<大蔵省財政金融研究所「フィナンシャル・レビュー」December－1989>

〔ノート〕

回帰分析で通常使用される決定係数の再検討

松　本　和　幸*
酒　井　英　昭**

Ⅰ．はじめに

経済分析において基本的なtoolとなっている決定係数に様々な問題があることは従来から指摘されてきた。しかし，回帰分析において，一応の「当てはまりの良さ」をみる場合には，決定係数が最も頻繁に参照されてきたように思われる。その理由は，①直観的にみても回帰式のまわりのデータの散らばり具合という意味で分かり易いこと，②回帰式のまわりの散らばり方が一定であれば，データ分布の座標軸に対する位置関係によっては決定係数がさほど変化しないものと考えていること，の２点によるものである。もちろん，決定係数の定義からも明らかなように，回帰平面と座標軸とがなす角度により，決定係数がある程度変化することは考慮されてきたと思われるが。

ところが，実際に簡単なシミュレーションを行って，データ分布の座標軸に対する位置変化に伴う決定係数の変化度合いを調べてみると，それが，相当大きく変化することが明らかとなった。そして，決定係数が，回帰平面のまわりの散らばり具合を，必ずしも正確には表していないと思われるようなケースも少なくないことがわかった（図１参照）。

決定係数（R^2）の定義からみて，そのようなことは当然とも言えるが，実際上は回帰平面のまわりでの分布の当てはまりの良さを測る参考指標として決定係数（R^2）が参照されることが一般的である。

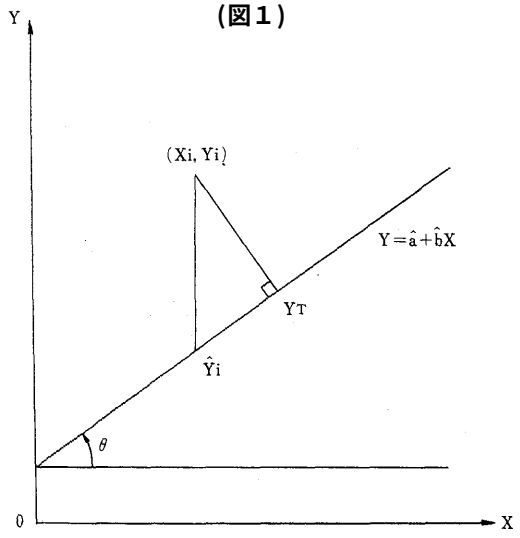
そこで本稿では，通常の決定係数（R^2）に加えて，同時に参照すべきと思われる，別の決定係数（M^2）を導入する。M^2はデータ分布と座標軸との位置関係には不変で，しかも，データ分布の形状にのみ依存するようなものである。

そのことについての定量的把握に加えて，本稿では，決定係数の大きさをデータの分布状態で直観的にとらえることも計測の目的とした。具体的には，データの分布を楕円，楕円体，ないし「ｎ次元楕円体」で近似した場合に，それらの分布図形の「細長さ」ないし「偏平度」が，どれくらいの大きさの決定係数と対応するかを調べた。たとえば，図２のようにデータがＸ－Ｙ平面上で楕円状に分布している場合に，あるｎに対して，縦横の比率が１：ｎの楕円に対応する決定係数はどの程度の大きさなのかを調べることは，実証家にとって興味深いものと思われるからである。

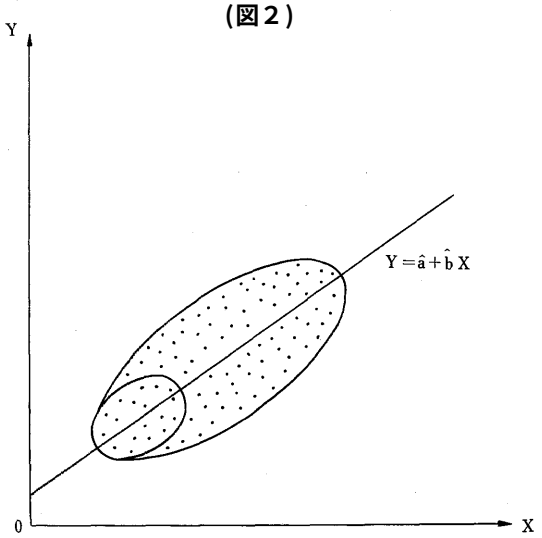
このような点に関する考察は，理工学的観点からは直交回帰等の形でいくつかの研究が行われてきたが，最近では，松本(1989)において若干異なった視点からの検討が加えら

