福山平成大学経営学部紀要
第 14 号（2018），45－64 頁

社会システム分析のための統合化プログラム３２
－多重共線性・ブートストラップ他－

福井 正康

福山平成大学経営学部経営学科

要旨：我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた。今回は、多重共線性を回避する手法についてのプログラムを作成し、非線形最小２乗法のプログラムを改訂した。また、ブートストラップ、代表的な多重比較検定、飛び離れたデータの棄却検定、相関と回帰の比較検定のプログラムを追加した。

キーワード：College Analysis、多重共線性、非線形最小２乗法、ブートストラップ、多重比較、棄却検定

URL：http://www.heisei-u.ac.jp/ba/fukui/

1. はじめに

我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた[1]。今回は全く新しいプログラムを作成するのではなく、これまでのプログラムの拡張や機能追加を中心に話をする。

我々はまず、これまでに作成した品質管理の異常検知プログラムの中から[2]、リッジ回帰分析、PLS 回帰分析を抜き出し、新たに主成分回帰分析を追加し、多重共線性の問題について考えるプログラムを作成した。

重回帰分析では、説明変数間に線形に近い関係がある場合に、多重共線性の問題が発生する可能性があり、重回帰式による予測が不安定となる。これに対して改善方法と考えられている代表的な手法がリッジ回帰分析、PLS 回帰分析、主成分回帰分析である。リッジ回帰分析は、多重共線性の元となる分散共分散行列に手を加える手法であり、PLS 回帰分析と主成分回帰分析は多重共線性を与える変数間の自由度を制約する手法である。我々のプログラムは 4 者を比較するように作成しており、その違いを理解し易くなっている。

次に、我々はこれまでに作成した非線形最小２乗法のプログラムを、初心者が利用しやすい形に変更した。その際、解を求める計算の初期値設定を、MCMC を用いる方法に変更した。これにより、これまでの多数回の初期値設定による試行過程はなくなり、実行メニューの見た目とプログラムの操作が大幅に簡易化された。しかし、初期値設定を手動で導入できた以前の形式にも良い点があるので、以前のメニューも残し、呼び出せるようにした。

多重比較の問題では、これまで実験計画法の中に Fisher の LSD 法だけしかなかったが、今回新しく、Bonferoni の方法、Turkey の方法、Scheffe の方法を導入した。

ブートストラップは解析的にパラメータの区間推定範囲が求められない場合に、モンテカルロ法によってそれを推定する手法である。これを導入するに当たり、我々は母平均の推定問題とパス解析の間接効果、擬似相関の区間推定問題に適用してみた。この結果を紹介する。

飛び離れたデータがある場合、それを分析に取り入れるかどうは問題である。我々は量的データの集計の中に Grubbs-Smirnov 棄却検定を組み込み、これを検定できるようにした。

また、相関と回帰分析の中に相関係数と回帰式の比較検定を追加した。

この報告には図が多いが、章を越えて示すことはないので、図の番号は章内で付ける。

2. 多重共線性

重回帰分析の多重共線性の問題やリッジ回帰分析、PLS 回帰分析については、すでに参考文献[2]の中で詳しく解説しているので、細かな定義はそれを参考にしてもらいたい。

2.1 重回帰分析と多重共線性を回避する手法

2.1.1 重回帰分析

重回帰分析の目的変数を y_λ （$\lambda = 1, 2, \cdots, N$）、説明変数を $x_{i\lambda}$ （$i = 1, 2, \cdots, p$）、誤差項を ε_λ として、それらの関係を以下とする。

$$y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0 + \varepsilon_\lambda$$

最小 2 乗法としての重回帰分析では、以下の値 D が最小になるように、パラメータ b_i, b_0 を決定する。

$$D = \sum_{\lambda=1}^{N} \left(y_\lambda - \sum_{i=1}^{p} b_i x_{i\lambda} - b_0 \right)^2 = {}^t(\mathbf{y} - \mathbf{Xb})(\mathbf{y} - \mathbf{Xb})$$

ここに、

$$(\mathbf{X})_{\lambda i} = \tilde{x}_{i\lambda} = x_{i\lambda} - \overline{x}_i \ , \ (\mathbf{y})_\lambda = \tilde{y}_\lambda = y_\lambda - \overline{y} \ , \ \mathbf{b} = {}^t(b_1, b_2, \cdots, b_p)$$

である。パラメータは以下で与えられる。

$$\mathbf{b} = ({}^t\mathbf{XX})^{-1}\,{}^t\mathbf{Xy} \ , \ b_0 = \overline{y} - \sum_{i=1}^{p} b_i \overline{x}_i$$

問題となる多重共線性は、行列 ${}^t\mathbf{XX}$ の非正則性から生じる。

多重共線性の判定については、i 番目の説明変数を、他の説明変数で予測して重相関係数 r_i を求め、以下の式で定義される VIF 指標を利用することが多い。

$$VIF_i = 1 / (1 - r_i^2)$$

一般にVIF指標が10以上であれば多重共線性の疑いがあるとみなされる。この式によるとVIFの値が10程度というのは、重相関係数が約0.95ということになる。

$$VIF_i \geq 10 \Leftrightarrow r_i \geq 0.95$$

これより、変数間の相関を調べて、どこかに0.9以上の値があれば問題とすることは1つの簡易的な方法と考えられる。但し、単純な2つの変数間の相関だけでなく、3つ以上の変数間に関係がある場合も考えられるので、単純に相関だけでは多重共線性は見抜けない。VIFがより重要な指標であると思われる。

2.1.2　リッジ回帰分析

リッジ回帰分析は重回帰分析の多重共線性の問題に対して、以下のように置くことによって正則性を確保しようとする手法である[2],[3]。

$$\mathbf{b}' = ({}^t\mathbf{XX} + \eta\mathbf{I})^{-1}\,{}^t\mathbf{Xy}$$

これは、以下を最小化する解でもある。

$$D' = {}^t(\mathbf{y} - \mathbf{Xb}')(\mathbf{y} - \mathbf{Xb}') + \eta\,{}^t\mathbf{b}'\mathbf{b}'$$

