v1,v2,v3,v4 ］　誤差変数［ e1 ］
　観測変数［ v1,v2,v3,v4 ］　潜在変数［ e1 ］
　外生変数［ v1,v2,v3,e1 ］　内生変数［ v4 ］
３）パラメータ（係数）の数はいくらか。[4]
４）パラメータの推定値を上の図に書き込め。



５）評価関数値はいくらか。［ 0.0000 ］
６）自由度の数はいくらか。［ 0 ］
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。［ 3 ］
８）適合度指標 GFI の値はいくらか。［ 1.0000 ］
９）良いモデルか。［良い・悪い］
10）パラメータの推定値は最小２乗法と最尤法で等しいか。［等しい・異なる］
11）パラメータの推定値は重回帰分析のどの変数と等しいか。［偏回帰係数・標準化係数］

問題３解答（因子分析）

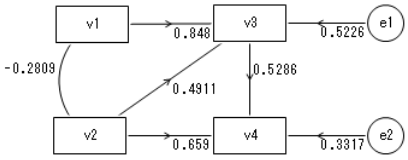
１）変数の数はいくらか。［ 7 ］
２）以下の変数名を書け。
　構造変数［ f1,v1,v2,v3 ］　誤差変数［ e1,e2,e3 ］
　観測変数［ v1,v2,v3 ］　潜在変数［ f1,e1,e2,e3 ］
　外生変数［ f1,e1,e2,e3 ］　内生変数［ v1,v2,v3 ］
３）パラメータ（係数）の数はいくらか。［ 6 ］
４）パラメータの推定値を上の図に書き込め。



５）評価関数値はいくらか。［ 0.0000 ］
６）自由度の数はいくらか。［ 0 ］
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。［ 3 ］
８）適合度指標 GFI の値はいくらか。［ 1.0000 ］
９）良いモデルか。［良い・悪い］

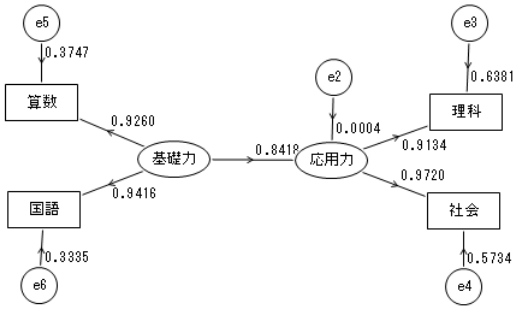
問題４解答

１）変数の数はいくらか。［ 6 ］
２）以下の変数名を書け。
　構造変数［ v1,v2,v3,v4 ］　誤差変数［ e1,e2 ］
　観測変数［ v1,v2,v3,v4 ］　潜在変数［ e1,e2 ］
　外生変数［ v1,v2,e1,e2 ］　内生変数［ v3,v4 ］
３）パラメータ（係数）の数はいくらか。［ 7 ］
４）パラメータの推定値を上の図に書き込め。



５）評価関数値はいくらか。［ 0.004 ］
６）自由度の数はいくらか。［ 1 ］
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。［ 4 ］
８）適合度指標 GFI の値はいくらか。［ 0.9950 ］
９）良いモデルか。［良い・悪い］

問題５解答
１）変数の数はいくらか。［ 11 ］
２）以下の変数名を書け。
　構造変数［ 算, 国, 理, 社, 基, 応 ］　誤差変数［ e2, e3, e4, e5, e6 ］
　観測変数［ 算, 国, 理, 社 ］　潜在変数［ 基, 応, e2, e3, e4, e5, e6 ］
　外生変数［ 基, e2, e3, e4, e5, e6 ］　内生変数［算, 国, 理, 社, 応 ］
３）パラメータ（係数）の数はいくらか。［ 10 ］
４）パラメータの推定値を上の図に書き込め。



５）評価関数値はいくらか。［ 0.109 ］
６）自由度の数はいくらか。［ 0 ］
７）有意に 0 と異なる誤差から以外のパラメータはいくつか。［ 2 ］
８）適合度指標 GFI の値はいくらか。［ 0.9155 ］
９）良いモデルか。［良い・悪い］誤差の影響が大きすぎる。

13.3　共分散構造分析の理論

1)　モデルの構造と方程式

ここでは図 1 の構造モデル（共分散構造分析 12（理論サンプル）.txt）を例として共分散構造分析の理論の説明をする。説明は 1 節と重複する部分がある。

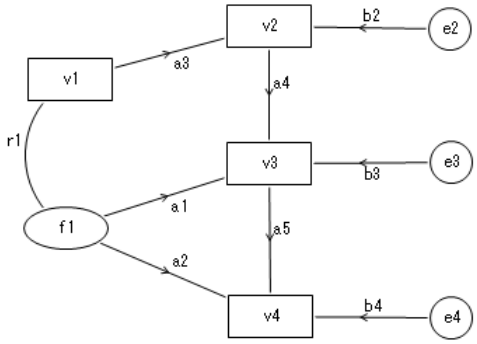


図１　構造モデル

四角や楕円や円で表される量はモデルに含まれる変数で､形によりその意味するところが異なり、それぞれラベルが付けられている。矢印は因果関係を表すパラメータで、これにもラベルが付けられている。また、双方向の矢印は相関を表すパラメータである。

このモデルをよく利用される影響行列の形で表現すると表 1 のようになる｡左側の変数が始点、上側の変数が終点である。

表 1　構造モデルの影響行列

	f_1	v_1	v_2	v_3	v_4	e_2	e_3	e_4
f_1		r		a_1	a_2			
v_1	r		a_3					
v_2				a_4				
v_3					a_5			
v_4								
e_2			b_2					
e_3				b_3				
e_4					b_4			

変数は通常、いくつかの視点から以下のように分けられる。

観測変数と潜在変数

観測変数とは実測値の分かっている変数であり、図 1 の構造モデルでは v_1, v_2, v_3, v_4 などの変数がこれに相当し､構造図では四角形で表現される。潜在変数とは直接には観測されない変数で､因子分析の因子や誤差などがこれに当り、構造図では楕円や円で表現される。図 1 の例では f_1, e_2, e_3, e_4 などの変数である。ここでは f_1 が因子、e_2, e_3, e_4 が誤差変数である。特に因子は楕円、誤差変数は円（または円なし）で表現される場合がある。

外生変数と内生変数

外生変数は構造モデルで相関を除いてどこからも影響を受けない（片側矢印が入らない）変数で、図 1 の構造モデルでは v_1, f_1, e_2, e_3, e_4 がこれに当る。内生変数はそれ以外の変数で v_2, v_3, v_4 などである。

構造変数と誤差変数

構造変数とはモデルの構成要素に使われる変数で、図 1 の構造モデルでは f_1, v_1, v_2, v_3, v_4 などがこれに当る。誤差変数とはモデルでは説明できないゆらぎの成分を表すもので e_2, e_3,

e_4がこれに当る。

これらの変数の関係は構造方程式と呼ばれる式で表現される。図 1 の構造モデルでは以下となる。

$$v_2 = a_3 v_1 + b_2 e_2$$
$$v_3 = a_1 f_1 + a_4 v_2 + b_3 e_3$$
$$v_4 = a_2 f_1 + a_5 v_3 + b_4 e_4$$

この方程式の左辺を構造変数に拡張し、以下のような式を考える

$$f_1 = f_1$$
$$v_1 = v_1$$
$$v_2 = a_3 v_1 + b_2 e_2$$
$$v_3 = a_1 f_1 + a_4 v_2 + b_3 e_3$$
$$v_4 = a_2 f_1 + a_5 v_3 + b_4 e_4$$

構造方程式の左辺には構造変数と呼ばれる変数を取るが､そのうちの内生変数は必ず誤差変数からの影響を受けるようにする。上の構造方程式を行列表示すると以下の形になる。

$$\begin{pmatrix} f_1 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & a_3 & 0 & 0 & 0 \\ a_1 & 0 & a_4 & 0 & 0 \\ a_2 & 0 & 0 & a_5 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_2 & 0 & 0 \\ 0 & 0 & 0 & b_3 & 0 \\ 0 & 0 & 0 & 0 & b_4 \end{pmatrix} \begin{pmatrix} f_1 \\ v_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

今、以下のように定義すると、

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & a_3 & 0 & 0 & 0 \\ a_1 & 0 & a_4 & 0 & 0 \\ a_2 & 0 & 0 & a_5 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_2 & 0 & 0 \\ 0 & 0 & 0 & b_3 & 0 \\ 0 & 0 & 0 & 0 & b_4 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} f_1 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} f_1 \\ v_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

構造方程式は(1)式のように表すことができる。

$$\mathbf{t} = \mathbf{A}\mathbf{t} + \mathbf{B}\mathbf{h} \qquad (1)$$

ここに$\mathbf{t}$は構造変数からなるベクトル、$\mathbf{h}$は外生変数からなるベクトルである。またパラメータは行列$\mathbf{A}$と$\mathbf{B}$に含まれる。

構造方程式は以下のように変形できる。

$$\mathbf{t} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{h} \qquad (2)$$

ここでベクトル$\mathbf{t}$のうち観測変数に注目し、観測変数で作られたベクトル$\mathbf{v}$とそれを取り出す行列$\mathbf{G}$を以下のように定義する。

$$\mathbf{v} = \mathbf{G}\mathbf{t}, \quad \text{ここに} \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

この関係を用いると、上式は(3)式のように変形される。

$$\mathbf{v} = \mathbf{G}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{h} \tag{3}$$

次に観測変数 $\mathbf{v}$ および外生変数 $\mathbf{h}$ の共分散行列を考える。簡単のため潜在変数は平均が 0、分散が 1 になるように標準化されているものとする。変数 $\mathbf{v}$ の共分散行列を $E(\mathbf{v}\mathbf{v}')$、変数 $\mathbf{h}$ の共分散行列を $E(\mathbf{h}\mathbf{h}')$ とするとそれらの関係は(4)式のようになる。

$$E(\mathbf{v}\mathbf{v}') = \mathbf{G}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}E(\mathbf{h}\mathbf{h}')\mathbf{B}'(\mathbf{I} - \mathbf{A})'^{-1}\mathbf{G}' \tag{4}$$

実際の計算では $E(\mathbf{v}\mathbf{v}')$ を標本から得られた不偏共分散行列（共分散行列の不偏推定量）で置き換え、$E(\mathbf{h}\mathbf{h}')$ についても観測変数部分は不偏共分散行列、潜在変数部分は分散を 1、共分散には必要に応じて共分散を表すパラメータを設定する。図 1 の構造モデルの場合は、潜在変数間または外生の観測変数と潜在変数間で、f_1 と v_1 の間だけに共分散 r を仮定しているので、以下の形となる。

$$E(\mathbf{v}\mathbf{v}') \Rightarrow \mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{23} & u_{33} & u_{34} \\ u_{41} & u_{24} & u_{34} & u_{44} \end{pmatrix}, \quad E(\mathbf{h}\mathbf{h}') \Rightarrow \mathbf{H} = \begin{pmatrix} 1 & r & 0 & 0 & 0 \\ r & u_{11} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

ここで $\mathbf{U}$ は不偏共分散行列であるが、標準化したデータの場合には相関行列となる。これを用いて(4)式を書き換えると以下のようになる。

$$\mathbf{G}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{H}\mathbf{B}'(\mathbf{I} - \mathbf{A})'^{-1}\mathbf{G}' = \mathbf{U} \tag{5}$$

これは観測値とパラメータを結びつける方程式である。この方程式を丁度方程式と呼び、一意的な解が存在する場合、その解を丁度解と呼ぶ。しかし丁度解が存在する場合はまれで、一般には解が不定（方程式数＜パラメータ数）になっていたり、不能（方程式数＞パラメータ数）になっていたりする。解が不定になっている場合をパラメータは識別不能という。不能になっている場合は最適近似解を求める。最適近似解を求める方法はいくつかあるが、ここでは主に利用される２つの方法について紹介する。

2) パラメータの推定

パラメータの推定は方程式の近似解を求めるための評価関数を作り、それを最小化する方法がとられるが、この節ではよく利用される２つの評価関数について説明する。

最小 2 乗法

方程式(5)の左辺と右辺の差の 2 乗和を最小化するために以下の評価関数を考える。

$$f_{MS}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=i}^{n}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} - u_{ij}\right)^2 \tag{6}$$

ここに $\boldsymbol{\theta}$ はパラメータを総称したものであり、n は観測変数の数、$\boldsymbol{\Sigma}(\boldsymbol{\theta})$ は以下のように(5)式の左辺を表す。

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{G}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{H}\mathbf{B}'(\mathbf{I}-\mathbf{A})'^{-1}\mathbf{G}'$$

丁度解の場合 $f_{MS}(\boldsymbol{\theta})$ の値は 0 である。

最尤法

我々はまず観測値を与える確率変数 $\mathbf{x}_\lambda$（$\lambda = 1, 2, \cdots, N$）がそれぞれ独立に n 変量正規分布に従うと考える。共分散行列を $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ とすると、$\mathbf{x}_\lambda$ の確率密度関数は以下で与えられる。

$$f(\mathbf{x}_\lambda, \boldsymbol{\theta}) = (2\pi)^{-p/2} \,|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_\lambda - \boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{x}_\lambda - \boldsymbol{\mu})\right]$$

N 回の独立な観測に関する確率密度関数は以下で与えられる。

$$f(\mathbf{x}, \boldsymbol{\theta}) = \prod_{\lambda=1}^{N} f(\mathbf{x}_\lambda, \boldsymbol{\theta})$$

最尤法ではこの確率密度関数に実測値 $\hat{\mathbf{x}}_\lambda$ を代入した尤度関数 $f(\boldsymbol{\theta})$ を最大化するようにパラメータを決定する。実際には計算の簡単化のため、尤度関数を対数変換した対数尤度関数の符号を変えたものを最小化する。符号を変えた対数尤度関数は以下で与えられる。

$$\begin{aligned} -\log f(\boldsymbol{\theta}) &= -\sum_{\lambda=1}^{N} \log f(\hat{\mathbf{x}}_\lambda, \boldsymbol{\theta}) \\ &= \frac{1}{2}\sum_{\lambda=1}^{N}(\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}}) - \frac{N}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}| + const. \\ &= \frac{N}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right) - \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}|\right) + const. \end{aligned}$$

ここに、

$$\mathbf{S} = \frac{1}{N}\sum_{\lambda=1}^{N}(\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}})(\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}})'$$

通常最尤法の評価関数としては、上の対数尤度関数に定数を加えた以下の式が用いられることが多い。丁度解の場合、この評価関数の値も 0 である。

$$f_{ML}(\theta) = \mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right) - \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}| - n \qquad (7)$$

これらの評価関数の最小化法には様々な方法が用いられるが、現在我々は、マルコフ連鎖モンテカルロ法（MCMC）によって初期値を与え、Newton-Raphson 法に類似の Levenberg-Marquart 法を用いて収束計算を行っている。

3) モデルの評価

ここではモデルの良し悪しを評価するいくつかの指標とその性質についてまとめておく。

パラメータの検定

最尤法の推定値 $\hat{\boldsymbol{\theta}}$ を用いると、以下のようになることが知られている。

$$z_i = \frac{\hat{\theta}_i}{\sigma_{\hat{\theta}_i}} \sim \mathrm{N}(0,1) \quad \text{ここに、} \quad \sigma^2_{\hat{\theta}_i} \equiv \left[\frac{N-1}{2}\frac{\partial^2}{\partial\theta_i^2} f_{ML}(\boldsymbol{\theta})\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

これを用いてパラメータの値を 0 と比較する検定を行うことができるが、この検定が有意

な結果を与えることが基本である。

解の検定

帰無仮説 H_0：構成されたモデルは正しい。

対立仮説 H_1：構成されたモデルは正しくない。

$$\chi^2 = (N-1) f_{ML} \sim \chi^2_{df}, \qquad df = \frac{1}{2} n(n+1) - p$$

ここに N はデータ数、n は観測変数の数、p は自由パラメータ数（外生観測変数数＋パス係数数＋誤差変数数＋共分散（相関）数）であり、df は χ^2 分布の自由度である。この検定はデータ数を増やして精度を上げるほど対立仮説である「モデルは正しくない」という結果が出易くなるという矛盾を含んでいる。

適合度指標

GFI（Goodness of Fit Index）

これは実測値による共分散行列とパラメータで表された共分散行列の類似の程度を見る指標で以下のように与えられる。

$$\mathrm{GFI} = 1 - \frac{\mathrm{tr}\left(\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{S} - \mathbf{I} \right)^2 \right)}{\mathrm{tr}\left(\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{S} \right)^2 \right)} \quad \text{ここに} \mathrm{tr}(\mathbf{A}^2) = \mathrm{tr}(\mathbf{A}\mathbf{A}')$$

この指標の値は 0.9 以上が良いとされるが、モデルの自由度が大きくなると値を大きくすることが難しくなる。ここで GFI の値が大きいからといってよいモデルとは限らないことを注意しておく。パラメータの有意性の検定は重要である。

AGFI（Adjusted Goodness of Fit Index）

これは GFI の自由度の問題を改善し、相関を加えて自由度を見かけ上小さくしても値が改善されるとは限らない指標である。

$$\mathrm{AGFI} = 1 - \frac{n(n+1)}{2df}(1 - \mathrm{GFI})$$

一般に AGFI≦GFI の関係がある。

情報量基準

AIC（Akaike’s Information Criterion）

これは一般の統計モデルの評価指標として有名であり、以下で定義される。

$\mathrm{AIC} = \chi^2 - 2df \ \left[= (N-1) f_{ML} + 2p + const. \right]$　　p は自由パラメータ数

この値が小さいほど良いモデルとされる。この指標には、標本数が多い場合、自由度が小さい（パラメータ数が多い）モデルが良いモデルと判断される傾向がある。

CAIC（Consistent Akaike’s Information Criterion）

これは AIC の標本数の影響を抑えた指標である。

$$\mathrm{CAIC} = \chi^2 - (\log(N)+1)df$$

自由度当りのモデルの分布と最尤モデルとの乖離を表す指標

RMSEA（Root Mean Square Error of Approximation）

RMSEA≦0.05 で当てはまりが良い、≧0.1 で当てはまりが悪いと言われている。

$$\mathrm{RMSEA} = \sqrt{\max\left\{ f_{ML}/df - 1/(N-1), 0 \right\}}$$

13.4　プログラムの利用法（詳細）

ここでは 4 節の図 1 で述べた例を用いてプログラムの動作を説明する。プログラムを起動すると図 1a のような実行画面が表示される。これはできるだけ簡易化した画面である。この中で拡張メニューボタンをクリックすると図 1b のような拡張実行画面が表示される。

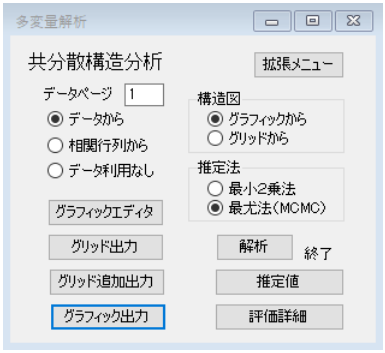


図 1a　実行画面

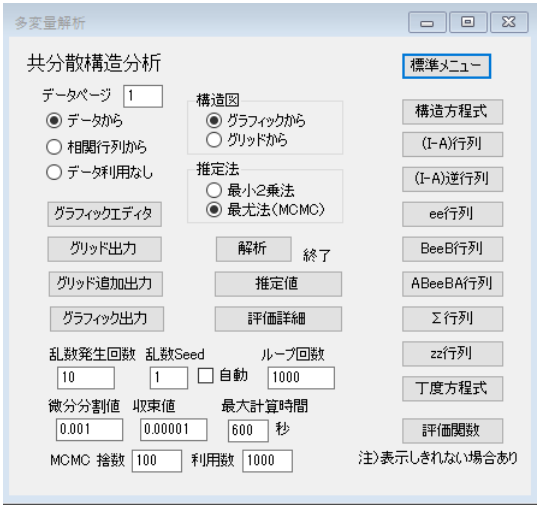


図 1b　拡張実行画面

拡張実行画面には細かな設定や、数式表示のためのボタンが含まれている。以後すべての機能が揃った拡張実行画面をもとに説明していく。この中の「グラフィックエディタ」、「グラフィック入力」、「グラフィック出力」ボタンについては、2 節や総合マニュアルのツールのところで説明しているので、ここでは触れない。

共分散構造分析のデータは基本的にデータ構造を記述したページと観測変数のデータ値を表すページに分かれる。前者を図 2 に示す（共分散構造分析 12（演習サンプル）.txt）。後者については通常の統計データの画面である。

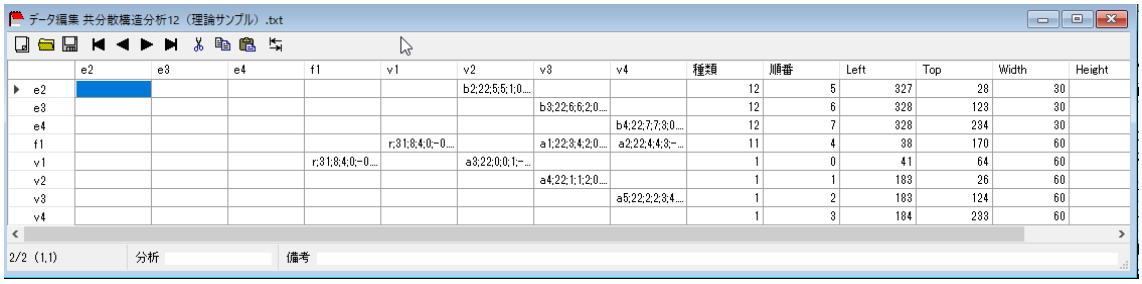


図 2　構造データ

分析は、メインメニュー左上の「データページ」テキストボックスに観測値のページ番号を記入して実行する。１つの観測データに複数のモデルを考える場合は、データを １ ページ目にして、2 ページ目以降を構造データにするのがよい。

最初に共分散構造分析の基礎となる数式について表示結果を説明する。図 1b の拡張分析実行画面の「構造方程式」ボタンをクリックすると構造方程式が図 3 のように表示される。

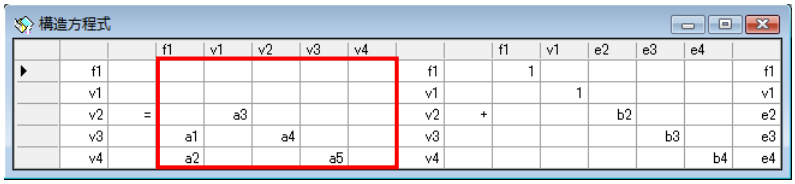


図 3　構造方程式

ここに 4 節（1）式中の行列 **A** は図の四角形で囲まれた部分である。実行画面の「(I-A) 行列」ボタンクリックすると図 4 のように **I**−**A** 行列の結果が表示される。

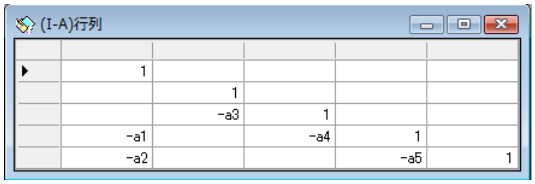


図 4　（I-A）行列

実行画面の「(I-A) 逆行列」ボタンクリックすると図 5 の **I**−**A** 逆行列が表示される。

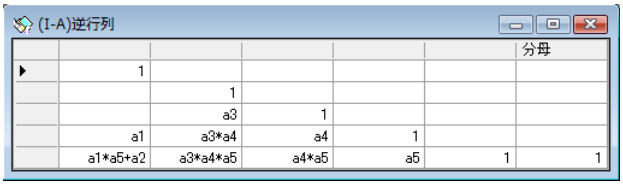


図 5　（I-A）逆行列

分母の列の最下行には **I**−**A** 行列の行列式の値が表示されている。実行画面の「ee 行列」ボタンをクリックすると図 6 のように行列 **H** が表示される。

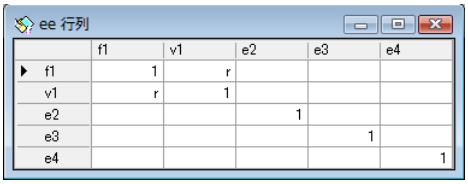


図 6　ee 行列

モデルで相関を仮定した場合はここにそのパラメータが残る。後は無相関と仮定される。

実行画面の「BeeB 行列」ボタンクリックすると図 7 のように行列 BHB′ が表示される。

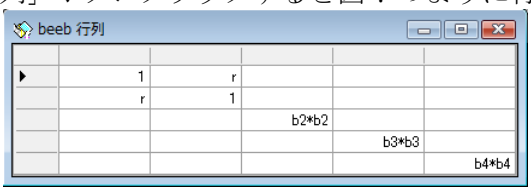


図 7　BHB’ 行列

「ABeeBA 行列」ボタンをクリックすると、図 8 のように行列 $(\mathbf{I}-\mathbf{A})^{-1}\mathbf{BHB}'(\mathbf{I}-\mathbf{A})'^{-1}$ が表示される。

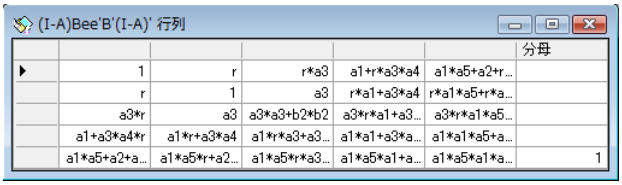


図 8　(I-A)$^{-1}$BHB’(I-A)’$^{-1}$ 行列

「Σ行列」ボタンクリックすると図 9 のように丁度方程式の左辺の $\boldsymbol{\Sigma(\theta)}$ が表示される。

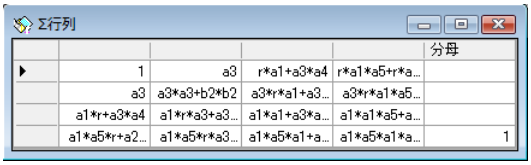


図 9　Σ行列

「zz 行列」ボタンクリックすると図 10 のように行列 $\mathbf{U}$ が表示される。これは観測変数の共分散行列（標準化解の場合は相関行列）である。

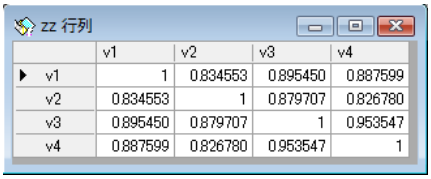


図 10　観測変数の相関行列

実行画面の「丁度方程式」ボタンクリックすると図 11 のように丁度方程式が表示される。

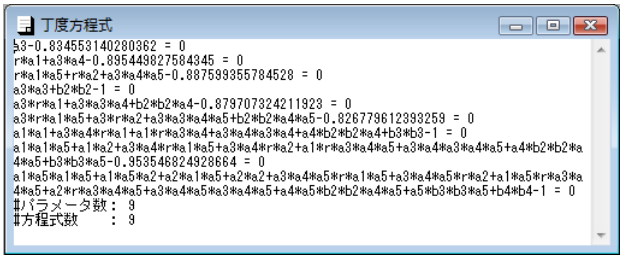


図 11　丁度方程式

「評価関数」ボタンをクリックすると図 12 のように評価関数が表示される。

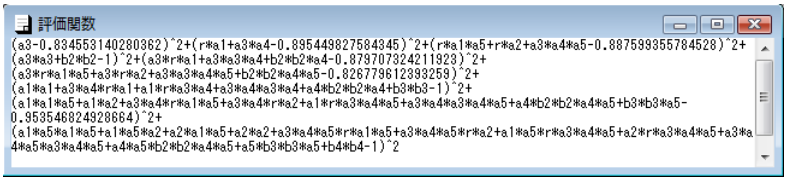


図 12　評価関数

これは最小 2 乗法における評価関数で、これを最小化するようにパラメータは選ばれる。最尤法の場合、表示が膨大になるのでかなり時間がかかる場合があるので注意を要する。

推定値については、「推定法」のグループの「最尤法」ラジオボタンを選択して、最初に「解析」ボタンクリックし、それから「推定値」をクリックすると図 13 のように表示される。

パラメータ行列　最尤法

	f1	v1	v2	v3	v4	e2	e3	e4
f1		0.9192		0.5781	0.3326			
v1	0.9192		0.8345					
v2				0.4361				
v3					0.6499			
v4								
e2			0.5509					
e3				0.2981				
e4					0.2692			
分散	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
評価関数値	0.0000							

図 13　最尤法の推定値

これは最尤法の推定値であるが、丁度方程式の解でもある。グラフィックエディタを用いて構造図を作成した場合は、構造図中にも推定値が表示される。さらに「評価詳細」ボタンをクリックすると、異なった形式の推定値と評価値が図 14a と図 14b のように表される。

パラメータ推定値

	変数		変数	推定値	標準誤差	検定値	両側確率
r	v1	<->	f1	0.9192	0.0620	14.8289	0.0000
a1	v3	<-	f1	0.5781	0.1353	4.2719	0.0000
a2	v4	<-	f1	0.3326	0.2424	1.3724	0.1699
a3	v2	<-	v1	0.8345	0.1023	8.1566	0.0000
a4	v3	<-	v2	0.4361	0.1265	3.4490	0.0006
a5	v4	<-	v3	0.6499	0.2319	2.8030	0.0051
b2	v2	<-	e2	0.5509	0.0723	7.6156	0.0000
b3	v3	<-	e3	0.2981	0.0849	3.5097	0.0004
b4	v4	<-	e4	0.2692	0.0514	5.2398	0.0000
評価関数値	0.0000						

図 14a　推定値の詳細表示

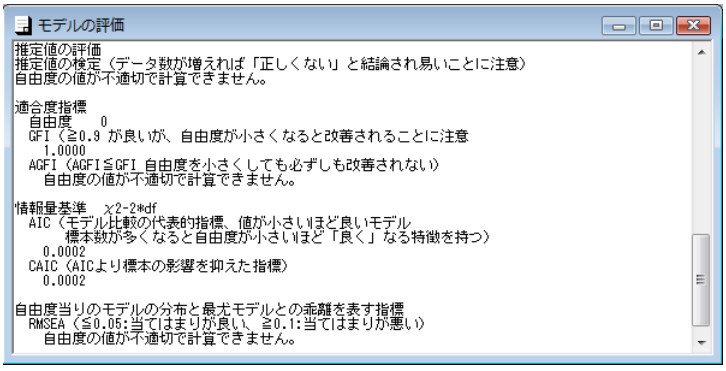


図 14b　モデルの評価（表示の後半部分）

最小 2 乗法の場合、推定値の検定部分や評価指標の適合度指標以外の部分は表示されない。

１４．パス解析

パス解析は観測変数間に線形の関係が仮定されるとき、因果関係の方向性を議論するために利用される手法で、共分散構造分析の特別な場合に相当する。ここではプログラムを実際に動かし、動きを見ながら、理論についても解説する。

メニュー［分析－多変量解析－パス解析］を選択すると、図１のようなパス解析実行画面が表示される。

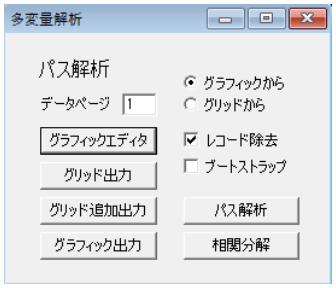


図１　パス解析実行画面

「グラフィックエディタ」ボタンで、グラフィックエディタを起動し、例えば図２のような構造図を描く。

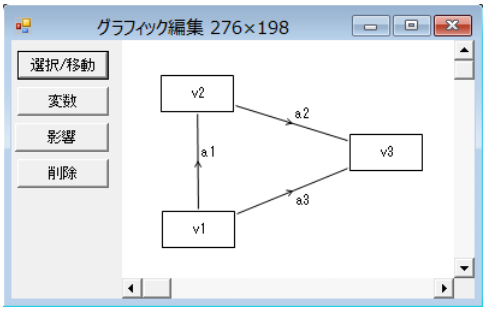


図２　構造図

共分散構造分析の構造図では誤差変数についても描画するが、パス解析では誤差変数の入り方は明らかであるため描画しない。図は単純に観測変数とそれらの間の影響だけで描かれる。但し、影響はすべての変数を結ぶものとし、影響のループは含まないものとする。

これらの変数名のデータは、グリッドエディタで、実行画面の「データページ」テキストボックスに指定されたページに含まれるものとする。プログラムはデータページの変数の中で、構造図の変数名に一致するデータを利用する。

変数間の構造データは、ラジオボタンにより、グリッドエディタとグラフィックエディタのどちらかを選ぶことができる。通常は、グラフィックエディタからの入力にしておき、良い構造が出来上がったら、「グリッド出力」か「グリッド追加出力」によって、構造データをグリッドエディタに移し、保存する。

実行画面で「パス解析」ボタンをクリックすると図３のように、構造間の影響の強さが表示される。

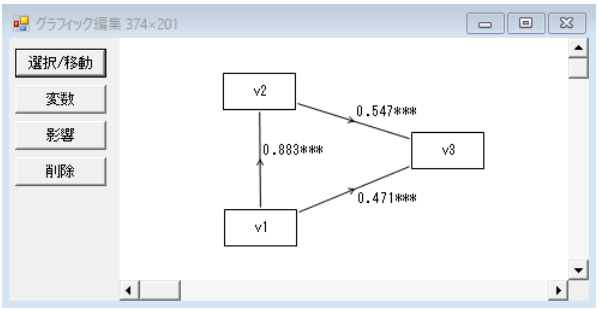


図３　パス解析結果

これを見て我々は影響の強さ、影響の方向の良し悪しを判定する。これらの影響の強さの数値は以下のような標準化した重回帰式から求められる。

v2=a1*v1+e2

v3=a3*v1+a2*v2+e3

ここで、e2 と e3 は誤差項であり、自分自身を除いて他の変数との相関はないものと考える。

これらの式から、各変数の相関について以下のような関係が分かる。

cov(v1,v2)=cov(v1,a1*v1+e2)=a1*cov(v1,v1)+cov(v1,e2)=a1

cov(v1,v3)=cov(v1,a3*v1+a2*v2+e3)=a3*cov(v1,v1)+a2*cov(v1,v2)=a3+a1*a2

cov(v2,v3)=cov(v2,a2*v2+a3*v1+e3)=a2*cov(v2,v2)+a3*cov(v2,v1)+cov(e3,v2)=a2+a1*a3

第 1 式について a1 を直接相関、第 2 式について、a3 を直接相関、a1*a2 を間接相関、第 3 式について、a2 を直接相関、a1*a3 を擬似相関と呼ぶ。直接相関は変数間を直接的に結ぶ関係、間接相関は変数間の影響を及ぼす方向通りにたどって行って 2 回以上でたどりつく関係、擬似相関は他の変数（ここでは v1）が両者に影響を及ぼしているような関係である。

左辺は相関係数であるので、これらの式は相関係数を、直接相関、間接相関、擬似相関に分解することに相当する。この関係は、実行画面の「相関分解」をクリックすることで示される。結果を図 4 に示す。


相関分解結果

	パス係数	t検定値	自由度	確率値	相関係数	直接効果	間接効果	擬似相関
▶ v1->v2	0.883	7.995	18	0.0000	0.883	0.883	0.000	0.000
v1->v3	0.471	5.950	17	0.0000	0.954	0.471	0.483	0.000
v2->v3	0.547	6.908	17	0.0000	0.963	0.547	0.000	0.416

図４　相関分解

直接効果、間接効果、擬似相関の合計が相関係数になっていることが分かる。

次に、もう少しだけ複雑なモデルを使って、これらの計算法を考えてみる。図 5 にモデルを示すが、ここではウィンドウの表示は省略する。

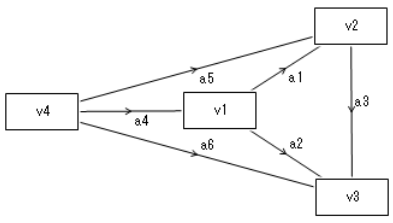


図５　パスの例２

ここではこの例を用いて v2, v3 への v4 の擬似相関を見てみよう。重回帰分析の計算を用いると、v2 と v3 の相関係数は以下のように与えられる。

cov(v2,v3)=cov(v2, a3*v2+a2*v1+a6*v4)

=a3+a1*a2+a5*a6+a5*a2*a4+a6*a1*a4

最初の項は直接相関、次の項は v1 からの擬似相関、下線の項は v4 からの擬似相関とみると、ある変数から影響をたどって行った道筋で、同一の変数を通る道筋を除いたものの総和となっている。この場合、v4 から v1 が v4 からの同一の道筋と考えると、そこを通る経路は v1 からの影響に置き換えると考える。これは v4 から v1 への影響が単純に係数の掛け算ではなく、

a1*a2*(a4*a4+cov(e1,e1))=a1*a2

のように回帰分析の際の誤差項の分散も含まれることから納得できる。

ブートストラップについて

様々な分布をするパラメータの区間推定や検定確率を求める手法としてブートストラップがある。これは n 個ある個体データから繰り返しを許して、n 個のデータを取り出す。これをブートストラップ標本という。ブートストラップ標本は取り出し方によって元のデータに近いものも、かなり異なっているものもある。これを 1 つのセットとして、十分な数のセットを取り出す（我々は 1000 セットから 2000 セット取り出している）。このデータセットを用いてモンテカルロシミュレーションを行い、パラメータの点推定値をセットの数だけ求める。この点推定値の平均がブートストラップで予想する平均である。さらにこれらの点推定値の分布から、上下 2.5%の値や、0 からのずれの検定確率などを求めることができる。

図 4 の相関分解のブートストラップによるパラメータ推定値を図 6 に示す。

相関分解ブートストラップ結果

	パス係数	t検定値	自由度	確率値	相関係数	BS直接効果	BS間接効果	BS擬似相関	直接効果Pr	間接効果Pr	擬似相関Pr
▶ v1->v2	0.883	7.995	18	0.0000	0.883	0.874	0.000	0.000	0.000	1.000	1.000
v1->v3	0.471	5.950	17	0.0000	0.954	0.476	0.479	0.000	0.002	0.000	1.000
v2->v3	0.547	6.908	17	0.0000	0.963	0.544	0.000	0.413	0.000	1.000	0.002

図 6　相関分解の係数のブートストラップ推定

例えば、直接効果は、点推定値と理論から求めた確率値、ブートストラップの平均とモンテカルロシミュレーションによる確率値の両方が表示されているので比較しやすい。

参考文献

[1] 多変量解析法入門, 永田靖, 棟近雅彦, サイエンス社, 2001.