＊　　大蔵省財政金融研究所主任研究官
＊＊　京都大学工学部理工学科助教授

(図1)



(図2)



れた。ただ，そこではM^2については２次元平面のケースしか考察されていない。そこで，本稿では，M^2を一般のｎ次元に拡張するとともに，さらに，そのように般化して定義されたM^2の特性についてのモンテカルロ・シミュレーションを追加した。

【決定係数とは何か】

問題の所在を直観的に示すために，決定係数の定義について改めて少し整理しよう。

２次元の場合（図１参照)，回帰直線の回りでの散らばり度を示す決定係数（R^2）とは，各点Y_iに関し，それのＹ軸に沿って回帰直線に降ろした点（Y_i）の分散が，Y_i自体の分散に比較してどれぐらい大きいかを表している。すなわち，

$$R^2=\frac{\Sigma(\hat{Y}_i-\bar{Y})^2}{\Sigma(Y_i-\bar{Y})^2} \quad (\hat{Y}_i \text{はY}_i\text{の推定値})$$

$$=1-\frac{\Sigma(Y_i-\hat{Y}_i)^2}{\Sigma(Y_i-\bar{Y})^2}$$

である。

つまり，下の式の分子はY_iのその推定値からの分散であって回帰直線に降ろした垂線の足(Y_T)についての分散ではないため，図１のθで示した角度によってR^2が変化し得るのである。また，ｎ次元に分布するデータについては，ｎ－１次元超平面(hyperplane)を回帰平面として同様に考えればよい。

Ⅱ．決定係数のシミュレーション

以下で紹介するシミュレーションは，変数が２変数と３変数のケースで，それぞれ,

$Y=\hat{a}+\hat{b}X+\varepsilon$ (εは誤差項)

$Y=\hat{a}+\hat{b}_1X_1+\hat{b}_2X_2+\varepsilon$ (εは誤差項)

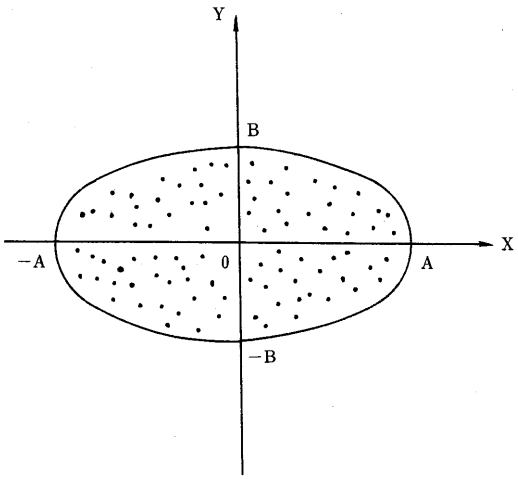
の形で回帰される。

ここで，筆者は，データが楕円（２変数の場合）ないし楕円体（３変数の場合）として分布する場合に，分布の「細長さ」および分布の原点に対する位置関係により，決定係数がどう変化するかを計測した。

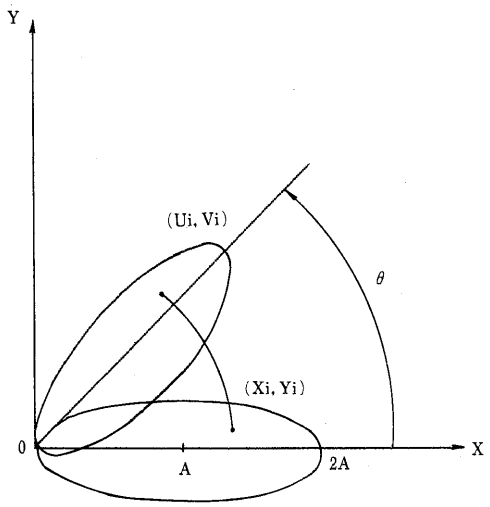
１．２変数の場合

まず，図３のような，原点を中心とする楕円を考える。図３のとおり，Ｘは－ＡからＡまで変化する。Ｙは－１から１まで変化する。なお，Ｙについても，たとえば，－ＢからＢまで変化するとしてもよいが，相似性を考慮すれば，Ｂについては－１から１までの変化

(図３)



(図４)