ここでパラメータηの値は、1個抜き交差検証より、平均2乗誤差が最小になるように選ばれる。多重共線性がある場合、重回帰分析の予測は、そのデータに対してだけは良い精度を与えるが、他の新しいデータを用いた場合、予測の精度が著しく低下する。そのため1個抜き交差検証が必要である。

2.1.3　ＰＬＳ回帰分析

PLS回帰分析ではまず、変数の線形結合を考える[2],[3]。

$$r_{i\lambda} = \sum_{j=1}^{p} u_{ij}\tilde{x}_{j\lambda} \qquad (i = 1, 2, \cdots, r\,;\, r < p)$$

この式を行列記号を用いて書くと以下となる。

$$\mathbf{R} = \mathbf{XU} \qquad \mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_r)$$

ここで、行列$\mathbf{U}$の各列ベクトルは直交し、順番に$\mathbf{Xu}_i$と$\mathbf{y}$との内積が最大化されるように選ばれる。

この新しい変数を用いて、目的変数を以下のように予測する。

$$\mathbf{y} = \mathbf{R\beta} + \boldsymbol{\varepsilon} = \mathbf{XU\beta} + \boldsymbol{\varepsilon}$$

最小2乗法を使い、以下の量を最小化するようにパラメータを決定する。

$$D'' = {}^t(\mathbf{y} - \mathbf{R\beta})(\mathbf{y} - \mathbf{R\beta})$$

その解は次のように与えられる。

$$\boldsymbol{\beta} = ({}^t\mathbf{RR})^{-1}\,{}^t\mathbf{Ry} = ({}^t\mathbf{U}\,{}^t\mathbf{XXU})^{-1}\,{}^t\mathbf{U}\,{}^t\mathbf{Xy}$$

これから、標準化偏回帰係数$\tilde{\mathbf{b}}$は以下となり、回帰係数も求められる。

$$\tilde{\mathbf{b}} = \mathbf{U}\boldsymbol{\beta}$$

多重共線性の改善の程度については、変数を $\mathbf{U}$ 行列で変換した後の i 番目の説明変数を、他の説明変数で予測して重相関係数 r_i を求め、以下の式で定義される VIF 指標を利用している。

$$VIF_i = 1/(1-r_i^2)$$

2.1.4 主成分回帰分析

主成分回帰分析ではまず、主成分分析によって、変数の線形結合を考える。

$$r_{i\lambda} = \sum_{j=1}^{p} u_{ij}\tilde{x}_{j\lambda} \qquad (i = 1,2,\cdots,r\,;\, r < p)$$

ここで $r_{i\lambda}$ は主成分得点である。この式について行列記号を用いて書くと以下となる。

$$\mathbf{R} = \mathbf{XU} \qquad \mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_r)$$

行列 $\mathbf{U}$ の各列ベクトルは、相関行列 $\tilde{\mathbf{R}}$ で与えられる以下の固有方程式から得られる正規化された固有ベクトルである。

$$\tilde{\mathbf{R}}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$$

どこまでの次元数を求めればよいかは、１つの方法として１個抜き交差検証法の重相関係数の大きさを元にして決めればよい。我々のプログラムではこの方法を用いている。

この新しい変数を用いて、目的変数を以下のように予測する。

$$\tilde{y}_\lambda = \sum_{j=1}^{r} \beta_j r_{j\lambda} + \varepsilon_\lambda$$

即ち、

$$\mathbf{y} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{XU}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

最小２乗法を使い、以下の量を最小化するようにパラメータを決定する。

$$D'' = {}^t(\mathbf{y} - \mathbf{R}\boldsymbol{\beta})(\mathbf{y} - \mathbf{R}\boldsymbol{\beta})$$

その解は次のように与えられる。

$$\boldsymbol{\beta} = ({}^t\mathbf{RR})^{-1}\,{}^t\mathbf{Ry} = ({}^t\mathbf{U}\,{}^t\mathbf{XXU})^{-1}\,{}^t\mathbf{U}\,{}^t\mathbf{Xy}$$

これから、標準化偏回帰係数 $\tilde{\mathbf{b}}$ は以下となる。

$$\tilde{\mathbf{b}} = \mathbf{U}\boldsymbol{\beta}$$

また、回帰係数は以下で与えられる。

$$b_i'' = \tilde{b}_i s_y / s_i\ , \quad b_0'' = \overline{y} - \sum_{i=1}^{p} b_i''\overline{x}_i$$

多重共線性の改善の程度については、変数を $\mathbf{U}$ 行列で変換した後の i 番目の説明変数を、他の説明変数で予測して重相関係数 r_i を求め、以下の式で定義される VIF 指標を利用するが、

$$VIF_i = 1/(1-r_i^2)$$

主成分分析では、主成分得点間の相関が 0 のために、この値は常に 1 になり、多重共線性の判定ができない。

2.2　プログラムの利用法

重回帰分析などの多重共線性の目安として、説明変数の相関係数が 0.9 とか、VIF の値が 10 以上ということが言われている。我々はこの多重共線性を回避すると考えられているリッジ回帰分析、PLS 回帰分析[2]、主成分回帰分析についてプログラムを作成した。ここではプログラムを実行しながら、多重共線性の問題点と、それをこれらの分析手法がどのように解決するのかを見て行く。

メニュー［分析－多変量解析等－予測手法－リッジ回帰分析他］を選択すると、図 1 のような分析実行メニューが表示される。

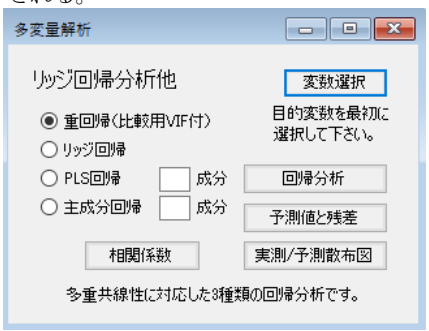


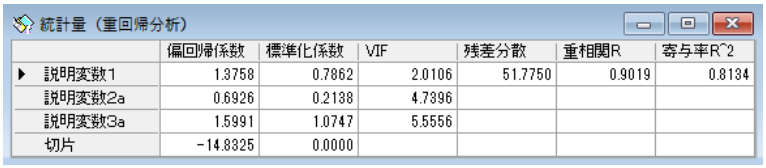
図 1　分析実行メニュー

ここでは図 2 のような形式のデータを用いて多重共線性と各種の分析の結果を見て行く。

図 2　多重共線性について調べるデータ

データは、目的変数と説明変数１を共通として、他の変数は、説明変数２a と説明変数３a、説明変数２b と説明変数３b、説明変数２c と説明変数３c という順番に選ぶ。説明変数２と説明変数３については、a, b, c となるに連れて、相関が大きくなる。c については、最初のレコードがほんの少し違っているだけで後は同じである。