１５．多次元尺度構成法

15.1　多次元尺度構成法とは

多次元尺度構成法（MDS: Multi Dimensional Scaling）は個体間に与えられた、類似度または非類似度（距離）を元に各個体の位置を求める手法である。個体間の非類似度がユークリッド空間上の距離として与えられる場合を計量 MDS、非類似度が順序のみ意味を持つ場合を非計量 MDS と呼ぶ。

個体iと個体 j （$1 \le i, j \le n$）の距離を d_{ij} とし、距離が以下の関係を満たすとき計量 MDS の手法が利用できる。

$$d_{ij} \ge 0, \quad d_{ij} = d_{ji}, \quad d_{ij} + d_{jk} \ge d_{ik}$$

今、 p 次元のユークリッド空間中の個体iの位置を $x_{i\alpha}$ （$1 \le \alpha \le p < n$）とする。個体iと個体 j との距離 d_{ij} は以下のように求められる。

$$d_{ij} = \left(\sum_{\alpha=1}^{p} (x_{i\alpha} - x_{j\alpha})^2 \right)^{1/2}$$

原点を個体の重心に設定するものとして、距離の関係から、元の位置座標を推測する手法を計量 MDS という。

非計量 MDS では、非類似度 s_{ij} を用いるが、これをディスパリティと呼ばれる量 $\hat{d}_{ij}$ に変換して利用する。これらは以下の関係を満たすようにする。

$$s_{ij} > s_{kl} \Rightarrow \hat{d}_{ij} \ge \hat{d}_{kl}, \quad s_{ij} = s_{kl} \Rightarrow \hat{d}_{ij} = \hat{d}_{kl}$$

ディスパリティ $\hat{d}_{ij}$ の導入は、矛盾を含む非類似度 s_{ij} を、矛盾なく求められる距離 d_{ij} を使って、順序関係を変えずにできるだけ実現可能な値に近づける操作と考えられる。

ディスパリティが求められたら、その値にできるだけ近づけるように d_{ij} を構成する。その基準をストレスと呼び、以下のように定義する。

$$S = \sqrt{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_{ij} - \hat{d}_{ij})^2 \Big/ \sum_{i=1}^{m} \sum_{j=i+1}^{m} d_{ij}^2} \qquad \text{ここに、} d_{ij} = \left(\sum_{\alpha=1}^{p} \left| x_{i\alpha} - x_{j\alpha} \right|^t \right)^{1/t}$$

一般にこの距離をミンコフスキー距離、tの値をミンコフスキー定数と呼ぶ。特にミンコフスキー定数が 2 の場合がユークリッド距離である。計算ではこのストレス S を最小化するように、Newton 法を実行している。

次元数 p を増やして行く際のストレスの変化を表す折れ線グラフ（ストレスプロット）を描き、どの次元から適合度が良くなるか調べる。また、s_{ij} の値を横軸に取り、縦軸にその値に対応する $\hat{d}_{ij}$ 及び d_{ij} の値を 2 種類のマーカーでポイントする。これをシェパードダイアグラムと呼ぶ。$\hat{d}_{ij}$ の値は同じ値を取るものがあるので、$\hat{d}_{ij}$ の上下に d_{ij} が散らばる傾向があるが、これらの点が $\hat{d}_{ij}$ に近く、s_{ij} の大きさによる逆転が起こらないほど適合度は

高いとされる。推測されたデータの点 $x_{i\alpha}$ の２つの次元について、平面上に点を描いて位置を確かめることも多い。

15.2　プログラムの利用法

メニュー［分析－多変量解析他－関係分析手法－多次元尺度構成法］を選択すると図１のような多次元尺度構成法実行画面が表示される。

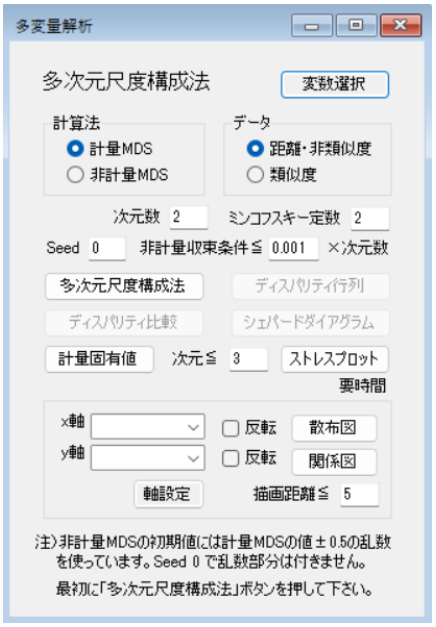
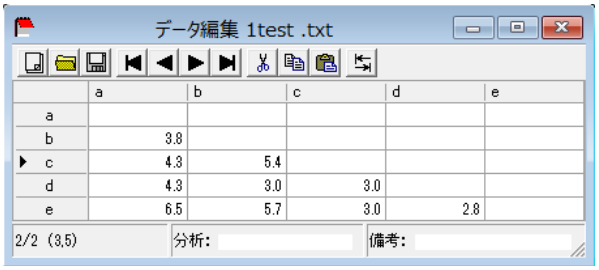


図１　多次元尺度構成法実行画面

データには、図２のように、類似度が低いほど大きな値を取る非類似度データ（または距離データ）か、類似度が高いほど大きな値を取る類似度データを用いる。これらの選択は「データ」グループボックスで指定する。



	a	b	c	d	e
a					
b	3.8				
c	4.3	5.4			
d	4.3	3.0	3.0		
e	6.5	5.7	3.0	2.8	

図２　非類似度データ

非類似度データの場合、対角成分は空欄か０にする。類似度データの場合、対角成分は空欄か、最も大きな値を取るものとする。類似度データの場合はこの最大の値から各セルの値を引いたものを非類似度データの値として用いている。データは図２のように三角データか、対称データを用いる。非対称データの場合の処理もできるが、我々のプログラムでは、２つの対応するデータの平均を取ることで対称化して利用している。

計量 MDS か非計量 MDS かは「計算法」グループボックスで指定する。計量 MDS の場合、ミンコフスキー定数は通常２で考える。このデータでは次元数を２として、「変数選択」

した後、「多次元尺度構成法」ボタンをクリックすると図３のような位置座標に関する実行結果が表示される。


座標

	1次元	2次元
a	2.844	-1.747
b	2.308	2.019
c	-1.463	-1.846
d	-0.473	0.974
e	-3.216	0.600

図３　計量 MDS の実行結果

計算途中の非類似度行列の固有値と固有ベクトルは、「計量固有値」ボタンをクリックすることで、図４のように表示される。


固有値と固有ベ...

	1次元	2次元
固有値	26.124	11.840
a	0.556	-0.508
b	0.452	0.587
c	-0.286	-0.536
d	-0.093	0.283
e	-0.629	0.174

図４　非類似度行列の固有値と固有ベクトル

「次元≦」を４で、「ストレスプロット」ボタンをクリックすると、図５のようなグラフが表示される。

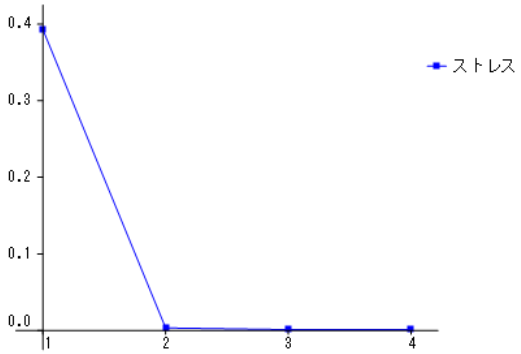


図５　ストレスプロット

ストレス値の評価は、0.2：良くない、0.1：悪くはない、0.05：良い、0.025：非常に良い、というように言われている。この例の場合だと、2 次元の段階で評価が良くなっているので、2 次元の結果を受け入れる。

2 次元の実行結果の位置を図として表示するために、「軸設定」ボタンで軸を選択し（この場合は自動的に 2 つの次元）、「散布図」ボタンをクリックすると、図６のような結果が表示される。

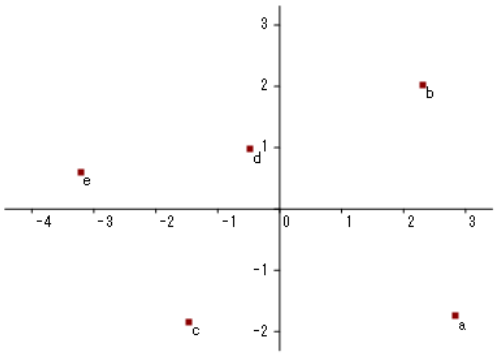


図６　位置関係結果

次に、参考文献１）にある例題を用いて非計量 MDS の操作を説明する。図 7 に距離行列を示す。これは類似度データである。

	クラウン	セドリック	サニー	マークⅡ	カローラ	スカイライン	マーチ	ヴィッツ	RAV4	パジェロ
クラウン	10									
セドリック	9	10								
▶ サニー	6	7	10							
マークⅡ	7	9	8	10						
カローラ	5	6	8	8	10					
スカイライン	2	3	6	3	6	10				
マーチ	2	3	5	4	7	6	10			
ヴィッツ	1	2	4	3	5	5	9	10		
RAV4	1	1	2	1	3	3	7	8	10	
パジェロ	2	3	3	4	5	2	5	5	4	10

図７　非計量、類似度データ

実行画面の「計算法」で「非計量 MDS」を選び、「データ」グループボックスで「類似度」を選ぶ。「次元数」を 2 にして、すべての変数を選択し、「多次元尺度構成法」ボタンをクリックすると、図 8 のような結果が表示される。

	1次元	2次元
▶ クラウン	4.057	-0.873
セドリック	3.639	-0.527
サニー	1.958	1.633
マークⅡ	3.080	-0.709
カローラ	0.744	0.570
スカイライン	-1.138	3.832
マーチ	-2.693	0.543
ヴィッツ	-3.737	0.072
RAV4	-4.563	-1.131
パジェロ	-1.608	-3.492

図８　非計量 MDS の実行結果

この解は一度計量 MDS の計算し、その解を Newton 法の初期値として非計量 MDS の計算を行ったものである。これは「Seed」テキストボックスがデフォルトの 0 のとき与えられる。計量 MDS の結果からずれた値を初期値として与える場合は、「Seed」の値を 0 以外の正の値とする。今はこのずれの範囲を±0.25×MDS の成分の絶対値の最大値としている。

実行画面の「ディスパリティ行列」ボタンをクリックすると、図 6 の類似度データに対応するディスパリティを図 9 のように表示する。但し、類似度データは、非類似度＝類似度最大値－類似度、によって、非類似度に変更されている。

ディスパリティ行列

	クラウン	セドリック	サニー	マークⅡ	カローラ	スカイライン	マーチ	ヴィッツ	RAV4	パジェロ
クラウン	0.0000	0.8391	3.8135	2.6355	4.6990	7.5870	7.5870	8.7703	8.7703	7.5870
セドリック	0.8391	0.0000	2.6355	0.8391	3.8135	6.7177	6.7177	7.5870	8.7703	6.7177
サニー	3.8135	2.6355	0.0000	2.2557	2.2557	3.8135	4.6990	5.6701	7.5870	6.7177
マークⅡ	2.6355	0.8391	2.2557	0.0000	2.2557	6.7177	5.6701	6.7177	8.7703	5.6701
カローラ	4.6990	3.8135	2.2557	2.2557	0.0000	3.8135	2.6355	4.6990	6.7177	4.6990
スカイライン	7.5870	6.7177	3.8135	6.7177	3.8135	0.0000	3.8135	4.6990	6.7177	7.5870
マーチ	7.5870	6.7177	4.6990	5.6701	2.6355	3.8135	0.0000	0.8391	2.6355	4.6990
ヴィッツ	8.7703	7.5870	5.6701	6.7177	4.6990	4.6990	0.8391	0.0000	2.2557	4.6990
RAV4	8.7703	8.7703	7.5870	8.7703	6.7177	6.7177	2.6355	2.2557	0.0000	5.6701
パジェロ	7.5870	6.7177	6.7177	5.6701	4.6990	7.5870	4.6990	4.6990	5.6701	0.0000

図 9　ディスパリティ行列

ディスパリティ行列の作り方や初期値の与え方については今後検討の余地が大きい。

「ディスパリティ比較」ボタンをクリックすると、図 10 のように非類似度、ディスパリティ、距離を非類似度の昇順に並べた表が表示される。

ディスパリティ比較

	S	D.P.	Dist
セドリック-クラウン	1.0000	0.8391	0.5424
マークⅡ-セドリック	1.0000	0.8391	0.5878
ヴィッツ-マーチ	1.0000	0.8391	1.1449
マークⅡ-サニー	2.0000	2.2557	2.5971
カローラ-サニー	2.0000	2.2557	1.6131
カローラ-マークⅡ	2.0000	2.2557	2.6629
RAV4-ヴィッツ	2.0000	2.2557	1.4592
マークⅡ-クラウン	3.0000	2.6355	0.9900
サニー-セドリック	3.0000	2.6355	2.7369
マーチ-カローラ	3.0000	2.6355	3.4378
RAV4-マーチ	3.0000	2.6355	2.5094

図 10　非類似度、ディスパリティ、距離比較表

この関係を図で表したものがシェパードダイアグラムである。「シェパードダイアグラム」ボタンをクリックすると図 11 のようなグラフが表示される。距離の点の散らばり方で、適合の良し悪しをみることができる。

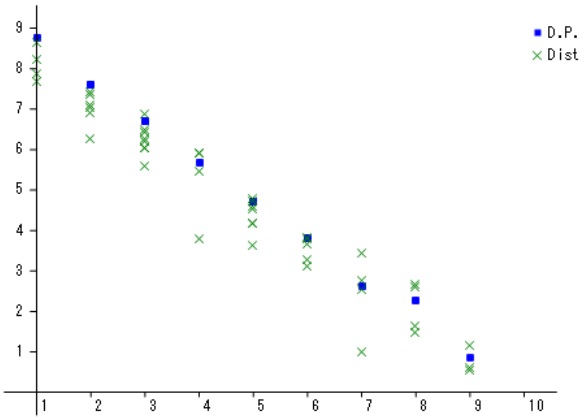


図 11　シェパードダイアグラム

軸を設定して「散布図」ボタンをクリックすると、図 12 のような位置表示のグラフが表示される。

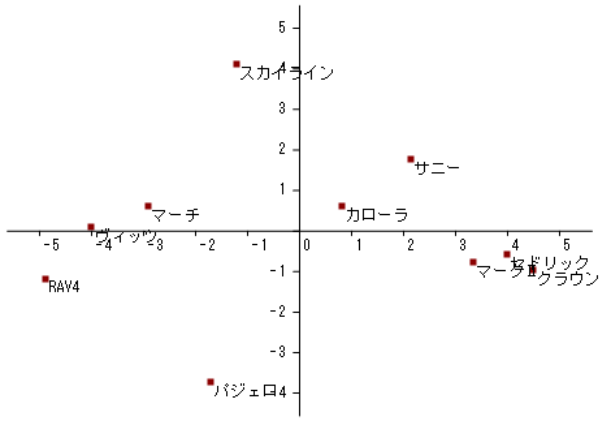


図 12　位置関係結果

次元を増やして行った際の、ストレスの変化は「ストレスプロット」ボタンをクリックすることで図 13 のように得られる。

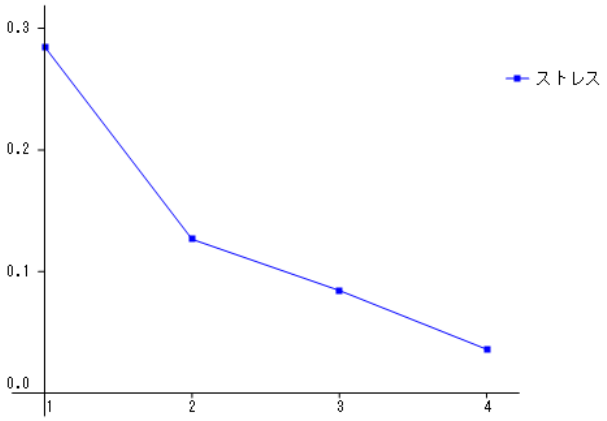


図 13　ストレスプロット

15.3　多次元尺度構成法の理論

多次元尺度構成法（MDS: Multi Dimensional Scaling）は個体間に与えられた、類似度または非類似度（距離）を元に各個体の位置（嗜好性等抽象的な位置関係も含む）を求める手法である。個体間の非類似度がユークリッド空間上の距離として与えられる場合を計量 MDS、非類似度が順序のみ意味を持つ場合を非計量 MDS と呼ぶ。我々はこれらの手法を順番に説明する。

1)　計量 MDS

個体iと個体j（$1 \le i, j \le n$）の距離をd_{ij}とし、距離が以下の関係を満たすとき計量 MDS の手法が利用できる。

$$d_{ij} \ge 0, \quad d_{ij} = d_{ji}, \quad d_{ij} + d_{jk} \ge d_{ik}$$

今、 p 次元のユークリッド空間中の個体 i の位置を $x_{i\alpha}$ （$1 \le \alpha \le p < n$）とする。個体 i と個体 j との距離 d_{ij} は以下のように求められる。

$$d_{ij} = \left(\sum_{\alpha=1}^{p} (x_{i\alpha} - x_{j\alpha})^2 \right)^{1/2}$$

この距離は原点の取り方に依存しないので、原点を個体の重心に設定するものとする。そのとき、

$$\bar{x}_\alpha = \frac{1}{n} \sum_{i=1}^{n} x_{i\alpha} = 0 \qquad (1)$$

である。原点から個体 i, j へのベクトルの内積を z_{ij} とすると、これは余弦定理により、以下のように与えられる。

$$z_{ij} = \sum_{\alpha=1}^{p} x_{i\alpha} x_{j\alpha} = \frac{1}{2} \left(d_{i0}^2 + d_{j0}^2 - d_{ij}^2 \right) \qquad (2)$$

ここに、d_{i0} は原点から個体 i までの距離である。(1) の関係式を使うと以下となるが、

$$\sum_{i=1}^{n} z_{ij} = \sum_{j=1}^{n} z_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij} = 0$$

例えば、この 2 番目の式に(2)式を代入して d_{i0} についての関係式を求め、これを(2)式に代入して z_{ij} を以下のように書き換えることができる。

$$z_{ij} = \frac{1}{2} \left(\sum_{k=1}^{n} \frac{d_{kj}^2}{n} + \sum_{k=1}^{n} \frac{d_{ik}^2}{n} - \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{d_{kl}^2}{n^2} - d_{ij}^2 \right) \qquad (3)$$

我々は求められた距離行列から、(3)式によってこの内積で作られた行列 $\mathbf{Z}$ を求め、(2)の最初の等号関係を用いて、以下に示す方法で位置 $x_{i\alpha}$ を求める。

我々はある次元 p を考え、以下の量 Q の最小化を考える。

$$Q = \sum_{k=1}^{n} \sum_{l=1}^{n} \left(z_{kl} - \sum\nolimits_{\beta=1}^{p} x_{k\beta} x_{l\beta} \right)^2 \qquad (4)$$

$\partial Q / \partial x_{i\alpha} = 0$ より、

$$\partial Q / \partial x_{i\alpha} = -2 \sum_{k=1}^{n} \sum_{l=1}^{n} \left(z_{kl} - \sum\nolimits_{\beta=1}^{p} x_{k\beta} x_{l\beta} \right) \sum\nolimits_{\gamma=1}^{p} (\delta_{ik} \delta_{\alpha\gamma} x_{l\gamma} + \delta_{il} \delta_{\alpha\gamma} x_{k\gamma})$$

$$= -4 \sum_{k=1}^{n} \sum_{l=1}^{n} \left(z_{kl} - \sum\nolimits_{\beta=1}^{p} x_{i\beta} x_{l\beta} \right) x_{l\alpha} = 0$$

これを行列の形に書き直すと以下を得る。

$$(\mathbf{Z} - \mathbf{X}\mathbf{X}')\mathbf{X} = \mathbf{0}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_p \end{pmatrix} \qquad (5)$$

ここで、$\mathbf{X}'\mathbf{X}(p \times p)$ は対称行列なので、直交行列 $\mathbf{U}(p \times p)$ を使って対角化できる。

$$\mathbf{U}'\mathbf{X}'\mathbf{X}\mathbf{U} = \mathbf{\Lambda} \qquad (6)$$

これを用いると $(\mathbf{Z} - \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}')\mathbf{X}\mathbf{U} = \mathbf{0}$ より、

$$\mathbf{Z}\mathbf{X}\mathbf{U} = \mathbf{X}\mathbf{U}\mathbf{\Lambda}$$

$\mathbf{H}=\mathbf{XU}$，$\mathbf{H}=\begin{pmatrix}\mathbf{h}_1 & \cdots & \mathbf{h}_p\end{pmatrix}$とすると、(5)と(6)から、

$$\mathbf{ZH}=\mathbf{H\Lambda},\quad \mathbf{H'H}=\mathbf{\Lambda} \qquad (7)$$

また、$\mathbf{XX'}=\mathbf{XUU'X'}=\mathbf{HH'}$であることから、(4)も(5)もすべて$\mathbf{H}$を使って書き換えられる。ここで$\mathbf{H}$を構成する縦ベクトルで(7)式を考えると、

$$\mathbf{Zh}_\alpha=\lambda_\alpha\mathbf{h}_\alpha,\quad \mathbf{h}'_\alpha\mathbf{h}_\alpha=\lambda_\alpha$$

これは行列$\mathbf{Z}$の固有方程式と固有ベクトル$\mathbf{h}_a$の規格化条件である。これを

$$\mathbf{h}_\alpha=\sqrt{\lambda_\alpha}\mathbf{y}_\alpha,\quad \mathbf{y}'_\alpha\mathbf{y}_\alpha=1,\quad \mathbf{H}=\mathbf{Y\Lambda}^{1/2}$$

のように規格化しなおすと、以下の関係を得る。

$$\mathbf{X}=\mathbf{HU'}=\mathbf{Y\Lambda}^{1/2}\mathbf{U'},\quad \mathbf{Y'Y}=\mathbf{I}$$

ここで$\mathbf{\Lambda}^{1/2}$は$\mathbf{\Lambda}$の対角要素を平方根に置き換えた行列である。

最後に、$\mathbf{U'}$は回転の行列であるから、回転の自由度を使って、

$$\mathbf{X}=\mathbf{Y\Lambda}^{1/2},\quad \mathbf{Y'Y}=\mathbf{I}$$

と考えてもよい。成分で書くと以下の関係となる。

$$x_{i\alpha}=\sqrt{\lambda_\alpha}y_{i\alpha},\quad \sum_{i=1}^{n}y_{i\alpha}^2=1$$

2)　非計量 MDS

非計量 MDS では、非類似度s_{ij}を用いるが、これをディスパリティと呼ばれる量$\hat{d}_{ij}$ に変換して利用する。これらは以下の関係を満たすようにする。

$$s_{ij}>s_{kl}\Rightarrow\hat{d}_{ij}\geq\hat{d}_{kl}$$

$$s_{ij}=s_{kl}\Rightarrow\hat{d}_{ij}=\hat{d}_{kl}$$

ディスパリティの生成は参考文献 2) に示された以下の手順で行う。ある手法で（我々のプログラムでは非類似度s_{jk}を用いた計量 MDS の手法)、位置が求まっているとする。その位置から距離d_{jk}を求める。非類似度s_{jk}を小さい順に並べ、それに$s_1, s_2, \cdots, s_l$ と番号を付ける。s_iに対応する非類似度s_{jk}に対応する距離d_{jk}についても同様に番号付けを行っておく。但し、s_iに同順位のものがある場合、それに対応するd_iについて、平均をとっておくものとする。

この準備を行った後、以下の手順を実行する。

１）$\hat{d}_1=d_1$とする。

２）$(k-1)$番目までの$\{\hat{d}_i\}$を作ったとする。

３）$d_k\geq\hat{d}_{k-1}$のとき、$\hat{d}_k=d_k$と定める。２）に行き、$\hat{d}_{k+1}$に移る。

４）$d_k<\hat{d}_{k-1}$のとき$\hat{d}_k$とその前の値を以下のように決定、変更する。２）に行き、$\hat{d}_{k+1}$に移る。

$i=1,2,\cdots,k-2$ と順に変えて、以下を満たす最小のiを見つける。

$$\hat{d}_{\cdot} = \frac{1}{i+1}\left(d_k + \sum_{j=1}^{i} \hat{d}_{k-j} \right) \geq \hat{d}_{k-i-1}$$

見つかれば、$\hat{d}_k = \hat{d}_{k-1} = \cdots = \hat{d}_{k-i} = \hat{d}_{\cdot}$ とする。

見つからなければ、$\hat{d}_k = \hat{d}_{k-1} = \cdots = \hat{d}_1 = \frac{1}{k}\sum_{j=0}^{k-1} \hat{d}_{k-j}$ とする。

５）$\hat{d}_n$ を定めたとき、プロセスを終了する。

ディスパリティ $\hat{d}_{ij}$ の導入は、矛盾を含む非類似度 s_{ij} を、矛盾なく求められる距離 d_{ij} を使って、順序関係を変えずにできるだけ実現可能な値に近づける操作と考えられる。

ディスパリティが求められたら、その値にできるだけ近づけるように再度 d_{ij} を構成しなおす。その基準をストレスと呼び、以下のように定義する。

$$S = \sqrt{\sum_{i=1}^{n}\sum_{j=i+1}^{n} (d_{ij} - \hat{d}_{ij})^2 \Big/ \sum_{i=1}^{m}\sum_{j=i+1}^{m} d_{ij}^2} \qquad \text{ここに、} d_{ij} = \left(\sum_{\alpha=1}^{p} \left| x_{i\alpha} - x_{j\alpha} \right|^t \right)^{1/t}$$

一般にこの距離をミンコフスキー距離、t の値をミンコフスキー定数と呼ぶ。特ミンコフスキー定数が 2 の場合がユークリッド距離である。我々のプログラムでは、S の最適化の方法は最急降下法を用い、$x_{i\alpha}$ の初期値には、元の計量 MDS から求めた値を使っている。ストレスの定義は参考文献 2）で別の定義を示しているが、ここでは参考文献 1）の定義に従っている。

次元数 p を増やして行く際のストレスの変化を表す折れ線グラフ（ストレスプロット）を描き、どの次元から適合度が良くなるか調べる。また、s_{ij} の値を横軸に取り、縦軸にその値に対応する $\hat{d}_{ij}$ 及び d_{ij} の値を 2 種類のマーカーでポイントする。これをシェパードダイアグラムと呼ぶ。$\hat{d}_{ij}$ の値は同じ値を取るものがあるので、$\hat{d}_{ij}$ の上下に d_{ij} が散らばる傾向があるが、これらの点が $\hat{d}_{ij}$ に近く、s_{ij} の大きさによる逆転が起こらないほど適合度は高い。推測されたデータの点 $x_{i\alpha}$ の α の２つの次元（$\alpha = 1, 2$ の場合が多いが）について平面上に点を描いて、位置を確かめることも多い。

参考文献

1）多変量解析法入門, 永田靖, 棟近雅彦, サイエンス社, 2001.

2）関連性データの解析法 多次元尺度構成法とクラスター分析法, 齋藤堯幸, 宿久洋, 共立出版, 2006.