を考えれば十分であることがわかる。

そこで，Ｘについては－ＡからＡまでの，Ｙについては－１から１までの一様乱数を発生させ，$X^2/A^2+Y^2 \leqq 1$ならば楕円内の点とし，そうでないならば，次の乱数を発生させる。こうして，楕円内に十分多くのランダムな点を求める。

次に，図４のように，点（Ｘ，Ｙ）をＸ方向にＡだけ平行移動した後，正の方向にθラディアン回転したものを，点（Ｕ，Ｖ）として，Ｕ，Ｖについて回帰分析して決定係数を求めることとする。ここで，

$U=(X+A)\cos\theta-Y\sin\theta$

$V=(X+A)\sin\theta+Y\cos\theta$

である。

周知のとおり，決定係数は，

$$R^2=\frac{\Sigma(\hat{V}_i-\bar{V})^2}{\Sigma(V_i-\bar{V})^2}$$

ここに，　$\bar{V}=A\sin\theta$

$\hat{V}_i=U_i\tan\theta$

である。

また，自由度調整済決定係数（$\bar{R}^2$）は，

$$\bar{R}^2=1-(1-R^2)\frac{n-1}{n-2}$$

として求められる。

２．3変数の場合

さて，３変数の場合には，図６のような，原点を中心とする楕円体を考える。相似性まで考慮して，Ｘが－ＡからＡ，Ｙが－ＢからＢ，Ｚが－ＣからＣ（Ｃ＝１）まで変化するものと考える。そうすれば，

$X^2/A^2+Y^2/B^2+Z^2 \leqq 1$

が楕円体の中にあるための条件である。さらに，ここでは簡単化のためにＢ＝１の場合のみ計測する。

そこで，Ｘについては－ＡからＡ，Ｙは－１から１，Ｚは－１から１の一様乱数を発生させ，楕円体の中にあるランダムな点を求める。

次に，図６のように，点（Ｘ，Ｙ，Ｚ）をＸ方向にＡだけ平行移動した後Ｘ－Ｙ平面に平行にθラデイアン，Ｘ－Ｙ平面に垂直にτラディアン回転した点を，（Ｕ，Ｖ，Ｗ）とすれば，

$$U=\{(X+A)\cos\tau-Z\sin\tau\}\cos\theta-Y\sin\theta$$

$$V=\{(X+A)\cos\tau-Z\sin\tau\}\sin\theta+Y\cos\theta$$

$$W=Z\cos\tau+(X+A)\sin\tau$$

となるが，決定係数という点に絞れば，θの変化はR^2には影響しないことから，$\theta\equiv 0$つまり，Ｘ－Ｙ平面に平行な回転は考えず，

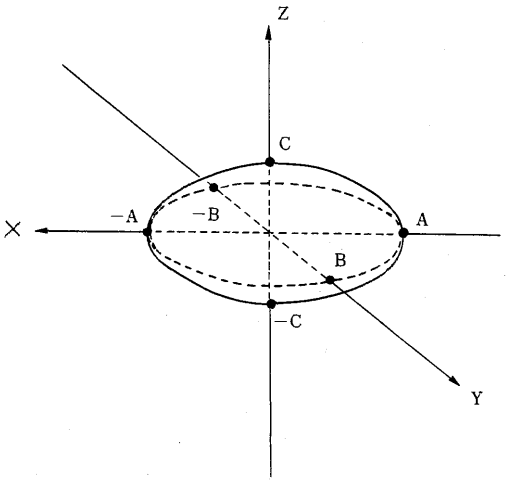
$U=(X+A)\cos\tau-Z\sin\tau$

$V=Y$

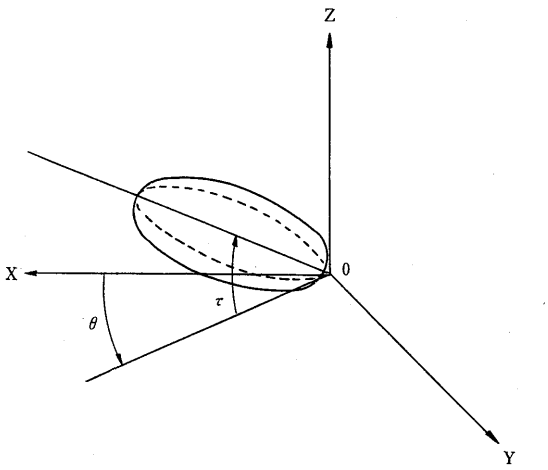
$W=Z\cos\tau+(X+A)\sin\tau$

で計測すればよい。

(図5)