「重回帰分析」ラジオボタンを選択し、説明変数を１,２a,３a、１,２b,３b、１,２c,３c と順番に選んで、「回帰分析」ボタンをクリックする。結果を図 3～図 5 に示す。


統計量（重回帰分析）

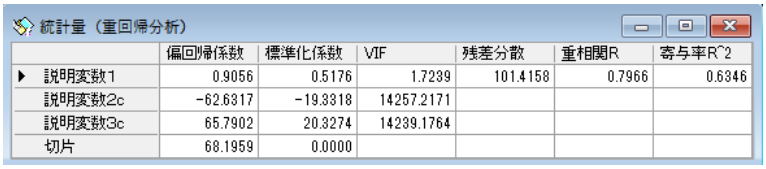
	偏回帰係数	標準化係数	VIF	残差分散	重相関R	寄与率R^2
▶ 説明変数1	1.3758	0.7862	2.0106	51.7750	0.9019	0.8134
説明変数2a	0.6926	0.2138	4.7396			
説明変数3a	1.5991	1.0747	5.5556			
切片	-14.8325	0.0000				

図３　ほぼ問題のない結果


統計量（重回帰分析）

	偏回帰係数	標準化係数	VIF	残差分散	重相関R	寄与率R^2
▶ 説明変数1	0.9422	0.5385	1.7112	86.9744	0.8286	0.6866
説明変数2b	6.1250	1.8905	11.3370			
説明変数3b	-1.4149	-0.9270	10.6017			
切片	67.8303	0.0000				

図４　問題のある結果


統計量（重回帰分析）

	偏回帰係数	標準化係数	VIF	残差分散	重相関R	寄与率R^2
▶ 説明変数1	0.9056	0.5176	1.7239	101.4158	0.7966	0.6346
説明変数2c	-62.6317	-19.3318	14257.2171			
説明変数3c	65.7902	20.3274	14239.1764			
切片	68.1959	0.0000				

図５　完全に問題のある結果

図５をみると、寄与率はそれほど低くなっていないが、偏回帰係数の値が大きくなって正と負で相殺している。これは、利用したデータでは予測値がほぼ良い値となるが、新しいデータで少し値が異なると予測が大きくずれる可能性があることを意味している。これが多重共線性の問題点である。

この図５のデータについて、リッジ回帰分析、PLS 回帰分析、主成分回帰分析がどのような結果を出すか見てみよう。但し、図１で PLS 回帰と主成分回帰のラジオボタンの右にある成分数のテキストボックスには２を入力した。

リッジ回帰分析の結果を図６、PLS 回帰分析の結果を図７、主成分回帰分析の結果を図８に示す。いずれも偏回帰係数の値は適正な値であり、交差検証の結果も良好である。


統計量（リッジ回帰分析）

	偏回帰係数	標準化係数	残差分散	重相関R	寄与率R^2	交差検証R	最良η
▶ 説明変数1	0.8760	0.5006	116.367	0.778	0.606	0.747	24.700
説明変数2c	1.5081	0.4655					
説明変数3c	1.6077	0.4967					
切片	73.8804	0.0000					

図６　リッジ回帰分析結果

統計量（PLS回帰分析）

	偏回帰係数	標準化係数	r-VIF	残差分散	重相関R	寄与率R^2	交差検証R	自由度
▶ 説明変数1	0.9397	0.5370	1.503	109.073	0.779	0.607	0.746	2
説明変数2c	1.6260	0.5019	1.503					
説明変数3c	1.6386	0.5063						
切片	60.2815	0.0000						

図７　PLS 回帰分析結果

統計量（主成分回帰分析）

	偏回帰係数	標準化係数	r-VIF	残差分散	重相関R	寄与率R^2	交差検証R	自由度
▶ 説明変数1	0.9397	0.5370	1.000	109.075	0.779	0.607	0.746	2
説明変数2c	1.6317	0.5036	1.000					
説明変数3c	1.6328	0.5045						
切片	60.2825	0.0000						

図８　主成分回帰分析結果

3.　非線形最小２乗法と非線形回帰分析

College Analysis には非線形最小２乗法のプログラムが含まれている。その分析実行画面を図１に示す。しかし、この分析実行画面は表示が多く複雑で、初心者が見た場合、分かりにくく感じるのではないかと思われる。解を求める際の初期値の設定は、乱数による設定と指定による設定があり、柔軟ではあるが、見た目に煩雑な印象を受ける。また、その他の設定も学習用には良いが初心者には理解できないであろう。

