因果関係を可変とする回帰分析(その２)
―シミュレーション検証―

日大生産工　　○篠原正明
情報システム研究所　　篠原健

１．はじめに

Ｎ事象に対して$3^{C(N,2)}$個の因果グラフが存在するが、その中で、結果事象数＝1、原因事象数＝Ｎ－１ の集中型星状の因果グラフのみに注目する。また、集中型星状因果グラフにおいては､Ｎ－１個の原因事象間に因果関係は無いとする。Ｎ個の個々の集中型星状因果グラフに対して、既存の最小２乗型回帰分析法を適用すると共に各種の適合度指標を計算することができる。そこで、ある事象を結果、その他の事象を原因とした線形モデルを仮定し、結果変量の観測に際して正規分布に従う観測誤差が発生するとしたデータ群を発生させ、そのデータに対して集中型星状因果グラフに限定した可変因果回帰分析を適用し、果たして、仮定した因果関係が可変因果回帰分析により検証されうるかのシミュレーション実験を以下に行う。

２．シミュレーション実験データ設定

３事象から構成され、各事象に変量１、変量２、変量３が対応する系を想定する。図１に示すように、事象２,３が事象１の原因となる因果関係を想定する。すなわち、変量１を被説明変数y_1 変量２、３を説明変数χ_2 、χ_3 とする。

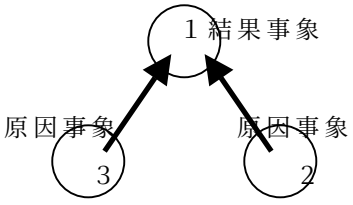

図１　想定する因果関係

y_1 とχ_2 、χ_3 の間に線形モデル式を仮定し、さらに、各変数は適当な確率分布に従い、平均値は１になるように正規化されているものとする。すなわち、線形モデル式は(１)式で与える。

$$y_1 = P_2 \chi_2 + P_3 \chi_3 + e \qquad (1)$$

ここで、$P_2 + P_3 = 1$、ｅは平均０の正規分布に従う観測誤差である。

データ設定１：$P_2 = 0.5$、$P_3 = 0.5$とし、χ_2 、χ_3 は(0,2)間の一様乱数、観測誤差ｅは平均値 0、標準偏差SDの正規分布の乱数で与え、(1)式に従って観測値y_1 を 10 標本分生成する。

データ設定２：$P_2 = -0.5$、$P_3 = 1.5$とし、χ_2 、χ_3 は(0,2)間の一様乱数、観測誤差ｅは平均値 0、標準偏差SDの正規分布の乱数で与え、(1)式に従って観測値y_1 を 10 標本分生成する。

Regression Analysis with Variable Causal Relationship － Part ２－

Masaaki SHINOHARA　and　Ken SHINOHARA

３．シミュレーション検証

[検証例 1:設定１、SD=0.1、標本数＝10]

表 1.1　標本データ

y1 標本値	x2 標本値	x3 標本値
0.799927	0.272326	1.180682
1.522283	0.940516	1.900295
1.159383	1.225517	1.223185
0.785036	0.228133	1.318295
1.255578	0.921252	1.401656
0.015587	0.038455	0.089338
0.780438	0.09085	1.808694
0.991108	0.663403	1.60357
1.535766	1.700644	1.085293
0.753092	0.450312	1.087784

表 1.2　y_1 を被説明変数とした場合(モデル式：$y_1 = S_2\ \chi_2 + S_3\ \chi_3$，定数項なし)の推定結果

R2	0.959347	パラメタ	推定値
補正 R2	0.954266	S2	0.600554
LogL	10.40092	S3	0.449255
AIC	-16.8018		
F 値	598.2733		

表 1.3　χ_2 を被説明変数とした場合(モデル式：$\chi_2 = q_3\ \chi_3 + q_1\ y_1$)

R2	0.928505	パラメタ	推定値
補正 R2	0.919568	Q3	-0.66351
LogL	5.659799	Q1	1.550104
AIC	-7.3196		
F 値	142.3444		

表 1.4　χ_3 を被説明変数とした場合(モデル式：$\chi_3 = r_1\ y_1 + r_3\ \chi_3$)

R2	0.849255	パラメタ	推定値
補正 R2	0.830412	R1	2.120461
LogL	2.641936	R2	-1.21332
AIC	-1.28387		
F 値	209.4028		

なお、回帰分析（最小二乗法）は文献〔1〕の Excel CD-ROM ソフトを使用した。

[検証例２:設定１、SD=0.2、標本数＝10]