(図6)



この場合（回帰式$W=\hat{b}_0+\hat{b}_1U+\hat{b}_2V$），決定係数は，

$$R^2=\frac{\Sigma(\hat{W}_i-\bar{W})^2}{\Sigma(W_i-\bar{W})^2}=\frac{\Sigma\hat{w}_i^2}{\Sigma w_i^2}=\frac{\hat{b}_1\Sigma w_i u_i+\hat{b}_2\Sigma w_i v_i}{\Sigma w_i^2}$$

ここに

$$\hat{b}_1=\frac{(\Sigma u_i w_i)(\Sigma v_i^2)-(\Sigma v_i w_i)(\Sigma u_i v_i)}{(\Sigma u_i^2)(\Sigma v_i^2)-(\Sigma u_i v_i)^2}$$

$$\hat{b}_2=\frac{(\Sigma v_i w_i)(\Sigma u_i^2)-(\Sigma u_i w_i)(\Sigma u_i v_i)}{(\Sigma u_i^2)(\Sigma v_i^2)-(\Sigma u_i v_i)^2}$$

小文字の変数は大文字の変数の平均からの偏差である

Ⅲ. 計測結果

表１は２次元の場合について，データ数を20,000個で，表２は３次元の場合について，データ数を25,000個でモンテカルロ・シミュレーションした結果である。

まず，２次元である楕円の場合からみると，Ａ＝１すなわち円状に分布しているときの決定係数は，θの値にかかわらず，ほぼ０であることがわかる。また，Ａがいずれの値であってもθ＝45°の場合に決定係数が最大となる。

各決定係数に対応するＡの値を，θ＝45°のケースでみると，R^2＝0.6はＡ＝３ぐらいに相当し，R^2＝0.9はＡ＝６，R^2＝0.99はＡ＝20に相当する。

３次元である楕円体の場合にも，τ＝45°で決定係数が最大となる。その他のR^2とＡとの関係については表１，２を参照されたい。

さて，次にθやτに注目してR^2をみてみよう。たとえば，２次元の楕円体でＡ＝７とすると，θ＝45°ならばR^2は0.92にも達するが，同じ形状に分布していてもθ＝５°であれば，R^2＝0.26，θ：10°であれば，R^2＝0.58に過ぎない。

そして，恐らく，機械的に回帰分析による説明変数選択を行っているようなときには，それらが同じ程度の説明力を持っていることに気付かず，R^2＝0.92なら可とし，R^2＝0.26や0.58なら不可とする可能性が高い。このように，説明変数の説明力はθやτによって大きく異なってみえる。３次元の場合についても同様であることは次の表２からもわかる。