１６．局所重回帰分析

16.1　局所重回帰分析とは

重回帰分析や数学の分冊の中で述べる非線形最小 2 乗法の予測手法は、パラメータを含んだ関数形を仮定し、最小 2 乗法によってパラメータの値を定め、予測関数を確定するものであった。しかし、局所重回帰分析は要求点を与えることによって、その近傍の点による重回帰分析の結果から直接予測値を求める方法で、関数形を必要としない興味深い予測手法である。

標準的な重回帰分析は、目的変数 $y_\lambda\ (\lambda=1,2,\cdots,N)$ と説明変数 $x_{i\lambda}\ (i=1,2,\cdots,p)$ の線形結合 $Y_\lambda=\sum_{i=1}^{p}b_i x_{i\lambda}+b_0$ との差の 2 乗の和 EV を最小にするようにパラメータ $b_i\ (i=0,1,2,\cdots,p)$ を決定する。ここに EV は以下で与えられる。

$$EV=\sum_{\lambda=1}^{N}\left(y_\lambda-Y_\lambda\right)^2=\sum_{\lambda=1}^{N}\left(y_\lambda-\sum_{i=1}^{p}b_i x_{i\lambda}-b_0\right)^2$$

これに対して局所重回帰分析は、各観測値に対してウェイト w_λ をかけて以下の EV' を最小化する。

$$EV'=\sum_{\lambda=1}^{N}w_\lambda\left(y_\lambda-Y_\lambda\right)^2=\sum_{\lambda=1}^{N}w_\lambda\left(y_\lambda-\sum_{i=1}^{p}b_i x_{i\lambda}-b_0\right)^2$$

ウェイト w_λ は以下のように求める。まず、説明変数についての要求点 x_i^r とバンド幅（調整パラメータ） $p\ (>0)$ を定める。要求点は局所重回帰分析のウェイトの中心を表す点である。次に標準化された観測点 $\tilde{x}_{i\lambda}=(x_{i\lambda}-\bar{x}_i)/\sigma_i$ と標準化された要求点 $\tilde{x}_i^r=(x_i^r-\bar{x}_i)/\sigma_i$ との間の以下のユークリッド距離を求める。

$$d_\lambda=\sqrt{\sum_{i=1}^{p}\left(\tilde{x}_{i\lambda}-\tilde{x}_i^r\right)^2}$$

但し、標準化の際の標準偏差は不偏分散からのものとする。

この距離 d_λ について、その平均を $\bar{d}$ 、不偏分散からの標準偏差を σ_d とし、これらを用いて、ウェイト w_λ を以下のように定義する。

$$w_\lambda=\exp\left[-\left(d_\lambda/p\sigma_\Gamma\right)^2\right]$$

これによって要求点の近傍の点にウェイトをかけて最小 2 乗法の解を求めることになる。

局所重回帰分析は要求点の近傍で成り立つ近似手法であるので、通常の RMSE（残差の 2 乗平均の平方根）や重相関係数の指標は使えず、その信頼性を求める指標は 1 個抜き交差検証法（HOOCV : Leave-One-Out Cross-Validation）を用いて与える。即ち、データ中の 1 点を抜き、その説明変数の座標 $x_{i\lambda}$ を要求点とし、残りの点で局所重回帰分析を行い、要求点の予測値 Y_λ を求める。元々この点には実測値 y_λ があるので予測の誤差が求められる。局所重回帰分析の精度の指標はこの実測値と予測値を利用し、通常の RMSE や重回帰分析の重相関係数等の定義を用いて与える。もちろんこの指標はバンド幅に影響される。

16.2　プログラムの利用法

メニュー［分析－多変量解析他－重回帰分析－局所重回帰分析］をクリックすると図 1 のような局所重回帰分析の実行画面が表示される。

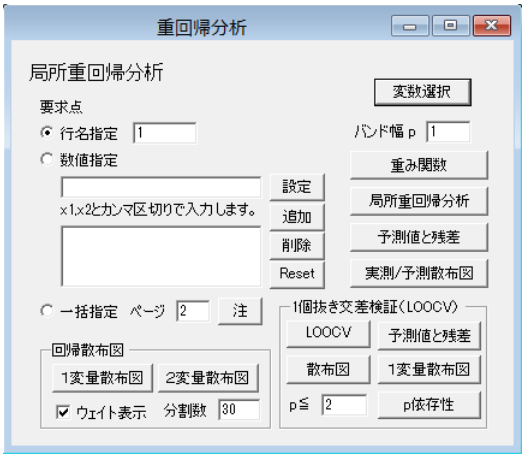


図 1　局所重回帰分析実行画面

通常の重回帰分析と同様に「変数選択」で、目的変数、説明変数の順に変数を選ぶ。要求点は、「行名指定」でデータから選択するか、「数値指定」で直接数値を入力する。行名指定は、データの行名（レコード名）の部分の表示で指定する。行名が見当たらない場合は、実行の際にメッセージが表示される。数値指定の場合は、テキストボックスに説明変数の値をカンマ区切りで入力する。複数の要求点を調べることが必要であるので、プログラムには入力した値を保存しておく機能が付いている。テキストボックスに書いた要求点のデータは、「追加」ボタンで下のリストボックスに追加保存される。リストボックスのデータは選択して、「設定」ボタンでテキストボックスに呼び戻すことができる。また、選択して「削除」ボタンで 1 つだけリストから削除でき、「Reset」ボタンですべて削除することができる。変数選択の場合と同じ要領で活用できる。

バンド幅を適当な値（ここでは 1）に設定し、適当な行名を指定して「局所重回帰分析」ボタンをクリックすると、図 2 のような分析結果が得られる。


偏回帰係数と要求点

	偏回帰係数	標準化係数	要求点	標準化点
▶ 説明1	0.3812	0.1982	31	-0.9408
説明2	0.4113	0.2146	19	-1.2370
切片/要求予測値	51.6478	-0.5032	71.2791	-0.9552

図 2　偏回帰係数の出力結果

重回帰式による推測結果と各観測点のウェイト値は「予測値と残差」ボタンで図 3 のように表示される。

予測値と残差

	実測値	予測値	残差	ウェイト
1	66	71.2791	-5.2791	1.0000
2	89	82.2936	6.7064	0.4037
3	73	75.1613	-2.1613	0.6670
4	80	76.4653	3.5347	0.5477
5	75	74.8603	0.1397	0.9383
6	79	67.4070	11.5930	0.8994
7	81	80.4278	0.5722	0.0679
8	66	72.2221	-6.2221	0.9224
9	70	75.4421	-5.4421	0.7975
10	78	77.7895	0.2105	0.6696

図３　実測値と予測値

実測値と予測値の関係は「実測/予測散布図」をクリックすると、図４のように表示される。

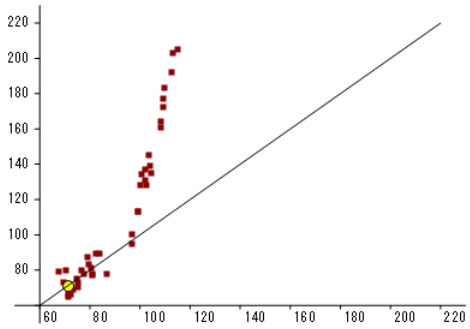


図４　実測/予測値散布図１

図中の黄色い点は要求点で、直線は実測と予測が同じであるとする直線である。要求点近傍の点の予測がうまく行っている状況が見える。

偏回帰係数は、要求点とバンド幅に大きく影響を受ける。要求点を変更したときの結果を図５に示す。今度は別の点の予測がうまく行っている。

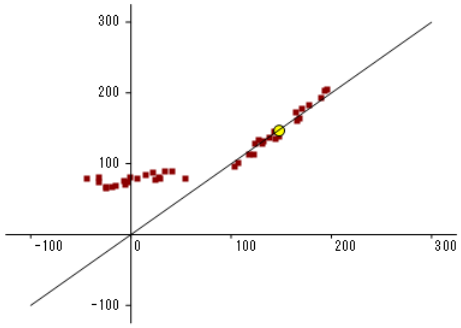


図５　実測/予測値散布図２

実際の x,y 軸の上で回帰直線を引いてみる。変数を目的変数と説明変数を１つにして、「1変量回帰散布図」を描くと図６のようになる。２つの図は要求点を変えて描いている。

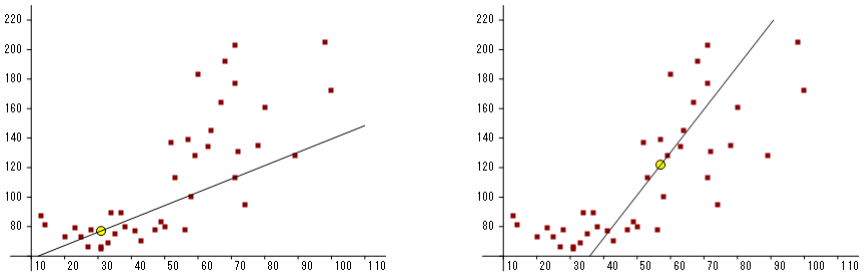


図６　１変量回帰散布図（p=1）

これは、データの散布図であり、図中の直線は回帰直線である。要求点によって回帰直線が変化しているのが分かる。

また、実際の x,y,z 軸上で回帰平面を描いてみる。変数を目的変数と説明変数を２つにして、「２変量回帰散布図」を描くと図 7 のようになる。2 つの図は要求点を変えて描いている。

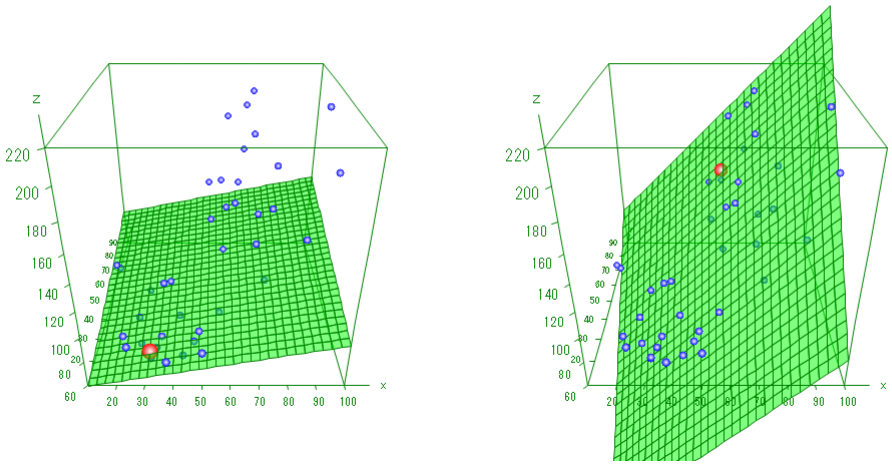


図７　２変量回帰散布図（p=1）

次にバンド幅を p=0.5 と p=5 にし、説明変数の数を１つにして、1 変数回帰散布図を描く。結果を図 8 に示す。

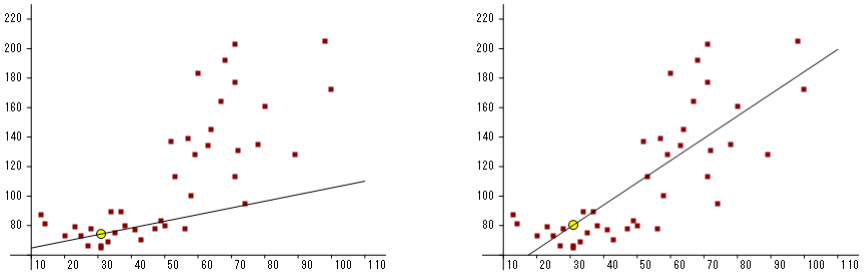


図８　図６左の要求点で p=0.5（左）と p=5（右）の１変量回帰散布図

バンド幅の値により、局所性が大きく変更を受けていることが分かる。右側の図は通常の回帰直線に近い。

分析メニューで「重み関数」ボタンをクリックすると 2 変数グラフ描画メニューが表示される。その中の「グラフ描画」ボタンをそのままクリックすると、図 9 左のような実際の重み関数のグラフ（この場合は 2 変数）が表示される。1 変数の場合は図 9 右のようなグラフになる。

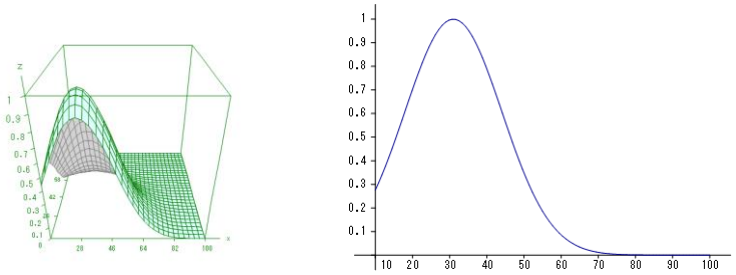
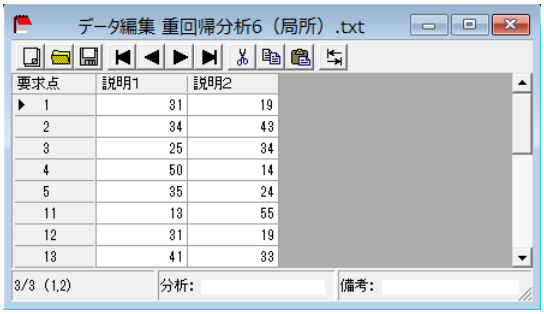


図 9　重み関数グラフ（左は 2 変数、右は 1 変数）

これまでは要求点を 1 点だけ指定したが、現実の分析では多くの要求点を一度に与えて予測値を求めることも考えられる。実行画面で、要求点の「一括指定」ラジオボタンを選択すると､別のページに与えられた複数の要求点のデータから一括で予測値を求めることもできる。要求点のページはラジオボタン右側の「ページ」テキストボックスに与える。デフォルトは 2 頁目になっているので必要なら変更する。要求点の頁の例を図 10 に示す。



要求点	説明1	説明2
1	31	19
2	34	43
3	25	34
4	50	14
5	35	24
11	13	55
12	31	19
13	41	33

図 10　要求点の一括指定

ここで注意することは、変数名を必ず正確に（全角半角や大文字小文字の区別を付けて）指定することである。分析では変数選択の数や順番が要求点の指定通りとは限らないので、プログラムでは変数名を探して順番等を合わせるようにしている。

一括指定した要求点を用いた場合は、重回帰式の偏相関係数などは重要でないので、結果は要求点と予測値を表形式で与える。要求点指定に空欄がある場合は、予測値の欄が空欄になる。予測値の出力例を図 11 に与える。

	予測値	説明1	説明2
1	71.279	31	19
2	84.700	34	43
3	75.704	25	34
4	78.734	50	14
5	74.550	35	24
11	82.866	13	55
12	71.279	31	19
13	80.777	41	33

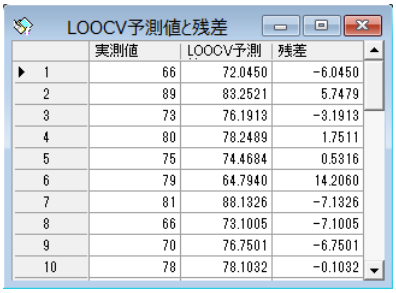
図 11　要求点一括指定の出力

局所重回帰分析の予測精度を与えるために、1 個抜き交差検証（LOOCV）を用いた RMSE と重相関係数を与える。「LOOCV」ボタンをクリックすると図 12 のような結果が表示される。

図 12　1 個抜き交差検証による RMSE と重相関係数

ここで採択率は、1 個抜いたデータで計算ができない場合があるので、計算できるデータ点の割合を示したものである。

この求めた予測値と実測値の具体的な値は 1 個抜き交差検証中の「予測値と残差」ボタンをクリックすることで図 13 のように与えられる。予測値が求められなかった部分は空白になっている。



	実測値	LOOCV予測	残差
1	66	72.0450	-6.0450
2	89	83.2521	5.7479
3	73	76.1913	-3.1913
4	80	78.2489	1.7511
5	75	74.4684	0.5316
6	79	64.7940	14.2060
7	81	88.1326	-7.1326
8	66	73.1005	-7.1005
9	70	76.7501	-6.7501
10	78	78.1032	-0.1032

図 13　1 個抜き交差検証による実測値と予測値

この関係は 1 個抜き交差検証中の「散布図」ボタンで、実測/予測散布図として図 14 のように与えられる。

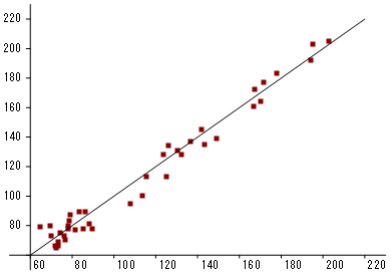


図 14　1 個抜き交差検証による実測/予測散布図

説明変数による予測値と実測値の関係は、1 変量の場合「1 変量散布図」をクリックして図 15 のように与えられる。この図の場合、特別に説明変数を 1 個だけにした。

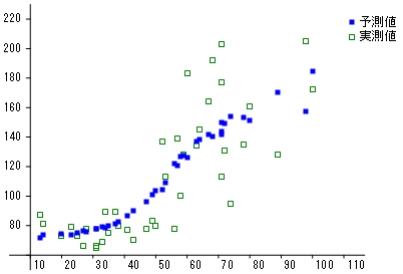


図 15　1 個抜き交差検証による 1 変量散布図

バンド幅によって、RMSE や重相関係数の値は変化する。「p 依存性」ボタンをクリックすると、RMSE のバンド幅 p の値による変化が図 16 のように示される。

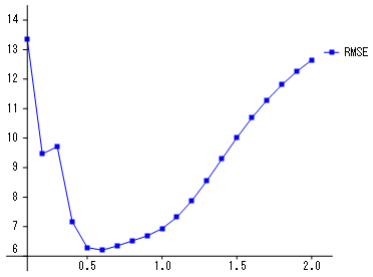


図 16　バンド幅の値による RMSE の変化

ここで、p=0.3 のところで値が急に大きくなっているが、この部分は 1 個抜き交差検証ですべての点が利用できなかった部分である。

前節で述べた局所重回帰分析の考え方とプログラムの利用法の要点をまとめておくと以下のようになる。

局所重回帰分析の目的

非線形な重回帰分析を、関数形を仮定せず直接予測値を求める方法で実現する。

予測式は？　→　要求点を与えた際の線形重回帰式　「局所重回帰分析」ボタン
（偏回帰係数の値　他の要求点では使えない）

予測値は？　→　要求点を与えて求める。「局所重回帰分析」ボタン

バンド幅 p とは？　→　どの程度要求点の近傍のデータを利用するかを決める値
大きいほど遠くまで利用する。∞で通常の重回帰分析

要求点を与えた予測の数値を見るには？　→　「予測値と残差」ボタン

要求点を与えた予測の状態を見るには？　→　「実測/予測散布図」ボタン
２変量までなら回帰散布図

１個抜き交差検証について

１個抜き交差検証（LOOCV）とは？　→　要求点を各データ点にし、その点を除いて予測する検証手法

予測の精度は　→　各要求点（=データ点）の実測値と予測値の相関係数（重相関係数）及び残差２乗平均の平方根（RMSE：小さいほど良い）

各点の予測と実測の値は？　→　１個抜き交差検証内の「予測値と残差」ボタン

各点の予測と実測の状態を見るには？　→　同「散布図」

予測へのバンド幅 p の依存性は？　→　「p 依存性」ボタン（最小のところが良い値）

問題

重回帰分析 6（局所）.txt の１頁目を読み込んで以下の問いに答えよ。

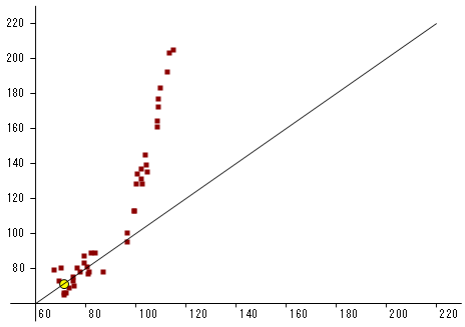
１）要求点を先頭のデータ（番号1）の位置にした際の、局所重回帰式を求めよ。
　目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
２）そのときの先頭のデータの実測値、予測値、残差、ウェイトを求めよ。
　実測値［　　］予測値［　　］残差［　　］ウェイト［　　］
３）そのときの実測／予測散布図を見よ。
４）そのときの２変量回帰散布図を見よ。
５）30番目のデータの実測値、予測値、残差、ウェイトを求めよ。
　実測値［　　］予測値［　　］残差［　　］ウェイト［　　］
６）要求点を30番目のデータの位置にした際の、局所重回帰式を求めよ。
　目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
７）そのときの先頭のデータの実測値、予測値、残差、ウェイトを求めよ。
　実測値［　　］予測値［　　］残差［　　］ウェイト［　　］
８）そのときの実測／予測散布図を見よ。
９）そのときの２変量回帰散布図を見よ。
10）要求点を（50,50）にしたときの局所重回帰式を求めよ。
　目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
11）そのときの実測／予測散布図を見よ。
12）バンド幅pを100にした場合の局所回帰式を求めよ。
　目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
13）そのときの実測／予測散布図を見よ。
14）予測値と残差のところで、ウェイトの大きさを見よ。

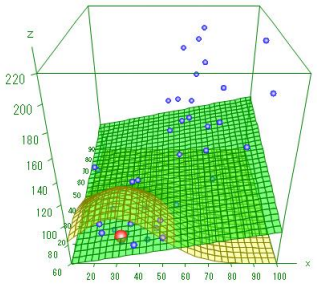
<u>バンド幅を元に戻して</u>

15）１個抜き交差検証をした場合のRMSEと重相関係数の値を求めよ。
　RMSE［　　］　重相関係数［　　］
16）１個抜き交差検証をした場合の1番の実測値、LOOCV予測値、残差を求めよ。
　実測値［　　］LOOCV予測値［　　］残差［　　］
17）１個抜き交差検証の予測の程度を散布図を使って見よ。
18）バンド幅pを動かして、RMSEが最小となるp値はどこか。P=［　　］

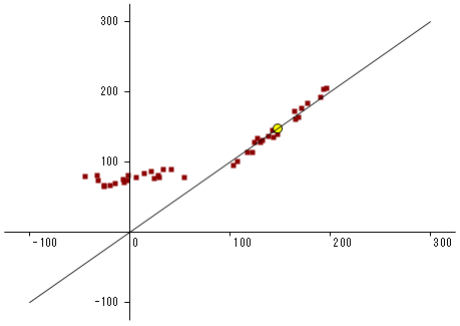
問題解答

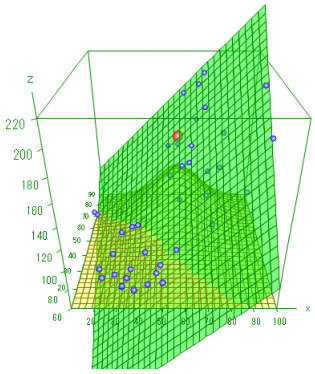
１）要求点を先頭のデータ（番号1）の位置にした際の、局所重回帰式を求めよ。
　目的変数＝［ 0.3812 ］説明変数１＋［ 0.4113 ］説明変数２＋［ 51.6478 ］
２）そのときの先頭のデータの実測値、予測値、残差、ウェイトを求めよ。
　実測値［ 66 ］予測値［ 71.279 ］残差［ -5.279 ］ウェイト［ 1.000 ］
３）そのときの実測／予測散布図を見よ。（左図）
４）そのときの２変量回帰散布図を見よ。（右図）



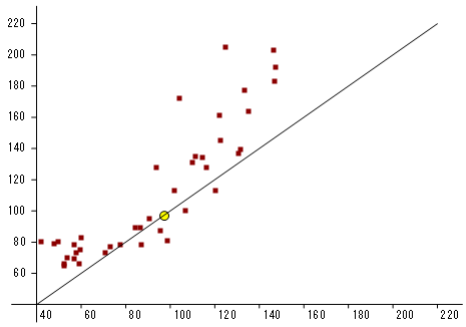


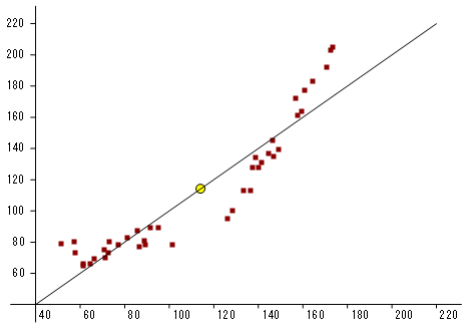
５）30 番目のデータの実測値、予測値、残差、ウェイトを求めよ。
実測値［ 139 ］予測値［ 104.222 ］残差［ 34.778 ］ウェイト［ 0.003 ］
６）要求点を 30 番目のデータの位置にした際の、局所重回帰式を求めよ。
目的変数＝［ 1.8413 ］説明変数１＋［ 2.2283 ］説明変数２＋［ -124.5259 ］
７）そのときの先頭のデータの実測値、予測値、残差、ウェイトを求めよ。
実測値［ 66 ］予測値［ -25.109 ］残差［ 91.109 ］ウェイト［ 0.000 ］
８）そのときの実測／予測散布図を見よ。（左図）
９）そのときの２変量回帰散布図を見よ。（右図）





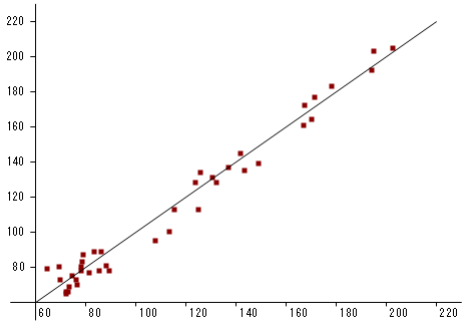
10）要求点を（50, 50）にしたときの局所重回帰式を求めよ。
目的変数＝［ 0.2164 ］説明変数１＋［ 1.3155 ］説明変数２＋［ 20.5844 ］
11）そのときの実測／予測散布図を見よ。（２つ下左）
12）バンド幅 p を 100 にした場合の局所回帰式を求めよ。
目的変数＝［ 0.9232 ］説明変数１＋［ 1.1434 ］説明変数２＋［ 10.8625 ］
13）そのときの実測／予測散布図を見よ。（下右）





14）予測値と残差のところで、ウェイトの大きさを見よ。
すべて１に近い値である。

15） １個抜き交差検証をした場合の RMSE と重相関係数の値を求めよ。
RMSE ［ 6.931 ］　重相関係数 ［ 0.987 ］
16） １個抜き交差検証をした場合の１番の実測値、LOOCV 予測値、残差を求めよ。
実測値 ［ 66 ］LOOCV 予測値 ［ 72.045 ］残差 ［ -6.045 ］
17） １個抜き交差検証の予測の程度を散布図を使って見よ。



18）バンド幅 p を動かして、RMSE が最小となる p 値はどこか。P=［ 0.6 ］

16.3　局所重回帰分析の理論

標準的な重回帰分析は、目的変数 $y_\lambda\ (\lambda=1,2,\cdots,N)$ と説明変数 $x_{i\lambda}\ (i=1,2,\cdots,p)$ の線形結合 $Y_\lambda=\sum_{i=1}^{p}b_i x_{i\lambda}+b_0$ との差の 2 乗和 L を最小にするようにパラメータ $b_i\ (i=0,1,\cdots,p)$ を決定する。ここに L は以下で与えられる。

$$L=\sum_{\lambda=1}^{N}\left(y_\lambda-Y_\lambda\right)^2=\sum_{\lambda=1}^{N}\left(y_\lambda-\sum_{i=1}^{p}b_i x_{i\lambda}-b_0\right)^2$$

これに対して局所重回帰分析は、各観測値に対してウェイト w_λ をかけて以下の L' を最小化する[1]。

$$L'=\sum_{\lambda=1}^{N}w_\lambda\left(y_\lambda-Y_\lambda\right)^2=\sum_{\lambda=1}^{N}w_\lambda\left(y_\lambda-\sum_{i=1}^{p}b_i x_{i\lambda}-b_0\right)^2$$

この解は、$\mathbf{b}={}^t\begin{pmatrix}b_0 & b_1 & b_2 & \cdots & b_p\end{pmatrix}$ として、以下のように求めることができる。

$$\mathbf{b}=\left({}^t\mathbf{\Omega\Pi\Omega}\right)^{-1}{}^t\mathbf{\Omega\Pi y} \qquad (1)$$

ここに、

$$\mathbf{y}={}^t\begin{pmatrix}y_1 & y_2 & \cdots & y_N\end{pmatrix},$$

$$\mathbf{\Omega}=\begin{pmatrix}1 & x_{11} & x_{21} & \cdots & x_{p1}\\ 1 & x_{12} & x_{22} & \cdots & x_{p2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{1N} & x_{2N} & \cdots & x_{pN}\end{pmatrix},\quad \mathbf{\Pi}=\begin{pmatrix}w_1 & 0 & \cdots & 0\\ 0 & w_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & w_N\end{pmatrix}$$

要求点 x_i^r の予測値 Y^r は、以下のように与えられる。

$$Y^r=\sum_{i=1}^{p}b_i x_i^r+b_0 \qquad (2)$$

ウェイト w_λ は以下のように求める。まず、説明変数についての要求点 x_i^r とバンド幅（調整パラメータ） $p\,(>0)$ を定める。要求点は局所重回帰分析のウェイトの中心を表す点である。次に標準化された観測点 $\tilde{x}_{i\lambda}=(x_{i\lambda}-\bar{x}_i)/\sigma_i$ と標準化された要求点 $\tilde{x}_i^r=(x_i^r-\bar{x}_i)/\sigma_i$ との間のユークリッド距離を求める。

$$d_\lambda=\sqrt{\sum_{i=1}^{p}\left(\tilde{x}_{i\lambda}-\tilde{x}_i^r\right)^2}$$

但し、標準化の際の標準偏差は不偏分散からのものとする。

この距離 d_λ について、その平均を $\bar{d}$ 、不偏分散からの標準偏差を σ_d とし、これらを用いて、ウェイト w_λ を以下のように定義する。

$$w_\lambda=\exp\left[-\left(d_\lambda/p\sigma_d\right)^2\right] \tag{3}$$

これによって要求点の近傍の点にウェイトをかけて最小２乗法の解を求めることになる。

標準化偏回帰係数については、標準化されたデータ $\tilde{y}_\lambda,\tilde{x}_{i\lambda}$ を用いて、以下のように求めることもできる。

$$\tilde{\mathbf{b}}=\left({}^t\tilde{\mathbf{\Omega}}\mathbf{\Pi}\tilde{\mathbf{\Omega}}\right)^{-1}{}^t\tilde{\mathbf{\Omega}}\mathbf{\Pi}\tilde{\mathbf{y}} \tag{4}$$

ここに、

$$\tilde{\mathbf{y}}={}^t\left(\tilde{y}_1\quad\tilde{y}_2\quad\cdots\quad\tilde{y}_N\right),\qquad \tilde{y}_\lambda=\frac{y_\lambda-\bar{y}}{\sigma_y}\quad\text{（不偏分散を用いた標準化）}$$

$$\tilde{\mathbf{\Omega}}=\begin{pmatrix}1&\tilde{x}_{11}&\tilde{x}_{21}&\cdots&\tilde{x}_{p1}\\1&\tilde{x}_{12}&\tilde{x}_{22}&\cdots&\tilde{x}_{p2}\\\vdots&\vdots&\vdots&\ddots&\vdots\\1&\tilde{x}_{1N}&\tilde{x}_{2N}&\cdots&\tilde{x}_{pN}\end{pmatrix},\qquad \mathbf{\Pi}=\begin{pmatrix}w_1&0&\cdots&0\\0&w_2&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&w_N\end{pmatrix}$$

別の書式で書くと以下となる。

$$\tilde{b}_i=\frac{\sigma_i}{\sigma_y}b_i,\qquad \tilde{b}_0=\frac{1}{\sigma_y}\left(b_0+\sum_{i=1}^{p}b_i\bar{x}_i-\bar{y}\right) \tag{5}$$

この関係は、以下のように求めることができる。

$$\begin{aligned}\frac{Y_\lambda-\bar{y}}{\sigma_y}&=\frac{1}{\sigma_y}\left(\sum_{i=1}^{p}b_ix_{i\lambda}+b_0-\bar{y}\right)\\&=\sum_{i=1}^{p}\frac{\sigma_i}{\sigma_y}b_i\,\frac{x_{i\lambda}-\bar{x}_i}{\sigma_i}+\frac{1}{\sigma_y}\left(b_0+\sum_{i=1}^{p}b_i\bar{x}_i-\bar{y}\right)\end{aligned}$$

通常の重回帰分析では $\bar{y}=\bar{Y}=\sum_{i=1}^{p}b_i\bar{x}_i+b_0$ であるから、標準化された定数項は０になるが、局所重回帰分析では一般に $\bar{y}\neq\bar{Y}$ であるので、標準化された定数項は０にならない。

偏回帰係数と標準化偏回帰係数の関係は、(5)式とは逆に以下のように書くこともできる。我々のプログラムではこの関係を利用している。

$$b_i = \frac{\sigma_y}{\sigma_i}\tilde{b}_i, \qquad b_0 = \sigma_y \tilde{b}_0 - \sum_{i=1}^{p} \frac{\sigma_y}{\sigma_i}\tilde{b}_i \bar{x}_i + \bar{y} \tag{6}$$

局所重回帰分析はバンド幅（調整パラメータ） p が無限大になるとウェイトがすべて 1 になり、通常の重回帰分析に近づく。

局所重回帰分析は要求点の近傍で成り立つ近似手法であるので、通常の RMSE や重相関係数の指標は使えず、その信頼性を求める指標は 1 個抜き交差検証法（HOOCV：Leave-One-Out Cross-Validation）を用いて与える。即ち、データ中の 1 点を抜き、その説明変数の座標 $x_{i\lambda}$ を要求点とし、残りの点で局所重回帰分析を行い、要求点の予測値 Y_λ を求める。元々この点には実測値 y_λ があるので予測の誤差が求められる。

局所重回帰分析の精度の指標はこの実測値と予測値を利用し、通常の重回帰分析の RMSE や重相関係数の定義を用いて以下のように与える。もちろんこの指標はバンド幅に影響される。

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{\lambda=1}^{N}\left(y_\lambda - Y_\lambda\right)^2}, \quad \text{重相関係数} = \frac{\sum_{\lambda=1}^{N}\left(y_\lambda - \bar{y}\right)\left(Y_\lambda - \bar{Y}\right)}{\sqrt{\sum_{\mu=1}^{N}\left(y_\mu - \bar{y}\right)^2 \sum_{\nu=1}^{N}\left(Y_\nu - \bar{Y}\right)^2}} \tag{7}$$

局所重回帰分析は、バンド幅や 1 個抜く点によって必ずしも予測値が求められるとは限らない。そのため、RMSE や重相関係数の値は求められた点だけを用いて計算することもある。

参考文献

[1] W.S.Cleveland and S.J.Delvin, Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, Vol.83, No.403, 596-610 (1988).