表２.1　標本データ

y1 標本値	x2 標本値	x3 標本値
1.077328	0.61366	1.02266
1.411968	1.174106	1.966687
1.160696	1.320264	1.257742
0.693732	0.737624	0.711238
1.142891	0.629331	1.487681
0.689625	1.084647	0.940201
0.3428	0.741497	0.131938
0.319057	0.66318	0.184995
1.692534	1.509608	1.258972
0.821993	0.278817	0.784737

表２.2　y_1 を被説明変数とした場合(モデル式：$y_1 = S_2\ \chi_2 + S_3\ \chi_3$，定数項なし)の推定結果

R2	0.783765	パラメタ	推定値
補正 R2	0.756736	S2	0.363206
LogL	2.105727	S3	0.618578
AIC	-0.21145		
F 値	105.5531		

表２.3　χ_2 を被説明変数とした場合（モデル式：$\chi_2 = q_3\ \chi_3 + q_1\ y_1$ ）

R2	0.150143	パラメタ	推定値
補正 R2	0.043911	Q3	-0.17271
LogL	-3.17505	Q1	1.044324
AIC	10.35009		
F 値	28.44248		

表２.4　χ_3 を被説明変数とした場合（モデル式：$\chi_3 = r_1\ y_1 + r_3\ \chi_3$ ）

R2	0.750063	パラメタ	推定値
補正 R2	0.71882	R1	1.147086
LogL	-0.98205	R2	-0.11139
AIC	5.964106		
F 値	65.33296		

[検証例３:設定１、SD=0.3、標本数＝10]

表３.1　標本データ

y1 標本値	x2 標本値	x3 標本値
0.708919	1.177332	0.999876
0.422585	0.111447	0.945352
1.428232	1.791184	1.639336
0.572781	0.68221	0.405321
0.809303	0.141181	0.963159
0.810158	0.625822	0.62265
1.387277	1.987092	1.652146
0.947053	1.103216	1.109884
1.120358	1.551023	1.016003
1.018029	0.89314	1.953137
9.224696	10.06365	11.30686

表３.2　y_1 を被説明変数とした場合(モデル式：$y_1 = S_2 \chi_2 + S_3 \chi_3$，定数項なし)の推定結果

R2	0.687496	パラメタ	推定値
補正 R2	0.648433	S2	0.370417
LogL	3.322923	S3	0.448997
AIC	-2.64585		
F 値	121.796		

表３.3　χ_2 を被説明変数とした場合（モデル式：$\chi_2 = q_3 \chi_3 + q_1 y_1$）

R2	0.672288	パラメタ	推定値
補正 R2	0.631324	Q3	-0.29039
LogL	-3.67391	Q1	1.501165
AIC	11.34782		
F 値	41.39019		

表３.4　χ_3 を被説明変数とした場合（モデル式：$\chi_3 = r_1 y_1 + r_3 \chi_3$）

R2	0.524086	パラメタ	推定値
補正 R2	0.464596	R1	1.478633
LogL	-2.63636	R2	-0.23598
AIC	9.272727		
F 値	55.95467		

[検証例４:設定２、SD=0.1、標本数＝10]

表４.1　標本データ

y1 標本値	x2 標本値	x3 標本値
1.14459	1.678746	1.386408
1.648651	0.006787	1.081236
1.361292	1.744129	1.515148
0.688341	0.559845	0.661683
1.682724	0.879641	1.377965
1.223719	1.580986	1.28553
1.715482	1.851884	1.773387
1.422716	1.975886	1.601328
-0.42966	1.693294	0.26461
-0.54379	1.830748	0.196023

表４.2　y_1 を被説明変数とした場合(モデル式：$y_1 = S_2 \chi_2 + S_3 \chi_3$，定数項なし)の推定結果

R2	0.995507	パラメタ	推定値
補正 R2	0.994945	S2	-0.48649
LogL	15.14461	S3	1.486744
AIC	-26.2892		
F 値	2274.515		

表４.3　χ_2 を被説明変数とした場合（モデル式：$\chi_2 = q_3 \chi_3 + q_1 y_1$）

R2	0.970258	パラメタ	推定値
補正 R2	0.96654	Q3	3.019515
LogL	8.030834	Q1	-2.01822
AIC	-12.0617		
F 値	779.0698		

表４.4　χ_3 を被説明変数とした場合（モデル式：$\chi_3 = r_1 y_1 + r_3 \chi_3$）

R2	0.995399	パラメタ	推定値
補正 R2	0.994824	R1	0.670600
LogL	19.12546	R2	0.328299
AIC	-34.2509		
F 値	4753.882		

