College Analysis 総合マニュアル

― 多変量解析３ ―

目次

２３．生存時間分析

23.1　生存時間分析とは

生存時間分析は時間に依存した生存数や死亡数から、死亡危険率や生存確率分布を予測する分析手法である。この分析は生物の生存時間だけでなく、機械の故障までの時間や達成目標への到達時間などにも利用できる。ここでは慣例に習って生存とか死亡とかの言葉を用いるが、状況に応じて解釈してもらいたい。

時刻$t=0$に$l(0)$個の個体があり、死亡や観測打ち切りなどで、時刻tに個体数が$l(t)$個、時刻$t+h$には$l(t+h)$個になっているものとする。この時間hの間の期間生存率$p(h,t)$は、以下のようになる。

$$p(h,t)=\frac{l(t+h)}{l(t)}$$

同様にして、期間死亡率$q(h,t)$も以下のように与えられる。

$$q(h,t)=1-p(h,t)=\frac{l(t)-l(t+h)}{l(t)}=\frac{d(h,t)}{l(t)}$$

ここに$d(h,t)=l(t+h)-l(t)$は期間死亡数を表す。特に、$h=1$とした期間生存率、期間死亡率を単に時刻tでの生存率$p(t)$、死亡率$q(t)$という。

時刻t以降の全個体の生存時間の合計$T(t)$を個体の数で割った量$e(t)$を平均余命という。

$$e(t)=T(t)/l(t)$$

また、$t=0$での平均余命を平均寿命という。

死亡の発生までの時間を確率変数Tとする確率分布を考え、その密度関数を$f(t)$、分布関数を$F(t)$とすると、分布関数$F(t)$は累積死亡関数で、時刻 t までに死亡する個体の割合である。これに対して、時刻 t まで生きる確率を表す関数$S(t)$を生存関数という。生存関数$S(t)$と分布関数$F(t)$の関係は $S(t)=1-F(t)$ のように与えられる。

時刻 t における死亡発生危険率をハザード関数（故障率関数）$\lambda(t)$といい、以下のように定義する。

$$\lambda(t)=\frac{f(t)}{S(t)}=-\frac{d}{dt}\log S(t)$$

生存時間分析では打ち切りのあるデータを扱う。打ち切りのあるデータの生存関数については、Kaplan-Meier の product-limit 推定法と呼ばれる方法がある。死亡が発生した時刻ごとに区切ったある時刻tで$l(t)$の生存が確認されており、次の区切りの時刻までの時間hの間に $d(t)$の死亡と$w(t)$の打ち切りがあったものとする。通常、期間死亡率$q(t)$は、死亡数$d(t)$をリスクにさらされた個体数で割って求めるが、product-limit 推定法では、リスクにさらされた個体数を、その時の生存数から打ち切り数の半分を引いた$l(t)-w(t)/2$とする。これを用いて期間死亡率$q(t)$及び期間生存率$p(t)$を以下のように与える。

$$q(t)=\frac{d(t)}{l(t)-w(t)/2},\quad p(t)=1-q(t)$$

生存関数は、期間生存率を時刻0から時間区切りごとに時刻tまで掛け合わせたものとする。

生存時間分布には、指数分布かワイブル分布が仮定されることが多い（$t \geq 0$）。

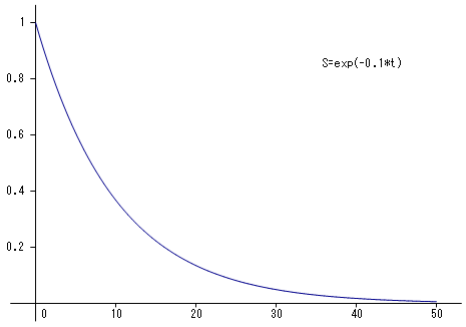
指数分布の密度関数は以下で与えられる。

$$f(t) = \lambda e^{-\lambda t}$$

生存関数は以下である。

$$S(t) = e^{-\lambda t}$$

指数分布の生存関数を図で表すと以下のようになる。



指数分布の生存関数

指数分布の場合、ハザード関数は一定である。

$$\lambda(t) = \frac{f(t)}{S(t)} = \lambda$$

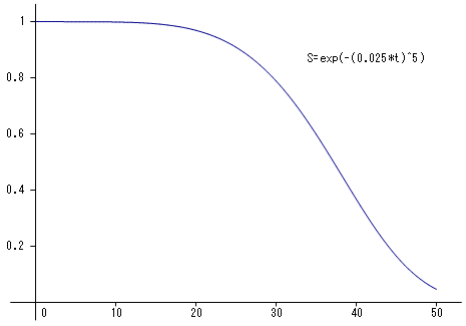
ワイブル分布の密度関数は以下で与えられる。

$$f(t) = (a/b)(t/b)^{a-1} \exp\left[-(t/b)^{a}\right]$$

生存関数は以下である。

$$S(t) = \exp\left[-(t/b)^{a}\right]$$

ワイブル分布の生存関数を図で表すと以下のようになる。



ワイブル分布の生存関数

ハザード関数は以下である。

$$\lambda(t) = \frac{f(t)}{S(t)} = at^{a-1}/b^{a}$$

指数分布は、ワイブル分布の $a = 1, b = 1/\lambda$ という特別な場合に相当する。

指数分布やワイブル分布の見極めは、生存関数に関する以下の関係を利用し、グラフが直

線になるか否かで判断することができる。

指数分布　　$-\log S(t) = \lambda t$

ワイブル分布　　$\log(-\log S) = a\log(t/b) = a\log t - a\log b$

次にワイブル分布を仮定した生存関数のパラメータの最尤推定法を簡単に述べる。今、i 番目の対象者が死亡するか、打ち切られる時刻を t_i とする。ここで、打ち切りデータと非打ち切りデータをそれぞれ $\delta_i = 0,1$ とする。

ワイブル分布の最尤推定で、尤度 $L(a,b)$ は以下で与えられることが知られている。

$$L(a,b) = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

ここに、$f(t)$ と $S(t)$ はワイブル分布の密度関数と生存関数である。この尤度を最大化してパラメータ a, b を求めるのが最尤推定法である。計算を取り扱い易くするために、尤度の対数を取った対数尤度を考える。実行にはコンピュータを利用する。

2 種類以上のワイブル分布を以下のように組み合わせる混合ワイブル分布を用いる方法もある。

$$f(t) = \sum_{k=1}^{K} \pi_k f_k(t)$$

計算には新しい EM アルゴリズムという方法を使う。この方法を組み込んだ生存時間分析はあまり聞かない。

比例ハザードモデルはハザード関数(死亡の危険率)を説明変数で予測するモデルである。即ち、ハザート関数に対して、それに関係する説明変数 $\mathbf{x} = {}^t(x_1, \cdots, x_p, 1)$ とパラメータ $\boldsymbol{\beta} = {}^t(\beta_1, \cdots, \beta_p, \beta_0)$ を用いて以下の仮定を行う。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t)\exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox の比例ハザードモデルでは $\lambda_0(t)$ と定数項 β_0 について議論しないが、ワイブルハザードモデルでは

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

$$b^{-a} = e^{\beta} \rightarrow \exp({}^t\mathbf{x}\boldsymbol{\beta})$$

として、時間に関してワイブル分布のハザード関数を仮定する。

混合ワイブルハザードモデルについては、通常のワイブル分布と比較すると、k 番目の分布について、以下の仮定をする。

$$b_k^{-a_k} = e^{\beta_k} \rightarrow \exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)$$

23.2　プログラムの利用法

メニュー［分析－多変量解析他－生存時間分析］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。

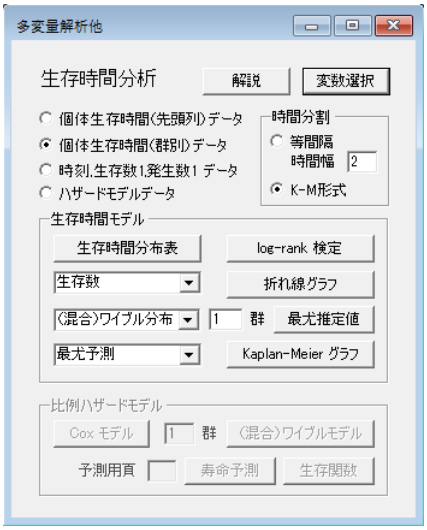


図１　生存時間分析実行メニュー

この分析のデータ形式は大きく分けて 3 種類ある。１つは個体の生存時間を元にしたデータで、先頭列で分類される形式とすでに群別に並べられている形式に分けられる。これらの形式は基本統計のデータ形式に類似している。次に、すでに生命表に近い形式になっているデータである。これは、観測時刻、その時点での生存個体数、その時点より後で次の時点までに死亡する期間発生数が、すでに表の形式になっているデータである。生存個体数と期間発生数は複数組入力が可能である。詳しくはサンプルを見てもらいたい。最後は、ハザードモデルデータで、重回帰分析などと同様の形式である。最初と最後の形式で、通常のデータと異なる部分は、観測の打ち切りデータが含まれる点である。打ち切りデータは、観測を打ち切られた時点の数値の後ろに+記号を付けて表す。観測が打ち切られた際の扱いは、生存数から打ち切られたデータ数の半分を引いて、死亡リスクに晒されたデータ数として処理している[1]。

最初に図 2 の単独データを元に説明をする。

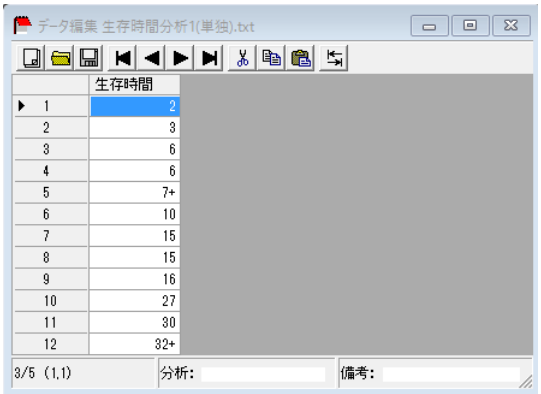
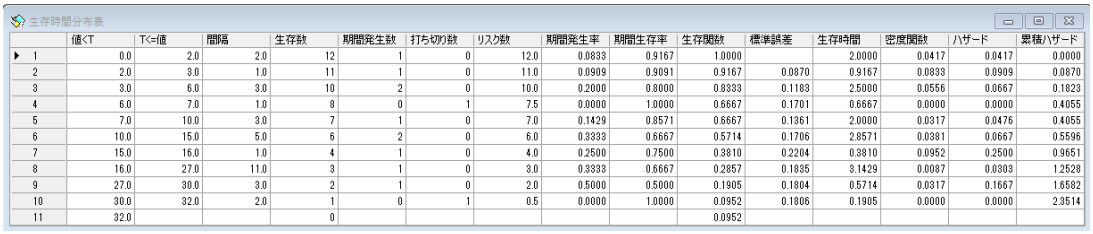


図 2　単独データ（生存時間分析 1(単独).txt 3 頁目）

このデータでは、2 個体が観測を打ち切られている。

「個体生存時間(群別)データ」ラジオボタンを選択し、変数選択を実行して、「生存時間分布表」ボタンをクリックすると図 3 のような結果が表示される。



生存時間分布表

	値<T	T<=値	間隔	生存数	期間発生数	打ち切り数	リスク数	期間発生率	期間生存率	生存関数	標準誤差	生存時間	密度関数	ハザード	累積ハザード
1	0.0	2.0	2.0	12	1	0	12.0	0.0833	0.9167	1.0000		2.0000	0.0417	0.0417	0.0000
2	2.0	3.0	1.0	11	1	0	11.0	0.0909	0.9091	0.9167	0.0870	0.9167	0.0833	0.0909	0.0870
3	3.0	6.0	3.0	10	2	0	10.0	0.2000	0.8000	0.8333	0.1183	2.5000	0.0556	0.0667	0.1823
4	6.0	7.0	1.0	8	0	1	7.5	0.0000	1.0000	0.6667	0.1701	0.6667	0.0000	0.0000	0.4055
5	7.0	10.0	3.0	7	1	0	7.0	0.1429	0.8571	0.6667	0.1361	2.0000	0.0317	0.0476	0.4055
6	10.0	15.0	5.0	6	2	0	6.0	0.3333	0.6667	0.5714	0.1706	2.8571	0.0381	0.0667	0.5596
7	15.0	16.0	1.0	4	1	0	4.0	0.2500	0.7500	0.3810	0.2204	0.3810	0.0952	0.2500	0.9651
8	16.0	27.0	11.0	3	1	0	3.0	0.3333	0.6667	0.2857	0.1835	3.1429	0.0087	0.0303	1.2528
9	27.0	30.0	3.0	2	1	0	2.0	0.5000	0.5000	0.1905	0.1804	0.5714	0.0317	0.1667	1.6582
10	30.0	32.0	2.0	1	0	1	0.5	0.0000	1.0000	0.0952	0.1806	0.1905	0.0000	0.0000	2.3514
11	32.0			0						0.0952					

図 3　生存時間分布表結果

図 3 では、様々な指標が区切られた時点毎に表示されている。ここで特に大切な指標は、「生存関数」と「ハザード」である。これらはそれぞれ、その時点まで生存している確率とその時点での死亡の危険率の意味を持つ。

図 3 の生存時間分布表の中で、生存数、生存関数、ハザード関数、累積ハザード関数については、コンボボックスで設定して、「折れ線グラフ」ボタンをクリックすると表示される。ここでは生存関数とハザード関数についてのグラフを図 4a と図 4b に示す。

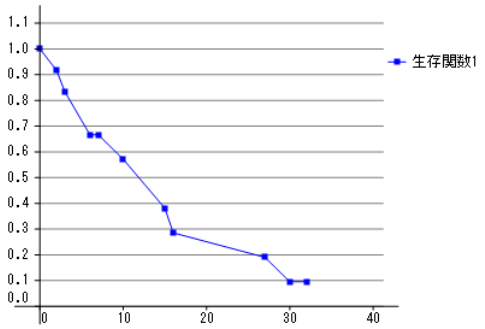


図 4a　生存関数

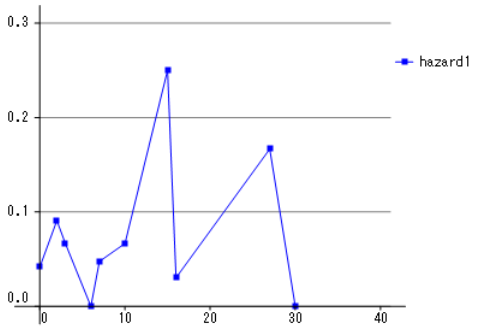


図 4b　ハザード関数

また、同じコンボボックスで「指数分布確認」または「ワイブル分布確認」を選択すると、図 5a と図 5b のような図が表示される。

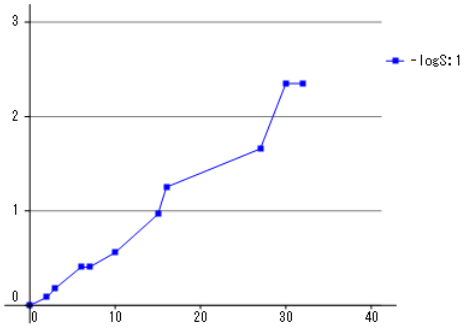


図 5a　指数分布確認

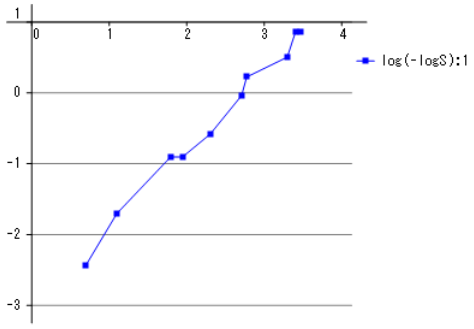


図 5b　ワイブル分布確認

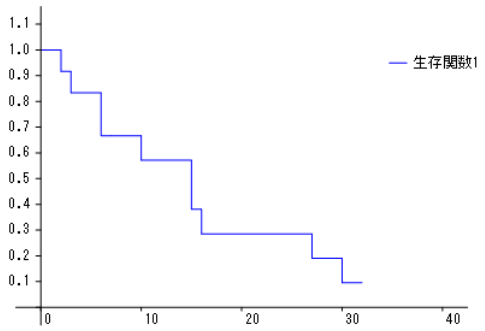
生存時間が指数分布またはワイブル分布に従うならば、それぞれの累積生存関数の時間依存性からこの点列は直線状に並ぶ。指数分布はワイブル分布の特殊な場合であるので、指数分布が成り立つ場合はワイブル分布も成り立つ。

分布の確認の場合、「折れ線グラフ」をクリックすると、上図と共に分布の当てはまりの良さを示す、図 6a や図 6b のような指標も表示される。

図 6a　指数分布の指標

図 6b　ワイブル分布の指標

生存関数の Kaplan-Meier 推定のグラフは、「Kaplan-Meier グラフ」ボタンをクリックして表示される。その際、左のコンボボックスで指定して、指数分布またはワイブル分布の予想曲線を描くこともできる。予想曲線のないグラフと、ワイブル分布の予想曲線を付けて描いたグラフを図 7a と図 7b に示す。



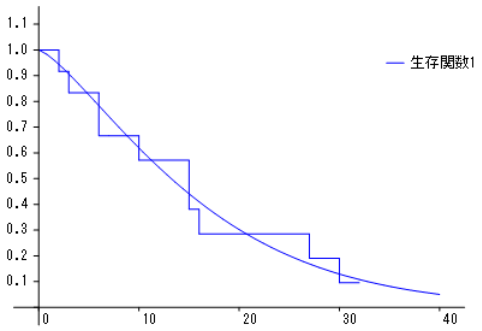


図 7a　Kaplan-Meier 生存関数グラフ　図 7b　予想曲線付き Kaplan-Meier グラフ

これらの予想曲線では最小 2 乗法によるものと最尤法によるものとが選択できる。上図は最尤法によるものである。

また、予想曲線は混合指数分布や混合ワイブル分布についても表示することができる。その際は分布を選んだコンボボックスの右のテキストボックスで混合する数を指定する。図 8 に 2 群の混合ワイブル分布による予測曲線を付けた Kaplan-Meier グラフを表示する。サンプルでは 2 つの時期に危険度が高くなっている。

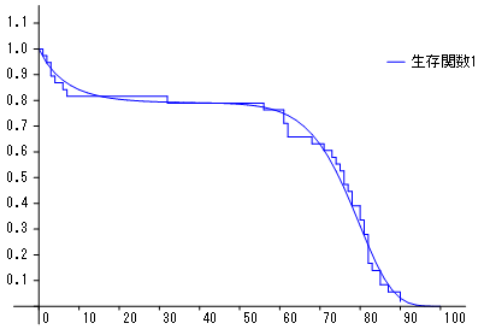
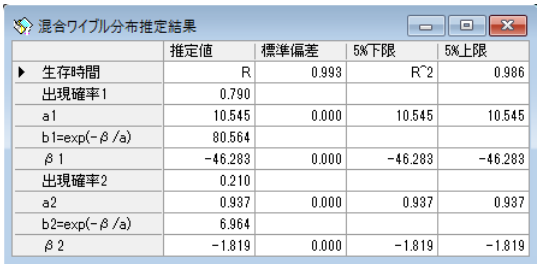


図 8　2 群混合分布による予測（生存時間分析 1(単独).txt 8 頁目）

このパラメータの値については、上と同じ設定で「最尤推定値」ボタンをクリックすると、図 9 のように表示される。



	推定値	標準偏差	5%下限	5%上限
生存時間	R	0.993	R^2	0.986
出現確率1	0.790			
a1	10.545	0.000	10.545	10.545
b1=exp(-β/a)	80.564			
β1	-46.283	0.000	-46.283	-46.283
出現確率2	0.210			
a2	0.937	0.000	0.937	0.937
b2=exp(-β/a)	6.964			
β2	-1.819	0.000	-1.819	-1.819

図 9　2 群混合ワイブル予測（生存時間分析 1(単独).txt 8 頁目）

ここでは表示されていないが、混合がない場合には、右端に最小 2 乗推定による推定値も表示される。

複数群の生存時間分布表は、先頭列で群分けデータ（生存時間分析 2(2 群比較).txt）または群別データを元に図 10 のように縦に並べて表示される。

	値<T	T<=値	間隔	生存数	期間発生数	打ち切り数	リスク数	期間発生率	期間生存率	生存関数	標準誤差	生存時間	密度関数	ハザード	累積ハザード
1	0.0	1.0	1.0	12	0	0	12.0	0.0000	1.0000	1.0000		1.0000	0.0000	0.0000	0.0000
2	1.0	2.0	1.0	12	1	0	12.0	0.0833	0.9167	1.0000	0.0000	1.0000	0.0833	0.0833	0.0000
3	2.0	3.0	1.0	11	1	0	11.0	0.0909	0.9091	0.9167	0.0870	0.9167	0.0833	0.0909	0.0870
4	3.0	6.0	3.0	10	2	0	10.0	0.2000	0.8000	0.8333	0.1183	2.5000	0.0556	0.0667	0.1823
5	6.0	7.0	1.0	8	1	0	8.0	0.1250	0.8750	0.6667	0.1701	0.6667	0.0833	0.1250	0.4055
6	7.0	9.0	2.0	7	0	0	7.0	0.0000	1.0000	0.5833	0.1627	1.1667	0.0000	0.0000	0.5390
7	9.0	10.0	1.0	7	1	0	7.0	0.1429	0.8571	0.5833	0.1423	0.5833	0.0833	0.1429	0.5390
8	10.0	15.0	5.0	6	2	0	6.0	0.3333	0.6667	0.5000	0.1604	2.5000	0.0333	0.0667	0.6931
9	15.0	16.0	1.0	4	1	0	4.0	0.2500	0.7500	0.3333	0.2041	0.3333	0.0833	0.2500	1.0986
10	16.0	22.0	6.0	3	0	0	3.0	0.0000	1.0000	0.2500	0.1667	1.5000	0.0000	0.0000	1.3863
11	22.0	27.0	5.0	3	1	0	3.0	0.3333	0.6667	0.2500	0.1250	1.2500	0.0167	0.0667	1.3863
12	27.0	30.0	3.0	2	1	0	2.0	0.5000	0.5000	0.1667	0.1614	0.5000	0.0278	0.1667	1.7918
13	30.0	32.0	2.0	1	1	0	1.0	1.0000	0.0000	0.0000	0.1596	0.0000	0.0000	0.5000	0.0000
14	32.0			0						0.0000					
1	0.0	1.0	1.0	9	4	0	9.0	0.4444	0.5556	1.0000		1.0000	0.4444	0.4444	0.0000
2	1.0	2.0	1.0	5	1	0	5.0	0.2000	0.8000	0.5556	0.2981	0.5556	0.1111	0.2000	0.5878
3	2.0	3.0	1.0	4	2	0	4.0	0.5000	0.5000	0.4444	0.2070	0.4444	0.2222	0.5000	0.8109
4	3.0	6.0	3.0	2	0	0	2.0	0.0000	1.0000	0.2222	0.2772	0.6667	0.0000	0.0000	1.5041
5	6.0	7.0	1.0	2	0	0	2.0	0.0000	1.0000	0.2222	0.1386	0.2222	0.0000	0.0000	1.5041
6	7.0	9.0	2.0	2	1	0	2.0	0.5000	0.5000	0.2222	0.1386	0.4444	0.0556	0.2500	1.5041
7	9.0	10.0	1.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.2095	0.1111	0.0000	0.0000	2.1972
8	10.0	15.0	5.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.1048	0.5556	0.0000	0.0000	2.1972
9	15.0	16.0	1.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.1048	0.1111	0.0000	0.0000	2.1972
10	16.0	22.0	6.0	1	1	0	1.0	1.0000	0.0000	0.0000	0.1048	0.0000	0.0000	0.1667	0.0000
11	22.0	27.0	5.0	0						0.0000					
12	27.0	30.0	3.0												
13	30.0	32.0	2.0												
14	32.0														

図 10　2 群の生存時間分布表

これ以外に、もっと群の違いを比較できる方法を考えて行きたい。

複数群の生存関数と Kaplan-Meire 生存関数グラフを図 11 と図 12 に示す。

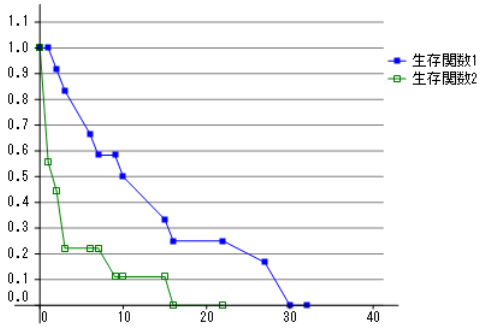


図 11　2 種類の生存関数グラフ

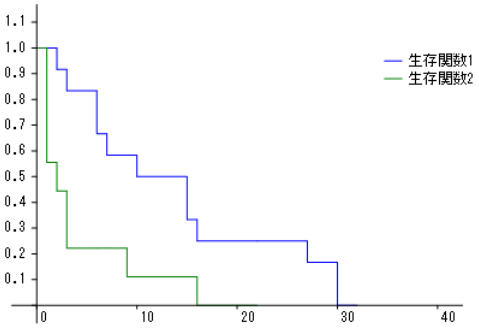


図 12　2 種類の Kaplan-Meier グラフ

複数群の生存関数間の差の log-rank 検定結果は、「log-rank 検定」ボタンをクリックすると図 13 のように表示される。

図 13　log-rank 検定結果

最後に、比例ハザードモデルの分析結果について示しておく。データは図 14 のような重回帰分析などと同じデータ形式である。

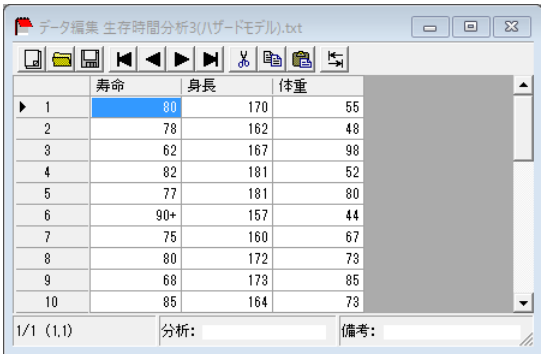


図 14　比例ハザードモデルデータ（生存時間分析 3(ハザードモデル).txt）

ハザードモデルでは Cox 比例ハザードモデルと Weibull 比例ハザードモデルを組み込んでいる。ハザード関数について、2 つのモデルとも以下の形を仮定する。

$$\lambda(t\,|\,\mathbf{x},\boldsymbol{\beta}) = \lambda_0(t)\exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox 比例ハザードモデルは $\lambda_0(t)$ や β_0 の推定は行わないが、分布の形に依存しない利点がある。Weibull ハザードモデルでは、時間部分にワイブル分布を仮定し、そのパラメータも説明変数で推定するという一般化線形モデルの形式を採用している。

$$\lambda(t\,|\,\mathbf{x},\boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

「Cox モデル」ボタンをクリックした結果を図 12 に、「Weibull モデル」ボタンをクリックした結果を図 15 に示す。

図 15　Cox 比例ハザードモデル結果

図 16　Weibull 比例ハザードモデル

最後に Weibull 比例ハザードモデルが予想する生存時間の平均値と実際の観測値との比較を行ってみる。「寿命予測」ボタンをクリックすると図 17a と図 17b の結果が示される。

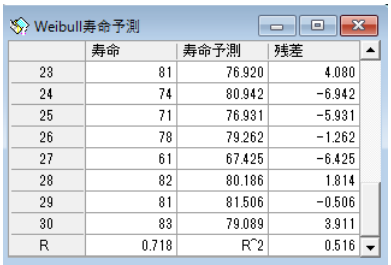


図 17a　寿命予測図

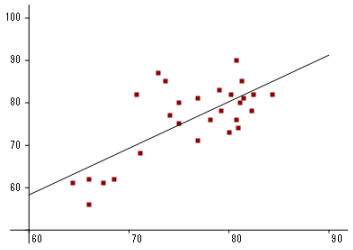


図 17b　実測/予測散布図

これには非打ち切りデータのみが用いられている。また、寿命予測の結果の最後に、予測値と実測値の相関係数の値とその 2 乗の値を表示している。

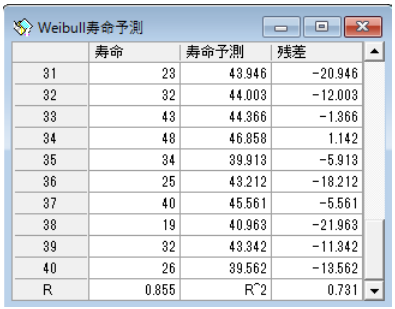
2 種混合ワイブルハザードモデルの場合、比例ハザードモデルの中の「群」テキストボックスに 2 を入れて、「(混合) ワイブルモデル」ボタンをクリックする。図 18 に結果を示す。


複合ワイブル比例ハザードモデル結果

	偏回帰係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
身長	0.0274	0.7407	0.9705	-1.4244	1.4792	1.028E00
体重	0.0526	0.7407	0.9434	-1.3992	1.5043	1.054E00
要因	5.2765	0.7407	0.0000	3.8248	6.7283	1.957E02
出現確率1	0.2497					
a1	29.7025	5.6218		18.6837	40.7212	
γ1	-140.1803	25.9021		-190.9484	-89.4123	
出現確率2	0.7503					
a2	7.3947	1.0264		5.3829	9.4065	
γ2	-40.1365	6.2861		-52.4574	-27.8157	

図 18　混合ワイブルハザードモデル（生存時間分析 3(ハザードモデル).txt 2 頁目）

このモデルによる実測・予測値と重相関係数 R の値、及びそのグラフを表示するには、「予測用頁」テキストボックスを空欄のまま、「寿命予測」ボタンをクリックする。結果は図 19 のようになる。


Weibull寿命予測

	寿命	寿命予測	残差
31	23	43.946	-20.946
32	32	44.003	-12.003
33	43	44.366	-1.366
34	48	46.858	1.142
35	34	39.913	-5.913
36	25	43.212	-18.212
37	40	45.561	-5.561
38	19	40.963	-21.963
39	32	43.342	-11.342
40	26	39.562	-13.562
R	0.855	R^2	0.731

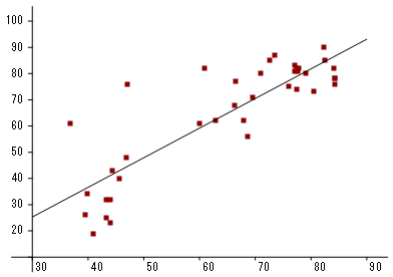


図 19　混合モデルによる実測・予測値

このモデルと混合ワイブル分布の Kaplan-Meier 推定とを比較してみる。寿命予測するページを現在のページ（空欄も可）にして「生存関数」ボタンをクリックし、各個体の生存関数を描画すると図 20 のようになる。また混合ワイブル分布を使った Kaplan-Meier 推定は図 21 のようになる。

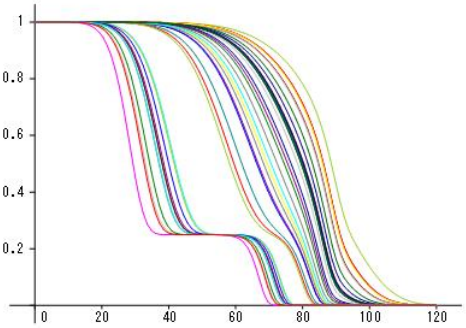


図 20　各個体の生存関数

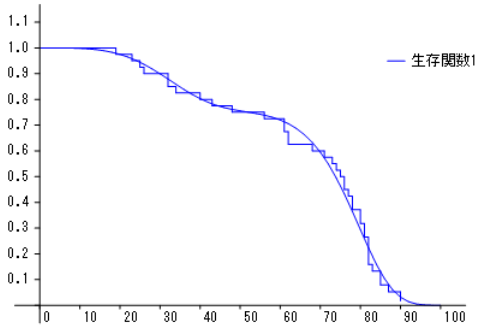


図 21　混合ワイブル分布による推定

このグラフの関係は、図 20 の曲線の平均を取ると、図 22 のように、図 21 に近い形になる。

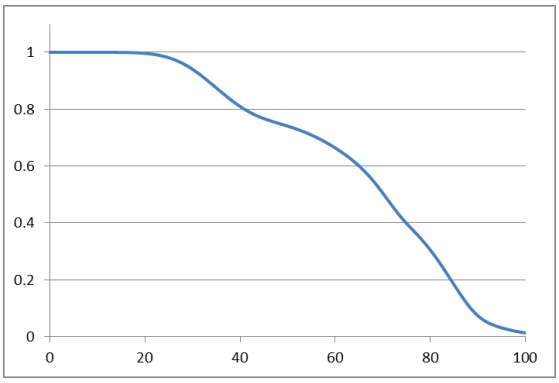
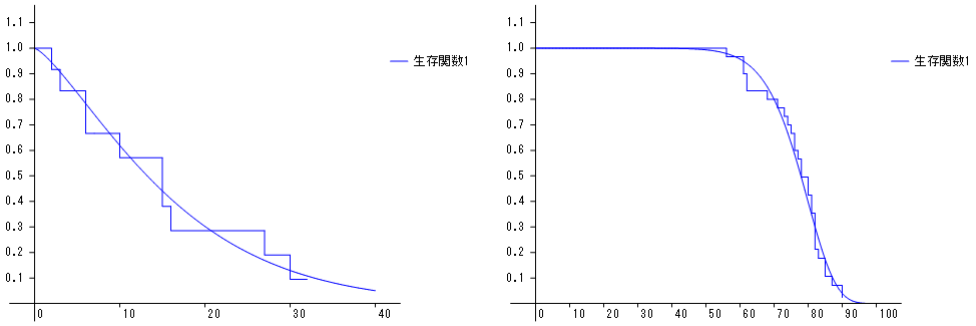


図 22　各個体の生存関数の平均

推定するデータを別頁にするときは、「予測用頁」テキストボックスにデータのある頁番号を入力し、「寿命予測」ボタンをクリックする。

問題１

１）生存時間分析１（単独）.txt の 3 頁目と 6 頁目のデータについて、以下のようなカプラン・マイヤー予測曲線とそのワイブル分布の推定グラフを描け。



２）それぞれのワイブル分布のパラメータの推定値を求めよ。

3 頁目　　a=［　　］, b=［　　］

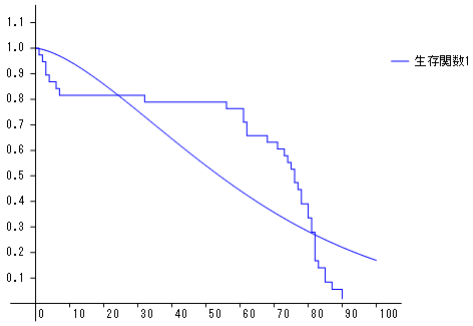
6 頁目　　a=［　　］, b=［　　］

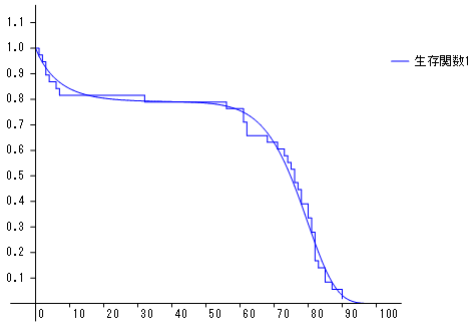
３）3 頁目のデータは指数分布とも考えられるか。指数分布確認の「折れ線グラフ」を書いて確かめよ。（前問のパラメータ a の数値からも分かる。）

グラフはほぼ直線になって［いる・いない］ので、

指数分布と考え［られる・られない］。

４）8 頁目のデータを用いて、以下のようなカプラン・マイヤー予測曲線とそのワイブル分布および 2 種混合ワイブル分布の推定グラフを描け。





５） ２種混合のワイブル分布のパラメータを求めよ。

１群　出現確率［　　］, a=［　　］, b=［　　］

２群　出現確率［　　］, a=［　　］, b=［　　］

寿命を長い群はどちらか。（ヒント b の大きさ）　　［１・２］群

問題２

生存時間分析 2(2 群比較).txt の 1 頁目のデータを用いて以下の問いに答えよ。

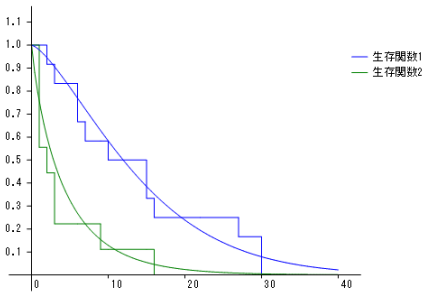
１） ２つの群のカプラン・マイヤー予測曲線を右図のように描け。

２） ２つの生存曲線は異なると言えるか。

Log-rank 検定を用いて判定せよ。

検定確率［　　］より、

異なると［いえる・いえない］。



生存時間分析 4(ハザードモデル).txt の 1 頁目のデータを用いて以下の問いに答えよ。

３）Cox 比例ハザードモデルを用いて各係数、検定確率、Exp(b)の値を求めよ。

	偏回帰係数	検定確率	EXP(b)
要因１			
要因２			

４）Exp(b)の値は 0/1 データの場合、2 つの場合の死亡確率の比になる。この場合、要因の無い人に比べてある人の死亡確率は何倍になるか。

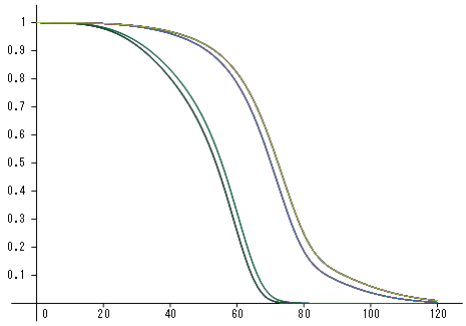
要因１［　　］倍、要因２［　　］倍

５）単独ワイブルモデルの場合、要因の無い人に比べてある人の死亡確率は何倍になるか。　　要因１［　　］倍、要因２［　　］倍

６）2 種混合ワイブルモデルの場合、要因の無い人に比べてある人の死亡確率は何倍になるか。　　要因１［　　］倍、要因２［　　］倍

７）２種混合ワイブルモデルの場合、
寿命に影響のある変数はどちらか。
［要因１・要因２］

８）２種混合ワイブルモデルについて、
各人の生存関数を比較した右図の
ようなグラフを描け。



問題１解答

２）それぞれのワイブル分布のパラメータの推定値を求めよ。
3頁目　　a=［ 1.319 ］, b=［ 17.453 ］
6頁目　　a=［ 10.622 ］, b=［ 80.587 ］

３）3頁目のデータは指数分布とも考えられるか。
グラフはほぼ直線になって［いる・いない］ので、指数分布と考え［られる・られない］。

５）２種混合のワイブル分布のパラメータを求めよ。
１群　出現確率［ 0.790 ］, a=［ 10.545 ］, b=［ 80.564 ］
２群　出現確率［ 0.210 ］, a=［ 0.937 ］, b=［ 6.964 ］
寿命を長い群はどちらか。（ヒントｂの大きさ）　　［1・２］群

問題２解答

２）２つの生存曲線は異なると言えるか。
検定確率［　　］より、異なると［いえる・いえない］。

３）Cox 比例ハザードモデルを用いて各係数、検定確率、Exp(b)の値を求めよ。

	偏回帰係数	検定確率	EXP(b)
要因１	1.9515	0.0000	7.0390
要因２	0.1838	0.6003	1.2017

４）この場合、要因の無い人に比べてある人の死亡確率は何倍になるか。
要因１［ 7.0390 ］倍、要因２［ 1.2017 ］倍

５）単独ワイブルモデルの場合、要因の無い人に比べてある人の死亡確率は何倍になるか。　　要因１［ 5.2562 ］倍、要因２［ 1.2125 ］倍

６）２種混合ワイブルモデルの場合、要因の無い人に比べてある人の死亡確率は何倍になるか。　　要因１［ 6.4308 ］倍、要因２［ 1.2416 ］倍

７）２種混合ワイブルモデルの場合、寿命に影響のある変数はどちらか。
［要因１・要因２］

23.3　生存時間分析の理論

1)　生存時間分析の基礎

時刻$t=0$に$l(0)$個の個体があり、死亡により時刻tに個体数が$l(t)$個になっているものとする。時刻tからの単位時間の間に死亡する割合$p(t)=-dl(t)/dt$は、以下で与えられると仮定する。

$$-\frac{dl(t)}{dt}=\mu(t)l(t)$$

ここに$\mu(t)$を時刻tにおける死力という。

上式を時刻tと時刻$t+h$の間で定積分すると以下の関係を得る。

$$\log l(t+h)-\log l(t)=-\int_t^{t+h}\mu(\tau)d\tau=-\int_0^h\mu(t+\tau)d\tau$$

これより、

$$l(t+h)=l(t)\exp\left[-\int_0^h\mu(t+\tau)d\tau\right]$$

ここで、$p(h;t)=\exp\left[-\int_0^h\mu(t+\tau)d\tau\right]$とおくと、$p(h;t)$は時刻$t$ から$t+h$の間の期間生存率と呼ばれる。この期間生存率は以下のようにも書ける。

$$p(h;t)=\frac{l(t+h)}{l(t)}$$

同様にして、期間死亡率$q(h;t)$も以下のように与えられる。

$$q(h;t)=1-p(h;t)=\frac{l(t)-l(t+h)}{l(t)}=\frac{d(h;t)}{l(t)}$$

ここに$d(h;t)=l(t)-l(t+h)$は期間死亡数を表す。特に、$h=1$とした期間生存率、期間死亡率を単に時刻tでの生存率$p(t)$、死亡率$q(t)$という。

時刻t以降の生存時間の合計$T(t)$を個体数で割った量を平均余命$e(t)$という。

$$e(t)=\int_t^\infty l(\tau)d\tau\Big/l(t)=T(t)/l(t)$$

また、$t=0$での平均余命を平均寿命という。

死亡の発生までの時間を確率変数Tとする確率分布を考え、その密度関数を$f(t)$、分布関数を$F(t)$とすると、これらには以下の関係がある。

$$F(t)=P(0\le T\le t)=\int_0^t f(\tau)d\tau$$

分布関数$F(t)$は累積死亡関数である。これに対して、時刻tまで生きる確率を表す関数$S(t)$を生存関数といい、以下で表される。

$$S(t)=P(T>t)=1-F(t)=\int_t^\infty f(t)dt$$

時刻tにおける死亡発生危険率をハザード関数（故障率関数）$\lambda(t)$といい、以下のように定義される。

$$\lambda(t)=\frac{f(t)}{S(t)}=-\frac{d}{dt}\log S(t)$$

このハザード関数を積分した累積ハザード関数$\Lambda(t)$は以下のように定義される。

$$\Lambda(t)=\int_0^t\lambda(\tau)d\tau=-\log S(t)$$

逆に生存関数は、以下のように表される。

$$S(t)=e^{-\Lambda(t)}$$

生存関数は$t\to\infty$で$S(t)\to 0$であるから、累積ハザード関数は$t\to\infty$で$\Lambda(t)\to\infty$でなければならない。

累積死亡分布には、指数分布や Weibull 分布が仮定される。指数分布の確率密度関数は以

下で与えられる。

$$f(t)=\lambda e^{-\lambda t} \quad (t\geq 0)$$

分布関数と生存関数はそれぞれ以下で与えられる。

$$F(t)=1-e^{-\lambda t}, \qquad S(t)=e^{-\lambda t} \qquad (t\geq 0)$$

確率変数の平均と分散はそれぞれ以下で与えられる。

$$E[T]=\frac{1}{\lambda}, \qquad V[T]=\frac{1}{\lambda^2}$$

ハザード関数は定数で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}=\lambda$$

Weibull 分布の確率密度関数は以下で与えられる。

$$f(t)=(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$

分布関数と生存関数はそれぞれ以下で与えられる。

$$F(t)=1-\exp\left[-\left(t/b\right)^{a}\right], \qquad S(t)=\exp\left[-\left(t/b\right)^{a}\right]$$

確率変数の平均と分散はそれぞれ以下で与えられる。

$$E[T]=b\Gamma\left(1+1/a\right), \qquad V[T]=b^2\left[\Gamma(2+1/a)-\Gamma(1+1/a)\right]^2$$

ここに、$\Gamma(x)$ はガンマ関数である。ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]}{\exp\left[-(t/b)^{a}\right]}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}/b^{a}$$

実際のハザード関数は、初期段階で値が大きく、しばらく時間が経つと安定期に入り、最終的な段階でまた値が大きくなる。安定期では指数分布が使われ、初期段階では Weibull 分布がよく利用される。最終段階ではあまり当てはまりが良くないと言われることもあるが、我々は Weibull 分布を当てはめてみる。全体への当てはめの分布としては、後に述べる混合 Weibull 分布を考えてみることにする。

2) Kaplan-Meier 推定と log-rank 検定

観測対象 $\lambda=1,\cdots,N$ に対して、生存時間を $t_\lambda=0$ から $t_\lambda=T_\lambda$（打ち切りのないデータ）、$t_\lambda=0$ から $t_\lambda=T_\lambda^+$（打ち切りのあるデータ、実際のデータでは 17+ 等と表記）とする。この終了時刻 T_λ を 0 から順番に並べた時刻を $t_0=0,t_1,\cdots,t_m$（同一のものもある）とし、t_m ですべて死亡および打ち切りが確認されたものとする。これに対して、一定の時間間隔で時刻を取る方法もある。各時点での生存数を l_i、$t_i<t\leq t_{i+1}$ の間に死亡した数を d_i、打ち切りになった数を w_i とする。これらを使って、死亡のリスクにさらされた数を $r_i=l_i-w_i/2$ とする。

死亡の期間発生率 q_i と期間生存率 p_i は以下で与えられる。

$q_i = d_i / r_i$,　　$p_i = 1 - q_i$

生存関数S_i、密度関数 f_i、ハザード関数λ_iは以下のように計算される。

$$S_i = \prod_{k=0}^{i-1} p_k \ , \qquad f_i = q_i S_i / (t_i - t_{i-1}) \ , \qquad \lambda_i = f_i / S_i = q_i / (t_i - t_{i-1})$$

このような生存関数の推定法を Kaplan-Meier の product-limit 推定法（以後 Kaplan-Meier 推定法と呼ぶ）という。生存関数S_iのばらつきを表す標準誤差$S.E.[S_i]$は近似的に以下で与えられることが知られている。

$$S.E.[S_i] = S_{i-1} \sqrt{\sum_{k=1}^{i-1} \frac{d_k}{l_k (l_k - d_k)}} \qquad (i \geq 2)$$

期間内の生存時間μ_iは以下で与えられる。

$$\mu_i = S_i (t_i - t_{i-1})$$

指数分布や Weibull 分布の見極めは、累積ハザード関数に関する以下の関係を利用し、グラフが直線になるか否かで判断することができる。

指数分布　　$-\log S(t) = \lambda t$

Weibull 分布　$\log(-\log S) = a \log(t/b) = a \log t - a \log b$

指数分布や Weibull 分布のパラメータの最小 2 乗推定は、以下の式によって与えられる。

指数分布　$S(t) = e^{-\lambda t}$

$$\lambda = -\sum_{i=0}^{m-1} t_i \log S_i \Big/ \sum_{i=0}^{m-1} t_i^2$$

Weibull 分布　$S(t) = \exp\left[-(t/b)^a \right]$

$t_i' = \log t_i$,　$S_i' = \log(-\log S_i)$　として、

$$a = \sum_{i=1}^{m-1} (t_i' - \bar{t}')(S_i' - \bar{S}') \Big/ \sum_{i=1}^{m-1} (t_i' - \bar{t}')^2 \quad , \quad b = \exp\left[-(\bar{S}' - a\bar{t}')/a \right]$$

分類数Gの個体群について、生存時間データの差の検定を行うには以下の性質を用いる。第r分類群のt_i時点での期間死亡数をd_i^r、生存数をl_i^rとして

$$O_r = \sum_{i=0}^{m-1} d_i^r \quad , \qquad E_r = \sum_{i=0}^{m-1} l_i^r (d_i / l_i) \quad , \qquad \text{ここに、} \ l_i = \sum_{r=1}^{G} l_i^r \ , \ \ d_i = \sum_{r=1}^{G} d_i^r$$

を計算し、以下の近似的な関係を用いて群間の差を検定する。

$$\chi^2 = \sum_{r=1}^{G} \frac{(O_r - E_r)^2}{E_r} \sim \chi^2_{G-1}$$

ここに、O_r は分類群 r の実測累積死亡数、E_r は分類群 r の予測累積死亡数である。この検定を Peto & Peto の log-rank 検定という。

3)　パラメータの最尤推定

①　指数分布に基づく最尤推定

最初に通常の指数分布の最尤推定を考える。指数分布の確率密度関数と生存関数は以下で与えられる。

$$f(t)=\lambda\exp(-\lambda t) \qquad (t\geq 0)$$
$$S(t)=\exp(-\lambda t) \qquad (t\geq 0)$$

指数分布の最尤推定で、尤度$L(\lambda)$は以下で与えられる。

$$L(\lambda)=\prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切り（死亡）データをそれぞれ$\delta_i=0,1$としている。

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\lambda$$

対数尤度は以下となる。

$$\log L(\lambda)=\sum_{i=1}^{N}\left[\delta_i \log\lambda(t_i)+\log S(t_i)\right]=\sum_{i=1}^{N}\left[\delta_i\log\lambda-\lambda t_i\right]$$

対数尤度を微分してスコアベクトルに相当するものを作成するが、この場合はスカラーである。これを仮にスコアと呼ぶ。

$$\frac{\partial}{\partial\lambda}\log L=\sum_{i=1}^{N}\left[\delta_i/\lambda-t_i\right]=\frac{1}{\lambda}\sum_{i=1}^{N}\delta_i-\sum_{i=1}^{N}t_i=0$$
$$\lambda=\sum_{i=1}^{N}\delta_i\Big/\sum_{i=1}^{N}t_i$$

スコアをもう一度微分して、情報行列$\Im$に相当するものを作成する。この場合もスカラーである。

$$\Im=-\frac{\partial^2}{\partial\lambda^2}\log L=\frac{1}{\lambda^2}\sum_{i=1}^{N}\delta_i$$

この逆数は、推定値の分散を与える。

② Weibull 分布に基づく最尤推定

最初に通常の Weibull 分布の最尤推定を考える。分布の確率密度関数と生存関数は以下で与えられる。

$$f(t)=(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$
$$S(t)=\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$

Weibull 分布の最尤推定で、尤度$L(a,b)$は以下で与えられる。

$$L(a,b)=\prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切り（死亡）データをそれぞれ$\delta_i=0,1$としている。

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]}{\exp\left[-(t/b)^{a}\right]}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}b^{-a}$$

対数尤度は以下となる。

$$\log L(a,b)=\sum_{i=1}^{N}\left[\delta_i\log\lambda(t_i)+\log S(t_i)\right]=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}b^{-a}\right)-t_i^{a}b^{-a}\right]$$

$$=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}e^{\beta}\right)-t_i^{a}e^{\beta}\right]=\sum_{i=1}^{N}\left[\delta_i\left(\log a+(a-1)\log t_i+\beta\right)-t_i^{a}e^{\beta}\right]$$