ただし，座標軸と回帰平面が成す角度が大体20°以上あり，しかも決定係数R^2が大体0.9以上あるような場合には，回転によってR^2

表１　２次元データ（楕円）のシミュレーション

（θの単位：度）

θ A	1	2	3	4	5
5	0.0000	0.0142	0.0527	0.0924	0.1473
15	0.0000	0.1163	0.3090	0.4613	0.5905
25	0.0000	0.2399	0.5104	0.6685	0.7722
35	0.0000	0.3246	0.6099	0.7524	0.8362
45	0.0000	0.3544	0.6384	0.7751	0.8526
55	0.0000	0.3284	0.6085	0.7529	0.8364
65	0.0000	0.2474	0.5072	0.6698	0.7727
75	0.0000	0.1251	0.3034	0.4644	0.5919
85	0.0000	0.0185	0.0485	0.0959	0.1497
θ A	6	7	8	9	10
5	0.2069	0.2625	0.3207	0.3700	0.4311
15	0.6833	0.7490	0.7960	0.8311	0.8622
25	0.8350	0.8754	0.9015	0.9205	0.9362
35	0.8839	0.9137	0.9323	0.9458	0.9566
45	0.8961	0.9231	0.9397	0.9518	0.9615
55	0.8839	0.9138	0.9322	0.9459	0.9566
65	0.8349	0.8759	0.9014	0.9208	0.9361
75	0.6828	0.7509	0.7955	0.8323	0.8618
85	0.2057	0.2689	0.3187	0.3769	0.4282
θ A	11	12	13	14	15
5	0.4716	0.5179	0.5601	0.5957	0.6235
15	0.8804	0.8978	0.9135	0.9239	0.9320
25	0.9452	0.9536	0.9612	0.9661	0.9698
35	0.9629	0.9686	0.9739	0.9772	0.9797
45	0.9671	0.9722	0.9769	0.9798	0.9821
55	0.9629	0.9686	0.9739	0.9771	0.9797
65	0.9451	0.9533	0.9612	0.9660	0.9698
75	0.8799	0.8967	0.9135	0.9235	0.9319
85	0.4677	0.5080	0.5602	0.5914	0.6223
θ A	16	17	18	19	20
5	0.6533	0.6781	0.7084	0.7259	0.7503
15	0.9400	0.9463	0.9525	0.9566	0.9615
25	0.9735	0.9764	0.9792	0.9810	0.9832
35	0.9822	0.9842	0.9861	0.9873	0.9888
45	0.9843	0.9861	0.9877	0.9888	0.9901
55	0.9823	0.9842	0.9860	0.9873	0.9888
65	0.9735	0.9765	0.9791	0.9811	0.9832
75	0.9400	0.9466	0.9523	0.9567	0.9615
85	0.6545	0.6833	0.7057	0.7275	0.7512
θ A	21	22	23	24	25
5	0.7704	0.7866	0.8000	0.8115	0.8275
15	0.9653	0.9684	0.9707	0.9729	0.9754
25	0.9849	0.9863	0.9873	0.9883	0.9894
35	0.9899	0.9909	0.9915	0.9922	0.9929
45	0.9911	0.9919	0.9925	0.9931	0.9937
55	0.9899	0.9909	0.9915	0.9922	0.9929
65	0.9849	0.9863	0.9873	0.9883	0.9894
75	0.9653	0.9685	0.9707	0.9729	0.9754
85	0.7700	0.7884	0.7995	0.8132	0.8268
θ A	26	27	28	29	30
5	0.8342	0.8447	0.8564	0.8638	0.8695
15	0.9765	0.9782	0.9802	0.9813	0.9822
25	0.9899	0.9906	0.9915	0.9919	0.9923
35	0.9932	0.9937	0.9943	0.9946	0.9949
45	0.9940	0.9944	0.9950	0.9952	0.9955
55	0.9932	0.9937	0.9943	0.9946	0.9949
65	0.9899	0.9905	0.9915	0.9919	0.9923
75	0.9765	0.9781	0.9802	0.9812	0.9822
85	0.8333	0.8424	0.8564	0.8628	0.8692

表２　３次元のデータ（楕円体）のシミュレーション（その１）

（τの単位：度）

τ A	1	2	3	4	5
5	0.0001	0.0175	0.0526	0.0933	0.1435
10	0.0001	0.0639	0.1749	0.2883	0.3975
15	0.0001	0.1268	0.3107	0.4653	0.5863
20	0.0001	0.1933	0.4262	0.5906	0.7014
25	0.0001	0.2537	0.5129	0.6724	0.7697
30	0.0001	0.3027	0.5734	0.7243	0.8105
35	0.0001	0.3381	0.6125	0.7559	0.8344
40	0.0001	0.3592	0.6343	0.7729	0.8471
45	0.0000	0.3662	0.6412	0.7784	0.8511
50	0.0000	0.3590	0.6339	0.7733	0.8473
55	0.0000	0.3375	0.6116	0.7566	0.8349
60	0.0000	0.3019	0.5719	0.7255	0.8114
65	0.0000	0.2527	0.5108	0.6744	0.7712
70	0.0000	0.1921	0.4232	0.5936	0.7039
75	0.0000	0.1256	0.3069	0.4699	0.5905
80	0.0000	0.0628	0.1709	0.2943	0.4041
85	0.0000	0.0168	0.0497	0.0985	0.1507
τ A	6	7	8	9	10
5	0.2053	0.2602	0.3207	0.3708	0.4245
10	0.5020	0.5778	0.6461	0.6958	0.7411
15	0.6834	0.7454	0.7958	0.8302	0.8595
20	0.7813	0.8288	0.8655	0.8899	0.9100
25	0.8354	0.8731	0.9014	0.9199	0.9349
30	0.8665	0.8979	0.9211	0.9362	0.9483
35	0.8843	0.9119	0.9322	0.9453	0.9558
40	0.8936	0.9192	0.9379	0.9499	0.9596
45	0.8965	0.9215	0.9396	0.9514	0.9608
50	0.8937	0.9192	0.9378	0.9499	0.9596
55	0.8845	0.9120	0.9321	0.9453	0.9558
60	0.8668	0.8980	0.9210	0.9362	0.9484
65	0.8359	0.8733	0.9012	0.9199	0.9349
70	0.7821	0.8292	0.8652	0.8899	0.9100
75	0.6849	0.7461	0.7952	0.8303	0.8596
80	0.5048	0.5792	0.6447	0.6961	0.7413
85	0.2091	0.2625	0.3180	0.3713	0.4250
τ A	11	12	13	14	15
5	0.4799	0.5101	0.5543	0.5933	0.6238
10	0.7807	0.8017	0.8290	0.8501	0.8653
15	0.8836	0.8963	0.9121	0.9238	0.9321
20	0.9261	0.9346	0.9450	0.9525	0.9577
25	0.9468	0.9530	0.9606	0.9661	0.9699
30	0.9579	0.9629	0.9690	0.9733	0.9763
35	0.9640	0.9683	0.9735	0.9772	0.9798
40	0.9671	0.9710	0.9758	0.9792	0.9815
45	0.9680	0.9719	0.9766	0.9798	0.9821
50	0.9671	0.9710	0.9759	0.9792	0.9815
55	0.9639	0.9683	0.9735	0.9772	0.9798
60	0.9578	0.9629	0.9690	0.9733	0.9763
65	0.9466	0.9530	0.9608	0.9661	0.9698
70	0.9258	0.9346	0.9452	0.9526	0.9577
75	0.8830	0.8964	0.9126	0.9240	0.9320
80	0.7789	0.8019	0.8304	0.8505	0.8650
85	0.4745	0.5108	0.5592	0.5905	0.6225