我々はこの欠点を解決するために新しくプログラムを作り直した。その分析実行画面を図２に示す。

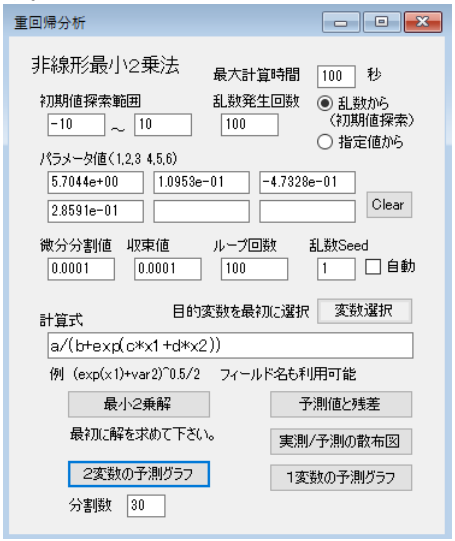


図１　旧分析実行画面

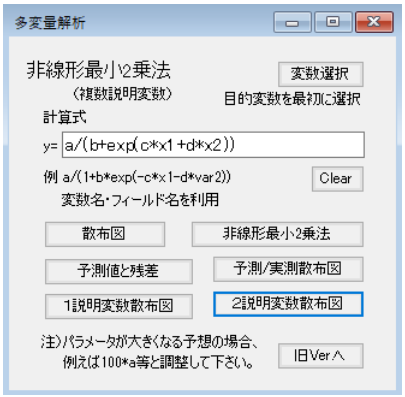


図２　新分析実行画面

図１に比べるとボタンの数はほぼ同じであるが、設定項目が殆どなくなっている。プログラムのアルゴリズムとしては、これまで乱数で初期設定をし、その中から最適値を選んでいたものが、新しいものではMCMCで初期値を設定するようになっている。そのため、解を求める操作は１回だけで、計算時間は短縮される。初期値の手動による指定については、非常に残念であったが、新しい画面では省略した。しかし大きなパラメータ値に対しては対応しきれていないので、旧バージョンの利用もできるようにしている。その他の設定については、必要なくなったものや故意に省いたものもある。必要があれば旧バージョンのものを使うこともできる。

追加した機能としては、説明変数が２つまでの場合は、最初にデータの散布図を見ることができる点である。例えば２つの説明変数の場合には図３のような散布図が表示される。

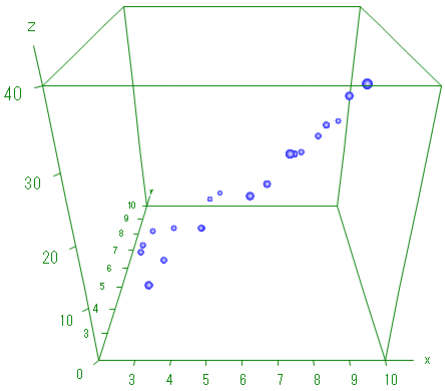


図３　３次元の散布図

上の変更とは別に、非線形最小２乗法の１変数の場合を特別に、非線形回帰分析として新しい分析とした。メニュー［分析－基本統計－非線形回帰分析］を選択すると図４のような分析実行画面が表示される。

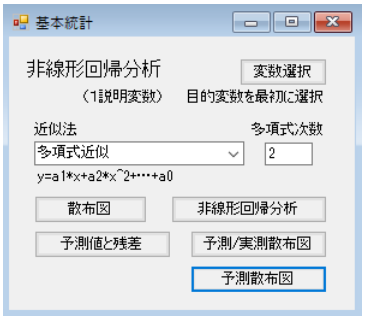


図４　非線形回帰分析実行画面

使い方は図２の非線形最小２乗法と同様であるが、この場合は、近似する関数として予めよく知られた関数が用意されている。例えば、多項式期近似で、「多項式次数」を２とした例を図５と図６に示す。

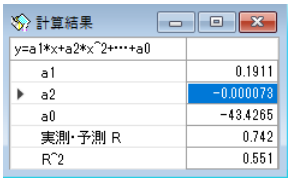


図５　非線形回帰分析実行結果

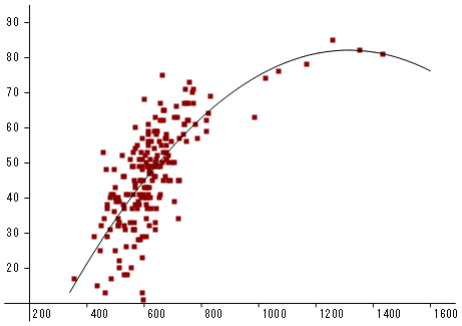


図６　予測散布図

ここで図５の反転の部分は、表示後桁数の再設定を行っている。

4.　ブートストラップの導入

ブートストラップとは、解析的には区間推定値の出せないもの、例えば、正規分布しない場合の平均値、中央値、複数の推定パラメータの関数などの区間推定を行う方法である。

N 個の測定データ（の組）から、重複を許してランダムに N 個のデータ（の組）を抜き出す操作を M 回行う。その取り出したデータを元に、分析を M 回行う。その分析による目的とする統計量の M 個のデータを昇順に並べ、両端から $\alpha/2\times100\%$ のデータを取って、$(1-\alpha)\times100\%$ の信頼区間とする。これがブートストラップの計算方法である。以下で応用問題を考えてみよう。

4.1　区間推定への応用

母平均と母分散の $(1-\alpha)\times100\%$ 信頼区間について、データに正規性が認められる場合には、以下のように与えられる。

母平均の推定（標本数 n，標本平均 $\bar{x}$，不偏分散 u^2）

$$\bar{x}-\frac{u}{\sqrt{n}}t_{n-1}(\alpha/2)\le\mu\le\bar{x}+\frac{u}{\sqrt{n}}t_{n-1}(\alpha/2)$$

母分散の推定（標本数 n，不偏分散 u^2）

$$\frac{(n-1)u^2}{\chi^2_{n-1}(\alpha/2)}\le\sigma^2\le\frac{(n-1)u^2}{\chi^2_{n-1}(1-\alpha/2)}$$

ここに、$t_{n-1}(\alpha/2)$は自由度$n-1$のt分布の上側確率$\alpha/2$の統計値、$\chi^2_{n-1}(\alpha/2)$は自由度$n-1$のχ^2分布の上側確率$\alpha/2$の統計値である。

しかし、データの分布が正規分布から外れた場合、上記の式は使えない。そのため、ブートストラップは有効な手段となる。メニュー［分析－基本統計－区間推定－平均と分散の推定］を選択すると、母平均と母分散の区間推定にブートストラップを加えた分析実行画面が図１のように表示される。

図１　区間推定実行画面

ブートストラップによる推定を実行する場合は、「ブートストラップ」チェックボックスにチェックを入れる。図２に正規母集団での通常の推定とブートストラップによる推定の結果の比較を示す。結果は非常に似通った範囲が推定されている。