ここで、$b^{-a}=e^{\beta}$（$b=e^{-\beta/a}$，$e^{\beta}=b^{-a}\rightarrow\exp({}^{t}\mathbf{x}\boldsymbol{\beta})$に相当）としている。

これを微分して、スコアベクトル$\mathbf{U}$と情報行列$\mathfrak{I}$をもとめると以下となる。

$$\boldsymbol{\beta}'=\begin{pmatrix}a\\ \beta\end{pmatrix},\qquad \mathbf{U}=\begin{pmatrix}\partial\log L/\partial a\\ \partial\log L/\partial\beta\end{pmatrix},\qquad \mathfrak{I}=-\begin{pmatrix}\partial^2\log L/\partial a^2 & \partial^2\log L/\partial a\partial\beta\\ \partial^2\log L/\partial a\partial\beta & \partial^2\log L/\partial\beta^2\end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a}\log L=\sum_{i=1}^{N}\left[\delta_i(1/a+\log t_i)-\log t_i\, t_i^{a}e^{\beta}\right]$$

$$\frac{\partial}{\partial\beta}\log L=\sum_{i=1}^{N}\left[\delta_i-t_i^{a}e^{\beta}\right]$$

$$\frac{\partial^2}{\partial a^2}\log L=-\sum_{i=1}^{N}\left[\delta_i/a^2+(\log t_i)^2t_i^{a}e^{\beta}\right]$$

$$\frac{\partial}{\partial a\partial\beta}\log L=-\sum_{i=1}^{N}\log t_i\, t_i^{a}e^{\beta}$$

$$\frac{\partial^2}{\partial\beta^2}\log L=-\sum_{i=1}^{N}t_i^{a}e^{\beta}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)}=\boldsymbol{\beta}'^{(m)}+(\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。この情報行列の逆行列の対角成分はパラメータの分散を与える。

上の推定法を用いると解は求まり、パラメータβの分散も計算できる。しかし、このままではパラメータbの分散は計算できない。そのため、解を上の方法で求め、求まった解を使ってパラメータbの分散を求めることにする。そのため、上の式をパラメータaとbでもう一度計算し直す。

$$\log L(a,b)=\sum_{i=1}^{N}\left[\delta_i\log\lambda(t_i)+\log S(t_i)\right]=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}b^{-a}\right)-t_i^{a}b^{-a}\right]$$

$$=\sum_{i=1}^{N}\left[\delta_i\left(\log a+(a-1)\log t_i-a\log b\right)-t_i^{a}b^{-a}\right]$$

$$\frac{\partial}{\partial a}\log L=\sum_{i=1}^{N}\left[\delta_i(1/a+\log t_i-\log b)-(\log t_i-\log b)t_i^a b^{-a}\right]$$

$$\frac{\partial}{\partial b}\log L=-\frac{a}{b}\sum_{i=1}^{N}\left[\delta_i-t_i^a b^{-a}\right]$$

$$\frac{\partial^2}{\partial a^2}\log L=-\sum_{i=1}^{N}\left[\delta_i(1/a^2)+(\log t_i-\log b)^2 t_i^a b^{-a}\right]$$

$$\frac{\partial^2}{\partial a\partial b}\log L=-\frac{1}{b}\sum_{i=1}^{N}\left[\delta_i-t_i^a b^{-a}-a(\log t_i-\log b)t_i^a b^{-a}\right]$$

$$\frac{\partial^2}{\partial b^2}\log L=\frac{a}{b^2}\sum_{i=1}^{N}\left[\delta_i-(a+1)t_i^a b^{-a}\right]$$

求まった解を上の式に代入し、情報行列を再度計算し直す。

$$\Im=-\begin{pmatrix}\partial^2\log L/\partial a^2 & \partial^2\log L/\partial a\partial b\\ \partial^2\log L/\partial a\partial b & \partial^2\log L/\partial b^2\end{pmatrix}$$

この情報行列の逆行列を用いてパラメータ b の分散を求める。この方法で実際に計算し、パラメータ a の分散を前の方法で計算した結果と比較してみると同一の値となっている。

③ 混合分布に基づく最尤推定

混合分布の最尤推定で、尤度 $L(\lambda)$ は以下で与えられる。

$$L(\lambda)=\prod_{i=1}^{N}f(t_i)^{\delta_i}S(t_i)^{1-\delta_i}$$

K 種混合分布では、それぞれの密度関数を $f_k(t)$、生存関数を $S_k(t)$ として、全体の密度関数と生存関数は以下となる。ここに、π_k は分布の重ね合わせの確率である。

$$f(t)=\sum_{k=1}^{K}\pi_k f_k(t),\qquad S(t)=\sum_{k=1}^{K}\pi_k S_k(t)$$

混合分布の最尤推定で、尤度 $L(\boldsymbol{\theta},\boldsymbol{\pi})$ は以下で与えられる。

$$L(\boldsymbol{\theta},\boldsymbol{\pi})=\prod_{i=1}^{N}f(t_i)^{\delta_i}S(t_i)^{1-\delta_i}=\prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k f_k(t_i)\right)^{\delta_i}\left(\sum_{k=1}^{K}\pi_k S_k(t_i)\right)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切り（死亡）データをそれぞれ $\delta_i=0,1$ としている。

対数尤度は以下となる。

$$\begin{aligned}
\log L(\boldsymbol{\theta},\boldsymbol{\pi}) &= \sum_{i=1}^{N}\left[\delta_i \log\sum_{k=1}^{K}\pi_k f_k(t_i) + (1-\delta_i)\log\sum_{k=1}^{K}\pi_k S_k(t_i)\right] \\
&= \sum_{i=1}^{N}\left[\delta_i \log\sum_{k=1}^{K}q_k^{(i)}\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + (1-\delta_i)\log\sum_{k=1}^{K}q_k^{(i)}\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right] \\
&\geq \sum_{i=1}^{N}\left[\sum_{k=1}^{K}q_k^{(i)}\delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + \sum_{k=1}^{K}q_k^{(i)}(1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}q_k^{(i)}\left[\delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + (1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}q_k^{(i)}\left[\delta_i \log f_k(t_i) + (1-\delta_i)\log S_k(t_i) + \log\pi_k - \log q_k^{(i)}\right]
\end{aligned}$$

上式の不等号は、$q_k^{(i)}$ の値によって、等号になることが知られている。

パラメータの推定には以下の手順 i) と ii) をパラメータ値が収束するまで繰り返す。このような 2 段階の推定法を EM アルゴリズムという。

i) パラメータ $q_k^{(i)}, \pi_k$ の最適化

この $q_k^{(i)}$ について、$\sum_{k=1}^{K}q_k^{(i)} = 1$ の条件をつけて右辺を最大化するために、ラグランジュの未定定数法を用いる。

$$\begin{aligned}
&\frac{\partial}{\partial q_k^{(i)}}\left[\log L(\boldsymbol{\theta},\boldsymbol{\pi}) - \sum_{i=1}^{N}\eta_i\left(\sum_{k=1}^{K}q_k^{(i)} - 1\right)\right] \\
&= \delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}} + (1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}} + 1 - \eta_i \\
&= \log\frac{\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i}}{q_k^{(i)}} - 1 - \eta_i = 0
\end{aligned}$$

これより、

$$q_k^{(i)} = e^{-(1+\eta_i)}\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i} = \frac{\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i}}{\sum_{k=1}^{K}\pi_k f_k(t_i)^{\delta_i} S_k(t_i)^{1-\delta_i}}$$

これを書き換えて、以下のようにすることもできる。

$$q_k^{(i)} = \pi_k f_k(t_i) \Big/ \sum_{k=1}^{K}\pi_k f_k(t_i) \quad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \pi_k S_k(t_i) \Big/ \sum_{k=1}^{K}\pi_k S_k(t_i) \quad \text{for } \delta_i = 0$$

この $q_k^{(i)}$ を群 k への帰属度という。

また、この尤度関数をパラメータ π_j で微分して 0 と置き、パラメータの推定を行うが、

$\sum_{k=1}^{K}\pi_k = 1$ の条件をつけるために、ラグランジュの未定定数法を用いる。

$$\frac{\partial}{\partial \pi_j}\left[\log L(\boldsymbol{\theta},\boldsymbol{\pi}) - \eta\left(\sum_{k=1}^{K}\pi_k - 1\right)\right] = \sum_{i=1}^{N} q_j^{(i)}/\pi_j - \eta = 0$$

より、

$$\pi_k = \frac{1}{\eta}\sum_{i=1}^{N} q_k^{(i)}, \qquad \sum_{k=1}^{K}\pi_k = \frac{1}{\eta}\sum_{k=1}^{K}\sum_{i=1}^{N} q_k^{(i)} = \frac{1}{\eta}\sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)} = \frac{1}{\eta}\sum_{i=1}^{N} 1 = \frac{N}{\eta} = 1$$

となり、以下の関係を得る。

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} q_k^{(i)}$$

ii) パラメータ $\boldsymbol{\theta}$ の推定

パラメータ $\boldsymbol{\theta}$ の最尤法による推定では、$q_k^{(i)}, \pi_k$ は i) の方法で求められた既知の定数として計算する。この部分の計算については具体的な関数形を用いて考える。

④ 混合指数分布に基づく最尤推定

指数分布の確率密度関数と生存関数の以下の具体的な表式を代入すると

$$f_k(t) = \lambda_k \exp(-\lambda_k t), \qquad S_k(t) = \exp(-\lambda_k t)$$

対数尤度は以下のようになる。

$$\begin{aligned}\log L(\boldsymbol{\lambda},\boldsymbol{\pi}) &\geq \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\left(\log\lambda_k - \lambda_k t_i\right) - (1-\delta_i)\lambda_k t_i + \log\pi_k - \log q_k^{(i)}\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log\lambda_k - \lambda_k t_i + \log\pi_k - \log q_k^{(i)}\right]\end{aligned}$$

これより、群 $\boldsymbol{k}$ への帰属度は以下となる。

$$q_k^{(i)} = \pi_k\lambda_k \exp(-\lambda_k t_i) \Big/ \sum_{k=1}^{K}\pi_k\lambda_k \exp(-\lambda_k t_i) \qquad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \pi_k \exp(-\lambda_k t_i) \Big/ \sum_{k=1}^{K}\pi_k \exp(-\lambda_k t_i) \qquad \text{for } \delta_i = 0$$

対数尤度を微分して、スコアベクトルを求め、それを 0 とする。

$$\frac{\partial}{\partial\lambda_j}\log L = \sum_{i=1}^{N} q_j^{(i)}\left(\delta_i/\lambda_j - t_i\right) = \frac{1}{\lambda_j}\sum_{i=1}^{N} q_j^{(i)}\delta_j - \sum_{i=1}^{N} q_j^{(i)} t_j = 0$$

これより、

$$\lambda_j = \sum_{i=1}^{N} q_j^{(i)}\delta_j \Big/ \sum_{i=1}^{N} q_j^{(i)} t_j$$

スコアをもう一度微分して、情報行列 $\mathfrak{I}$ に相当するものを作成する。

$$\mathfrak{I}_{jk} = -\frac{\partial^2}{\partial\lambda_j\partial\lambda_k}\log L = \frac{\delta_{jk}}{\lambda_j^2}\sum_{i=1}^{N} q_j^{(i)}\delta_i$$

この逆行列の対角成分は、推定値の分散を与える。

⑤ 混合 Weibull 分布に基づく最尤推定

K 種混合 Weibull では、以下となる。

$$f(t)=\sum_{k=1}^{K}\pi_k f_k(t)=\sum_{k=1}^{K}\pi_k a_k t^{a_k-1}b_k^{-a_k}\exp(-t^{a_k}b_k^{-a_k})$$
$$=\sum_{k=1}^{K}\pi_k a_k t^{a_k-1}e^{\beta_k}\exp(-t^{a_k}e^{\beta_k})$$
$$S(t)=\sum_{k=1}^{K}\pi_k S_k(t)=\sum_{k=1}^{K}\pi_k\exp(-t^{a_k}b_k^{-a_k})=\sum_{k=1}^{K}\pi_k\exp(-t^{a_k}e^{\beta_k})$$
$$S(t)=\sum_{k=1}^{K}\pi_k S_k(t)=\sum_{k=1}^{K}\pi_k\exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k))$$

混合 Weibull 分布の対数尤度は以下となる。

$$\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi})\geq\sum_{i=1}^{N}\sum_{k=1}^{K}q_k^{(i)}\Big[\delta_i\left(\log a_k+(a_k-1)\log t_i+\beta_k\right)$$
$$-t_i^{a_k}e^{\beta_k}+\log\pi_k-\log q_k^{(i)}\Big]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)}=\frac{\pi_k a_k t^{a_k-1}e^{\beta_k}\exp(-t^{a_k}e^{\beta_k})}{\sum_{k=1}^{K}\pi_k a_k t^{a_k-1}e^{\beta_k}\exp(-t^{a_k}e^{\beta_k})}\qquad\text{for}\quad\delta_i=1$$

$$q_k^{(i)}=\frac{\pi_k\exp(-t^{a_k}e^{\beta_k})}{\sum_{k=1}^{K}\pi_k\exp(-t^{a_k}e^{\beta_k})}\qquad\text{for}\quad\delta_i=0$$

ここで、$b_k^{-a_k}=e^{\beta_k}$ （$b_k=e^{-\beta_k/a_k}$ に相当）としている。

$$\frac{\partial}{\partial a_j}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi})=\sum_{i=1}^{N}q_j^{(i)}\Big[\delta_i\left(1/a_j+\log t_i\right)-\log t_i\,t_i^{a_j}e^{\beta_j}\Big]$$

$$\frac{\partial}{\partial\beta_j}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi})=\sum_{i=1}^{N}q_j^{(i)}\Big[\delta_i-t_i^{a_j}e^{\beta_j}\Big]$$

$$\frac{\partial^2}{\partial a_j\partial a_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi})=-\delta_{jk}\sum_{i=1}^{N}q_j^{(i)}\Big[\delta_i/a_j^2+(\log t_i)^2\,t_i^{a_j}e^{\beta_j}\Big]$$

$$\frac{\partial^2}{\partial a_j\partial\beta_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi})=-\delta_{jk}\sum_{i=1}^{N}q_j^{(i)}\log t_i\,t_i^{a_j}e^{\beta_j}$$

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi})=-\delta_{jk}\sum_{i=1}^{N}q_j^{(i)}t_i^{a_j}e^{\beta_j}$$

4) 比例ハザードモデル

比例ハザードモデルはハザート関数に対して、説明変数 $\mathbf{x} = {}^t(x_1, x_2, \cdots, x_p, 1)$ とパラメータ $\boldsymbol{\beta} = {}^t(\beta_1, \beta_2, \cdots, \beta_p, \beta_0)$ を用いて、以下の仮定を行う。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t) \exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i \beta_i + \beta_0$$

Cox の比例ハザードモデルでは $\lambda_0(t)$ と定数項 β_0 について議論しないが、Weibull ハザードモデルでは

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

として、時間に関して Weibull 分布のハザード関数を仮定する。

a) Cox の比例ハザードモデル

Cox の比例ハザードモデルでは、尤度関数に対して近似的な部分尤度関数を考えて処理を行う。その対数尤度は以下で与えられる[3]。

$$\log L'(\boldsymbol{\beta}) = \sum_{i=0}^{m-1}\left[\sum_{j \in D_i} {}^t\mathbf{x}_j\boldsymbol{\beta} - d_i \log \sum_{j \in R_i} \exp({}^t\mathbf{x}_j\boldsymbol{\beta})\right]$$

ここに、$\boldsymbol{\beta}$ は定数項を除いた偏回帰係数ベクトル、D_i は $t_i < t \le t_{i+1}$ で亡くなった個体の集合、R_i は時刻 t_i で生存が確認されている個体の集合である。これを最大化するようにニュートン・ラフソン法を使って $\boldsymbol{\beta}$ を求める。ここではそのための準備として以下の値を示しておく。

$$\mathbf{U} \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log L'(\boldsymbol{\beta}) = \sum_{i=1}^{m-1}\left[\sum_{i \in D_i} \mathbf{x}_j - d_i \sum_{j \in R_i} w_j \mathbf{x}_j \Big/ \sum_{j \in R_i} w_j\right]$$

$$\mathfrak{I} \equiv -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial {}^t\boldsymbol{\beta}} \log L'(\boldsymbol{\beta})$$

$$= \sum_{i=1}^{m-1} d_i \left[\sum_{j \in R_i} w_j \mathbf{x}_j \, {}^t\mathbf{x}_j \Big/ \sum_{j \in R_i} w_j - \sum_{j \in R_i} w_j \mathbf{x}_j \sum_{j \in R_i} w_j \, {}^t\mathbf{x}_j \Big/ \left(\sum_{j \in R_i} w_j\right)^2\right]$$

ここに $w_j = \exp({}^t\mathbf{x}_j\boldsymbol{\beta})$

この $\mathbf{U}$ をスコアベクトル、$\mathfrak{I}$ を情報行列という。$\boldsymbol{\beta}$ の推定値は以下の計算を繰り返して求める。

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

b) Weibull ハザードモデル

Weibull ハザードモデルは、ハザード関数に対して以下の仮定を行う。

$$\lambda(t)=\frac{f(t)}{S(t)}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}/b^{a}=at^{a-1}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})$$

通常の Wei 分布との関係は以下である。

$$b^{-a}=e^{\beta}\rightarrow\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\quad\left(\beta\rightarrow{}^{t}\mathbf{x}\boldsymbol{\beta}\equiv\sum_{i=1}^{p}x_i\beta_i+\beta_0\right)$$

これより、$b=\exp(-{}^{t}\mathbf{x}\boldsymbol{\beta}/a)$ であるから、$\mu\equiv E[T]=b\Gamma\left(1+1/a\right)$ より、

$$\eta\equiv{}^{t}\mathbf{x}\boldsymbol{\beta}=-a\log b=-a\log\left(\mu/\Gamma(1+1/a)\right)$$

となり、右辺が一般化線形モデルの連結関数となる。

この関係を用いて、密度関数と生存関数を求めると以下となる。

$$f(t)=at^{a-1}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\exp\left[-t^{a}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\right]$$

$$S(t)=\exp\left[-\left(t/b\right)^{a}\right]=\exp\left[-t^{a}b^{-a}\right]=\exp\left[-t^{a}\exp({}^{t}\mathbf{x}\boldsymbol{\beta})\right]$$

打ち切りデータと非打ち切り（死亡）データをそれぞれ $\delta_i=0,1$ と区別し、尤度を求めると以下となる。添え字 i について、ここでは個体の番号として使っている。

$$L(\alpha,\boldsymbol{\beta})=\prod_{i=1}^{N}f(t_i)^{\delta_i}S(t_i)^{1-\delta_i}$$

さらに、対数尤度は以下となる。

$$\begin{aligned}\log L(\alpha,\boldsymbol{\beta})&=\sum_{i=1}^{N}\left[\delta_i\log\lambda(t_i)+\log S(t_i)\right]\\&=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right)-t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right]\\&=\sum_{i=1}^{N}\left[\delta_i\left(\log a+(a-1)\log t_i+{}^{t}\mathbf{x}_i\boldsymbol{\beta}\right)-t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right]\end{aligned}$$

対数尤度を微分してスコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ を求めると以下となる。

$$\boldsymbol{\beta}'=\begin{pmatrix}a\\\boldsymbol{\beta}\end{pmatrix},\quad\mathbf{U}=\begin{pmatrix}\partial\log L/\partial a\\\partial\log L/\partial\boldsymbol{\beta}\end{pmatrix},$$

$$\mathfrak{I}=-\begin{pmatrix}\partial^2\log L/\partial a^2 & \partial^2\log L/\partial a\partial{}^{t}\boldsymbol{\beta}\\\partial^2\log L/\partial a\partial\boldsymbol{\beta} & \partial^2\log L/\partial\boldsymbol{\beta}\partial{}^{t}\boldsymbol{\beta}\end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a}\log L=\sum_{i=1}^{N}\left[\delta_i(1/a+\log t_i)-\log t_i t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right]$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\log L=\sum_{i=1}^{N}\left[\delta_i\mathbf{x}_i-\mathbf{x}_i t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right]$$

$$\frac{\partial^2}{\partial a^2}\log L=\sum_{i=1}^{N}\left[-\delta_i/a^2-(\log t_i)^2 t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})\right]$$

$$\frac{\partial^2}{\partial a\partial\boldsymbol{\beta}}\log L=-\sum_{i=1}^{N}(\log t_i)\mathbf{x}_i t_i^{a}\exp({}^{t}\mathbf{x}_i\boldsymbol{\beta})$$

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial^t\boldsymbol{\beta}}\log L=-\sum_{i=1}^{N}\mathbf{x}_i\,{}^t\mathbf{x}_i t_i^a\exp({}^t\mathbf{x}_i\boldsymbol{\beta})$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)}=\boldsymbol{\beta}'^{(m)}+(\Im^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

求められたパラメータを使って、個人の予想寿命を以下のように求めることができる。

$$\mu\equiv E[T]=b\,\Gamma\left(1+1/a\right)=\exp(-{}^t\mathbf{x}\boldsymbol{\beta}/a)\Gamma\left(1+1/a\right)$$

この値を実際の寿命と比較することで相関係数等を求めることもできる。

c）混合 Weibull ハザードモデル

K 種混合 Weibull ハザードモデルでは以下を仮定する。

$$f(t)=\sum_{k=1}^{K}\pi_k f_k(t)=\sum_{k=1}^{K}\pi_k a_k t^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k))$$

$$S(t)=\sum_{k=1}^{K}\pi_k S_k(t)=\sum_{k=1}^{K}\pi_k\exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k))$$

通常の Weibull 分と比較すると、ここでは以下を仮定している。

$$b_k^{-a_k}=e^{\beta_k}\rightarrow\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\qquad(\beta_k\rightarrow{}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k\equiv\sum_{i=1}^{p}x_i\beta_i+\gamma_k)$$

これより、$b_k\rightarrow\exp\left[-({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)/a_k\right]$ であるから、

$$\mu\equiv E[T]=\sum_{k=1}^{K}\pi_k b_k\Gamma\left(1+1/a_k\right)$$

となる。連結関数については、以下の関数の逆関数である。

$$\mu=\sum_{k=1}^{K}\pi_k\exp\left[-({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)/a_k\right]\Gamma\left(1+1/a_k\right)$$

$$=\sum_{k=1}^{K}\pi_k\exp\left[-(\eta+\gamma_k)/a_k\right]\Gamma\left(1+1/a_k\right)$$

混合 Weibull 分布の対数尤度は以下となる。

$$\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\pi})\geq\sum_{i=1}^{N}\sum_{k=1}^{K}q_k^{(i)}\Big[\delta_i\left(\log a_k+(a_k-1)\log t_i+{}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k\right)$$

$$-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k)+\log\pi_k-\log q_k^{(i)}\Big]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)}=\frac{\pi_k a_k t_i^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k))}{\sum_{k=1}^{K}\pi_k a_k t_i^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k))}\qquad\text{for}\ \ \delta_i=1$$

$$q_k^{(i)} = \frac{\pi_k \exp(-t_i^{a_k} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k))}{\sum_{k=1}^{K} \pi_k \exp(-t_i^{a_k} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k))} \qquad \text{for} \quad \delta_i = 0$$

ここで、$b_k^{-a_k} \to \exp({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)$ （$b_k \to \exp\left[-({}^t\mathbf{x}\boldsymbol{\beta} + \gamma_k)/a_k\right]$）としている。

対数尤度を微分してスコアベクトル$\mathbf{U}$と情報行列$\mathfrak{I}$を求めると以下となる。

$$\boldsymbol{\beta}' = \begin{pmatrix} \mathbf{a} \\ \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \partial \log L/\partial \mathbf{a} \\ \partial \log L/\partial \boldsymbol{\gamma} \\ \partial \log L/\partial \boldsymbol{\beta} \end{pmatrix},$$

$$\mathfrak{I} = -\begin{pmatrix} \partial^2 \log L/\partial \mathbf{a}\partial^t\mathbf{a} & \partial^2 \log L/\partial \mathbf{a}\partial^t\boldsymbol{\gamma} & \partial^2 \log L/\partial \mathbf{a}\partial^t\boldsymbol{\beta} \\ \partial^2 \log L/\partial \boldsymbol{\gamma}\partial^t\mathbf{a} & \partial^2 \log L/\partial \boldsymbol{\gamma}\partial^t\boldsymbol{\gamma} & \partial^2 \log L/\partial \boldsymbol{\gamma}\partial^t\boldsymbol{\beta} \\ \partial^2 \log L/\partial \boldsymbol{\beta}\partial^t\mathbf{a} & \partial^2 \log L/\partial \boldsymbol{\beta}\partial^t\boldsymbol{\gamma} & \partial^2 \log L/\partial \boldsymbol{\beta}\partial^t\boldsymbol{\beta} \end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a_j}\log L = \sum_{i=1}^{N} q_j^{(i)} \left[\delta_i \left(1/a_j + \log t_i\right) - \log t_i \, t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j) \right]$$

$$\frac{\partial}{\partial \gamma_j}\log L = \sum_{i=1}^{N} q_j^{(i)} \left[\delta_i - t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j) \right]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}\log L = \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)} \left[\delta_i \mathbf{x}_i - \mathbf{x}_i t_i^{a_k} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j) \right]$$

$$\frac{\partial^2}{\partial a_j \partial a_k}\log L = -\delta_{jk} \sum_{i=1}^{N} q_j^{(i)} \left[\delta_i / a_j^2 + (\log t_i)^2 \, t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j) \right]$$

$$\frac{\partial^2}{\partial a_j \partial \gamma_k}\log L = -\delta_{jk} \sum_{i=1}^{N} q_j^{(i)} \log t_i \, t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial a_j \partial \boldsymbol{\beta}}\log L = -\sum_{i=1}^{N} q_j^{(i)} \log t_i \, \mathbf{x}_i t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial \gamma_j \partial \gamma_k}\log L = -\delta_{jk} \sum_{i=1}^{N} q_j^{(i)} t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial \gamma_j \partial \boldsymbol{\beta}}\log L = -\sum_{i=1}^{N} q_j^{(i)} \mathbf{x}_i t_i^{a_j} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial^t \boldsymbol{\beta}}\log L = -\sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)} \mathbf{x}_i \, {}^t\mathbf{x}_i t_i^{a_k} \exp({}^t\mathbf{x}_i\boldsymbol{\beta} + \gamma_k)$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)} = \boldsymbol{\beta}'^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

参考文献

[1] 打波守, Excel で学ぶ生存時間解析, オーム社, 2005.

[2] 柳井晴夫, 高木廣文編著, 多変量解析ハンドブック, 現代数学社, 1986.

[3] Annete J. Dobson, 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

２４．リッジ回帰分析他

この章ではこれまでに作成した、品質管理の異常検知プログラム（リファレンスマニュアルーＯＲ２ー）の中から、リッジ回帰分析、PLS 回帰分析を抜き出し、統計分析に使い易い形に変更し、新たに主成分回帰分析を追加して、重回帰分析における多重共線性の回避法について考察した。

重回帰分析では、入力変数が多くその値が似通っている場合に、多重共線性の問題が発生する可能性があり、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。

24.1　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 [1]、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検討して行く。

メニュー［分析ー多変量解析等ー予測手法ーリッジ回帰分析等］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。

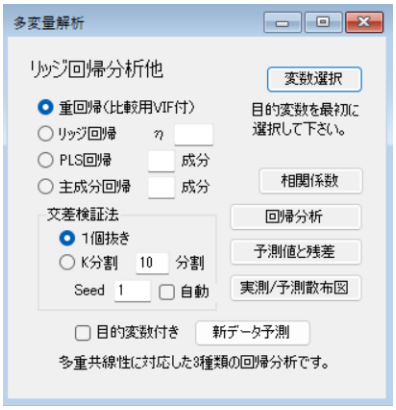


図 1　分析実行メニュー

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。

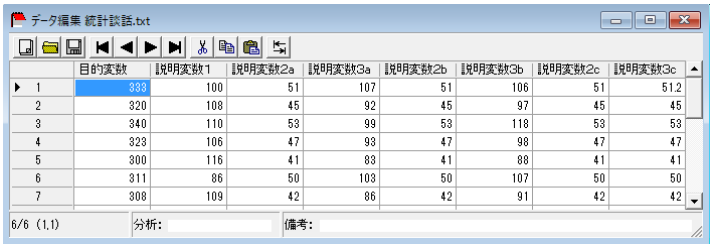
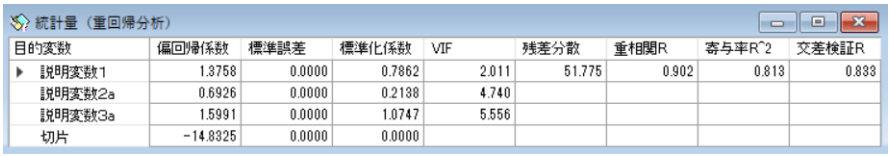


図 2　データ

このデータは、目的変数と説明変数１を共通として、残りの変数は、説明変数２a と３a、説明変数２b と３b、説明変数２c と３c という組み合わせで利用する。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、後は全く同じデータで、相関はほぼ１である。以後、説明変数の取り方は、この a, b, c を使って指定する。

「重回帰分析」ラジオボックスを選択し、目的関数と説明変数１を選んだ次に、説明変数２と３を a, b, c の順番に選んで、「回帰分析」ボタンをクリックする。結果を図 3a、図 3b、図 3c に示す。


統計量（重回帰分析）

目的変数	偏回帰係数	標準誤差	標準化係数	VIF	残差分散	重相関R	寄与率R^2	交差検証R
説明変数1	1.3758	0.0000	0.7862	2.011	51.775	0.902	0.813	0.833
説明変数2a	0.6926	0.0000	0.2138	4.740				
説明変数3a	1.5991	0.0000	1.0747	5.556				
切片	-14.8325	0.0000	0.0000					

図 3a　ほぼ問題のない結果


統計量（重回帰分析）

目的変数	偏回帰係数	標準誤差	標準化係数	VIF	残差分散	重相関R	寄与率R^2	交差検証R
説明変数1	0.9422	0.2472	0.5385	1.711	86.974	0.829	0.687	0.709
説明変数2b	6.1250	1.1779	1.8905	11.337				
説明変数3b	-1.4149	0.5366	-0.9270	10.602				
切片	67.8303	42.7350	0.0000					

図 3b　問題のある結果


統計量（重回帰分析）

目的変数	偏回帰係数	標準誤差	標準化係数	VIF	残差分散	重相関R	寄与率R^2	交差検証R
説明変数1	0.9463	0.3513	0.5408	1.712	101.151	0.797	0.636	0.728
説明変数2c	-57.2248	53.6122	-17.6629	11630.778				
説明変数3c	60.4874	53.6077	18.6736	11633.330				
切片	58.8225	60.6105	0.0000					

図 3c　完全に問題のある結果

図 3c を見ると、寄与率は高くなっているが、偏回帰係数の値が非常に大きくなって正と負で相殺している。これは、新しいデータで、説明変数２と３の値が少し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である。実際、図 3b については交差検証（1 個抜き検証）の値はかなり下がっている。

次に、特に問題のある b と c の場合について、3 つの分析を比較する。まず、b の場合、3 つの分析結果を図 4、図 5、図 6 に示す。


統計量（リッジ回帰分析）

目的変数	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	0.8813	0.5037	102.186	0.825	0.680	0.709	0.460
説明変数2b	5.0419	1.5562				dη	0.0100
説明変数3b	-0.9581	-0.6277					
切片	79.1583	0.0000					

図 4　b についてのリッジ回帰分析結果


統計量（PLS回帰分析）

目的変数	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	0.9044	0.5168	1.550	128.443	0.733	0.537	0.645	2
説明変数2b	2.0689	0.6386	1.550					
説明変数3b	0.4873	0.3192						
切片	71.3585	0.0000						

図５　b についての PLS 回帰分析


統計量（主成分回帰分析）

目的変数	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	0.7920	0.4526	1.000	140.951	0.702	0.492	0.605	2
説明変数2b	1.3960	0.4309	1.000					
説明変数3b	0.7265	0.4760						
切片	90.4773	0.0000						

図６　b についての主成分回帰分析

ここに、PLS 回帰分析と主成分回帰分析の変数変換後の独立成分の数（以後自由度と呼ぶ）は 2 にしている。この結果を見ると、リッジ回帰の交差検証 R の値が最も高くなっているが、他のデータからの予測を考えると、偏回帰係数の値が少し大きくなっていることが問題である。最良の分析は PLS 回帰分析ではないかと思われる。主成分回帰分析の r-VIF の値が 1 になっているのは、主成分得点の相関が 0 になっているからである。

次に、c の場合、3 つの分析結果を図 7、図 8、図 9 に示す。


統計量（リッジ回帰分析）

目的変数	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	0.8700	0.4972	117.143	0.779	0.607	0.713	1.380
説明変数2c	1.5025	0.4638				dη	0.0100
説明変数3c	1.5995	0.4938					
切片	75.1410	0.0000					

図７　c についてのリッジ回帰分析結果


統計量（PLS回帰分析）

目的変数	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	0.9409	0.5377	1.504	109.025	0.779	0.607	0.715	2
説明変数2c	1.6277	0.5024	1.504					
説明変数3c	1.6399	0.5063						
切片	60.0052	0.0000						

図８　c についての PLS 回帰分析


統計量（主成分回帰分析）

目的変数	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	0.9409	0.5377	1.000	109.026	0.779	0.607	0.715	2
説明変数2c	1.6342	0.5044	1.000					
説明変数3c	1.6334	0.5043						
切片	60.0076	0.0000						

図９　c についての主成分回帰分析

ここに、PLS 回帰分析と主成分回帰分析の変数変換後の独立成分の数（以後自由度と呼ぶ）は前と同様 2 にしている。この結果を見ると、PLS 回帰と主成分回帰の結果はほぼ同一である。

多重共線性が問題にならない場合は、もちろん重回帰分析を使うが、PLS 回帰分析と主成分回帰分析で、変換後の変数数を元の変数数に設定すると、当然重回帰分析と同じ結果を

得る。リッジ回帰分析については、多少違う結果が出る。

最後に、「予測値と残差」ボタンと「実測/予測散布図」ボタンを押すと、重回帰分析のメニューにあるような予測値と残差のグリッド出力とグラフ出力が得られる。特に新しいものではないので、ここでは省略する。

このプログラムを統計的なモデル作成として使うには問題がある。プログラムの中では偏回帰係数の検定が行われていないからである。例えば、PLS 回帰などの場合、変換後の説明変数で重回帰分析を行った際の係数の検定はできるが、元の変数の回帰式の係数の検定については､作者の理解不足でどのように対処すべきか不明である｡これを読まれた方で、詳しい方に教えていただければ幸いです。

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、これは一度リッジ回帰分析他を実行した後、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 R^2 の値で示してくれる。

24.2　リッジ回帰等の理論

この節では、異常検知プログラムの理論の部分と重複するが、比較を分かり易くするために重回帰分析、リッジ回帰分析、PLS 回帰分析の理論を再掲する。主成分回帰分析の理論については PLS 回帰分析の理論と大きく変わらない。

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 $x_{i\lambda}$ （$i = 1,2,\cdots,p$）とし、それらの関係を ε_λ を誤差項として以下とする。

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

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、

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

である。パラメータは以下で与えられる。

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

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 指標を利用している。

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

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。

$$VIF_i \simeq 10 \Leftrightarrow r_i \simeq 0.95$$

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 つ以上の変数間に相関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) 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。（注：$\mathbf{X}, \mathbf{y}$ の定義は前節と同じ）

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

これは、以下を最小化する解でもある。

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

ここでパラメータηの値は以下のようにして求められる。λ番目の個体を抜いた 1 個抜き交差検証のリッジ回帰パラメータを$\mathbf{b}'^{(-\lambda)}$とすると、そのときの平均 2 乗誤差$e(\eta)$は以下で与えられる。

$$e(\eta) = \frac{1}{N}\sum_{\lambda=1}^{N}(\tilde{y}_\lambda - \sum_{i=1}^{p}\tilde{x}_{i\lambda} b_i'^{(-\lambda)})^2 , \qquad \tilde{y}_\lambda = y_\lambda - \bar{y} , \ \tilde{x}_{i\lambda} = x_{i\lambda} - \bar{x}$$

これは、近似的に以下のように書くこともできる[1]。

$$e(\eta) = \frac{1}{N}\,{}^t\mathbf{AA}$$

ここに、

$$\mathbf{A} = diag(\mathbf{I}-\mathbf{H})^{-1}(\mathbf{I}-\mathbf{H})\mathbf{y} \ , \quad \mathbf{H}(N\times N) = \mathbf{X}({}^t\mathbf{XX} + N\eta\mathbf{I})^{-1}\,{}^t\mathbf{X}$$

また、$diag(\mathbf{I}-\mathbf{H})^{-1}$は対角要素が$(1-(\mathbf{H})_{ii})^{-1}$となる対角行列である。運用上はパラメータ$\eta$の値を変化させて、この$e(\eta)$が最小になるようなパラメータ$\eta$を選ぶ。

もう少し安全性を考えて、以下の一般化交差確認検証法と呼ばれる方法から与えられる誤差$e_{GCV}(\eta)$を最小化する場合もある。

$$e_{GCV}(\eta) = \frac{1}{N}\,{}^t\mathbf{A}'\mathbf{A}'$$

ここに、$\mathbf{A}' = (\mathbf{I} - \mathbf{H})\mathbf{y}/[1 - tr\mathbf{H}/N]$ である。我々のプログラムでは前者の判定法を利用している。

上で述べた近似的方法は OR の異常検知では使っているが、ここでは素直にデータを１つずつ抜きながら計算を行っている。そのため、データ数が多くなると近似的方法に比べて時間がかかる。また、例えばデータ数が 10,000 を超えるような場合では、近似的方法でも計算時間が非常に長くかかる。そのような場合、プログラムには、効果的な k 分割法も加えられている。これは、データをほぼ数の等しい k 個の組に分類し、1 つの組をテストデータに、他の k-1 個の組を解析データとして予測値などを求め、実測値と比較する方法である。これをすべての組がテストデータになるように繰り返し、予測の精度を残差分散の平均値で求めるものである。k 分割法の分割数がデータ数と等しい場合には 1 個抜き交差検証法と同じ結果になる。

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。そのため交差検証は必須である。

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。

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

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。詳細は後に示す。

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、

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

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$\tilde{\mathbf{b}}$は以下となる。

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

また、回帰係数は、以下で与えられる。

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

多重共線性の改善の程度については、変数を **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。

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

最後に行列 $\mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_r)$ の決定法について述べる。この行列の 1 列目 $\mathbf{u}_1$ は $\mathbf{Xu}_1$ が最も **y** の方向に向くように、以下のように求める。

$$L_1 = {}^t\mathbf{y}\mathbf{X}\mathbf{u}_1 - \mu_1({}^t\mathbf{u}_1\mathbf{u}_1 - 1) \quad \rightarrow \qquad \text{最大化}$$

この解は以下で与えられる。

$$\mathbf{u}_1 = {}^t\mathbf{X}\mathbf{y} / \left\| {}^t\mathbf{X}\mathbf{y} \right\|$$

次の $\mathbf{u}_2$ については、**X** から $\mathbf{u}_1$ 方向の成分を取り除き、以下のように求める。

$$L_2 = {}^t\mathbf{y}(\mathbf{X} - \mathbf{d}_1\,{}^t\mathbf{d}_1\mathbf{X})\mathbf{u}_2 - \mu_2({}^t\mathbf{u}_2\mathbf{u}_2 - 1) \quad \rightarrow \qquad \text{最大化}$$

ここに、$\mathbf{d}_1 = \mathbf{Xu}_1 / \left\| \mathbf{Xu}_1 \right\|$ である。確かに $\mathbf{X} - \mathbf{d}_1\,{}^t\mathbf{d}_1\mathbf{X}$ は $\mathbf{u}_1$ 方向の成分を取り除いている。

$$(\mathbf{X} - \mathbf{d}_1\,{}^t\mathbf{d}_1\mathbf{X})\mathbf{u}_1 = \mathbf{Xu}_1 - \mathbf{Xu}_1\,{}^t(\mathbf{Xu}_1)\mathbf{Xu}_1 / \left\| \mathbf{Xu}_1 \right\|^2 = \mathbf{0}$$

この解は以下で与えられる。

$$\mathbf{u}_2 = {}^t(\mathbf{X} - \mathbf{d}_1\,{}^t\mathbf{d}_1\mathbf{X})\mathbf{y} / \left\| {}^t(\mathbf{X} - \mathbf{d}_1\,{}^t\mathbf{d}_1\mathbf{X})\mathbf{y} \right\|$$

このベクトル $\mathbf{u}_2$ は $\mathbf{u}_1$ と直交する。

$${}^t\mathbf{u}_1\mathbf{u}_2 \propto {}^t\mathbf{u}_1\,{}^t(\mathbf{X} - \mathbf{d}_1\,{}^t\mathbf{d}_1\mathbf{X})\mathbf{y} = {}^t(\mathbf{Xu}_1 - \mathbf{Xu}_1\,{}^t(\mathbf{Xu}_1)\mathbf{Xu}_1 / \left\| \mathbf{Xu}_1 \right\|^2)\mathbf{y} = 0$$

これを続けると、k 番目の係数ベクトル $\mathbf{u}_k$ は以下のように求められることが分かる。

$$L_k = {}^t\mathbf{y}\left(\mathbf{X} - \sum_{i=1}^{k-1} \mathbf{d}_i\,{}^t\mathbf{d}_i\mathbf{X} \right)\mathbf{u}_k - \mu_k({}^t\mathbf{u}_k\mathbf{u}_k - 1) \quad \rightarrow \qquad \text{最大化}$$

$$\mathbf{u}_k = {}^t\left(\mathbf{X} - \sum_{i=1}^{k-1} \mathbf{d}_i\,{}^t\mathbf{d}_i\mathbf{X} \right)\mathbf{y} \Bigg/ \left\| {}^t\left(\mathbf{X} - \sum_{i=1}^{k-1} \mathbf{d}_i\,{}^t\mathbf{d}_i\mathbf{X} \right)\mathbf{y} \right\|$$

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または同じことであるが、残差分散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。

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。

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

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。

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

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。

参考文献

[1] 井出剛, 入門機械学習による異常検知, コロナ社, 2015.

[2] ホームページ http://www.heisei-u.ac.jp/ba/fukui/analysis.html 内のサンプルデータ Samples.zip 内のファイル

２５．直交表実験計画法とコンジョイント分析

25.1　直交表実験計画法とは

通常の分散分析では、水準を設定する要因（配置）の数によって、実験回数が幾何級数的に増加し、現実的な実験計画を行うことが困難になる。この問題に対して、比較的少ない実験回数で、各要因の効果や要因間の交互作用の効果を測定できるようにした方法が直交表実験計画法である[1]。

直交表は、ある要因の１つの水準に対して、他の要因の各水準が同数だけ現れるように配置した表で、品質管理の田口メソッドでは中心となる表である。我々はこの表を用いて要因の水準を設定し、他の要因の影響は同数であるのですべて同じとみなして、ある要因の水準ごとの測定値の差を求める。

例えば、2 水準を持つ要因 a,b,c の直交表は表 1 である。但し、ここでは計算が分かり易いように水準は 0,1 で表している。通常は 1 を足して 1,2 を利用する。

表１　L8(2^7)直交表（0,1 表示）

	a	b	c	ab	ac	bc	abc
1	0	0	0	0	0	0	0
2	0	0	1	0	1	1	1
3	0	1	0	1	0	1	1
4	0	1	1	1	1	0	0
5	1	0	0	1	1	0	1
6	1	0	1	1	0	1	0
7	1	1	0	0	1	1	0
8	1	1	1	0	0	0	1

ここで a 列が 0 または 1 の場合の他の列の 0 と 1 の数は 2 つずつである。a,b,c 列は 0,1 を 2 つに分けるように作られている。その他の列、例えば ab 列の値は a 列と b 列の同じ行の値から、ab=(a+b) mod 2（mod 2 は 2 で割った余り）の式を使って求める。また、abc の列は、ab 列と c 列から、abc=(ab+c) mod 2、または a 列と bc 列から、abc=(a+bc) mod 2 として求められる。また、aa=bb=cc=0 という性質から、abbc, aab, bcc 等はそれぞれ ac, b, b となり、表で表される組み合わせ以上はない。

要因 a と要因 b の相互作用は、ab 列に表れる。ab の 0 には、(a,b)の(0,0)と(1,1)が、1 には、(0,1)と(1,0)が対応している。それぞれ、a,b の 0,1 が違った組み合わせで 1 つずつ表れている。どこかに特別に強め合う組み合わせが存在する場合、ab の 0 と 1 の状態でデータの平均（または合計）は異なるはずである。それがなければ、2 つの状態は似た平均を持つ。これにより a,b 要因の交互作用が明らかになるということになる。

次に 3 水準を持つ要因 a,b の直交表を考えてみる。これは L9(3^4)直交表として、表 2 で与えられる。ここでも要素は 0,1,2 で与えられているが、実際は 1 を足して、1,2,3 で表示される。

表 2　L9(3^4)直交表

	A	b	ab	abb
1	0	0	0	0
2	0	1	1	2
3	0	2	2	1
4	1	0	1	1
5	1	1	2	0
6	1	2	0	2
7	2	0	2	2
8	2	1	0	1
9	2	2	1	0

この場合も、b 列は a 列の 0,1 を 3 つに分けるように作られている。また、ab 列は a,b 列を使って、ab=(a+b) mod 3 で作られ、abb 列は abb=(ab+b) mod 3 で作られている。ここで、aab や aabb も計算できるが、aaa=bbb=0 より、(aab+abb) mod 3=0, (aabb+ab) mod 3=0 となり、表の値と独立ではなくなる。

交互作用は、ab と abb の両方に表れる。例えば ab が 0 の場合(a,b)は(0,0),(1,2),(2,1)、1 の場合(0,1),(1,0),(2,2)、2 の場合(0,2),(2,0),(1,1)となり、a,b の 0,1 が違った組み合わせで 1 つずつ表れている。abb についても、abb が 0 の場合(a,b)は(0,0),(1,1),(2,2)、1 の場合(0,2),(1,0),(2,1)、2 の場合(0,1),(1,2),(2,0)となり、同様に a,b の 0,1 が違った組み合わせで 1 つずつ表れている、よって要因単独の効果は相殺され、どこかに特別に強め合う組み合わせが存在する場合にのみ、ab と abb の 0,1,2 の状態でデータの平均は異なるはずである。それがなければ、3 つの状態は似た平均を持つ。これにより a,b 要因の交互作用が明らかになるということになる。

これらのことを前提にして、r 水準直交表実験計画法の理論を考えてみる。今実験 i のデータを x_i （$i=1,\cdots,n$）とする。 j 列の水準値が a のとき、 j 列の水準 a による平均 μ からのずれを c_{ja} として、データ x_i は以下のように書けると仮定する。

$$x_i = \mu + \sum_{j=1}^{p} c_{j[i]} + \varepsilon_i \qquad \text{ここに } \varepsilon_i \sim N(0,\sigma^2)$$

ここで $j[i]$ は j 列の i 番目のデータの水準を表す。これが a の場合、列 j について $[i]=a$ である。ここで、$c_{j[i]}$ は単独の効果の場合もあるし、交互作用の一部である場合もある。水準による影響 c_{ja} については、以下を仮定しておく。

$$\sum_{a=1}^{r} c_{ja} = 0$$

また直交表の性質から、すべての i を取ることはすべての水準を同数（m 回とする）取ることになるので、以下の関係も与えられる。

$$\sum_{i=1}^{n} c_{j[i]} = 0, \quad n = mr$$

r 水準直交表の場合、交互作用は自由度が $(r-1)\times(r-1)$ となることから、$r-1$列で表すことができる。例えば 2 水準直交表で 1 列、3 水準直交表で 2 列である。

このデータを使って列 j の値が a である合計を T_{ja} とする。

$$T_{ja}=\sum_{[i]=a\ for\ j} x_i \sim N(m(\mu+c_{ja}),m\sigma^2)$$

これを使うとデータの合計 T は以下のように書ける。

$$T=\sum_{i=1}^{N} x_i=\sum_{a=1}^{r} T_{ja} \sim N(n\mu,n\sigma^2)$$

この関係から $c_{ja}=0$ のときには、

$$S_j=m\sum_{a=1}^{r}(T_{ja}/m-T/n)^2=\sum_{a=1}^{r}T_{ja}^2/m-T^2/n,\qquad S_j/\sigma^2 \sim \chi^2_{r-1}$$

さらに、独立な列のときには、

$$(S_{j_1}+S_{j_2})/\sigma^2 \sim \chi^2_{2(r-1)}$$

特に交互作用を表す複数列は独立であり、S_j を合計して検定を行えばよい。

実際の検定では、要因や交互作用を割り当てていない列を $j_1, j_2, \cdots, j_d$ とすると、

$$S_e/\sigma^2=\sum_{l=1}^{d} S_{j_l}\Big/\sigma^2 \sim \chi^2_{d(r-1)}$$

の性質を用いて、列 j による寄与について検定する。

$$F_j=\frac{S_j/(r-1)}{S_e/d(r-1)} \sim F_{r-1,d(r-1)}$$

同様に、列 j_1 と j_2 による寄与についての検定は以下を用いる。

$$F_{j_1+j_2}=\frac{(S_{j_1}+S_{j_2})/(2r-2)}{S_e/d(r-1)} \sim F_{2r-2,d(r-1)}$$

これによって交互作用の検定も可能となる。

また、j 列の水準 a について、$c_{ja}=0$ のときは、$1/n_e=1/m-1/mr$ として、

$$\begin{aligned}T_{ja}/m-T/mr&=\frac{1}{m}(1-1/r)\sum_{[i]=a}\varepsilon_i-\frac{1}{mr}\sum_{[i]\neq a}\varepsilon_i\\&\sim N\left(0,\frac{(r-1)^2}{m^2r^2}m\sigma^2+\frac{1}{m^2r^2}(mr-m)\sigma^2\right)\\&\sim N\left(0,\frac{(r-1)}{mr}\sigma^2\right)\sim N\left(0,\sigma^2/n_e\right)\end{aligned}$$

よって σ^2 の推定量 $V_e=S_e/d(r-1)$ を用いて以下となり、

$$\frac{(T_{ja}/m-T/mr)\sqrt{n_e}}{\sqrt{V_e}}=\frac{T_{ja}/m-T/mr}{\sqrt{V_e/n_e}} \sim t_{d(r-1)}$$

水準 a におけるデータの区間推定も可能である。

多水準法と擬水準法

多水準法や擬水準法は異なった水準数の要因を混ぜ合わせるときに利用される。多水準法は、例えば 2 水準の要因と 4 水準の要因が混在する場合に使われる。4 水準の要因の自由度は 3 であり、2 水準の要因の自由度は 1 であるため、4 成分を表すのに、2 水準の直交表の 3 列を利用する。その 3 列は、例えば a, b, ab のように、2 つの独立な列とその相互作用の列になる。また、4 水準と 2 水準の交互作用では、同じく 3 列を利用する。これはすでに見たように、3 つの列の分散 $S_{j_1}, S_{j_2}, S_{j_3}$ を合計することにより要因及び交互作用の効果を検定することができる。