１７．数量化Ⅳ類

17.1　数量化Ⅳ類とは

林の数量化Ⅳ類はデータ間の親近性を仮定し、その中に内在するパターンをデータの空間配置として表現する手法であり、多次元尺度構成法の一種である。r 次元ユークリッド空間中に m 個のデータがあり、データ i とデータ j との親近性（類似度）を e_{ij} とする。親近性には正負の符号の制限はないが、親近性が高いほど大きな値を取るものとする。また一般に対称性 $e_{ij}=e_{ji}$ を仮定しない。r 次元の空間中のデータ i の位置座標を $x_{i\alpha}$ （$\alpha=1,2,\cdots,r$）とする（これは後から求まる）。

データ i とデータ j の距離 d_{ij} (≥ 0) を位置座標 $x_{i\alpha}$ と $x_{j\alpha}$ を使って、以下のように定義する。

$$d_{ij}^{(r)2}=\sum_{\alpha=1}^{r}(x_{i\alpha}-x_{j\alpha})^2$$

今、親近性の高いデータ同士は近い距離に位置するように配置したいが、これを実現するために、$\sum_{i=1}^{m}x_{i\alpha}^2=1$ の条件を付けて以下の量 Q を最大化することを考える。

$$Q=-\sum_{i=1}^{m}\sum_{j=1}^{m}e_{ij}d_{ij}^{(r)2}=-\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\alpha=1}^{r}e_{ij}(x_{i\alpha}-x_{j\alpha})^2$$

これを解くと $\lambda_\alpha\neq 0$ の場合、以下のような条件の付いた、$r-1$ 個の自明でない解が求まる。

$$\sum_{i=1}^{m}x_{i\alpha}=0$$

この解を 2 次元空間上に配置すると、それぞれのデータの関係が分かる。

17.2　プログラムの利用法

数量化Ⅳ類のデータは、数間の親近性（類似度）または距離（非類似度）を表すデータである。その例を図 1 に示す。

図 1　距離を表すデータ

メニュー［分析－多変量解析他－関係分析手法－数量化Ⅳ類］を選択すると図 2 のような数量化Ⅳ類分析メニューが表示される。

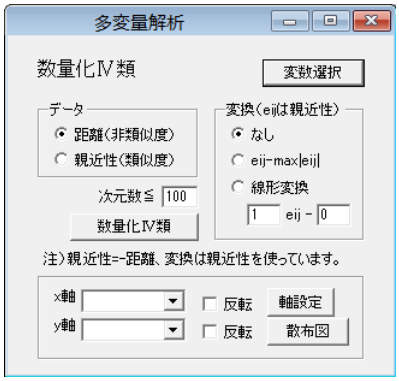


図２　数量化Ⅳ類分析メニュー

変数選択ですべての変数を選択し、データによって「距離」か「親近性」を選択する。距離の場合はデータの符号を変えて親近性にして分析を進める。変数の変換が必要な場合は変換ラジオボタンで指定する。特に「eij-max|eij|」は固有値をすべて正にするための設定であり、「線形変換」は他の多次元尺度構成法と合わせるための設定である。「次元数」大きな値を設定しておけば、変数数-1 の値になる。もちろん見やすくするため小さな値に設定することもできる。

「数量化Ⅳ類」ボタンをクリックすると、図３のような実行結果が示される。


座標

	1次元	2次元	3次元	4次元	5次元
固有値	39.788	36.280	33.808	33.202	32.522
みかん	0.111	-0.160	0.564	-0.070	-0.687
りんご	0.242	-0.275	0.130	0.734	0.378
いちご	0.218	0.539	-0.585	0.185	-0.346
ぶどう	0.158	0.495	0.372	-0.418	0.501
なし	0.179	-0.603	-0.426	-0.494	0.111
メロン	-0.908	0.004	-0.056	0.062	0.043
間隔比	0.000	0.483	0.823	0.907	1.000
ed順位相関	0.570	0.953	0.770	0.511	

図３　分析結果

固有値、固有ベクトルが表示され、その下に固有値の間隔比と親近性と予測距離との順位相関が表示される。

「軸設定」をして、「散布図」ボタンをクリックすると、パラメータ（固有ベクトル）の値が散布図として図 4 のように表示される。軸の向きは「反転」チェックボックスによって変更できる。

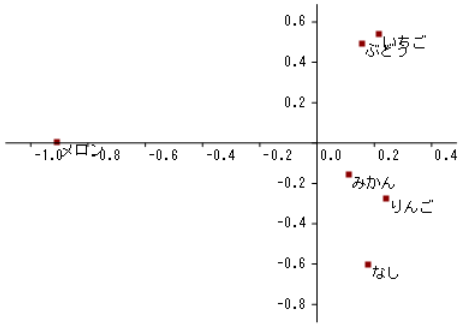


図４　パラメータ散布図

多次元尺度構成法の計量 MDS と数量化Ⅳ類の違いを見るために、図 5 の対称データを使って結果の散布図を比較する。

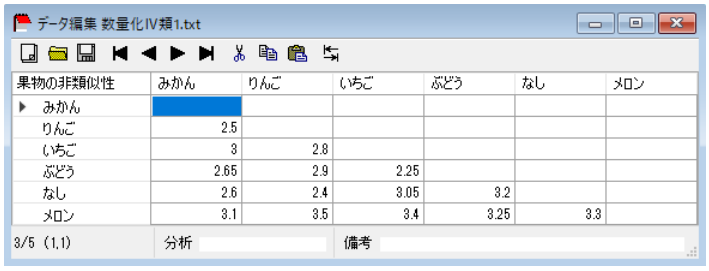


図 5　比較用データ

図 6a に計量 MDS の結果、図 6b に数量化Ⅳ類の結果を示す。

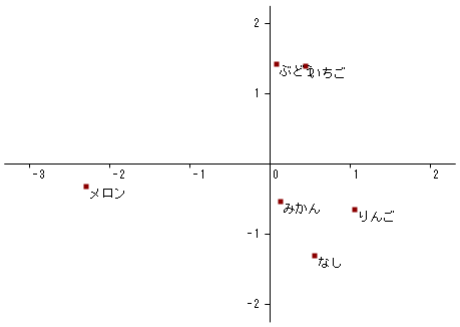


図 6a　計量 MDS の結果

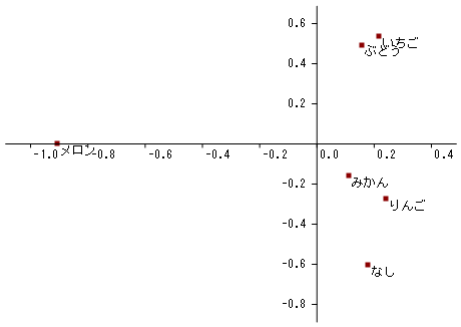


図 6b　数量化Ⅳ類の結果

但し、計量 MDS の軸の符号については、数量化Ⅳ類と合わせるようにしている。配置としてはよく似ている。

17.3　数量化Ⅳ類の理論

林の数量化Ⅳ類はデータ間の親近性を仮定し、その中に内在するパターンをデータの空間配置として表現する手法である。r 次元ユークリッド空間中に m 個のデータがあり、データ i とデータ j との親近性（類似度）を e_{ij} とする。親近性には正負の符号の制限はないが、親近性が高いほど大きな値を取るものとする。また一般に対称性 $e_{ij}=e_{ji}$ を仮定しない。同一のデータ同士の親近性 e_{ii} は、後の議論から定義する必要はないが、取り敢えず 0 としておく。r 次元の空間中のデータ i の位置座標を $x_{i\alpha}$ （$\alpha=1,2,\cdots,r$）とし、これをベクトルで表し $\mathbf{x}_{\alpha}=(x_{1\alpha},x_{2\alpha},\cdots,x_{m\alpha})'$ とする。

データ i とデータ j の距離 d_{ij} (≥ 0) を位置座標 $x_{i\alpha}$ と $x_{j\alpha}$ を使って、以下のように定義する。

$$d_{ij}^{(r)2}=\sum_{\alpha=1}^{r}(x_{i\alpha}-x_{j\alpha})^2 \qquad (1)$$

今、親近性の高いデータ同士は近い距離に位置するように配置したいが、これを実現するために、以下の量 Q を最大化することを考える。

$$Q=-\sum_{i=1}^{m}\sum_{j=1}^{m}e_{ij}d_{ij}^{(r)2}=-\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\alpha=1}^{r}e_{ij}(x_{i\alpha}-x_{j\alpha})^2 \tag{2}$$

ここで、

$$g_{ij}=h_{ij}-\delta_{ij}\sum_{k=1}^{m}h_{ik}\ ,\qquad h_{ij}=e_{ij}+e_{ji} \tag{3}$$

と定義とすると、Qは以下のように書ける。

$$Q=\sum_{\alpha=1}^{r}\sum_{i=1}^{m}\sum_{j=1}^{m}g_{ij}x_{i\alpha}x_{j\alpha}=\sum_{\alpha=1}^{r}\mathbf{x}'_{\alpha}\mathbf{G}\mathbf{x}_{\alpha} \tag{4}$$

ここで、$\mathbf{x}_{\alpha}$の値によってQの値はいくらでも大きくできるため、以下の条件を付けることにする。

$$\sum_{i=1}^{m}x_{i\alpha}^2=1 \tag{5}$$

制約条件を付けたラグランジュの未定定数法を用いて、Qの式を以下のように変更する。

$$L=\sum_{\alpha=1}^{r}\mathbf{x}'_{\alpha}\mathbf{G}\mathbf{x}_{\alpha}-\sum_{\alpha=1}^{r}\lambda_{\alpha}(\mathbf{x}'_{\alpha}\mathbf{x}_{\alpha}-1) \tag{6}$$

これを$\mathbf{x}_{\alpha}$で微分して以下の固有値方程式を得る。

$$\mathbf{G}\mathbf{x}_{\alpha}=\lambda_{\alpha}\mathbf{x}_{\alpha} \tag{7}$$

固有値方程式を成分で書き換えると以下のようになる。

$$\sum_{k=1}^{m}g_{ik}x_{k\alpha}=\lambda_{\alpha}x_{i\alpha} \tag{8}$$

これより以下となる。

$$\lambda_{\alpha}\sum_{i=1}^{m}x_{i\alpha}=\sum_{i=1}^{m}\sum_{k=1}^{m}g_{ik}x_{k\alpha}=0 \tag{9}$$

ここで定義式によって成り立つ以下の関係を使った。

$$\sum_{j=1}^{m}g_{ij}=0 \tag{10}$$

(9)式より$\lambda_{\alpha}\neq 0$の場合、以下となる

$$\sum_{i=1}^{m}x_{i\alpha}=0 \tag{11}$$

また、(10)式が成り立つことから方程式の１つの解として

$$\lambda_{\alpha}=0,\quad x_{i\alpha}=1/\sqrt{m} \tag{12}$$

を持つことも分かる。この場合(9)式の関係から、(11)式は成り立たなくてもよい。

最後に、方程式(8)を用いると、(4)の定義と(5)の制約より以下となる。

$$Q=\sum_{\alpha=1}^{r}\sum_{i=1}^{m}\lambda_{\alpha}x_{i\alpha}^2=\sum_{\alpha=1}^{r}\lambda_{\alpha} \tag{13}$$

親近性e_{ij}の線形変換に対する固有値と固有ベクトルの変化を調べてみる。

$$e'_{ij} = ae_{ij} + b \tag{14}$$

の変換に対して、

$$h'_{ij} = ah_{ij} + 2b$$

$$g'_{ij} = ag_{ij} - 2b(m\delta_{ij} - 1), \qquad \sum_{j=1}^{m} g'_{ij} = 0 \tag{15}$$

これにより、固有方程式は以下となる。

$$\sum_{k=1}^{m} (ag_{ik} + 2b) y_{k\alpha} = (\lambda'_\alpha + 2mb) y_{i\alpha} \tag{16}$$

これは$y_{k\alpha} = x_{k\alpha}$とすると以下の関係を得る。

$$\lambda'_\alpha = a\lambda_\alpha - 2mb \qquad \text{for } \lambda_\alpha \neq 0, \sum_{i=1}^{m} x_{i\alpha} = 0$$

$$\lambda'_\alpha = 0 \qquad \text{for } \lambda_\alpha = 0,\ x_{i\alpha} = const. \tag{17}$$

即ち、固有値も線形の変換を受ける。これより、0 でない固有値の分布の間隔比

$$\gamma(\alpha) = \frac{\lambda_{\max} - \lambda_\alpha}{\lambda_{\max} - \lambda_{\min}} \tag{18}$$

は変換(14)に対して不変である。これにより、データに固有の親近性の特徴を調べることができると考えられる。最後に、数量化の適合度の１つの指標として、距離$-e_{ij}$と(1)で与えられるr次元の距離$d_{ij}^{(r)}$との順位相関係数を考えることもある。しかし、これは次元数を増やせば必ず適合度が上がるとは限らず、注意が必要である。

参考文献

[1] 齋藤堯幸・宿久洋，関連性データの解析法，共立出版，2006

１８．パネル時系列分析

18.1　パネル時系列分析とは

パネルデータを使った時系列分析の中で、時刻tの時系列データy_t, x_tがあるとき、その中から時刻tを含めて２期分のそれ以前のデータを取り出す。それらのデータを説明変数とし、例えば時刻$t+1$のデータy_{t+1}を目的変数として予測する重回帰分析をパネル重回帰分析と呼ぶことにする。予測値をY_{t+1}とすると１期先の予測式は以下のように与えられる。

$$Y_{t+1} = b_{11}y_t + b_{12}y_{t-1} + b_{21}x_t + b_{22}x_{t-1} + b_0$$

これを表で描くと表１のようになる。

表１　時系列データ

時刻	変数１	変数２
$t+1$	y_{t+1}	
t	y_t	x_t
$t-1$	y_{t-1}	x_{t-1}
$t-2$	y_{t-2}	x_{t-2}
$t-3$	y_{t-3}	x_{t-3}
$t-4$	y_{t-4}	x_{t-4}
:	:	:

この予測を実現するためには表１を書き換えた表２のデータを使った重回帰分析を考える。

表２　パネル時系列分析データ

目的変数	説明変数１	説明変数２	説明変数３	説明変数４
y_t	y_{t-1}	y_{t-2}	x_{t-1}	x_{t-2}
y_{t-1}	y_{t-2}	y_{t-3}	x_{t-2}	x_{t-3}
y_{t-2}	y_{t-3}	y_{t-4}	x_{t-3}	x_{t-4}
:	:	:	:	:

パネル時系列分析には、これらの変数の他に他の分析で予測した結果を組み込むことができる。そこで我々は通常の時系列分析の結果をデータとして組み込むことを考えてみた。時系列分析は、傾向変動と周期変動を分解するモデルを考える。データの不規則な大きな変動も考える必要があるので、傾向変動には自然に傾向を求めることができる局所回帰分析を採用した。そのためバンド幅によって局所的な回帰式に影響を与える範囲を限定することができる。また周期変動については、分解する周期（周波数）を複数指定できるようにしている。この予測された変数を$\tilde{y}_t$とするとパネル時系列分析のデータは表３のようになる。

表　時系列分析を加えたパネル時系列分析データ

目的変数	説明変数１	説明変数２	説明変数３	説明変数４	説明変数５
y_t	y_{t-1}	y_{t-2}	x_{t-1}	x_{t-2}	$\tilde{y}_t$
y_{t-1}	y_{t-2}	y_{t-3}	x_{t-2}	x_{t-3}	$\tilde{y}_{t-1}$

y_{t-2}	y_{t-3}	y_{t-4}	x_{t-3}	x_{t-4}	$\tilde{y}_{t-2}$
:	:	:	:	:	:

時系列分析ではデータが時間の経過とともに明らかになっていくので、現在のすべてのデータから求めたパラメータを使って、過去の各時間の予測を行うことはその時点のデータの影響を強く受け過ぎるという難点がある。そこで、過去の予測を行う際には、その時点までのデータから計算されたパラメータを用いることとし、これによって実測値と予測値の相関を求めることにする。これは交差検証と呼ばれる手法の１つであるが、プログラムにはこの手法を付け加えている。

18.2　プログラムの利用法

パネル時系列分析のデータは複数変数の時系列データである。その例を図１に示す。

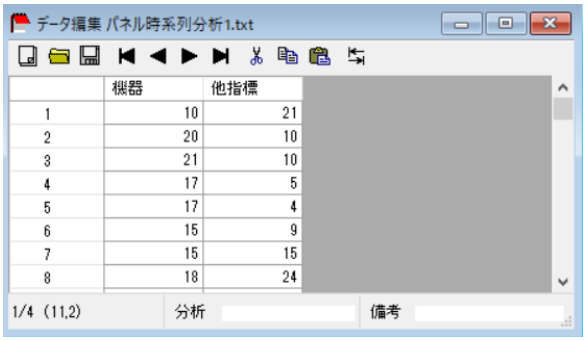


図１　パネル時系列分析のデータ

メニュー［分析－多変量解析他－予測手法－パネル時系列分析］を選択すると図 2 のようなパネル時系列分析実行画面が表示される。

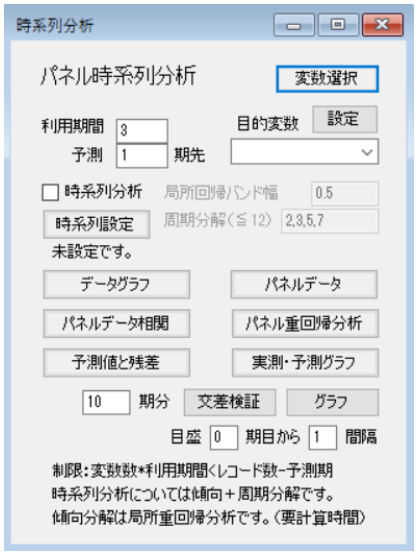
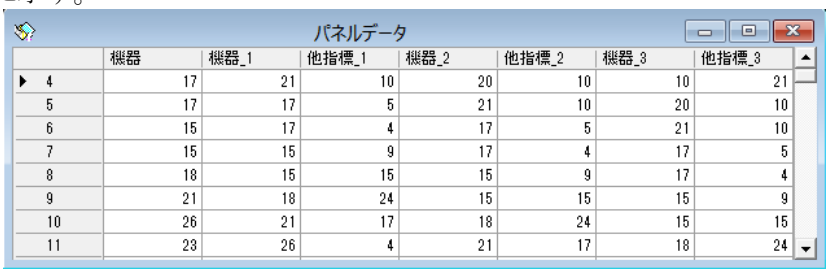


図２　分析実行画面

使用するデータをすべて「変数選択」ボタンで選ぶが、変数間の時間的な影響を調べるツールとして使うことも考えているため、通常の重回帰分析のように目的変数を最初に選択す

ることはしない。目的変数は、変数選択した候補をコンボボックスに読み込んだ後で、その中から「設定ボタン」で選択する。選択肢の中には単独の変数の他に「すべて」というものがあり、選択したすべての変数を目的変数にして、素早く結果を求めるときに利用する。ボタンによってはこれが使えないものもある。

この分析では、何期分のデータを利用するか、何期先の予測をするかを設定することができる。それに応じて、「パネルデータ」ボタンでは時系列データを通常の重回帰分析の形式に変形して出力する。出力結果をそのまま重回帰分析のデータとして利用することもできる。変数「機器」を目的変数とし、3 期分のデータを利用し、1 期先の予測をする場合の出力データを図 3 に示す。



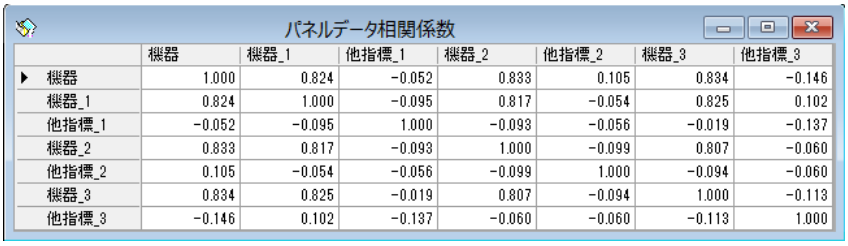
パネルデータ

	機器	機器_1	他指標_1	機器_2	他指標_2	機器_3	他指標_3
▶ 4	17	21	10	20	10	10	21
5	17	17	5	21	10	20	10
6	15	17	4	17	5	21	10
7	15	15	9	17	4	17	5
8	18	15	15	15	9	17	4
9	21	18	24	15	15	15	9
10	26	21	17	18	24	15	15
11	23	26	4	21	17	18	24

図 3　計算用データ

この中で「機器」は目的変数で、左に月単位で与えられているデータとする。また、例えば「機器_2」は変数「機器」の 2 期前のデータを表している。

図 3 の計算用データの各変数間の相関係数は、「パネルデータ相関」ボタンをクリックすることで図 4 のように与えられる。

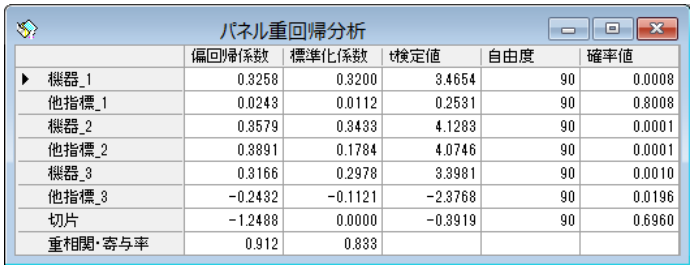


パネルデータ相関係数

	機器	機器_1	他指標_1	機器_2	他指標_2	機器_3	他指標_3
▶ 機器	1.000	0.824	−0.052	0.833	0.105	0.834	−0.146
機器_1	0.824	1.000	−0.095	0.817	−0.054	0.825	0.102
他指標_1	−0.052	−0.095	1.000	−0.093	−0.056	−0.019	−0.137
機器_2	0.833	0.817	−0.093	1.000	−0.099	0.807	−0.060
他指標_2	0.105	−0.054	−0.056	−0.099	1.000	−0.094	−0.060
機器_3	0.834	0.825	−0.019	0.807	−0.094	1.000	−0.113
他指標_3	−0.146	0.102	−0.137	−0.060	−0.060	−0.113	1.000

図 4　パネルデータ相関出力結果

このデータを使った重回帰分析の詳細は、「パネル時系列分析」ボタンで図 5 のように与えられる。

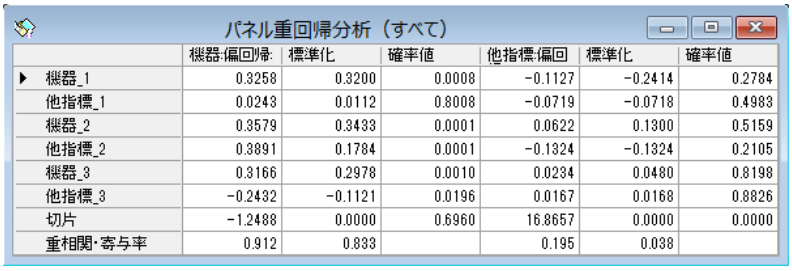


パネル重回帰分析

	偏回帰係数	標準化係数	t検定値	自由度	確率値
▶ 機器_1	0.3258	0.3200	3.4654	90	0.0008
他指標_1	0.0243	0.0112	0.2531	90	0.8008
機器_2	0.3579	0.3433	4.1283	90	0.0001
他指標_2	0.3891	0.1784	4.0746	90	0.0001
機器_3	0.3166	0.2978	3.3981	90	0.0010
他指標_3	−0.2432	−0.1121	−2.3768	90	0.0196
切片	−1.2488	0.0000	−0.3919	90	0.6960
重相関・寄与率	0.912	0.833			

図 5　目的変数を「機器」とした場合のパネル時系列分析結果

目的変数を「すべて」に設定すると、「パネル時系列分析」ボタンで図６のような結果になる。



パネル重回帰分析（すべて）

	機器:偏回帰	標準化	確率値	他指標:偏回	標準化	確率値
▶ 機器_1	0.3258	0.3200	0.0008	-0.1127	-0.2414	0.2784
他指標_1	0.0243	0.0112	0.8008	-0.0719	-0.0718	0.4983
機器_2	0.3579	0.3433	0.0001	0.0622	0.1300	0.5159
他指標_2	0.3891	0.1784	0.0001	-0.1324	-0.1324	0.2105
機器_3	0.3166	0.2978	0.0010	0.0234	0.0480	0.8198
他指標_3	-0.2432	-0.1121	0.0196	0.0167	0.0168	0.8826
切片	-1.2488	0.0000	0.6960	16.8657	0.0000	0.0000
重相関･寄与率	0.912	0.833		0.195	0.038	

図６　目的変数をすべてとした場合のパネル時系列分析結果

これは各変数を目的変数にして、偏回帰係数、標準化偏回帰係数、確率値、重相関係数、寄与率を出力している。どの変数の何期前のデータが重要であるか、標準化係数や確率値を見ることで知ることができる。

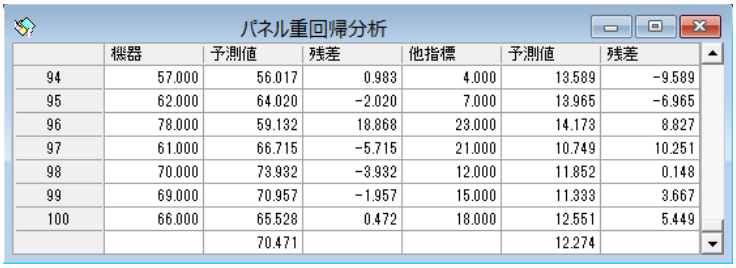
目的変数を「機器」とした場合の実測値、予測値、残差は、「予測値と残差」ボタンをクリックすることで図７のように求められる。ここで一番下の予測値は、１期先（設定で変更可能）の予測値で、実測値はまだない。

予測値と残差

	機器	予測値	残差
94	57.000	56.017	0.983
95	62.000	64.020	-2.020
96	78.000	59.132	18.868
97	61.000	66.715	-5.715
98	70.000	73.932	-3.932
99	69.000	70.957	-1.957
100	66.000	65.528	0.472
1期先		70.471	

図７　目的変数を「機器」とした場合の予測値と残差結果

また、目的変数を「すべて」とした場合の実測値、予測値、残差は、同様にして図８のように求められる。



パネル重回帰分析

	機器	予測値	残差	他指標	予測値	残差
94	57.000	56.017	0.983	4.000	13.589	-9.589
95	62.000	64.020	-2.020	7.000	13.965	-6.965
96	78.000	59.132	18.868	23.000	14.173	8.827
97	61.000	66.715	-5.715	21.000	10.749	10.251
98	70.000	73.932	-3.932	12.000	11.852	0.148
99	69.000	70.957	-1.957	15.000	11.333	3.667
100	66.000	65.528	0.472	18.000	12.551	5.449
		70.471			12.274	

図８　目的変数をすべてとした場合の予測値と残差結果

実測値と予測値について、結果をグラフで表示するためには、「実測・予測グラフ」ボタンをクリックする。実行結果は図９に示す。

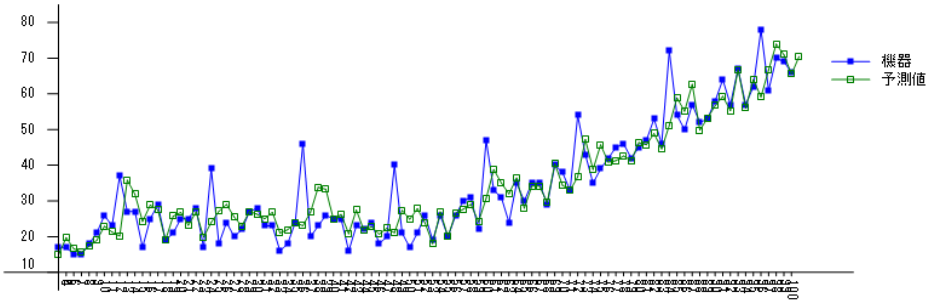


図 9　実測値と予測値グラフ

ここにデータの名前（年月）は縦表示にしてある。

我々がこれまで求めてきた各時点の予測値は、全体の結果を使って求めた係数から計算して得られた値である。それゆえ、この係数には各時点の実測値の結果が含まれている。そのためこれらのデータは厳密には予測値ではない。これを補正するためには、予測値は各時点のそれより過去のデータから求めるべきであろう。この考え方は交差検証の考え方に通じる。「期分」のテキストボックスに予測したい期間の数値を入れ、「交差検証」ボタンをクリックすると、過去のデータからだけで作られた予測値と残差が図 10 のように表示される。但し、表示期間を 50 期分にしている。


交差検証

	機器	予測値	残差
94	57.000	54.447	2.553
95	62.000	63.934	-1.934
96	78.000	57.643	20.357
97	61.000	68.532	-7.532
98	70.000	74.934	-4.934
99	69.000	71.341	-2.341
100	66.000	65.481	0.519
R･R^2	0.907	0.824	

図 10　目的変数を「機器」とした場合の 50 期分の交差検証結果

目的変数をすべてにして同様の結果を得ることもできる。「グラフ」ボタンをクリックすると、図 10 の結果をグラフ化することができる。結果を図 11 に示す。

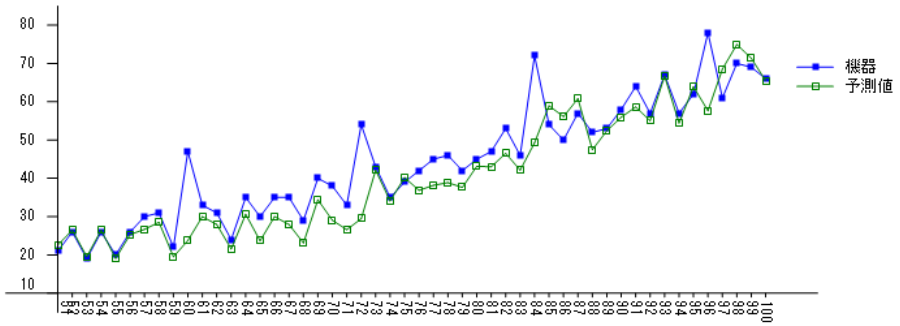


図 11　交差検証での実測値と予測値

純粋なパネル時系列分析の結果は以上であるが、我々はさらに予測精度を上げるために、

傾向変動や周期変動の分解を考える従来の時系列分析の予測値をパネルデータに加え、２つの分析の良い部分を組み合わせることにした。ここで、傾向変動には局所回帰分析を用いている。図２の分析実行画面の時系列分析チェックボックスにチェックを入れると、「局所回帰バンド幅」と「周期分解（≦12)」のテキストボックスが利用できるようになる。バンド幅の値はデフォルトでほぼ良い結果が得られるが、例えば 12 ヶ月周期が明らかな場合には、周期分解に 12 を含める。周期分解のためのデータ数は最低でも最大周期の２倍必要なので、周期は適当に小さくという意味で「(≦12)」の指摘を加えてある。しかし、この範囲に縛られる必要はない。ここでは 12 を加えている。

時系列分析を加えた場合、データの数によっては計算時間がかかる場合があるので、最初に「時系列設定」のボタンをクリックする。「計算が終わりました。」の表示が出たら、以後はすぐに表示される。「パネルデータ」ボタンをクリックすると、図 12 のように最後の列に時系列分析の予測値が追加される。但し、計算が可能な途中からの挿入となる。プログラムはこの部分を利用して計算をする。

パネルデータ

	機器	機器_1	他指標_1	機器_2	他指標_2	機器_3	他指標_3	機器_ts
22	28	25	1	25	22	21	17	0.000000
23	17	28	5	25	1	25	22	0.000000
24	39	17	4	28	5	25	1	0.000000
25	18	39	15	17	4	28	5	31.967064
26	24	18	9	39	15	17	4	24.869150
27	20	24	13	18	9	39	15	25.120598
28	22	20	23	24	13	18	9	18.966236
29	27	22	25	20	23	24	13	18.279529
30	28	27	16	22	25	20	23	21.529595
31	23	28	15	27	16	22	25	26.493811

図 12　時系列分析を加えた計算用データ

重回帰分析では、変数の数が増えると寄与率の値は増加するので、前以上の結果は期待できるが、増加の程度は、元のデータの性質による。例えば周期性が強いデータならば、時系列分析の変数の効果が強く効いてくる。

これ以降の分析は時系列分析を含めない場合と同様であるので、図 13 と図 14 に交差検証の結果のみを示しておく。データがそろってきた最後の方の数値はよく合っている。

交差検証

	機器	予測値	残差
94	57.000	58.364	-1.364
95	62.000	58.160	3.840
96	78.000	75.271	2.729
97	61.000	64.673	-3.673
98	70.000	71.280	-1.280
99	69.000	72.271	-3.271
100	66.000	67.021	-1.021
R･R^2	0.947	0.896	