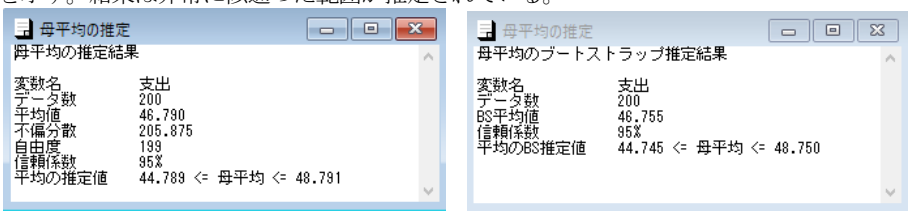


図２　正規母集団における母平均の推定とブートストラップによる推定

4.2　パス解析への応用

パス解析は観測変数間に線形の関係を仮定し、因果関係の方向性を議論するために利用される手法で、共分散構造分析の観測変数だけが現れる特別な場合に相当する[4]。結果としては、直接効果の他に、間接効果や擬似相関といった指標も表示される。間接効果や擬似相関といった指標は、共分散構造分析ではあまり表示されることはなく、信頼区間なども簡単には分からない。しかし、パス解析では使う計算が重回帰分析の行列計算だけであるので、計算時間が短く、ブートストラップに必要な繰り返しを短時間で実行できる。そのため、我々はパス解析にブートストラップを導入することにした。図３にメニュー［分析－多変量解析他－共分散構造分析－パス解析］を選択した場合のブートストラップを加えたパス解析の実行画面を示す。

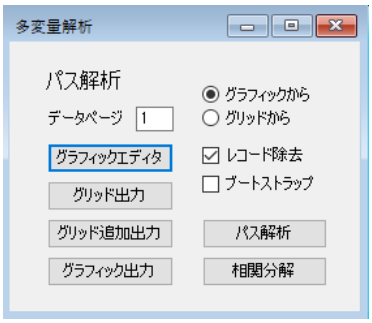


図３　パス解析実行画面

我々のプログラムで図４のようなパス図に通常のパス解析を行った場合の結果とブートストラップを行った結果をそれぞれ図５と図６に示す。

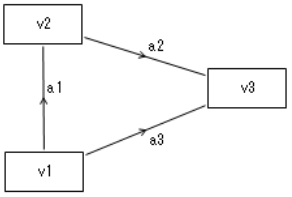
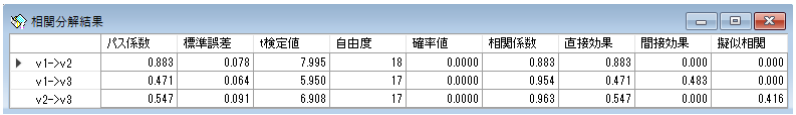


図４　パス図



	パス係数	標準誤差	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	0.078	7.995	18	0.0000	0.883	0.883	0.000	0.000
v1->v3	0.471	0.064	5.950	17	0.0000	0.954	0.471	0.483	0.000
v2->v3	0.547	0.091	6.908	17	0.0000	0.963	0.547	0.000	0.416

図５　一般的なパス解析結果

	パス係数	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	直2.5%下限	直
v1->v2	0.883	0.0000	0.883	0.874	0.000	0.000	0.000	1.000	1.000	0.702	
v1->v3	0.471	0.0000	0.954	0.476	0.479	0.000	0.002	0.000	1.000	0.221	
v2->v3	0.547	0.0000	0.963	0.544	0.000	0.413	0.000	1.000	0.002	0.318	

図６　ブートストラップによる推定結果

ブートストラップの場合、パラメータの検定確率や範囲が出力されるので横に長くなっており、図は途中で切っている。

5.　多重比較

多重比較の手法は様々な文献やホームページ上で数多く紹介されており、どの手法が良いか意見も分かれるところである。これまで我々のプログラムには Fisher の LSD 法だけが用意されていたが、LSD 法は４分類以上では使用してはいけない等の意見もあり、他の手法も用意しておく必要性が生じた。そこで、これまでの LSD 法の他に、図１右下の四角で囲まれたボタン

ように、比較的よく利用される Bonferroni の方法、Turkey-Kramer の方法、Sheffe の方法の 3 つの手法を追加しておいた。

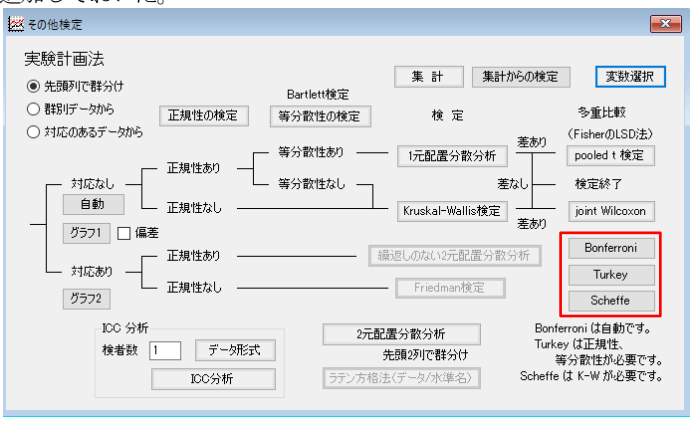


図 1　実験計画法での多重比較

これらの手法の中で、Bonferroni の方法は、データの正規性や等分散性から pooled t 検定または joint Wilcoxon 検定を自動選択して、それを元に実行する。この検定確率は通常の検定確率に比較回数をかけて出力されている。Turkey-Kramer の方法は正規性・等分散性がある場合の方法である。また、Sheffe の方法は非正規分布の場合にも適用可能であるが、LSD 法と同じく、最初に Kruskal-Wallis 検定を必要とする。

それぞれの結果を図 2 から図 4 に示す。Turkey-Kramer の方法では確率の値が数値では与えられていない。

pooled Bonferroniの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率*回数(両側)			
1-不良品率	1.0000	0.0058	0.0297
2-不良品率	0.0058	1.0000	1.0000
3-不良品率	0.0297	1.0000	1.0000

図 2　Bonferroni の方法による実行結果

Tukey-Kramerの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率範囲(両側)			
1-不良品率	1.0000	p<0.01	p<0.01
2-不良品率	p<0.01	1.0000	n.s.
3-不良品率	p<0.01	n.s.	1.0000

図 3　Turkey-Kramer の方法による実行結果


Scheffeの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率(片側)			
1-不良品率	1.0000	0.0070	0.0323
2-不良品率	0.0070	1.0000	0.7301
3-不良品率	0.0323	0.7301	1.0000