擬水準法は多水準法のように直交表を変更することなく割り当てできる場合以外に使われる。例えば、3 水準直交表の中に 2 水準の要因を割り当てる場合などに使われる。擬水準法では 3 水準の 1,2,3 の 3 つの水準のうちの 1 つ、例えば 3 を 1 に置き直すことが行われる。例えば今 j 列にこの置き換えを適用したとする。これにより、他の要因のある水準から見た場合、 j 列は常に水準 1 が水準 2 の 2 倍生じることになるが、この条件は常に同じであり、平均を比較するので見ている要因の値への影響は同一である。この場合、列 j の分散 S_j が以下のように拡張されるだけである。

$$S_j = \sum_{a=1}^{r} m_{ja}(T_{ja}/m_{ja} - T/n)^2 = \sum_{a=1}^{r} T_{ja}^2/m_{ja} - T^2/n, \qquad \sum_{a=1}^{r} m_a = n$$

$$S_j/\sigma^2 \sim \chi_{r-1}^2$$

しかし、交互作用についてはこのようにはならない。表 2 の直交表を要因 a を擬水準として書き直したものが表 3 の直交表である。

表 3　擬水準直交表

	A	b	ab	abb
1	0	0	0	0
2	0	1	1	2
3	0	2	2	1
4	1	0	1	1
5	1	1	2	0
6	1	2	0	2
7	0	0	0	0
8	0	1	1	2
9	0	2	2	1

この直交表で見ると、例えば ab が 0 の場合(a,b)は(0,0),(1,2),(0,0)、1 の場合(0,1),(1,0),(0,1)、2 の場合(0,2),(1,1),(0,2)となり、a,b 各水準が同じ数だけ現れていない。abb についても、abb が 0 の場合(a,b)は(0,0),(1,1),(0,0)、1 の場合(0,2),(1,0),(0,2)、2 の場合(0,1),(1,2),(0,1)となり、同じく a,b の各水準が同じ数だけ現れていない。これでは要因の単独の効果が除去されず交互作用は検証できない。そのため以下のような処理になる。まず、2 列の全分散 $S_{j_1 j_2}$ を以下のようにする。

$$S_{j_1 j_2} = \sum_{a=1}^{r_1}\sum_{b=1}^{r_2} T^2_{j_1 a j_2 b} \Big/ m_{j_1 a j_2 b} - T^2/n\,, \qquad \sum_{a=1}^{r_1}\sum_{b=1}^{r_2} m_{j_1 a j_2 b} = n$$

$$S_{j_1 j_2} \Big/ \sigma^2 \sim \chi^2_{r_1 r_2 - 1}$$

ここに列 j_1 の水準数を r_1、列 j_2 の水準数を r_2、$j_1 a j_2 b$ のデータ数を $m_{j_1 a j_2 b}$ としている。交互作用はこれから単独の分散を引いたものとして定義する。

$$S_{j_1 \times j_2} = S_{j_1 j_2} - S_{j_1} - S_{j_2}$$

$$S_{j_1 \times j_2} \Big/ \sigma^2 \sim \chi^2_{(r_1-1)(r_2-1)}$$

これによって交互作用の効果が計算できる。

25.2　プログラムの利用法

メニュー［分析－多変量解析他－実験計画法－直交表実験計画法］を選択すると図 1 のような直交表実験計画法の実行画面が表示される。

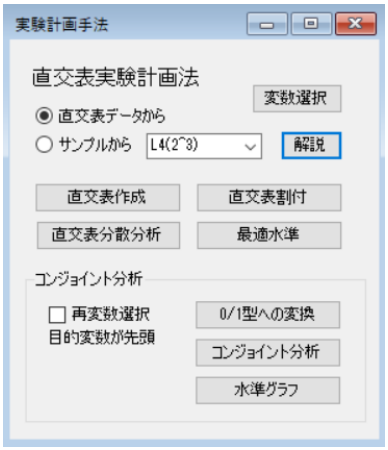


図 1　直交表実験計画法実行画面

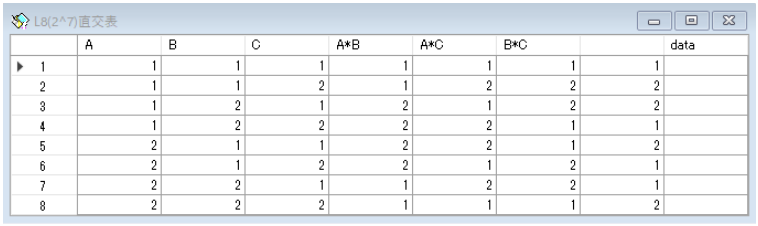
このプログラムには大きく分けて 3 つの機能が含まれている。1 つは与えられた変数構成から、直交表への変数の割り当てを行う機能、2 つ目は割り当てられた直交表に具体的なデータを代入した場合の分散分析の結果を表示する機能である。3 つ目は直交表によって作られたアンケート結果を分析するコンジョイント分析である。ここではまず、直交表への変数を割り当てる問題から解説をする。コンジョイント分析については少し詳しく解説する必要があるため、新しく節を変えて説明する。

要因の割り当てについてのデータは図 2a のような形式である。これは要因すべてが 2 水準の割り当ての例である。

図 2a　直交表割当データ 1

ここに要因名は実験を設定する際の要因の名前と交互作用の候補を表している。ここでは、要因名として A, B, C を用いているが、もちろん一文字である必要はなく、日本語でも構わない。分類数はその要因の水準数である。単独の要因の後ろには必ず記入する。また、ここで用いた「要因名」、「分類数」もどんな名称でも構わない。A*B, A*C, B*C は各要因の交互作用を表している。単独の要因名で半角の"*"を挟んだ名前が交互作用の名前である。直交表のどこの列に交互作用を割り当てるかは重要な問題であるので、それを補助する機能は必要である。

デフォルトの「直交表データから」ラジオボタンを選び、2 つの変数を選択して、「直交表割付」ボタンをクリックすると図 2b のような結果が表示される。

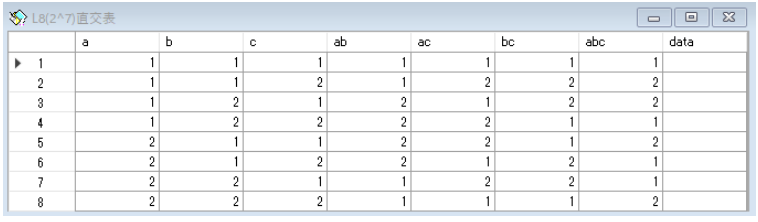


	A	B	C	A*B	A*C	B*C		data
1	1	1	1	1	1	1	1	
2	1	1	2	1	2	2	2	
3	1	2	1	2	1	2	2	
4	1	2	2	2	2	1	1	
5	2	1	1	2	2	1	2	
6	2	1	2	2	1	2	1	
7	2	2	1	1	2	2	1	
8	2	2	2	1	1	1	2	

図 2b　直交表割付結果 1

この直交表は L8(2^7)と呼ばれる直交表である。直交表の列の並びは伝統的なものがあるかも知れないが、このプログラムでは最初に単独のものから始め、相互作用の列が続く。7 列目の変数名の空欄は、ここを誤差列にするための空欄である。誤差列はいくつあってもよい。また、「data」の部分は実験結果を記入する欄である。実験の割付は単独項の水準値によって決める。この出力結果はグリッドエディタにコピーしてデータ入力用にできる。

複雑な要因の割付は手で実行するとガイドが必要となる。そのため「直交表作成」ボタンをクリックすると図 2c のような直交表の候補が出力される。



	a	b	c	ab	ac	bc	abc	data
1	1	1	1	1	1	1	1	
2	1	1	2	1	2	2	2	
3	1	2	1	2	1	2	2	
4	1	2	2	2	2	1	1	
5	2	1	1	2	2	1	2	
6	2	1	2	2	1	2	1	
7	2	2	1	1	2	2	1	
8	2	2	2	1	1	1	2	

図 2c　直交表候補 1

この変数名の a, b, c, ab, ac, bc, abc は各桁の掛け算を表しており、交互作用を見る場合は最適である。例えば要因 a と要因 bc の交互作用は要因 abc の列に表れる。これを見ると図

2 と図 3 が対応していることが分かる。これは独立な要因数が直交表の独立な要因数と一致している場合である。

次に 2 水準直交表で独立な要因数が直交表の単独な要因数より多い場合を考える。図 3 に直交表割付データを示す。

図 3a　直交表割付データ 2

ここでは要因数が 4 つで L8(2^7)直交表の独立な要因数 3 を超えている。割付結果を図 3b に示す。

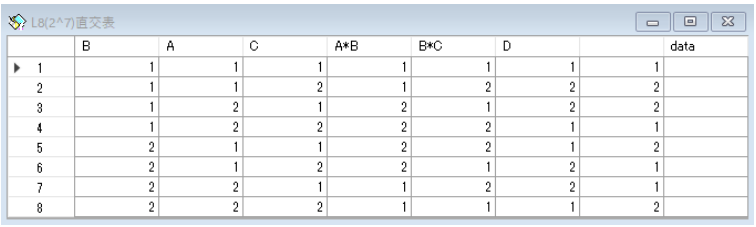


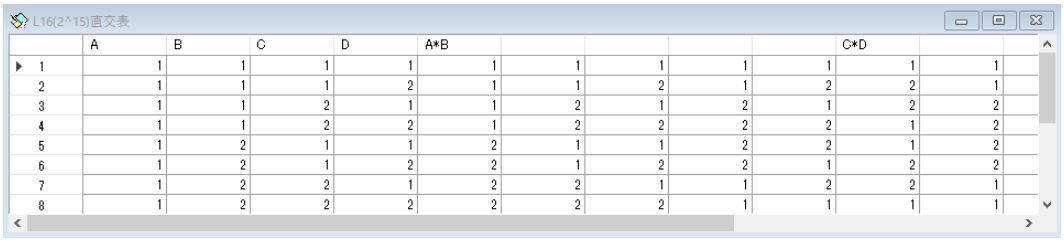
図 3b　直交表割付結果 2

要因 D は直交表の空いている列に割り付けられている。

直交表にはある大きさの直交表で割り付けられない場合がある。例えば図 4a の割付データである。

図 4a　直交表割付データ 3

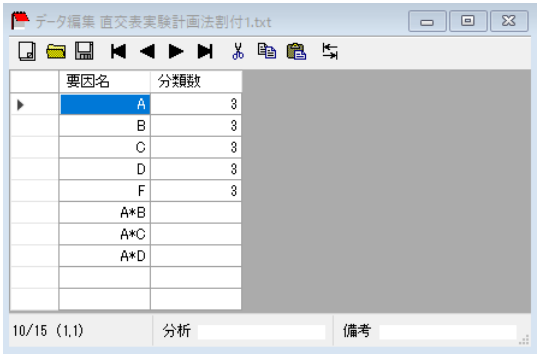
この要因は L8(2^7)直交表には割り付けられず、可能なところまで割り付けて、割り付けられない旨のメッセージを表示する。もちろん直交表を自動で拡張することもできるが、このような場合は手動で拡張するようにしている。この場合、「サンプルから」ラジオボタンを選択し「L16(2^15)」直交表を選び、再度「直交表割付」ボタンをクリックする。割付に大きな直交表が使われる。



	A	B	C	D	A*B				C*D		
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	2	1	1	2	1	2	2	1
3	1	1	2	1	1	2	1	2	1	2	2
4	1	1	2	2	1	2	2	2	2	1	2
5	1	2	1	1	2	1	1	2	2	1	2
6	1	2	1	2	2	1	2	2	1	2	2
7	1	2	2	1	2	2	1	1	2	2	1
8	1	2	2	2	2	2	2	1	1	1	1

図 4b　直交表割付結果 3

次に 3 水準直交表の割付について考えてみる。3 水準の直交表割付データを図 5a に示す。

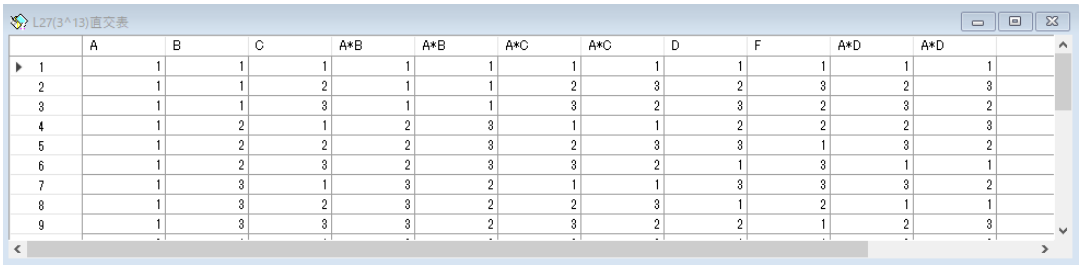


	要因名	分類数
	A	3
	B	3
	C	3
	D	3
	F	3
	A*B	
	A*C	
	A*D	

10/15 (1,1)　分析　備考

図 5a　直交表割付データ 4

この割付結果を図 5b に示す。

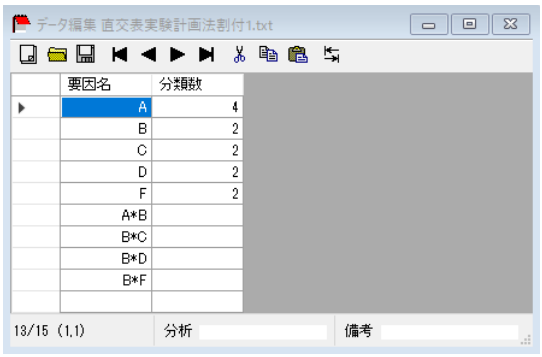


	A	B	C	A*B	A*B	A*C	A*C	D	F	A*D	A*D
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	2	1	1	2	3	2	3	2	3
3	1	1	3	1	1	3	2	3	2	3	2
4	1	2	1	2	3	1	1	2	2	2	3
5	1	2	2	2	3	2	3	3	1	3	2
6	1	2	3	2	3	3	2	1	3	1	1
7	1	3	1	3	2	1	1	3	3	3	2
8	1	3	2	3	2	2	3	1	2	1	1
9	1	3	3	3	2	3	2	2	1	2	3

図 5b　直交表割付結果 4

3 水準直交表では交互作用は 2 列で表されると述べたが、A*B, A*C, A*D は 2 つずつ列名がある。

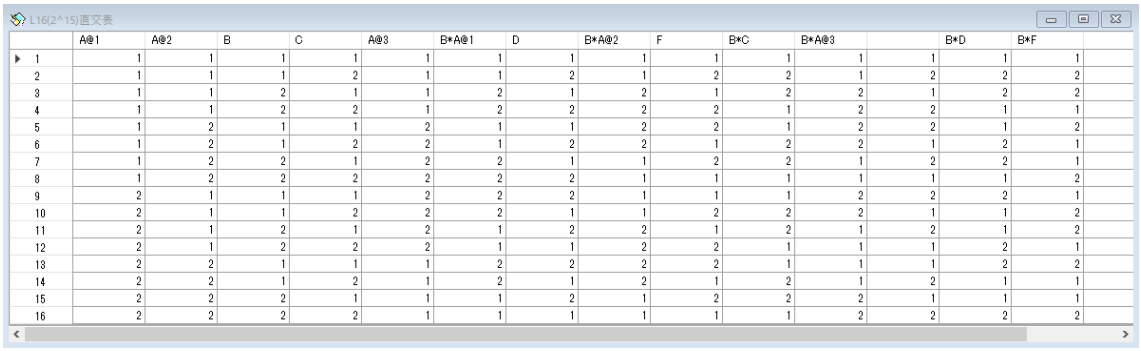
次に多水準法の割付について見てみる。図 6a に割付データ、図 6b に割付結果、図 6c に説明のための割付候補を示す。



	要因名	分類数
	A	4
	B	2
	C	2
	D	2
	F	2
	A*B	
	B*C	
	B*D	
	B*F	

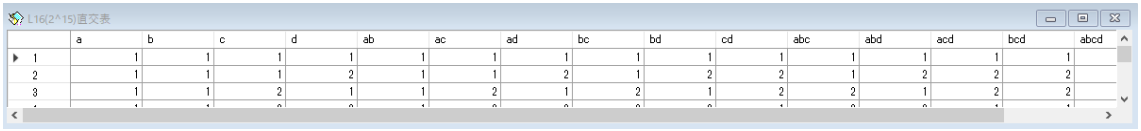
13/15 (1,1)　分析　備考

図 6a　直交表割付データ 5



	A@1	A@2	B	C	A@3	B*A@1	D	B*A@2	F	B*C	B*A@3		B*D	B*F
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	2	1	1	2	1	2	2	1	2	2	2
3	1	1	2	1	1	2	1	2	1	2	2	1	2	2
4	1	1	2	2	1	2	2	2	2	1	2	2	1	1
5	1	2	1	1	2	1	1	2	2	1	2	2	1	2
6	1	2	1	2	2	1	2	2	1	2	2	1	2	1
7	1	2	2	1	2	2	1	1	2	2	1	2	2	1
8	1	2	2	2	2	2	2	1	1	1	1	1	1	2
9	2	1	1	1	2	2	2	1	1	1	2	2	2	1
10	2	1	1	2	2	2	1	1	2	2	2	1	1	2
11	2	1	2	1	2	1	2	2	1	2	1	2	1	2
12	2	1	2	2	2	1	1	2	2	1	1	1	2	1
13	2	2	1	1	1	2	2	2	2	1	1	1	2	2
14	2	2	1	2	1	2	1	2	1	2	1	2	1	1
15	2	2	2	1	1	1	2	1	2	2	2	1	1	1
16	2	2	2	2	1	1	1	1	1	1	2	2	2	2

図 6b　直交表割付結果５



	a	b	c	d	ab	ac	ad	bc	bd	cd	abc	abd	acd	bcd	abcd
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	2	1	1	2	1	2	2	1	2	2	2	
3	1	1	2	1	1	2	1	2	1	2	2	1	2	2	

図 6c　直交表候補５

割付データは、2 水準の中に 4 水準が含まれている。割付結果を見ると、4 水準の要因 A が、A@1, A@2, A@3 に分けられて割付されている。その位置は、図 6c における a, b, ab の位置である。また B との交互作用も、B の位置が c の位置であるので、ac, bc, abc の位置になる。@1,@2,@3 の記号は@1 と@2 をかけて@3 を作るという意味で、分かり易くするために付けられている。直交表実験計画法を実行するときには、プログラムで消して実行するので、そのまま残しておいてもよいし、消して同じ要因名として実行してもよい。

次に擬水準法の割付について見てみる。図 7a に割付データ、図 7b に割付結果を示す。

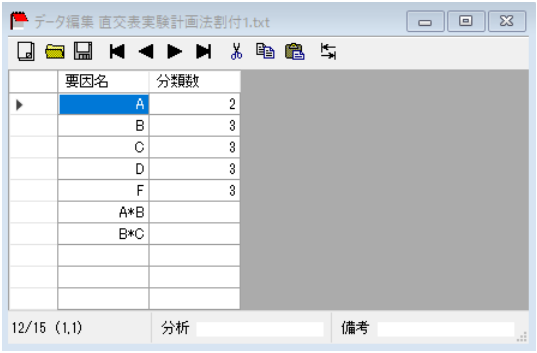
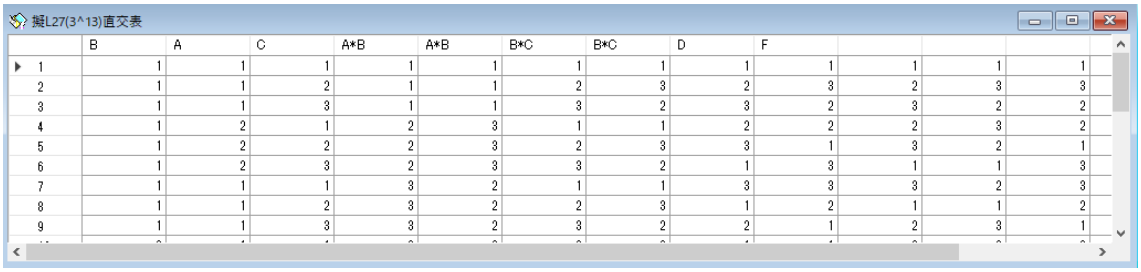


図 7a　直交表割付データ 6



	B	A	C	A*B	A*B	B*C	B*C	D	F				
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	2	1	1	2	3	2	3	2	3	3	
3	1	1	3	1	1	3	2	3	2	3	2	2	
4	1	2	1	2	3	1	1	2	2	2	3	2	
5	1	2	2	2	3	2	3	3	1	3	2	1	
6	1	2	3	2	3	3	2	1	3	1	1	3	
7	1	1	1	3	2	1	1	3	3	3	2	3	
8	1	1	2	3	2	2	3	1	2	1	1	2	
9	1	1	3	3	2	3	2	2	1	2	3	1	

図 7b　直交表割付結果６

3 水準への 2 水準の割付であるので、要因 A の本来の水準 3 の位置が水準 1 に強制的に変わっている。また、交互作用については、一般的な直交表の方法では計算できないので、水準値自体は変更されないままである。そのため、以下のようなメッセージが表示される。

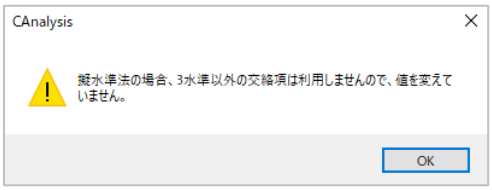


図 7c　擬水準法でのメッセージ

2 水準法の直交表に 3 水準の要因を割り当てる場合、多水準法と擬水準法を合わせた方法が用いられる。その割当データを図 8a、割当結果を図 8b に示す。

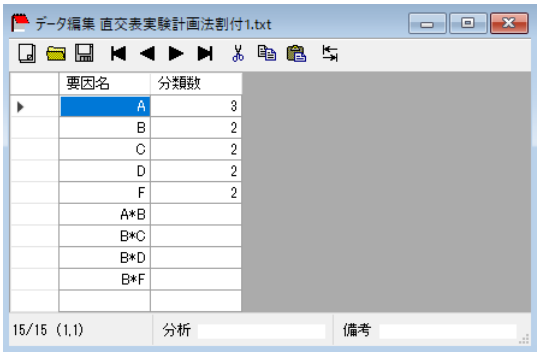


図 8a　直交表割当データ 7

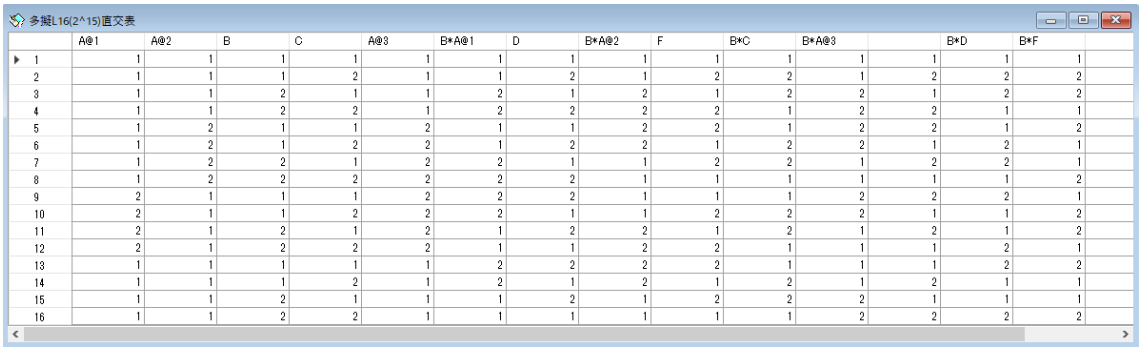


図 8b　直交表割当結果 7

この方法は、3 水準の要因を一度擬水準法を用いて仮想的な 4 水準にし、その 4 水準を多水準法を用いて 2 水準に割り付ける方法である。

ここでプログラムの制約について述べておく。まず、使える直交表は以下の種類に限られる。

L4(2^3)，L8(2^7)，L16(2^15)，L9(3^4)，L27(3^13)

割り付ける要因数については、上の直交表による制約の他に、多水準法を使った 4 水準要因の個数及び、擬水準法と多水準法を併用した 3 水準要因の個数は 1 個に限られる。これらの制約については、必要があれば拡張することも考える予定である。

次に直交表分散分析について説明する。データの形式は例えば L8(2^7)では図 9a の通りである。

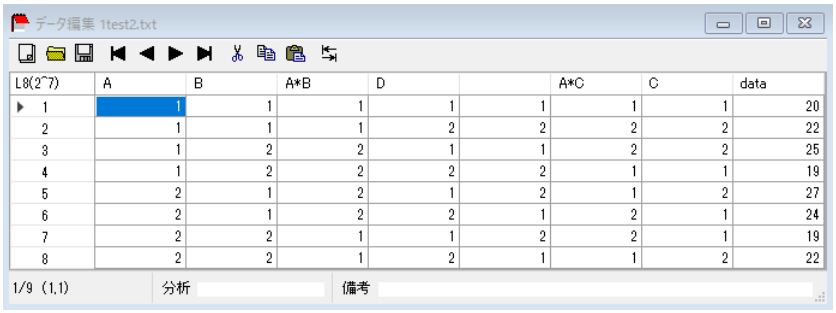


図 9a　直交表分散分析用データ 1

このデータを元に「直交表分散分析」ボタンをクリックすると図 9b のような結果が表示される。

図 9b　直交表分散分析実行結果 1

これでは有意なものが見られないので、例えば、D, A*C, C を取り除いて、図 10a のようなデータにし、実行すると図 10a のような結果になる。このように不要なデータを取り除く作業をデータのプーリングという。

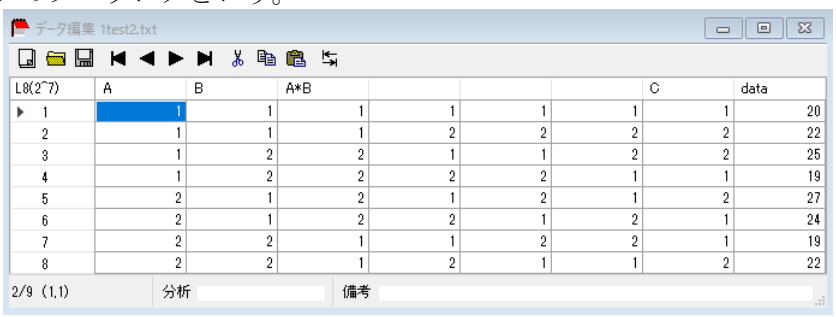
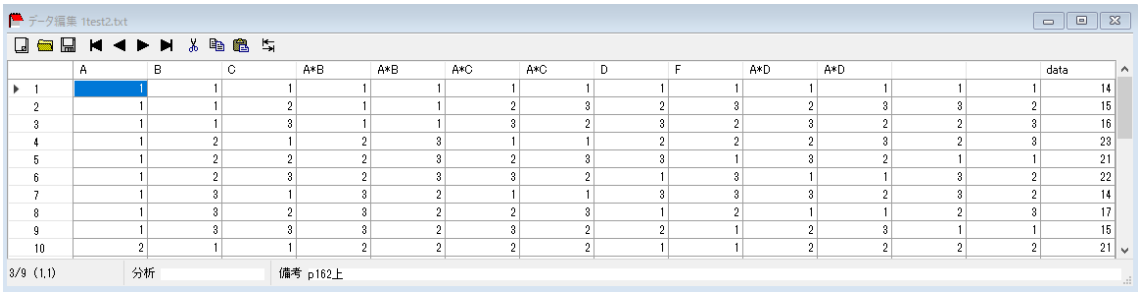


図 10a　直交表分散分析データ 2

図 10b　直交表分散分析実行結果 2

このデータでは要因 A と要因 B の交互作用と要因 C で差があることが分かる。

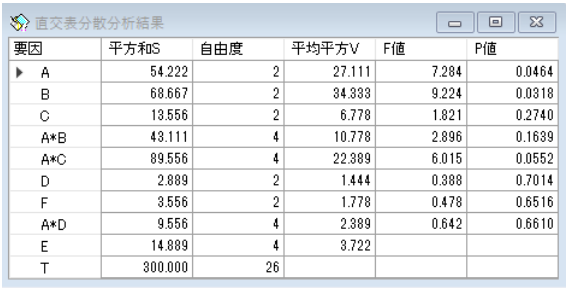
次に、複数列を使う要因の場合の例を示す。図 11a は 3 水準直交表に基づくデータで、交互作用は 2 列を使って表示される。このデータの実行結果を図 11b に示す。2 列はまとめられ計算されていることが分かる。


データ編集 1test2.txt

	A	B	C	A*B	A*B	A*C	A*C	D	F	A*D	A*D			data
1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
2	1	1	2	1	1	2	3	2	3	2	3	3	2	15
3	1	1	3	1	1	3	2	3	2	3	2	2	3	16
4	1	2	1	2	3	1	1	2	2	2	3	2	3	23
5	1	2	2	2	3	2	3	3	1	3	2	1	1	21
6	1	2	3	2	3	3	2	1	3	1	1	3	2	22
7	1	3	1	3	2	1	1	3	3	3	2	3	2	14
8	1	3	2	3	2	2	3	1	2	1	1	2	3	17
9	1	3	3	3	2	3	2	2	1	2	3	1	1	15
10	2	1	1	2	2	2	2	1	1	2	2	2	2	21

3/9 (1,1)　分析　備考 p162上

図 11a　直交表分散分析データ 3

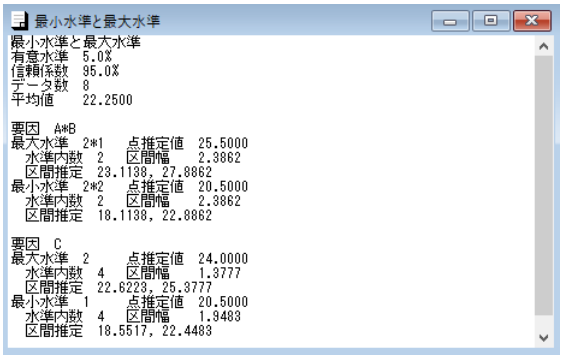

直交表分散分析結果

要因	平方和S	自由度	平均平方V	F値	P値
A	54.222	2	27.111	7.284	0.0464
B	68.667	2	34.333	9.224	0.0318
C	13.556	2	6.778	1.821	0.2740
A*B	43.111	4	10.778	2.896	0.1639
A*C	89.556	4	22.389	6.015	0.0552
D	2.889	2	1.444	0.388	0.7014
F	3.556	2	1.778	0.478	0.6516
A*D	9.556	4	2.389	0.642	0.6610
E	14.889	4	3.722		
T	300.000	26			

図 11b　直交表分散分析実行結果 3

このデータでは、要因 A、要因 B に差が見られる。

ここでこれらの要因によるデータの平均の最大及び最小推定値を求めてみる。例えば図 10a のデータで、「最適水準」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。


最小水準と最大水準

```
最小水準と最大水準
有意水準  5.0%
信頼係数  95.0%
データ数  8
平均値    22.2500

要因  A*B
最大水準  2*1    点推定値  25.5000
  水準内数  2    区間幅    2.3862
  区間推定  23.1138, 27.8862
最小水準  2*2    点推定値  20.5000
  水準内数  2    区間幅    2.3862
  区間推定  18.1138, 22.8862

要因  C
最大水準  2      点推定値  24.0000
  水準内数  4    区間幅    1.3777
  区間推定  22.6223, 25.3777
最小水準  1      点推定値  20.5000
  水準内数  4    区間幅    1.9483
  区間推定  18.5517, 22.4483
```

図 12　最適水準値

ここでは、検定で有意差があった項目ごとに最大と最低の要因の組を与えている。これをまとめて表示することも可能だが、水準が重複したり、相反する水準が現れたりすることがあるのでこのような表示に留めている。今後議論の必要がある。

ここで求めた結果の有意差と信頼係数の指定は変数選択の画面の中で変更可能である。図 11 の結果で有意水準を 10%、それと連動させて信頼係数 90%にすると最適水準値の結果は図 13 のようになる。

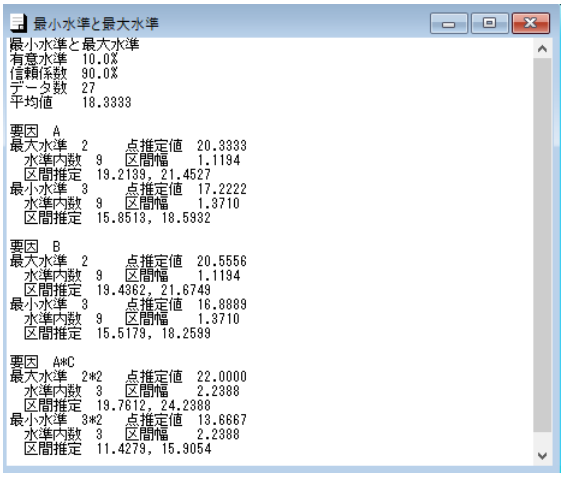


図 13　有意水準と信頼係数を 10%にした最適水準値

25.3　コンジョイント分析

コンジョイント分析は直交表分散分析と数量化 I 類を合わせた分析である。直交表分散分析では直交表によって実験の組み合わせを考え実験計画を立てるが、コンジョイント分析ではアンケートの中で商品の特徴を効率よく組み合わせるために直交表が使われる。回答者はこのように特徴が組み合わされた商品に対して効用値（点数でも好きな順位でもよい）を付ける。分析では、効用値を目的変数にして数量化 I 類を実行し、各質問項目がどのように効用値に影響しているかを見る。コンジョイント分析が数量化 I 類と異なるところは、交互作用を容易に取り入れられることである。

回答者が複数の場合、直交表を含めたデータを下に加えて行くが、通常の直交表分散分析では同じ質問項目の平均を使って分析を実行するため、多くのデータを集めることは意味がない。実際、直交表分散分析は実験回数を減らすことが目的なので、複数組のデータということは考えない。しかし、コンジョイント分析は数量化 I 類（0/1 データに変換して重回帰分析）の処理を行うので、パラメータの精度はデータ数に応じて良くなり、特徴が見つけやすくなる。

ここでは参照を容易にするために、前節の分析実行画面を図 1 に再掲する。

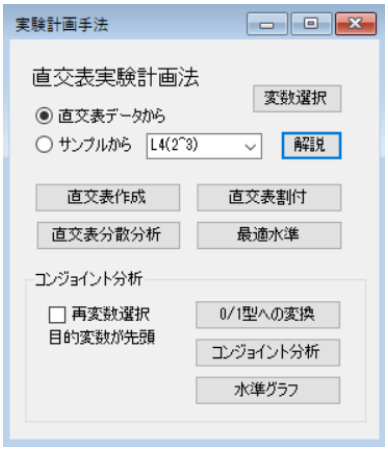
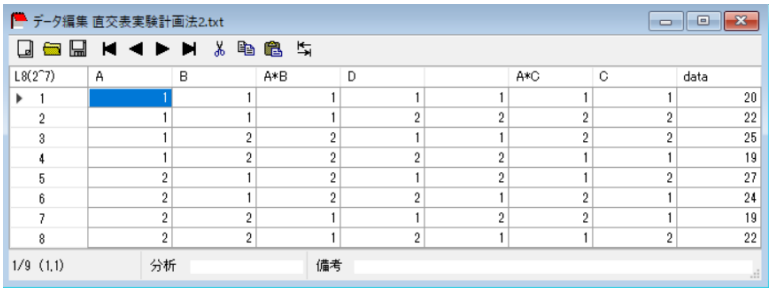


図 1　直交表実験計画法実行画面

前節で用いた直交表実験計画法 2.txt のデータ（図 9a）を利用して、コンジョイント分析のプログラムを見て行こう。改めて図 2 にそのデータを表示する。



L8(2^7)	A	B	A*B	D	A*C	C	data
1	1	1	1	1	1	1	20
2	1	1	1	2	2	2	22
3	1	2	2	1	1	2	25
4	1	2	2	2	2	1	19
5	2	1	2	1	2	1	27
6	2	1	2	2	1	2	24
7	2	2	1	1	2	2	19
8	2	2	1	2	1	1	22

1/9 (1,1)　分析　備考

図 2　直交表実験計画法のデータ 1（直交表実験計画法 2.txt）

この場合、データ（目的変数）の選択順は直交表分散分析と同じく最後にする。直交表分散分析の結果を図 3 に、コンジョイント分析の結果を図 4 に示す。直交表実験計画法の結果を図 3 に、コンジョイント分析の結果を図 4 に示す。

要因	平方和S	自由度	平均平方V	F値	P値
A	4.500	1	4.500	2.250	0.3743
B	8.000	1	8.000	4.000	0.2952
A*B	18.000	1	18.000	9.000	0.2048
D	2.000	1	2.000	1.000	0.5000
A*C	0.500	1	0.500	0.250	0.7048
C	24.500	1	24.500	12.250	0.1772
誤差	2.000	1	2.000		
Total	59.500	7			

図 3　直交表分散分析の結果

	重回帰ウェイ	重回帰確率	基準化ウェイ	基準化確率	結合変数	結合確率
A:1	0.0000		-0.7500	0.3743	A	0.3743
A:2	1.5000	0.3743	0.7500	0.3743	B	0.2952
B:1	0.0000		1.0000	0.2952	A*B	0.2048
B:2	-2.0000	0.2952	-1.0000	0.2952	D	0.5000
A*B:1	0.0000		-1.5000	0.2048	A*C	0.7048
A*B:2	3.0000	0.2048	1.5000	0.2048	C	0.1772
D:1	0.0000		0.5000	0.5000		
D:2	-1.0000	0.5000	-0.5000	0.5000		
A*C:1	0.0000		-0.2500	0.7048		
A*C:2	0.5000	0.7048	0.2500	0.7048		
C:1	0.0000		-1.7500	0.1772		
C:2	3.5000	0.1772	1.7500	0.1772		
定数項	19.5000	0.0431	22.2500	0.0143		
重相関	0.983	寄与率	0.966			
有効性F値	4.7917	自由度	6,1	p値	0.1006	

図 4　コンジョイント分析の結果

コンジョイント分析の結果は、数量化 I 類の重回帰ウェイトと基準化ウェイトを用いており、どの選択肢が結果を上げるのか、下げるのか、などがよく分かる。また「結合確率」によって、直交表分散分析と同様に、各変数の重要性も分かる。ここで、基準化ウェイトの係数が 0 となることを検定する「基準化確率」や 1 つの変数を構成する分けられた複数変数の係数が同時に 0 になることを検定する「結合確率」は、結合仮説の検定の以下の関係式を用いて求めている。

$$F = \frac{(\mathbf{R}\hat{\mathbf{b}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{b}} - \mathbf{r})/q}{EV/(N-k-1)} \sim F_{q,N-k-1}$$

「水準グラフ」ボタンをクリックすると、この基準化ウェイトの変動を図 5 のようにグラフ化することができる。

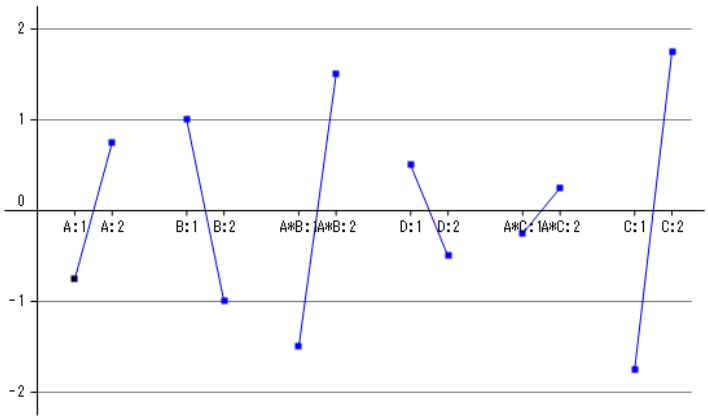


図 5　基準化ウェイト

次にもう少し複雑な例として、1 つの変数を 2 つ以上の列で表す例を考える。図 6 にそのデータを与えておく（前節図 11a）。

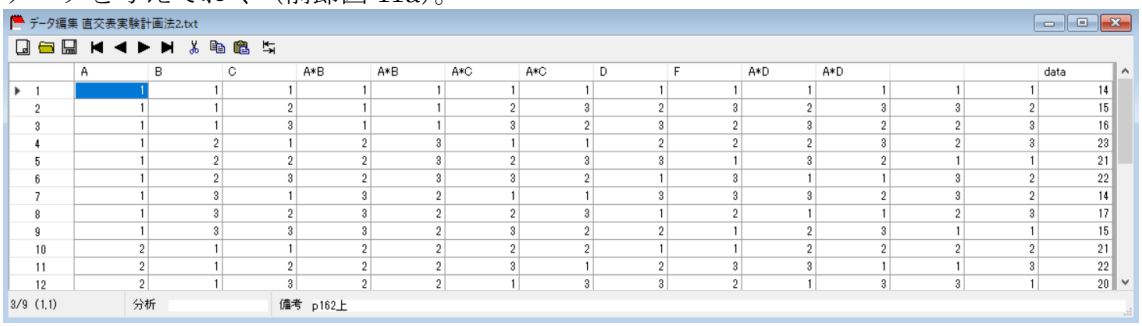


図 6　直交表実験計画法のデータ 2（直交表実験計画法 2.txt）

このデータから求められたコンジョイント分析の結果を図 7 に示す。

	重回帰ウェイ	重回帰確率	基準化ウェイ	基準化確率	結合変数	結合確率
A:1	0.0000		−0.8889	0.1657	A	0.0464
A:2	2.8889	0.0337	2.0000	0.0190	B	0.0318
A:3	−0.2222	0.8190	−1.1111	0.1018	C	0.2740
B:1	0.0000		−0.7778	0.2127	A*B	0.1639
B:2	3.0000	0.0300	2.2222	0.0133	A*C	0.0552
B:3	−0.6667	0.5042	−1.4444	0.0513	D	0.7014
C:1	0.0000		1.0000	0.1296	F	0.6516
C:2	−1.5556	0.1624	−0.5556	0.3497	A*D	0.6610
C:3	−1.4444	0.1874	−0.4444	0.4450		
A*B:1	0.0000		−1.0000	0.1296		
A*B:2	2.5556	0.0483	1.5556	0.0415		
A*B:3	0.4444	0.6507	−0.5556	0.3497		
A*B:1	0.0000		−0.6667	0.2731		
A*B:2	0.5556	0.5743	−0.1111	0.8428		

図 7　コンジョイント分析結果

また基準化ウェイトをグラフにしたものを図 8 に示す。

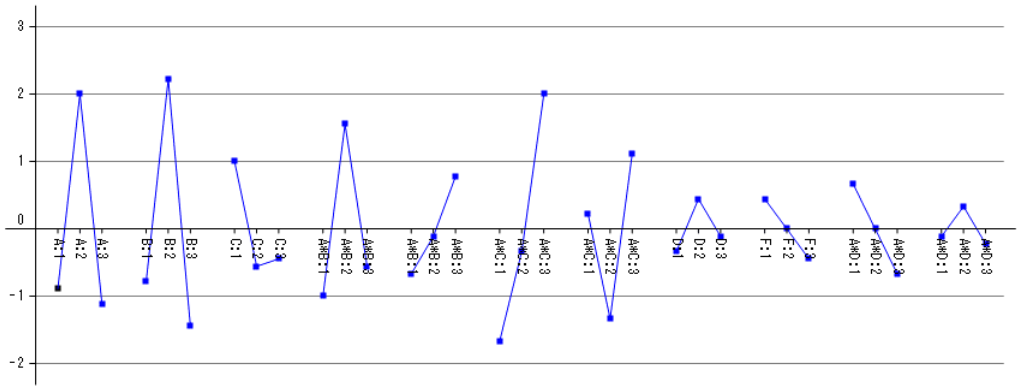


図 8　基準化ウェイト

この分析メニューでは、不要な変数を取り除いて選択し、通常の数量化 I 類のように、目的変数を先頭に選んで分析することもできる。その際は、コンジョイント分析グループボックス内の「再変数選択」チェックボックスにチェックを入れる。

参考文献

[1] Excel でここまでできる実験計画法　一元配置実験から直交配列表実験まで, 森田浩・今里健一郎・奥村清志, 日本規格協会, 2011.