[検証例５:設定２、SD=0.2、標本数＝10]

表５.1　標本データ

y1 標本値	x2 標本値	x3 標本値
-0.44576	0.60173	0.0169
-0.25675	0.96138	0.135491
2.339178	0.943019	1.811113
2.204062	1.773296	1.966714
1.054625	1.353308	1.260157
0.395483	1.426361	0.713851
0.297038	1.033712	0.505325
1.668765	1.63253	1.496115
0.751346	0.516472	0.689999
0.755221	0.293071	0.672047

表５.2　y_1 を被説明変数とした場合(モデル式：$y_1 = S_2 \chi_2 + S_3 \chi_3$，定数項なし)の推定結果

R2	0.983917	パラメタ	推定値
補正 R2	0.981906	S2	-0.50811
LogL	7.457907	S3	1.552579
AIC	-10.9158		
F 値	477.8575		

表５.3　χ_2 を被説明変数とした場合（モデル式：$\chi_2 = q_3 \chi_3 + q_1 y_1$）

R2	0.802521	パラメタ	推定値
補正 R2	0.777836	Q3	2.71563
LogL	1.551016	Q1	-1.65585
AIC	0.897967		
F 値	119.6535		

表５.4　χ_3 を被説明変数とした場合（モデル式：$\chi_3 = r_1 y_1 + r_3 \chi_3$ ）

R2	0.986935	パラメタ	推定値
補正 R2	0.985302	R1	0.630761
LogL	11.96163	R2	0.338545
AIC	-19.9233		
F 値	944.0377		

[検証例６:設定２、SD=0.3、標本数＝10]

表６.1　標本データ

y1 標本値	x2 標本値	x3 標本値
2.04141	1.102386	1.738426
1.626164	1.506913	1.819552
2.071978	0.181323	1.198183
0.814243	1.705592	1.041936
1.671339	1.791769	1.387789
0.486571	1.69991	1.237314
1.297347	1.156789	1.248405
0.886344	1.280162	0.896144
0.037247	1.621847	0.578406
3.484119	0.030884	1.732367

表６.2　y_1 を被説明変数とした場合(モデル式：$y_1 = S_2 \chi_2 + S_3 \chi_3$ ，定数項なし)の推定結果

R2	0.869114	パラメタ	推定値
補正 R2	0.852754	S2	-0.73991
LogL	-3.28139	S3	1.78812
AIC	10.56277		
F 値	100.2245		

表６.3　χ_2 を被説明変数とした場合（モデル式：$\chi_2 = q_3 \chi_3 + q_1 y_1$ ）

R2	0.569027	パラメタ	推定値
補正 R2	0.515156	Q3	2.00569
LogL	-4.80212	Q1	-1.00292
AIC	13.60425		
F 値	43.42185		

表６.4　χ_3 を被説明変数とした場合（モデル式：$\chi_3 = r_1 y_1 + r_3 \chi_3$ ）

R2	0.76592	パラメタ	推定値
補正 R2	0.73666	R1	0.527623
LogL	2.821309	R2	0.436617
AIC	-1.64262		
F 値	212.3033		

４．考察

（4.1）データ設定１の検証例においては全ての場合で、想定したモデル式　$y_1 = P_2 \chi_2 + P_3 \chi_3$ に対応した　y_1 を被説明変数とした因果関係の場合に、決定決数 R2 が大きく、AIC が小さいことが確認できた。

（4.2）データ設定２の検証例では、SD＝0.1, 0.2, 0.3 の場合に χ_3 を被説明変数とすると AIC が小さくなっている。決定係数 R2 では、SD＝0.3 の場合に、y_1 を被説明変数とすると大きいと言える。その他の場合は、微妙である。

５．おわりに

相関大が必ずしも因果関係を示すわけではないが、本来の因果モデルに従って、小さな観測誤差のもとで、相当数の標本に対する観測値データが得られるならば、想定した因果関係に対して回帰分析によりパラメタ推定する場合が最も適合度が増すと考えられないだろうか？　この考えの妥当性を検証するために、シミュレーション検証実験を行った。データ設定１については、妥当性が示されたと考えられる。

集中型星状因果グラフ以外のトポロジの場合、因果関係と相関関係の数量的関連性、様々な因果グラフのトポロジに共通した適合性指標の選択、説明変数の数が多い場合の様々なデータ設定でのシミュレーション検証、等々は今後の課題である。

参考文献

〔１〕縄田和満：Excel 統計解析ボックスによるデータ解析、朝倉書店(2001)