図 13　時系列分析を加えた交差検証結果

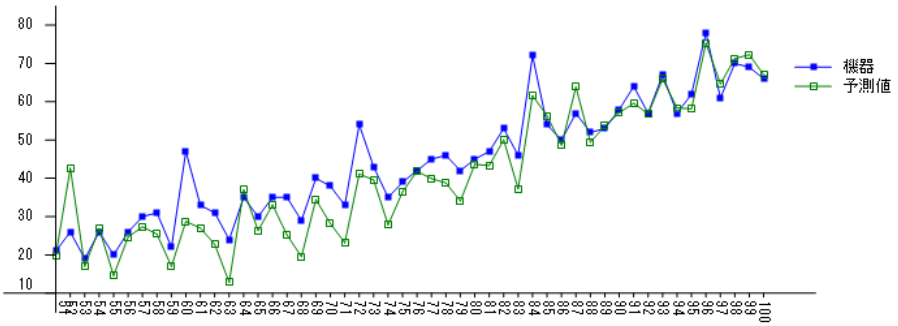
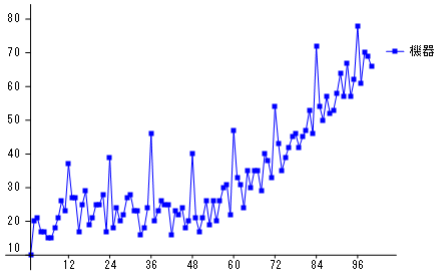


図 14　時系列分析を加えた交差検証での実測値と予測値

問題１

パネル時系列分析 1.txt、１頁目のデータを用いて以下の問いに答えよ。

１）機器のデータの変化の以下のグラフを描け（目盛はメニューの下部で設定)。



2 つの変数を選び、目的変数を機器にする。利用期間を 3 期、予測を 1 期先にする。

２）説明変数は何個になるか。

［　　］個

３）最初の変数の組は、何期目から始まり、そのときの目的変数の値はいくらか。

先頭の期［　　］期目、目的変数の値［　　］

４）目的変数と最も相関の高いのは、何期前のどちらの変数か。

［　　］期前の［機器・他指標］

５）時系列分析を加えないパネル重回帰分析の式はどうなるか。

機器＝　［　　］機器_1＋［　　］他指標_1

＋［　　］機器_2＋［　　］他指標_2

＋［　　］機器_3＋［　　］他指標_3＋［　　］

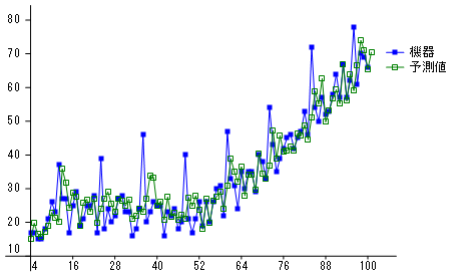
６）このパネル重回帰分析の寄与率はいくらか。

寄与率［　　］

７）先頭のデータの実測値、予測値、残差を求めよ。

実測値［　　］　予測値［　　］　残差［　　］

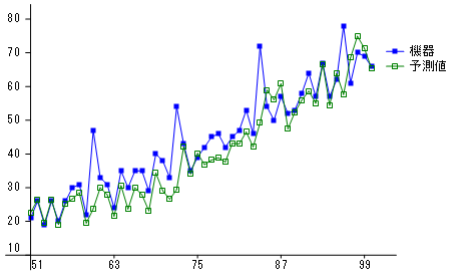
８）以下の実測予測グラフを描け。



９）50 期分の交差検証を実行し、寄与率 R^2 を求めよ。（実測・予測の最終行参照）

寄与率 R^2［　　］

10）同じ設定の交差検証で、以下のような実測・予測グラフを描け。これでは 12 ヶ月毎の変動がうまく予測されていない。



次に、周期 12 の時系列分析のデータを付け加える。

11）説明変数は何個になるか。

［　　］個

12）新しい設定で重回帰式はどうなるか。

機器＝　［　　］機器_1＋［　　］他指標_1
　　＋［　　］機器_2＋［　　］他指標_2
　　＋［　　］機器_3＋［　　］他指標_3
　　＋［　　］機器_ts＋［　　］

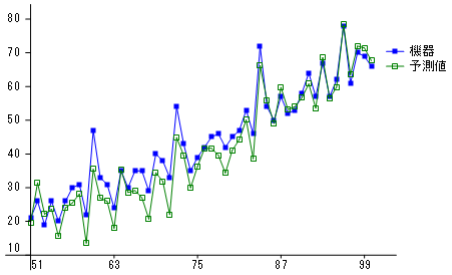
13）新しいパネル重回帰分析の寄与率はいくらか。

寄与率［　　］

14）50 期分の交差検証を実行し、寄与率 R^2 を求めよ。（実測・予測の最終行参照）

寄与率 R^2［　　］

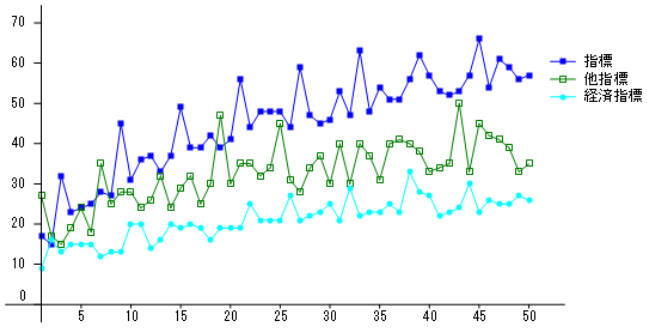
15）以下のような実測・予測グラフを描け。この場合 12 ヶ月毎の変動はうまく予測されている。



問題２

パネル時系列分析 3.txt はある会社の予測したい指標、その他の指標、経済指標の 3 種類のデータである。3 期分のデータを使った 1 期先を予測するパネル重回帰分析を用いて予測の可能性を以下に従って考えよ。

１）３つのデータの時系列グラフを示せ（目盛は５間隔）。



以後は３期分のデータを使って、「指標」についての１期先の予測を考える。

２）これらの３期分のデータの中で、予測値との相関が一番大きいものはどれか。

［指標・他指標・経済指標］の［　　］期前のデータで、相関係数［　　］

３）回帰式を求めよ。

予測値＝［　　］指標１期前＋［　　］他指標１期前

＋［　　］経済指標１期前＋…＋［　　］経済指標３期前＋［　　］

４）最も影響力のある変数は何か。

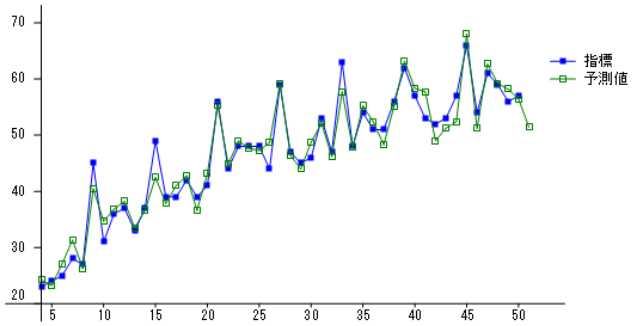
［指標・他指標・経済指標］の［　　］期前のデータで、

標準化偏回帰係数の値［　　］、偏回帰係数の検定確率値［　　］

５）予測の寄与率はいくらか。［　　］

６）１期先はいくらに予測されたか。［　　］

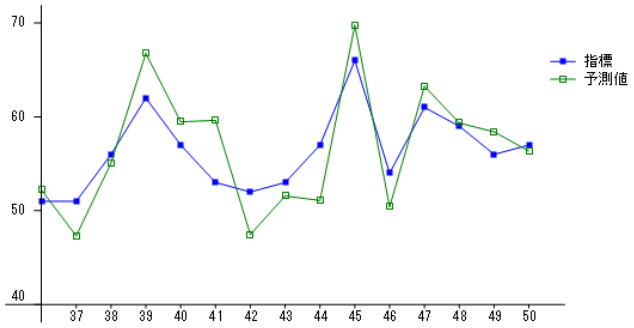
７）３）の回帰式を用いた実測値と予測値のグラフを描け。



８）これまでの予測は、予測式を求めるときに予測される数値が使われていた。これを防ぐために、予測値の値を求める際、それまでの値しか使わないようにした。そのようにした場合の予測の精度を見たい。50 期目の実測値と予測値を比較せよ。但し、交差検証には 15 期分のデータを使うことにする。

実測値	予測値	交差検証予測値

９）交差検証による実測、予測の 15 期分のグラフを描け。



10）１期先のデータ予測について、2 頁目のデータでの 50 期目の予測と 1 頁目のデータでの交差検証 50 期目の予測の値は同じか。

2 頁目 50 期目予測	1 頁目交差検証同予測

［同じ・同じでない］

データを 1 頁目に戻すこと。

11）1 期先と 5 期先との予測の寄与率（R^2）を交差検証の結果を元に比較せよ。

1 期先	5 期先

12）交差検証の 1 期先予測の寄与率（R^2）を、周期 12 期の時系列分析を加えない場合と加えた場合で比較せよ。

時系列分析を加えない	時系列分析を加える

問題１解答

2 つの変数を選び、目的変数を機器にする。利用期間を 3 期、予測を 1 期先にする。

２）説明変数は何個になるか。　　［ 6 ］個

３）最初の変数の組は、何期目から始まり、そのときの目的変数の値はいくらか。
先頭の期［ 4 ］期目、目的変数の値［ 17 ］

４）目的変数と最も相関の高いのは、何期前のどちらの変数か。
［ 3 ］期前の［機器・他指標］

５）時系列分析を加えないパネル重回帰分析の式はどうなるか。
機器＝　［ 0.3258 ］機器_1＋［0.0243］他指標_1
＋［ 0.3579 ］機器_2＋［ 0.3891 ］他指標_2
＋［ 0.3166 ］機器_3＋［ -0.2432 ］他指標_3＋［ -1.2488 ］

６）このパネル重回帰分析の寄与率はいくらか。　　寄与率［ 0.833 ］

７）先頭のデータの実測値、予測値、残差を求めよ。
実測値［ 17 ］　予測値［ 14.945 ］　残差［ 2.055 ］

９）50 期分の交差検証を実行し、寄与率 R^2 を求めよ。（実測・予測の最終行参照）
寄与率 R^2［ 0.824 ］

11）説明変数は何個になるか。　　［ 7 ］個

12）新しい設定で重回帰式はどうなるか。
機器＝　［ -0.0984 ］機器_1＋［ -0.0092 ］他指標_1
＋［ -0.0734 ］機器_2＋［ 0.5453 ］他指標_2
＋［ 0.0755 ］機器_3＋［ -0.1212 ］他指標_3
＋［ 1.0905 ］機器_ts＋［ -5.3908 ］
13）新しいパネル重回帰分析の寄与率はいくらか。　寄与率［ 0.950 ］
14）50 期分の交差検証を実行し、寄与率 R^2 を求めよ。（実測・予測の最終行参照）
寄与率 R^2［ 0.952 ］

問題２解答

以後は 3 期分のデータを使って、「指標」についての 1 期先の予測を考える。
２）これらの 3 期分のデータの中で、予測値との相関が一番大きいものはどれか。
［指標・他指標・経済指標の［ 2 ］期前のデータで、相関係数［　　］
３）回帰式を求めよ。
予測値＝［ 0.2379 ］指標 1 期前＋［ 0.0755 ］他指標 1 期前
＋［ 0.5685 ］経済指標 1 期前＋…＋［0.3131］経済指標 3 期前＋［ -4.0462 ］
４）最も影響力のある変数は何か。
［指標・他指標・経済指標］の［ 2 ］期前のデータで、
標準化偏回帰係数の値［ 0.6001 ］、偏回帰係数の検定確率値［ 0.0000 ］
５）予測の寄与率はいくらか。［ 0.952 ］
６）1 期先はいくらに予測されたか。［ 51.427 ］
８）50 期目の実測値と予測値を比較せよ。但し、交差検証には 15 期分のデータを使うことにする。

実測値	予測値	交差検証予測値
57	56.448	56.359

10）1 期先のデータ予測について、2 頁目のデータでの 50 期目の予測と 1 頁目のデータでの交差検証 50 期目の予測の値は同じか。

2 頁目 50 期目予測	1 頁目交差検証同予測
56.359	56.359

［同じ・同じでない］
11）1 期先と 5 期先との予測の寄与率（R^2）を交差検証の結果を元に比較せよ。

1 期先	5 期先
0.765	0.281

12）交差検証の 1 期先予測の寄与率（R^2）を、周期 12 期の時系列分析を加えない場合と加えた場合で比較せよ。

時系列分析を加えない	時系列分析を加える
0.765	0.583

18.3　パネル時系列分析の理論

変数i（$i=1,\cdots,p$）、時刻t（$t=T,\cdots,0$）の時系列データ$x_{i,t}$があるとき、その中から時刻tを含めてr期分のそれ以前のデータを取り出す。それらのデータを説明変数とし、時刻$t+a$（$a\geq 1$）のある変数dのデータ$x_{d,t+a}$を目的変数として予測する重回帰分析をパネル重回帰分析と呼ぶことにする。これはa期先の予測である。

予測値を$X_{d,t+a}$とすると予測式は以下のように与えられる。

$$X_{d,t+a} = \sum_{i=1}^{p}\sum_{j=0}^{r-1} b_{i,j} x_{i,t-j} + b_0 \tag{1}$$

係数$b_{i,j}, b_0$は以下の量Lを最小化することによって求める。

$$L = \sum_{t=a+r-1}^{T} \left(x_{d,t} - \sum_{i=1}^{p}\sum_{j=0}^{r-1} b_{i,j+1} x_{i,t-a-j} - b_0 \right)^2 \tag{2}$$

今、目的変数と説明変数をそれぞれ以下のように定義し、

$y_\lambda = x_{d,\lambda+a+r-2}$ （$\lambda = 1, \cdots, T-a-r+2$）

$z_{\alpha,\lambda} = z_{i+pj,\lambda} = x_{i+pj,\lambda+r-2-j}$ （$i = 1, \cdots, p,\ j = 0, \cdots, r-1, \alpha = 1, \cdots, pr$）

係数をb_αにして(2)式を書き変えると、以下のような式になる。

$$L = \sum_{\lambda=1}^{T-a-r+2} \left(y_\lambda - \sum_{\alpha=1}^{pr} b_\alpha z_{\alpha,\lambda} - b_0 \right)^2 \tag{3}$$

これから、偏回帰係数 $\mathbf{b} = {}^t\begin{pmatrix} b_0 & b_1 & b_2 & \cdots & b_s \end{pmatrix}$，$s = pr$ は以下のように求めることができる。

$$\mathbf{b} = \left({}^t\mathbf{\Omega}\mathbf{\Omega} \right)^{-1} {}^t\mathbf{\Omega}\mathbf{y} \tag{4}$$

ここに、

$\mathbf{y} = {}^t\begin{pmatrix} y_1 & y_2 & \cdots & y_N \end{pmatrix}$，$N = T-a-r+2$

$$\mathbf{\Omega} = \begin{pmatrix} 1 & z_{11} & z_{21} & \cdots & z_{s1} \\ 1 & z_{12} & z_{22} & \cdots & z_{s2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{1N} & z_{2N} & \cdots & z_{sN} \end{pmatrix}$$

時系列分析ではデータが時間の経過とともに明らかになっていくので、現在のすべてのデータから求めたパラメータを使って、過去の各時間の予測を行うことはその時点のデータの影響を強く受け過ぎるという難点がある。そこで、過去の予測を行う際には、その時点までのデータから計算されたパラメータを用いることとし、これによって実測値と予測値の相関を求めることにする。これは一種の交差検証になっている。プログラムにはこの交差検証を付け加えている。

パネル重回帰分析には、他の分析で予測した結果を組み込むことができる。そこで時系列分析の結果をデータとして組み込むことを考えてみた。時系列分析は、傾向変動と周期変動を分解するモデルを考える。データの不規則な大きな変動も考える必要があるので、傾向変動には自然に傾向を求めることができる局所回帰分析を採用した。そのためバンド幅によって局所的な回帰式に影響を与える範囲を限定することができる。また周期変動については、分解する周期（周波数）を複数指定できるようにしている。

１９．メタ分析

19.1　メタ分析とは

メタ分析は、多くの研究資料から同一の調査内容を選び出し、それらを再度集計して結果をより強固なものにしようとする分析手法である。１つの研究資料からは、効果量と呼ばれる統計量とその分散及び､データ数を取り出す｡代表的な効果量には標準化された平均値差、オッズ比、相関係数などがある。しかし、研究資料ごとにこれらが同じである保証はないので、必要があれば、これらを統一的な効果量に変換する。その後、各研究資料にデータ数でウェイトをかけて、研究で与えられた結果が保証されるかどうか検討する。この一連の手法をメタ分析という。

我々がプログラムの中で扱う効果量は以下で述べる通りである。種々の資料には効果量（または検定確率）とデータ数は記載されているが、効果量の分散が記載されていないことが多い。また、参考文献[1] では、後の統計的分析のために分散は記載されているが、データ数が記載されていない。これらの状況に対処するために、結果を求めるために必要なデータは何か、またそれを得るためにはどのようなデータが必要かを検討した。結論は、比較的良い近似として、結果表示に必要なデータは、効果量と全データ数または、効果量と分散であった。プログラムにはこれらの量を相互に変換する機能が必要である。

変換の対象となる効果量には以下のものがある。ここでは効果量の表式のみ示す。

１）標準化平均値差 D

効果量　$d=\dfrac{\bar{x}_1-\bar{x}_2}{u_{pooled}}$,　$u_{pooled}=\sqrt{\dfrac{(n_1-1)u_1^2+(n_2-1)u_2^2}{n_1+n_2-2}}$

２）バイアス修正平準化平均値差 G

効果量　$g=J\times d$,　$J=1-\dfrac{3}{4(n_1+n_2-2)-1}$

３）対数オッズ比 LOR

効果量　$LOR=\ln\left(\dfrac{a/b}{c/d}\right)=\ln\left(\dfrac{ad}{bc}\right)$

	効果あり	効果なし	合計
介入群	a	b	$a+b$
統制群	c	d	$c+d$

４）相関係数

効果量　$r=s_{xy}/(s_x s_y)$

５）ｔ統計量

効果量　$t=\sqrt{\dfrac{n_1n_2}{n_1+n_2}}\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\dfrac{(n_1-1)u_1^2+(n_2-1)u_2^2}{n_1+n_2-2}}}=\sqrt{\dfrac{n_1n_2}{n_1+n_2}}d\simeq\sqrt{\dfrac{N}{4\alpha}}d$

但し、プログラムではｔ統計量は変換後の対象とはしていない。

様々な効果量、分散、データ数から、特定の効果量、分散、データ数に変換したとする。次に、これらの統一されたデータを用いて研究をまとめて有意差を見る分析を行う。これには研究間に差があるか否かによって２つの方法がある。「研究間比較」の検定を行ったうえで、差がない場合は「固定効果モデル」、差がある場合は「変量効果モデル」という手法を利用する。この一連の手法が「メタ分析」である。メタ分析を用いて求めた結果を表した図がフォレストプロットである。

何らかの指標の違いにより、研究がいくつかのグループに分けられるとする。そのグループ間の効果量の差を検定するには、「研究群間比較」検定を用いる。これはメタ分析を用いた１元配置分散分析に相当する検定である。

19.2　プログラムの利用法

メニュー［分析－多変量解析等－メタ分析］を選択すると、メタ分析の分析実行画面が図２のように表示される。

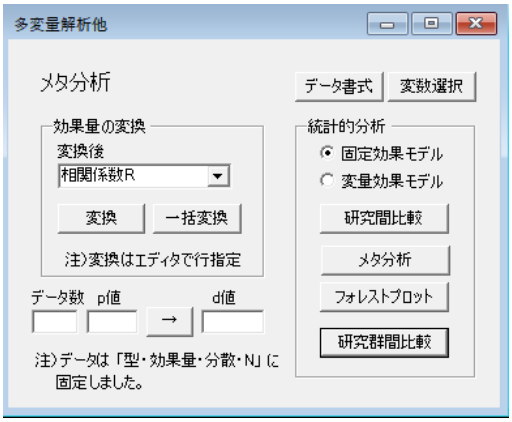


図２　分析実行画面

ここでは、様々な形式の混じった図３のデータを元にプログラムの利用法を説明する。

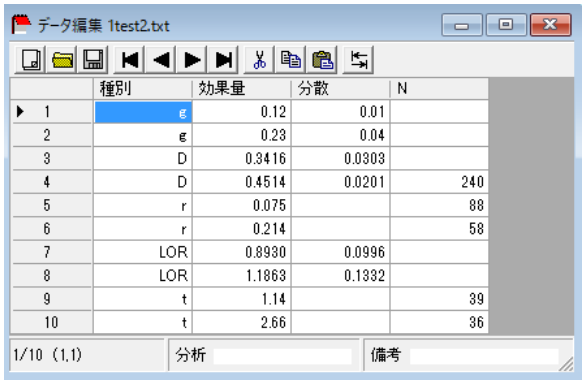


図３　メタ分析データ

このデータではデータ型、有効量は必須で、分散またはデータ数はどちらかが必要である。どちらも求められている場合は両方に記入する。一般には分散が与えられる場合は少なく、むしろ、データ型、有効量、データ数が与えられることが多いと思われる。データ型には、G：バイアス修正平準化平均値差、D：標準化平均値差、LOR：対数オッズ比、R：相関係

数、T：ｔ検定値、が指定できる。なお、指定する文字は大文字でも小文字でも同じである。

また、2 群の差の検定などでは、検定統計量を省略し、検定確率だけを表示している場合もあるので、その際には、標準化平均値差 D の値を簡易的に計算できる機能をメニューの下に設けている。その他の対数オッズ比や相関係数では、殆どの場合、値を記述するので、ここでは標準化平均値差 D に限定している。また、ノンパラメトリック検定の確率から近似的に D を求めても、少し乱暴ではあるが、経験上特に大きな差は出ないように思う。

一般に各研究では効果量が同一とは限らない。異なる効果量の場合は、効果量の変換を行い、同じ効果量に合わせて分析する。そのためにプログラムには効果量の変換機能を付けている。変数選択で 3 つの変数を選択し、「変換後」コンボボックスで変換先の型を選び、「一括返還」ボタンをクリックすると、図 4 のような結果が得られる。ここでは、標準化平均値差 D として出力している。元のデータで空欄であった部分は補完されている。

効果量変換

	種別	効果量	分散	N
1	D	0.1202	0.0100	476
2	D	0.2315	0.0405	119
3	D	0.3416	0.0303	159
4	D	0.4514	0.0201	240
5	D	0.1504	0.0457	88
6	D	0.4382	0.0723	58
7	D	0.4923	0.0303	226
8	D	0.6540	0.0405	169
9	D	0.3978	0.1238	39
10	D	0.9660	0.1448	36

図 4　効果量の標準化平均値差 D への変換

変換後の型は、入力の型の中からｔ検定値を除いた型を選ぶことができる。

グリッドの一部分のデータについて変換をしたい場合は、種別・効果量・分散・N の必要な行を連続的に選択して「変換」ボタンをクリックする。出力結果は省略する。

すべての研究結果を統合して検定を行いたい場合、研究間の効果量の値にばらつきがあるかどうか知らなければならない。それを調べる場合は、入力を型の揃ったデータにして（例えば図 4 で与えたデータ）、「研究間比較」ボタンをクリックする。結果を図 5 に示す。

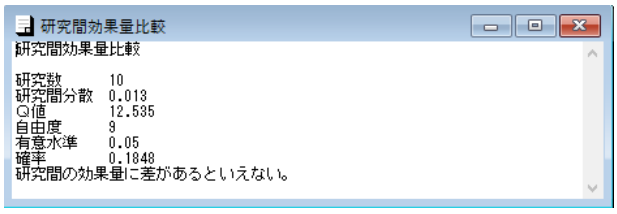


図 5　研究間効果量の差の比較検定

この結果から、研究間の効果量に差が見られないので、分析には「固定効果モデル」を用いる。これは「変量効果モデル」に比べて差が検出されやすい検定である。

固定効果モデルを用いた最終的な分析結果を得るには、「固定効果モデル」ラジオボタンを選択し、「メタ分析」ボタンをクリックする。結果を図 6 に示す。

フォレストデータ

	種別	効果量	分散	N	標準誤差	p値	2.5%下限	2.5%上限
1	D	0.120	0.010	476	0.100	0.229	−0.076	0.316
2	D	0.232	0.041	119	0.201	0.250	−0.163	0.626
3	D	0.342	0.030	159	0.174	0.050	0.000	0.683
4	D	0.451	0.020	240	0.142	0.001	0.174	0.729
5	D	0.150	0.046	88	0.214	0.482	−0.269	0.569
6	D	0.438	0.072	58	0.269	0.103	−0.089	0.965
7	D	0.492	0.030	226	0.174	0.005	0.151	0.833
8	D	0.654	0.041	169	0.201	0.001	0.260	1.048
9	D	0.398	0.124	39	0.352	0.258	−0.292	1.087
10	D	0.966	0.145	36	0.381	0.011	0.220	1.712
結合		0.326	0.003	1610	0.056	0.000	0.216	0.437

図６　変量効果モデルを用いた分析結果

各研究の結果がまとめて表示され、一番下の行に結合された結果が表示されている。

さらに、この結果を分かり易く表す図がフォレストプロットである。「フォレストプロット」ボタンをクリックすると、図７のような結果が表示される。

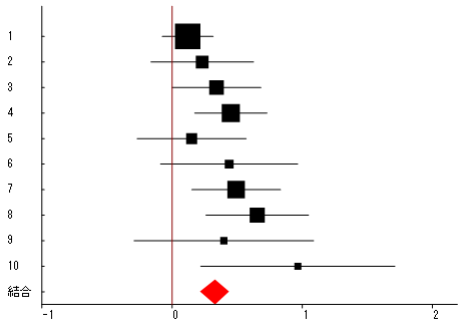


図７　フォレストプロット

一番下のひし形が、0 をまたいでいないことから、この結果では、有意に差があるといえるということになる。

次に、研究がいくつかの特徴に分かれ、その研究群間に差があるかどうか調べてみたいと考えたとする。その際には、先頭列に分類変数を加えた図８のようなデータを用いる。

データ編集 1test2.txt

	研究群	種別	効果量	分散	N
1	1	g	0.12	0.01	
2	1	g	0.23	0.04	
3	1	g	0.34	0.03	
4	1	g	0.45	0.02	
5	1	g	0.42	0.01	
6	2	g	0.39	0.02	
7	2	g	0.49	0.03	
8	2	g	0.65	0.04	
9	2	g	0.76	0.02	
10	2	g	0.87	0.01	

10/11 (1,1)　分析　備考

図８　２つの研究群による比較データ

すべてのデータを並んだ順に選択し、分析実行画面の「研究群間比較」ボタンをクリックすると、図９のような結果が得られる。

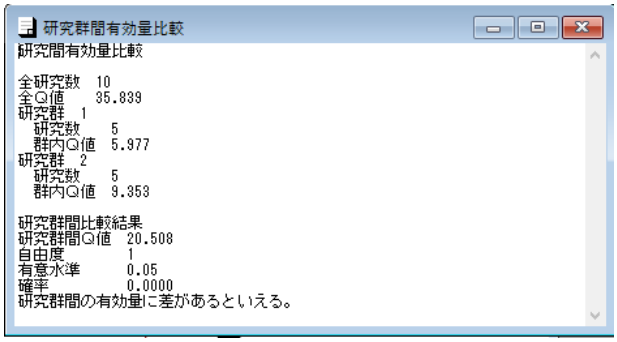


図 9　研究群間の比較結果

19.3　メタ分析の理論

メタ分析は、多くの研究資料から同一の調査内容を選び出し、それらを再度集計して結果をより強固なものにしようとする分析手法である。１つの研究資料からは、効果量と呼ばれる統計量とその分散及び､データ数を取り出す｡代表的な効果量には標準化された平均値差、オッズ比、相関係数などがある。しかし、研究資料ごとにこれらが同じである保証はないので、必要があれば、これらを統一的な効果量に変換する。その後、各研究資料にデータ数でウェイトをかけて、研究で与えられた結果が保証されるかどうか検討する。この一連の手法をメタ分析という。

我々はこの一連の過程を計算するプログラムの開発を考えた。ここでは、参考文献[1] に従い、効果量の入力、効果量の変換、統計的分析に分けて、理論的にどのような式が使われているのかをまとめて紹介する。

1）効果量とその入力

我々がプログラムの中で扱う効果量は以下で述べる通りである。種々の資料には効果量（または検定確率）とデータ数は記載されているが、効果量の分散が記載されていないことが多い。また、参考文献[1] では、後の統計的分析のために分散は記載されているが、データ数が記載されていない。これらの状況に対処するために、我々は結果表示に必要なデータは何か、またそれを得るためにはどのようなデータが必要かを検討した。結論は、比較的良い近似として、結果表示に必要なデータは、効果量と全データ数または、効果量と分散であった。ここでは、効果量と、全データ数または分散のどちらかが分かっているものとして、他方を求める近似式を与えておく。但しこの結果には $y=1/x(1-x)$ のグラフの性質を利用している。

a）標準化平均値差 d （ヘッジスの g とも呼ばれる）

対応のない場合

効果量　　$d=\dfrac{\bar{x}_1-\bar{x}_2}{u_{pooled}}$,　　$u_{pooled}=\sqrt{\dfrac{(n_1-1)u_1^2+(n_2-1)u_2^2}{n_1+n_2-2}}$　　（pooled 標準偏差）

分散　　$V_d = \dfrac{n_1+n_2}{n_1 n_2} + \dfrac{d^2}{2(n_1+n_2)}$

全データ数　$N = n_1 + n_2$

$d, N \to V_d$ のとき、$V_d \simeq \dfrac{4\alpha + d^2/2}{N}$

$d, V_d \to N$ のとき、$N \simeq \dfrac{4\alpha + d^2/2}{V_d}$

ここで、分散の $(n_1+n_2)/n_1 n_2 = N/n_1(N-n_1)$ の項については、例えば、$N=100$ とすると、図１のようなグラフとなる。

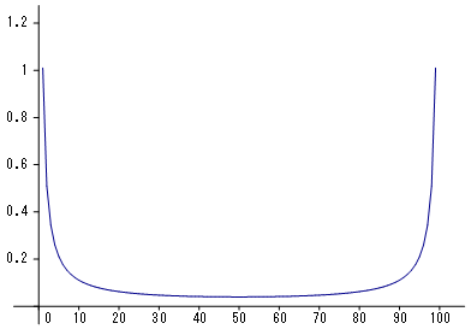


図１　$y = 100/x(100-x)$

このグラフは、中央部で $4/N$ に近いほぼ安定な値を取っており、この項による変動は少ないと考えられる。そこで我々は、この関数の $x = N/2$ の値を中心とした正規分布による加重平均を考え、その結果を $(n_1+n_2)/n_1 n_2 \simeq 4\alpha/N$ とした。

α の値については、以下のように計算した。

$$\alpha = \frac{1}{4A}\int_{0.1}^{0.9} \frac{1}{x(1-x)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-0.5)^2}{2\sigma^2}\right] dx$$

$$A = \frac{1}{\sqrt{2\pi}\sigma}\int_{0.1}^{0.9} \exp\left[-\frac{(x-0.5)^2}{2\sigma^2}\right] dx$$

この場合、例えば、$\sigma = 0.2$ とすると、$\alpha = 1.187$ となる。我々はこの値を利用する。

標準化平均値差の代わりに、資料で t 統計量が使われている場合は、簡単に標準化平均値差に変換することができる。

$$t = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{(n_1-1)u_1^2 + (n_2-1)u_2^2}{n_1+n_2-2}}} = \sqrt{\frac{n_1 n_2}{n_1+n_2}} d \simeq \sqrt{\frac{N}{4\alpha}} d$$

$$V_t \simeq \frac{N V_d}{4\alpha} \simeq \frac{4\alpha + d^2/2}{4\alpha} = 1 + \frac{d^2}{8\alpha}$$

これを利用すると、以下の変換も可能になる。

$t, N \to V_t$ のとき、$V_t \simeq 1+\dfrac{t^2}{2N}$

$t, V_t \to N$ のとき、$N \simeq \dfrac{t^2}{2(V_t-1)}$

b）**バイアス修正平準化平均値差 g**

効果量　$g = J \times d$,　$J = 1-\dfrac{3}{4(n_1+n_2-2)-1}$

分散　$V_g = J^2 \times V_d$

全データ数　$N = n_1+n_2$

$N \to V_g$ のとき、$J = 1-\dfrac{3}{4(N-2)-1}$ として、$V_g = J^2 V_d \simeq \dfrac{4\alpha J^2 + g^2/2}{N}$

$V_g \to N$ のとき、$J \simeq 1$ であると考え、$N \simeq \dfrac{4\alpha + g^2/2}{V_g}$

c）**対数オッズ比**

以下の 2 次元分割表を考える。

	効果あり	効果なし	合計
介入群	a	b	$a+b$
統制群	c	d	$c+d$

効果量　$LOR = \ln\left(\dfrac{a/b}{c/d}\right) = \ln\left(\dfrac{ad}{bc}\right)$

分散　$V_{LOR} = \dfrac{1}{a}+\dfrac{1}{b}+\dfrac{1}{c}+\dfrac{1}{d}$

全データ数　$N = a+b+c+d = n_1+n_2$

$N \to V_{LOR}$ のとき、$V_{LOR} \simeq \dfrac{16\alpha^2}{N}$　$V_{LOR} = \dfrac{1}{a}+\dfrac{1}{b}+\dfrac{1}{c}+\dfrac{1}{d} \simeq \dfrac{4\alpha}{n_1}+\dfrac{4\alpha}{n_2} \simeq \dfrac{16\alpha^2}{N}$