２６．パネルデータ分析

26.1　パネルデータ分析とは

重回帰分析で、除外された変数がある場合、時系列データが与えられていると、その影響を取り除くことが可能になることがある。例えばそれが個体に依存し、時間には依存しない固定効果の場合、各個体の測定値の時間平均を引くと固定効果を消し去ることができる。またそれが時間に依存し、個体には依存しない時間効果の場合、各時間の測定値の個体平均を引くと時間効果を消し去ることができる。パネルデータ分析は、この性質を利用して個体や時間に依存する特殊な影響を取り除いて重回帰分析を行う手法である。

パネルデータ回帰分析は、以下のような形の回帰分析である。

$$y_{it} = \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + c_i + d_t + u_{it}$$

ここに、$i = 1, \cdots, n$は個体の識別記号、$t = 1, \cdots, T$ は時間の識別番号である。通常の回帰分析と異なるところは、時間について変化しない固定効果c_iと個体について変化しない時間効果d_tを含むことである。但し、固定効果、時間効果、誤差項については以下の仮定を置き、固定効果と時間効果については直接には観測されないものとする。

$$\bar{c} = \bar{d} = 0,\ \bar{u}_i = \bar{u}_t = \bar{u} = 0$$

最初の式から観測されないc_iとd_tを消すために、次の変換を実行し、

$$\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}, \quad \tilde{x}_{ait} = x_{ait} - \bar{x}_{ai} - \bar{x}_{at} + \bar{x}_a, \quad \tilde{u}_{it} = u_{it} - \bar{u}_i - \bar{u}_t + \bar{u} = u_{it}$$

結局以下の関係を得る。

$$\tilde{y}_{it} = \sum_{a=1}^{p} \beta_a \tilde{x}_{ait} + u_{it}$$

この関係を使って回帰分析を実行し、回帰係数$\hat{\beta}_a$を推定する。最後に、推定された回帰係数を使って、定数項、固定効果、時間効果について、以下のように求める。

$$\hat{\beta}_0 = \bar{y} - \sum_{a=1}^{p} \hat{\beta}_a \bar{x}_a, \quad \hat{c}_i = \bar{y}_i - \sum_{a=1}^{p} \hat{\beta}_a \bar{x}_{ai} - \hat{\beta}_0, \quad \hat{d}_t = \bar{y}_t - \sum_{a=1}^{p} \hat{\beta}_a \bar{x}_{at} - \hat{\beta}_0$$

定義や式の詳細については、最後の節にまとめておく。

ここで述べるモデルはまだプロトタイプの域を出ていない。今後追加の分析や使い易くする機能を加えて行く予定である。

26.2　プログラムの利用法

メニュー［分析－多変量解析他－経済・経営手法－パネルデータ分析］を選択すると、図1のような分析実行画面が表示される。

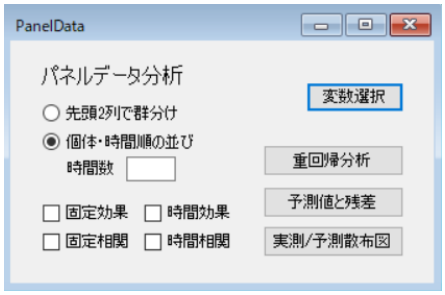


図１　分析実行画面

これに対してデータは通常の重回帰分析と同じである。但し、データのレコード配置は、「個体・時間順の並び」の場合、図２のように、個体１・年次１，個体１・年次２，…，個体ｎ・年次１，…，個体ｎ・年次Ｔの順番で目的変数，説明変数１，説明変数２，… が順番に並び、「先頭２列で群分け」の場合、図３のように、個体分類，時間分類，目的変数，説明変数１，説明変数２，… のようにデータが並ぶ（パネルデータ分析 1.txt）。

図２　「個体・時間順の並び」のデータ形式

図３　「先頭２列で群分け」のデータ形式

図２の形式では、データを欠損値なく、順序通りに並べる必要があるが、図３の形式では、データの順序が変わっていても、欠損値があっても対応可能である。ただ、時間平均や個体平均を取る場合、欠損値を除いた個数で平均をとるため、偏りがあるような欠損値の場合（例えば、時間と共に増大するデータの最初の値や最後の値の欠損など）、結果の正当性に問題が出る恐れもある。

図 2 の場合、変数選択ですべての列を選び、「時間数」を 4 と入力して、その他の設定は何もせず、「重回帰分析」ボタンをクリックした結果を図 4 に示す。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数1	0.1410	0.8489	0.0130	10.8450	0.0000	0.1155	0.1665
説明変数2	2.4932	0.2642	0.6087	4.0956	0.0000	1.3001	3.6863
説明変数3	−0.2353	−0.0436	0.2818	−0.8349	0.4038	−0.7876	0.3170
切片	−0.0319	0.0000	12.0772	−0.0026	0.9979	−23.7028	23.6391
R	0.957	R^2	0.916				

図 4　効果も相関も考えない場合の結果

これは、通常の重回帰分析の結果に一致している。

次に、時間に関してデータ間に相関がある場合について分析を実行する。「時間相関」にチェックを入れて、「重回帰分析」ボタンをクリックすると図 5 に示す結果となる。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数1	0.1410	0.8489	0.0064	22.0403	0.0000	0.1285	0.1535
説明変数2	2.4932	0.2642	0.5638	4.4218	0.0000	1.3881	3.5983
説明変数3	−0.2353	−0.0436	0.3494	−0.6735	0.5007	−0.9200	0.4495
切片	−0.0319	0.0000	11.5072	−0.0028	0.9978	−22.5855	22.5218
R	0.957	R^2	0.916				

図 5　時間相関がある場合の結果

次に、固定効果と時間相関がある場合の例を示す。「固定効果」チェックボックスと「時間相関」チェックボックスにチェックを入れて「重回帰分析」ボタンをクリックすると、図 6 のような結果となる。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数1	0.1505	0.9061	0.0174	8.6690	0.0000	0.1165	0.1845
説明変数2	2.6666	0.2826	0.7125	3.7425	0.0002	1.2701	4.0632
説明変数3	−0.1792	−0.0332	0.3377	−0.5308	0.5955	−0.8410	0.4826
切片	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
R	0.962	R^2	0.925				

図 6　固定効果と時間相関がある場合の結果

固定効果と時間相関がある場合の、予測値について示す。「予測値と残差」ボタンをクリックすると、図 7 のような結果を得る。

予測値と残差

目的変数	実測値	部分予測値	固定効果	時間効果	定数	予測値	残差
1	127	142.5975	−1.8880	0.0000	−10.8833	129.8262	−2.8262
2	153	157.1725	−1.8880	0.0000	−10.8833	144.4012	8.5988
3	183	184.6594	−1.8880	0.0000	−10.8833	171.8881	11.1119
4	177	206.6558	−1.8880	0.0000	−10.8833	193.8844	−16.8844
5	83	98.9225	4.6512	0.0000	−10.8833	92.6904	−9.6904
6	106	121.0912	4.6512	0.0000	−10.8833	114.8591	−8.8591
7	135	127.3684	4.6512	0.0000	−10.8833	121.1362	13.8638
8	147	148.5464	4.6512	0.0000	−10.8833	142.3143	4.6857
9	100	111.8435	−1.7283	0.0000	−10.8833	99.2319	0.7681
10	125	142.9235	−1.7283	0.0000	−10.8833	130.3119	−5.3119

図 7　固定効果と時間相関がある場合の予測値と残差

図 7 により、本章 1 節の最初に述べた回帰式の構造がよく分かる。部分予測値は、個体効果や時間効果を除いた予測値で、それにこれらの効果や定数を含めると実際の予測値となる。これは最後から 2 列目に表示されている。

最後に、「実測・予測散布図」ボタンをクリックして実測値と予測値をグラフに描くと図

8 のようになる。

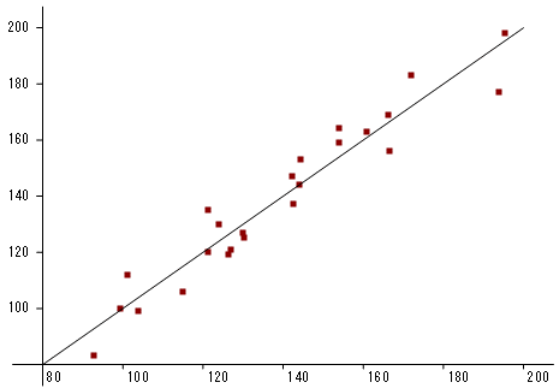


図 8　実測値/予測値グラフ

26.3　パネルデータ分析の理論

パネルデータの目的変数を y_{it}（$i=1,\cdots,n,\ \ t=1,\cdots,T$）、説明変数を x_{ait}（$a=1,\cdots,p$）、個体の固定効果を c_i、時間効果を d_t、定数項を β_0、誤差を u_{it} として、以下のモデルを考える。

$$Y_{it}=\sum_{a=1}^{p}\beta_a x_{ait}+\beta_0+\delta_f c_i+\delta_t d_t$$

$$y_{it}=Y_{it}+u_{it}=\sum_{a=1}^{p}\beta_a x_{ait}+\beta_0+\delta_f c_i+\delta_t d_t+u_{it}$$

ここに、δ_f は固定効果を考える場合は 1、考えない場合は 0 を与える定数で、δ_t は時間効果を考える場合は 1、考えない場合は 0 を与える定数である。また、固定効果、時間効果、誤差について、以下を仮定する。

$$\bar{c}=\bar{d}=0,\ \bar{u}_i=\bar{u}_t=\bar{u}=0$$

また、変数については以下のような変換を考える（固定効果がある場合は時間平均を引き、時間効果がある場合は個体平均を引く）。

$$\tilde{y}_{it}=y_{it}-\delta_f\bar{y}_i-\delta_t\bar{y}_t+\delta_f\delta_t\bar{y},\quad \tilde{x}_{ait}=x_{ait}-\delta_f\bar{x}_{ai}-\delta_t\bar{x}_{at}+\delta_f\delta_t\bar{x}_a,$$

$$\tilde{u}_{it}=u_{it}-\delta_f\bar{u}_i-\delta_t\bar{u}_t+\delta_f\delta_t\bar{u}=u_{it}$$

ここに、

$$\bar{y}_i=\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{a=1}^{p}\beta_a x_{ait}+\delta_f c_i+\delta_t d_t+\beta_0+u_{it}\right]=\sum_{a=1}^{p}\beta_a\bar{x}_{ai}+\delta_f c_i+\beta_0$$

$$\bar{y}_t=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{a=1}^{p}\beta_a x_{ait}+\delta_f c_i+\delta_t d_t+\beta_0+u_{it}\right]=\sum_{a=1}^{p}\beta_a\bar{x}_{at}+\delta_t d_t+\beta_0$$

$$\bar{y}=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\sum_{a=1}^{p}\beta_a x_{ait}+\delta_f c_i+\delta_t d_t+\beta_0+u_{it}\right]=\sum_{a=1}^{p}\beta_a\bar{x}_a+\beta_0$$

この変換によって、固定効果と時間効果の項は消え、以下のような関係を得る。

$$\tilde{y}_{it} = \sum_{a=1}^{p} \beta_a \tilde{x}_{ait} + \delta_f c_i - \delta_f^2 c_i + \delta_t d_t - \delta_t^2 d_t + \beta_0 (1 - \delta_f - \delta_t + \delta_f \delta_t) + u_{it}$$

$$= \sum_{a=1}^{p} \beta_a \tilde{x}_{ait} + (1 - \delta_f)(1 - \delta_t)\beta_0 + u_{it}$$

これにより、予測値は以下のようになる。

$$\tilde{Y}_{it} = \sum_{a=1}^{p} \beta_a \tilde{x}_{ait} + (1 - \delta_f)(1 - \delta_t)\beta_0$$

ここで実際のプログラムでは、$(1-\delta_f)(1-\delta_t)\beta_0$ 部分を改めて β_0 として計算している。後に述べるように、固定効果や時間効果がある場合、最小 2 乗法の計算から自動的に予測値 $\hat{\beta}_0$ は 0 になる。

実測値と予測値の差の 2 乗を以下のように L とする。

$$L = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \tilde{Y}_{it} \right)^2$$

これを最小化することから、

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \sum_{b=1}^{p} \hat{\beta}_b \tilde{x}_{bit} - \hat{\beta}_0 \right) = 0$$

$$\frac{\partial L}{\partial \beta_a} = -2 \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{ait} \left(\tilde{y}_{it} - \sum_{b=1}^{p} \hat{\beta}_b \tilde{x}_{bit} - \hat{\beta}_0 \right) = 0$$

ここで、固定効果や時間効果がある場合、

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{y}_{it} = 0, \quad \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{bit} = 0$$

であることから、第 1 式より自動的に $\hat{\beta}_0 = 0$ が示される。

最小化の方程式は、まとめて、

$$\tilde{\mathbf{y}} = \begin{pmatrix} \tilde{y}_{11} \\ \tilde{y}_{12} \\ \vdots \\ \tilde{y}_{nT} \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} 1 & \tilde{x}_{111} & \cdots & \tilde{x}_{p11} \\ 1 & \tilde{x}_{112} & \cdots & \tilde{x}_{p12} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{x}_{1nT} & \cdots & \tilde{x}_{pnT} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{nT} \end{pmatrix}$$

のように定義すると、以下のように書かれる。

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

これを解いて、以下を得る。

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

この計算過程から、残差 $\hat{u}_{it}$ に対する以下の制約が得られる。

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{u}_{it} = 0, \qquad \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{ait} \hat{u}_{it} = 0$$

これは、重回帰分析における誤差項の一般的な性質である。

固定効果や時間効果がある場合、以下の回帰式から、

$$\tilde{y}_{it} = \sum_{a=1}^{p} \hat{\beta}_a \tilde{x}_{ait} + \hat{\beta}_0 + \hat{u}_{it}$$

次の制約も追加される。

固定効果から $\bar{\hat{u}}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{it} = 0$,　　時間効果から $\bar{\hat{u}}_t = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_{it} = 0$

すべての仮定を考えると、残差$\hat{u}_{it}$の自由度は以下となる。

$$D = nT - p - \delta_f (n-1) - \delta_t (T-1) - 1$$

次に、パラメータの標準誤差について考える。回帰式 $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u}$ より、

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{u}$$

これらを成分で表示すると以下となるが、

$$(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})_{ab} = \sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{x}_{ait}\tilde{x}_{bit}, \quad (\tilde{\mathbf{X}}'\mathbf{u})_a = \sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{x}_{ait} u_{it}$$

$\tilde{\mathbf{w}} = \tilde{\mathbf{X}}'\mathbf{u}$ とすると、平均と共分散は以下となる。

$$E[\tilde{w}_a] = E\left[(\tilde{\mathbf{X}}'\mathbf{u})_a\right] = \sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{x}_{ait} E[u_{it}] = 0$$

$$Cov[\tilde{w}_a, \tilde{w}_b] = Cov\left[(\tilde{\mathbf{X}}'\mathbf{u})_a, (\tilde{\mathbf{X}}'\mathbf{u})_b\right] = E\left[\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{x}_{ait} u_{it} \sum_{j=1}^{n}\sum_{t'=1}^{T} \tilde{x}_{bjt'} u_{jt'}\right]$$

$$= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T} \tilde{x}_{ait} u_{it} \tilde{x}_{bjt'} u_{jt'}\right]$$

個体に対する独立性を仮定した場合（時間に関しては仮定しない）、

$E[u_{it}u_{jt'}] = E[u_{it}u_{it'}]\delta_{ij}$, $\tilde{v}_{ait} = \tilde{x}_{ait}u_{it}$, $\eta_{ai} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \tilde{v}_{ait}$　とすると

$$Cov[\tilde{w}_a, \tilde{w}_b] = E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{t'=1}^{T} \tilde{x}_{ait} u_{it} \tilde{x}_{bit'} u_{it'}\right] = E\left[\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{v}_{ait} \sum_{t'=1}^{T} \tilde{v}_{ait'}\right]$$

$$= TE\left[\sum_{i=1}^{n} \eta_{ai}\eta_{bi}\right] = nT\sigma_{\eta_a\eta_b}$$

$$\to \frac{nT}{n-1}\sum_{i=1}^{n} \hat{\eta}_{ai}\hat{\eta}_{bi} = \frac{n}{n-1}\sum_{i=1}^{n}\left(\sum_{t=1}^{T} \tilde{v}_{ait}\right)\left(\sum_{t'=1}^{T} \tilde{v}_{ait'}\right)$$

時間に関する独立性を仮定した場合（個体に関しては仮定しない）

$E[u_{it}u_{jt'}] = E[u_{it}u_{jt}]\delta_{tt'}$, $\tilde{v}_{ait} = \tilde{x}_{ait}u_{it}$, $\gamma_{at} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{v}_{ait}$　とすると

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{x}_{ait}u_{it}\tilde{x}_{bjt}u_{jt}\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{v}_{ait}\sum_{j=1}^{n}\tilde{v}_{bjt}\right]$$

$$=nE\left[\sum_{t=1}^{T}\gamma_{at}\gamma_{bt}\right]=nT\sigma_{\gamma_a\gamma_b}$$

$$\rightarrow\frac{nT}{T-1}\sum_{t=1}^{T}\hat{\gamma}_{at}\hat{\gamma}_{bt}=\frac{T}{T-1}\sum_{t=1}^{T}\left(\sum_{i=1}^{n}\tilde{v}_{ait}\right)\left(\sum_{j=1}^{n}\tilde{v}_{ajt}\right)$$

個体と時間に関する独立性を仮定した場合（通常の不均一分散）

$E\left[u_{it}u_{jt'}\right]=E\left[u_{it}u_{it}\right]\delta_{ij}\delta_{tt'}$，　$\tilde{v}_{ait}=\tilde{x}_{ait}u_{it}$　とすると

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{x}_{ait}u_{it}\tilde{x}_{bit}u_{it}\right]=E\left[\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{v}_{ait}\tilde{v}_{bit}\right]=nT\sigma_{\tilde{v}_a\tilde{v}_b}$$

$$\rightarrow\frac{nT}{nT-p-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\tilde{v}_{ait}\tilde{v}_{bit}$$

個体も時間も独立性を仮定しない場合

$\tilde{v}_{ait}=\tilde{x}_{ait}u_{it}$，　$\phi_{at}=\dfrac{1}{\sqrt{nT}}\displaystyle\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{v}_{ait}$　とすると

$$Cov\left[\tilde{w}_a,\tilde{w}_b\right]=E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{ait}u_{it}\sum_{j=1}^{n}\sum_{t'=1}^{T}\tilde{x}_{bjt'}u_{jt'}\right]=nT\sigma_{\phi_a\phi_b}$$

$$\rightarrow nT\hat{\phi}_a\hat{\phi}_b=\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\tilde{v}}_{ait}\sum_{j=1}^{n}\sum_{t'=1}^{T}\hat{\tilde{v}}_{bjt'}$$

上で述べたことを利用して、改めてパラメータ $\hat{\boldsymbol{\beta}}$ の共分散を求める。

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{u}$$

より、$\mathbf{G}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ とおくと $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\mathbf{G}\tilde{\mathbf{w}}$

これから以下の関係を得る。

$$Cov\left[\hat{\beta}_a,\hat{\beta}_b\right]=Cov\left[(\mathbf{G}\tilde{\mathbf{w}})_a,(\mathbf{G}\tilde{\mathbf{w}})_b\right]=E\left[\sum_{c=1}^{p}g_{ac}\tilde{w}_c\sum_{d=1}^{p}g_{bd}\tilde{w}_d\right]$$

$$=\sum_{c=1}^{n}\sum_{d=1}^{n}g_{ac}g_{bd}Cov\left[\tilde{w}_c,\tilde{w}_d\right]$$

個体に対する独立性を仮定した場合（時間に関しては仮定しない）

$$Cov\left[\hat{\beta}_a,\hat{\beta}_b\right]\rightarrow\frac{nT}{n-1}\sum_{c=1}^{n}\sum_{d=1}^{n}g_{ac}g_{bd}\sum_{i=1}^{n}\hat{\eta}_{ci}\hat{\eta}_{di}$$

$$=\frac{n}{n-1}\sum_{c=1}^{n}\sum_{d=1}^{n}g_{ac}g_{bd}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\hat{\tilde{v}}_{cit}\right)\left(\sum_{t'=1}^{T}\tilde{v}_{dit'}\right)$$

他も同様であるのでここでは省略する。

ここで、$p=1$の固定効果モデルで、個体に対する独立性を仮定すると

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}=(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})_{11}=\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{1it}^2 \text{ より、} \mathbf{G}=g=\frac{1}{(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})_{11}}=\frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{1it}^2}=\frac{1}{nTQ_{\tilde{x}}}$$

$$Var\left[\hat{\beta}\right]=Var\left[g\tilde{w}_1\right]=\frac{1}{n^2T^2Q_{\tilde{x}}^2}Var\left[\tilde{w}_1\right]$$

$$\rightarrow \frac{nT}{n-1}\frac{1}{n^2T^2\hat{Q}_{\tilde{x}}^2}\sum_{i=1}^{n}\hat{\eta}_{1i}^2=\frac{n}{n-1}\frac{1}{n^2T^2\hat{Q}_{\tilde{x}}^2}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\hat{\tilde{v}}_{1it}\right)^2$$

注）有効性の検定及び係数の結合仮説検定には以下の関係を用いるが、

$$F=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\sim F_{k,\infty}$$

個体と時間に関する独立性を仮定しない方法で作られた共分散行列は正則ではなく、有効性の検定には利用できない。

固定効果と時間効果の求め方

$$y_{it}=\sum_{a=1}^{p}\beta_a x_{ait}+c_i+d_t+\beta_0+u_{it}$$

回帰分析の結果は以下の式になる。

$$\tilde{y}_{it}=\sum_{a=1}^{p}\hat{\beta}_a\tilde{x}_{ait}+\hat{u}_{it}$$

左辺と右辺第１項の関係から、

$$\bar{\hat{u}}_i=\bar{\hat{u}}_t=\bar{\hat{u}}=0$$

この関係を元の式に用いると

$$y_{it}=\sum_{a=1}^{p}\hat{\beta}_a x_{ait}+\hat{\beta}_0+c_i+d_t+\hat{u}_{it}$$

これより

$$\bar{y}=\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_a+\hat{\beta}_0 \quad\rightarrow\quad \hat{\beta}_0=\bar{y}-\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_a$$

$$\bar{y}_i=\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{ai}+\hat{\beta}_0+\hat{c}_i \quad\rightarrow\quad \hat{c}_i=\bar{y}_i-\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{ai}-\hat{\beta}_0$$

$$\bar{y}_t=\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{at}+\hat{\beta}_0+\hat{d}_t \quad\rightarrow\quad \hat{d}_t=\bar{y}_t-\sum_{a=1}^{p}\hat{\beta}_a\bar{x}_{at}-\hat{\beta}_0$$

第１式の関係を用いると、$\bar{\hat{c}}=\bar{\hat{d}}=0$を得る。

固定効果と時間効果の直接解法との比較

ここまでは固定効果と時間効果を回帰式から消して重回帰分析を実行し、その後、これらを求める方法を説明してきたが、これらの効果を直接回帰式に含めて計算する方法と結果を比較してみよう。これらの効果を回帰式に含めると回帰式は以下のように書かれる。

$$
\begin{aligned}
y_{it} &= \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + c_i + d_t + u_{it} \\
&= \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 \cdot 1 + \sum_{k=2}^{n} c_k \delta_{ki} + \sum_{t'=2}^{T} d_{t'} \delta_{t't} + u_{it} \\
&= \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 x_{0it} + \sum_{k=2}^{n} c_k C_{kit} + \sum_{t'=2}^{T} d_{t'} D_{t'it} + u_{it}
\end{aligned}
$$

ここで、

$$x_{0it} = 0, \quad C_{kit} = \delta_{ki}, \quad D_{t'it} = \delta_{t't}$$

これを用いて最小２乗法を計算してみよう。

$$
L = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \sum_{b=1}^{p} \beta_b x_{bit} - \beta_0 - \sum_{l=1}^{n} c_l \delta_{li} - \sum_{m=1}^{T} d_m \delta_{mt} \right)^2
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \beta_0} &= -2 \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \sum_{b=1}^{p} \beta_b x_{bit} - \beta_0 - \sum_{l=1}^{n} c_l \delta_{li} - \sum_{m=1}^{T} d_m \delta_{mt} \right) \\
&= -2nT \left(\bar{y} - \sum_{b=1}^{p} \beta_b \bar{x}_b - \beta_0 - \bar{c} - \bar{d} \right) \\
&= -2nT \left(\bar{y} - \sum_{b=1}^{p} \beta_b \bar{x}_b - \beta_0 \right) = 0
\end{aligned}
$$

$$\beta_0 = \bar{y} - \sum_{b=1}^{p} \beta_b \bar{x}_b$$

$$
\begin{aligned}
\frac{\partial L}{\partial c_k} &= -2 \sum_{i=1}^{n} \sum_{t=1}^{T} \delta_{ki} \left(y_{it} - \sum_{b=1}^{p} \beta_b x_{bit} - \beta_0 - \sum_{l=1}^{n} c_l \delta_{li} - \sum_{m=1}^{T} d_m \delta_{mt} \right) \\
&= -2 \sum_{t=1}^{T} \left(y_{kt} - \sum_{b=1}^{p} \beta_b x_{bkt} - \beta_0 - c_k - d_t \right) \\
&= -2 \sum_{t=1}^{T} \left(y_{kt} - \sum_{b=1}^{p} \beta_b x_{bkt} - \beta_0 - c_k \right) = 0
\end{aligned}
$$

$$c_k = \bar{y}_k - \sum_{b=1}^{p} \beta_b \bar{x}_{bk} - \beta_0$$

同様に、

$$d_t = \bar{y}_t - \sum_{b=1}^{p} \beta_b \bar{x}_{bt} - \beta_0$$

$$\frac{\partial L}{\partial \beta_a} = -2 \sum_{i=1}^{n} \sum_{t=1}^{T} x_{ait} \left(y_{it} - \sum_{b=1}^{p} \beta_b x_{bit} - \beta_0 - c_i - d_t \right)$$

$$= -2 \left[\sum_{i=1}^{n} \sum_{t=1}^{T} x_{ait} \left(\tilde{y}_{it} - \sum_{b=1}^{p} \beta_b \tilde{x}_{bit} \right) \right]$$

$$= -2 \left[\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{ait} \left(\tilde{y}_{it} - \sum_{b=1}^{p} \beta_b \tilde{x}_{bit} \right) \right] = 0$$

これで結果が一致した。

実際の計算では、c_i の第 1 成分を 0 にして計算するため、$nc = \sum_{i=2}^{n} c_i$ として、$c_i' = c_i - c$ がここでの計算の c_i に相当する。d_t についても同様である。これにより、$\beta_0' = \beta_0 + c + d$ がここでの計算の β_0 に相当する。

補遺　パネルデータ分析の欠損値について

ここでは、これまで一定とみなしてきた、個体あたりの観測時間数（合わせて時間当たりの観測個体数も）が一定でない場合の計算法を考えてみる。我々は以前述べた方法の平均の取り方を変えることにより実現可能と考える。

パネルデータの目的変数を y_{it} （$i = 1, \cdots, n_t \le n$， $t = 1, \cdots, T_i \le T$ ）、説明変数を x_{ait} （$a = 1, \cdots, p$）、個体の固定効果を c_i、時間効果を d_t、定数項を β_0、誤差を u_{it} として、以下のモデルを考える。これまでは個体からみた時点の数や時点からみた個体の数は同じとしていたが、ここでは異なる場合も含むふぞろいなデータと仮定してみる。

$$\sum_{i=1}^{n} T_i = \sum_{t=1}^{T} n_t = N$$

$$Y_{it} = \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + \delta_f c_i + \delta_t d_t$$

$$y_{it} = Y_{it} + u_{it} = \sum_{a=1}^{p} \beta_a x_{ait} + \beta_0 + \delta_f c_i + \delta_t d_t + u_{it}$$

ここに、δ_f は固定効果を考える場合は 1、考えない場合は 0 を与える定数で、δ_t は時間効果を考える場合は 1、考えない場合は 0 を与える定数である。また、固定効果、時間効果、誤差について、以下を仮定する（ふぞろいなデータではここが問題かも知れない）。

$$\bar{c} = \bar{d} = 0, \ \bar{u}_i = \bar{u}_t = \bar{u} = 0$$

また、変数については以下のような変換を考える。

$$\tilde{y}_{it} = y_{it} - \delta_f \bar{y}_i - \delta_t \bar{y}_t + \delta_f \delta_t \bar{y}, \quad \tilde{x}_{ait} = x_{ait} - \delta_f \bar{x}_{ai} - \delta_t \bar{x}_{at} + \delta_f \delta_t \bar{x}_a,$$

$$\tilde{u}_{it} = u_{it} - \delta_f \bar{u}_i - \delta_t \bar{u}_t + \delta_f \delta_t \bar{u} = u_{it}$$

ここに、

$$\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\left[\sum_{a=1}^{p}\beta_a x_{ait} + \delta_f c_i + \delta_t d_t + \beta_0 + u_{it}\right] = \sum_{a=1}^{p}\beta_a \bar{x}_{ai} + \delta_f c_i + \beta_0$$

$$\bar{y}_t = \frac{1}{n_t}\sum_{i=1}^{n_t}\left[\sum_{a=1}^{p}\beta_a x_{ait} + \delta_f c_i + \delta_t d_t + \beta_0 + u_{it}\right] = \sum_{a=1}^{p}\beta_a \bar{x}_{at} + \delta_t d_t + \beta_0$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left[\sum_{a=1}^{p}\beta_a x_{ait} + \delta_f c_i + \delta_t d_t + \beta_0 + u_{it}\right] = \sum_{a=1}^{p}\beta_a \bar{x}_a + \beta_0$$

この変換によって、固定効果と時間効果の項は消え、以下のような関係を得る。

$$\tilde{y}_{it} = \sum_{a=1}^{p}\beta_a \tilde{x}_{ait} + \delta_f c_i - \delta_f^2 c_i + \delta_t d_t - \delta_t^2 d_t + \beta_0(1-\delta_f-\delta_t+\delta_f\delta_t) + u_{it}$$

$$= \sum_{a=1}^{p}\beta_a \tilde{x}_{ait} + (1-\delta_f)(1-\delta_t)\beta_0 + u_{it}$$

これにより、予測値は以下のようになる。

$$\tilde{Y}_{it} = \sum_{a=1}^{p}\beta_a \tilde{x}_{ait} + (1-\delta_f)(1-\delta_t)\beta_0$$

ここで実際のプログラムでは、$(1-\delta_f)(1-\delta_t)\beta_0$ 部分を改めて β_0 として計算している。後に述べるように、固定効果や時間効果がある場合、最小 2 乗法の計算から自動的に予測値 $\hat{\beta}_0$ は 0 になる。

これ以後は、通常のパネルデータ分析と同じような処理になる。

参考文献

[1] J.H.Stock,M.W.Watson,宮尾龍蔵訳, 入門計量経済学, 共立出版, 2016.

２７．テキスト CR 分析

27.1　プログラムについて

文書の出現単語を行、文書名を列として、単語の出現数の 2 次元分割表を作り、コレスポンデンス分析を用いて、文書を分類する分析が行われることがあるが、我々はこれをテキスト CR 分析と呼ぶことにする。テキスト CR 分析は、通常のコレスポンデンス分析に比べて以下のような特徴がある。１つは単語の出現数をそのまま使うかどうか、もう１つは出現単語のすべてを取って分析するのか一部を利用するのかである。これらの問題に対して我々は参考文献[1]で、一応以下のような結論を得た。前者に対しては文書の長さを変えると単語数も変わり、分析結果も変わることから、単語数は文書ごとにある一定の数に標準化して利用する方がよい。また、後者に対してはある程度安定的な答えが出る必要性から、分割表の中で 0 の占める割合の 0 比率というものを考えて、これが、0.2 程度以下がよいと結論した。また、同じ文献の中で新しい標準化の方法も提案した。これらの結果を元に、我々はこのテキスト CR 分析に特化した分析を College Analysis の中に組み込むことにした。

メニュー「分析－多変量解析他－分類手法－テキスト CR 分析」を選択すると図 1 のような分析実行画面が表示される。

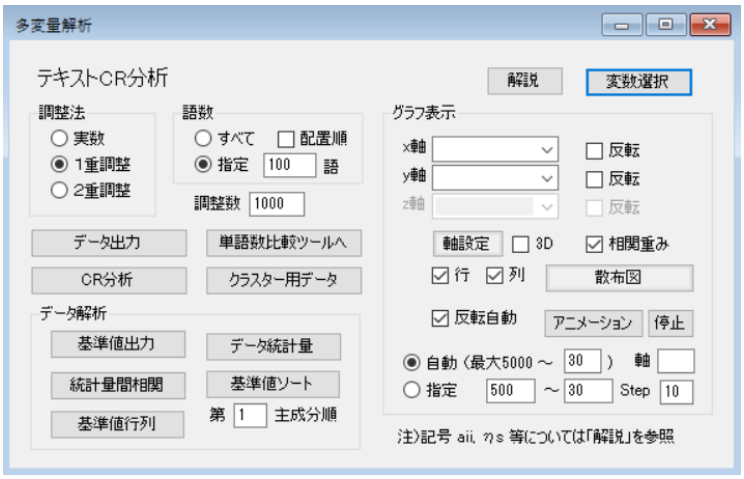


図１　分析実行画面

この画面は、大きく３つの部分に分かれている。左上は基本的な分析ツールであり、この部分がテキスト CR 分析の本体である。右側は結果をグラフやアニメーションで表示する部分である。左下は分析結果に現れる成分やグラフの軸について考察を加えるためのデータ解析の部分である。この分析実行画面について、次節の単語比較ツールに続いて、順を追って機能別にプログラムの動きを見て行くことにする。

27.2　単語比較ツール

テキスト CR 分析では、まず複数の文書から単語の数を取り出し、テキスト間で共通する単語について１つにまとめ、すべての文書の語数の合計順に並べ替えるという前処理が必

要である。この処理を簡単に行うために、ここでは以前に作成したツールについて紹介する。

メニュー［ツールー単語比較ツール］を選択するか、1 節図 1 の「単語比較ツールへ」ボタンをクリックすると、図 1 のような「単語比較ツール」実行画面が表示される。

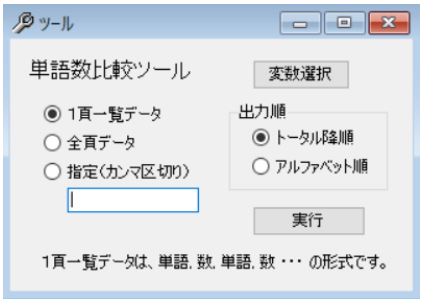
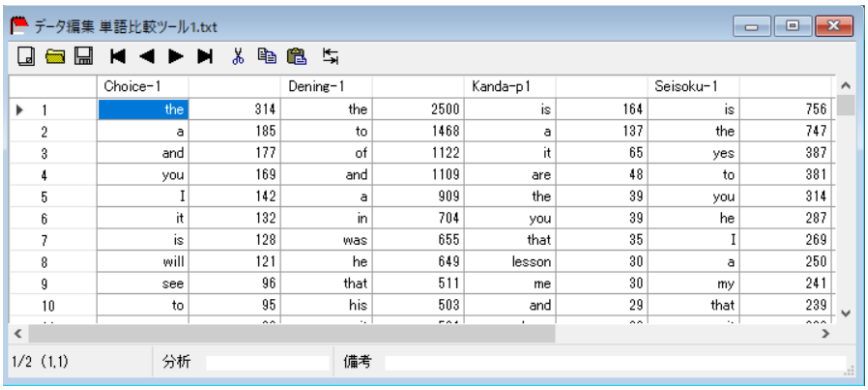


図 1　単語比較ツール実行画面

単語比較のためには、図 2 のように 1 頁に単語とその数、単語とその数、…と並んだデータか、各頁に単語とその数が与えられたデータか、どちらか必要である。単語の並びについては図 2 では文書ごとに降順になっているが、特に指定はない。

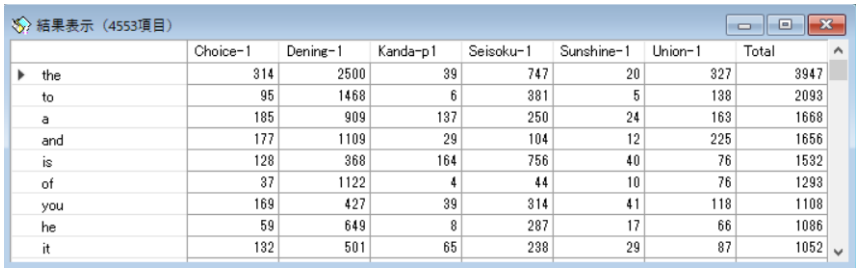


	Choice-1		Dening-1		Kanda-p1		Seisoku-1	
1	the	314	the	2500	is	164	is	756
2	a	185	to	1468	a	137	the	747
3	and	177	of	1122	it	65	yes	387
4	you	169	and	1109	are	48	to	381
5	I	142	a	909	the	39	you	314
6	it	132	in	704	you	39	he	287
7	is	128	was	655	that	35	I	269
8	will	121	he	649	lesson	30	a	250
9	see	96	that	511	me	30	my	241
10	to	95	his	503	and	29	that	239

図 2　単語比較のデータ（単語比較ツール 1.txt）

図 2 で与えられた前者の場合は、単語比較ツール実行画面の「1 頁一覧データ」を選択し、変数選択で、利用する文書の単語と数の組を指定する。後者の 1 頁 1 文書の場合は、「全ページ」ラジオボタンか、「指定（カンマ区切り）」ラジオボタンを選択し、後者の場合はそのページ番号を下のテキストボックスにカンマ区切りで入れておく。

出力は、選択文書全体の語数合計降順の「トータル降順」か「アルファベット順」が選べる。通常、データ形式は「1 頁一覧データ」、出力順は「トータル降順」がよい。この後「実行」ボタンをクリックすると図 3 に示す実行結果が表示される。この結果は単語が頻度順に並べられている。



	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	Total
the	314	2500	39	747	20	327	3947
to	95	1468	6	381	5	138	2093
a	185	909	137	250	24	163	1668
and	177	1109	29	104	12	225	1656
is	128	368	164	756	40	76	1532
of	37	1122	4	44	10	76	1293
you	169	427	39	314	41	118	1108
he	59	649	8	287	17	66	1086
it	132	501	65	238	29	87	1052

図 3　単語比較ツール出力結果

我々のテキスト CR 分析プログラムは、図 3 の形式のデータを用いるが、単語数の合計を表す「Total」の欄は、分析に不要である。しかし、後に変数選択の中で落とすことができるので、あっても問題はない。このデータは新規に作成されたデータとしても、既存のデータの最後の頁に追加して使うこともできる。後者の場合は、グリッド出力メニュー［編集ーエディタ頁追加］を利用すると便利である。

27.3　基本分析ツール

2 章の図 1、分析実行画面の基本分析ツールの部分を切り取って図 1 に再掲する。

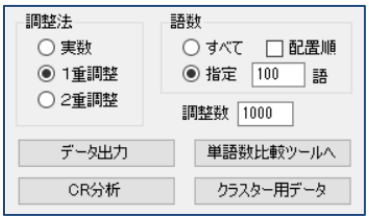


図 1　分析実行画面中の基本分析ツール

テキスト CR 分析では単語数の調整を行うが、このプログラムでは、単語の頻度をそのまま利用する「実数」、単語の頻度をそろえる「１重調整」、単語の頻度をそろえた上で分析に利用する単語数を設定し再度頻度をそろえる「２重調整」の方法を扱うことができる。利用する単語数は「すべて」か、後ろに語数を指定した「指定」を選択できる。このメニューではデフォルトとして、調整法は「１重調整」、語数は「指定」100 語にしている。語数の「調整数」は分析に直接影響を与えないが、「データ出力」の際には値が変わってくるので、見た目が良い程度で記入しておく。デフォルトは 1000 になっている。

「変数選択」で Total を除くすべての変数（文書）を選択し、図 4 の「データ出力」ボタンをクリックすると、図 2 のような出力結果を得る。

データ出力

	出現text数	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	合計	0比率	順位
old	4	1.777	0.885	8.979	0.000	0.000	1.644	13.283	0.077	82
many	6	1.421	0.572	4.489	2.659	3.328	0.730	13.200	0.076	83
run	5	6.751	0.312	0.000	3.090	0.832	2.009	12.994	0.077	84
when	5	1.066	5.152	0.000	1.222	0.832	4.565	12.836	0.078	85
please	5	0.711	0.442	8.418	0.000	2.496	0.730	12.797	0.079	86
day	5	1.599	2.680	0.000	0.719	2.496	5.113	12.606	0.080	87
pat	1	0.000	0.000	0.000	0.000	12.479	0.000	12.479	0.089	88
rat	4	3.731	0.000	4.489	3.450	0.000	0.730	12.400	0.092	89
mother	5	0.711	1.639	2.806	4.168	0.000	2.739	12.063	0.093	90
oh	3	0.000	0.000	0.000	0.647	9.983	1.278	11.908	0.097	91
had	4	0.355	7.285	0.000	0.216	0.000	4.018	11.873	0.100	92
come	6	2.487	1.119	2.245	2.443	0.832	2.739	11.865	0.099	93
out	4	2.665	3.304	0.000	1.150	0.000	4.565	11.684	0.101	94
program	1	0.000	0.000	0.000	0.000	11.647	0.000	11.647	0.109	95
people	3	0.000	2.290	0.000	0.000	9.151	0.183	11.624	0.113	96

図 2　データ出力結果

この結果は一度 1000 語に調整を実行して、その中で頻度の上位から指定語数を選択して表示したものである。これが分析に使うデータである。この中には、参考のために、調整後の単語の合計数や 0 比率などが表示されている。ここでは例として、総頻度が 82 位から 96 位までを表示しているが、この中で水色の網掛けの単語がある。これは１つの文書以外では頻度が 0 の単語である。0 比率が低いところの網掛けの単語では、本来利用しない固有名詞

などが残っている場合があり、そのような場合にはデータから削除する。データの削除にはエディタのメニュー［ツールー検索］で表示される検索画面で、「行名検索」機能を用いるとよい。

ここで単語の並び順に対して、１つだけ例外を述べておく。単語を「すべて」選択した場合、「配置順」チェックボックスにチェックを入れると、頻度順ではなく、元の単語の並び順に出力される。これは、特別な単語を入れてその振る舞いを観察する 5 節のデータ解析の際に利用する。

「CR 分析」ボタンをクリックすると、指定された調整法で、指定された語数でコレスポンデンス分析を実行する。但し、単語数は文書数より多くする必要がある。実行結果を図３に示す。

コレスポンデンス分析結果

0比率: 0.112	群	第1成分	第2成分	第3成分	第4成分	第5成分	重み1成分	重み2成分	重み3成分	重み4成分	重み5成分
固有値		0.182	0.146	0.079	0.057	0.016					
相関係数		0.427	0.383	0.281	0.239	0.127					
寄与率		0.379	0.305	0.164	0.119	0.034					
累積寄与率		0.379	0.684	0.848	0.966	1.000					
Choice-1	2	0.161	0.286	0.203	1.787	-1.140	0.069	0.109	0.057	0.427	-0.145
Dening-1	2	1.268	0.731	0.775	-1.513	-0.995	0.541	0.280	0.218	-0.361	-0.126
Kanda-p1	2	-1.856	0.604	0.472	-0.453	0.141	-0.792	0.231	0.133	-0.108	0.018
Seisoku-1	2	0.094	0.052	-2.134	-0.331	0.004	0.040	0.020	-0.600	-0.079	0.001
Sunshine-1	2	-0.019	-2.245	0.449	-0.230	0.031	-0.008	-0.859	0.126	-0.055	0.004
Union-1	2	0.841	0.502	0.442	0.515	2.008	0.359	0.192	0.124	0.123	0.255
the	1	0.908	0.691	-0.112	0.008	-0.132	0.388	0.264	-0.031	0.002	-0.017
is	1	-1.441	0.049	-0.620	-0.631	0.231	-0.615	0.019	-0.174	-0.151	0.029
a	1	-0.945	0.574	0.798	-0.148	0.403	-0.403	0.220	0.224	-0.035	0.051

図３　コレスポンデンス分析結果

同じ処理を通常のコレスポンデンス分析のメニューで実施すると、最初に単語（行名）が表れるようになっているが、ここでは文書の類似性の方が重要であるので、文書名（列名）が最初に並ぶように設定している。表示の項目の意味については、補遺を参照してもらいたいが、特に寄与率と累積寄与率は重要である。

コレスポンデンス分析の結果を用いてクラスター分析を行い、すべての次元を参照して分類することも可能である。その際、クラスター分析では相関の重み付き成分を利用する方が現実的であるため、「クラスター用データ」ボタンをクリックすると図３の四角で囲んだ部分を出力するようにしている。結果を図４に示す。

クラスター分析用CRデータ

	重み1成分	重み2成分	重み3成分	重み4成分	重み5成分
Choice-1	0.069	0.109	0.057	0.427	-0.145
Dening-1	0.541	0.280	0.218	-0.361	-0.126
Kanda-p1	-0.792	0.231	0.133	-0.108	0.018
Seisoku-1	0.040	0.020	-0.600	-0.079	0.001
Sunshine-1	-0.008	-0.859	0.126	-0.055	0.004
Union-1	0.359	0.192	0.124	0.123	0.255

図４　クラスター用データ出力

これをクラスター分析のプログラムのデータとしてデンドログラムを描くことになるが、距離測定法は重み付けをしたことを考慮して、平方ユークリッド距離、クラスター構成法は標準的なウォード法が適していると考える。これらの設定での結果を図８に示す。

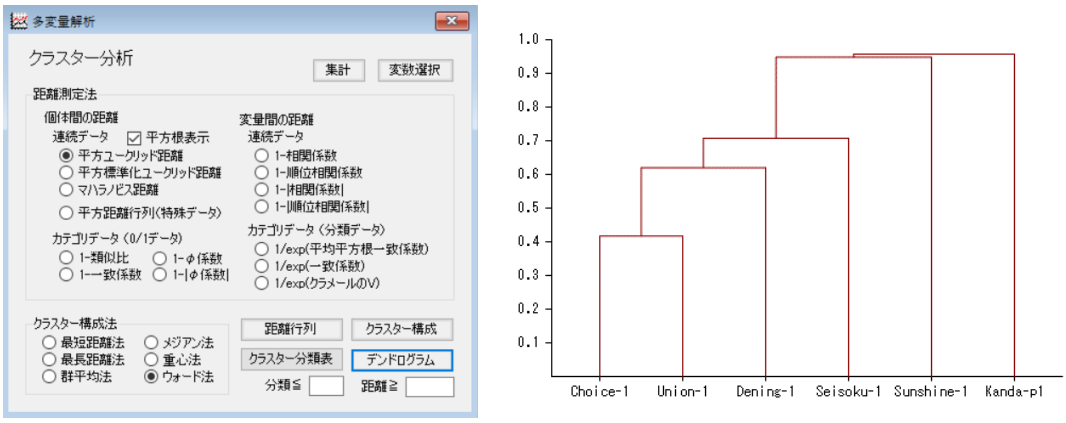


図８　クラスター分析の実行画面とデンドログラム

27.4　グラフ描画とアニメーションツール

次にテキスト CR 分析の結果のグラフ表示を考える。図 1 に分析実行画面のグラフに関する部分を切り取って表示した。

図１　分析実行画面中のグラフ表示

分析結果を表示するには、「軸設定」ボタンをクリックして、成分を各軸に割り当てる。例えば、x 軸を第 1 成分に、y 軸を第 2 成分にし、「相関重み」を加え、その他の設定をデフォルトの設定にして、「散布図」ボタンをクリックした結果を図 2 に示す。

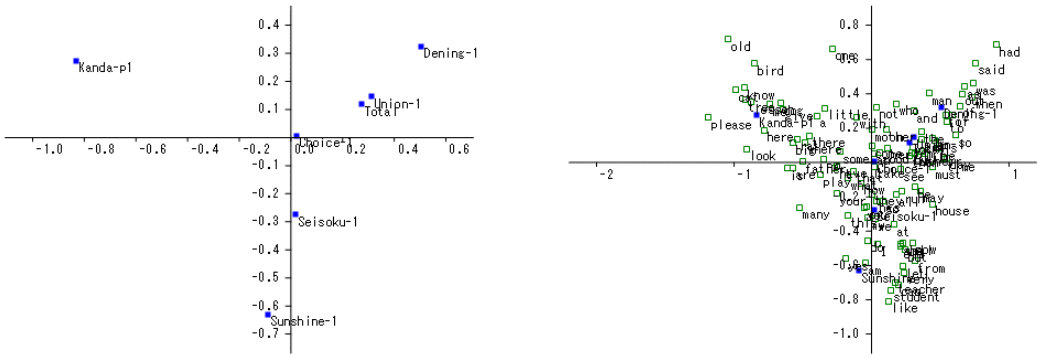


図２　コレスポンデンス分析による散布図

左が「列」成分だけの表示、右が「行」成分も含めた表示である。

同様に、「3D」チェックボックスをチェックし、z 軸を第 3 成分にして、その他の設定を

図 2 と同じにした散布図を図 3 に示す。但し、分かりにくいのでここでは「列」成分だけにしている。

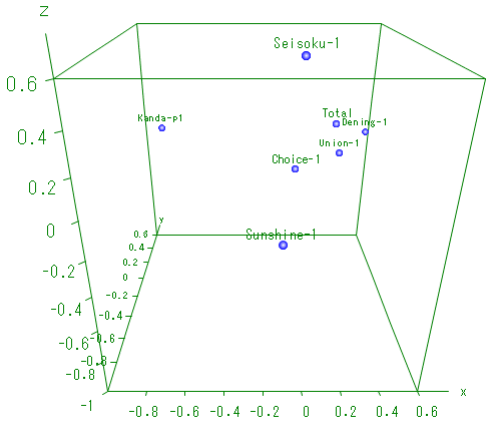


図 3　コレスポンデンス分析による散布図（３次元表示）

我々は利用する語数を 100 語に固定してこれまでの計算を行ってきたが、これは 0 比率の値を参考にしながら決めた値である。しかし、語数を決定するとき結果の安定性は重要である。そこで、結果が語数によってどのように変化するかをアニメーションで表示する試みを思い付いた。これは指定された最大語数から、徐々に選択語数を減らして行き、最終的に指定された最小語数まで、散布図が変わって行く様子をアニメーションのように表示する機能である。この動きは紙面上で表現できないが、変化の過程の文書と単語の配置の安定性によってコレスポンデンス分析の正当性を確認する方法である。

この設定では、単語数の変化を「自動」にするか、「指定」にするか設定できる。「軸」に数値を設定すると絶対値がその数値までの範囲が表示される。図 4 にその過程を簡単に示す。実際に動かしてみると大変興味深いので試してもらいたい。

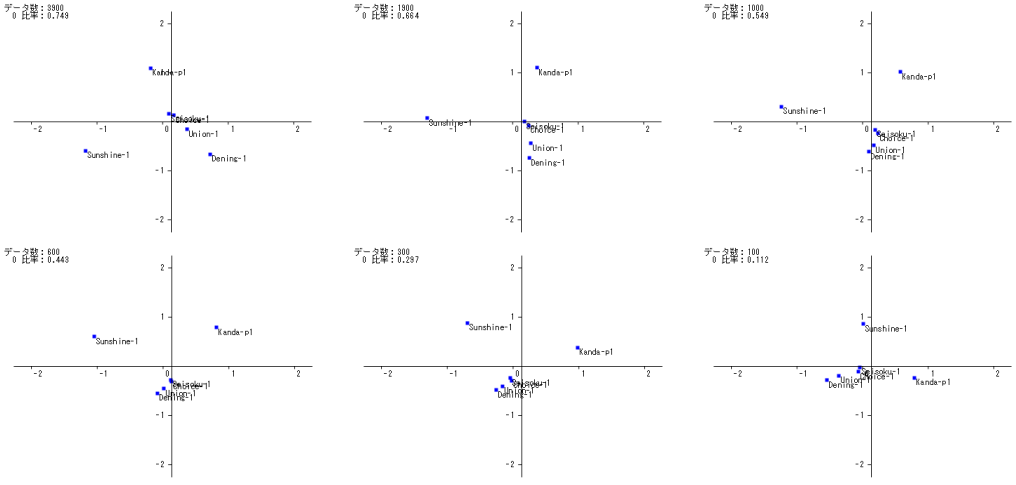


図 4　アニメーション表示の例

27.5　データ解析ツールと成分の解釈

CR 分析では成分の意味が明確でない。これは因子分析などと異なる CR 分析の特徴である。特に、テキスト CR 分析では教科書（この節では文書の代わりに教科書を使う）によって単語の数が極端に違う場合があり、この単語の数が教科書の大きな特徴になっている。しかし、この単語数にしてもどの成分が単語数を表しているのか明確ではなく、単語数の似た教科書どうしの比較では、単語数と成分にはあまり関係の見られないこともある。では、これらを調べるには何を見ればよいのか。ここでは定性的な議論であるが、３つの教科書の組についてテキスト CR 分析の特徴を見て行くことにする。

３つの教科書の組としては、１）語数の適度に異なる現代の教科書の組、２）語数の極端に異なる明治期と現代の教科書の組、３）語数の揃った明治期の教科書の組を考える。これらについて１）ではテキスト CR 分析 2.txt、２）ではテキスト CR 分析 1.txt (p1)、３）ではテキスト CR 分析 1.txt (p2) を利用する。

分かり易いように、図 1 に分析実行画面からデータ解析ツールの部分を切り抜いた画面を示しておく。

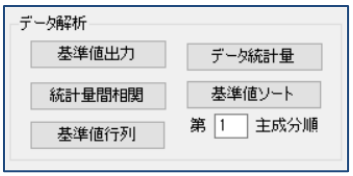


図１　分析実行画面中のデータ解析

１）の場合

テキスト CR 分析では単語をある語数で切り取って分析する。そのため、教科書ごとの頻度が 0 の単語の比率である「文書 0 比率」が重要である。文書 0 比率は単語数の少ない教科書では大きくなる傾向がある。表 1 に実際の結果を示す。切り取る単語数によらず、全単語数との相関係数に大きな変動はない。これは我々の予想と異なり、単語数が適当に異なる教科書間の興味ある特徴である。

表１　1 重調整法による文書 0 比率と全単語数との相関係数

語数	50	100	300	500	1000	全単語数
C5	0.140	0.220	0.470	0.548	0.696	1046
C6	0.120	0.210	0.410	0.508	0.678	1085
NH5	0.120	0.290	0.530	0.634	0.767	993
NH6	0.140	0.210	0.407	0.542	0.698	1494
SS5	0.180	0.300	0.453	0.524	0.666	1369
SS6	0.080	0.210	0.403	0.482	0.639	1844
NC1	0.000	0.000	0.053	0.148	0.305	7266
NC2	0.000	0.000	0.050	0.130	0.236	9954
NC3	0.000	0.020	0.083	0.164	0.278	10322
NH1	0.000	0.010	0.043	0.128	0.304	8778
NH2	0.000	0.020	0.067	0.172	0.326	10714
NH3	0.000	0.010	0.107	0.200	0.320	9922
SS1	0.000	0.010	0.083	0.172	0.345	6252

SS2	0.000	0.020	0.120	0.232	0.353	6499
SS3	0.000	0.010	0.070	0.166	0.299	9435
相関係数	-0.912	-0.924	-0.943	-0.943	-0.959	

切り取られたデータから作られた基準値（補遺(A3)式を参照）を $x_{i\lambda}$ とすると、以下のような関係が見られる。

$$\frac{1}{m}\sum_{\lambda=1}^{m} x_{i\lambda} = \frac{1}{m}\sum_{\lambda=1}^{m} n_{i\lambda} \Big/ \sqrt{n_{i\cdot} n_{\cdot\lambda}} \simeq c_m \qquad (1)$$

ここに c_m は教科書の種類 i によらず、切り取った数 m だけによる定数である。これは標準化の操作を行ったテキスト CR 分析の特徴かも知れない。我々はこの指標を「基準値平均」と名付けることにする。

この関係を実際のデータで見てみよう。表 2 に結果を示す。

表 2　基準値平均とその標準偏差

語数	50	100	300	500	1000
C5	0.0334	0.0253	0.0131	0.0108	0.0068
C6	0.0359	0.0235	0.0128	0.0098	0.0063
NH5	0.0389	0.0250	0.0122	0.0086	0.0054
NH6	0.0359	0.0228	0.0132	0.0095	0.0060
SS5	0.0322	0.0213	0.0140	0.0113	0.0074
SS6	0.0327	0.0224	0.0138	0.0107	0.0070
NC1	0.0333	0.0233	0.0129	0.0095	0.0066
NC2	0.0310	0.0213	0.0120	0.0091	0.0066
NC3	0.0305	0.0216	0.0119	0.0089	0.0062
NH1	0.0362	0.0239	0.0128	0.0097	0.0063
NH2	0.0323	0.0234	0.0126	0.0094	0.0063
NH3	0.0325	0.0222	0.0123	0.0090	0.0064
SS1	0.0343	0.0231	0.0125	0.0093	0.0065
SS2	0.0327	0.0231	0.0123	0.0091	0.0065
SS3	0.0315	0.0220	0.0120	0.0093	0.0065
標準偏差	0.0023	0.0012	0.0006	0.0008	0.0004

これを見ると標準偏差は値の 10%以下であり、近似は良い結果を与えている。

次に、$a_{ii} = \sum_{\lambda=1}^{m} x_{i\lambda}^2$ で与えられる基準値で作られた基準値行列（補遺(A2)式参照）の対角成分と文書 0 比率の関係を見てみよう。大まかではあるが、以下の関係が見られるようである。

$$(1-\eta_i)a_{ii} = (1-\eta_i)\sum_{\lambda=1}^{m} x_{i\lambda}^2 \simeq d \qquad (2)$$

ここに d は教科書の種類 i にも切り取った単語数 m にもよらない定数である。我々はこの指標を「対角指標」と名付けることにする。表 3 でこの関係を見てみよう。