図４　Sheffe の方法による実行結果

次に質的データにおける多重比較について追加した残差分析を紹介する。残差分析はχ^2検定後に行う多重比較の一種である。ここでは、標準的な Haberman の残差分析を用いている。これはセルi, j に対する基準化残差e_{ij} の以下の性質を利用している。

$$e_{ij} = \frac{n_{ij} - n_{i.}n_{.j}/n}{\sqrt{(n_{i.}n_{.j}/n)(1-n_{i.}/n)(1-n_{.j}/n)}} \sim N(0,1)$$

残差分析を図５に示されるような分割表のデータで行ってみよう。

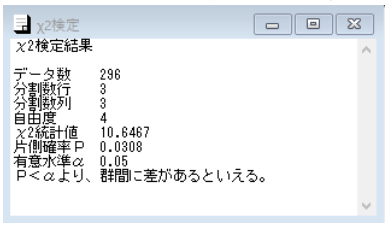

データ編集 残差分析.txt

	賛成	中立	反対
年齢群1	24	32	43
年齢群2	35	30	39
年齢群3	42	29	22

1/1 (1,1)　分析　備考

図５　残差分析分割表

まず、この分割表に対してχ^2検定を実施した結果を図６に示す。結果に有意差があることから、残差分析を行うと図７に示すような結果になる。


χ2検定

χ2検定結果

データ数	296
分割数行	3
分割数列	3
自由度	4
χ2統計値	10.6467
片側確率 P	0.0308
有意水準α	0.05

P＜αより、群間に差があるといえる。

図６　χ^2分析結果


残差分析

基準化残差	賛成	中立	反対
年齢群1	-2.5414	0.4176	2.1203
年齢群2	-0.1249	-0.5206	0.6273
年齢群3	2.7115	0.1109	-2.8001
両側確率			
年齢群1	0.0110	0.6762	0.0340
年齢群2	0.9006	0.6027	0.5305
年齢群3	0.0067	0.9117	0.0051

図７　残差分析結果

これによると、群間に差があるが、その位置は年齢群１の賛成と反対、年齢群３の賛成と反対にあることが分かる。

6.　棄却検定

データの中に飛び離れた値があり、これを分析から除くべきかどうか調べる必要がある場合、Grubbs-Smirnov 棄却検定が利用できる。飛び離れたデータが最大値$x_{\max}$ である場合、まず、それを除いたデータが正規分布に従うかどうか確認する。正規分布に従う場合、以下の統計量

$T_{\max}$ を求め、それと全データ数を用いて参考文献[5]で与えられた数表から検定確率を調べる。

$$T_{\max} = \frac{x_{\max} - \overline{x}}{u}$$

ここに、$\overline{x}$ と u はそれぞれ全データを用いた平均値と、不偏分散からの標準偏差である。

データが正規分布でない場合、対数正規分布も確認する。対数正規分布の場合は、データに対数変換を行って上と同様の検定を行う。正規分布でも対数正規分布でもない場合は、一応元データを用いて検定を行ってはいるが、信頼性は乏しい。

飛び離れたデータが最小値 $x_{\min}$ である場合も全く同様に、以下の統計量 $T_{\min}$ を利用する。

$$T_{\min} = \frac{\overline{x} - x_{\min}}{u}$$

「棄却検定」のボタンは、量的データの集計の実行画面の中にある。

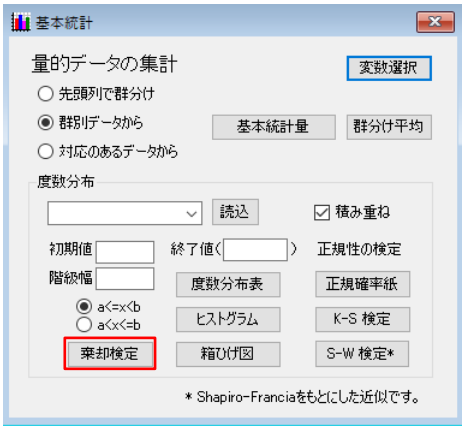


図 1　量的データの集計実行画面

図 2a と図 2b にデータが正規分布の場合と非正規分布の場合のプログラムの実行結果を示す。

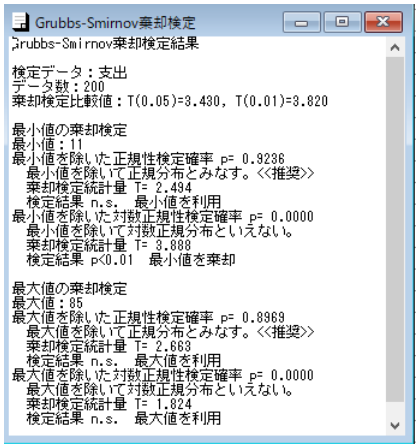


図 2a　実行結果（正規分布の場合）

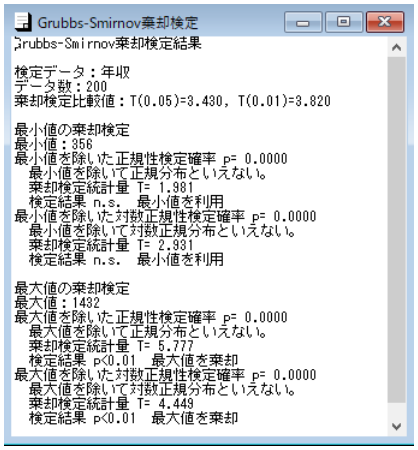


図 2b　実行結果（非正規分布の場合）

7. 相関と回帰の比較検定

これまで College Analysis の相関と回帰分析では、相関係数と回帰係数は 0 との比較の場合だけを考えてきた。しかし、相関係数や回帰式が同じかどうかを調べることも多くなると考え、検定を加えることにした。

相関係数と母相関係数の比較では、データ数を n 、標本相関係数を r 、母相関係数を ρ として、以下の関係を利用する。

$$T = \frac{\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}}{1/\sqrt{n-3}} \sim N(0,1)$$