$V_{LOR} \to N$ のとき、$N \simeq \dfrac{16\alpha^2}{V_{LOR}}$

効果量の代わりに、資料で χ^2 統計量が使われていた場合は、簡単に効果量に変換することができない。

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

分割表の度数から効果量を計算する必要がある。

d）**相関係数**

効果量　$r = \dfrac{s_{xy}}{s_x s_y}$　　分散　$V_r = \dfrac{(1-r^2)^2}{n-1}$

データ数　$N = n$

$N \to V_r$ のとき、$V_r = \dfrac{(1-r^2)^2}{N-1}$

$V_g \to N$ のとき、$N = \dfrac{(1-r^2)^2}{V_r} + 1$

2）効果量の変換

効果量は相互に変換可能である。ここではプログラムで用いられる変換について式を与える。

$d \leftrightarrow g$

効果量：$g = J \times d$　　分散：$V_g = J^2 \times V_d$

ここに、$J = 1 - \dfrac{3}{4(N-2)-1}$　（入力の際に N は設定済みとする）

$LOR \leftrightarrow d$

効果量：$d = LOR \times \dfrac{\sqrt{3}}{\pi}$　　分散：$V_d = V_{LOR} \times \dfrac{3}{\pi^2}$

$r \to d$

効果量：$d = \dfrac{2r}{\sqrt{1-r^2}}$　　分散：$V_d = \dfrac{4V_r}{(1-r^2)^3}$

$d \to r$

効果量：$r = \dfrac{d}{\sqrt{d^2 + a}}$　　分散：$V_r = \dfrac{a^2 V_d}{(d^2 + a)^3}$

ここに、$a = \dfrac{(n_1 + n_2)^2}{n_1 n_2} \simeq 4\alpha$

3）統計的分析

a）固定効果モデル

固定効果モデルでは、研究間の差はなく、研究 i の効果量 d_i は独立に $d_i \sim N(0, V_i)$ に従うと仮定し、以下の集計を考える。

$$d = \sum_{i=1}^{n} d_i / V_i \Big/ \sum_{i=1}^{n} 1/V_i \sim N\left(0, \sum_{i=1}^{n} V_i / V_i^2 \Big/ \left(\sum_{i=1}^{n} 1/V_i\right)^2\right) = N\left(0, 1 \Big/ \sum_{i=1}^{n} 1/V_i\right)$$

ここで、$w_i = 1/V_i$ として、これをウェイトと考え、$w = \sum_{i=1}^{n} w_i$ とすると、以下となる。

$$d_i \sim N(0, 1/w_i), \qquad d = \sum_{i=1}^{n} w_i d_i \Big/ w \sim N\left(0, 1/w\right)$$

この性質より、研究を結合した検定は、検定統計量 $z = d \Big/ \sqrt{1/w} \sim N(0,1)$ を使って行う。

b）変量効果モデル

変量効果モデルでは、研究間に差があり、研究 i の効果量 d_i は広く拡がり、$d_i \sim N(0, V_i + \sigma^2)$ に従うと考える。$w_i' = 1/(V_i + \sigma^2)$ とおくと、$d_i \sim N(0, 1/w_i')$ より、$\sqrt{w_i'} d_i \sim N(0,1)$ となり、以下を得る。

$$Q' = \sum_{i=1}^{n} w_i'(d_i - d)^2 = \sum_{i=1}^{n} \frac{(d_i - d)^2}{V_i + \sigma^2} \sim \chi^2_{n-1}$$

一方、

$$Q = \sum_{i=1}^{n} w_i(d_i - d)^2 = \sum_{i=1}^{n} \frac{(d_i - d)^2}{V_i}$$

は元の分散で測った量である。その差は、以下で与えられる。

$$Q - Q' = \sum_{i=1}^{n} \frac{(d_i - d)^2 \sigma^2}{V_i(V_i + \sigma^2)} = \sum_{i=1}^{n} (d_i - d)^2 w_i' w_i \sigma^2 = C\sigma^2$$

ここに、Q' と C には、期待値を使って、

$$E(Q') = n - 1$$

また、$C = E[\sum_{i=1}^{n} (d_i - d)^2 w_i' w_i]$ は、

$$E[(d_i - d)^2] = E[d_i^2] - 2E[d_i d] + E[d^2]$$

$$E[d_i^2] = V_i' = 1/w_i'$$

$$E[d_i d] = E[d_i \sum_{j=1}^{n} d_j w_j' / w'] = V_i' w_i' / w' = 1/w'$$

$$E[d^2] = E[\sum_{i=1}^{n} d_i w_i' / w' \sum_{j=1}^{n} d_j w_j' / w'] = \sum_{i=1}^{n} V_i w_i'^2 / w'^2 = \sum_{i=1}^{n} w_i' / w'^2 = 1/w'$$

より、

$$C = \sum_{i=1}^{n} (1/w_i' - 1/w') w_i' w_i = \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i' w_i / w' \simeq \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i^2 / w$$

これらより、σ^2 が以下のように求められる。

$$\sigma^2 = \frac{Q - (n-1)}{C}$$

以後、ウェイトとして、$w_i' = \dfrac{1}{V_i + \sigma^2}$，$w' = \sum_{i=1}^{n} w_i'$ を用いて、計算を行えばよい。即ち、研究を結合した検定は、検定統計量 $z' = d/\sqrt{1/w'} \sim N(0,1)$ を使って行う。

4）研究群間の比較

何らかの指標の違いにより、研究が k 個のグループに分けられるとする。各グループの研究の数を n_i，全体の研究の数を n とするとき、そのグループ間の効果量の差を検定するには、以下の性質を用いる。

$Q_{Total} \sim \chi^2_{n-1}$ ， $Q_i \sim \chi^2_{n_i-1}$， $n = \sum_{i=1}^{k} n_i$ より、

$$Q_{Total} - \sum_{i=1}^{k} Q_i \sim \chi^2_{df} \quad ， \quad df = (n-1) - \sum_{i=1}^{k} (n_i - 1) = k - 1$$

この計算には、固定効果モデルではウェイト $w_i = 1/V_i$ を用い、変量効果モデルではウェイト $w_i' = 1/V_i'$ を用いる。

参考文献

[1] 山田剛史, 井上俊哉編, メタ分析入門 心理・教育研究の系統的レビューのために, 東京大学出版会, 2012.

２０．２値ロジスティック回帰

ここでは事象の出現する確率や出現しない確率を 2 項分布に基づくモデルで解析する 2 項分布モデルと呼ばれる手法を解説する。その中で、最もよく利用されるのがロジスティックモデルという方法であることから、章のタイトルを 2 値ロジスティック回帰という名称にした。ロジスティック回帰分析というのは一般によく利用される用語である。

20.1　2 項分布モデルとは

2 項分布モデルとは、2 項分布の出現確率を説明変数の線形結合で推測するモデルである。ケース i（$i=1,\cdots,n$）の試行回数を n_i、事象の出現回数を y_i、出現確率を p_i とすると、出現回数の密度関数、平均、分散はそれぞれ以下のようになる。但し、事象が起きたかどうかという、2 群の判別分析に相当する場合は、$n_i=1,\ y_i=\{0,1\}$ である。

$$f(y_i;p_i)={}_{n_i}C_{y_i}\,p_i^{y_i}(1-p_i)^{n_i-y_i}$$

平均　$\mu_i=n_i p_i$

分散　$\sigma_i^2=n_i p_i(1-p_i)$

2 項分布モデルは、この確率 p_i を説明変数の線形関数で予測する手法である。それらの関係は、以下のような連結関数と呼ばれる関数 $g(x)$ によって与えられる。

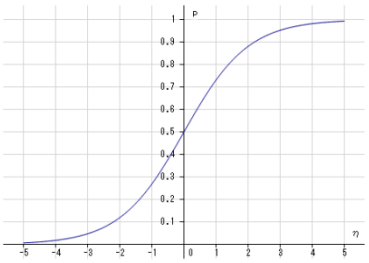
$$\eta_i=g(\mu_i)=g(np_i)=\sum_{j=1}^{p-1}\beta_j x_{ij}+\beta_0$$

連結関数の仮定でよく利用されるモデルが、ロジスティックモデル、プロビットモデル、極値モデル等である。

ロジスティックモデル（最もよく使われる）

$$\eta_i=\log\frac{p_i}{1-p_i}=\sum_{j=1}^{p-1}\beta_j x_{ij}+\beta_0$$

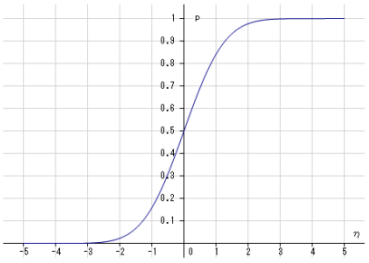
$$\left(p=e^{\eta}/(1+e^{\eta})\right)$$



プロビットモデル（正規分布の密度関数を利用）

$$\eta_i=\Phi^{-1}(p_i)=\sum_{j=1}^{p-1}\beta_j x_{ij}+\beta_0$$

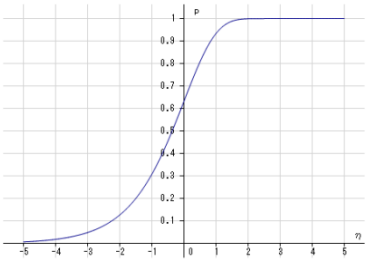
$(p=\Phi(\eta))$　$\Phi(x)$ は正規分布の確率値



極値モデル

$$\eta_i=\log[-\log(1-p_i)]=\sum_{j=1}^{p-1}\beta_j x_{ij}+\beta_0$$

$$\left(p=1-e^{-e^{\eta}}\right)$$



p_i に含まれるパラメータ β_0, β_j は、起きている現実が最も起こり易いという考えで、以下の尤度 L を最大化するように求められる。これを最尤法という。

$$L = \prod_{i=1}^{N} f(y_i; p_i) \qquad \rightarrow \quad 最大化$$

実際の計算では尤度の対数を取った対数尤度を最大化する。

$$\log L = \sum_{i=1}^{N} \log f(y_i; p_i) \qquad \rightarrow \quad 最大化$$

モデルの当てはまりの良さは対数尤度値、逸脱度、ピアソンの χ^2 値、尤度比、R^2 等の値で確認できる。対数尤度値は大きな値になるほどよい。逸脱度はモデルの最適値からのずれを表す統計量で、小さな値になるほどよく、検定は P>0.05 で良好と解釈される。逸脱度と同様に最適値からのずれを表す統計量にピアソンの χ^2 統計量がある。モデルに意味があるかどうか調べる尤度比は最小モデルからのずれを表す統計量で、大きな値になるほどよく、検定は p<0.05 で良好（モデルに意味がある）と解釈される。

20.2　プログラムの利用法

メニュー［分析ー多変量解析他ー判別手法ー2 値ロジスティック回帰］を選択すると、図 1 のような、2 値ロジスティック回帰分析の実行画面が表示される。

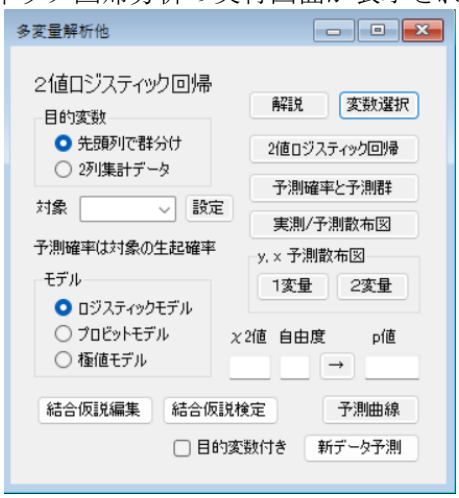


図 1　分析実行画面

利用するデータの形式は、「2 列集計データ」と「先頭列で群分け」があり、それぞれ図 2a と図 2b のように、目的変数が 2 列で表されるか、1 列で表されるかの違いである。

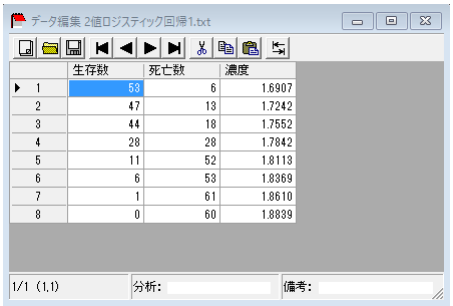


図 2a　2 列集計データ

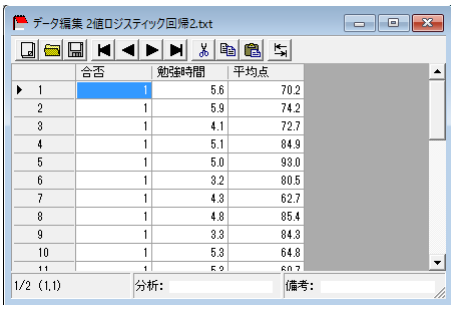
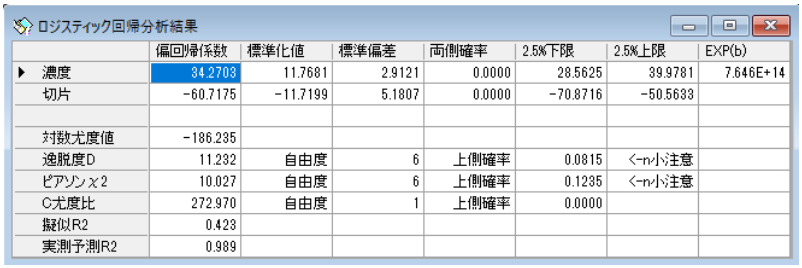


図 2b　先頭列で群分けデータ

目的変数が 2 列で表される場合は、事象 1 が何回起きて、事象 2 が何回起きたかの集計データで、1 列で表される場合は、1 回の試行で事象が起きるかどうかの個体ごとのデータである。2 列の場合、対象変数と非対象変数を入力し、対象変数をコンボボックスで選択しておく。1 列のデータを起きない回数と起きた回数にして 2 列で表現することも可能である。目的変数が 1 列の場合は、2 列の特別な場合と考えてもよい。以後データ形式を分けて、プログラムの出力について説明する。

図 2a のデータのとき、「ロジスティックモデル」ラジオボタンを選択し、「2 値ロジスティック回帰」ボタンをクリックすると図 3 の結果が表示される。


ロジスティック回帰分析結果

	偏回帰係数	標準化値	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
濃度	34.2703	11.7681	2.9121	0.0000	28.5625	39.9781	7.646E+14
切片	-60.7175	-11.7199	5.1807	0.0000	-70.8716	-50.5633	
対数尤度値	-186.235						
逸脱度D	11.232	自由度	6	上側確率	0.0815	<-n小注意	
ピアソンχ2	10.027	自由度	6	上側確率	0.1235	<-n小注意	
C尤度比	272.970	自由度	1	上側確率	0.0000		
擬似R2	0.423						
実測予測R2	0.989						

図 3　2 値ロジスティック回帰結果

ここでは回帰パラメータの値とその検定値、対数尤度値、逸脱度、目的変数と予測値との相関係数の 2 乗値が表示される。

また、「予測確率と予測値」ボタンをクリックすると、個別の実測値、予測確率、予測値が図 4 のように表示される。


予測確率と予測値

	実測値	予測確率	予測値
1	6	0.059	3.457
2	13	0.164	9.842
3	18	0.362	22.451
4	28	0.605	33.898
5	52	0.795	50.096
6	53	0.903	53.291
7	61	0.955	59.222
8	60	0.979	58.743

図 4　予測確率と予測値

「実測/予測散布図」をクリックすると、この実測値と予測値が、図 5 のようにプロットされる。

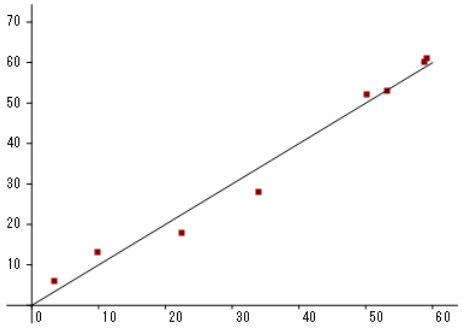


図 5　実測/予測散布図

予測の説明変数が 1 つまたは 2 つの場合、実測値と確率の予測関数（連結関数の逆関数）

の関係を表示することができる。ここでは説明変数が 1 つであるので、「y, x 予測散布図」グループボックス内の「1 変量」ボタンをクリックする。結果は図 6 のようになる。

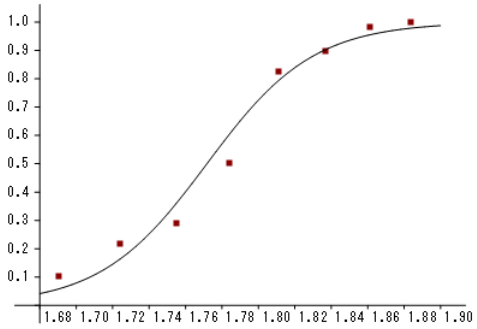


図 6　予測関数とデータ（ロジスティックモデル）

但し、ここでは軸設定を使ってグラフの軸を変更している。

この図と同様に、プロビットモデルと極値モデルの予測関数についても図 7a と図 7b で当てはまりを見てみる。

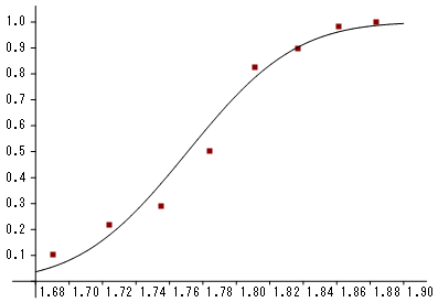


図 7a　プロビットモデル

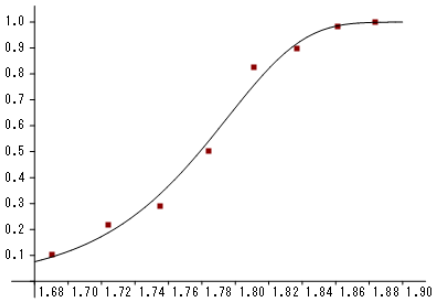


図 7b　極値モデル

これらを比べると極値モデルの当てはまりが良いことが分かる。このことは、「2 値ロジスティック回帰」ボタンで表示される、対数尤度値、逸脱度 D、R^2 の値でも確認できる。

次に図 2b のデータを用いた場合のロジスティックモデルの実行結果を示す。目的変数は「0/1 形式（1 列）」を選択し、「2 値ロジスティック回帰」ボタンをクリックすると図 8 のような結果が表示される。

ロジスティック回帰分析結果

	偏回帰係数	標準化値	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 勉強時間	5.0765	2.0267	2.5049	0.0427	0.1670	9.9861	160.2202
平均点	0.4581	2.0532	0.2231	0.0400	0.0208	0.8955	1.5811
切片	-52.7491	-2.1184	24.9005	0.0341	-101.5540	-3.9442	
対数尤度値	-5.569						
逸脱度D	11.137	自由度	27	上側確率	0.9969	<-n小注意	
ピアソンχ2	13.422	自由度	27	上側確率	0.9863	<-n小注意	
C尤度比	29.917	自由度	2	上側確率	0.0000		
擬似R2	0.729						
実測予測R2	0.864						
誤判別確率	1を他と	0.077	0を他と	0.059			

図 8　2 値ロジスティック回帰結果

このデータ形式では、以下に述べる、予測による 0/1 の判別についての誤判別確率が追加されている。

また、「予測確率と予測値」ボタンをクリックすると、個体別の実測値、予測確率、予測値が図 9 のように表示される。

	実測値	予測確率	予測値
9	1	0.932	1.000
10	1	0.979	1.000
11	1	0.877	1.000
12	1	1.000	1.000
13	1	0.991	1.000
14	0	0.000	0.000
15	0	0.374	0.000
16	0	0.070	0.000
17	0	0.005	0.000
18	0	0.000	0.000

図 9　予測確率と予測値

ここでは、予測値として、予測確率が 0.5 未満なら 0、予測確率が 0.5 以上なら 1 が与えられている。この予測値と実測値との違いを表すのが、図 8 の誤判別確率である。

「実測/予測散布図」をクリックすると、この実測値と予測確率が、図 10 のように表示される。ここに、図 10 の図 5 との違いは、実測値と予測値の代わりに実測値と予測確率を用いているところである。

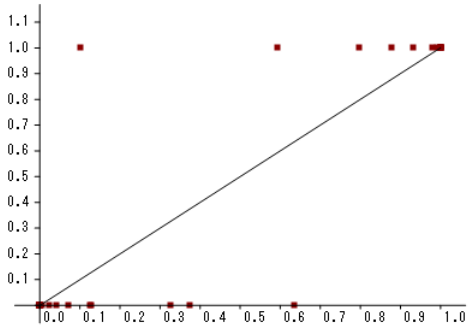


図 10　実測/予測確率散布図

このデータでは説明変数が 2 つであるので、「y, x 予測散布図」グループボックス内の「2 変量」ボタンをクリックする。結果は図 11 のようなグラフになる。

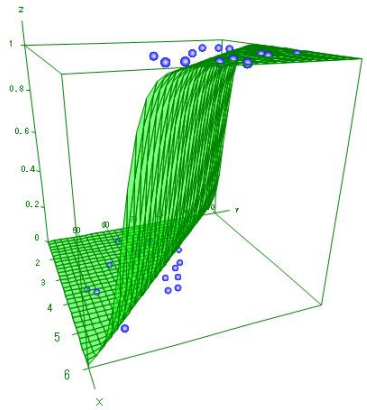


図 11　予測関数とデータ（ロジスティックモデル）

最後に、分析実行画面の下部に、利用する可能性のある χ^2 分布の確率を求めるボタンを追加しておいた。専用のメニューもあるが、必要に応じて利用してもらいたい。

分析実行画面の一番下にある「結合仮説」チェックボックスは、結合仮説検定を行うかどうかのチェックボックスである。結合仮説検定では、複数の係数が同時に 0 であるかどうかや係数の線形結合がある値となるかどうかなどを調べる。使い方の詳細は重回帰分析のプログラムの解説のところにあるので参照してもらいたい。

予測について

機械学習などでは、訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の一番下に「新データ予測」ボタンがあるが、これは一度 2 値ロジスティック分析を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は「目的変数付き」チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を正解率の値で示してくれる。これは「先頭列で群分け」のとき有効である。

問題 1

2 値ロジスティック回帰 1.txt は毒物の濃度による生存数と死亡数の結果である。この（恐ろしい）データを用いて以下の問いに答えよ。

目的変数は y1,y2 形式(2 列)で、対象を死亡数、ロジスティック回帰とすること。

１）偏回帰係数の値を求めよ。（p は予測死亡確率である）

$$\eta = \log\frac{p}{1-p} =$$ ［　　］濃度＋［　　］

２）これらの２つの係数は 0 でないといえるか。

濃度係数　検定確率［　　］　0 と異なると［いえる・いえない］

切片　　検定確率［　　］　0 と異なると［いえる・いえない］

３）最適値からのずれを表す逸脱度、最小モデルからのずれを表す尤度比の値はいくらか。これらの値から、このモデルは有効と考えられるか。

逸脱度［　　］　検定確率［　　］　モデルは［有効・有効でない］

尤度比［　　］　検定確率［　　］　モデルは［有効・有効でない］

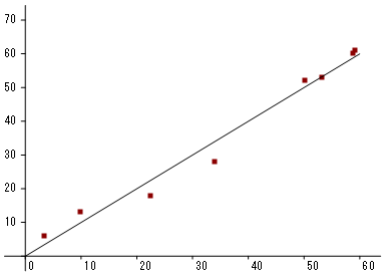
注）逸脱度は小さいほど良い（$p>0.05$）、尤度比は大きいほど良い（$p<0.05$）。

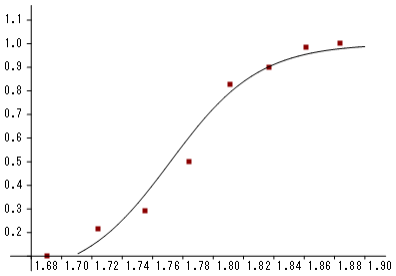
４）先頭データの死亡数の実測値、予測確率、予測値を求めよ。

実測値［　　］，予測確率［　　］，予測値［　　］

５）実測値（縦軸）と予測値（横軸）の実測/予測散布図を描け。下図左

６）濃度を横軸に取った実測値、予測関数のグラフ（予測散布図）を描け。下図右





７）実測値と予測値の相関係数の 2 乗の値はいくらか。［　　］

８）3 つのモデルの中で最も適合の良いモデルは何か。

［ロジスティックモデル・プロビットモデル・極値モデル］

問題２

Samples¥判別分析 1.txt は勉強時間と模擬テストの平均点で入試の合否を予測するためのデータである。対象を 1（合格群）として 2 値ロジスティック回帰と判別分析を比較しながら以下の問いに答えよ。

１）偏回帰係数の値を求めよ。（p は予測死亡確率である）

$$\log\frac{p}{1-p} =$$［　　］勉強時間＋［　　］平均点＋［　　］

２）これら 3 つの係数は 0 でないといえるか。

勉強時間係数　検定確率［　　］　0 と異なると［いえる・いえない］

平均点係数　　検定確率［　　］　0 と異なると［いえる・いえない］

切片　　検定確率［　　］　0 と異なると［いえる・いえない］

３）標準化係数や上の検定確率から、判別に対してより影響の強い変数は何か。

［勉強時間・平均点］

４）最適値からのずれを表す逸脱度、最小モデルからのずれを表す尤度比の値はいくらか。これらの値から、このモデルは有効と考えられるか。

逸脱度［　　］　検定確率［　　］　モデルは［有効・有効でない］

尤度比［　　］　検定確率［　　］　モデルは［有効・有効でない］

注）逸脱度は小さいほど良い（$p>0.05$）、尤度比は大きいほど良い（$p<0.05$）。

５）誤判別確率を判別分析と比較せよ。

ロジスティック回帰　1 群を他群と［　　］　2 群を他群と［　　］

判別分析　　1 群を他群と［　　］　2 群を他群と［　　］

６）5，6，7 番の人の合格予測確率を求めよ。

5 番［　　］　6 番［　　］　7 番［　　］

７）5，6，7 番の人はそれぞれどのように予測されたか。

5 番［合格・不合格］　6 番［合格・不合格］　7 番［合格・不合格］

判別の分点は予測確率がいくらのところか。［　　］

８）平均点が１点上がると、合否のリスク比は何倍になるか。（但し、これによる回帰係数の変化は十分小さいとする）　　［　　］倍

注）判別分析は誤判別確率でしか予測を表せないが、ロジスティック回帰分析では予測確率を求めることができる。ただ、ロジスティックモデル・プロビットモデル・極値モデル等、確率と線形の式を結ぶ「連結関数」に任意性が残る。（逆にこれをうまく選ぶことによって良い予測につながるとも言える）

問題３

2 値ロジスティック回帰 2.txt (5 頁目) のデータを用いて以下の問いに答えよ。データは、先頭列で群分け形式で、対象を 1（発症群）、モデルはロジスティックモデルとすること。

１）対数オッズを予測する回帰式の偏回帰係数の値を求めよ。（p は予測発症確率である）

$$\log\frac{p}{1-p} = [\quad] \text{要因１} + [\quad] \text{要因２} + [\quad]$$

２）これら３つの係数は 0 でないといえるか。

要因１係数　　検定確率［　　］　0 と異なると［いえる・いえない］

要因２係数　　検定確率［　　］　0 と異なると［いえる・いえない］

切片　　検定確率［　　］　0 と異なると［いえる・いえない］

３）各要因の有無による発症オッズの比（罹患危険率の比）は EXP(b)の欄で与えられているが、２つの要因でそれぞれいくらか。

要因１［　　］　要因２［　　］

４）最適値からのずれを表す逸脱度、最小モデルからのずれを表す尤度比の値はいくらか。これらの値から、このモデルは有効と考えられるか。

逸脱度［　　］　モデルは［有効・有効でない］

尤度比［　　］　モデルは［有効・有効でない］

注）逸脱度は小さいほど良い（p>0.05）、尤度比は大きいほど良い（p<0.05）。

５）所属群の判定で、誤判別確率はいくらか。

1 群（合格群）を他と　［　　］

0 群（不合格群）を他と［　　］

６）判別の分点は予測確率がいくらのところか。［　　］

７）4 番目の人の実測値、予測確率、予測値を求めよ。

実測値［　　］，予測確率［　　］，予測値［　　］

問題１解答

目的変数は y1, y2 形式(2 列)で、対象を死亡数、ロジスティック回帰とすること。

１）偏回帰係数の値を求めよ。（p は予測死亡確率である）

$$\eta = \log\frac{p}{1-p} = [\ 34.2703\ ] \text{濃度} + [\ -60.7175\ ]$$

２）これらの２つの係数は 0 でないといえるか。
濃度係数　検定確率［ 0.0000 ］　0 と異なると［いえる・いえない］
切片　　検定確率［ 0.0000 ］　0 と異なると［いえる・いえない］
３）最適値からのずれを表す逸脱度、最小モデルからのずれを表す尤度比の値はいくらか。これらの値から、このモデルは有効と考えられるか。
逸脱度［ 11.232 ］　検定確率［ 0.0815 ］　モデルは［有効・有効でない］
尤度比［ 272.970 ］　検定確率［0.0000］　モデルは［有効・有効でない］
注）逸脱度は小さいほど良い（p>0.05)、尤度比は大きいほど良い（p<0.05)。
４）先頭データの死亡数の実測値、予測確率、予測値を求めよ。
実測値［ 18 ］, 予測確率［ 0.362 ］, 予測値［ 22.451 ］
５）実測値（縦軸）と予測値（横軸）の実測/予測散布図を描け。下図左
７）実測値と予測値の相関係数の 2 乗の値はいくらか。［ 0.989 ］

８）3 つのモデルの中で最も適合の良いモデルは何か。
［ロジスティックモデル・プロビットモデル・極値モデル］

問題２解答

１）偏回帰係数の値を求めよ。(p は予測死亡確率である)

$$\log\frac{p}{1-p} =$$［ 5.0765 ］勉強時間＋［0.4581］平均点＋［ -52.7491 ］

２）これら３つの係数は 0 でないといえるか。
勉強時間係数　検定確率［ 0.0427 ］　0 と異なると［いえる・いえない］
平均点係数　　検定確率［ 0.0400 ］　0 と異なると［いえる・いえない］
切片　　検定確率［ 0.0341 ］　0 と異なると［いえる・いえない］
３）標準化係数や上の検定確率から、判別に対してより影響の強い変数は何か。
［勉強時間・平均点］
４）最適値からのずれを表す逸脱度、最小モデルからのずれを表す尤度比の値はいくらか。これらの値から、このモデルは有効と考えられるか。
逸脱度［ 11.137 ］　検定確率［ 0.9969 ］　モデルは［有効・有効でない］
尤度比［ 29.917 ］　検定確率［ 0.0000 ］　モデルは［有効・有効でない］
注）逸脱度は小さいほど良い（p>0.05)、尤度比は大きいほど良い（p<0.05)。
５）誤判別確率を判別分析と比較せよ。
ロジスティック回帰　1 群を他群と［ 0.077 ］　2 群を他群と［ 0.059 ］
判別分析　　1 群を他群と［ 0.077 ］　2 群を他群と［ 0.059 ］
６）5，6，7 番の人の合格予測確率を求めよ。
5 番［ 1.000 ］　6 番［ 0.593 ］　7 番［ 0.100 ］
７）5，6，7 番の人はそれぞれどのように予測されたか。
5 番［合格・不合格］　6 番［合格・不合格］　7 番［合格・不合格］
判別の分点は予測確率がいくらのところか。［ 0.5 ］
８）平均点が 1 点上がると、合否のリスク比は何倍になるか。(但し、これによる回帰係数の変化は十分小さいとする)　　［ 1.581 ］倍

問題３解答

１）対数オッズを予測する回帰式の偏回帰係数の値を求めよ。(p は予測発症確率である)

$$\log\frac{p}{1-p} =$$［ 2.0953 ］要因１＋［ 0.1324 ］要因２＋［ -0.5654 ］

２）これら３つの係数は 0 でないといえるか。
要因１係数　　検定確率［ 0.0076 ］　0 と異なると［いえる・いえない］
要因２係数　　検定確率［ 0.8578 ］　0 と異なると［いえる・いえない］
切片　　検定確率［ 0.2032 ］　0 と異なると［いえる・いえない］
３）各要因の有無による発症オッズの比（罹患危険率の比）は EXP(b)の欄で与えられているが、２つの要因でそれぞれいくらか。
要因１［ 8.1278 ］　要因２［ 1.1416 ］
４）最適値からのずれを表す逸脱度、最小モデルからのずれを表す尤度比の値はいくらか。これらの値から、このモデルは有効と考えられるか。
逸脱度［ 51.782 ］　モデルは［有効・有効でない］
尤度比［ 10.044 ］　モデルは［有効・有効でない］
注）逸脱度は小さいほど良い（p>0.05）、尤度比は大きいほど良い（p<0.05）。
５）所属群の判定で、誤判別確率はいくらか。
1 群（合格群）を他と　［ 0.400 ］
0 群（不合格群）を他と［ 0.150 ］
６）判別の分点は予測確率がいくらのところか。［ 0.5 ］
７）4 番目の人の実測値、予測確率、予測値を求めよ。
実測値［ 1 ］，予測確率［ 0.822 ］，予測値［ 1 ］

20.3　一般化線形モデルの理論

２値ロジスティック回帰分析は２項分布の確率を説明変数の１次式の関数で予測する分析手法である。この分析は、確率の大きさによって事象の出現、非出現を区別することにも使え、質的データの予測手法としても利用できる。同様の分析に判別分析があるが、これは説明変数によるマハラノビス距離を用いた判別方法で、理論的な根拠という点ではロジスティック回帰分析の方が優れている。ここでは今後のために、参考文献[2] に従って、一般化線形モデルの基礎からロジスティック回帰分析について説明しておく。