表 3　対角指標

語数	50	100	300	500	1000
C5	0.0776	0.0971	0.0820	0.0944	0.0694
C6	0.0878	0.0824	0.0739	0.0706	0.0549
NH5	0.1258	0.1089	0.0792	0.0645	0.0467

NH6	0.0950	0.0881	0.0895	0.0791	0.0592
SS5	0.0734	0.0692	0.0902	0.0919	0.0758
SS6	0.0817	0.0783	0.0869	0.0891	0.0725
NC1	0.0707	0.0777	0.0821	0.0774	0.0745
NC2	0.0671	0.0701	0.0755	0.0749	0.0769
NC3	0.0693	0.0717	0.0746	0.0727	0.0714
NH1	0.0768	0.0778	0.0805	0.0861	0.0740
NH2	0.0709	0.0776	0.0833	0.0827	0.0764
NH3	0.0765	0.0802	0.0839	0.0790	0.0806
SS1	0.0801	0.0799	0.0804	0.0763	0.0713
SS2	0.0714	0.0755	0.0760	0.0704	0.0705
SS3	0.0750	0.0778	0.0810	0.0841	0.0811

この指標についての全体の平均は 0.0784、標準偏差は 0.0106 である。

次に、これらの指標を含めて、テキスト CR 分析の成分の性質、特に単語数に結び付いた成分を調べる際に重要と思われる指標について考える。図 2 に分析実行画面の「データ統計量」ボタンをクリックした結果を示す。ここではデータ数を 300 にしている。

CR分析統計量（注：単語数と文書0比率には比例関係があります。他の分析に使うときにはご注意下さい。）

	単語数	文書0比率	頻度合計	頻度平均	頻度偏差	基準値平均	基準値偏差	aii	(1−αs)aii	第1成分	第2成分	第3成分
C5	159	0.470	733.270	2.444	5.402	0.0131	0.0185	0.1547	0.0820	−0.814	−1.244	−2.526
C6	177	0.410	764.055	2.547	5.674	0.0128	0.0159	0.1253	0.0739	−0.509	0.819	−0.178
NH5	141	0.530	846.928	2.823	7.672	0.0122	0.0203	0.1684	0.0792	−1.436	−0.237	2.068
NH6	178	0.407	781.124	2.604	5.938	0.0132	0.0182	0.1508	0.0895	−0.723	2.089	0.433
SS5	164	0.453	685.172	2.284	4.434	0.0140	0.0188	0.1649	0.0902	−1.827	−2.034	0.300
SS6	179	0.403	719.631	2.399	4.998	0.0138	0.0171	0.1456	0.0869	−0.868	1.761	−1.537
NC1	284	0.053	673.961	2.247	4.147	0.0129	0.0110	0.0868	0.0821	0.311	−0.152	−0.090
NC2	285	0.050	615.933	2.053	4.122	0.0120	0.0109	0.0795	0.0755	0.927	−0.118	0.081
NC3	275	0.083	604.825	2.016	4.142	0.0119	0.0114	0.0814	0.0746	1.111	−0.326	0.121
NH1	287	0.043	719.526	2.398	4.624	0.0128	0.0108	0.0842	0.0805	0.226	−0.028	0.269
NH2	280	0.067	666.044	2.220	4.259	0.0126	0.0118	0.0893	0.0833	1.099	−0.222	0.104
NH3	268	0.107	639.790	2.133	4.254	0.0123	0.0127	0.0939	0.0839	1.281	−0.221	0.137
SS1	275	0.083	678.983	2.263	4.706	0.0125	0.0117	0.0877	0.0804	0.182	−0.121	0.245
SS2	264	0.120	655.332	2.184	4.323	0.0123	0.0117	0.0863	0.0760	0.899	−0.086	0.193
SS3	279	0.070	616.958	2.057	4.338	0.0120	0.0121	0.0871	0.0810	1.239	−0.322	0.150

図 2　「データ統計量」実行結果

ここでは、開発者が重要であると考える指標が教科書ごとに並んでいるが、教科書ごとの文書 0 比率は単語数と関係のある重要な指標であろう。また、基準値から作られる基準値行列 a_{ij} は、固有方程式を与えることから重要な要素であるが、特に対角成分 a_{ii} は各教科書のデータのばらつきを与えるものである。またこの指標は(2)式から文書 0 比率と関係しているとも考えられる。同様にして、教科書ごとの基準値の標準偏差も意味を持つかも知れない。これに、各教科書の固有ベクトル成分を 3 つまで加え、検討すべき指標と考えた。これらの指標については、青色に網掛がされており、簡単に教科書ごとの相関を見ることができるようになっている。

これに対して、上で述べた基準値平均や対角指標は、あまり教科書による変動が期待されないので、確認をするためのデータである。また、基準値の元となる頻度については、直接固有方程式の行列を与えるものではないので、網掛けが行われていない。もちろん相関を求めることが必要な場合は、図 2 のデータをグリッドエディタにそのままコピーし、相関を調べることもできる。

次に、先に述べた網掛けの指標の相関を求めてみよう。図 1 のメニューの中の「統計量間相関」ボタンをクリックすると、図 3 のような主要統計量間の相関行列が得られる。

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.984	0.977	−0.898	0.088	−0.123
基準値偏差	0.984	1.000	0.994	−0.888	0.053	−0.066
aii	0.977	0.994	1.000	−0.922	0.037	−0.091
第1成分	−0.898	−0.888	−0.922	1.000	0.006	0.042
第2成分	0.088	0.053	0.037	0.006	1.000	−0.007
第3成分	−0.123	−0.066	−0.091	0.042	−0.007	1.000

図３　主要統計量間の相関行列

ここでは文書 0 比率と第 1 成分とが強い相関を持っているので、第 1 成分が単語数を通じた難易度を表しているものと解釈できる。

テキスト CR 分析の固有方程式の行列を与える基準値行列a_{ij}については、図 1 のメニューで「基準値行列」ボタンをクリックすると、図 4 のように与えられる。

基準値行列出力

基準値行列	C5	C6	NH5	NH6	SS5	SS6	NC1	NC2	NC3	NH1	NH2	NH3	SS1	SS2	SS3
C5	0.155	0.068	0.070	0.058	0.083	0.074	0.064	0.053	0.054	0.064	0.058	0.053	0.062	0.055	0.053
C6	0.068	0.125	0.078	0.081	0.066	0.080	0.066	0.060	0.055	0.071	0.056	0.057	0.065	0.062	0.053
NH5	0.070	0.078	0.168	0.085	0.095	0.069	0.062	0.052	0.052	0.074	0.055	0.051	0.070	0.059	0.051
NH6	0.058	0.081	0.085	0.151	0.060	0.089	0.062	0.057	0.053	0.067	0.058	0.054	0.064	0.058	0.052
SS5	0.083	0.066	0.095	0.060	0.165	0.062	0.060	0.050	0.047	0.059	0.046	0.043	0.061	0.047	0.043
SS6	0.074	0.080	0.069	0.089	0.062	0.146	0.061	0.054	0.049	0.061	0.052	0.050	0.060	0.055	0.048
NC1	0.064	0.066	0.062	0.062	0.060	0.061	0.087	0.063	0.062	0.070	0.066	0.062	0.073	0.065	0.064
NC2	0.053	0.060	0.052	0.057	0.050	0.054	0.063	0.079	0.070	0.064	0.073	0.073	0.060	0.070	0.069
NC3	0.054	0.055	0.052	0.053	0.047	0.049	0.062	0.070	0.081	0.062	0.074	0.077	0.060	0.070	0.075
NH1	0.064	0.071	0.074	0.067	0.059	0.061	0.070	0.064	0.062	0.084	0.067	0.066	0.075	0.068	0.066
NH2	0.058	0.056	0.055	0.058	0.046	0.052	0.066	0.073	0.074	0.067	0.089	0.077	0.064	0.075	0.077
NH3	0.053	0.057	0.051	0.054	0.043	0.050	0.062	0.073	0.077	0.066	0.077	0.094	0.063	0.073	0.077
SS1	0.062	0.065	0.070	0.064	0.061	0.060	0.073	0.060	0.060	0.075	0.064	0.063	0.088	0.063	0.063
SS2	0.055	0.062	0.059	0.058	0.047	0.055	0.065	0.070	0.070	0.068	0.075	0.073	0.063	0.086	0.071
SS3	0.053	0.053	0.051	0.052	0.043	0.048	0.064	0.069	0.075	0.066	0.077	0.077	0.063	0.071	0.087

図４　300 語での基準値行列

この行列の対角成分には黄色、各行の最も小さな値には緑色の網掛けがしてある。さらに、この表示にはまだ下があり、そこには教科書の基準値を 2 組掛け合わせた場合の 0 比率が表示されている。この 0 比率が非対角成分の下がり方に影響を与えている。

このデータの場合、第 1 成分の意味は分かったが、第 2 成分以降は単語との関係で意味が決まる。それを見るための機能が「基準値ソート」ボタンである。このボタンの下のテキストボックスに成分の番号を入力し、「基準値ソート」ボタンをクリックすると図 5 の結果が得られる。ここでは第 2 成分についての結果を表示している。

基準値ソート出力

	C5	C6	NH5	NH6	SS5	SS6	NC1	NC2	NC3	NH1	NH2	NH3	SS1	SS2	SS3	第2成	単語0	データ
memory	0.000	0.018	0.000	0.046	0.000	0.061	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.005	4.746	0.600	186
join	0.000	0.000	0.000	0.038	0.000	0.045	0.005	0.012	0.005	0.001	0.001	0.001	0.006	0.002	0.000	3.958	0.333	250
sea	0.009	0.000	0.000	0.095	0.000	0.005	0.007	0.003	0.003	0.008	0.002	0.002	0.000	0.006	0.008	3.859	0.267	153
enjoyed	0.000	0.030	0.000	0.049	0.000	0.050	0.010	0.005	0.000	0.006	0.006	0.000	0.003	0.004	0.002	3.809	0.333	118
ate	0.000	0.009	0.000	0.045	0.000	0.043	0.006	0.006	0.001	0.012	0.001	0.000	0.011	0.002	0.003	3.656	0.267	177
trip	0.000	0.025	0.000	0.053	0.000	0.030	0.008	0.001	0.002	0.010	0.003	0.002	0.005	0.004	0.007	3.531	0.200	149
swimming	0.000	0.046	0.000	0.027	0.000	0.039	0.001	0.001	0.003	0.005	0.000	0.002	0.003	0.002	0.004	3.515	0.267	195
curry	0.000	0.008	0.000	0.064	0.000	0.025	0.005	0.008	0.000	0.003	0.023	0.004	0.010	0.001	0.002	3.415	0.267	144
live	0.000	0.000	0.000	0.081	0.000	0.005	0.005	0.002	0.007	0.015	0.002	0.011	0.007	0.010	0.007	3.348	0.267	145

図５　基準値ソート結果

第 2 成分の大きい順に単語が表示され、基準値の値が示されている。上位 5 つの単語については（現在は 10 個）、最も大きな基準値の 90%以上の教科書の位置が青色に網掛けされている。これらの単語と教科書は互いに似た位置にあり、これを用いて利用者は第 2 成分

として影響力の大きな単語及びそれに近い教科書を知ることができる。同様に、第 2 成分の小さい（負の）単語についても基準値の値を知ることができる。

２）の場合

ここでは１つの教科書の単語数が多く、他も不揃いな場合を考える。語数調整した場合の文書 0 比率と全単語数との関係を表 4 に与える。

表 4　1 重調整法による文書 0 比率と全単語数との相関係数

語数	50	100	300	500	1000	全単語数
Choice-1	0.000	0.070	0.257	0.388	0.582	466
Dening-1	0.000	0.050	0.150	0.212	0.272	3844
Kanda-p1	0.100	0.260	0.517	0.656	0.800	200
Seisoku-1	0.020	0.090	0.280	0.414	0.581	736
Sunshine-1	0.120	0.180	0.420	0.528	0.662	338
Union-1	0.000	0.020	0.157	0.242	0.399	935
相関係数	-0.489	-0.489	-0.637	-0.701	-0.833	

これによると、利用する単語数が多くなると相関は高くなるが、単語数が少ないと相関が低くなり、0 比率を単語数と関連付けることは次第に難しくなる。ただ、0 比率は切り取られた単語の中でどれだけ満遍なく単語を使っているかを表す指標であり、教科書の「標準性」を表す指標のように考えられる。以下には異論があると思われるが、標準的な教科書は比較的やさしいとも考えられ、0 比率は難易度とも関係しているように思われる。ここでは 0 比率を教科書の単語数や標準性を通して難易度と関係する指標と考えて先に進む。

次に、基準値平均について１）の場合に述べたことが成立するか調べてみる。基準値平均については、表 5 の通りである。

表 5　基準値平均とその標準偏差

	50	100	300	500	1000
Choice-1	0.0579	0.0380	0.0199	0.0145	0.0091
Dening-1	0.0502	0.0333	0.0165	0.0127	0.0095
Kanda-p1	0.0555	0.0384	0.0208	0.0138	0.0075
Seisoku-1	0.0569	0.0379	0.0197	0.0146	0.0086
Sunshine-1	0.0524	0.0383	0.0216	0.0164	0.0106
Union-1	0.0514	0.0361	0.0195	0.0146	0.0105
標準偏差	0.0032	0.0020	0.0017	0.0012	0.0012

これによると教科書による標準偏差は基準値平均のほぼ 10%以内に収まっている。また、対角指標については表 6 の関係が得られた。

表 6　対角指標

	50	100	300	500	1000
Choice-1	0.2093	0.2016	0.1885	0.1684	0.1271
Dening-1	0.2172	0.2118	0.1993	0.2004	0.2156
Kanda-p1	0.2645	0.2407	0.1960	0.1436	0.0853
Seisoku-1	0.2204	0.2194	0.2120	0.1890	0.1421
Sunshine-1	0.1924	0.2289	0.2189	0.1972	0.1550
Union-1	0.1842	0.1916	0.1872	0.1808	0.1721

この指標についての全体の平均は 0.1920、標準偏差は 0.0350 である。

次に、主要統計量間の相関行列を求めてみよう。図 6a に 100 語の場合、図 6b に 500 語の場合を与える。

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.908	0.991	-0.906	-0.331	0.103
基準値偏差	0.908	1.000	0.948	-0.718	-0.210	0.168
aii	0.991	0.948	1.000	-0.881	-0.289	0.064
第1成分	-0.906	-0.718	-0.881	1.000	0.054	0.049
第2成分	-0.331	-0.210	-0.289	0.054	1.000	0.003
第3成分	0.103	0.168	0.064	0.049	0.003	1.000

図 6a　主要統計量間の相関行列（100 語）

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.925	0.937	-0.206	0.925	0.258
基準値偏差	0.925	1.000	0.976	-0.071	0.927	0.274
aii	0.937	0.976	1.000	0.066	0.965	0.270
第1成分	-0.206	-0.071	0.066	1.000	0.021	-0.007
第2成分	0.925	0.927	0.965	0.021	1.000	0.041
第3成分	0.258	0.274	0.270	-0.007	0.041	1.000

図 6b　主要統計量間の相関行列（500 語）

100 語では文書 0 比率と第 1 成分とが強い相関を持っているので、第 1 成分が難易度を表しているものと解釈できる。しかし 500 語ではむしろ第 2 成分の相関が高い。第 3 成分についてはどちらも相関が高くない。そこで、文書 0 比率を第 1 成分と第 2 成分で重回帰分析することを試みる。図 7a は 100 語、図 7b は 500 語の場合である。いずれも重回帰分析の結果と CR 分析による散布図を上下に示している。Dening-1, Kanda-p1, Sunshine-1 の位置を考えるとこれらの結果から、軸が回転している（反転も含む）ことが分かる。

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	-0.0753	-0.8908	0.0154	-4.8871	3	0.0164
第2成分	-0.0228	-0.2823	0.0147	-1.5485	3	0.2193
切片	0.1175	0.0000	0.0151	7.7761	3	0.0044
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.949	0.901	0.913	0.834	13.5932	0.0313

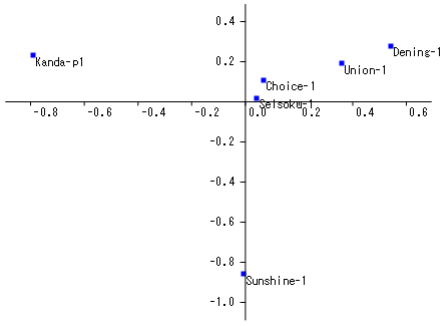


図 7a　重回帰分析と CR 分析の散布図（100 語）

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	-0.0359	-0.2263	0.0279	-1.2849	3	0.2890
第2成分	0.1442	0.9299	0.0273	5.2804	3	0.0132
切片	0.4163	0.0000	0.0272	15.3180	3	0.0006
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.952	0.907	0.919	0.845	14.6286	0.0284

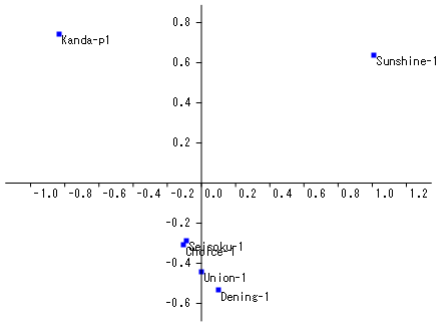


図 7b　重回帰分析と CR 分析の散布図（500 語）

重回帰分析の結果より、第 1 成分と第 2 成分の役割を変えると文書 0 比率をかなりの精度で説明していることが分かる。ではこの回転はなぜ起きるのだろうか。「基準値ソート」ボタンの下のテキストボックスを第「1」成分順にして、「基準値ソート」ボタンをクリックした結果を図 8a（100 語）と図 8b（500 語）に示す。

基準値ソート出力

	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	第1成分	単語0比率	データ順位
had	0.004	0.090	0.000	0.002	0.000	0.048	2.504	0.333	92
was	0.035	0.126	0.000	0.035	0.000	0.038	1.947	0.333	35
when	0.011	0.061	0.000	0.013	0.009	0.052	1.942	0.167	85
as	0.047	0.069	0.000	0.014	0.000	0.057	1.762	0.333	57
out	0.030	0.041	0.000	0.013	0.000	0.054	1.717	0.333	94
of	0.032	0.156	0.010	0.015	0.042	0.071	1.690	0.000	16
so	0.023	0.057	0.000	0.009	0.026	0.043	1.631	0.167	73

図 8a　基準値ソート（100 語）

基準値ソート出力

	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	第1成分	単語0比率	データ順位
pat	0.000	0.000	0.000	0.000	0.117	0.000	3.284	0.833	88
program	0.000	0.000	0.000	0.000	0.113	0.000	3.284	0.833	95
jim	0.000	0.000	0.000	0.000	0.095	0.000	3.284	0.833	129
m	0.000	0.000	0.000	0.000	0.095	0.000	3.284	0.833	129
hi	0.000	0.000	0.000	0.000	0.085	0.000	3.284	0.833	157
oka	0.000	0.000	0.000	0.000	0.080	0.000	3.284	0.833	174
doesn	0.000	0.000	0.000	0.000	0.067	0.000	3.284	0.833	214

図 8b　基準値ソート（500 語）

これを見ると、100 語では標準的な単語が上位を占めているが、500 語では Sunshine-1 で使われている現代的な単語が上位を占めている。一般的な単語は殆どの教科書で使われるので、100 語の場合は「標準性」即ち 0 比率が変動の主流になり、500 語の場合のように特別な単語が特定の教科書で使われている場合は、それらの単語と教科書が変動の主流になる。これが第 1 成分と第 2 成分の交代が起きる理由である。このことから、成分の意味にとって単語の選択数は重要な意味を持っていることが分かる。

３）の場合

ここでは教科書の単語数にほとんど違いがない場合を考える。語数調整した場合の文書 0 比率と全単語数との相関関係を表 7 に与える。

表 7　1 重調整法による文書 0 比率と全単語数との相関係数

語数	50	100	300	500	1000	全単語数
Choice-1	0.020	0.030	0.187	0.332	0.547	466
Drill-1	0.020	0.050	0.200	0.350	0.549	505
J&B-1	0.000	0.080	0.257	0.362	0.506	613
National-1	0.020	0.040	0.200	0.350	0.580	426
Taisho-1	0.000	0.030	0.190	0.316	0.495	633
Tsuda-p1	0.020	0.090	0.260	0.406	0.601	469
相関係数	-0.953	0.049	0.136	-0.352	-0.897	

これによると、利用する単語数が多くなるとやはり相関は高くなるが、そうでない場合、文書 0 比率は単語数にほとんどよらないようである。

次に、基準値平均について１）の場合に述べたことが成立するか調べてみる。結果は表 8 の通りである。

表 8　基準値平均とその標準偏差

	50	100	300	500	1000
Choice-1	0.0551	0.0398	0.0203	0.0146	0.0087
Drill-1	0.0527	0.0357	0.0187	0.0135	0.0087
J&B-1	0.0540	0.0337	0.0173	0.0134	0.0096
National-1	0.0550	0.0397	0.0212	0.0152	0.0089
Taisho-1	0.0514	0.0334	0.0187	0.0141	0.0095
Tsuda-p1	0.0510	0.0353	0.0199	0.0148	0.0093
標準偏差	0.0018	0.0028	0.0014	0.0007	0.0004

教科書による標準偏差は基準値平均の 10%以内に収まっている。また、対角指標については表 9 の関係が得られる。

表 9　対角指標

	50	100	300	500	1000
Choice-1	0.1861	0.2075	0.1853	0.1620	0.1172
Drill-1	0.1997	0.2063	0.1910	0.1658	0.1297
J&B-1	0.2014	0.1854	0.1708	0.1643	0.1501
National-1	0.1851	0.2048	0.1927	0.1660	0.1165
Taisho-1	0.1903	0.1858	0.1823	0.1700	0.1417
Tsuda-p1	0.1826	0.2002	0.2029	0.1775	0.1326
標準偏差	0.0079	0.0102	0.0108	0.0055	0.0133

この指標についての全体の平均は 0.1751、標準偏差は 0.0258 である。

以上の結果から、基準値平均についてはほぼ近似が成り立っていると考えることができるが、対角指標については今の段階では何とも言えない。。一般に標準化を行わない場合、このようなことはなく、アニメーションで見た結果の安定性も十分ではない。これらの指標と安定性の問題について今後もう少し考察を進める必要があるだろう。

次に、主要統計量間の相関行列を求めてみよう。図 9a に 100 語の場合、図 9b に 500 語

の場合を与える。

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.754	0.342	0.694	0.276	0.058
基準値偏差	0.754	1.000	0.002	0.986	-0.024	0.111
aii	0.342	0.002	1.000	-0.149	0.152	0.809
第1成分	0.694	0.986	-0.149	1.000	0.015	-0.008
第2成分	0.276	-0.024	0.152	0.015	1.000	0.024
第3成分	0.058	0.111	0.809	-0.008	0.024	1.000

図 9a　主要統計量間の相関行列（100 語）

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.672	0.813	-0.085	0.214	-0.149
基準値偏差	0.672	1.000	0.784	0.671	0.062	-0.166
aii	0.813	0.784	1.000	0.293	0.555	-0.305
第1成分	-0.085	0.671	0.293	1.000	0.000	0.011
第2成分	0.214	0.062	0.555	0.000	1.000	-0.010
第3成分	-0.149	-0.166	-0.305	0.011	-0.010	1.000

図 9b　主要統計量間の相関行列（500 語）

100 語では文書 0 比率と第 1 成分とがある程度相関を持っているが、500 語ではもはやどの成分とも相関は低い。そこで、文書 0 比率を第 1 成分と第 2 成分で重回帰分析することを試みる。図 10a は 100 語、図 10b は 500 語の場合である。いずれも重回帰分析の結果と CR 分析による散布図を上下に示している。

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	0.0164	0.6898	0.0092	1.7850	3	0.1723
第2成分	0.0063	0.2653	0.0091	0.6864	3	0.5418
切片	0.0546	0.0000	0.0091	5.9768	3	0.0094
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.743	0.552	0.503	0.253	1.8483	0.2998

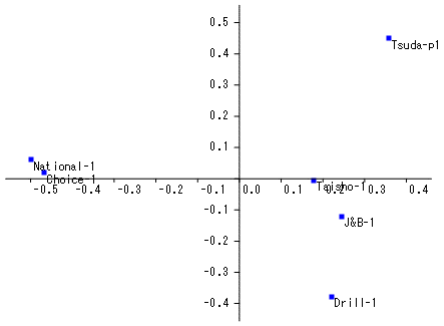


図 10a　重回帰分析と CR 分析の散布図（100 語）

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	-0.0024	-0.0848	0.0159	-0.1510	3	0.8896
第2成分	0.0060	0.2144	0.0158	0.3816	3	0.7282
切片	0.3528	0.0000	0.0158	22.3695	3	0.0002
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.231	0.053	0.000	0.000	0.0842	0.9213

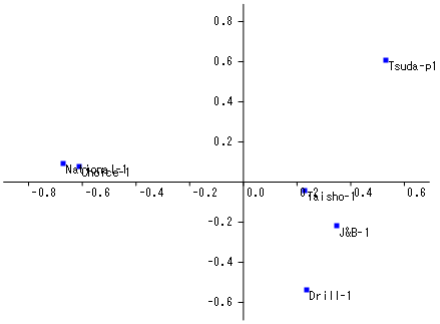


図 10b　重回帰分析と CR 分析の散布図（500 語）

100 語ではある程度の寄与率はあるが、500 語では重回帰式は全く意味がない。以上のように単語数に差がない場合は、文書 0 比率と単語数の相関もないし、成分との関係も得られない。

これまでの分析で、教科書分類の大きな 1 つの要素は単語数の多さや教科書の標準性に関係する文書 0 比率であった。しかし、この指標も殆ど同じレベルの教科書間では分類に影響を与えない。CR 分析で意味のあることは 0 比率がどの程度分析に影響を与えているのか、また影響を与えているならどの成分が 0 比率を表しているのかを知り、その他の成分の役割を検討することであると思われる。

補遺　テキスト CR 分析の理論

教科書ごと単語ごとの出現数のデータを $n_{i\lambda}$（$1 \le i \le p$，$1 \le \lambda \le m$，$p \ll m$）とする（調整済みを含む）。ここに p は教科書の数、m は利用する単語の数である。

各文書にパラメータ u_i、各単語にパラメータ v_λ を与え、これを用いて文書と単語の相関係数 ρ を以下のように定義する。

$$\rho = \frac{S_{uv}}{S_u S_v}$$

ここに、

$$S_{uv} = \frac{1}{n}\sum_{i=1}^{p}\sum_{\lambda=1}^{m} n_{i\lambda} u_i v_\lambda, \quad S_u^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i^2, \quad S_v^2 = \frac{1}{n}\sum_{\lambda=1}^{m} n_{\cdot\lambda} v_\lambda^2$$

$$n_{i\cdot} = \sum_{\lambda=1}^{m} n_{i\lambda}, \quad n_{\cdot\lambda} = \sum_{i=1}^{p} n_{i\lambda}, \quad n = \sum_{i=1}^{p}\sum_{\lambda=1}^{m} n_{i\lambda}$$

であり、パラメータについては以下を仮定する。

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i = 0, \quad \bar{v} = \frac{1}{n}\sum_{\lambda=1}^{m} n_{\cdot\lambda} v_\lambda = 0$$

この相関係数 ρ について、$S_u^2 = 1$，$S_v^2 = 1$ とする制約条件を付けて最大値を求める。そのために Lagrange の未定乗数法を用いる。

$$L = S_{uv} - \alpha\left(S_u^2 - 1\right) - \beta\left(S_v^2 - 1\right)$$

ここにαとβは未定乗数である。このLをu_iとv_λで微分して、以下の方程式を得る。

$$\sum_{\lambda=1}^{m} n_{i\lambda} v_\lambda - 2\alpha n_{i\cdot} u_i = 0, \qquad \sum_{i=1}^{p} n_{i\lambda} u_i - 2\beta n_{\cdot\lambda} v_\lambda = 0$$

左の式にu_iをかけてiについて和をとると$\rho = 2\alpha$、右の式にv_λをかけてλについて和をとると$\rho = 2\beta$を得る。すなわち、

$$\sum_{\lambda=1}^{m} n_{i\lambda} v_\lambda - \rho n_{i\cdot} u_i = 0, \qquad \sum_{i=1}^{p} n_{i\lambda} u_i - \rho n_{\cdot\lambda} v_\lambda = 0$$

次に、右式をv_λについて解いて、

$$v_\lambda = \frac{1}{\rho n_{\cdot\lambda}} \sum_{j=1}^{p} n_{j\lambda} u_j$$

これを左式に代入すると、

$$\sum_{j=1}^{p} \sum_{\lambda=1}^{m} \frac{n_{i\lambda} n_{j\lambda}}{n_{i\cdot} n_{\cdot\lambda}} u_j - \rho^2 u_i = 0$$

さらに、$u_i = \sqrt{n/n_{i\cdot}}\, z_i$とすると、$\sum_{i=1}^{p} z_i^2 = \frac{1}{n} \sum_{i=1}^{p} n_{i\cdot} u_i^2 = S_u^2 = 1$となり、以下を得る。

$$\sum_{j=1}^{p} a_{ij}\, z_j - \rho^2 z_i = 0 \qquad \text{(A1)}$$

ここにa_{ij}は以下となる。

$$a_{ij} = \sum_{\lambda=1}^{m} \left(\frac{n_{i\lambda}}{\sqrt{n_{i\cdot} n_{\cdot\lambda}}} \frac{n_{j\lambda}}{\sqrt{n_{j\cdot} n_{\cdot\lambda}}} \right) = \sum_{\lambda=1}^{m} x_{i\lambda} x_{j\lambda} \qquad \text{(A2)}$$

ここに、

$$x_{i\lambda} \equiv \frac{n_{i\lambda}}{\sqrt{n_{i\cdot} n_{\cdot\lambda}}} \qquad \text{(A3)}$$

今後$x_{i\lambda}$をデータ$n_{i\lambda}$に対する基準値、a_{ij}が与える行列$\mathbf{A}$を基準値行列と呼ぶ。一般に基準値行列a_{ij}には以下の関係がある。

$$a_{ij} = \sum_{\lambda=1}^{m} x_{i\lambda} x_{j\lambda} \le \frac{1}{2} \left(\sum_{\lambda=1}^{m} x_{i\lambda}^2 + \sum_{\lambda=1}^{m} x_{j\lambda}^2 \right) = \frac{1}{2} \left(a_{ii} + a_{jj} \right)$$

これらの関係を使うとv_λは、z_jを用いて以下のようにも書ける。

$$v_\lambda = \frac{1}{\rho n_{\cdot\lambda}} \sum_{j=1}^{p} n_{j\lambda} u_j = \frac{1}{\rho} \sum_{j=1}^{p} \sqrt{\frac{n_{j\cdot}}{n_{\cdot\lambda}}} x_{j\lambda} u_j = \frac{1}{\rho n_{\cdot\lambda}} \sum_{j=1}^{p} \sqrt{\frac{n}{n_{j\cdot}}} n_{j\lambda} z_j = \frac{1}{\rho} \sqrt{\frac{n}{n_{\cdot\lambda}}} \sum_{j=1}^{p} x_{j\lambda} z_j$$

(A1) 式は行列$\mathbf{A}$の固有方程式である。但し、a_{ij}にはその形に起因した以下の制約がある。

$$\sum_{j=1}^{p} a_{ij}\sqrt{n_{j\cdot}} = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\cdot}n_{\cdot\lambda}}}\sum_{j=1}^{p}\frac{n_{j\lambda}\sqrt{n_{j\cdot}}}{\sqrt{n_{j\cdot}n_{\cdot\lambda}}}\right) = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\cdot}n_{\cdot\lambda}}}\sum_{j=1}^{p}\frac{n_{j\lambda}}{\sqrt{n_{\cdot\lambda}}}\right)$$

$$= \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}\sqrt{n_{\cdot\lambda}}}{\sqrt{n_{i\cdot}n_{\cdot\lambda}}}\right) = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\cdot}}}\right) = \sqrt{n_{i\cdot}}$$

よって、**A** には固有値 1 の自明な固有ベクトル

$${}^{t}\mathbf{z} = \left(\sqrt{n_{1\cdot}/n} \quad \sqrt{n_{2\cdot}/n} \quad \cdots \quad \sqrt{n_{p\cdot}/n}\right)$$

が存在する。これは $\mathbf{u}$ にすると ${}^{t}\mathbf{u} = (1 \quad 1 \quad \cdots \quad 1)$ になり、$\bar{u} = (1/n)\sum_{i=1}^{p} n_{i\cdot}u_i = 1 \neq 0$ であり、平均が 0 の条件を満たさない。また、v_λ についても以下となり、全く特徴を表さない。

$$v_\lambda = \frac{1}{\rho n_{\cdot\lambda}}\sum_{j=1}^{p} n_{j\lambda}u_j = \frac{1}{n_{\cdot\lambda}}\sum_{j=1}^{p} n_{j\lambda} = 1$$

そのため、CR 分析ではこの解は省いて表示する。

参考文献

[1] コレスポンデンス分析を用いた英文テキスト分類における語数調整法と単語の選択基準，福井正康、渡辺清美，福山平成大学経営研究，第 15 号，(2019) 63-78

[2] テキストコレスポンデンス分析専用プログラムの開発， 日本言語教育 ICT 学会研究紀要，第 7 号， (2020) 49-58

２８．操作変数回帰分析

28.1 操作変数回帰分析とは

ある目的変数が２つの説明変数で表される重回帰分析を考える。２つの説明変数のうち１つは目的変数に直接影響を及ぼし、他からの影響は考えられないものとする。しかし、もう１つの説明変数は重回帰式に取り入れられないある要因から影響を受け、同時に目的変数もその要因から影響を受けるものとする。この要因により、目的変数とこの説明変数の間には直接的な影響の他に擬似相関が発生する。我々は前者の直接的な影響を与える説明変数を外生変数と呼び、後者のある要因から影響を受ける説明変数を内生変数と呼ぶ。また、この内生変数に影響を及ぼす要因を欠落バイアス要因（変数）と呼ぶ。図１にこれらの関係を示す。

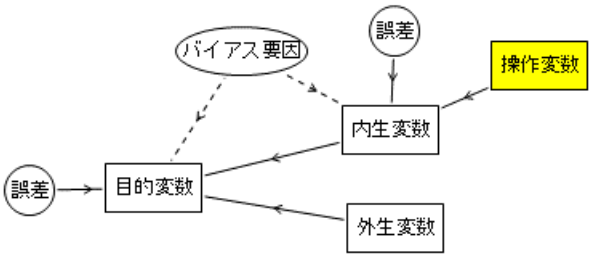


図１　操作変数回帰分析の関係

このようなバイアスが存在する場合、２つの説明変数による重回帰分析は正しい直接影響を与えない。本来重回帰分析では誤差と説明変数の間に相関がないように考えて、直接的な影響を計算するからである。また、この潜在変数を考えるようなモデルは共分散構造分析の対象のように思われるかも知れないが、これは解の識別が不可能なモデルで、残念ながら安定した解を求めることはできない。この状況を解決する方法が操作変数回帰分析である。

操作変数回帰分析では内生変数だけに影響を与え、目的変数に直接影響を与えないような観測可能な変数を考える。この変数を操作変数と呼ぶ。分析方法は、まず内生変数を操作変数と他の外生変数で回帰し、その予測値を内生変数の実測値の代わりにして重回帰分析を行うというものである。もちろん内生変数や操作変数は複数あっても構わない。ただ、操作変数の数は内生変数の数以上である必要がある。

ここで、外生変数は考えず、内生変数が１つだけの場合について、少し数式を使って操作変数回帰分析の原理を考えてみよう。今バイアスによって、目的変数と説明変数の間に相互の関係があるものとする。

ある変数 y_λ が他の変数 x_λ によって、誤差項 u_λ を含めて、(1)式のように線形に予測されるとき、パラメータの推定に回帰分析は有効である。ここで $\lambda\,(=1,\cdots,N)$ は個体を表す記号とする。

$$y_\lambda = a_1 x_\lambda + a_0 + u_\lambda \qquad (1)$$

しかし、これと同じような関係がもう１つあって、y_λ と x_λ がその連立方程式の解として与えられる場合や、x_λ も y_λ によって逆に予測される場合などは、２つの誤差項の影響により、回帰分析を用いてそのまま係数を求めることは困難になる。この状況を数式で表現すると以下のような連立方程式になる。

$$\begin{aligned} y_\lambda &= a_1 x_\lambda + a_0 + u_\lambda \\ y_\lambda &= b_1 x_\lambda + b_0 + v_\lambda \end{aligned} \quad \text{，または、} \quad \begin{aligned} y_\lambda &= a_1 x_\lambda + a_0 + u_\lambda \\ x_\lambda &= b_1 y_\lambda + b_0 + v_\lambda \end{aligned} \qquad (2)$$

但し、これら２つの連立方程式は、下の２番目の式の x_λ と y_λ を交換すると形式的に同じものである。この状況の例としては、図２のような価格と需要及び、価格と供給の均衡の問題がよく紹介される。

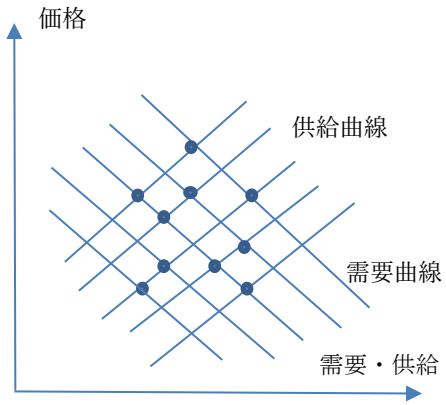


図２　需要と供給の均衡

さて、例えば(2)式左側の場合、解は以下となる。

$$x_\lambda = [(b_0 + v_\lambda) - (a_0 + u_\lambda)]/(a_1 - b_1)$$
$$y_\lambda = [a_1(b_0 + v_\lambda) - b_1(a_0 + u_\lambda)]/(a_1 - b_1)$$

この解では u_λ と v_λ の変動により、図１のように、x_λ と y_λ に(1)式のような１つの関係を与えることはできず、パラメータを回帰分析などで決定することは困難になる。しかし、例えば x_λ に近い別の x'_λ が存在し、それが u_λ に関係しなれば、

$$x'_\lambda \simeq [(b_0 + v_\lambda)]/(a_1 - b_1)$$
$$y_\lambda = [a_1(b_0 + v_\lambda) - b_1(a_0 + u_\lambda)]/(a_1 - b_1)$$

より、

$$y_\lambda \simeq a_1 x'_\lambda + a_0 - b_1 u_\lambda/(a_1 - b_1)$$

となり、図３のように(2)式上側のパラメータの値を回帰分析を用いて求めることが可能となるはずである。この x'_λ を作り出す分析が操作変数回帰分析であり、それに用いる以下で述べる変数 z_λ を操作変数と呼ぶ。

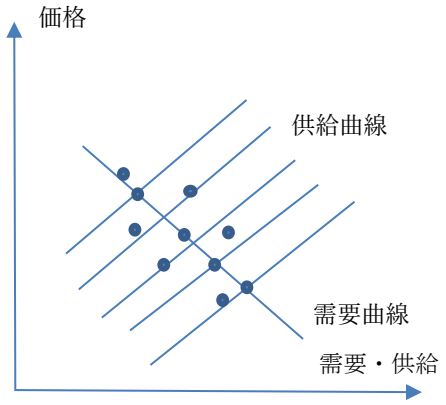


図３　操作変数を用いた場合の均衡

今、x_λと関係が比較的強く、誤差u_λと無相関な、$x_\lambda = c_1 z_\lambda + c_0 + e_\lambda = \hat{x}_\lambda + e_\lambda$となる操作変数$z_\lambda$を（もしあれば）考える。この操作変数$z_\lambda$による予測値$\hat{x}_\lambda$を上式に代入すると、

$$y_\lambda = a_1(\hat{x}_\lambda + e_\lambda) + a_0 + u_\lambda$$
$$y_\lambda = b_1(\hat{x}_\lambda + e_\lambda) + b_0 + v_\lambda$$

これを解くと

$$\hat{x}_\lambda = [(b_0 + v_\lambda + b_1 e_\lambda) - (a_0 + u_\lambda + a_1 e_\lambda)]/(a_1 - b_1)$$
$$y_\lambda = [a_1(b_0 + v_\lambda + b_1 e_\lambda) - b_1(a_0 + u_\lambda + a_1 e_\lambda)]/(a_1 - b_1)$$

ここで、$\hat{x}_\lambda$はz_λから作られ、u_λやe_λと無相関であるから、$u_\lambda + (a_1 - b_1)e_\lambda \simeq 0$となっているはずである。

$$\hat{x}_\lambda \simeq [(b_0 - a_0) + v_\lambda]/(a_1 - b_1)$$
$$y_\lambda \simeq [(a_1 b_0 - a_0 b_1) + (a_1 v_\lambda - b_1 u_\lambda)]/(a_1 - b_1)$$

これより、前に述べた通り、以下となる。

$$y_\lambda \simeq a_1 \hat{x}_\lambda + a_0 - b_1 u_\lambda / (a_1 - b_1)$$

すなわちx_λの代わりに操作変数を使った予測値$\hat{x}_\lambda$を用いると、回帰分析によってパラメータの値を求めることができる。

これまで操作変数回帰分析の基本的な考え方を学んだので、少し一般的な手順を示しておこう。操作変数回帰分析では目的変数を以下のような線形の式で予測する。ここに、変数$x_{i\lambda}\ (i = 1, \cdots, k)$は誤差項$u_\lambda$と相関する内生変数と呼ばれる変数で、変数$w_{i\lambda}\ (i = 1, \cdots, p)$は誤差項と相関しない外生変数と呼ばれる変数である。

$$y_\lambda = \sum_{i=1}^{k} b_i x_{i\lambda} + \sum_{i=1}^{p} b_{k+i} w_{i\lambda} + b_0 + u_\lambda \qquad (3)$$

一般の回帰分析では、変数はすべて外生変数である。

操作変数回帰分析では一般に２段階法という手法が使われる。すなわちまず、各内生変数$x_{i\lambda}$を操作変数と呼ばれる誤差項u_λと無相関な変数$z_{i\lambda}\ (i = 1, \cdots, m \geq k)$と外生変数$w_{i\lambda}$で予測する。

$$x_{i\lambda} = \sum_{j=1}^{m} c_j^{(i)} z_{j\lambda} + \sum_{j=1}^{p} c_{r+j}^{(i)} w_{j\lambda} + c_0^{(i)} + v_{i\lambda} = \hat{x}_{i\lambda} + v_{i\lambda}$$

ここで、$\hat{x}_{i\lambda}$ は回帰の予測値である。次にこの予測値で(3)式の内生変数を置き換え、回帰分析を実行する。

$$y_\lambda = \sum_{i=1}^{k} b_i' \hat{x}_{i\lambda} + \sum_{i=1}^{p} b_{k+i}' w_{i\lambda} + b_0' + u_\lambda' \qquad (4)$$

この推定されたパラメータ b_i', b_{k+i}', b_0' は $N \to \infty$ で、真の値 b_i, b_{k+i}, b_0 に一致することが知られている。ここで、パラメータの区間推定などに利用される誤差項 u_λ'' については、以下の式から求められる。

$$y_\lambda = \sum_{i=1}^{k} b_i' x_{i\lambda} + \sum_{i=1}^{p} b_{k+i}' w_{i\lambda} + b_0' + u_\lambda'' \qquad (5)$$

ここでは得られたパラメータの値はそのままで、内生変数の予測値を元の値に変えている。これは、各パラメータが $N \to \infty$ で真の値に近づくことから、誤差 u_λ'' も真の誤差 u_λ に近づくことによる。これ以上の詳細は後の理論的解説のところで紹介する。

28.2 プログラムの利用法

メニュー［分析－多変量解析他－経済・経営手法－操作変数回帰分析］を選択すると、図３に示す操作変数回帰分析の実行メニューが表示される。この分析用のデータは図 4 の形式で与えられる。

図３　操作変数回帰分析実行メニュー

図４　操作変数回帰分析のデータ形式

操作変数回帰分析の変数選択では、最初に目的変数を選択するが、その後の順番は特に決まっていない。しかし、内生変数の変数名の先頭には「@」記号、操作変数の変数名の先頭には「$」記号を付けて区別する。先頭にこれらの記号が付いていない変数は外生変数と解釈される。ここではすべての変数を選択すると、内生変数２つ、外生変数１つ、操作変数２つのモデルである。

まず、「第１段階回帰」ボタンをクリックして、操作変数の妥当性を調べる。操作変数の妥当性は第１段階回帰での F 値が 2 以上のとき妥当であると判断される。結果を図 5 に示す。

第１段階回帰分析結果

@内生変数1	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
$操作変数1	1.5112	0.4555	0.2140	7.0627	0.0000	1.0919	1.9306	0.9541
$操作変数2	0.8758	0.5747	0.1366	6.4120	0.0000	0.6081	1.1435	0.9663
外生変数	0.0325	0.0181	0.1065	0.3054	0.7601	-0.1762	0.2412	-0.6825
切片	-11.2521	0.0000	16.2512	-0.6924	0.4887	-43.1039	20.5998	
R	0.989	R^2	0.978	調整済R	0.987	調整済R^2	0.973	
F値	439.9434	自由度	3,∞	検定確率	0.0000			
@内生変数2	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
$操作変数1	0.2155	0.1089	0.5647	0.3817	0.7027	-0.8913	1.3224	0.7436
$操作変数2	0.9600	1.0559	0.2416	3.9733	0.0001	0.4865	1.4336	0.8199
外生変数	0.5024	0.4699	0.2315	2.1704	0.0300	0.0487	0.9560	-0.3481
切片	28.2781	0.0000	35.6531	0.7931	0.4277	-41.6006	98.1569	
R	0.885	R^2	0.783	調整済R	0.862	調整済R^2	0.742	
F値	28.1095	自由度	3,∞	検定確率	0.0000			

図５　第１段階回帰分析結果

２つの内生変数に対して、それぞれ２つの操作変数と１つの外生変数による回帰式が示されている。ここでの標準誤差やそれに基づく指標の計算には、不均一分散を考慮した方法を用いている。この場合、操作変数の妥当性は満たされている。

「第１段階予測値」ボタンをクリックすると内生変数ごとの第１段階回帰分析の予測値が図６のように表示される。この値が第２段階の操作変数回帰分析に使われる。

第１段階回帰予測値

	@内生変数1	@内生変数2
1	162.780	192.228
2	140.836	180.553
3	159.121	190.002
4	144.669	180.939
5	127.169	175.070
6	157.310	181.139
7	131.080	174.649
8	133.908	172.066
9	139.850	179.898

図６　第１段階回帰分析の予測値

操作変数回帰分析では、内生変数の実測値の代わりに第１段階回帰分析の予測値が用いられる。結果を図７に示す。

操作変数回帰分析結果

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
@内生変数1	0.2538	0.2599	1.0072	0.2520	0.8011	-1.7202	2.2278	0.700
@内生変数2	1.4342	0.8763	1.3737	1.0440	0.2965	-1.2583	4.1267	0.735
外生変数	0.6470	0.3698	0.8311	0.7785	0.4363	-0.9819	2.2760	-0.113
切片	-52.5328	0.0000	87.7868	-0.5984	0.5496	-224.5917	119.5262	
R	0.763	R^2	0.582	調整済R	0.709	調整済R^2	0.503	
F値	14.476	自由度	3,∞	検定確率	0.000			

図７　操作変数回帰分析の結果

ここでの標準誤差及びそれに基づく指標の導出には、操作変数回帰分析のパラメータと内生変数の実測値を用いた予測値からの残差を利用し、不均一分散を考慮した方法を用いている。

「予測値と残差」ボタンをクリックすると、操作変数回帰分析から推定されたパラメータの値と内生変数の実測値を用いて計算された予測値と残差の値が図８のように表示される。


予測値と残差

	実測値	予測値	残差
1	333	331.516	1.484
2	320	317.528	2.472
3	340	318.834	21.166
4	323	313.366	9.634
5	300	315.103	-15.103
6	311	293.012	17.988
7	308	297.666	10.334
8	296	289.989	6.011
9	316	315.218	0.782

図 8　操作変数回帰分析の予測値と残差

この予測値と実測値の関係は「実測/予測散布図」ボタンをクリックすることで図 9 のような散布図として表示される。ここで、グラフは重回帰分析などと同じく、縦軸が実測値、横軸が予測値である。

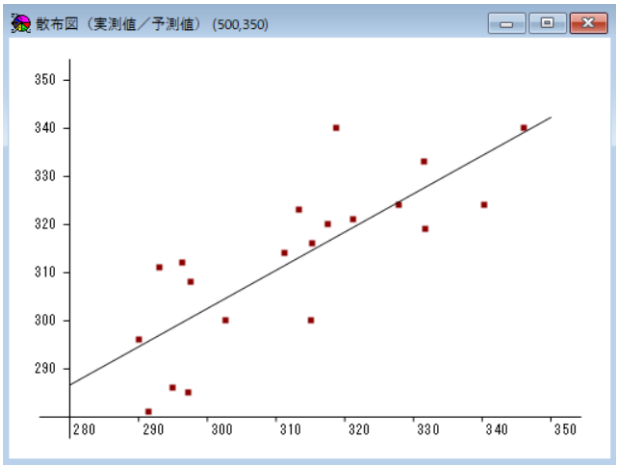


図 9　実測/予測散布図

操作変数の外生性の問題については、「過剰識別の検定」ボタンで調べることができる。但し、内生変数数より操作変数が多い場合のみ検定可能であるため、この例のように、内生変数２つ、操作変数２つの場合は検定を行うことができない。このソフトではこの条件を満たさない場合は、図 10 のような結果が表示される。

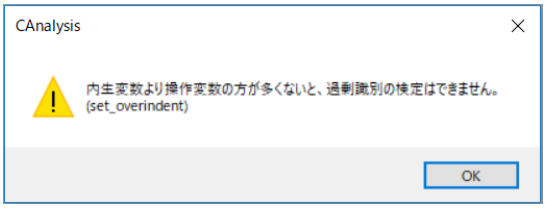


図 10　過剰識別の検定の注意メッセージ

今の場合、例えば、内生変数を１つ減らして内生変数１だけにすれば調べることが可能となる。結果は図 11 のように表示される。特に最下行の「過識別 F 値」のところが過剰識別の検定の部分である。「過識別確率」の値が有意水準の確率より大きければ（例えば 0.05 より大きければ）操作変数の外生性に問題はないと考える。


過剰識別の検定結果

IV回帰残差	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限	相関係数
$操作変数1	-1.1211	-0.6740	0.8429	-1.3301	0.1835	-2.7731	0.5309	-0.084
$操作変数2	0.5480	0.7174	0.5087	1.0773	0.2813	-0.4490	1.5451	0.067
外生変数	0.0660	0.0735	0.2958	0.2232	0.8234	-0.5137	0.6457	0.000
切片	-5.7737	0.0000	53.2953	-0.1083	0.9137	-110.2307	98.6832	
R	0.324	R^2	0.105	調整済R		調整済R^2		
F値	0.596	自由度	3,∞	検定確率	0.618			
過識別F値	0.888	過識別自由...	2,∞	過識別確率	0.412	≧αで良		