２群の相関係数の比較では、データ数を n_1, n_2 、標本相関係数を r_1, r_2 として、以下の関係を利用する。

$$T = \frac{\frac{1}{2}\log\frac{1+r_1}{1-r_1} - \frac{1}{2}\log\frac{1+r_2}{1-r_2}}{\sqrt{1/(n_1-3)+1/(n_2-3)}} \sim N(0,1)$$

回帰係数と母回帰係数の比較では、データ数を n 、標本回帰式を $y = ax + b$ 、母回帰式を $y = \alpha x + \beta$ として、以下の関係を利用する。

勾配係数の比較　$T_a = (a-\alpha)\sqrt{SS_x/V_E} \sim t_{n-2}$

定数係数の比較　$T_b = \dfrac{b-\beta}{\sqrt{V_E\left(1/n + \overline{x}^2/SS_X\right)}} \sim t_{n-2}$

ここに、以下の関係を用いている。

$$\overline{x} = \frac{1}{n}\sum_{\lambda=1}^{n} x_\lambda ,\quad \overline{y} = \frac{1}{n}\sum_{\lambda=1}^{n} y_\lambda$$

$$SS_x = \sum_{\lambda=1}^{n} x_\lambda^2 - n\overline{x}^2 ,\quad SS_y = \sum_{\lambda=1}^{n} y_\lambda^2 - n\overline{y}^2 ,\quad SS_{xy} = \sum_{\lambda=1}^{n} x_\lambda y_\lambda - n\overline{xy}$$

$$V_E = \frac{1}{n-2}\left[SS_y - (SS_{xy})^2/SS_x\right]$$

2 群の回帰係数の比較では、データ数を n_1, n_2 、標本回帰式を $y = a_1 x + b_1$, $y = a_2 x + b_2$ として、まず、以下の関係を利用して勾配係数の比較を行う。

$$F_a = \left[(\Delta_2/\Delta_1) - 1\right](n_1 + n_2 - 4) \sim F_{1,\, n_1+n_2-4}$$

勾配係数が異なるとすると、回帰式はそのまま使われ、勾配係数が等しいとすると、以下の関係を利用して定数係数の比較を行う。

$$F_b = \left[(\Delta_3/\Delta_2)-1\right](n_1+n_2-3) \sim F_{1,\,n_1+n_2-3}$$

ここで、定数係数が異なるとすると $a=(SS_{xy1}+SS_{xy2})/(SS_{x1}+SS_{x2})$, $b_i=\overline{y}_i-a\overline{x}_i$ として、回帰式は以下を与える。

$$y=ax+b_1,\quad y=ax+b_2$$

定数係数が同じとすると $a=SS_{xy}/SS_x$, $b=\overline{y}-a\overline{x}$ として、回帰式は同一に以下で与えられる。

$$y=ax+b$$

ここに、$i=1,2$ として以下の関係を用いた。

$$\overline{x}_i=\frac{1}{n_i}\sum_{\lambda=1}^{n}x_{i\lambda},\quad \overline{y}_i=\frac{1}{n_i}\sum_{\lambda=1}^{n}y_{i\lambda}$$

$$SS_{xi}=\sum_{\lambda=1}^{n}x_{i\lambda}^2-n_i\overline{x}_i^2,\quad SS_{yi}=\sum_{\lambda=1}^{n}y_{i\lambda}^2-n_i\overline{y}_i^2,\quad SS_{xyi}=\sum_{\lambda=1}^{n}x_{i\lambda}y_{i\lambda}-n_i\overline{x}_i\overline{y}_i$$

$$\Delta_1=\left[SS_{y1}-(SS_{xy1})^2/SS_{x1}\right]+\left[SS_{y2}-(SS_{xy2})^2/SS_{x2}\right]$$

$$\Delta_2=SS_{y1}+SS_{y2}-\frac{(SS_{xy1}+SS_{xy2})^2}{SS_{x1}+SS_{x2}}$$

$$\Delta_3=SS_y-(SS_{xy})^2/SS_x$$

メニュー［分析－基本統計－相関と回帰分析］を選択して表示される新しい実行画面を図１に示す。

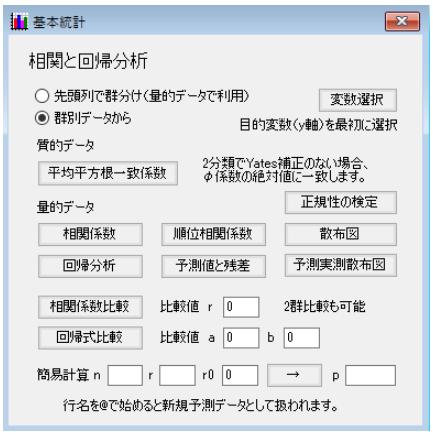


図１　分析実行画面

実行画面の下半分が新しく作成した部分である。以下の図２のようなデータを用いて、「先頭列で群分け」として「相関係数比較」ボタンをクリックすると図３のような分析結果が表示される。