1）指数型分布族

ある単一のパラメータ θ を持つ確率変数 Y が以下の密度関数に従うとき、その分布を指数型分布族という。

$$f(y;\theta)=\exp[a(y)b(\theta)+c(\theta)+d(y)]$$

指数型分布族には、ポアソン分布、正規分布、2 項分布等が含まれる。特に $a(y)=y$ のとき分布は正準形であると言われ、$b(\theta)$ は分布の自然パラメータと呼ばれる。

確率変数 $a(\mathrm{Y})$ については

$$\frac{d}{d\theta}\int f(y;\theta)dy=\int[a(y)b'(\theta)+c'(\theta)]f(y;\theta)dy$$
$$=E[a(Y)]b'(\theta)+c'(\theta)=0$$

より、

$$E[a(Y)]=-c'(\theta)/b'(\theta)$$

$$\frac{d^2}{d\theta^2}\int f(y;\theta)dy$$
$$=\int[a(y)b''(\theta)+c''(\theta)]f(y;\theta)dy+\int[a(y)b'(\theta)+c'(\theta)]^2 f(y;\theta)dy$$
$$=E[a(Y)^2]b'(\theta)^2+E[a(Y)][b''(\theta)+2b'(\theta)c'(\theta)]+c''(\theta)+c'(\theta)^2$$
$$=E[a(Y)^2]b'(\theta)^2-\frac{c'(\theta)}{b'(\theta)}[b''(\theta)+2b'(\theta)c'(\theta)]+c''(\theta)+c'(\theta)^2$$
$$=E[a(Y)^2]b'(\theta)^2-E[a(Y)]^2b'(\theta)^2+\frac{1}{b'(\theta)}[-b''(\theta)c'(\theta)+c''(\theta)b'(\theta)]$$
$$=V[a(Y)]b'(\theta)^2+\frac{1}{b'(\theta)}[-b''(\theta)c'(\theta)+c''(\theta)b'(\theta)]=0$$

より、

$$V[a(Y)]=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{b'(\theta)^3}$$

という性質がある。

対数密度関数 $l(y;\theta)=\log f(y;\theta)$ の θ に関する微分の y を確率変数とみなした

$$U(Y;\theta)=a(Y)b'(\theta)+c'(\theta)$$

は、スコア統計量とも呼ばれ、その分布の期待値と分散は上式を使うと以下となる。

$$E[U]=0$$
$$V[U]=V[a(Y)]b'(\theta)^2=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{b'(\theta)}$$

さらに、

$$V[\mathrm{U}]=E[U^2]-E[U]^2=E[U^2]$$
$$E[U']=E[a(Y)]b''(\theta)+c''(\theta)=-\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{b'(\theta)}=-V[U]$$

の関係より、以下も成り立つ。

$$V[U]=E[U^2]=-E[U']$$

スコア統計量の分散$V[U]$は情報量とも呼ばれる。

2）正準形の一般化線形モデル

正準形の指数型分布族の分布に従う確率変数$Y_i\ (i=1,2,\cdots,N)$が、パラメータθ_iの同じ形の以下の独立な密度関数の分布に従うと考える。

$$f(y_i;\theta_i)=\exp[y_i b(\theta_i)+c(\theta_i)+d(y_i)]$$

対数密度関数$l(y_i;\theta_i)$は以下で与えられる。

$$l(y_i;\theta_i)=y_i b(\theta_i)+c(\theta_i)+d(y_i)$$

確率変数Y_iの平均と分散は前節の議論より、以下のように与えられる。

$$E[Y_i]=-c'(\theta_i)/b'(\theta_i)\equiv\mu_i$$

$$V[Y_i]=\frac{b''(\theta_i)c'(\theta_i)-c''(\theta_i)b'(\theta_i)}{b'(\theta_i)^3}$$

ここで、θ_i は μ_i の関数であるとみることができる。

我々はこの μ_i に対して、ある説明変数 ${}^t\mathbf{x}_i=(1,x_{i1},\cdots,x_{ip})$（$i=1,\cdots,N$）とパラメータ ${}^t\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)$ を用いて以下のような仮定をする。

$$\eta_i\equiv g(\mu_i)=\sum_{j=1}^{p-1}\beta_j x_{ij}+\beta_0={}^t\boldsymbol{\beta}\mathbf{x}_i$$

この仮定により、θ_i は $\boldsymbol{\beta}$ の関数と見ることができる。またこの関係を与える関数 $\eta_i=g(\mu_i)$ を連結関数という。

確率変数 $\boldsymbol{Y}_i$ の同時密度関数（尤度関数）は以下で与えられる。

$$f(\mathbf{y};\boldsymbol{\theta})=\exp\left[\sum_{i=1}^{N}\{y_i b(\theta_i)+c(\theta_i)+d(y_i)\}\right]$$

また対数尤度関数 $l(\mathbf{y};\boldsymbol{\theta})$ は以下のようになる。

$$l(\mathbf{y};\boldsymbol{\theta})=\sum_{i=1}^{N}l_i=\sum_{i=1}^{N}[y_i b(\theta_i)+c(\theta_i)+d(y_i)] \qquad (1)$$

この対数尤度関数の β_j による微分をスコアベクトルと呼び、$\boldsymbol{U}_j$ とすると、スコアベクトル $\boldsymbol{U}_j$ は以下のようになる。

$$U_j=\sum_{i=1}^{N}\frac{\partial l_i}{\partial\beta_j}=\sum_{i=1}^{N}\frac{\partial\mu_i}{\partial\beta_j}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial l_i}{\partial\theta_i}$$

ここで、

$$\frac{\partial l_i}{\partial\theta_i}=y_i b'(\theta_i)+c'(\theta_i)=b'(\theta_i)\left[y_i+c'(\theta_i)/b'(\theta_i)\right]$$

$$=b'(\theta_i)(y_i-E[Y_i])=b'(\theta_i)(y_i-\mu_i)$$

$$\frac{\partial\theta_i}{\partial\mu_i}=\frac{b'(\theta_i)^2}{-c''(\theta_i)b'(\theta_i)+c'(\theta_i)b''(\theta_i)}=\frac{1}{b'(\theta_i)V[Y_i]}$$

$$\frac{\partial\mu_i}{\partial\beta_j}=\frac{\partial\eta_i}{\partial\beta_j}\frac{\partial\mu_i}{\partial\eta_i}=x_{ij}\frac{\partial\mu_i}{\partial\eta_i}$$

となることから、以下の関係を得る。

$$U_j=\sum_{i=1}^{N}\frac{\partial l_i}{\partial\beta_j}=\sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{V[Y_i]}\frac{\partial\mu_i}{\partial\eta_i} \qquad (2)$$

また、以下も成り立つ。

$$E[U_j]=0 \qquad (3)$$

さらに $\boldsymbol{U}_j$ の β_k による微分を U_{jk} とすると、U_{jk} は以下のようになる。

$$U_{jk} \equiv \frac{\partial U_j}{\partial \beta_k} = \frac{\partial}{\partial \beta_k}\sum_{i=1}^{N}\frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N}\frac{\partial \mu_i}{\partial \beta_k}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial}{\partial \theta_i}\left(\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial l_i}{\partial \theta_i}\right)$$

$$= \sum_{i=1}^{N}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2\frac{\partial}{\partial \theta_i}\left(\frac{\partial l_i}{\partial \theta_i}\right) + \sum_{i=1}^{N}\frac{\partial \mu_i}{\partial \beta_k}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial l_i}{\partial \theta_i}\frac{\partial}{\partial \theta_i}\left(\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \theta_i}{\partial \mu_i}\right)$$

$$= \sum_{i=1}^{N}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2[y_i b''(\theta_i) + c''(\theta_i)]$$

$$+ \sum_{i=1}^{N}\frac{(y_i - \mu_i)x_{ik}}{V[Y_i]}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial}{\partial \theta_i}\left(\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \theta_i}{\partial \mu_i}\right)$$

また、

$$E[Y_i b''(\theta_i) + c''(\theta_i)] = E[Y_i]b''(\theta_i) + c''(\theta_i) = -c'(\theta_i)b''(\theta_i)/b'(\theta_i) + c''(\theta_i)$$

$$= \frac{-c'(\theta_i)b''(\theta_i) + b'(\theta_i)c''(\theta_i)}{b'(\theta_i)} = -b'(\theta_i)^2 V[Y_i]$$

$$E\left[\sum_{i=1}^{N}\frac{(Y_i - \mu_i)x_{ik}}{V[Y_i]}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial}{\partial \theta_i}\left(\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \theta_i}{\partial \mu_i}\right)\right] = 0$$

であることから、(2) 式を求める際の計算により、U_{jk} の変数の値を確率変数で置き換えて計算すると以下となる。

$$E[U_{jk}] = -\sum_{i=1}^{N}\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_k}\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2 b'(\theta_i)^2 V[U_i] = -\sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

一方、(2) の表式より、

$$E[U_j U_l] = \sum_{i=1}^{N}\sum_{k=1}^{N}\frac{E[(y_i - \mu_i)(y_k - \mu_k)]x_{ij}x_{kl}}{V[Y_i]V[Y_k]}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \mu_k}{\partial \eta_k}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{N}\frac{V[Y_i]\delta_{ik}x_{ij}x_{kl}}{V[Y_i]V[Y_k]}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \mu_k}{\partial \eta_k} = \sum_{i=1}^{N}\frac{x_{ij}x_{il}}{V[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

であることから、以下となる。

$$E[U_{jk}] = -E[U_j U_k] \tag{4}$$

ここで、$(\mathfrak{I})_{jk} = -E[U_{jk}] = E[U_j U_k]$ とすると、行列 $\mathfrak{I}$ は情報行列と呼ばれる。

$$(\mathfrak{I})_{jk} = E(U_j U_k) = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{5}$$

今、(1) で与えられる対数尤度関数が最大となる $\boldsymbol{\beta}$ の値を求めてみよう。これには、

$$\frac{\partial l}{\partial \beta_j} = U_j = 0$$

という方程式を解くことになる。

（$f(x) = 0$ を解くには $y_m = f'(x_{m-1})(x_m - x_{m-1}) - f(x_{m-1}) = 0$ を計算することから）

解はニュートン・ラフソン法によると、

$$U_j^{(m)} = \sum_{k=1}^{p} U_{jk}^{(m-1)}(\beta_k^{(m)} - \beta_k^{(m-1)}) + U_j^{(m-1)} = 0$$

のように、$\beta_k^{(m)}$ の値を逐次求めて行くことになるが、実際の計算では $U_{jk}^{(m-1)}$ の代わりに、期待値を取った情報行列 $-(\Im^{(m-1)})_{jk}$ を用いる。この式を書き変えると、以下となる。

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + (\Im^{(m-1)})^{-1}\mathbf{U}^{(m-1)}$$

(3)式と(5)式を元にして、大標本においては、スコアベクトルの分布は漸近的に

$$\mathbf{U} \sim \mathbf{N}(\mathbf{0}, \Im), \quad {}^t\mathbf{U}\Im^{-1}\mathbf{U} \sim \chi^2(p)$$

であることも示される。

最尤推定量 $l(\boldsymbol{\beta})$ の推定値 $\mathbf{b}$ の近傍でのテイラー展開近似は以下となり、

$$l(\boldsymbol{\beta}) = l(\mathbf{b}) + {}^t(\boldsymbol{\beta}-\mathbf{b})\mathbf{U}(\mathbf{b}) - \frac{1}{2}{}^t(\boldsymbol{\beta}-\mathbf{b})\Im(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b})$$

スコアベクトルの推定値 $U(\mathbf{b})$ の近傍でのテイラー展開近似は以下となる。

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}(\mathbf{b}) - \Im(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b}) = -\Im(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b})$$

ここでは $E[\partial U_j/\partial\beta_k] = -\Im_{jk}$ や $\mathbf{U}(\mathbf{b}) = \mathbf{0}$ を使っている。これらより、

$${}^t(\boldsymbol{\beta}-\mathbf{b})\Im(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b}) \sim \chi^2(p)$$

も示される。また、同様にして以下も示される。

$$\mathbf{b} = \boldsymbol{\beta} + \Im^{-1}\mathbf{U} \sim N(\boldsymbol{\beta}, \Im^{-1}E(\mathbf{U}\mathbf{U}')\Im^{-1}) = N(\boldsymbol{\beta}, \Im^{-1}\Im\Im^{-1}) = N(\boldsymbol{\beta}, \Im^{-1})$$

後に説明するが、モデルの最適値からのずれを表す**逸脱度** D を以下のように定義する。

$$D = 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\hat{\mathbf{b}};\mathbf{y})] \sim \chi^2(N-p)$$

ここに、$l(\mathbf{b}_{\max};\mathbf{y})$ はパラメータ数 N の飽和モデルでの対数尤度、$l(\hat{\mathbf{b}};\mathbf{y})$ は現在考えているパラメータ数 p のモデルでの対数尤度である。同じパラメータ数では、この値が小さい連結関数のモデルほど適合が良いと判断する。但し、分布は漸近的に成り立つものである。

モデルに意味があるかどうかの検定では、以下の**尤度比** χ^2 統計量が使われる。

$$C = 2[l(\hat{\mathbf{b}};\mathbf{y}) - l(\mathbf{b}_{\min};\mathbf{y})] \sim \chi^2(p-1)$$

ここに $l(\mathbf{b}_{\min};\mathbf{y})$ は定数パラメータ 1 つの最小モデルの対数尤度、$l(\hat{\mathbf{b}};\mathbf{y})$ は現在考えている、パラメータ数 p のモデルでの対数尤度である。これは、帰無仮説として最小モデルが正しい（回帰式は意味がない）とする検定である。

2 項分布の場合、対数尤度関数は以下で与えられる。

$$l(\boldsymbol{\beta};\mathbf{y}) = \sum_{i=1}^{N}\left[y_i \log p_i + (n_i - y_i)\log(1-p_i) + \log\binom{n_i}{y_i}\right]$$

$l(\mathbf{b}_{\max};\mathbf{y})$ と $l(\hat{\mathbf{b}};\mathbf{y})$ は、確率 p_i の中に、それぞれ実測値を用いた $p_i = y_i/n_i$ と予測値を用いた $\hat{p}_i = \hat{y}_i/n_i$ を代入して求める。$l(\mathbf{b}_{\min};\mathbf{y})$ は確率 p_i の中に同じ以下の値を代入して求める。

$$\tilde{p} = \sum_{i=1}^{N} y_i \Big/ \sum_{i=1}^{N} n_i$$

実測値と推測値の関係を与える指標として、決定係数からの類推である以下の擬似 R^2 も利用される。

$$\tilde{R}^2 = \frac{l(\mathbf{b}_{\min};\mathbf{y}) - l(\hat{\mathbf{b}};\mathbf{y})}{l(\mathbf{b}_{\min};\mathbf{y})}$$

さらにプログラムでは、実測値と予測値の相関係数も求めている。

3）2 項分布モデル

2 項分布のパラメータを説明変数の線形結合で推測する場合、密度関数、密度関数の対数、逸脱度、目的変数の平均と分散は以下のようになる。ここで、密度関数の対数の最後の項はパラメータに依存していないので、計算上は考えないことにする（参考文献[2] の数値に従っている）。

$$f(y_i;p_i) = {}_{n_i}C_{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}$$

$$\begin{aligned} l(y_i;p_i) &= \log[{}_{n_i}C_{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}] \\ &= y_i \log p_i + (n_i - y_i)\log(1-p_i) + \log {}_{n_i}C_{y_i} \\ &= y_i[\log p_i - \log(1-p_i)] + n_i \log(1-p_i) + \log {}_{n_i}C_{y_i} \\ &\to y_i[\log p_i - \log(1-p_i)] + n_i \log(1-p_i) \end{aligned}$$

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i p_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i p_i}\right)\right] \sim \chi^2(N-p)$$

$$C = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{\hat{y}_i}{n_i \tilde{p}}\right) + (n_i - y_i)\log\left(\frac{n_i - \hat{y}_i}{n_i - n_i \tilde{p}}\right)\right] \sim \chi^2(p-1)$$

$$\tilde{p} = \sum_{i=1}^{N} y_i \Big/ \sum_{i=1}^{N} n_i\,, \qquad E[Y_i] = n_i p_i \equiv \mu_i\,,\ \ V[Y_i] = n_i p_i(1-p_i)$$

ここでは n_i 回の試行に対して、y_i 回の事象が起こったとしているが、１回の試行で起こったか起こらないかにする場合は、$n_i = 1,\ y_i = \{0,1\}$ とすればよい。

逸脱度と同様に最適値からのずれを表す統計量に以下の**ピアソン χ^2 統計量**がある。

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{n_i \hat{p}_i(1-\hat{p}_i)} \sim \chi^2(N-p)$$

これは逸脱度と漸近的に同じ指標であるが、逸脱度と比べてこちらの方が分布によく適合するという意見もある。

これまでは、2 項分布に基づく一般論であったが、これ以降は、説明変数との関係を与える連結関数の部分に仮定が入る。連結関数の仮定でよく利用されるモデルが、ロジスティックモデル、プロビットモデル、極値モデル等である。以下に最終的な計算で用いられる式を与えておく。

ロジスティックモデル

$$\eta_i = \log\frac{p_i}{1-p_i} = \sum_{j=1}^{p-1}\beta_j x_{ij} + \beta_0$$

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}}, \quad 1-p_i = \frac{1}{1+e^{\eta_i}}$$

$$\mu_i = n_i p_i = \frac{n_i e^{\eta_i}}{1+e^{\eta_i}}, \qquad \frac{\partial\mu_i}{\partial\eta_i} = \frac{n_i e^{\eta_i}}{(1+e^{\eta_i})^2} = n_i p_i(1-p_i)$$

$$U_j = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{V[Y_i]}\frac{\partial\mu_i}{\partial\eta_i} = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{n_i p_i(1-p_i)} n_i p_i(1-p_i) = \sum_{i=1}^{N}(y_i-\mu_i)x_{ij}$$

$$(\Im)_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}n_i^2 p_i^2(1-p_i)^2}{n_i p_i(1-p_i)} = \sum_{i=1}^{N} x_{ij}x_{ik}n_i p_i(1-p_i)$$

プロビットモデル

$$\eta_i = \Phi^{-1}(p_i) = \sum_{j=1}^{p-1}\beta_j x_{ij} + \beta_0$$

$$p_i = \Phi(\eta_i)$$

$$\mu_i = n_i p_i = n_i\Phi(\eta_i), \qquad \frac{\partial\mu_i}{\partial\eta_i} = \frac{n_i}{\sqrt{2\pi}}\exp(-\eta_i^2/2)$$

$$U_j = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{V[Y_i]}\frac{\partial\mu_i}{\partial\eta_i} = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{p_i(1-p_i)}\frac{\exp(-\eta_i^2/2)}{\sqrt{2\pi}}$$

$$(\Im)_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{p_i(1-p_i)}\frac{n_i\exp(-\eta_i^2)}{2\pi}$$

極値モデル

$$\eta_i = \log[-\log(1-p_i)] = \sum_{j=1}^{p-1}\beta_j x_{ij} + \beta_0$$

$$p_i = 1-\exp[-\exp(\eta_i)] \qquad (1-p_i = \exp[-\exp(\eta_i)])$$

$$\mu_i = n_i p_i, \qquad \frac{\partial\mu_i}{\partial\eta_i} = n_i\frac{\partial p_i}{\partial\eta_i} = n_i\exp(\eta_i)\exp[-\exp(\eta_i)]$$

$$U_j = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{V[Y_i]}\frac{\partial\mu_i}{\partial\eta_i} = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{p_i(1-p_i)}\exp(\eta_i)\exp[-\exp(\eta_i)]$$

$$= \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{p_i}\exp(\eta_i)$$

$$(\Im)_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{p_i(1-p_i)}n_i\exp(2\eta_i)\exp[-2\exp(\eta_i)]$$

$$= \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{p_i}n_i(1-p_i)\exp(2\eta_i)$$

極値モデルの計算には以下の性質を利用する。

$$\lim_{p\to 0}\exp(\eta)/p = 1$$

プロビットモデルと極値モデルの場合、 $p_i \to 0$ や $p_i \to 1$ のときに、計算機のまるめ誤差や分布関数の近似誤差から、除算のエラーが生じることがある。そのため、プログラムではある程度のところで、これらの極限を止めるようにしている。また最終結果でも対数尤度の計算で同様のことが起こる可能性があるので、同じように極端な値を避けるようにしている。現在のプログラムでは、$0.000001 \le p_i \le 0.999999$ の範囲に設定している。

20.4　一般化線形モデルについて（自分なりの説明）

準備

$$\int\prod_{\lambda=1}^{N} f(y_\lambda\,|\,\theta_\lambda)dy_1\cdots dy_N = 1,\quad \theta_\lambda = \theta_\lambda(\mathbf{x}_\lambda,\boldsymbol{\beta})\text{ として}$$

$$\begin{aligned}
0 &= \frac{\partial^2}{\partial\beta_i\partial\beta_j}\int\prod_{\lambda=1}^{N} f(y_\lambda\,|\,\theta_\lambda)\mathbf{dy}\\
&= \frac{\partial^2}{\partial\beta_i\partial\beta_j}\int\exp\left[\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right]\mathbf{dy}\\
&= \frac{\partial}{\partial\beta_i}\int\exp\left[\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right]\frac{\partial}{\partial\beta_j}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\mathbf{dy}\\
&= \frac{\partial}{\partial\beta_i}\int\prod_{\lambda=1}^{N} f(y_\lambda\,|\,\theta_\lambda)\frac{\partial}{\partial\beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\mathbf{dy}\\
&= \int\prod_{\lambda=1}^{N} f(y_\lambda\,|\,\theta_\lambda)\left[\frac{\partial}{\partial\beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right]\left[\frac{\partial}{\partial\beta_j}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right]\mathbf{dy}\\
&\quad + \int\prod_{\lambda=1}^{N} f(y_\lambda\,|\,\theta_\lambda)\frac{\partial^2}{\partial\beta_i\partial\beta_j}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\mathbf{dy}
\end{aligned}$$

これらを書き換えて、以下を得る。

$$E\left[\frac{\partial}{\partial\beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right] = 0$$

$$E\left[\left\{\frac{\partial}{\partial\beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right\}\left\{\frac{\partial}{\partial\beta_j}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right\}\right] = -E\left[\frac{\partial^2}{\partial\beta_i\partial\beta_j}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda)\right] \tag{1}$$

また、

$$u_i(\boldsymbol{\beta}) \equiv \frac{\partial}{\partial\beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda) \qquad U_{jk}(\boldsymbol{\beta}) \equiv \frac{\partial^2}{\partial\beta_j\partial\beta_k}\sum_{\lambda=1}^{N}\log f(y_\lambda\,|\,\theta_\lambda) \tag{2}$$

のように定義すると以下となる。

$$E[u_i(\boldsymbol{\beta})] = 0$$
$$Cov[u_i(\boldsymbol{\beta}), u_j(\boldsymbol{\beta})] = E[u_i(\boldsymbol{\beta})u_j(\boldsymbol{\beta})] = -E[U_{ij}(\boldsymbol{\beta})] \equiv \mathfrak{I}_{ij}(\boldsymbol{\beta}) \quad (3)$$

ここに$u_i(\boldsymbol{\beta})$をスコアベクトル、$\mathfrak{I}_{ij}(\boldsymbol{\beta})$を情報行列と呼ぶ。

正準系指数型分布族の最尤法

正準形の指数型分布族の仮定

$$f(y_\lambda \mid \theta_\lambda) = \exp[y_\lambda b(\theta_\lambda) + c(\theta_\lambda) + d(y_\lambda)]$$

$\int f(y_\lambda \mid \theta_\lambda) dx_\lambda = 1$　より、

$$\begin{aligned}\frac{\partial}{\partial \beta_j}\int f(y_\lambda \mid \theta_\lambda) dy_\lambda &= \int \frac{\partial}{\partial \beta_j}\exp[y_\lambda b(\theta_\lambda) + c(\theta_\lambda) + d(y_\lambda)]dy_\lambda \\ &= \int \left[y_\lambda \frac{\partial b(\theta_\lambda)}{\partial \beta_j} + \frac{\partial c(\theta_\lambda)}{\partial \beta_j} \right] f(y_\lambda \mid \theta_\lambda) dy_\lambda \\ &= E[y_\lambda]\frac{\partial b(\theta_\lambda)}{\partial \beta_j} + \frac{\partial c(\theta_\lambda)}{\partial \beta_j} = 0\end{aligned}$$

よって、

$$E[y_\lambda] = -\frac{\partial c(\theta_\lambda)}{\partial \beta_j} \Big/ \frac{\partial b(\theta_\lambda)}{\partial \beta_j} \equiv \mu_\lambda(\theta_\lambda) \quad (4)$$

この$\mu_\lambda(\theta_\lambda)$について以下の仮定をする。

$$\eta_\lambda = g(\mu_\lambda) = {}^t\boldsymbol{\beta}\mathbf{x}_\lambda$$

一方、

$$\frac{\partial}{\partial \beta_i}\log f(y_\lambda \mid \theta_\lambda) = y_\lambda \frac{\partial b(\theta_\lambda)}{\partial \beta_i} + \frac{\partial c(\theta_\lambda)}{\partial \beta_i} = \frac{\partial b(\theta_\lambda)}{\partial \beta_i}(y_\lambda - \mu_\lambda) \quad (5)$$

であるので、対数尤度関数の微分を 0 にすることは、

$$\frac{\partial}{\partial \beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda \mid \theta_\lambda) = \sum_{\lambda=1}^{N}\frac{\partial b(\theta_\lambda)}{\partial \beta_i}(y_\lambda - \mu_\lambda) = 0 \quad (6)$$

(6)の解を$\hat{\boldsymbol{\beta}}$とすると、(2)の定義から、

$$\mathbf{u}(\hat{\boldsymbol{\beta}}) = 0 \quad (7)$$

ここで、多変数の Taylor 展開より、

$$f(\mathbf{x}) \simeq f(\mathbf{x}_0) + \sum_j (x_j - x_{0j})\left[\frac{\partial f(\mathbf{x})}{\partial x_j}\right]_{\mathbf{x}=\mathbf{x}_0} \equiv f(\mathbf{x}_0) + \sum_j (x_j - x_{0j})\frac{\partial f(\mathbf{x}_0)}{\partial x_j}$$

$f(\mathbf{x})$を$\frac{\partial}{\partial \beta_i}\sum_{\lambda=1}^{N}\log f(y_\lambda \mid \theta_\lambda) = u_i(\boldsymbol{\beta})$に置き換えると

$$u_i(\boldsymbol{\beta}) \simeq u_i(\boldsymbol{\beta}_0) + \sum_j (\beta_j - \beta_{0j}) \frac{\partial u_i(\boldsymbol{\beta}_0)}{\partial \beta_j} = u_i(\boldsymbol{\beta}_0) + \sum_j (\beta_j - \beta_{0j}) U_{ij}(\boldsymbol{\beta}_0)$$

これをベクトル表示して以下の関係が得られる。

$$\begin{aligned} \mathbf{u}(\boldsymbol{\beta}) &\simeq \mathbf{u}(\boldsymbol{\beta}_0) + \mathbf{U}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \simeq \mathbf{u}(\boldsymbol{\beta}_0) + E[\mathbf{U}(\boldsymbol{\beta}_0)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ &= \mathbf{u}(\boldsymbol{\beta}_0) - \mathfrak{I}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \end{aligned} \tag{8}$$

ニュートン・ラフソン法では $\mathbf{u}(\boldsymbol{\beta}_1) = \mathbf{0}$ とおいて、新しい位置 $\boldsymbol{\beta}_1$ を得る。

$$\mathbf{0} = \mathbf{u}(\boldsymbol{\beta}_1) \simeq \mathbf{u}(\boldsymbol{\beta}_0) - \mathfrak{I}(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

これを解くと以下となるので、

$$\boldsymbol{\beta}_1 \simeq \boldsymbol{\beta}_0 + \mathfrak{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{u}(\boldsymbol{\beta}_0) \tag{9}$$

上の操作を繰り返すことによって、$\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ を与える推定値 $\hat{\boldsymbol{\beta}}$ に近づいて行く。すなわち、

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + \mathfrak{I}(\boldsymbol{\beta}_n)^{-1} \mathbf{u}(\boldsymbol{\beta}_n) \xrightarrow{n \to \infty} \hat{\boldsymbol{\beta}}$$

上の式を元に、

$$\begin{aligned} &\boldsymbol{\beta}_n \xrightarrow{n \to \infty} \hat{\boldsymbol{\beta}} \\ &Cov(\beta_{ni}, \beta_{nj}) \simeq Cov[(\mathfrak{I}(\boldsymbol{\beta}_n)^{-1} \mathbf{u}(\boldsymbol{\beta}_n))_i, (\mathfrak{I}(\boldsymbol{\beta}_n)^{-1} \mathbf{u}(\boldsymbol{\beta}_n))_j] \\ &\qquad = \sum_{k,l} \mathfrak{I}(\boldsymbol{\beta}_n)^{-1}_{ik} E[u(\boldsymbol{\beta}_n)_k u(\boldsymbol{\beta}_n)_l] \mathfrak{I}(\boldsymbol{\beta}_n)^{-1}_{jl} \\ &\qquad \simeq \sum_{k,l} \mathfrak{I}(\boldsymbol{\beta}_n)^{-1}_{ik} \mathfrak{I}(\boldsymbol{\beta}_n)_{kl} \mathfrak{I}(\boldsymbol{\beta}_n)^{-1}_{jl} = \mathfrak{I}(\boldsymbol{\beta}_n)^{-1}_{ij} \xrightarrow{n \to \infty} \mathfrak{I}(\hat{\boldsymbol{\beta}})^{-1}_{ij} \end{aligned}$$

となり、推定値の分散などが求められる。

参考文献

[1] Annette J. Dobson 著, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.