図 11　過剰識別の検定結果

28.3 操作変数回帰分析の制約条件

操作変数回帰分析では目的変数を以下のような線形の式で予測する。ここに、変数 $x_{i\lambda}\ (i=1,\cdots,k)$ は誤差項 u_λ と相関する内生変数と呼ばれる変数で、変数 $w_{i\lambda}\ (i=1,\cdots,p)$ は誤差項と相関しない外生変数と呼ばれる変数である。これらの変数に基づく回帰式を以下とする。

$$y_\lambda=\sum_{i=1}^{k}b_i x_{i\lambda}+\sum_{i=1}^{p}b_{k+i}w_{i\lambda}+b_0+u_\lambda \qquad (1)$$

最小 2 乗法に基づく誤差に対しては以下が成り立つ。

$$\sum_{\lambda=1}^{N}x_{i\lambda}u_\lambda=0,\quad \sum_{\lambda=1}^{N}w_{i\lambda}u_\lambda=0,\quad \sum_{\lambda=1}^{N}u_\lambda=0$$

操作変数回帰分析の 2 段階法では、最初に、各内生変数 $x_{i\lambda}$ を操作変数と呼ばれる誤差項 u_λ と無相関な変数 $z_{i\lambda}\ (i=1,\cdots,m\geq k)$ と外生変数 $w_{i\lambda}$ で予測する。

$$x_{i\lambda}=\sum_{j=1}^{m}c_j^{(i)}z_{j\lambda}+\sum_{j=1}^{p}c_{r+j}^{(i)}w_{j\lambda}+c_0^{(i)}+v_{i\lambda}=\hat{x}_{i\lambda}+v_{i\lambda}\quad(1\leq i\leq k)$$

ここで、$\hat{x}_{i\lambda}$ は回帰の予測値である。上と同様に最小 2 乗法に基づく誤差に対して、

$$\sum_{\lambda=1}^{N}z_{j\lambda}v_{i\lambda}=0,\quad \sum_{\lambda=1}^{N}w_{j\lambda}v_{i\lambda}=0,\quad \sum_{\lambda=1}^{N}v_{i\lambda}=0\quad \text{これより、}\sum_{\lambda=1}^{N}\hat{x}_{j\lambda}v_{i\lambda}=0$$

次にこの予測値で(1)式の内生変数を置き換え、回帰分析を実行する。

$$y_\lambda=\sum_{i=1}^{k}b_i'\hat{x}_{i\lambda}+\sum_{i=1}^{p}b_{k+i}'w_{i\lambda}+b_0'+u_\lambda' \qquad (3)$$

制約条件：$\sum_{\lambda=1}^{N}\hat{x}_{i\lambda}u_\lambda'=0,\quad \sum_{\lambda=1}^{N}w_{i\lambda}u_\lambda'=0,\quad \sum_{\lambda=1}^{N}u_\lambda'=0$

ここで、

$$x_{0\lambda}=1 \qquad \tilde{x}_{0\lambda}=1$$

$$x_{i\lambda}=x_{i\lambda}\quad 1\leq i\leq k\ , \qquad \tilde{x}_{i\lambda}=\hat{x}_{i\lambda}\quad 1\leq i\leq k$$

$$x_{k+i\lambda}=w_{i\lambda}\quad 1\leq i\leq p \qquad \tilde{x}_{k+i\lambda}=w_{i\lambda}\quad 1\leq i\leq p$$

$$(\mathbf{X})_{\lambda i}=x_{i\lambda},\quad \mathbf{M}=\mathbf{X}'\mathbf{X} \qquad (\tilde{\mathbf{X}})_{\lambda i}=\tilde{x}_{i\lambda},\quad \tilde{\mathbf{M}}=\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$$

のような表記を用いると、(1),(3)式は以下のように書かれる。

$$y_\lambda = \sum_{i=0}^{k+p} b_i x_{i\lambda} + u_\lambda,\ \text{制約条件：}\ \sum_{\lambda=1}^{N} x_{i\lambda} u_\lambda = 0 \qquad (4)$$

$$y_\lambda = \sum_{i=0}^{k+p} b'_i \tilde{x}_{i\lambda} + u'_\lambda,\ \text{制約条件：}\ \sum_{\lambda=1}^{N} \tilde{x}_{i\lambda} u'_\lambda = 0 \qquad (5)$$

重回帰分析の最小２乗法の理論より、パラメータは以下のように与えられることが知られているが、

$$b'_i = \sum_{j=0}^{k+p} \tilde{M}_{ij}^{-1} \sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} y_\lambda$$

これに(3)式を代入すると、以下となる。

$$\begin{aligned} b'_i &= \sum_{j=0}^{k+p} \tilde{M}_{ij}^{-1} \sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} y_\lambda = \sum_{j=0}^{k+p} \tilde{M}_{ij}^{-1} \sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} \left(\sum_{l=0}^{k+p} b_l x_{l\lambda} + u_\lambda \right) \\ &= \sum_{j=0}^{k+p} \tilde{M}_{ij}^{-1} \sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} \left(\sum_{l=0}^{k+p} b_l (\tilde{x}_{l\lambda} + v_{l\lambda}) + u_\lambda \right) \\ &= \sum_{j=0}^{k+p} \sum_{l=0}^{k+p} \tilde{M}_{ij}^{-1} \tilde{M}_{jl} b_l + \sum_{j=0}^{k+p} \tilde{M}_{ij}^{-1} \sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} u_\lambda = b_i + \sum_{j=0}^{k+p} \tilde{M}_{ij}^{-1} \sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} u_\lambda \end{aligned} \qquad (6)$$

ここで、計算の途中で以下を用いた。

$$\sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} v_{l\lambda} = 0$$

この表式で、$\tilde{x}_{j\lambda}$ と u_λ は $N \to \infty$ で無相関であることから、$b'_i \xrightarrow[N\to\infty]{} b_i$ となり、一致性が示される。これより、b'_i の標準誤差は以下となる。

$$S_{b'_i}^2 = \sum_{j=0}^{k+p} \sum_{k=0}^{k+p} \tilde{M}_{ij}^{-1} \left(\sum_{\lambda=1}^{N} \tilde{x}_{j\lambda} \tilde{x}_{k\lambda} u_\lambda^2 \right) \tilde{M}_{ik}^{-1}$$

次に、(6)式の u_λ はどのような値で推定されるだろうか。パラメータ b'_i は $N \to \infty$ で、真の値 b_i に一致することが分かったので、誤差項 u_λ については、以下の回帰式の u''_λ から求められる。

$$y_\lambda = \sum_{i=1}^{k} b'_i x_{i\lambda} + \sum_{i=1}^{p} b'_{k+i} w_{i\lambda} + b'_0 + u''_\lambda \qquad (7)$$

ここでは得られたパラメータの値はそのままで、内生変数の予測値を元の値に変えている。これは左辺が与えられて、各パラメータが $N \to \infty$ で真の値に近づくことから、誤差 u''_λ も真の誤差 u_λ に近づくとして理解できる。(7)式と(3)式を比べると u''_λ と u'_λ の間には以下の関係がある。

$$u''_\lambda = \sum_{i=1}^{k} b'_i v_{i\lambda} + u'_\lambda \qquad (8)$$

この関係を使うと、以下のような関係も示すことができる。

$$\sum_{\lambda=1}^{N} \hat{x}_{i\lambda} u''_\lambda = 0,\ \ \sum_{\lambda=1}^{N} w_{i\lambda} u''_\lambda = 0,\ \ \sum_{\lambda=1}^{N} u''_\lambda = 0 \qquad (9)$$

次に、過剰識別の検定について少し細かく説明する。過剰識別の検定では、まず(7)式の誤差項u_λ''を以下のように操作変数と外生変数で回帰する。

$$u_\lambda'' = \sum_{i=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 + e_\lambda \tag{10}$$

ここで、誤差e_λについての右辺の制約は以下である。

$$\sum_{\lambda=1}^{N} z_{j\lambda} e_\lambda = 0,\ \sum_{\lambda=1}^{N} w_{j\lambda} e_\lambda = 0,\ \sum_{\lambda=1}^{N} e_\lambda = 0 \qquad \text{制約} m+p+1 \text{個} \tag{11}$$

ここで気にかかる問題は、上の制約以外に左辺のu_λ''の制約からe_λに新たに制約が付かないかという点である。これについて考えてみよう。u_λ''についての左辺の制約(9)を考える。最初の制約は以下である。

$$\begin{aligned}\sum_{\lambda=1}^{N} \hat{x}_{i\lambda} u_\lambda'' &= \sum_{\lambda=1}^{N} \hat{x}_{i\lambda} \left[\sum_{j=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 + e_\lambda \right] \\ &= \sum_{\lambda=1}^{N} \hat{x}_{i\lambda} \left[\sum_{j=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 \right] = 0\end{aligned} \tag{12}$$

(12)式は左辺の誤差に関する制約が右辺のパラメータに対する条件になっている。また、

$$\begin{aligned}\sum_{\lambda=1}^{N} w_{j\lambda} u_\lambda'' &= \sum_{\lambda=1}^{N} w_{i\lambda} \left[\sum_{j=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 + e_\lambda \right] \\ &= \sum_{\lambda=1}^{N} w_{i\lambda} \left[\sum_{j=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 \right] = 0\end{aligned} \tag{13}$$

(13)式も右辺のパラメータに対する条件である。さらに、

$$\begin{aligned}\sum_{\lambda=1}^{N} u_\lambda'' &= \sum_{\lambda=1}^{N} \left[\sum_{j=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 + e_\lambda \right] \\ &= \sum_{\lambda=1}^{N} \left[\sum_{j=1}^{m} d_j z_{j\lambda} + \sum_{j=1}^{p} d_{m+j} w_{j\lambda} + d_0 \right] \\ &= N \left[\sum_{j=1}^{m} d_j \bar{z}_j + \sum_{j=1}^{p} d_{m+j} \bar{w}_j + d_0 \right] = 0\end{aligned} \tag{14}$$

これは右辺の性質より自動的に成り立つ。これらのことから、u_λ''の制約によってはe_λに制約が追加されることはなく、e_λについての制約は$m-k-1$個のままである。よってe_λの自由度は$N-m-p-1$である。

次に、u_λ''を外生変数だけで回帰すると、

$$u_\lambda'' = \sum_{j=1}^{p} d'_{m+j} w_{j\lambda} + d'_0 + e'_\lambda \tag{15}$$

e'_λについての右辺の制約は以下である。

$$\sum_{\lambda=1}^{N} w_{j\lambda} u''_{\lambda} = 0, \quad \sum_{\lambda=1}^{N} u''_{\lambda} = 0 \qquad \text{制約数：} \ p+1 \tag{16}$$

前と同様にu''_{λ}についての左辺の制約を考えてみよう。

$$\sum_{\lambda=1}^{N} \hat{x}_{i\lambda} u''_{\lambda} = \sum_{\lambda=1}^{N} \hat{x}_{i\lambda} \left[\sum_{j=1}^{p} d'_{m+j} w_{j\lambda} + d'_0 + e'_{\lambda} \right] = 0 \tag{17}$$

(17)式はu''_{λ}についての制約となる。これによって、e'_{λ}にk個の制約が追加される。

$$\begin{aligned} \sum_{\lambda=1}^{N} w_{j\lambda} u''_{\lambda} &= \sum_{\lambda=1}^{N} w_{i\lambda} \left[\sum_{j=1}^{p} d'_{m+j} w_{j\lambda} + d'_0 + e'_{\lambda} \right] \\ &= \sum_{\lambda=1}^{N} w_{i\lambda} \left[\sum_{j=1}^{p} d'_{m+j} w_{j\lambda} + d'_0 \right] = 0 \end{aligned} \tag{18}$$

(18)式はパラメータに対する条件である。最後に、

$$\sum_{\lambda=1}^{N} u''_{\lambda} = \sum_{\lambda=1}^{N} \left[\sum_{j=1}^{p} d'_{m+j} w_{j\lambda} + d'_0 + e'_{\lambda} \right] = N \left[\sum_{j=1}^{p} d'_{m+j} \bar{w}_j + d'_0 \right] = 0 \tag{19}$$

これは回帰式から自動的に求められる関係である。以上のことからe'_{λ}についての制約は元々の$p+1$個に加え、u''_{λ}からの制約がk個追加され、$p+k+1$個となる。よってe'_{λ}の自由度は$N-k-p-1$である。

これより結合仮説の検定を利用すると以下のような統計量が導出される。

$$\sum_{\lambda=1}^{N} e'_{\lambda} - \sum_{\lambda=1}^{N} e_{\lambda} \to \chi^2_{(N-k-p-1)-(N-m-p-1)} = \chi^2_{m-k}$$

この関係を使って、すべての$d_{ij} = 0$（m個）のF統計量を求める。一般的には変数数から、$(m+p)-p=m$を自由度とするため

$$F = \frac{\left(\sum_{\lambda=1}^{N} e'^2_{\lambda} - \sum_{\lambda=1}^{N} e^2_{\lambda} \right) \Big/ m}{\sum_{\lambda=1}^{N} e^2_{\lambda} \Big/ (N-m-p-1)}$$

であるが、上に述べた計算より以下となる。

$$mF = \frac{\sum_{\lambda=1}^{N} e'^2_{\lambda} - \sum_{\lambda=1}^{N} e^2_{\lambda}}{\sum_{\lambda=1}^{N} e^2_{\lambda} \Big/ (N-m-p-1)} \xrightarrow[N\to\infty]{} \chi^2_{m-k} \tag{20}$$

28.4　操作変数回帰分析の理論

行列と分布の公式

この補遺で使う統計の公式をまとめておく。$\mathbf{u}$が確率変数である。

【公式１】　$Cov(\mathbf{Au}, \mathbf{Bu}) = \mathbf{A\Sigma_u B'}$

【公式２】　$\mathbf{A\Sigma_u B'} = \mathbf{0}$ならば、$\mathbf{Au}$と$\mathbf{Bu}$は独立した分布である。

【公式３】 $\mathbf{u}(m\times 1)\sim N(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{u}})$ のとき、

$\mathbf{d}+\mathbf{A}\mathbf{u}\sim N(\mathbf{d},\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{A}')$

$\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{u}\sim \chi_m^2$

【公式４】 $\mathbf{u}\sim N(0,\mathbf{I}_m)$ で $\mathbf{C}(m\times m)$ がべき等行列（$\mathbf{CC}=\mathbf{C}$）のとき、

$\mathbf{u}'\mathbf{C}\mathbf{u}\sim \chi_r^2$ 但し、$rank(\mathbf{C})=r$

操作変数回帰分析の行列を用いた理論

除外されたバイアスの影響により、回帰式の説明変数と誤差項との間に相関があり、通常の回帰分析では回帰係数の推定値に問題がある場合、一部の説明変数に影響を与え、誤差項と無相関な変数を見つけられれば、その変数を利用して、正しい回帰係数を求めることができる可能性がある。これが操作変数回帰分析である。説明変数の中で誤差項と無相関な変数を外生変数、誤差項と相関がある変数を内生変数と呼ぶ。内生変数だけに影響を与え、誤差項と無相関な変数を操作変数と呼ぶ。ここでは目的変数を $y_\lambda\ (\lambda=1,\cdots,N)$、内生変数を $x_{i\lambda}\ (i=1,\cdots,k)$、外生変数を $w_{i\lambda}\ (i=1,\cdots,r)$、操作変数を $z_{i\lambda}\ (i=1,\cdots,m)$ とする。ここでは、これらの変数を以下のように行列表示する。ここで個体間の相関はないものと仮定する。

$$\mathbf{y}(N\times 1)=\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix},\quad \mathbf{X}(N\times(k+r+1))=\begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} & w_{11} & \cdots & w_{r1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{kN} & w_{1N} & \cdots & w_{rN} \end{pmatrix},$$

$$\mathbf{Z}(N\times(m+r+1))=\begin{pmatrix} 1 & z_{11} & \cdots & z_{m1} & w_{11} & \cdots & w_{r1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{1N} & \cdots & z_{mN} & w_{1N} & \cdots & w_{rN} \end{pmatrix}$$

また、正しい回帰係数と誤差を $\boldsymbol{\beta}((k+r+1)\times 1)$，$\boldsymbol{\varepsilon}(N\times 1)$ として、回帰式は以下のように書ける。

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \qquad (1)$$

この式を用いてそのまま最小２乗法を実行すると、$\boldsymbol{\varepsilon}$ の推定値 $\hat{\boldsymbol{\varepsilon}}$ と $\mathbf{X}$ は結果として独立となり、正しい回帰係数の推定値 $\hat{\boldsymbol{\beta}}$ は得られない。これ以降、ある量の真の値を a とするとき、その推定値を $\hat{a}$ と書くことにする。

この問題を解決するために、操作変数回帰分析の２段階法では以下のようなことを考える。まず、第１段階として、説明変数の中の内生変数について、すべての操作変数と外生変数で回帰する。これを行列表示すると以下となる。

$$\mathbf{X}=\mathbf{Z}\mathbf{C}+\mathbf{V} \qquad (2)$$

ここで $\mathbf{C}$ は係数行列、$\mathbf{V}$ は誤差行列である。但し、$\mathbf{X}$ に含まれる定数と外生変数については恒等式とするので、$\mathbf{C}$ と $\mathbf{V}$ については以下のようにする。

$$\mathbf{C}((m+r+1)\times(k+r+1))=\begin{pmatrix} 1 & \mathbf{c_0}(1\times k) & \mathbf{0}(1\times r) \\ \mathbf{0}(m\times 1) & \mathbf{c_z}(m\times k) & \mathbf{0}(m\times r) \\ \mathbf{0}(r\times 1) & \mathbf{c_w}(r\times k) & \mathbf{I}(r\times r) \end{pmatrix}$$

$$\mathbf{V}(N\times(k+r+1))=\begin{pmatrix}0 & v_{11} & \cdots & v_{k1} & 0 & \cdots & 0\\ 0 & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & v_{1N} & \cdots & v_{kN} & 0 & \cdots & 0\end{pmatrix}$$

これらの推定値を $\hat{\mathbf{C}}$ と $\hat{\mathbf{V}}$ とすると、推定結果は以下のように書ける。

$$\mathbf{X}=\mathbf{Z}\hat{\mathbf{C}}+\hat{\mathbf{V}}=\hat{\mathbf{X}}+\hat{\mathbf{V}}$$

次に第２段階として、予測値 $\hat{\mathbf{X}}$ を使って以下のような回帰式を考える。

$$\mathbf{y}=\hat{\mathbf{X}}\mathbf{b}+\mathbf{u} \qquad (3)$$

推定値の関係は以下となる。

$$\mathbf{y}=\hat{\mathbf{X}}\hat{\mathbf{b}}+\hat{\mathbf{u}}$$

この回帰係数の推定値 $\hat{\mathbf{b}}$ が $N\to\infty$ で真の回帰係数 $\boldsymbol{\beta}$ に収束することが知られている。これを証明する。

(2)式の回帰係数の推定値については、

$$\hat{\mathbf{C}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

と書けることから、以下を得る。

$$\hat{\mathbf{X}}=\mathbf{Z}\hat{\mathbf{C}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}=\mathbf{P}_{\mathbf{Z}}\mathbf{X},\quad \mathbf{P}_{\mathbf{Z}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}',\quad \hat{\mathbf{X}}'\hat{\mathbf{X}}=\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}$$

ここに、ある行列 $\mathbf{A}$ に対して、$\mathbf{P}_{\mathbf{A}}\equiv\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ と定義する。この $\mathbf{P}_{\mathbf{A}}$ はべき等行列（$\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{A}}=\mathbf{P}_{\mathbf{A}}$）である。この関係を使うと $\hat{\mathbf{b}}$ は以下のように書ける。

$$\hat{\mathbf{b}}=(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}=(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}$$
$$=(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})=\boldsymbol{\beta}+(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\boldsymbol{\varepsilon}$$

上の関係より、【公式 3】を使うと

$$\hat{\mathbf{b}}\sim N(\boldsymbol{\beta},(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\boldsymbol{\Sigma}_{\varepsilon}\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1})$$

ここで、外生変数は誤差項 $\boldsymbol{\varepsilon}$ と無相関であるという仮定から、

$$\mathbf{Z}'\boldsymbol{\varepsilon}\xrightarrow[N\to\infty]{}\mathbf{0}\text{，即ち、}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\Sigma}_{\varepsilon}\mathbf{P}_{\mathbf{Z}}\xrightarrow[N\to\infty]{}\mathbf{0}$$

よって 2 段階法で求めた回帰係数は $N\to\infty$ で真の回帰係数に一致する。

$$\hat{\mathbf{b}}\xrightarrow[N\to\infty]{}\boldsymbol{\beta}$$

この関係を使うと以下となり、

$$\hat{\mathbf{e}}\equiv\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\xrightarrow[N\to\infty]{}\mathbf{y}-\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\varepsilon}$$

この $\hat{\mathbf{e}}$ を用いて $\boldsymbol{\varepsilon}$ の代替とできることが分かる。ここで注意することは、$\hat{\mathbf{e}}$ の定義では $\hat{\mathbf{X}}$ ではなく、$\mathbf{X}$ の値をそのまま使うことである。以上より、

$$\hat{\mathbf{b}}\sim N(\boldsymbol{\beta},(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\boldsymbol{\Sigma}_{\varepsilon}\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1})$$
$$\to N(\boldsymbol{\beta},(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}\hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1})$$

有効性の検定では、重回帰分析で利用した以下の式を用いる。

$$F=(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}\hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})\big/q\sim F_{q,\infty}$$

ここに、

$\mathbf{R}((k+r)\times(k+r+1)) = [\mathbf{0}((k+r)\times 1) \quad \mathbf{I}((k+r)\times(k+r))]$, $\mathbf{r}((k+r)\times 1) = \mathbf{0}$

プログラムには組み込んでいないが、これは$\mathbf{R}$と$\mathbf{r}$の取り方によって、結合仮説の検定にも利用できる。

以下の関係を使うと、

$$\hat{\mathbf{b}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})$$
$$= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'((\hat{\mathbf{X}}+\hat{\mathbf{V}})\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{\beta}+(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\boldsymbol{\varepsilon}$$

$\hat{\mathbf{b}}$の分布は次のように書くこともできる。

$$\hat{\mathbf{b}} \sim N(\boldsymbol{\beta}, (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}) \to N(\boldsymbol{\beta}, (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{e}}\hat{\mathbf{e}}'\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1})$$

我々のプログラムではこの関係を利用している。

また、説明変数、操作変数共に１つの場合、

$$\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{Z}\mathbf{Z}'/\mathbf{Z}'\mathbf{Z}$$

より、以下となる。

$$\hat{\mathbf{b}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$
$$= \mathbf{Z}'\mathbf{y}/\mathbf{X}'\mathbf{Z} = \boldsymbol{\beta}+\mathbf{Z}'\boldsymbol{\varepsilon}/\mathbf{X}'\mathbf{Z} \to \boldsymbol{\beta}+\mathbf{Z}'\hat{\mathbf{e}}/\mathbf{X}'\mathbf{Z} \sim N(\boldsymbol{\beta}, (\mathbf{Z}'\hat{\mathbf{e}}/\mathbf{X}'\mathbf{Z})^2)$$

さて、2 段階法で求めた形が十分な説明力を持っているかどうかの問題を過剰識別制約の問題という。もし、十分な説明力があれば、2 段階法の誤差$\hat{\mathbf{e}}$を操作変数$\mathbf{Z}$で回帰しても係数がすべて 0 になるはずである。それを調べる問題を考えてみる。

回帰係数を$\mathbf{d}$、誤差を$\mathbf{u}_e$として以下のような回帰式を考える。

$$\hat{\mathbf{e}} = \mathbf{Z}\mathbf{d}+\mathbf{u}_e \tag{4}$$

この回帰式で定数項を含むすべての係数が 0 になるかどうか調べる検定はRVとEVを回帰変動と誤差変動、df_{RV}とdf_{EV}をそれらの自由度として、以下のように与えられる。

$$F = \frac{RV/df_{RV}}{EV/df_{EV}} \sim F_{df_{RV},df_{EV}} \tag{5}$$

ここで、推定値を用いて$\hat{\mathbf{e}} = \mathbf{Z}\hat{\mathbf{d}}+\hat{\mathbf{u}}_e$として、誤差変動と回帰変動は以下で与えられる。

$$EV = \hat{\mathbf{u}}_e'\hat{\mathbf{u}}_e = (\hat{\mathbf{e}}-\mathbf{Z}\hat{\mathbf{d}})'(\hat{\mathbf{e}}-\mathbf{Z}\hat{\mathbf{d}}) = (\hat{\mathbf{e}}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}})'(\hat{\mathbf{e}}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}})$$
$$= \hat{\mathbf{e}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\hat{\mathbf{e}} = \hat{\mathbf{e}}'(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\hat{\mathbf{e}}$$
$$RV = \hat{\mathbf{e}}'\hat{\mathbf{e}}-EV = \hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}$$

また、このRVの表現は、最小 2 乗法で求めた結合仮説の検定の統計量を使って以下のようにも求められる。

$$(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r}) = \hat{\mathbf{d}}'\mathbf{Z}'\mathbf{Z}\hat{\mathbf{d}}$$
$$= ((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}})'(\mathbf{Z}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}}$$
$$= \hat{\mathbf{e}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}} = \hat{\mathbf{e}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}} = \hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}$$

但し、$\mathbf{R} = \mathbf{I}_{m+r+1}$, $\mathbf{r} = \mathbf{0}$である。

最後に、(5)式を示しておく。

i) まず $\mathbf{Z}'\mathbf{Z}$ の正則性と対称性を用いると、ある正方行列 $\mathbf{F}$ で $\mathbf{Z}'\mathbf{Z}=\mathbf{F}'\mathbf{F}$ のように書けることから、$\mathbf{P}_{\mathbf{Z}}$ はある行列 $\mathbf{B}(N\times(m+r+1))$ を用いて以下のように書ける。

$$\mathbf{P}_{\mathbf{Z}}=\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'=(\mathbf{Z}\mathbf{F}^{-1})(\mathbf{Z}\mathbf{F}^{-1})'=\mathbf{B}\mathbf{B}',$$

以下の関係を用いると

$$\hat{\mathbf{e}}=\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}=\mathbf{X}(\boldsymbol{\beta}-\hat{\mathbf{b}})+\boldsymbol{\varepsilon}=[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}]\boldsymbol{\varepsilon}$$

回帰変動 RV は以下のように書ける。

$$\begin{aligned}RV&=\hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}=\boldsymbol{\varepsilon}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}]'\mathbf{P}_{\mathbf{Z}}[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}]\boldsymbol{\varepsilon}\\&=\boldsymbol{\varepsilon}'[\mathbf{P}_{\mathbf{Z}}-\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}]\boldsymbol{\varepsilon}\\&=\boldsymbol{\varepsilon}'[\mathbf{B}\mathbf{B}'-\mathbf{B}\mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}\mathbf{B}']\boldsymbol{\varepsilon}\\&=\boldsymbol{\varepsilon}'\mathbf{B}[\mathbf{I}_{m+r+1}-\mathbf{B}'\mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{B}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}]\mathbf{B}'\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}'\mathbf{B}[\mathbf{I}_{m+r+1}-\mathbf{P}_{\mathbf{B}'\mathbf{X}}]\mathbf{B}'\boldsymbol{\varepsilon}\end{aligned}$$

ここで、

$$\begin{aligned}\mathbf{B}'\boldsymbol{\varepsilon}&\sim N(\mathbf{0},\sigma^2\mathbf{B}'\mathbf{B})=N(\mathbf{0},\sigma^2\mathbf{F}'^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{F}^{-1})\\&=N(\mathbf{0},\sigma^2\mathbf{F}'^{-1}\mathbf{F}'\mathbf{F}\mathbf{F}^{-1})=N(\mathbf{0},\sigma^2\mathbf{I}_{m+r+1})\end{aligned}$$

であることから、以下となる。

$$\mathbf{z}\equiv\mathbf{B}'\boldsymbol{\varepsilon}/\sigma\sim N(\mathbf{0},\mathbf{I})$$

また $(\mathbf{B}'\mathbf{X})((m+r+1)\times(k+r+1))$ で $k\le m$ であることから、

$$rank(\mathbf{P}_{\mathbf{B}'\mathbf{X}})=k+r+1,\quad rank(\mathbf{I}_{m+r-1}-\mathbf{P}_{\mathbf{B}'\mathbf{X}})=m-k$$

よって、【公式 4】より、

$$RV/\sigma^2=\hat{\mathbf{e}}'\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}/\sigma^2=\mathbf{z}'[\mathbf{I}_{m+r+1}-\mathbf{P}_{\mathbf{B}'\mathbf{X}}]\mathbf{z}\sim\chi^2_{m-k}$$

ii) 次に、$N\to\infty$ を仮定すると、

$$rank(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})=N-m-r-1$$

と【公式 4】より、

$$EV/\sigma^2=(\hat{\mathbf{e}}/\sigma)'(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})(\hat{\mathbf{e}}/\sigma)\xrightarrow[N\to\infty]{}(\boldsymbol{\varepsilon}/\sigma)'(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})(\boldsymbol{\varepsilon}/\sigma)\sim\chi^2_{N-m-r-1}$$

iii) また、$\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{e}}$ と $(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\hat{\mathbf{e}}$ は

$$\begin{aligned}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\Sigma}_{\hat{\mathbf{e}}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})&=\mathbf{P}_{\mathbf{Z}}[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}]\boldsymbol{\Sigma}_{\varepsilon}[\mathbf{I}-\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'](\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\\&=\sigma^2\mathbf{P}_{\mathbf{Z}}[\mathbf{I}-\mathbf{P}_{\mathbf{Z}}\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\\&\qquad+\mathbf{X}(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'](\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\\&=\sigma^2\mathbf{P}_{\mathbf{Z}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})=\mathbf{0}\end{aligned}$$

であるから、【公式 2】より、独立である。

これらのことから、以下の関係を得る。

$$F=\frac{RV/(m-k)}{EV/(N-m-r-1)}=\frac{(\hat{\mathbf{e}}'\hat{\mathbf{e}}-\hat{\mathbf{u}}_e'\hat{\mathbf{u}}_e)/(m-k)}{\hat{\mathbf{u}}_e'\hat{\mathbf{u}}_e/(N-m-r-1)}\underset{N\to\infty}{\sim}F_{m-k,\infty}$$

参考文献

[1] J.H.Stock,M.W.Watson,宮尾龍蔵訳, 入門計量経済学, 共立出版, 2016.

２９．トービット回帰分析

29.1　トービット回帰分析とは

トービット回帰分析は目的変数のデータに切断のある場合の回帰分析である[1]。例えば、車への支出と所得との関係を考えると、車を持っている人、または購入する人は車にお金をかけるが、車を持っていない人の車への支出は０円である。このデータのように、ある所に下限や上限があるデータを切断されたデータと呼ぶ。このデータの切断はデータの床効果や天井効果とも呼ばれる。データの切断では、切断された値以下または以上のデータはすべて切断された値で置き換えられ、それ以上のデータの情報は不明なものとして扱われる。

例えば、図１は０で切断のあるデータである。左のグラフの直線は通常の回帰分析で、切断された値はそのままの値として最小２乗法で計算されている。それに対して右のグラフでは同じデータにトービット回帰分析を適用した場合の回帰直線が引かれている。

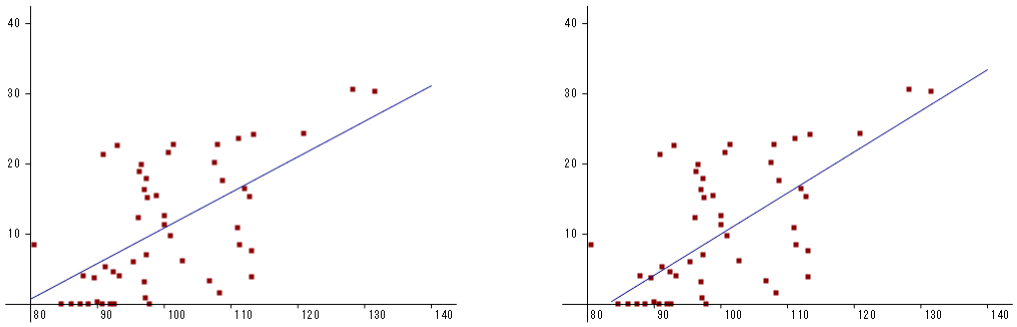


図１　通常の回帰直線とトービット回帰分析の回帰直線

トービット回帰分析の回帰直線は、通常の回帰直線に比べて、切断データがあたかもマイナスの位置にあるかのように、傾きが少し急になっている。これは、下からの切断の場合、同じ切断の点でも、横軸の負の方向に進むほどウェイトが小さく与えられるというトービット回帰分析の特徴である。

トービット回帰分析は、誤差分布が正規分布に従うという、最尤法を使ってパラメータの値を求める。切断データについては、密度関数の値を使わず、切断値を取る確率だけが分かるものとして尤度関数に取り入れられる。これは、生存時間分析の打ち切りデータの扱いと同じである。具体的に、下からの切断の場合、尤度関数L_lは正規分布の密度関数$f(x)$と分布関数$F(x)$を用いて以下で与えられる。

$$L_l=\prod_{\lambda=1}^{N} f(y_\lambda-Y_\lambda)^{\delta_\lambda} F(a-Y_\lambda)^{1-\delta_\lambda}\ ,\quad Y_\lambda=\sum_{i=1}^{p} b_i x_{i\lambda}+b_0$$

ここで、δ_λはλ番目の個体が切断値の場合に$\delta_\lambda=0$、切断値でない場合$\delta_\lambda=1$をとる。また、y_λは目的関数の実測値、Y_λは目的関数の予測値、aは切断値である。

下からの切断と同様に上からの切断の場合も考えられる。計算の詳細は３節トービット回帰分析の理論のところで述べる。

29.2　プログラムの利用法

メニュー［分析－多変量解析他－予測手法－トービット回帰分析］を選択すると、図１で与えられる分析実行画面が表示される。

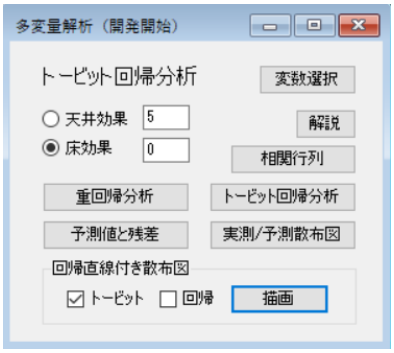


図１　トービット回帰分析実行画面

データは通常の重回帰分析と同様のデータ形式で、例えば図２のように与えられる。

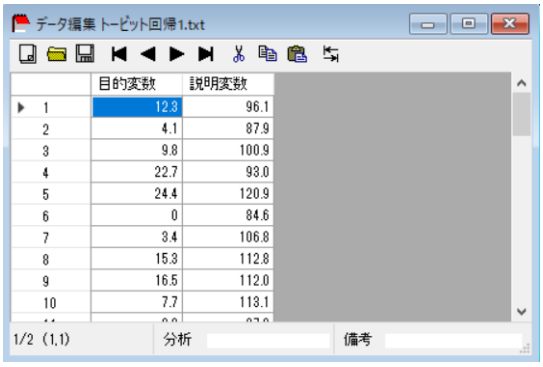


図２　トービット回帰分析データ（トービット回帰 1.txt）

ここでは説明変数が１つだけの場合であるが、複数の場合も重回帰分析と同じである。

「重回帰分析」ボタンをクリックすると比較のための通常の重回帰分析の結果が表示される。但し、トービット回帰の計算に合わせて、誤差項については均一分散として計算されている。結果を図３に示す。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
説明変数	0.5066	0.6045	5.2569	48	0.0000	0.604	0.604
切片	-39.7512	0.0000	-4.0868	48	1.9998		
R	0.604	R^2	0.365	調整済R	0.593	調整済R^2	0.352
F値	27.6347	自由度	1, 48	検定確率	0.0000		

図３　重回帰分析結果

トービット回帰分析の計算は、まず目的変数に天井効果があるか床効果があるかを選択し、切断値を右のテキストボックスに入力する。その後、変数を選んで、「トービット回帰」ボタンをクリックすると図４で与えられる結果が表示される。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95%下限	95%上限
説明変数	0.5855	0.6986	0.1112	5.2671	0.0000	0.3676	0.8033
切片	-48.5193	0.0000	11.2870	-4.2987	0.0000	-70.6414	-26.3973
実測予測R	0.608	R^2	0.369				

図４　トービット回帰分析結果

ここで、通常の回帰分析の場合、回帰分散と目的変数の分散との比が重相関係数の２乗（寄与率）に一致するが、トービット回帰分析の場合は一般に一致しない。どのようにすべきか考えたが､意味がはっきり分かる方が良いと考え､目的変数値と予測値との相関を採用した。また、予測値は「予測値と残差」ボタンで図５のように表示されるが、例えば下からの切断の場合、予測値が切断値より小さい場合は、切断値に置き換えている。

予測値と残差

	実測値	予測値	残差
1	12.3	7.744	4.556
2	4.1	2.944	1.156
3	9.8	10.555	-0.755
4	22.7	5.929	16.771
5	24.4	22.264	2.136
6	0	1.011	-1.011
7	3.4	14.009	-10.609
8	15.3	17.522	-2.222
9	16.5	17.053	-0.553

図５　予測値と残差

「実測/予測散布図」は、この実測値と予測値を使って、図６のように表される。

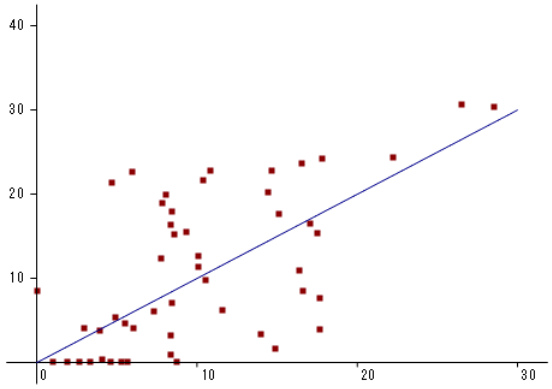


図６　実測/予測散布図

「回帰直線付き散布図」グループボックス内の「描画」ボタンでは、図７のような散布図が表示される。

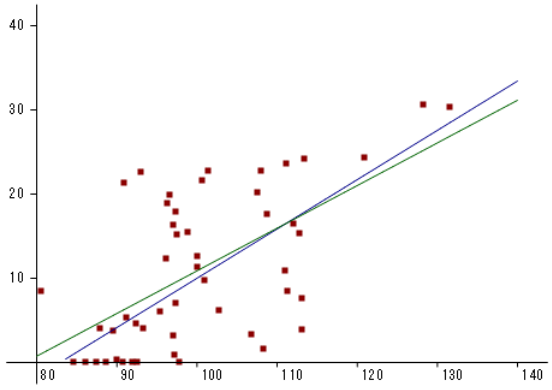


図７　回帰直線付き散布図

ここでは、「トービット」及び「回帰」チェックボックスにチェックを入れ、両方の回帰直線を表示している。傾きが大きい方がトービット回帰分析の結果である。

29.3　トービット回帰分析の理論

トービット回帰分析のモデルは、回帰式の誤差項が正規分布することが元になっている。まず、以下の回帰式を考える。

$$y_\lambda = \sum_{j=0}^{p} b_j x_{j\lambda} + u_\lambda \quad （x_{0\lambda}=1;\lambda=1,\cdots,N）$$

誤差項を$u_\lambda \sim N(0,\sigma^2)$で独立とすると、密度関数は以下になる。

$$f(u_\lambda) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-u_\lambda^2/2\sigma^2] = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(y_\lambda - \sum_{j=0}^{p} b_j x_{j\lambda}\right)^2\right]$$

ここで、標準正規分布の密度関数を$\phi(x)$とすると、以下のようにも表される。

$$f(u_\lambda) = \frac{1}{\sigma}\phi(u_\lambda/\sigma) = \frac{1}{\sigma}\phi\left[\frac{1}{\sigma}\left(y_\lambda - \sum_{j=0}^{p} b_j x_{j\lambda}\right)\right]$$

また、分布関数$F(x)$は標準正規分布の分布関数$\Phi(x)$を用いて以下のようにも表される。

$$F(u_\lambda) = \int_{-\infty}^{u_\lambda} f(t)dt = \int_{-\infty}^{u_\lambda/\sigma} \phi(x)dx$$

$$= \Phi(u_\lambda/\sigma) = \Phi\left[\frac{1}{\sigma}\left(y_\lambda - \sum_{j=0}^{p} b_j x_{j\lambda}\right)\right]$$

ここで、各データに対して、切断データの場合$\delta_\lambda = 0$、そうでないデータの場合$\delta_\lambda = 1$とすると、例えば下からの$y_\lambda = a$の切断（床効果）の場合、尤度関数Lは以下となる。

$$L_l = \prod_{\lambda=1}^{N} f(u_\lambda)^{\delta_\lambda} F(u_\lambda)^{1-\delta_\lambda}$$

また、上からの$y_\lambda = a$の切断（天井効果）の場合、尤度関数Lは以下となる。

$$L_u = \prod_{\lambda=1}^{N} f(u_\lambda)^{\delta_\lambda} [1-F(u_\lambda)]^{1-\delta_\lambda}$$

$y_\lambda = a$のとき、u_λの値は以下である。

$$u_\lambda = a - \sum_{j=0}^{p} b_j x_{j\lambda}$$

下からの切断の場合、尤度関数L_lは以下で与えられる。

$$L_l = \prod_{\lambda=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(y_\lambda - \sum_{j=0}^{p} b_j x_{j\lambda}\right)^2\right\}\right]^{\delta_\lambda} \Phi\left[\frac{1}{\sigma}\left(a - \sum_{j=0}^{p} b_j x_{j\lambda}\right)\right]^{1-\delta_\lambda}$$

同様に上からの切断の場合、尤度関数L_uは以下となる。

$$L_u = \prod_{\lambda=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(y_\lambda - \sum_{j=0}^{p} b_j x_{j\lambda}\right)^2\right\}\right]^{\delta_\lambda} \left\{1-\Phi\left[\frac{1}{\sigma}\left(a - \sum_{j=0}^{p} b_j x_{j\lambda}\right)\right]\right\}^{1-\delta_\lambda}$$

対数尤度関数はそれぞれ以下となる。

$$\log L_l = -\log\sqrt{2\pi}\sum_{\lambda=1}^{N}\delta_\lambda - \log\sigma\sum_{\lambda=1}^{N}\delta_\lambda - \frac{1}{2\sigma^2}\sum_{\lambda=1}^{N}\delta_\lambda\left(y_\lambda - \sum_{j=0}^{p}b_j x_{j\lambda}\right)^2$$
$$+\sum_{\lambda=1}^{N}(1-\delta_\lambda)\log\Phi\left[\frac{1}{\sigma}\left(a-\sum_{j=0}^{p}b_j x_{j\lambda}\right)\right]$$

$$\log L_u = -\log\sqrt{2\pi}\sum_{\lambda=1}^{N}\delta_\lambda - \log\sigma\sum_{\lambda=1}^{N}\delta_\lambda - \frac{1}{2\sigma^2}\sum_{\lambda=1}^{N}\delta_\lambda\left(y_\lambda - \sum_{j=0}^{p}b_j x_{j\lambda}\right)^2$$
$$+\sum_{\lambda=1}^{N}(1-\delta_\lambda)\log\left\{1-\Phi\left[\frac{1}{\sigma}\left(a-\sum_{j=0}^{p}b_j x_{j\lambda}\right)\right]\right\}$$

この後、ニュートン・ラフソン法を用いてパラメータの推定を行うが、その際、パラメータを以下のように設定する。

$$\tilde{\mathbf{b}} = \begin{pmatrix}\mathbf{b}\\ \sigma\end{pmatrix},\quad (\mathbf{b})_i = b_i\ (i=0,\cdots,p)$$

対数尤度をパラメータで微分してスコアベクトル$\mathbf{U}$と情報行列$\mathfrak{I}$を求めると以下となる。

$$\mathbf{U} = \begin{pmatrix}\partial\log L/\partial\mathbf{b}\\ \partial\log L/\partial\sigma\end{pmatrix},\quad \mathfrak{I} = -\begin{pmatrix}\partial^2\log L/\partial\mathbf{b}\partial^t\mathbf{b} & \partial^2\log L/\partial^t\mathbf{b}\partial\sigma\\ \partial^2\log L/\partial\mathbf{b}\partial\sigma & \partial^2\log L/\partial\sigma^2\end{pmatrix}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\tilde{\mathbf{b}}^{(m+1)} = \tilde{\mathbf{b}}^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここで右肩の(m)や$(m+1)$は、ニュートン・ラフソン法の計算のステップを表す。この情報行列の逆行列の対角成分はパラメータの分散を与える。

ニュートン・ラフソン法のために、定義と計算式を書いておく。

下からの切断の場合

$$\sum_{\lambda=1}^{N}\delta_\lambda = m,\quad Y_\lambda = \sum_{j=1}^{p}b_j x_{j\lambda}$$

$$\frac{\partial}{\partial b_i}\log L_l = \frac{1}{\sigma^2}\sum\nolimits_{\lambda=1}^{N}\delta_\lambda x_{i\lambda}\left(y_\lambda - Y_\lambda\right)$$
$$-\frac{1}{\sigma}\sum\nolimits_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]/\Phi\left[(a-Y_\lambda)/\sigma\right]$$

$$\frac{\partial}{\partial\sigma}\log L_l = -\frac{m}{\sigma} + \frac{1}{\sigma^3}\sum\nolimits_{\lambda=1}^{N}\delta_\lambda\left(y_\lambda - Y_\lambda\right)^2$$
$$-\frac{1}{\sigma^2}\sum\nolimits_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)\,\phi\left[(a-Y_\lambda)/\sigma\right]/\Phi\left[(a-Y_\lambda)/\sigma\right]$$

$$\frac{\partial^2}{\partial b_i \partial b_k}\log L_l = -\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}\delta_\lambda x_{i\lambda}x_{k\lambda}$$
$$-\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}x_{k\lambda}(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]/\Phi\left[(a-Y_\lambda)/\sigma\right]$$
$$-\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}x_{k\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]^2/\Phi\left[(a-Y_\lambda)/\sigma\right]^2$$

$$\frac{\partial^2}{\partial b_i \partial \sigma}\log L_l = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}\delta_\lambda x_{i\lambda}\left(y_\lambda - Y_\lambda\right)$$
$$+\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]/\Phi\left[(a-Y_\lambda)/\sigma\right]$$
$$-\frac{1}{\sigma^4}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)^2 x_{i\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]/\Phi\left[(a-Y_\lambda)/\sigma\right]$$
$$-\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]^2/\Phi\left[(a-Y_\lambda)/\sigma\right]^2$$

$$\frac{\partial^2}{\partial \sigma^2}\log L_l = \frac{m}{\sigma^2}-\frac{3}{\sigma^4}\sum_{\lambda=1}^{N}\delta_\lambda\left(y_\lambda - Y_\lambda\right)^2$$
$$+\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]/\Phi\left[(a-Y_\lambda)/\sigma\right]$$
$$-\frac{1}{\sigma^5}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)^3\phi\left[-Y_\lambda/\sigma\right]/\Phi\left[-Y_\lambda/\sigma\right]$$
$$-\frac{1}{\sigma^4}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)^2\phi\left[(a-Y_\lambda)/\sigma\right]^2/\Phi\left[(a-Y_\lambda)/\sigma\right]^2$$

上からの切断の場合

$$\frac{\partial}{\partial b_i}\log L_l = \frac{1}{\sigma^2}\sum_{\lambda=1}^{N}\delta_\lambda x_{i\lambda}\left(y_\lambda - Y_\lambda\right)$$
$$+\frac{1}{\sigma}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]/\left\{1-\Phi\left[\left(a-Y_\lambda\right)/\sigma\right]\right\}$$

$$\frac{\partial}{\partial \sigma}\log L_l = -\frac{m}{\sigma}+\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}\delta_\lambda\left(y_\lambda - Y_\lambda\right)^2$$
$$+\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]/\left\{1-\Phi\left[\left(a-Y_\lambda\right)/\sigma\right]\right\}$$

$$\frac{\partial^2}{\partial b_i \partial b_k}\log L_l = -\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}\delta_\lambda x_{i\lambda}x_{k\lambda}$$
$$+\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}x_{k\lambda}(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]/\left\{1-\Phi\left[\left(a-Y_\lambda\right)/\sigma\right]\right\}$$
$$-\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}x_{k\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]^2/\left\{1-\Phi\left[\left(a-Y_\lambda\right)/\sigma\right]\right\}^2$$

$$\frac{\partial^2}{\partial b_l \partial \sigma}\log L_l = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}\delta_\lambda x_{i\lambda}\left(y_\lambda - Y_\lambda\right)$$
$$-\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]\Big/\left\{1-\Phi\left[(a-Y_\lambda)/\sigma\right]\right\}$$
$$+\frac{1}{\sigma^4}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)^2 x_{i\lambda}\phi\left[(a-Y_\lambda)/\sigma\right]\Big/\left\{1-\Phi\left[(a-Y_\lambda)/\sigma\right]\right\}$$
$$-\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(1-\delta_\lambda)x_{i\lambda}(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]^2\Big/\left\{1-\Phi\left[(a-Y_\lambda)/\sigma\right]\right\}^2$$

$$\frac{\partial^2}{\partial \sigma^2}\log L_l = \frac{m}{\sigma^2}-\frac{3}{\sigma^4}\sum_{\lambda=1}^{N}\delta_\lambda\left(y_\lambda - Y_\lambda\right)^2$$
$$-\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)\phi\left[(a-Y_\lambda)/\sigma\right]\Big/\left\{1-\Phi\left[(a-Y_\lambda)/\sigma\right]\right\}$$
$$+\frac{1}{\sigma^5}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)^3\phi\left[-Y_\lambda/\sigma\right]\Big/\left\{1-\Phi\left[(a-Y_\lambda)/\sigma\right]\right\}$$
$$-\frac{1}{\sigma^4}\sum_{\lambda=1}^{N}(1-\delta_\lambda)(a-Y_\lambda)^2\phi\left[(a-Y_\lambda)/\sigma\right]^2\Big/\left\{1-\Phi\left[(a-Y_\lambda)/\sigma\right]\right\}^2$$

参考文献

[1] 浅野晳, 中村二朗, 計量経済学（y21）第 2 版, 有斐閣, 2009.