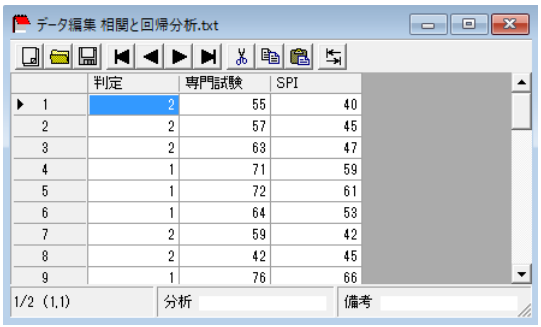


図２　相関係数と回帰係数の比較データ

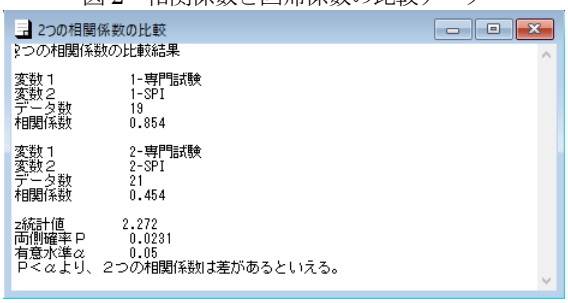


図３　相関係数の比較分析結果

また、「回帰式比較」のボタンをクリックすると、図４のような結果が表示される。

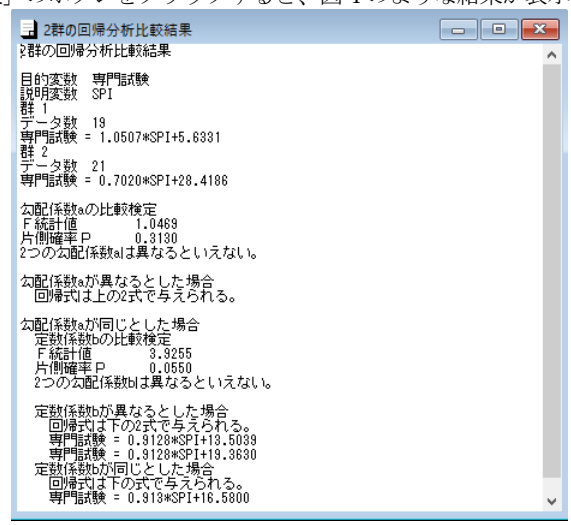


図４　回帰式の比較結果

8. おわりに

リッジ回帰分析や PLS 回帰分析などは、多重共線性が生じた場合、基本的に自動的に補修を行う手法である。我々が分析で使うとすると、そのまま使うのではなく、多重共線性の元を考えて、まずデータから手を加えるはずである。そのため、分析手法として独立させることは、実用上さほど意味があるように思われない。しかし、教育用としては非常に分かり易く、それぞれの差も理解し易いので有効であると思う。

次に、非線形最小 2 乗法についてメニュー画面を見直した。元の画面はかなり複雑であったので、統計に不慣れな利用者には使いづらかったと思われるが、新しい画面は相当簡単な見た目で使い易いので、あまり抵抗なく使えるように思う。分析方法も、初期値を一様乱数で発生させて収束計算を何度も試みる方法に変えて、MCMC の Metropolis-Hastings 法を使って現実的な初期値に設定しているので収束計算は 1 回で済む。そのためあまり多くの設定は必要ない。ただ、MCMC の Metropolis-Hastings 法の特徴として、値の範囲が限られた多くのパラメータを決めるのは得意であるが、パラメータは少数であっても MCMC の初期値から遠く離れたところに解がある場合は、その位置まで移動する際に問題が生じることがある。そのため、不本意ではあるが、メニューの中にパラメータの初期値が設定可能な元のメニューを呼び出せるボタンを用意している。このような例の場合は初期値から離れたところまで飛べる Hamiltonian モンテカルロ法を利用すればよいと聞く[6]。しかしこれについては試作的にプログラムに組み込んではいるが、応用としてまだ利用していない。

ブートストラップは解析的にパラメータの区間推定範囲が求められない際に、モンテカルロ法によってそれを推定する手法である。我々はこの手法を母平均の推定問題とパス解析に応用した。母平均の推定問題で、正規分布する母集団で理論値と比較したところ、類似する結果を得た。また、パス解析では通常求められない、間接効果や擬似相関の区間推定に利用できることも分かった。これらはいずれも 1 回の計算が短時間で処理される分析である。計算時間を要する、例えばニュートン・ラフソン法を用いる、共分散構造分析や非線形最小 2 乗法、非計量多次元尺度構成法などでの利用は難しい。

多重比較の問題では、多変量解析の「実験計画法」の中に、これまでの Fisher の LSD 法に加え、新しく Bonferroni の方法、Turkey の方法、Scheffe の方法を導入した。Bonferroni の方法はどんな場合にも使える方法であるので、ここでは正規性や等分散性がある場合とない場合に自動で分けて、結果を表示するようにした。この良し悪しには問題があるが、基本的には pooled t 検定と joint rank Wilcoxon 検定と同じであるので（検定確率が多重比較する回数倍になっている）、Fisher の LSD 法のところを利用して答えを求めるのもよい。Turkey の方法は正規性、等分散性が必要であり、1 元配置分散分析を予め実行することなく直接多重比較が可能な方法である。最後に Scheffe の方法については、Kruskal-Wallis 検定で差があると判定された場合について検定を行うように言われている。

実験計画法についてはその他の ICC 分析やグラフ描画機能も追加しているので、多少メニューが見にくくなっているように思う。今後もう少し改良するようにしてみたい。

飛び離れたデータの棄却検定は基本統計の「量的データの集計」の中に組み込んだが、ヒストグラムや箱ひげ図との関係もあり、良い選択だったと思う。これは最大と最小のデータについて判定するので、１つずつ消して再度検定をし直すようにして使う。

基本統計の「相関と回帰分析」の中に相関係数と回帰式の比較検定を追加したが、実験計画法のところと同じく、メニューとしては少し複雑になったように思う。基本統計は初心者の利用を考えて作っているので、もう少し工夫が必要かもしれない。

参考文献

[1] 福井正康, College Analysis リファレンスマニュアル, http://www.heisei-u.ac.jp/mi/fukui/

[2] 福井正康, 大山知之, 織田望, 社会システム分析のための統合化プログラム３０ －異常検知－, 福山平成大学経営研究, 第 13 号, (2017) 121-138.

[3] 井出剛, 入門機械学習による異常検知, コロナ社, 2015.

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

[5] 丹後俊郎, 古川俊之監修, 新版 医学への統計学, 朝倉書店, 1993.

[6] 豊田秀樹, 基礎からのベイズ統計学：ハミルトニアンモンテカルロ法による実践的入門, 浅倉書店, 2015.

Multi-purpose Program for Social System Analysis 32 - Multicollinearity, Bootstrap Method and Others -

Masayasu FUKUI

Department of Business Administration, Faculty of Business Administration, Fukuyama Heisei University

Abstract: We have been constructing a unified program on the social system analysis for the purpose of education. This time, we make a program on methods to avoid multicollinearity, and revise a program on nonlinear least squares method. We also introduce bootstrap method, representative multiple comparisons tests, rejection test of exceptional data and comparison test on correlation and regression.

Keywords: College Analysis, multicollinearity, nonlinear least squares method, bootstrap, multiple comparisons test, rejection test

URL: http://www.heisei-u.ac.jp/ba/fukui/