２１．多値ロジスティック回帰

ここでは事象の出現する確率や出現しない確率を多項項分布に基づくモデルで解析する多項分布モデルと呼ばれる手法を解説する。その中で、最もよく利用されるのが名義及び順序ロジスティックモデルという方法であることから、章のタイトルを多値ロジスティック回帰という名称にした。

21.1　多値ロジスティック分析とは

３つ以上のカテゴリーの各出現確率を、多項分布モデルを使って予測する手法を、2 値ロジスティック分析との類似で多値ロジスティック分析と呼ぶ。ここに多項分布モデルとは、多項分布の出現確率を説明変数の線形結合の関数で推測するモデルである。

今、ケースi（$i=1,\cdots,n$）、試行回数をn_i、カテゴリーα（$\alpha=1,\cdots,J$）の出現回数を$y_{i\alpha}$、出現確率を$p_{i\alpha}$とすると、多項分布の密度関数は以下で与えられる。

$$f(y_i;p_i)=n_i!\prod_{\alpha=1}^{J}\frac{p_{i\alpha}^{y_{i\alpha}}}{y_{i\alpha}!}\qquad \text{ここに、}\ \sum_{\alpha=1}^{J}y_{i\alpha}=n_i,\ \ \sum_{\alpha=1}^{J}p_{i\alpha}=1$$

これより、$y_{i\alpha}$及び$p_{i\alpha}$の中の１つは他の変数で規定される。一般性を失わず、他の変数で規定されるカテゴリーを仮に$\alpha=1$とする。

多項分布のカテゴリーは、その性質によって、名義尺度と順序尺度に分けられる。名義尺度とは純粋な分類で、例えば性別や単純な分類のように、その順番に意味を持たないものをいう。順序尺度は、例えば５段階評価などのように、その分類に順番が付くものをいう。これらの性質により、多項分布モデルの分析法は異なり、分けて考えなければならない。

名義尺度ロジスティックモデルは、基準となるカテゴリーの出現確率に対する他のカテゴリーのオッズ比の対数（ロジット）を説明変数の線形結合で推測する。

$$\eta_{i\alpha}=\log\frac{p_{i\alpha}}{p_{i1}}=\sum_{k=1}^{p-1}\beta_{k\alpha}x_{ik}+\beta_{0\alpha},\quad \alpha=2,\cdots,J$$

適合度指標については、最適値からのずれを表す、逸脱度（小さいほど良い：P>0.05）、ピアソンのχ^2統計量（小さいほど良い：p>0.05）及び、最小モデルからのずれを表す、尤度比χ^2統計量（大きいほど良い：p<0.05）などがある。

ここではn_i回の試行に対して、$y_{i\alpha}$回のカテゴリーが起こったとしているが、１回の試行で起こったか起こらないかにする場合は、$n_i=1,\ y_{i\alpha}=\{0,1\}$とする。

順序尺度ロジスティックモデルには、累積ロジットモデル、隣接カテゴリ―ロジットモデル、連続比ロジットモデル等があるが、ここでは最も扱いやすく、プログラムで取り入れている累積ロジットモデルについて説明する。

累積ロジットモデルでは、以下の比の対数を線形関数で予測する。

$$\eta_{i1}=\log\frac{p_{i1}}{p_{i2}+\cdots+p_{iJ}}=\sum_{k=1}^{p-1}\beta_{k1}x_{ik}+\beta_{01},$$

$$\vdots$$

$$\eta_{ij} = \log \frac{p_{i1} + \cdots + p_{ij}}{p_{ij+1} + \cdots + p_{iJ}} = \sum_{k=1}^{p-1} \beta_{kj} x_{ik} + \beta_{0j}$$

$$\vdots$$

$$\eta_{i(J-1)} = \log \frac{p_{i1} + \cdots + p_{i(J-1)}}{p_{iJ}} = \sum_{k=1}^{p-1} \beta_{k(J-1)} x_{ik} + \beta_{0(J-1)}$$

これは、連続した複数のカテゴリーの出現確率と残りの事象の出現確率の対数オッズ比（ロジット）を説明変数の線形関数で予測することに相当する。

これらより、$q_\alpha = p_1 + p_2 + \cdots + p_\alpha$ と定義すると、各事象α について独立に、q_α と$1-q_\alpha$ の 2 項分布として β_α の値を推定できることが分かる。そのためこれは 2 値ロジスティック回帰の拡張として捉えることができ、各事象 p_α（$\alpha = 1, 2, \cdots, J$）については以下のように与えることができる。

$$p_1 = q_1,\ p_\alpha = q_\alpha - q_{\alpha-1},\ p_J = 1 - q_{J-1}$$

注意すべきこととして、この分析を実行するに当たり、ある変数 k について、分類が完全に x_{ik} の大きさで分かれるとき、累積ロジットモデルの計算は不可能である。

C.Analysis の中では触れていないが、順序尺度の場合、累積ロジットモデル以外に以下のようなモデルも考えられている。

比例オッズモデル

$$\eta_{ij} = \log \frac{p_{i1} + \cdots + p_{ij}}{p_{ij+1} + \cdots + p_{iJ}} = \sum_{k=1}^{p-1} \beta_k x_{ik} + \beta_{0j} \qquad （\beta_{0j} 以外は共通）$$

隣接カテゴリロジットモデル

$$\eta_{ij} = \log \frac{p_{ij+1}}{p_{ij}} = \sum_{k=1}^{p-1} \beta_{kj} x_{ik} + \beta_{0j}$$

連続比ロジットモデル（どちらか）

$$\eta_{ij} = \log \frac{p_{ij+1} + \cdots + p_{iJ}}{p_{ij}} = \sum_{k=1}^{p-1} \beta_{kj} x_{ik} + \beta_{0j}$$

$$\eta_{ij} = \log \frac{p_{ij}}{p_{i1} + \cdots + p_{ij-1}} = \sum_{k=1}^{p-1} \beta_{kj} x_{ik} + \beta_{0j}$$

21.2　プログラムの利用法

メニュー［分析－多変量解析等－判別手法－多値ロジスティック回帰］を選択すると図 1 のような多値ロジスティック回帰分析の分析実行画面が表示される。

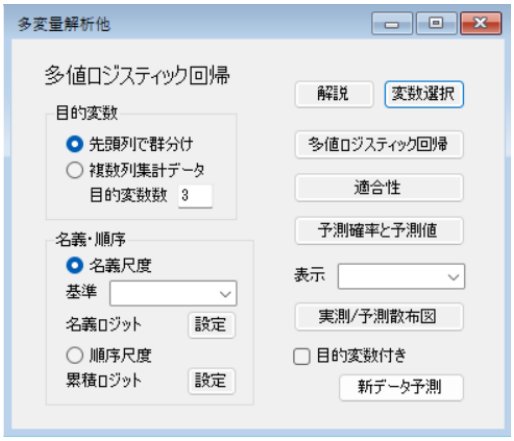


図１　分析実行画面

「複数列集計データ」の例を図２に示す。

図２　複数列集計データ

「目的変数」グループボックスの「複数列集計データ」を選択し、変数選択ですべての変数を選択し、「名義ロジスティック」の設定から図３のように基準に「重要でない」を選択する。

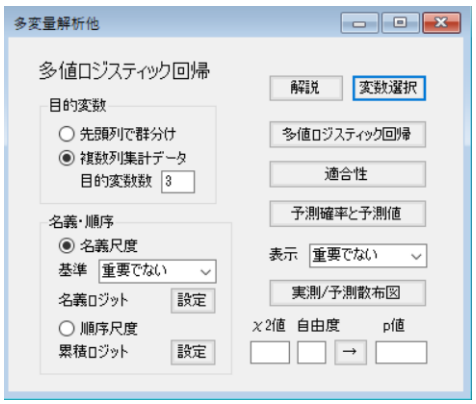


図３　複数列集計データの名義ロジスティック設定

ここでは、「重要でない」カテゴリーの確率で、他のカテゴリーの確率を割った対数オッズを説明変数の線形関数で推定することになる。

「多値ロジスティック回帰」ボタンをクリックすると図４のような分析結果が表示される。

log(確率/重要でない確率)の線形予測

	偏回帰係数	標準化値	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 重要/重要でない							
性別	-0.3881	-1.5355	0.2528	0.1247	-0.8835	0.1073	0.6783
年齢	1.1283	3.6819	0.3064	0.0002	0.5277	1.7289	3.0903
職業	1.5877	5.1083	0.3108	0.0000	0.9785	2.1969	4.8925
切片	-0.5908	-2.2770	0.2595	0.0228	-1.0993	-0.0823	
とても重要/重要でない							
性別	-0.8130	-2.9999	0.2710	0.0027	-1.3442	-0.2818	0.4435
年齢	1.4781	4.0607	0.3640	0.0000	0.7647	2.1916	4.3846
職業	2.9167	8.4227	0.3463	0.0000	2.2380	3.5955	18.4805
切片	-1.0391	-3.3824	0.3072	0.0007	-1.6412	-0.4370	

図４　対数オッズの推定

ここでは、オッズ比推定の偏回帰係数、標準化偏回帰係数、偏回帰係数の標準誤差、偏回帰係数が０となる検定確率、偏回帰係数の下限と上限、説明変数単位量の変化によるオッズ比の変化量が表示される。

「適合性」ボタンをクリックすると、図５のように各種の適合性指標が表示される。

適合性

▶ 対数尤度値	-290.351				
逸脱度D	3.939	自由度	4	上側確率	0.4144
ピアソンχ2	3.927	自由度	4	上側確率	0.4160
C尤度比	77.842	自由度	6	上側確率	0.0000
擬似R2	0.118				
実測予測R2	0.981				

図５　適合性指標

「予測確率と予測値」ボタンをクリックすると、図６のような結果が表示される。

予測確率と予測値

	重要でない	予測確率	予測値	重要	予測確率	予測値	とても重要	予測確率	予測値
▶	26	0.524	23.589	12	0.290	13.066	7	0.185	8.345
	9	0.235	10.556	21	0.402	18.069	15	0.364	16.375
	5	0.098	5.855	14	0.264	15.865	41	0.638	38.280
	40	0.652	42.411	17	0.245	15.934	8	0.102	6.655
	17	0.351	15.444	15	0.408	17.931	12	0.241	10.625
	8	0.174	7.145	15	0.320	13.134	18	0.505	20.721

図６　予測確率と予測値

これには３つのカテゴリーについての実測値、予測確率、予測値が表示される。「表示変数」を１つ選んで、「実測/予測散布図」ボタンをクリックすると、図７のような散布図が表示される。

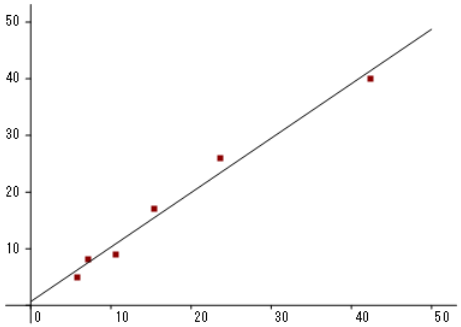


図７　実測/予測散布図

同じデータを順序尺度として､順序ロジスティックの累積ロジットモデルで分析すると図8のような結果を得る。

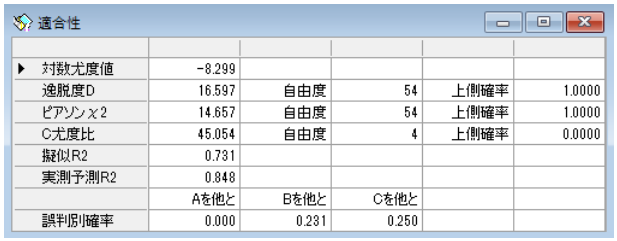
log(累積確率/他累積確率)の線形予測

	偏回帰係数	標準化値	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 重要でない/重要～							
性別	0.5723	2.1135	0.2708	0.0346	0.0416	1.1031	1.7724
年齢	-1.2597	-4.1112	0.3064	0.0000	-1.8602	-0.6591	0.2838
職業	-2.2490	-6.2871	0.3577	0.0000	-2.9501	-1.5479	0.1055
切片	0.0746	0.2993	0.2492	0.7647	-0.4139	0.5631	
～重要/とても重要							
性別	0.5930	2.1886	0.2710	0.0286	0.0619	1.1241	1.8095
年齢	-0.9711	-2.6696	0.3638	0.0076	-1.6841	-0.2581	0.3787
職業	-2.1137	-6.1048	0.3462	0.0000	-2.7923	-1.4351	0.1208
切片	1.5266	4.9453	0.3087	0.0000	0.9216	2.1317	

図８　累積ロジットモデルでの結果

これは最初が「重要でない」を「重要」と「とても重要」を足したカテゴリーで割った対数オッズ、次が「重要でない」と「重要」を足したカテゴリーを「とても重要」で割った対数オッズについての説明変数の線形関数での推定である。

最後に目的変数が同じファイル２頁目の「先頭列で群分け」（ファイルは異なる）で与えられる場合、「適合性」の結果に図９のように誤判別確率の値が表示される。



適合性

▶ 対数尤度値	-8.299				
逸脱度D	16.597	自由度	54	上側確率	1.0000
ピアソンχ2	14.657	自由度	54	上側確率	1.0000
C尤度比	45.054	自由度	4	上側確率	0.0000
擬似R2	0.731				
実測予測R2	0.848				
	Aを他と	Bを他と	Cを他と		
誤判別確率	0.000	0.231	0.250		

図９　先頭列で群分けの場合の適合性結果

予測について

機械学習などでは､訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の右下に「新データ予測」ボタンがあるが、これは一度多値ロジスティック分析を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は｢目的変数付き｣チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を正解率の値で示してくれる。これは「先頭列で群分け」のとき有効である。

問題

多値ロジスティック回帰 1.txt は性別（男性 0, 女性 1)、年齢区分による意見の集約データである。多値ロジスティック回帰を用いて以下の問いに答えよ。

<u>基準カテゴリーを「重要でない」にした名義ロジスティックについて</u>

１）$\log(q_k/q_{\text{重要でない}})$を説明変数の線形関数で予測した場合の重回帰式を求めよ。

重要の対数オッズ＝［　　］男性＋［　　］年齢２

＋［　　］年齢３＋［　　］

とても重要の対数オッズ＝［　　］男性＋［　　］年齢２
＋［　　］年齢３＋［　　］

２）上の重要のパラメータは有効（0 でない）といえるか。
男性　の係数　検定確率［　　］　0 と異なると［いえる・いえない］
年齢２の係数　検定確率［　　］　0 と異なると［いえる・いえない］
年齢３の係数　検定確率［　　］　0 と異なると［いえる・いえない］
切片　　検定確率［　　］　0 と異なると［いえる・いえない］

３）男性と女性を比べた場合の（男性/女性）
重要/重要でない　　のオッズ比　［　　］
とても重要/重要でないのオッズ比　［　　］

４）最良モデルからのずれを表す、逸脱度とピアソンのχ^2値はいくらか。
逸脱度　　［　　］　検定確率［　　］
ピアソンのχ^2比　［　　］　検定確率［　　］
最良モデルとの差があると［いえる・いえない］←　差がない方がよい

５）パラメータ数が最小のモデルからのずれを表す、C 尤度比の値はいくらか。
C 尤度比　［　　］　検定確率［　　］
最小のモデルと差があると［いえる・いえない］←　差がある方がよい

６）実測値と予測値の相関係数の 2 乗の値はいくらか。　［　　］

７）5 番目のデータの実測値・予測確率・予測値を求めよ。
重要でない　実測値［　　］予測確率［　　］予測値［　　］
重要　　実測値［　　］予測確率［　　］予測値［　　］
とても重要　実測値［　　］予測確率［　　］予測値［　　］

順序ロジスティックについて

８）$\log(q_{k\text{ 以下}}/q_{k+1\text{ 以上}})$を濃度の線形関数で予測した場合の重回帰式を求めよ。
重要でないの対数オッズ＝［　　］男性＋［　　］年齢２
＋［　　］年齢３＋［　　］
重要以下の対数オッズ＝［　　］男性＋［　　］年齢２
＋［　　］年齢３＋［　　］

９）男性と女性を比べた場合の（男性/女性）
重要でない/その他のオッズ比　［　　］
重要/その他の　　オッズ比　［　　］

10）最良モデルからのずれを表す、逸脱度とピアソンのχ^2値はいくらか。
逸脱度　　［　　］　検定確率［　　］
ピアソンのχ^2比　［　　］　検定確率［　　］
最良モデルとの差があると［いえる・いえない］←　差がない方がよい

11）パラメータ数が最小のモデルからのずれを表す、C 尤度比の値はいくらか。

C 尤度比　［　　］　検定確率［　　］
最小のモデルと差があると［いえる・いえない］←　差がある方がよい

12）実測値と予測値の相関係数の 2 乗の値はいくらか。　［　　］

13）5 番目のデータの実測値・予測確率・予測値を求めよ。
重要でない　実測値［　　］予測確率［　　］予測値［　　］
重要　　実測値［　　］予測確率［　　］予測値［　　］
とても重要　実測値［　　］予測確率［　　］予測値［　　］

問題解答

基準カテゴリーを「重要でない」にした名義ロジスティックについて

１）$\log(q_k/q_{\text{重要でない}})$を説明変数の線形関数で予測した場合の重回帰式を求めよ。
重要の対数オッズ＝［ -0.3881 ］男性＋［ 1.1283 ］年齢 2
＋［ 1.5877 ］年齢 3＋［ -0.5908 ］
とても重要の対数オッズ＝［ -0.8130 ］男性＋［ 1.4781 ］年齢 2
＋［ 2.9167 ］年齢 3＋［ -1.0391 ］

２）上の重要のパラメータは有効（0 でない）といえるか。
男性　の係数　検定確率［ 0.1247 ］　0 と異なると［いえる・いえない］
年齢２の係数　検定確率［ 0.0002 ］　0 と異なると［いえる・いえない］
年齢３の係数　検定確率［ 0.0000 ］　0 と異なると［いえる・いえない］
切片　　検定確率［ 0.0228 ］　0 と異なると［いえる・いえない］

３）男性と女性を比べた場合の（男性/女性）
重要/重要でない　　のオッズ比　［ 0.6783 ］<- EXP(b)の値
とても重要/重要でないのオッズ比　［ 0.4435 ］<- EXP(b)の値

４）最良モデルからのずれを表す、逸脱度とピアソンのχ^2値はいくらか。
逸脱度　　［ 3.939 ］　検定確率［ 0.4144 ］
ピアソンのχ^2比　［ 3.927 ］　検定確率［ 0.4160 ］
最良モデルとの差があると［いえる・いえない］←　差がない方がよい

５）パラメータ数が最小のモデルからのずれを表す、C 尤度比の値はいくらか。
C 尤度比　［ 77.842 ］　検定確率［ 0.0000 ］
最小のモデルと差があると［いえる・いえない］←　差がある方がよい

６）実測値と予測値の相関係数の 2 乗の値はいくらか。　［ 0.981 ］

７）5 番目のデータの実測値・予測確率・予測値を求めよ。
重要でない　実測値［ 17 ］予測確率［ 0.351 ］予測値［ 15.444 ］
重要　　実測値［ 15 ］予測確率［ 0.408 ］予測値［ 17.931 ］
とても重要　実測値［ 12 ］予測確率［ 0.241 ］予測値［ 10.625 ］

順序ロジスティックについて

８）$\log(q_{k\text{ 以下}}/q_{k+1\text{ 以上}})$を濃度の線形関数で予測した場合の重回帰式を求めよ。
重要でないの対数オッズ＝［ 0.5723 ］男性＋［ -1.2597 ］年齢 2
＋［ -2.2490 ］年齢 3＋［ 0.0746 ］
重要以下の対数オッズ＝［ 0.5930 ］男性＋［ -0.9711 ］年齢 2
＋［ -2.1137 ］年齢 3＋［ 1.5266 ］

９）男性と女性を比べた場合の（男性/女性）
重要でない/その他のオッズ比　［ 1.772 ］
重要/その他の　　オッズ比　［ 1.809 ］注）重要性の定義が上と逆

10）最良モデルからのずれを表す、逸脱度とピアソンのχ^2値はいくらか。

逸脱度　　［ 3.858 ］　　検定確率［ 0.4255 ］
ピアソンのχ2比　　［ 3.843 ］　　検定確率［ 0.4276 ］
最良モデルとの差があると［いえる・いえない］←　差がない方がよい

11）パラメータ数が最小のモデルからのずれを表す、C 尤度比の値はいくらか。
C 尤度比　［ 77.922 ］　　検定確率［ 0.0000 ］
最小のモデルと差があると［いえる・いえない］←　差がある方がよい

12）実測値と予測値の相関係数の 2 乗の値はいくらか。　［ 0.981 ］

13）5 番目のデータの実測値・予測確率・予測値を求めよ。
重要でない　実測値［ 17 ］予測確率［ 0.351 ］予測値［ 15.463 ］
重要　　実測値［ 15 ］予測確率［ 0.408 ］予測値［ 17.943 ］
とても重要　実測値［ 12 ］予測確率［ 0.241 ］予測値［ 10.593 ］

21.3　多項分布モデル

多項分布の密度関数、対数尤度関数は以下で与えられる。ここで、対数尤度関数の最後の項はパラメータに依存していないので、計算上は考えないことにする（参考文献[1]の数値に従っている）。

密度関数

$$f(y_i;p_i)=n_i!\prod_{\alpha=1}^{J}\frac{p_{i\alpha}^{y_{i\alpha}}}{y_{i\alpha}!},\quad \sum_{j=1}^{J}y_{ij}=n_i,\quad \sum_{j=1}^{J}p_{ij}=1$$

これより、y_i 及び p_i の中の 1 つは他の変数で規定される。

対数尤度関数

$$l(y_i;p_i)=\log f(y_i;p_i)=\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}+\log n_i!-\sum_{\alpha=1}^{J}\log y_{i\alpha}!\to\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}$$

以下この関係を利用して計算過程を考えてみる。

21.4　名義ロジスティック回帰

一般性を失わず、他から規定されるカテゴリを J $(p_J=1-p_1-\cdots-p_{J-1})$ とすると、

$$\frac{\partial l_i}{\partial p_{i\alpha}}=\frac{\partial}{\partial p_{i\alpha}}\left(\sum_{\beta=1}^{J-1}y_{i\beta}\log p_{i\beta}+y_{iJ}\log p_{iJ}\right)=\frac{y_{i\alpha}}{p_{i\alpha}}-\frac{y_{iJ}}{p_{iJ}}\qquad(\alpha\neq J)$$

$$E[Y_{i\alpha}]=n_i p_{i\alpha}\equiv\mu_{i\alpha}$$

$$Cov[Y_{i\alpha}Y_{i\beta}]=n_i p_{i\alpha}(\delta_{\alpha\beta}-p_{i\beta})$$

名義尺度ロジスティックモデルは、基準となるカテゴリに対する他のカテゴリのロジットを説明変数の線形結合で推測する。

$$\eta_{i\alpha}=\log\frac{p_{i\alpha}}{p_{iJ}}=\sum_{k=1}^{p-1}\beta_{k\alpha}x_{ik}+\beta_{0\alpha},\quad j=1,\cdots,J-1$$

$$p_{i\alpha}=p_{iJ}e^{\eta_{i\alpha}}$$

$$1=\sum_{\alpha=1}^{J}p_{i\alpha}=p_{iJ}+\sum_{\alpha=1}^{J-1}p_{i\alpha}=p_{iJ}\left(1+\sum_{\alpha=1}^{J-1}e^{\eta_{i\alpha}}\right)\quad \text{より、}$$

$$p_{iJ}=\frac{1}{1+\sum_{\beta=1}^{J-1}e^{\eta_{i\beta}}},\quad p_{i\alpha}=\frac{e^{\eta_{i\alpha}}}{1+\sum_{\beta=1}^{J-1}e^{\eta_{i\beta}}}$$

対数尤度関数

$$l(y_i;p_i)=\log f(y_i;p_i)=\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}+\log n_i!-\sum_{\alpha=1}^{J}y_{i\alpha}!\to\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}$$

以下この関係を利用して計算過程を考えてみる。

$$\frac{\partial p_{i\beta}}{\partial\eta_{i\alpha}}=\frac{e^{\eta_{i\beta}}\delta_{\alpha\beta}\left(1+\sum_{\gamma=1}^{J-1}e^{\eta_{i\gamma}}\right)-e^{\eta_{i\alpha}}e^{\eta_{i\beta}}}{\left(1+\sum_{\gamma=1}^{J-1}e^{\eta_{i\gamma}}\right)^2}=p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\qquad(\alpha\neq J,\beta\neq J)$$

$$\frac{\partial p_{iJ}}{\partial\eta_{i\alpha}}=\frac{-e^{\eta_{i\alpha}}}{\left(1+\sum_{\gamma=1}^{J-1}e^{\eta_{i\gamma}}\right)^2}=-p_{i\alpha}p_{iJ}\qquad(\alpha\neq J)$$

より、

$$\frac{\partial p_{i\beta}}{\partial\beta_{j\alpha}}=\frac{\partial\eta_{i\alpha}}{\partial\beta_{j\alpha}}\frac{\partial p_{i\beta}}{\partial\eta_{i\alpha}}=x_{ij}\frac{\partial p_{i\beta}}{\partial\eta_{i\alpha}}=x_{ij}p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\qquad(\alpha\neq J,\beta\neq J)$$

$$\frac{\partial p_{iJ}}{\partial\beta_{j\alpha}}=\frac{\partial\eta_{i\alpha}}{\partial\beta_{j\alpha}}\frac{\partial p_{iJ}}{\partial\eta_{i\alpha}}=x_{ij}\frac{\partial p_{iJ}}{\partial\eta_{i\alpha}}=-x_{ij}p_{i\alpha}p_{iJ}\qquad(\alpha\neq J)$$

以上より、

$$U_j^{\alpha}=\sum_{i=1}^{N}\frac{\partial l_i}{\partial\beta_{j\alpha}}=\sum_{i=1}^{N}\sum_{\beta=1}^{J}\frac{\partial p_{i\beta}}{\partial\beta_{j\alpha}}\frac{\partial l_i}{\partial p_{i\beta}}=\sum_{i=1}^{N}\left(\sum_{\beta=1}^{J-1}\frac{\partial p_{i\beta}}{\partial\beta_{j\alpha}}\frac{\partial l_i}{\partial p_{i\beta}}+\frac{\partial p_{iJ}}{\partial\beta_{j\alpha}}\frac{\partial l_i}{\partial p_{iJ}}\right)$$

$$=\sum_{i=1}^{N}x_{ij}\left(\sum_{\beta=1}^{J-1}p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\frac{y_{i\beta}}{p_{i\beta}}-p_{i\alpha}p_{iJ}\frac{y_{iJ}}{p_{iJ}}\right)$$

$$=\sum_{i=1}^{N}x_{ij}\left(y_{i\alpha}-p_{i\alpha}\sum_{\beta=1}^{J-1}y_{i\beta}-p_{i\alpha}y_{iJ}\right)=\sum_{i=1}^{N}x_{ij}(y_{i\alpha}-n_ip_{i\alpha})$$

$$\mathfrak{I}_{jk}^{\alpha}=-\frac{\partial U_j^{\alpha}}{\partial\beta_{k\alpha}}=-\frac{\partial}{\partial\beta_{k\alpha}}\sum_{i=1}^{N}x_{ij}(y_{i\alpha}-n_ip_{i\alpha})=\sum_{i=1}^{N}x_{ij}n_i\frac{\partial p_{i\alpha}}{\partial\beta_{k\alpha}}$$

$$=\sum_{i=1}^{N}x_{ij}x_{ik}n_ip_{i\alpha}(1-p_{i\alpha})$$

これらのスコアベクトルと情報ベクトルより、$\beta_{j\alpha}$（$\alpha\neq J$）は推定される。

最適値からのずれを表す、逸脱度、ピアソンのχ^2統計量及び、最小モデルからのずれを表す、尤度比χ^2統計量は以下のようになる。

$$D=2\sum_{i=1}^{N}\sum_{\alpha=1}^{J}y_{i\alpha}\log\frac{y_{i\alpha}}{n_i\hat{p}_{i\alpha}}\sim\chi^2((J-1)(N-p))$$

$$\chi^2 = \sum_{i=1}^{N}\sum_{\alpha=1}^{J}\frac{(y_{i\alpha}-\hat{y}_{i\alpha})^2}{n_i\hat{p}_{i\alpha}} \sim \chi^2((J-1)(N-p))$$

$$C = 2\sum_{i=1}^{N}\sum_{\alpha=1}^{J} y_{i\alpha}\log\frac{\hat{y}_{i\alpha}}{n_i\tilde{p}_\alpha} \sim \chi^2((J-1)(p-1)),\quad \tilde{p}_\alpha = \sum_{i=1}^{N} y_{i\alpha}\Big/\sum_{i=1}^{N} n_i$$

ピアソンのχ^2統計量は逸脱度と漸近的に同じ指標であるが、逸脱度と比べてこちらの方が分布によく適合するという意見もある。ここではn_i回の試行に対して、y_i回の事象が起こったとしているが、1回の試行で起こったか起こらないかにする場合は、$n_i=1,\ y_i=\{0,1\}$とする。但し、分布はデータ数が無限大のときの極限であるので、注意が必要である。

21.5　順序ロジスティック回帰

順序ロジスティック回帰には、累積ロジットモデル、隣接カテゴリ―ロジットモデル、連続比ロジットモデルなどがあるが、ここでは最も扱いやすく、プログラムで取り入れている累積ロジットモデルについて説明する。他のモデルについては、プログラムに導入次第報告する。

累積ロジットモデル

累積ロジットモデルでは、以下の比の対数を線形関数で予測する。

$$\frac{p_1}{p_2+\cdots+p_J}=e^{\eta_1},\quad \frac{p_1+p_2}{p_3+\cdots+p_J}=e^{\eta_2},\quad \cdots,\quad \frac{p_1+\cdots+p_{J-1}}{p_J}=e^{\eta_{J-1}}$$

これは、連続した複数のカテゴリーの出現確率と残りのカテゴリーの出現確率のオッズ比を説明変数の線形関数で予測することに相当する。

上の関係を以下のように書き換え、

$$\frac{p_2+\cdots+p_J}{p_1}=e^{-\eta_1},\quad \frac{p_3+\cdots+p_J}{p_1+p_2}=e^{-\eta_2},\quad \cdots,\quad \frac{p_J}{p_1+\cdots+p_{J-1}}=e^{-\eta_{J-1}}$$

$q_\alpha = p_1+p_2+\cdots+p_\alpha$と定義すると、以下の関係が示される。

$1-p_1=p_2+\cdots+p_J=p_1e^{-\eta_1}$　より、　$p_1=\dfrac{e^{\eta_1}}{1+e^{\eta_1}}=q_1$

$p_2=p_1e^{-\eta_1}-(p_1+p_2)e^{-\eta_2}$　より、

$$p_2=\frac{1}{1+e^{-\eta_2}}-\frac{1}{1+e^{-\eta_1}},\quad p_1+p_2=\frac{e^{\eta_2}}{1+e^{\eta_2}}=q_2$$

$p_3=(p_1+p_2)e^{-\eta_2}-(p_1+p_2+p_3)e^{-\eta_3}$　より、

$$p_3=\frac{1}{1+e^{-\eta_3}}-\frac{1}{1+e^{-\eta_2}},\quad p_1+p_2+p_3=\frac{e^{\eta_3}}{1+e^{\eta_3}}=q_3$$

同様にして、

$$p_{J-1}=\frac{1}{1+e^{-\eta_{J-1}}}-\frac{1}{1+e^{-\eta_{J-2}}},\quad p_1+\cdots+p_{J-1}=\frac{e^{\eta_{J-1}}}{1+e^{\eta_{J-1}}}=q_{J-1}$$

$$p_J = 1-(p_1+\cdots+p_{J-1}) = 1-\frac{1}{1+e^{-\eta_{J-1}}} = \frac{1}{1+e^{\eta_{J-1}}} = 1-q_{J-1}$$

これらより、q_α について考えれば、各カテゴリーα について独立に、q_α と$1-q_\alpha$ の 2 項分布としてβ_α の値を推定できることが分かる。そのためこれは 2 値ロジスティック回帰の拡張として捉えることができ、各カテゴリ p_α（$\alpha = 1,2,\cdots,J$）については以下のように与えることができる。

$$p_1 = q_1,\ p_\alpha = q_\alpha - q_{\alpha-1},\ p_J = 1-q_{J-1}$$

比例オッズモデル

参考文献

[1] Annette J. Dobson 著, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.

２２．K 平均法

K 平均法は、非階層的なクラスター分析の代表的な手法の１つで、多数のデータでも高速に分類できる特徴を持っている。データ $x_{i\lambda}$ は λ 番目（$\lambda=1,\cdots,N$）の個体の i 番目（$i=1,\cdots,p$）の変数を表している。K 平均法はこの個体をある決められた K 個のクラスターに分類する。ここではプログラム中で使ったこの手法の手順を示しておく。

データはそのままでも標準化してもよいが､データの大きさや単位が異なる場合は標準化して使用する方がすべての変数を同等に扱える。ここでは標準化したデータも $x_{i\lambda}$ で表すことにする。

K 平均法は以下の方法によってクラスター構成を行う。

① データの中から K 個のデータを乱数によって選び出し、それをクラスターの代表点にして、他のデータを最も近い代表点に配置し、K 個のクラスターを構成する。

② 各クラスターの重心を新たなクラスターの代表点として、クラスターを再配置する。前回のクラスターと新しいクラスターの構成が異なれば再配置をもう一度繰り返し、同じならば終了する。

この方法は簡単で、高速であるが、結果は最初の乱数に依存することが多い。そのため、階層的クラスター分析の Ward 法で用いられる within group error（群内誤差最小）の考え方を取り入れ、その総和 E の最も小さいものを最良の結果としている。さらに、最初の代表点の選び方を効率的にするために、乱数で定めた最初の１つの代表点からの距離の 2 乗に応じた確率で次の代表点を選び、さらに２つの代表点からの近い方の距離の 2 乗に応じた確率で 3 番目の代表点を選ぶ、という方法を繰り返す K means++法を用いて、within group error が大きくなる選択を防いでいる。

メニュー［分析－多変量解析－分類手法－K 平均法］を選択すると図 1 のような分析メニューが表示される。

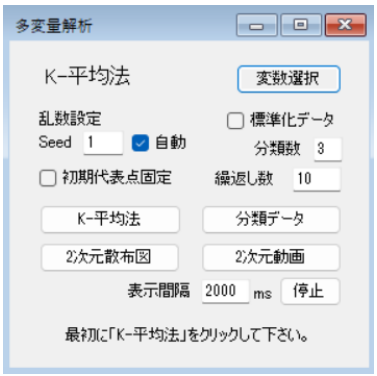


図１　分析メニュー

例として k 平均法 1.txt のデータを用いて、「分類数」を 3 にし、「K 平均法」のボタンをクリックすると結果が図 2 ように表示される。


データ数と代表点

分類名	データ数	英語	数学	国語	理科	社会
1	4	0.198	0.152	-0.222	0.951	-0.185
2	5	-1.259	-1.272	-1.033	-0.920	-0.857
3	6	0.918	0.959	1.009	0.133	0.838
群内誤差合計	20.548					

図２　K 平均法結果表示（k 平均法 1.txt）

これは、各群に割り振られたデータ（個体）数、群内誤差の総和、代表点の値を表示した結果である。これを求める際に得られた群別の個体の割り当て結果については、「分類データ」ボタンをクリックすると図３のように与えられる。


分類(デ...

	分類
亀本	2
藤田	1
三井	3
松井	1
村社	3
田中	3
佐藤	3
増川	2

図３　分類結果

この結果は、元のグリッドエディットのデータに貼り付けて、分類結果をさらに調べて行く際に利用される。

変数を２つ選んだ場合、この結果を散布図として表すことができる。データファイルを k 平均法 2 に変えて２つの変数を選択し、一度「K 平均法」ボタンをクリックして「2 次元散布図」ボタンをクリックすると図４のような散布図が表示される。

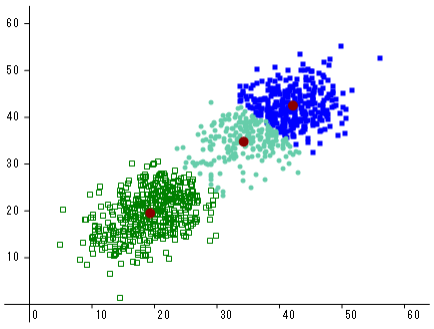


図４　２次元散布図（k 平均法 2.txt）

この分類は最初の「K 平均法」で得られた結果と同じである。続けて「2 次元動画」ボタンをクリックすると、図４の分類になるまでの過程を動画として見ることができる。

最後に、ここで得られた結果は乱数の値によるもので再現性がない。そのため、図 1 のメニューには乱数の「Seed」を固定する方法と、「初期代表点固定」チェックボックスによって同じ結果が繰り返される方法とが提供されている。