３０．産業連関分析

30.1　産業連関分析とは

産業連関分析は、国民経済の構造を生産技術的な連結関係で表す重要な手法である。産業構造は、投入と産出、輸出を含む最終需要、輸入、粗付加価値等を用いて、以下の産業連関表で記述される。

表１　競争輸入型産業連関表

	中間需要	最終需要	輸出	輸入	合計
中間投入	$\mathbf{X}(n\times n)$	$\mathbf{F}(n\times d)$	$\mathbf{E}(n\times e)$	$-\mathbf{M}(n\times p)$	$\mathbf{T}(n\times 1)$
粗付加価値	$\mathbf{V}(r\times n)$				
合計	${}^{t}\mathbf{T}(1\times n)$				

表２　非競争輸入型産業連関表

	中間需要	最終需要	輸出	輸入	合計
中間投入国内	$\mathbf{X}^{d}(n\times n)$	$\mathbf{F}^{d}(n\times d)$	$\mathbf{E}(n\times e)$	$\mathbf{0}\,(n\times p)$	$\mathbf{T}(n\times 1)$
中間投入輸入	$\mathbf{X}^{i}(n\times n)$	$\mathbf{F}^{i}(n\times d)$	$\mathbf{0}(n\times e)$	$-\mathbf{M}(n\times p)$	$\mathbf{0}\,(n\times 1)$
粗付加価値	$\mathbf{V}(r\times n)$				
合計	${}^{t}\mathbf{T}(1\times n)$				

ここに、それぞれの項目枠内は行列形式で表されており、行列の行数と列数は、$\mathbf{X}(n\times n)$ の形で右側の括弧の中に記述されている。即ち産業は n 部門、最終需要が d 部門、輸出が e 部門、輸入が p 部門、粗付加価値が r 部門あることになる。列ベクトル $\mathbf{T}$ は産業毎の国内での総産出量を表す。また、${}^{t}\mathbf{T}$ は $\mathbf{T}$ の転置行列である。輸入は通常１部門として列ベクトルで表すことが多いと思われるが、ここでは複数部門として部門合計を求められるようにしている。

表 1 と表 2 では産業連関表の重要な２つの形式を表示したが、この他に一部の主要な輸入品についてのみ非競争輸入扱いにした競争・非競争混合輸入型もある。混合輸入型については、一般的取扱いが困難であるので、指標の計算はすべて競争輸入型に直して行なうことにした。競争輸入型の場合、これらの表式間には以下のような関係がある。

$$\Sigma\mathbf{X}+\Sigma\mathbf{F}+\Sigma\mathbf{E}-\Sigma\mathbf{M}=\Sigma{}^{t}\mathbf{X}+\Sigma{}^{t}\mathbf{V}=\mathbf{T}.$$

図１に産業連関分析の分析メニューを示す。

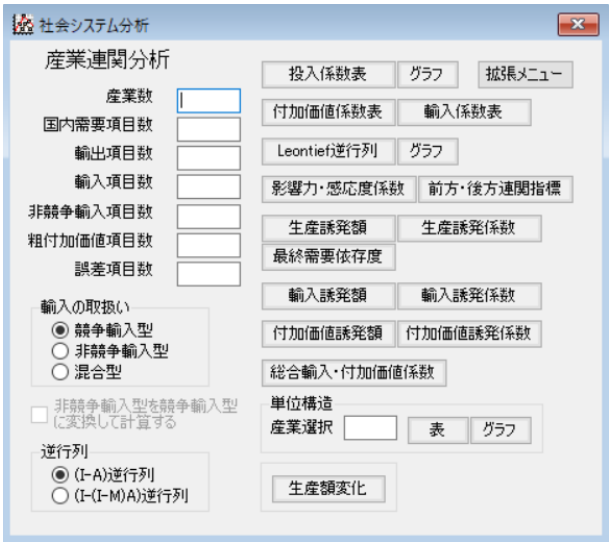


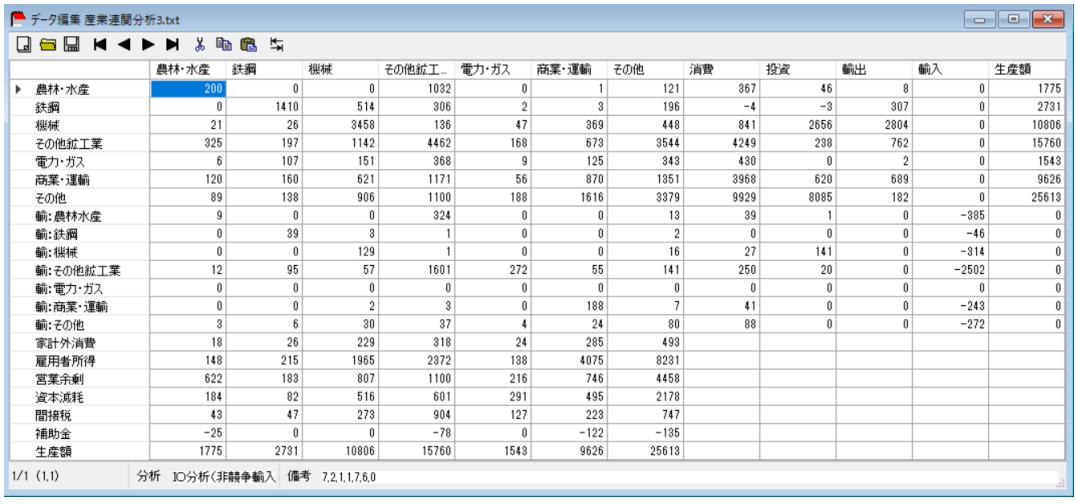
図１ 産業連関分析分析メニュー

図２と図３に競争輸入型と非競争輸入型の具体的なデータ画面を示す。

	農林・水産	鉄鋼	機械	他鉱工業	電力・ガス	商業・運輸	その他	消費	投資	輸出	輸入	生産額
農林・水産	209	0	0	1356	0	1	133	406	47	8	-385	1775
鉄鋼	0	1450	518	307	2	3	198	-4	-3	307	-46	2731
機械	21	26	3588	137	47	369	465	868	2797	2804	-314	10806
他鉱工業	337	291	1199	6063	440	729	3684	4499	258	762	-2502	15760
電力・ガス	6	107	151	368	9	125	343	430	0	2	0	1543
商業・運輸	120	161	624	1174	56	1058	1358	4009	620	689	-243	9626
その他	92	144	936	1137	192	1640	3459	10017	8085	182	-272	25613
家計外消費	18	26	229	318	24	285	493					
雇用者所得	148	215	1965	2372	138	4075	8231					
営業余剰	622	183	807	1100	216	746	4458					
資本減耗	184	82	516	601	291	495	2178					
間接税	43	47	273	904	127	223	747					
補助金	-25	0	0	-78	0	-122	-135					
生産額	1775	2731	10806	15760	1543	9626	25613					

1/1 (1,1) 分析 IO分析(競争輸入型 備考 7,2,1,1,0,6,0

図２ 競争輸入型産業連関表のデータ



	農林・水産	鉄鋼	機械	その他鉱工...	電力・ガス	商業・運輸	その他	消費	投資	輸出	輸入	生産額
農林・水産	200	0	0	1032	0	1	121	367	46	8	0	1775
鉄鋼	0	1410	514	306	2	3	196	-4	-3	307	0	2731
機械	21	26	3458	136	47	369	448	841	2656	2804	0	10806
その他鉱工業	325	197	1142	4462	168	673	3544	4249	238	762	0	15760
電力・ガス	6	107	151	368	9	125	343	430	0	2	0	1543
商業・運輸	120	160	621	1171	56	870	1351	3968	620	689	0	9626
その他	89	138	906	1100	188	1616	3379	9929	8085	182	0	25613
輸:農林水産	9	0	0	324	0	0	13	39	1	0	-385	0
輸:鉄鋼	0	39	3	1	0	0	2	0	0	0	-46	0
輸:機械	0	0	129	1	0	0	16	27	141	0	-314	0
輸:その他鉱工業	12	95	57	1601	272	55	141	250	20	0	-2502	0
輸:電力・ガス	0	0	0	0	0	0	0	0	0	0	0	0
輸:商業・運輸	0	0	2	3	0	188	7	41	0	0	-243	0
輸:その他	3	6	30	37	4	24	80	88	0	0	-272	0
家計外消費	18	26	229	318	24	285	493					
雇用者所得	148	215	1965	2372	138	4075	8231					
営業余剰	622	183	807	1100	216	746	4458					
資本減耗	184	82	516	601	291	495	2178					
間接税	43	47	273	904	127	223	747					
補助金	-25	0	0	-78	0	-122	-135					
生産額	1775	2731	10806	15760	1543	9626	25613					

1/1 (1,1) 分析 IO分析(非競争輸入 備考 7,2,1,1,7,6,0

図３ 非競争輸入型産業連関表のデータ

これらは参考文献の巻末に記載されている例をこのプログラムに合うように入力したもの

である。

分析を実行するために必要な入力項目は、「産業数」、「国内需要項目数」、「輸出項目数」、「輸入項目数」、「非競争輸入項目数」、「粗付加価値項目数」である。図 2 のデータの「産業数」は 7、「消費・投資項目数」は 2、「輸出項目数」は 1、「輸入項目数」は 1、「粗付加価値項目数」は 6 である。

「非競争輸入項目数」は非競争輸入型及び混合輸入型の産業連関表の場合用いるもので、粗付加価値方向に輸入項目がある場合の項目数である。競争輸入型の場合これは 0 になる。また、「誤差項目数」は輸入項目の右隣りに配置される項目であるが、産業連関表はあくまで行と列の合計が一致することが原則であるので通常この項目は 0 である。特別な事情のある場合のみ利用することもありえると考えて設けている。取扱いについては今後の経験の中から決めて行きたい。

項目「基準ページ」は複数年次（複数ページ）の産業連関表を入力している場合、どのページを利用するかということを指定する項目である。データが、競争輸入型であるか、非競争輸入型であるか、混合輸入型であるかは、「輸入の取扱い」のオプションボタンで選択する。レオンチェフ逆行列の形式は考えるバランスモデルによって変わってくる。ここではよく利用される２つの形式を「逆行列」のオプションボタンで選択する。非競争輸入型や混合輸入型のデータを競争輸入型に変えて計算するためのチェックボックスも用意されている。特に混合輸入型の場合、このプログラムでは競争輸入型で計算する以外の方式は作成していない。その他の入力データやコマンドボタンについては、分析の解説と共に説明する。

30.2　産業連関分析の理論

ここでは、表示の簡単化のために以下の表式を用いる。

- 任意の列または行ベクトル $\mathbf{C}$ の各要素を対角成分として作られる対角行列を $diag(\mathbf{C})$、その逆行列を $diag^{-1}(\mathbf{C})$ と表すことにする。
- 行列 $\mathbf{A}(m\times n_a)$ と $\mathbf{B}(m\times n_b)$ を列方向に並べて作られる m 行 n_a+n_b 列の行列を $\mathbf{A}\oplus\mathbf{B}$ と表すことにする。
- 行列 $\mathbf{A}$ の行和をとって得られる列ベクトルを $\Sigma\mathbf{A}$、列和をとって得られる行ベクトルを $\Sigma'\mathbf{A}$ とする。
- 行列の成分は括弧付きで添え字を付けて表すか、イタリック文字に添え字を付けて表すかどちらかにする。即ち、行列 $\mathbf{X}$ の (i,j) 成分は $(\mathbf{X})_{ij}=X_{ij}=x_{ij}$ である。

産業の生産技術構造は競争輸入型、非競争輸入型それぞれ、以下の投入係数行列を用いて表されるが、今後表式の最初に a), b) を付けて、それぞれ競争輸入型、非競争輸入型とする。

$$\text{a)}\ \mathbf{A}=(a_{ij})=(X_{ij}/T_j)=\mathbf{X}diag^{-1}(\mathbf{T}) \quad \text{（競争輸入型）}$$

$$\text{b)}\ \mathbf{A}^d=(a^d_{ij})=(X^d{}_{ij}/T_j)=\mathbf{X}^d diag^{-1}(\mathbf{T}) \quad \text{（非競争輸入型）}. \qquad (1)$$

図 1 の分析メニューでは「投入係数表」のコマンドボタンをクリックすることにより求め

られる。結果は表形式で与えられるが、求まった表のセル幅、桁数合わせや文字の配置は結果の表示画面の中で設定する。また、グラフも右隣の「グラフ」のコマンドボタンで、３次元立体棒グラフとして表示される。グラフの簡単な設定は、結果グラフ表示フォームの中で行うことが出来る。表示関係のこれらの機能についても、必要なものを追加して行かなければならない。

図１の「付加価値係数」$\tilde{\mathbf{V}}$は、競争輸入型も非競争輸入型も以下で表す。

$$\tilde{\mathbf{V}} = \mathbf{V}diag^{-1}(\mathbf{T}). \qquad (2)$$

付加価値の全ての項目の和をとった付加価値係数ベクトル$\tilde{\mathbf{V}}_0$も以下のように計算出来る。

$$\tilde{\mathbf{V}}_0 = (\Sigma'\mathbf{V})diag^{-1}(\mathbf{T}). \qquad (3)$$

「輸入係数」$\tilde{\mathbf{M}}$は「輸入／国内投入」の意味を持っており、輸入の扱いに応じて以下のようになる。

a) $\tilde{\mathbf{M}} = diag^{-1}(\mathbf{AT} + \Sigma\mathbf{F})\Sigma\mathbf{M}$

b) $\tilde{\mathbf{M}} = diag^{-1}(\mathbf{A}^d\mathbf{T} + \Sigma\mathbf{F}^d + \Sigma\mathbf{M})\Sigma\mathbf{M}.$ (4)

次に、Leontief 逆行列により生産の波及構造を調べるが、輸入の取り扱いによりバランス式が異なり、それによって逆行列の表式が異なってくる。特に競争輸入型の場合に注意すると、バランス式は以下のようになる。

a) $\mathbf{T} = \mathbf{AT} + \Sigma\mathbf{F} + \Sigma\mathbf{E} - \Sigma\mathbf{M}$

$= \mathbf{AT} + \Sigma\mathbf{F} + \Sigma\mathbf{E} - \overline{\mathbf{M}}(\mathbf{AT} + \Sigma\mathbf{F})$

$= (\mathbf{I} - \overline{\mathbf{M}})\mathbf{AT} + (\mathbf{I} - \overline{\mathbf{M}})\Sigma\mathbf{F} + \Sigma\mathbf{E}$

b) $\mathbf{T} = \mathbf{A}^d\mathbf{T} + \Sigma\mathbf{F}^d + \Sigma\mathbf{E}.$ (5)

ここに$\overline{\mathbf{M}}$は輸入係数ベクトルの成分を対角成分として得られた正方行列で、$\overline{\mathbf{M}} = diag(\tilde{\mathbf{M}})$である。これより、国内総生産を求めると、以下のようになる。

a) $\mathbf{T} = (\mathbf{I} - \mathbf{A})^{-1}(\Sigma\mathbf{F} + \Sigma\mathbf{E} - \Sigma\mathbf{M})$

$= [\mathbf{I} - (\mathbf{I} - \overline{\mathbf{M}})\mathbf{A}]^{-1}[(\mathbf{I} - \overline{\mathbf{M}})\Sigma\mathbf{F} + \Sigma\mathbf{E}]$

b) $\mathbf{T} = (\mathbf{I} - \mathbf{A}^d)^{-1}(\Sigma\mathbf{F}^d + \Sigma\mathbf{E}).$ (6)

これより Leontief 逆行列は、競争輸入型の場合は２通り、非競争輸入型の場合は１通り考えることにする。

a) $\mathbf{B} = (b_{ij}) = (\mathbf{I} - \mathbf{A})^{-1}$ または、$\mathbf{B} = (b_{ij}) = [\mathbf{I} - (\mathbf{I} - \overline{\mathbf{M}})\mathbf{A}]^{-1}$

b) $\mathbf{B} = (b_{ij}) = (\mathbf{I} - \mathbf{A}^d)^{-1}.$ (7)

図１のメニューでは「Leontief 逆行列」のボタンをクリックする際に、「輸入の取扱い」と「逆行列」のオプションボタンの選択によって、これらを選択出来るようになっている。

メニュー中の「影響力・感応度係数」ボタンのクリックによって、産業別の影響力係数$n\Sigma'\mathbf{B}/\Sigma'\Sigma\mathbf{B}$と感応度係数 $n\Sigma\mathbf{B}/\Sigma'\Sigma\mathbf{B}$を表示する。ここに$\mathbf{B}$は（2.7）のそれぞれの表式を用いる。また、同様の指標として、前方連関指標$\Sigma\mathbf{B}$と後方連関指標$\Sigma'\mathbf{B}$も「前方・後方連関指標」ボタンのクリックによって求めることが出来る。

さて、最終需要項目別の生産誘発額$\mathbf{T}'$は、「生産誘発額」ボタンをクリックすることによ

り求めることが出来る。式 (6), (7) より、容易にその計算式の意味が理解出来るであろう。

$$\begin{aligned} &\text{a)}\ \mathbf{T}'(n\times(d+e)) = [\mathbf{I}-(\mathbf{I}-\overline{\mathbf{M}})\mathbf{A}]^{-1}[(\mathbf{I}-\overline{\mathbf{M}})\mathbf{F}\oplus\mathbf{E}] \\ &\qquad\qquad = \mathbf{B}[(\mathbf{I}-\overline{\mathbf{M}})\mathbf{F}\oplus\mathbf{E}] \\ &\text{b)}\ \mathbf{T}'(n\times(d+e)) = (\mathbf{I}-\mathbf{A}^d)^{-1}(\mathbf{F}^d\oplus\mathbf{E}) \\ &\qquad\qquad = \mathbf{B}(\mathbf{F}^d\oplus\mathbf{E}). \end{aligned} \tag{8}$$

最終需要の各項目別産業合計に対する生産誘発額の割合を表す生産誘発係数$\tilde{\mathbf{T}}'$は、(8)より以下のように与えられる。

$$\begin{aligned} &\text{a)}\ \tilde{\mathbf{T}}' = \mathbf{B}\Gamma\mathbf{F}\,diag^{-1}(\Sigma'\mathbf{F})\oplus\mathbf{B}\mathbf{E}\,diag^{-1}(\Sigma'\mathbf{E}) \\ &\text{b)}\ \tilde{\mathbf{T}}' = \mathbf{B}\mathbf{F}^d\,diag^{-1}(\Sigma'\mathbf{F}^d)\oplus\mathbf{B}\mathbf{E}\,diag^{-1}(\Sigma'\mathbf{E}). \end{aligned} \tag{9}$$

ここに、$\Gamma = \mathbf{I}-\overline{\mathbf{M}}$ である。全需要による生産誘発係数ベクトル $\tilde{\mathbf{T}}'_0$ は、以下で与えられる。

$$\begin{aligned} &\text{a)}\ \tilde{\mathbf{T}}'_0 = \mathbf{B}(\Gamma\Sigma\mathbf{F}+\Sigma\mathbf{E})/\Sigma'(\Sigma\mathbf{F}+\Sigma\mathbf{E}) \\ &\text{b)}\ \tilde{\mathbf{T}}'_0 = \mathbf{B}(\Sigma\mathbf{F}^d+\Sigma\mathbf{E})/\Sigma'(\Sigma\mathbf{F}^d+\Sigma\mathbf{E}). \end{aligned} \tag{10}$$

これらの結果は「生産誘発係数」ボタンにより、表示することが出来る。

また、特に競争輸入型の場合、輸入 $\Sigma\mathbf{M}$ は(2.7)から求まる $\mathbf{B}-\mathbf{I}=\Gamma\mathbf{A}\mathbf{B}$ の表式及び Γ の定義から、(11)のように書き換えることが出来る。

$$\begin{aligned} \Sigma\mathbf{M} &= -(\mathbf{I}-\mathbf{A})\mathbf{T}+\Sigma\mathbf{F}+\Sigma\mathbf{E} \\ &= -(\mathbf{I}-\mathbf{A})\mathbf{B}(\Gamma\Sigma\mathbf{F}+\Sigma\mathbf{E})+\Sigma\mathbf{F}+\Sigma\mathbf{E} \\ &= \overline{\mathbf{M}}\Gamma^{-1}\mathbf{B}\Gamma\Sigma\mathbf{F}+\overline{\mathbf{M}}\mathbf{A}\mathbf{B}\Sigma\mathbf{E}. \end{aligned} \tag{11}$$

これより各最終需要項目による輸入誘発額$\mathbf{M}'$として以下を得る。

$$\mathbf{M}' = \overline{\mathbf{M}}\Gamma^{-1}\mathbf{B}\Gamma\mathbf{F}\oplus\overline{\mathbf{M}}\mathbf{A}\mathbf{B}\mathbf{E}. \tag{12}$$

これは「輸入誘発額」ボタンにより求めることが出来る。

輸入誘発額の最終需要に対する割合として定義される輸入誘発係数$\tilde{\mathbf{M}}'$は、以下のように与えられる。

$$\tilde{\mathbf{M}}' = \overline{\mathbf{M}}\Gamma^{-1}\mathbf{B}\Gamma\mathbf{F}\,diag^{-1}(\Sigma'\mathbf{F})\oplus\overline{\mathbf{M}}\mathbf{A}\mathbf{B}\mathbf{E}\,diag^{-1}(\Sigma'\mathbf{E}). \tag{13}$$

同様にして全需要項目による輸入誘発係数$\tilde{\mathbf{M}}'_0$は以下で与えられる。

$$\tilde{\mathbf{M}}'_0 = \overline{\mathbf{M}}(\Gamma\mathbf{B}\Gamma^{-1}\Sigma\mathbf{F}+\mathbf{A}\mathbf{B}\Sigma\mathbf{E})/\Sigma'(\Sigma\mathbf{F}+\Sigma\mathbf{E}). \tag{14}$$

これらは「輸入誘発係数」ボタンにより求めることが出来る。

付加価値係数の(3)式より、付加価値は以下のように与えられ、

$$\Sigma^t\mathbf{V} = {}^t\tilde{\mathbf{V}}_0\,diag(\mathbf{T}) = diag(\tilde{\mathbf{V}}_0)\mathbf{T}, \tag{15}$$

産出額の式より、これは競争輸入型と非競争輸入型に分けて、以下のように書き換えられる。

$$\begin{aligned} &\text{a)}\ \Sigma^t\mathbf{V} = diag(\tilde{\mathbf{V}}_0)\mathbf{B}(\Gamma\Sigma\mathbf{F}+\Sigma\mathbf{E}) \\ &\text{b)}\ \Sigma^t\mathbf{V} = diag(\tilde{\mathbf{V}}_0)\mathbf{B}(\Sigma\mathbf{F}^d+\Sigma\mathbf{E}). \end{aligned} \tag{16}$$

これより、それぞれの最終需要項目が付加価値を誘発する額を表す、付加価値誘発額ベクトル$\mathbf{V}'$は以下の式で与えられる。

$$\begin{aligned} &\text{a)}\ \mathbf{V}' = diag(\tilde{\mathbf{V}}_0)\mathbf{B}(\Gamma\mathbf{F}\oplus\mathbf{E}) \\ &\text{b)}\ \mathbf{V}' = diag(\tilde{\mathbf{V}}_0)\mathbf{B}(\mathbf{F}^d\oplus\mathbf{E}) \end{aligned} \tag{17}$$

これは「付加価値誘発額」ボタンにより求められる。

付加価値誘発額の最終需要に対する割合として定義される付加価値誘発係数$\tilde{\mathbf{V}}'$は、以下のように定義される。

a) $\tilde{\mathbf{V}}' = diag(\tilde{\mathbf{V}}_0)\mathbf{B}[\Gamma\mathbf{F}diag^{-1}(\Sigma'\mathbf{F}) \oplus \mathbf{E}diag^{-1}(\Sigma'\mathbf{E})]$

b) $\tilde{\mathbf{V}}' = diag(\tilde{\mathbf{V}}_0)\mathbf{B}[\mathbf{F}^d diag^{-1}(\Sigma'\mathbf{F}^d) \oplus \mathbf{E}diag^{-1}(\Sigma'\mathbf{E})]$. (18)

また、全付加価値項目による付加価値誘発係数は、以下で与えられる。

a) $\tilde{\mathbf{V}}' = diag(\tilde{\mathbf{V}}_0)\mathbf{B}(\Gamma\Sigma\mathbf{F} + \Sigma\mathbf{E})/\Sigma'(\Sigma\mathbf{F} + \Sigma\mathbf{E})$

b) $\tilde{\mathbf{V}}' = diag(\tilde{\mathbf{V}}_0)\mathbf{B}(\Sigma\mathbf{F}^d + \Sigma\mathbf{E})/\Sigma'(\Sigma\mathbf{F}^d + \Sigma\mathbf{E})$. (19)

これは「付加価値係数」ボタンにより求められる。

競争輸入型の場合全輸入額と全付加価値額は、以下の形で与えられるが、

$$\Sigma'\Sigma\mathbf{M} = \Sigma'\overline{\mathbf{M}}\Gamma^{-1}\mathbf{B}\Gamma\Sigma\mathbf{F} + \Sigma'\overline{\mathbf{M}}\mathbf{AB}\Sigma\mathbf{E} \quad (20)$$

$$\Sigma'\Sigma^t\mathbf{V} = \Sigma' diag(\tilde{\mathbf{V}}_0)\mathbf{B}\Gamma\Sigma\mathbf{F} + \Sigma' diag(\tilde{\mathbf{V}}_0)\mathbf{B}\Sigma\mathbf{E}, \quad (21)$$

総合輸入係数・総合付加価値係数はこれらの輸入額・付加価値額の合計の、$\Sigma\mathbf{F}$及び、$\Sigma\mathbf{E}$に係る係数ベクトルをそれぞれ、消費・投資に係る係数、輸出に係る係数と呼んだものである。また、最終需要合計に係る係数は、最終需要合計に占める各産業別の消費・投資と輸出の割合を掛けて、それぞれの係数を足して、$\Sigma'\overline{\mathbf{M}}\Gamma^{-1}\mathbf{B}\Gamma\mathbf{W}_f + \Sigma'\overline{\mathbf{M}}\mathbf{AB}\mathbf{W}_e$のように定義する。ここに、$\mathbf{W}_f = diag(\Sigma\mathbf{F})/\Sigma'(\Sigma\mathbf{F} + \Sigma\mathbf{E})$，$\mathbf{W}_e = diag(\Sigma\mathbf{E})/\Sigma'(\Sigma\mathbf{F} + \Sigma\mathbf{E})$である。これらは、行ベクトルであるので、表示に際しては転置を取ったものを用いる。

産業iの需要を１単位だけ満たすための各産業の投入産出関係は、unit structure と呼ばれる。ここで、$\mathbf{e}_i = {}^t(0,0,\cdots,1,\cdots,0)$は第 i 成分のみ 1 でその他は 0 の縦ベクトルとすると、競争輸入型の(2.6)の関係式$\mathbf{T} = (\mathbf{I} - \mathbf{A})^{-1}(\Sigma\mathbf{F} + \Sigma\mathbf{E} - \Sigma\mathbf{M})$で、$(\Sigma\mathbf{F} + \Sigma\mathbf{E} - \Sigma\mathbf{M})$の代わりに$\mathbf{e}_i$を用いて、これを(1)から求まる関係式$\mathbf{X} = \mathbf{A}diag(\mathbf{T})$の中に代入して、unit structure $\mathbf{U}_i$が以下のように求められる。

$$\mathbf{U}_i = \mathbf{A}diag((\mathbf{I} - \mathbf{A})^{-1}\mathbf{e}_i). \quad (22)$$

これは、図１の「単位構造」フレームで「産業選択」のテキストボックスに産業番号を書き込み、「表」か「グラフ」のボタンをクリックすることによって求められる。

投入係数の変化の問題を扱うには、RAS 法がよく用いられる。RAS 法では、$\mathbf{A}$から$\mathbf{A}'$への投入係数の変化を$\mathbf{A}' = \hat{\mathbf{R}}\mathbf{A}\hat{\mathbf{S}}$のように、代替変化乗数ベクトル$\mathbf{R}$と加工度変化乗数ベクトル$\mathbf{S}$を用いて記述する。ここに、$\hat{\mathbf{R}} = diag(\mathbf{R})$，$\hat{\mathbf{S}} = diag(\mathbf{S})$である。RAS 法を用いた分析には、基準時点と比較時点の２時点の産業連関表と２時点間の年数が必要であり、これらのデータをテキストボックスに入力した後、「代替・加工度変化」ボタンをクリックする。結果は１年間当たりのそれぞれのベクトルの値が表示される。直接２時点間の差が見たい場合には、２時点間の年数を１にすればよい。また、比較時点の投入係数行列に、１年間の代替変化乗数ベクトルと加工度変化乗数ベクトルより作られる対角行列を複数回掛けて、将来の予測をすることも出来る。これは、予測年次のテキストボックスに何年後かの値を入れて、「予測投入係数表」のボタンをクリックすることによって求めることが出来る。(RAS 法

については、新しいバージョンで削除している）

新しいバージョンでは３次元棒グラフを３Ｄビューアを用いたものに変えている。図４は Leontief 逆行列の棒グラフ表示である。

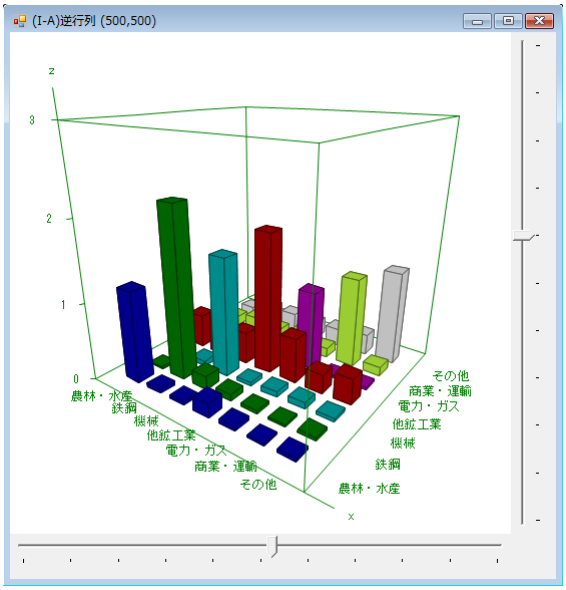


図 4　レオンチェフ逆行列

参考文献

[1] 宮沢健一編，産業連関分析入門，日本経済新聞社, 1991.

３１．経済時系列分析

時系列分析は分野によって様々な手法が考えられているが、C.Analysis にはこれまで２つの手法を取り入れてきた。１つは、データを傾向変動、振幅変動、季節変動に分解して予測する変動の分解モデルで、基本的に単一のデータの過去のふるまいから未来を予測するパターン分析モデルである。しかし、時系列データは自身の過去のデータだけに影響されるわけではなく、他の指標からの影響も受けている。そのため、我々は自己回帰モデルに他の変数を加えた自己回帰・分布ラグモデルを考え、それを元にして簡単な変動の分解モデルを加えたパネル時系列分析というプログラムを作った。しかし、このプログラムは、ラグ次数の設定がすべての変数で固定されていたり、統計的な処理が単純であったりと、経済の分野で使用するには機能が不十分であった。そのため、今回新しく経済時系列分析で使われる手法に特化したプログラムの開発を進めることにした。

1.　自己回帰・分布ラグモデル

経済分野で使われる時系列分析では自己回帰に基づく重回帰分析が基本で、その利用には変数の階差変換や変数ごとの時間的なラグ次数の設定など、柔軟な対応を必要とする。メニュー［分析－多変量解析他－経済・経営手法－経済時系列分析］を選択すると、図１に与えられる分析実行画面が表示される。

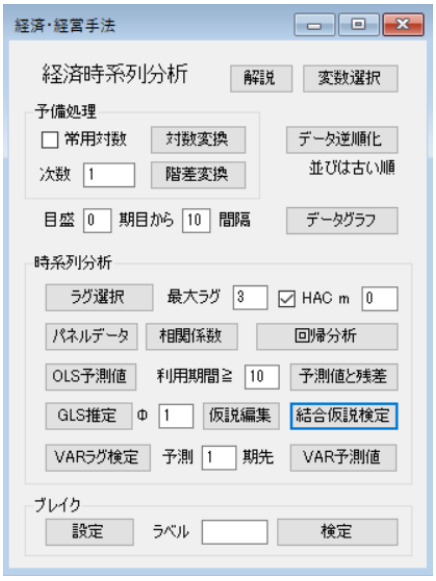


図１　経済時系列分析実行画面

この分析の変数選択は通常のものと異なり、変数を選択する部分とそのラグを設定する部分とに分かれる。右上の「変数選択」ボタンでは利用する変数を選択し、そのラグは「時系列分析」グループボックス内の「ラグ選択」ボタンで設定する。

まずここでは変数自体の変換について述べ、その後ラグの選択について説明する。時系列分析の変数変換では、対数変換がよく利用される。これは変数の変化率についての変動をみる場合に用いられる変換である。以後図２のデータを用いて説明を行う。

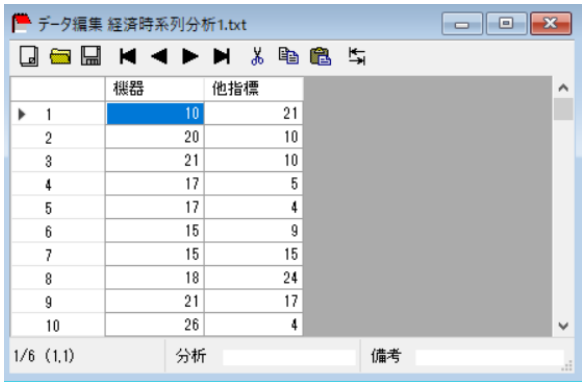


図２　データ（経済時系列分析 1.txt）

変数「機器」は、自己回帰を与える部分で、「他指標」は自己回帰以外を代表した「機器」に影響を与える変数として理解してもらいたい。

変数選択ですべての変数を選択し、「対数変換」ボタンをクリックすると図３のような結果を、次数を１として「階差変換」をクリックすると図４のような結果を得る。

図３　対数変換結果

図４　階差変換結果

これらの結果は、結果グリッドのメニュー「編集－エディタ全列追加」または「編集－エディタ指定列追加」を用いて、図 2 のグリッドエディットのデータに図 5 のように簡単に追加できる。

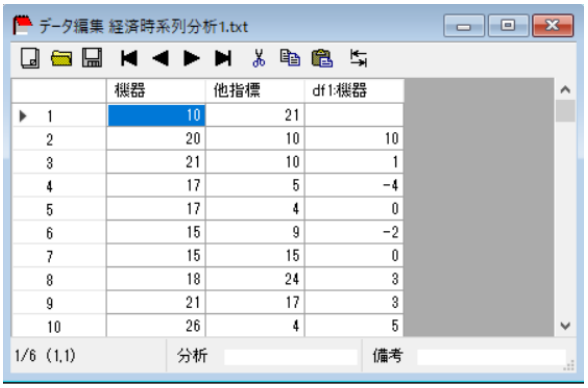


図５　機器の１階階差データの追加

このデータの中から例えば「機器」を選択し、「データグラフ」ボタンをクリックすると、図６のようなグラフが表示される。

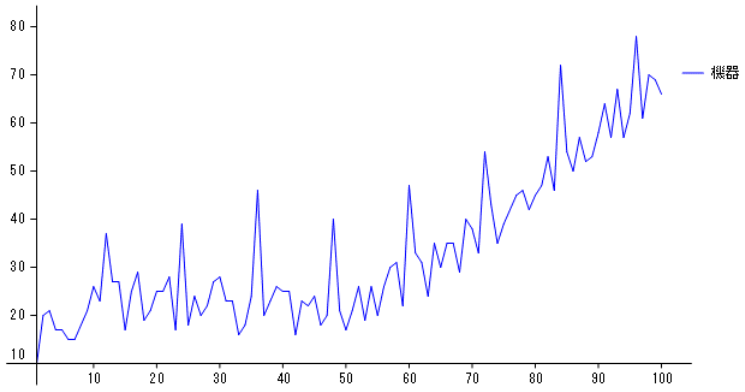


図６　グラフ表示結果

時系列分析は多くのデータを扱うことから、軸の目盛りは間隔をあけて表示することもできるし、凡例の表示・非表示、補助線の描画、グラフポイントの描画など、表示には一応の機能が付いている。

次に「機器」と「他指標」を選択し、「時系列分析」グループボックス内の最大ラグを 3 にして、図 7 のようにラグを選択する。ここでは最初に選択した変数について、最大ラグまでのすべてのラグを個別に選択できるようになっている。

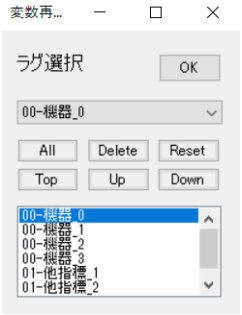
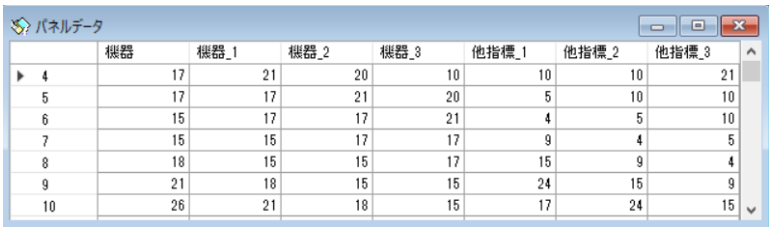


図７　ラグ選択画面

ラグの選択後「パネルデータ」ボタンをクリックすると、図 8 のようなパネルデータが出力される。

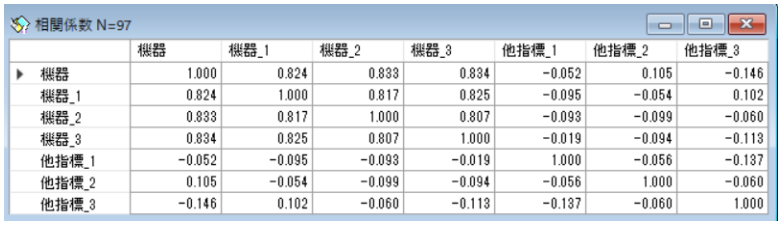


	機器	機器_1	機器_2	機器_3	他指標_1	他指標_2	他指標_3
4	17	21	20	10	10	10	21
5	17	17	21	20	5	10	10
6	15	17	17	21	4	5	10
7	15	15	17	17	9	4	5
8	18	15	15	17	15	9	4
9	21	18	15	15	24	15	9
10	26	21	18	15	17	24	15

図８　分析用パネルデータ

これが時系列分析の元となるデータである。「変数名_番号」の番号がラグを表している。また、ラグの付いていない変数はラグなしの元データである。ラグを取ったために、データはすべて使えるわけではなく、最大ラグの次の位置から始まっている。

「相関係数」ボタンをクリックすると図 8 に与えられるパネルデータの相関係数が図 9 のように表示される。

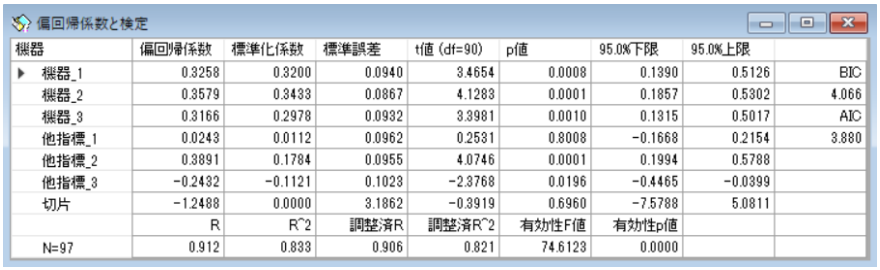


相関係数 N=97

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

図 9　パネルデータ相関係数

パネルデータを使った重回帰分析は、「回帰分析」ボタンをクリックすると図 10a の結果が表示される。これは均一分散の結果であるが、単純な不均一分散の場合の結果は「HAC」チェックボックスをクリックし、テキストボックス「m」を 0 にすると図 10b のようになる。「m」の値を変えると 4 節に述べるように時系列相関が考慮される。



偏回帰係数と検定

機器	偏回帰係数	標準化係数	標準誤差	t値 (df=90)	p値	95.0%下限	95.0%上限	
▶ 機器_1	0.3258	0.3200	0.0940	3.4654	0.0008	0.1390	0.5126	BIC
機器_2	0.3579	0.3433	0.0867	4.1283	0.0001	0.1857	0.5302	4.066
機器_3	0.3166	0.2978	0.0932	3.3981	0.0010	0.1315	0.5017	AIC
他指標_1	0.0243	0.0112	0.0962	0.2531	0.8008	−0.1668	0.2154	3.880
他指標_2	0.3891	0.1784	0.0955	4.0746	0.0001	0.1994	0.5788	
他指標_3	−0.2432	−0.1121	0.1023	−2.3768	0.0196	−0.4465	−0.0399	
切片	−1.2488	0.0000	3.1862	−0.3919	0.6960	−7.5788	5.0811	
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		
N=97	0.912	0.833	0.906	0.821	74.6123	0.0000		

図 10a　重回帰分析結果（均一分散の場合）

偏回帰係数と検定

機器	偏回帰係数	標準化係数	標準誤差	z値	p値	95.0%下限	95.0%上限	
▶ 機器_1	0.3258	0.3200	0.0968	3.3657	0.0008	0.1361	0.5156	BIC
機器_2	0.3579	0.3433	0.0893	4.0081	0.0001	0.1829	0.5329	4.066
機器_3	0.3166	0.2978	0.0941	3.3652	0.0008	0.1322	0.5010	AIC
他指標_1	0.0243	0.0112	0.0912	0.2670	0.7895	−0.1544	0.2031	3.880
他指標_2	0.3891	0.1784	0.0973	3.9985	0.0001	0.1984	0.5798	
他指標_3	−0.2432	−0.1121	0.0929	−2.6176	0.0089	−0.4253	−0.0611	
切片	−1.2488	0.0000	3.0134	−0.4144	0.6786	−7.1549	4.6572	
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		
N=97	0.912	0.833	0.906	0.821	88.5922	0.0000		

図 10b　重回帰分析結果（不均一分散の場合）

重回帰分析に用いたラグ次数について適正な値を得るために、分析結果の図 10 には BIC（Bayes Information Criterion）と AIC（Akaike Information Criterion）の値を示してある。これらは値が小さくなるほど良いモデルとして知られた指標である。一般に BIC>AIC となる。ここで用いたこれらの指標の定義は、モデルに使われた係数を k（定数項を含む）、利用した時刻数を T、回帰式の残差の 2 乗和を SSR として以下で与えられる。この他に全体に T を掛けたものなどもある。

$$BIC = \ln(SSR/T) + k(\ln T/T)$$

$$AIC = \ln(SSR/T) + k(2/T)$$

データの予測については、時系列分析の場合、通常の重回帰分析のようにすべての計算データを使った予測とその時点までのデータを使った予測を分けて考えなければならない。前者を推測または「OLS 予測」と呼び、後者を単に予測または「準サンプル外予測」と呼ぶ。「OLS 予測」ボタンをクリックすると、図 11 のように OLS 予測値と OLS 残差が表示される。

	実測値	OLS予測値	OLS残差
4	17	14.9450	2.0550
5	17	19.7193	-2.7193
6	15	16.6345	-1.6345
7	15	15.6650	-0.6650
8	18	17.2839	0.7161
9	21	18.9658	2.0342
10	26	22.8891	3.1109
11	23	21.3128	1.6872
12	37	20.1817	16.8183
13	27	35.8314	-8.8314
14	27	31.8665	-4.8665

図 11　OLS 予測値と OLS 残差

その時点までのデータを使った予測値は、「予測値と残差」ボタンで求められる。その際、過去にデータのないところでは予測が計算できないので、最低限必要な過去のデータを「データ期間」で指定しておく必要がある。例えばこれを 10 とすると、10 番目の古いデータから計算をはじめるので、最低限過去の 9 個のデータを用いて計算することになる。結果を図 12 に示す。結果は 4 番目から 12 番目が計算されないため空欄になっている。但し、あまり形式はよくないが、4 列目に残差の 2 乗平均の平方根である RMSFE や実測と予測の相関係数 R などの結果を加えている。

	実測値	予測値	残差	
4	17			RMSFE
5	17			6.013
6	15			R
7	15			0.933
8	18			R^2
9	21			0.871
10	26			
11	23			
12	37			
13	27	27.0820	-0.0820	
14	27	27.9062	-0.9062	
15	17	20.3265	-3.3265	
16	25	25.5641	-0.5641	
17	29	22.7334	6.2666	

図 12　予測値と残差

RMSFE は予測誤差の推定値であるので、以下の式を用いて予測の 95%信頼区間を求めることもできる。

$$\hat{y}_{t+1} - 1.96 \times RMSFE \le y_{t+1} \le \hat{y}_{t+1} + 1.96 \times RMSFE$$

さて、ここで目的変数のラグと他指標のラグを用いた回帰分析を考えたが、他指標はどれが必要であろうか。これを調べる検定は「グレンジャーの因果性テスト」と呼ばれる。この検定は他指標のラグに関する係数がすべて 0 かどうかを調べる結合仮説検定に帰着する。結合仮説検定は他の場面でも利用されるため、ここでも重回帰分析やロジスティック回帰分析と同様のメニューを追加している。ここでも簡単にその利用法を説明する。

「仮説編集」ボタンをクリックすると図 13 のような結合仮説編集画面が表示される。

	機器_1	機器_2	機器_3	他指標_1	他指標_2	他指標_3	切片		
制約1	1							=	0
制約2		1						=	0
制約3			1					=	0
制約4				1				=	0
制約5					1			=	0
制約6						1		=	0
制約7									

図 13　結合仮説編集画面

結合仮説は、各制約行で与えられた線形制約の積事象の検定となる。例えば、グレンジャーの因果性テストでは、他指標の係数がすべて 0 を検定するので、結合仮説編集画面は、図 14 のようになる。

図 14　グレンジャーの因果性テストの編集画面

ここで、制約は制約 4 のところから始まっているが、これは全体を上にあげて、制約 1 のところから始めてもよいし、このままでもよい。この編集画面を表示したまま「結合仮説検定」ボタンをクリックすると、図 15 のような結果が得られる。この結果によると「他指標」は回帰式に必要であることになる。

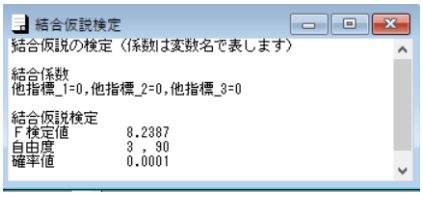


図 15　結合仮説検定（グレンジャーの因果性テスト）結果

2.　augmented Dickey-Fuller（ADF）テスト

ここでは時系列のトレンドの問題について考える。時系列データのトレンドには決定論的トレンドと確率トレンドがある。決定論的トレンドは時間に関して線形の関数で表され、ある時系列データはこの直線の周りで変動することになる。これに対して確率トレンドは、変動に定常性はなく、時間と共に変化する。例えばある時期まで上昇でそれ以後下降に変わるなどである。経済現象に関するトレンドは多くの場合、確率トレンドである。

確率トレンドの最も単純なモデルはランダムウォークモデルである。最も単純なランダムウォークモデルは以下の形式である。

$$y_t = y_{t-1} + u_t, \qquad u_t \text{ は i.i.d.}$$

乱数を変えた 2 つの例を図 1 に示す。このような不規則な変動が確率トレンドである。

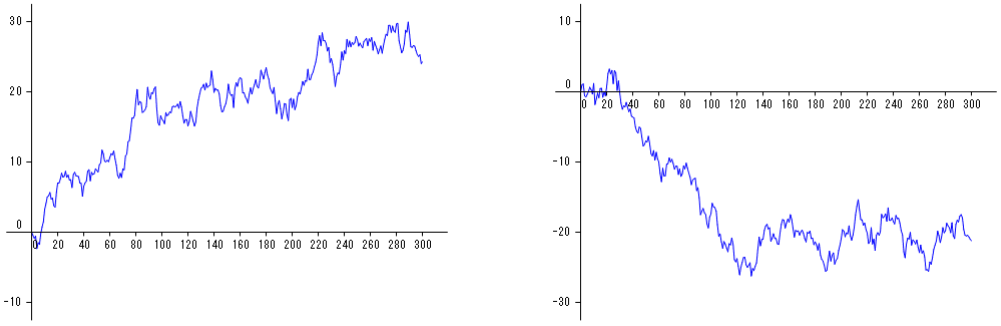


図 1　ランダムウォークモデル

このモデルでは、y_t の分散は時間と共に変化し、際限なく増加する。これは右辺の y_{t-1} の係数が 1 になっていることが原因である。このようなランダムウォークモデルが採択された場合、データそのものより、データの階差を用いて分析を進める。

ラグ次数 1 の自己回帰モデル（AR(1)モデル）は以下の形で与えられるが、

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

この係数 β_1 が 1 より小さいことが y_t の分散が発散しない（y_t が定常である）条件である。$\beta_1 = 1.5$ と $\beta_1 = 0.5$ の例を図 2 に示す。前者は期間を 1/10 に設定してある。結果の違いは明らかだろう。

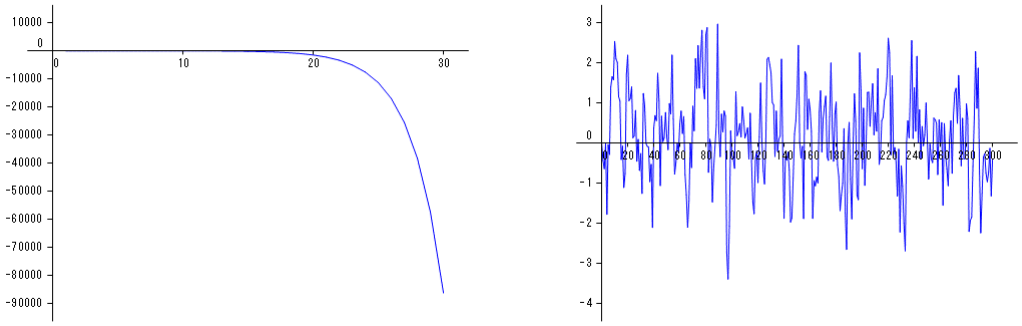


図 2　AR(1)モデル（$\beta_1 = 1.5, \beta_1 = 0.5$）

一般のラグ次数 p の自己回帰モデル（AR(p)モデル）は以下の形で与えられる。

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \cdots + \beta_p y_{t-p} + \beta_0 + u_t$$

このモデルが定常であるための条件は、以下の方程式の解 z が $|z| > 1$ となることである。

$$1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_p z^p = 0$$

以後、この問題を理解しやすくするため、$p = 3$ とする。モデルは以下となる。

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \beta_0 + u_t \qquad (1)$$

この式を変形すると以下となる。

$$y_t + b_1 y_{t-1} + b_2 y_{t-2} + b_0 = c(y_{t-1} + b_1 y_{t-2} + b_2 y_{t-3} + b_0) + u_t \qquad (2)$$

この変形はいつでも可能である。例えば、$p = 3$ の場合以下の手順で行う。

$$cb_0 - b_0 = \beta_0 \rightarrow b_0 = \beta_0/(c-1)$$
$$cb_2 = \beta_3 \rightarrow b_2 = \beta_3/c$$
$$cb_1 - b_2 = \beta_2 \rightarrow \quad b_1 = (b_2 + \beta_2)/c = \beta_3/c^2 + \beta_2/c$$
$$c - b_1 = \beta_1 \rightarrow c = b_1 + \beta_1 = \beta_3/c^2 + \beta_2/c + \beta_1$$

これより、c は、$\beta_3/c^3 + \beta_2/c^2 + \beta_1/c = 1$ の解であることが分かる。

改めて、$Y_t = y_t + b_1 y_{t-1} + b_2 y_{t-2} + b_0$ とおくと、(2)式は以下のように書ける。

$$Y_t = cY_{t-1} + u_t \qquad (3)$$

これを書き下すと、

$$Y_t = c^n Y_{t-n} + \sum_{i=0}^{n-1} c^i u_{t-i}$$

となり、$n \to \infty$ で Y_t の分散が収束するためには $|c|<1$ が必要である。今、$z = 1/c$ とおくと、収束条件は以下となる。

$\beta_1 z + \beta_2 z^2 + \beta_3 z^3 = 1$　の解が　$|z|>1$　　(4)

これを一般の p に拡張するのは容易である。

上の(4)式は $z=1$ として結合仮説の検定によって $|z|\neq 1$ を調べることもできるが、$|z|<1$ の可能性は捨てられない。しかし、以下のような変形によって結合仮説を用いず、直接調べることもできる。(2)式より、

$$Y_t - Y_{t-1} = (c-1)Y_{t-1} + u_t$$

となるが、左辺と右辺を分けて展開すると、$(c-1)b_0 = \beta_0$ として、

左辺：$Y_t - Y_{t-1} = \Delta y_t + b_1 \Delta y_{t-1} + b_2 \Delta y_{t-2}$

右辺：$(c-1)Y_{t-1} + u_t = (c-1)(y_{t-1} + b_1 y_{t-2} + b_2 y_{t-3} + b_0) + u_t$

$$= (c-1)(y_{t-1} + b_1 y_{t-2} + b_2 y_{t-3}) + \beta_0 + u_t$$

$$= (c-1)y_{t-1} + (c-1)\{b_1 \Delta y_{t-2} + (b_1 + b_2)y_{t-3}\} + \beta_0 + u_t$$

以上より、

$$\Delta y_t + b_1 \Delta y_{t-1} + b_2 \Delta y_{t-2} = (c-1)y_{t-1} + (c-1)\{b_1 \Delta y_{t-2} + (b_1 + b_2)y_{t-3}\} + \beta_0 + u_t$$

これをまとめて、

$$\Delta y_t = (c-1)y_{t-1} - b_1 \Delta y_{t-1} + \{(c-1)b_1 - b_2\}\Delta y_{t-2} + (c-1)(b_1 + b_2)y_{t-3} + \beta_0 + u_t$$

$$= \delta y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \gamma_3 y_{t-3} + \beta_0 + u_t$$

として $\delta = 0$ を検定すればよいことになる。

しかし、上の式はそのまま使うことができない。なぜなら、上の式に以下の多重共線性が見られるからである。

$$y_{t-1} - \Delta y_{t-1} - \Delta y_{t-2} - y_{t-3} = 0$$

そのため、$\gamma_3 y_{t-3}$ の項は捨てて以下の式とする。

$$\Delta y_t = \delta y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \beta_0 + u_t$$

この式には Δy_{t-2} 項があるため、y_{t-3} の情報も含まれている。

最後に、ここで注意することは、検定では対立仮説が $\delta < 0$ となる片側検定を用いることである（結果として $-2 < \hat{\delta} < 0$ なら問題ないであろう）。検定確率は誤差の相関などから、通常の t 検定の検定確率を用いることができず、後に表 1 で示す、参考文献 [1] に与えられた数値を用いる。プログラムでは解説の中に示している。

最後に、以下の関係を実際のデータで調べてみる。

検定１

$\beta_1 z + \beta_2 z^2 + \beta_3 z^3 = 1$　の解が　$|z|=1$　　即ち、$\beta_1 + \beta_2 + \beta_3 = 1$ の結合仮説検定

検定２

$\Delta y_t = \delta y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \beta_0 + u_t$　　として、$\delta = 0$ の検定

ここでは図１のデータを解析し、2 つの検定の一致性を見てみよう。このデータでは、「機器」とその階差である「df1_機器」を利用する。階差データは、実行メニューの「階差変換」ボタンで簡単に求められる。

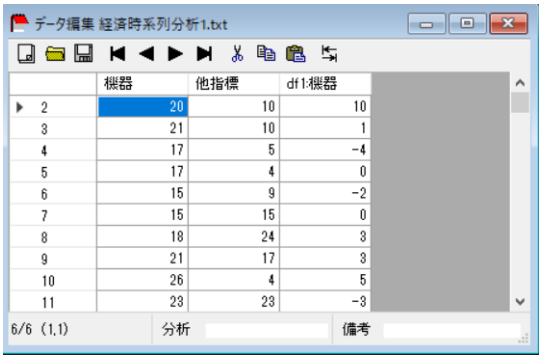


図１　データ（経済時系列分析 1.txt）

最初に検定１を調べる。

まず変数選択で、「機器」を選択する。次にラグを設定するために、「ラグ選択」で、「機器_0, 機器_1, 機器_2, 機器_3」を順番に選択する。その後、「結合仮説編集」ボタンで、出てきた結果を図 2 のように書き換える。これを開いたままで実行メニューの「結合仮説検定」をクリックすると、図 3 のような結果が得られる。

図２　結合仮説編集

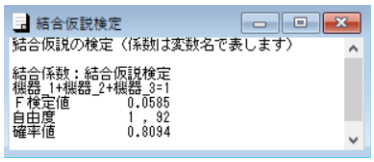


図３　結合仮説検定結果

ここで注目するのは、F 検定値（0.0585）と検定確率値（0.8094）である。検定確率値は定常性がない場合には利用できないが、比較には使える。

次に検定２を調べる。

変数選択で、「df1_機器と機器」を選択する。次にラグを設定するために、「変数ラグ選択」で、「df1_機器_0, 機器_1, df1_機器_1, df1_機器_2」を順番に選択する。そのまま、「回帰分析」ボタンをクリックすると図 4 の結果が表示される。

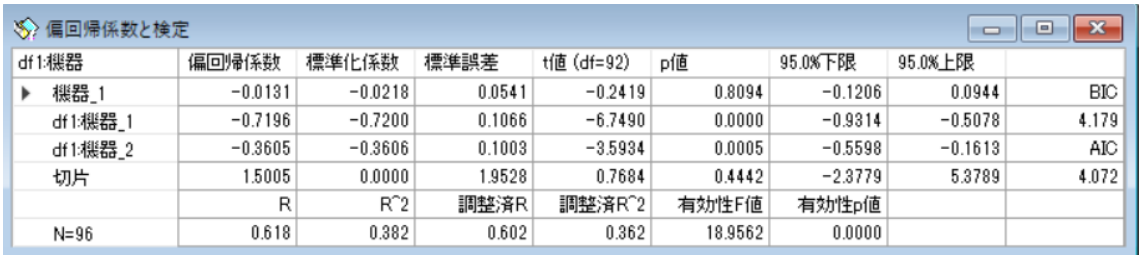


図４　ADF テスト結果

ここでは、「機器_1」の t 値（-0.2419）と p 値（0.8094）に注目する。結果としては、確率トレンドがないと言えないとなるが、t 値の 2 乗（0.0585）と p 値はともに結合仮説検定の結果に一致する。

ADF 検定の検定確率は分布のずれのため通常の t 検定の検定確率とは異なる。利用者は ADF テスト結果の t 値を参考文献［1］に掲載されている表 1 と比較して判定する。

表 1　ADF テストの検定値

	10%	5%	1%
定数項のみ	-2.57	-2.86	-3.43
定数項と時間トレンド	-3.12	-3.41	-3.96

3. Quandt Likelihood Ratio （QLR）統計量

時系列データのブレイクとは、ある時点で回帰直線の傾きまたは切片が変化する現象である。これには制度や政策または経済情勢などの変化が原因する。このブレイクの有無は以下のようにして調べることができる。

ある時点 τ でブレイクが起こっているかどうか調べたい場合、以下のようなダミー変数 $D_\tau(t)$ を利用する。

$$D_\tau(t)=\begin{cases}0 & t<\tau\\ 1 & t\geq\tau\end{cases}$$

目的変数 y_t に対して、例えば以下のような回帰式を考える。$D_\tau(t)$ のかかった項は交差項である。

$$y_t=\beta_1 y_{t-1}+\cdots+\beta_p y_{p-1}+\beta_0+\gamma_1 y_{t-1}D_\tau(t)+\cdots+\gamma_p y_{t-p}D_\tau(t)+\gamma_0 D_\tau(t)+u_t$$

この式は交差項を持った自己回帰モデルであるが、別の変数のラグを加えても構わない。この回帰分析において、以下の結合仮説が棄却されるとき、

$$\gamma_1=0,\cdots,\gamma_p=0,\gamma_0=0$$

回帰式にはブレイクがあると判定される。

ブレイクが起こっている時点が特定できない場合、分析期間 T の中で $0.15T$ と $0.85T$ までの間で、結合仮説検定最大の F 値を与える時刻 τ を取って検定を行う。その際、最大の F 値は正規性を持たないため、参考文献[1]で示された表から臨界値を求めるが、プログラム中でも対応する値が表示される。

プログラムの動きを見てみよう。変数選択では「機器」、変数ラグ選択ではラグ 3 までのすべての変数を選び、分析実行メニューの「ブレイク」グループボックス内の「設定」ボタンをクリックすると、図 1 のようなブレイク編集画面が表示される。

図 1　ブレイク編集画面

図 1 で 1 が入力された変数の交差項がすべて 0 である検定を行うことになる。この場合制約数は 4 である。ラベルテキストボックスに調べたい時点のラベルを入力するとその時点

について、何も入力しないと期間 $0.15T$ から $0.85T$ の間で最大の F 値を探してブレイクの検定を行う。入力しない場合の結果は、テキスト表示と F 値の変化を示したグラフ表示になる。結果を図 2 と図 3 に示す。

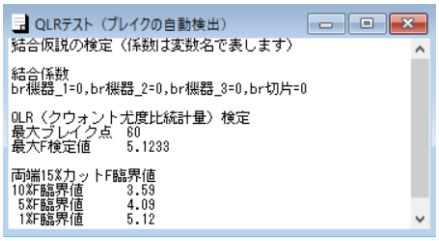


図 2　QLR 推定結果

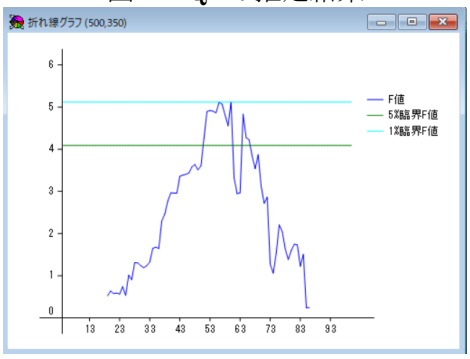


図 3　QLR 推定グラフ

結果は時刻 60 でブレイクが起こっていることになる。図 3 では F 値の変化がよく分かる。軸の目盛は分析実行画面の「データグラフ」の目盛の設定と共用にしている。使い勝手によって今後修正していく。

4.　Heteroscedasticity and Autocorrelation Constant（HAC）標準誤差の導入

時系列データを用いた回帰分析において、誤差項は説明変数に依存した時間的な変化と時間についての系列相関を持つ可能性が高い。ここでは誤差が不均一分散で系列相関がある場合の重回帰分析について説明する。これは参考文献[1]で述べられた説明変数が 1 つの場合からの拡張である。プログラム上でこの機能を利用するためには、プログラム実行画面の「HAC」チェックボックスにチェックを入れるが、後に述べる trancation parameter について、右のテキストボックスを空欄にしておくと自動で値が設定される。必要な場合は空欄に数値を入れ自分で設定することもできる。特に 0（または 1 も）と設定すると、通常の不均一分散の処理になる。

目的変数を p 個の説明変数と定数項で回帰する重回帰式を以下のように仮定する。

$$\mathbf{y}=\mathbf{Z}\mathbf{d}+\mathbf{u}$$

ここに、$\mathbf{y}(T\times 1)$，$\mathbf{Z}=(\mathbf{1}\quad \mathbf{X}(T\times p))$，$\mathbf{d}'=(b_0\quad \mathbf{b}'(1\times p))$，$\mathbf{u}(T\times 1)\sim N(\mathbf{0},\Sigma)$

最小 2 乗法で以下の量の最小化を考える。

$$L=(\mathbf{y}-\mathbf{Z}\mathbf{d})'(\mathbf{y}-\mathbf{Z}\mathbf{d})$$

回帰係数 $\mathbf{d}$ で微分して、回帰係数の推定値 $\hat{\mathbf{d}}$ を求めると以下となる。

$$\frac{\partial L}{\partial \mathbf{d}} = -2\mathbf{Z}'(\mathbf{y}-\mathbf{Z}\mathbf{d}) = \mathbf{0} \quad \text{より、} \hat{\mathbf{d}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

この $\hat{\mathbf{d}}$ を書き換えると、

$$\hat{\mathbf{d}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Z}\mathbf{d}+\mathbf{u}) = \mathbf{d}+(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}$$

となり、これを用いると $\hat{\mathbf{d}}$ の平均と分散は以下となる。

$$E[\hat{\mathbf{d}}] = \mathbf{d}+(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E[\mathbf{u}] = \mathbf{d}$$

$$Cov[\hat{\mathbf{d}},\hat{\mathbf{d}}] = E[(\hat{\mathbf{d}}-\mathbf{d})(\hat{\mathbf{d}}-\mathbf{d})'] = (\mathbf{Z}'\mathbf{Z})^{-1}E[\mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z}](\mathbf{Z}'\mathbf{Z})^{-1}$$

ここで、$v_{i\lambda} = Z_{i\lambda}u_\lambda$ として、

$$\begin{aligned} E[\mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z}]_{ij} &= E\left[\sum_{\lambda=1}^{T}\sum_{\lambda'=1}^{T} Z_{i\lambda}Z_{j\lambda'}u_\lambda u_{\lambda'}\right] = E\left[\sum_{\lambda=1}^{T}\sum_{\lambda'=1}^{T} v_{i\lambda}v_{j\lambda'}\right] \\ &= E\left[\sum_{\lambda=1}^{T} v_{i\lambda}v_{j\lambda} + \sum_{\lambda=1}^{T}\sum_{k=1}^{\lambda-1}(v_{i\lambda}v_{j\lambda-k}+v_{j\lambda}v_{i\lambda-k})\right] \\ &= \sum_{\lambda=1}^{T} Cov[v_{i\lambda},v_{j\lambda}] + \sum_{k=1}^{T-1}\sum_{\lambda=k+1}^{T}\{Cov[v_{i\lambda},v_{j\lambda-k}]+Cov[v_{i\lambda},v_{j\lambda-k}]\} \end{aligned}$$

時間について系列相関は一定であると仮定すると、以下と考えられる。

$$\sum_{\lambda=1}^{T} Cov[v_{i\lambda},v_{j\lambda}] = TCov[v_i,v_j]$$

$$\sum_{k=1}^{T-1}\sum_{\lambda=k+1}^{T}\{Cov[v_{i\lambda},v_{j\lambda-k}]+Cov[v_{j\lambda},v_{i\lambda-k}]\} = \sum_{k=1}^{T-1}(T-k)\{Cov[v_i,v_{j(k)}]+Cov[v_j,v_{i(k)}]\}$$

ここで、$Cov[v_i,v_{j(k)}]$は、v_i とその k 期前の v_j との共分散を表すものとする。この推定値としては Newey-West が与えた以下の形を採用する。

$$E[\mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z}]_{ij} \to \frac{T}{T-p-1}\left[\sum_{\lambda=1}^{T}\hat{v}_{i\lambda}\hat{v}_{j\lambda} + \sum_{k=1}^{m-1}\frac{m-k}{m}\sum_{\lambda=k+1}^{T}(\hat{v}_{i\lambda}\hat{v}_{j\lambda-k}+\hat{v}_{j\lambda}\hat{v}_{i\lambda-k})\right] \equiv T(\tilde{\mathbf{\Sigma}}_{\hat{\mathbf{v}}})_{ij}$$

ここに m は、truncation parameter と呼ばれ、ガイドラインとして以下の値が使われる。

$$m = 0.75T^{1/3}$$

但しこの m は系列相関の強弱によって変更してもよい。

以上より、回帰係数は以下の分布となる。

$$\hat{\mathbf{d}}-\mathbf{d} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \sim N(\mathbf{0},\mathbf{\Sigma}_{\hat{\mathbf{d}}}), \qquad \mathbf{\Sigma}_{\hat{\mathbf{d}}} \equiv (\mathbf{Z}'\mathbf{Z})^{-1}E[\mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z}](\mathbf{Z}'\mathbf{Z})^{-1}$$

計算には上の推定値を用いて、

$$\hat{\mathbf{d}}-\mathbf{d} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \sim N(\mathbf{0},\tilde{\mathbf{\Sigma}}_{\hat{\mathbf{d}}}), \qquad \tilde{\mathbf{\Sigma}}_{\hat{\mathbf{d}}} \equiv T(\mathbf{Z}'\mathbf{Z})^{-1}\tilde{\mathbf{\Sigma}}_{\hat{\mathbf{v}}}(\mathbf{Z}'\mathbf{Z})^{-1}$$

特に１変数回帰の場合は以下となり、参考文献 [1]に与えられた形となる。

$$E[\sum_{\lambda=1}^{T}\sum_{\lambda'=1}^{T}(x_\lambda-\bar{x})u_\lambda u_{\lambda'}(x_\lambda-\bar{x})] \to T\hat{\sigma}_{\hat{v}}^2\left[1+2\sum_{k=1}^{m-1}\frac{m-k}{m}\tilde{\rho}_{(k)}\right] \equiv T\hat{\sigma}_{\hat{v}}^2 f_T \equiv T\tilde{\sigma}_{\hat{v}}^2$$

$$\hat{b}_1 - b_1 = \frac{1}{T\sigma_x^2}(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{u} \sim N(0,\sigma_{\hat{b}_1}^2), \qquad \tilde{\sigma}_{\hat{b}_1}^2 \equiv \frac{\hat{\sigma}_{\hat{v}}^2}{T(\hat{\sigma}_x^2)^2}f_T$$

ここに、

$$\hat{\sigma}_{\hat{v}}^2 \equiv \frac{1}{T-2}\sum_{\lambda=1}^{T}\hat{v}_\lambda^2, \quad \tilde{\rho}_{(k)} \equiv \sum_{\lambda=k+1}^{T}\hat{v}_\lambda\hat{v}_{\lambda-k} \Big/ \sqrt{\sum_{\lambda=1}^{T}\hat{v}_\lambda^2\sum_{\lambda=1}^{T}\hat{v}_\lambda^2}, \quad m = 0.75T^{1/3}$$

ここからはプログラムの動作を紹介する。分布ラグモデルの OLS 推定と（HAC 標準誤差用いた）HAC 推定を図 1 のデータを元に比較してみる。

図 1　OLS 推定と HAC 推定の比較用データ（経済時系列分析 2(GLS).txt）

「最大ラグ」を 1 にして、1 次のラグまで考えた結果は、「変数選択」で 2 つの変数を選び、「ラグ選択」で、目的変数の 0 次と説明変数の 0 次と 1 次を選んで、「回帰分析」ボタンをクリックすると図 2 の OLS 推定結果が表示される。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	t値 (df=194)	p値	95.0%下限	95.0%上限	
説明変数	0.8139	0.5542	0.0905	8.9954	0.0000	0.6355	0.9924	BIC
説明変数_1	0.1367	0.0937	0.0899	1.5202	0.1301	−0.0407	0.3141	4.874
切片	40.3300	0.0000	6.7671	5.9598	0.0000	26.9836	53.6765	AIC
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		4.824
N=197	0.593	0.352	0.588	0.345	52.6751	0.0000		

図 2　OLS 推定結果

次に同じ設定で、「HAC」チェックボックスにチェックを入れて（m は空欄）、「回帰分析」ボタンをクリックすると図 3 の HAC 推定結果が表示される。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z値	p値	95.0%下限	95.0%上限	
説明変数	0.8139	0.5542	0.0792	10.2780	0.0000	0.6587	0.9691	BIC
説明変数_1	0.1367	0.0937	0.0939	1.4552	0.1456	−0.0474	0.3208	4.874
切片	40.3300	0.0000	6.6225	6.0898	0.0000	27.3501	53.3099	AIC
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値	m(HAC)	4.824
N=197	0.593	0.352	0.588	0.345	64.1339	0.0000	4	

図 3　HAC 推定結果

ここで、２つの方法は偏回帰係数の値が同じで標準誤差の値が異なる。また、Newey-West 推定量の truncation parameter m の値は自動で与えられて 4 に設定されている。この値は利用者が設定することも可能である。

次に 3 次までのラグを取るとすると、「最大ラグ」を 3 にして、説明変数の 3 次のラグまで選択すると、OLS 推定結果と HAC 推定結果はそれぞれ図 4 と図 5 のように与えられる。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	t値 (df=190)	p値	95.0%下限	95.0%上限	
▶ 説明変数	0.8111	0.5499	0.0917	8.8415	0.0000	0.6301	0.9920	BIC
説明変数_1	0.1338	0.0907	0.0965	1.3864	0.1673	−0.0566	0.3243	4.934
説明変数_2	0.0509	0.0347	0.0966	0.5271	0.5988	−0.1396	0.2414	AIC
説明変数_3	−0.0817	−0.0555	0.0919	−0.8886	0.3753	−0.2631	0.0997	4.850
切片	42.6845	0.0000	8.8005	4.8502	0.0000	25.3253	60.0437	
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		
N=195	0.593	0.352	0.582	0.339	25.8225	0.0000		

図 4　OLS 回帰分析結果

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	z値	p値	95.0%下限	95.0%上限	
▶ 説明変数	0.8111	0.5499	0.0789	10.2789	0.0000	0.6564	0.9657	BIC
説明変数_1	0.1338	0.0907	0.0945	1.4169	0.1565	−0.0513	0.3190	4.934
説明変数_2	0.0509	0.0347	0.0903	0.5634	0.5732	−0.1262	0.2280	AIC
説明変数_3	−0.0817	−0.0555	0.0915	−0.8931	0.3718	−0.2610	0.0976	4.850
切片	42.6845	0.0000	9.5783	4.4564	0.0000	23.9114	61.4576	
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値	m(HAC)	
N=195	0.593	0.352	0.582	0.339	31.9821	0.0000	4	

図 5　HAC 回帰分析結果

一般に OLS 標準誤差と HAC 標準誤差を比較すると HAC 標準誤差の方が大きくなると思いがちであるが、一概にそうとも限らない。

5.　分布ラグモデルの ADL 及び GLS アプローチ

時系列分析において、誤差項に系列相関がある分布ラグモデルでは、OLS 推定量は一致性を持つが、OLS 標準誤差は一致性を持たない。標準誤差の一致性を保証するため、Newey-West による HAC 標準誤差を導入したが、データに強い外生性がある場合は他の手法も考えられる。その中から我々は自己回帰・分布ラグモデル（autoregressive distributed lag model (ADL)）による方法と一般化最小 2 乗（generalized least squares (GLS)）による方法を説明する。

r 期のラグと誤差相関を持つ分布ラグモデルは以下のように表される。

$$y_t = \sum_{k=0}^{r} b_{k+1} x_{t-k} + b_0 + u_t \qquad (1)$$

ここで、説明変数は一般に複数変数であるが、説明の簡単化のために 1 種類としている。このモデルに対する OLS 推定は、誤差項に対して、以下の $AR(p)$ モデルが適用されると仮定する。ここで、$\tilde{u}_t$ には系列相関がないものとする。

$$u_t = \sum_{k=1}^{p} \phi_k u_{t-k} + \tilde{u}_t \qquad (2)$$

(2)の回帰式を(1)に代入し、u_{t-k} に(1)のラグを取った式を代入すると以下になる。

$$
\begin{aligned}
y_t &= \sum_{k=0}^{r} b_{k+1}x_{t-k} + b_0 + \sum_{j=1}^{p}\phi_j u_{t-j} + \tilde{u}_t \\
&= \sum_{k=0}^{r} b_{k+1}x_{t-k} + b_0 + \sum_{j=1}^{p}\phi_j \left(y_{t-j} - \sum_{k=0}^{r} b_{k+1}x_{t-j-k} - b_0 \right) + \tilde{u}_t \\
&= \sum_{j=1}^{p}\phi_j y_{t-j} + \sum_{k=0}^{r} b_{k+1}x_{t-k} - \sum_{k=0}^{r} b_{k+1}\phi_1 x_{t-1-k} - \cdots - \sum_{k=0}^{r} b_{k+1}\phi_p x_{t-p-k} + b_0\left(1 - \sum_{j=1}^{p}\phi_j\right) + \tilde{u}_t \\
&= \sum_{j=1}^{p}\phi_j y_{t-j} + b_1 x_t + (b_2 - b_1\phi_1)x_{t-1} + (b_3 - b_2\phi_1 - b_1\phi_2)x_{t-2} \\
&\qquad + \cdots - b_{r+1}\phi_p x_{t-p-r} + b_0\left(1 - \sum_{j=1}^{p}\phi_j\right) + \tilde{u}_t \\
&= \sum_{j=1}^{p}\phi_j y_{t-j} + \sum_{j=0}^{p+r}\delta_{j+1}x_{t-j} + \delta_0 + \tilde{u}_t
\end{aligned}
$$

ここに、

$b_1 = \delta_1$, $b_2 = \delta_2 + \delta_1\phi_1$ $(\delta_2 = b_2 - b_1\phi_1)$,

$b_3 = \delta_3 + (\delta_2 + \delta_1\phi_1)\phi_1 + \delta_1\phi_2$ $(\delta_3 = b_3 - b_2\phi_1 - b_1\phi_2)$, …

以上の操作により、系列相関のある分布ラグモデルは、系列相関のない以下のような自己回帰・分布ラグモデルになる。

$$
y_t = \sum_{j=1}^{p}\phi_j y_{t-j} + \sum_{j=0}^{p+r}\delta_{j+1}x_{t-j} + \delta_0 + \tilde{u}_t \qquad (3)
$$

しかし、この回帰式は元の分布ラグモデルの回帰式と異なる。この回帰係数ϕ_j, δ_kの推定値から元の回帰係数b_kを推定することはできるが、その標準誤差を推定することはできない。そのためこれを書き直す。即ち、(3)式の右辺のy_{t-1}に(3)式のラグを代入すると以下となる。

$$
\begin{aligned}
y_t &= \sum_{j=1}^{p}\phi_j y_{t-j} + \sum_{j=0}^{p+r}\delta_{j+1}x_{t-j} + \delta_0 + \tilde{u}_t \\
&= \phi_1 y_{t-1} + \sum_{j=2}^{p}\phi_j y_{t-j} + \delta_1 x_t + \sum_{j=1}^{p+r}\delta_{j+1}x_{t-j} + \delta_0 + \tilde{u}_t \\
&= \phi_1\left(\sum_{j=1}^{p}\phi_j y_{t-j-1} + \sum_{j=0}^{p+r}\delta_{j+1}x_{t-j-1} + \delta_0 + \tilde{u}_{t-1} \right) + \sum_{j=2}^{p}\phi_j y_{t-j} \\
&\quad + \delta_1 x_t + \sum_{j=1}^{p+r}\delta_{j+1}x_{t-j} + \delta_0 + \tilde{u}_t \\
&= (\phi_1^2 + \phi_2)y_{t-2} + \sum_{j=2}^{p-1}(\phi_1\phi_j + \phi_{j+1})y_{t-j-1} + \phi_1\phi_p y_{t-p-1} + \delta_1 x_t + (\delta_2 + \phi_1\delta_1)x_{t-1} \\
&\quad + \sum_{j=1}^{p+r-1}(\delta_{j+2} + \phi_1\delta_{j+1})x_{t-j-1} + \phi_1\delta_{p+r+1}x_{t-p-r-1} + \delta_0(1+\phi_1) + \tilde{u}_t + \phi_1\tilde{u}_{t-1}
\end{aligned}
$$

同様にして、上のy_{t-2}にまた(3)式のラグを代入し、それを繰り返すと以下となる。

$$y_t = (\phi_1^2 + \phi_2)\left(\sum_{j=1}^{p} \phi_j y_{t-j-2} + \sum_{j=0}^{p+r} \delta_{j+1} x_{t-j-2} + \delta_0 + \tilde{u}_{t-2} \right) + \sum_{j=2}^{p-1} (\phi_1\phi_j + \phi_{j+1}) y_{t-j-1}$$
$$+ \phi_1\phi_p y_{t-p-1} + \delta_1 x_t + (\delta_2 + \phi_1\delta_1) x_{t-1} + \sum_{j=1}^{p+r-1} (\delta_{j+2} + \phi_1\delta_{j+1}) x_{t-j-1} + \phi_1\delta_{p+r+1} x_{t-p-r-1}$$
$$+ \delta_0(1+\phi_1) + \tilde{u}_t + \phi_1\tilde{u}_{t-1}$$
$$= \delta_1 x_t + (\delta_2 + \phi_1\delta_1) x_{t-1} + \{(\delta_3 + \phi_1\delta_2) + (\phi_1^2 + \phi_2)\delta_1\} x_{t-2} + \cdots + \delta_0\{1 + \phi_1 + (\phi_1^2 + \phi_2)\} + \cdots$$
$$= b_1 x_t + b_2 x_{t-1} + b_3 x_{t-2} + \cdots + \delta_0\{1 + \phi_1 + (\phi_1^2 + \phi_2) + \cdots\}$$

最後の式において、r 期よりも長い部分の係数は r 期ラグモデルが正確に成り立っていれば 0 になる。しかし、現実はそのようにならない。ただ、係数は有意でないものになり、無視することが可能であると考える。このように、ラグ次数を長く取っておき、係数がある位置より有意でなくなることを確かめてラグを決めてもよい。その際、係数の標準誤差は OLS 標準誤差とすることができる。

このようなアプローチは自己回帰・分布ラグモデル（autoregressive distributed lag model (ADL)）による方法と呼ばれている。我々はこれを ADL 推定と呼ぶ。

次に GLS 法について説明する。(1),(2)式を合わせたものは以下のようにも書き換えられる。

$$y_t = \sum_{k=0}^{r} b_{k+1} x_{t-k} + b_0 + \sum_{j=1}^{p} \phi_j u_{t-j} + \tilde{u}_t$$
$$= \sum_{k=0}^{r} b_{k+1} x_{t-k} + b_0 + \sum_{j=1}^{p} \phi_j \left(y_{t-j} - \sum_{k=0}^{r} b_{k+1} x_{t-j-k} - b_0 \right) + \tilde{u}_t$$
$$= \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{k=0}^{r} b_{k+1} \left(x_{t-k} - \sum_{j=1}^{p} \phi_j x_{t-k-j} \right) + b_0 \left(1 - \sum_{j=1}^{p} \phi_j \right) + \tilde{u}_t$$

これは、

$$\tilde{y}_t \equiv y_t - \sum_{j=1}^{p} \phi_j y_{t-j}, \quad \tilde{x}_t \equiv x_t - \sum_{j=1}^{p} \phi_j x_{t-j}$$

と置き換えれば、以下となる。

$$\tilde{y}_t = \sum_{k=0}^{r} b_{k+1} \tilde{x}_{t-k} + \tilde{b}_0 + \tilde{u}_t \qquad (4)$$

この解を求めるには、まず $\tilde{y}_t = y_t$, $\tilde{x}_t = x_t$ として、(4)式を通常の OLS 回帰で求め、その誤差を $\hat{u}_t^{(1)}$ とする。この誤差を目的変数に使って以下のように(2)式を OLS 推定して $\hat{\phi}_k^{(1)}$ を得る。

$$\hat{u}_t^{(1)} = \sum_{k=1}^{p} \hat{\phi}_k^{(1)} \hat{u}_{t-k}^{(1)} + \tilde{u}_t^{(2)}$$

この推定値 $\hat{\phi}_k^{(1)}$ を用いて以下のように $\tilde{y}_t^{(1)}$ や $\tilde{x}_t^{(1)}$ を求める。

$$\tilde{y}_t^{(1)} = y_t - \sum_{j=1}^{p} \hat{\phi}_j^{(1)} y_{t-j}\,, \quad \tilde{x}_t^{(1)} = x_t - \sum_{j=1}^{p} \hat{\phi}_j^{(1)} x_{t-j}$$

これをまた(4)式に代入して、OLS 回帰を用いて誤差$\hat{u}_t^{(2)}$を求める。この誤差を目的変数に使って以下のように$\hat{\phi}_k^{(2)}$を得る。

$$\hat{u}_t^{(2)} = \sum_{k=1}^{p} \hat{\phi}_k^{(2)} \hat{u}_{t-k}^{(2)} + \tilde{u}_t^{(3)}$$

この作業を回帰係数b_{k+1}が収束するまで繰り返す。以上の手順で回帰係数を求める手法を繰り返しコクレンーオーカット（Cochrane-Orcutt）推定法とよぶ。これは、一般化最小２乗（generalized least squares (GLS)）法の特殊な解法である。

ここでは、前節で用いたデータを使って、ADL 法と GLS 法を比較してみよう。モデルは、簡単であるが以下と考える。

$$y_t = b_1 x_t + b_0 + u_t\,, \quad u_t = \phi_1 u_{t-1} + \tilde{u}_t$$

まず単純に OLS 推定した結果を図１に示す。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	t値 (df=192)	p値	95.0%下限	95.0%上限	
▶ 説明変数	0.8640	0.5859	0.0863	10.0171	0.0000	0.6939	1.0342	BIC
切片	46.0167	0.0000	5.6867	8.0920	0.0000	34.8002	57.2332	4.871
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		AIC
N=194	0.586	0.343	0.583	0.340	100.3415	0.0000		4.837

図１　OLS 推定結果

ADL モデルでは、ラグ次数を上げて考えるが、ここでは４次まで考えるとする。「最大ラグ次数」を４として、説明変数のすべての次数のラグを選択する。結果を図２に示す。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	t値 (df=188)	p値	95.0%下限	95.0%上限	
▶ 説明変数	0.8117	0.5504	0.0919	8.8341	0.0000	0.6304	0.9930	BIC
説明変数_1	0.1414	0.0958	0.0968	1.4599	0.1460	−0.0497	0.3324	4.958
説明変数_2	0.0401	0.0271	0.0971	0.4123	0.6806	−0.1516	0.2317	AIC
説明変数_3	−0.1166	−0.0788	0.0976	−1.1947	0.2337	−0.3092	0.0760	4.857
説明変数_4	0.0813	0.0551	0.0921	0.8825	0.3786	−0.1004	0.2630	
切片	39.9050	0.0000	9.6440	4.1378	0.0001	20.8806	58.9293	
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		
N=194	0.597	0.357	0.583	0.340	20.8730	0.0000		

図２　ADL 推定結果

ADL 推定の結論としては図２の０次のところだけを考える。それ以降の説明変数のラグの p 値は p>0.05 である。

次に GLS 推定であるが、ラグ次数は０次、誤差のラグ次数は１次と仮定する。最大ラグ次数を図２のデータ数と合わせるために３とし、その中で０次のラグだけを選択する。また GLS 推定のΦの右のテキストボックスに誤差のラグ次数１を代入し、「GLS 推定」のボタンをクリックすると結果は以下のようになる。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	t値 (df=192)	p値	95.0%下限	95.0%上限	
▶ 説明変数	0.8097	0.5588	0.0867	9.3375	0.0000	0.6387	0.9808	BIC
切片	41.4861	0.0000	4.8009	8.6413	0.0000	32.0168	50.9554	4.793
Φ1	0.1628	0.1620	0.0716	2.2750	0.0240	0.0217	0.3040	AIC
収束:101	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		4.760
N=194	0.559	0.312	0.556	0.309	87.1891	0.0000		

図３　GLS 推定結果（誤差ラグ次数１次）

ここでは、左端下から２行目に「収束:101」とあるが、これは上で述べた繰り返しコクレンーオーカット法の繰り返し数である。ここでは収束になっているが、モデルに問題がある場合は非収束となる場合もある。このプログラムではその判断を 10000 回の繰り返しで収束するか否かとしている。

同様にして、誤差のラグ次数を２次まで取った結果を調べてみる。ここではデータ数を合わせるために、最大ラグを２に変えている。実行結果を図４に示す。

偏回帰係数と検定

目的変数	偏回帰係数	標準化係数	標準誤差	t値 (df=192)	p値	95.0%下限	95.0%上限	
▶ 説明変数	0.8112	0.5596	0.0867	9.3557	0.0000	0.6402	0.9822	BIC
切片	41.6120	0.0000	4.8238	8.6263	0.0000	32.0975	51.1265	4.795
Φ1	0.1588	0.1605	0.0718	2.2100	0.0283	0.0171	0.3005	AIC
Φ2	-0.0001	-0.0001	0.0718	-0.0015	0.9988	-0.1417	0.1415	4.761
収束:839	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値		
N=194	0.560	0.313	0.556	0.310	87.5287	0.0000		

図４　GLS 推定結果（誤差ラグ次数２次）

ϕ_2 が有意に 0 と異ならないことから、誤差の自己回帰の次数は１次が妥当と思われる。以上のことから、図１の単純な OLS 推定の場合に比べて、ADL 推定と GLS 推定の結果はよく似ていることが分かる。

6.　ベクトル自己回帰（VAR）モデルによる多期間の繰り返し予測

VAR（Vector AutoRegressin）モデルは２つ以上の変数を同じ変数のラグによって回帰するモデルである。例えば２次元のベクトルモデルでは、回帰式を以下のように仮定する。

$$\begin{aligned} x_t &= c_1 + a_{11}x_{t-1} + \cdots + a_{1r}x_{t-r} + b_{11}y_{t-1} + \cdots + b_{1s}y_{t-s} + u_{1t} \\ y_t &= c_2 + a_{21}x_{t-1} + \cdots + a_{2r}x_{t-r} + b_{21}y_{t-1} + \cdots + b_{2s}y_{t-s} + u_{2t} \end{aligned} \qquad (1)$$

回帰係数は回帰式単独で与えた場合と同じである。また、回帰係数の誤差も単独で与えた場合と同じとすることが多い[2]。予測において、１期先は以下のようになり、現時点の値だけで計算可能である。ここに $\hat{a}_{\alpha i}, \hat{b}_{\alpha i}$ は、上の回帰分析で求めた推定値を利用している。

$$\hat{x}_{t+1} = \hat{c}_1 + \hat{a}_{11}x_t + \cdots + \hat{a}_{1r}x_{t-r+1} + \hat{b}_{11}y_t + \cdots + \hat{b}_{1s}y_{t-s+1}$$

$$\hat{y}_{t+1} = \hat{c}_2 + \hat{a}_{21}x_t + \cdots + \hat{a}_{2r}x_{t-r+1} + \hat{b}_{21}y_t + \cdots + \hat{b}_{2s}y_{t-s+1}$$

しかし、２期先以上は、繰り返し予測が使われることが多く、上で予測された１期先の値を以下のように利用することになる。

$$\hat{x}_{t+2} = \hat{c}_1 + \hat{a}_{11}\hat{x}_{t+1} + \cdots + \hat{a}_{1r}x_{t-r+2} + \hat{b}_{11}\hat{y}_{t+1} + \cdots + \hat{b}_{1s}y_{t-s+2}$$

$$\hat{y}_{t+2} = \hat{c}_2 + \hat{a}_{21}\hat{x}_{t+1} + \cdots + \hat{a}_{2r}x_{t-r+2} + \hat{b}_{21}\hat{y}_{t+1} + \cdots + \hat{b}_{2s}y_{t-s+2}$$

このため、複数の変数による１つだけの変数の予測は不可能で、ベクトル自己回帰モデルは必須である。さらに先の予測については、この操作を繰り返す。

これを実際に実行するのが分析実行画面下方の以下の部分である。

図１　VAR モデルによる繰り返し予測と検定

例えば、「予測」を 5 期先に変え、「ラグ選択」でラグのない変数 2 つを最初に選び、「VAR 予測値」ボタンをクリックすると、以下のような結果が表示される。

	実測機器	予測値	残差	実測他指標	予測値	残差
94	57	56.0175	0.9825	4	13.5891	-9.5891
95	62	64.0200	-2.0200	7	13.9649	-6.9649
96	78	59.1320	18.8680	23	14.1729	8.8271
97	61	66.7145	-5.7145	21	10.7487	10.2513
98	70	73.9322	-3.9322	12	11.8519	0.1481
99	69	70.9572	-1.9572	15	11.3334	3.6666
100	66	65.5279	0.4721	18	12.5505	5.4495
予測1		70.4712			12.2742	
予測2		70.8359			11.6248	
予測3		68.6319			12.6463	
予測4		70.6247			12.9391	
予測5		71.1632			12.4189	

図２　VAR 予測値

この予測の過去の値は系列全体のデータを使った OLS 予測値で、真の未来の予測ではないが、予測 1～予測 5 の部分は未来の予測である。

次に、予測のラグ次数について、与えたラグ次数が正当かどうかを調べてみる。これには、与えた最大ラグ次数の係数がすべて 0 かどうかの結合仮説検定と、情報量基準を用いた方法がある。これらについて説明する。

簡単のために、VAR（Vector AutoRegression）モデルとして、(1)式で与えられた 2 次元のベクトルモデルを考える。ここでは、変数 x_t のラグ次数を r 、変数 y_t のラグ次数を s（$p=\max(r,s)$）としているが、通常 VAR(p)モデルとして、 $p=r=s$、のように同じ場合を考える。一般には異なっていても問題はないと思われるので、プログラムでも異なっているものとして扱う。回帰式の変数と係数を以下のように定義すると、

$$\mathbf{X}=\begin{pmatrix} x_{p+1} & y_{p+1} \\ \vdots & \vdots \\ x_{T-1} & y_{T-1} \\ x_T & y_T \end{pmatrix},\quad \mathbf{Z}=\begin{pmatrix} 1 & x_p & \cdots & x_{p-r+1} & y_p & \cdots & y_{p-s+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T-2} & \cdots & x_{T-r-1} & y_{T-2} & \cdots & y_{T-s-1} \\ 1 & x_{T-1} & \cdots & x_{T-r} & y_{T-1} & \cdots & y_{T-s} \end{pmatrix},$$

$$\mathbf{U}=\begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{pmatrix}=\begin{pmatrix} u_{1p+1} & u_{2p+1} \\ \vdots & \vdots \\ u_{1T-1} & u_{2T-1} \\ u_{1T} & u_{2T} \end{pmatrix},\quad \mathbf{D}'=\begin{pmatrix}\mathbf{d}_1' \\ \mathbf{d}_2'\end{pmatrix}=\begin{pmatrix} c_1 & a_{11} & \cdots & a_{1r} & b_{11} & \cdots & b_{1s} \\ c_2 & a_{21} & \cdots & a_{2r} & b_{21} & \cdots & b_{2s} \end{pmatrix}$$

(1)式は次のように表示される。

$$\mathbf{X}=\mathbf{ZD}+\mathbf{U}$$

最小２乗法を用いて、パラメータの予測値$\hat{\mathbf{D}}$を以下のように求める。

$$\hat{\mathbf{D}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}=\mathbf{D}+(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}$$

ここで、個々の回帰式のパラメータは以下のように表される。

$$\hat{\mathbf{d}}_\alpha=\mathbf{d}_\alpha+(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}_\alpha$$

ここで誤差の分布を次のように仮定すると、

$$\mathbf{u}_\alpha\sim N(\mathbf{0},\boldsymbol{\Sigma}_{\alpha\alpha})$$

パラメータ$\hat{\mathbf{d}}_\alpha$の分布は、補遺（公式 3）より以下となる。

$$\hat{\mathbf{d}}_\alpha\sim N(\mathbf{d}_\alpha,(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{\alpha\alpha}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1})$$

ここでパラメータをまとめたベクトル

$$\hat{\mathbf{d}}=\begin{pmatrix}\hat{\mathbf{d}}_1\\ \hat{\mathbf{d}}_2\end{pmatrix}$$

の分布について、誤差が説明変数の値によらない均一分散の場合と、説明変数の値による不均一分散の場合に分けてみてみよう。

均一分散誤差の場合

各回帰式で説明変数の値によらず誤差分散が均一の場合、以下の関係式を得る。

$$\hat{\mathbf{d}}_\alpha\sim N(\mathbf{d}_\alpha,(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{\alpha\alpha}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1})=N(\mathbf{d}_\alpha,\sigma_{\alpha\alpha}^2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{I}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1})$$
$$=N(\mathbf{d}_\alpha,\sigma_{\alpha\alpha}^2(\mathbf{Z}'\mathbf{Z})^{-1})$$

これを回帰式間に拡張すると次のようになる。

$$\hat{\mathbf{d}}=\begin{pmatrix}\hat{\mathbf{d}}_1\\ \hat{\mathbf{d}}_2\end{pmatrix}\sim N\left[\begin{pmatrix}\mathbf{d}_1\\ \mathbf{d}_2\end{pmatrix},\begin{pmatrix}\sigma_{11}^2(\mathbf{Z}'\mathbf{Z})^{-1} & \sigma_{12}(\mathbf{Z}'\mathbf{Z})^{-1}\\ \sigma_{12}(\mathbf{Z}'\mathbf{Z})^{-1} & \sigma_{22}^2(\mathbf{Z}'\mathbf{Z})^{-1}\end{pmatrix}\right]\equiv N\left(\mathbf{d},\boldsymbol{\Sigma}_{\mathbf{d}}\right)$$

不均一分散誤差の場合

各回帰式で説明変数の値に依存して誤差分散が不均一の場合、以下の関係を利用する。

$$\hat{\mathbf{d}}_\alpha\sim N(\mathbf{d}_\alpha,(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{\alpha\alpha}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1})$$

これを回帰式間に拡張すると、右辺の一部は次のようになる。

$$(\mathbf{Z}'\boldsymbol{\Sigma}_{\alpha\beta}\mathbf{Z})_{ij}=\sum_{\lambda=1}^{N}\sum_{\mu}^{N}z_{i\lambda}z_{j\mu}u_{\alpha\lambda}u_{\beta\mu}$$

これを利用して以下の関係を得る。

$$\hat{\mathbf{d}}=\begin{pmatrix}\hat{\mathbf{d}}_1\\ \hat{\mathbf{d}}_2\end{pmatrix}\sim N\left[\begin{pmatrix}\mathbf{d}_1\\ \mathbf{d}_2\end{pmatrix},\begin{pmatrix}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{11}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} & (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{12}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\\ (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{12}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} & (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_{22}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\end{pmatrix}\right]$$
$$=N\left(\mathbf{d},\boldsymbol{\Sigma}_{\mathbf{d}}\right)$$

次に結合仮説の検定を考える。（公式 3）を使うと以下の関係を得る。

$$F = (\mathbf{R}\hat{\mathbf{d}} - \mathbf{r})'[\mathbf{R}\boldsymbol{\Sigma}_{\mathbf{d}}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{d}} - \mathbf{r})/q \sim F_{q,\infty} \qquad (2)$$

これは通常の結合仮説の検定統計量を拡張したものである。それぞれの変数のラグの最高次数がすべて 0 であるかどうかの検定をした場合、ベクトルの次数の 2 乗が同時制約の数であり、この場合$q = 2^2 = 4$となる。

最後に、ベクトル自己回帰モデルの BIC と AIC を一般的に示しておく。ベクトルの次元をkとして、誤差共分散$(\boldsymbol{\sigma})_{\alpha\beta}$とラグ次数の和$l_p$を以下のように定義する。

$$(\boldsymbol{\sigma})_{\alpha\beta} = \sigma_{\alpha\beta} = \frac{1}{N}\sum_{t=1}^{N} u_{\alpha t}u_{\beta t} \quad （\alpha, \beta = 1, \cdots, k）, \; l_p = \sum_{i=1}^{k} l_i$$

ここにl_iは変数iのラグ次数である。特に(1)の場合は、$k = 2$，$l_p = r + s$。

これらを用いて、BICとAICは以下となる。

$$BIC = \log|\boldsymbol{\sigma}| + k(l_p + 1)\frac{\log N}{N}$$

$$AIC = \log|\boldsymbol{\sigma}| + k(l_p + 1)\frac{2}{N}$$

設定をした後、図 1 の「VAR ラグ検定」ボタンをクリックすると図 3 が表示される。左は最大ラグ次数 1 次、右は 3 次の場合である。

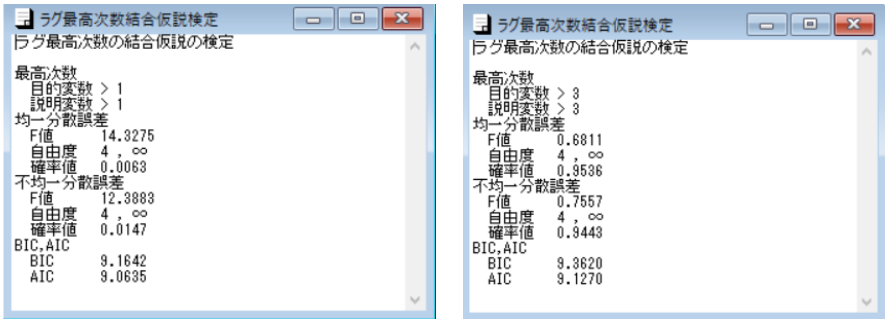


図 3　VAR ラグ検定

これには各変数の最大ラグの係数がすべて 0 かどうかの検定と、BIC と AIC の値が表示されている。BIC と AIC は他のモデルと比べて小さい方が良しとされるが、最大ラグ次数 1 次の方が良いことが分かる。

最後に、この章で述べた理論は分析実行画面の「解説」ボタンをクリックすることで、図 4 のように要約されて表示されることを示しておく。

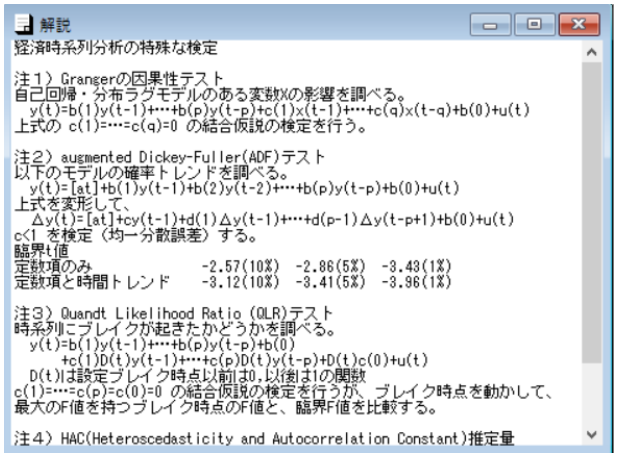


図 4　簡単な解説表示画面

補遺

重複になるが、この章で使う公式を示しておく。

(公式 1) $Cov(\mathbf{Au},\mathbf{Bu}) = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{B}'$

(公式 2) $\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{B}' = \mathbf{0}$ ならば、 $\mathbf{Au}$ と $\mathbf{Bu}$ は独立した分布

(公式 3) $\mathbf{u}(m\times 1) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}})$ のとき、

$\mathbf{d} + \mathbf{Au} \sim N(\mathbf{d}, \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{A}')$

$\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{u} \sim \chi_m^2$

(公式 4) $\mathbf{u} \sim N(0, \mathbf{I}_m)$ で $\mathbf{C}(m\times m)$ がべき等行列（$\mathbf{CC} = \mathbf{C}$）のとき、

$\mathbf{u}'\mathbf{Cu} \sim \chi_r^2$　但し、$rank(\mathbf{C}) = r$

参考文献

[1]　J.H.Stock,M.W.Watson,宮尾龍蔵訳, 入門計量経済学, 共立出版, 2016.

[2]　沖本竜義, 経済・ファイナンスデータの計量時系列分析, 朝倉書店, 2010.