τ^A	16	17	18	19	20
5	0.6540	0.6808	0.7046	0.7321	0.7485
10	0.8800	0.8926	0.9027	0.9133	0.9201
15	0.9400	0.9468	0.9520	0.9574	0.9609
20	0.9628	0.9671	0.9704	0.9737	0.9760
25	0.9735	0.9766	0.9790	0.9814	0.9830
30	0.9792	0.9816	0.9835	0.9853	0.9866
35	0.9823	0.9844	0.9860	0.9875	0.9886
40	0.9838	0.9857	0.9872	0.9886	0.9896
45	0.9843	0.9862	0.9876	0.9890	0.9899
50	0.9838	0.9858	0.9872	0.9886	0.9896
55	0.9823	0.9844	0.9860	0.9875	0.9886
60	0.9792	0.9817	0.9835	0.9853	0.9866
65	0.9736	0.9767	0.9790	0.9813	0.9829
70	0.9629	0.9672	0.9705	0.9736	0.9759
75	0.9401	0.9470	0.9522	0.9571	0.9608
80	0.8801	0.8934	0.9032	0.9124	0.9198
85	0.6543	0.6845	0.7068	0.7273	0.7469

τ^A	21	22	23	24	25
5	0.7691	0.7821	0.7987	0.8104	0.8263
10	0.9281	0.9330	0.9391	0.9433	0.9485
15	0.9650	0.9675	0.9705	0.9727	0.9752
20	0.9785	0.9801	0.9820	0.9833	0.9848
25	0.9848	0.9859	0.9872	0.9882	0.9893
30	0.9880	0.9889	0.9900	0.9907	0.9916
35	0.9898	0.9906	0.9915	0.9921	0.9928
40	0.9907	0.9914	0.9922	0.9928	0.9935
45	0.9910	0.9917	0.9925	0.9930	0.9937
50	0.9907	0.9914	0.9922	0.9928	0.9935
55	0.9898	0.9906	0.9915	0.9921	0.9928
60	0.9880	0.9889	0.9900	0.9907	0.9916
65	0.9848	0.9859	0.9873	0.9882	0.9893
70	0.9785	0.9801	0.9820	0.9833	0.9848
75	0.9649	0.9675	0.9706	0.9727	0.9751
80	0.9279	0.9331	0.9392	0.9435	0.9483
85	0.7682	0.7824	0.7996	0.8120	0.8250

τ^A	26	27	28	29	30
5	0.8354	0.8454	0.8547	0.8621	0.8707
10	0.9518	0.9551	0.9580	0.9605	0.9632
15	0.9769	0.9785	0.9799	0.9811	0.9824
20	0.9859	0.9869	0.9877	0.9885	0.9893
25	0.9900	0.9907	0.9913	0.9919	0.9924
30	0.9922	0.9927	0.9932	0.9936	0.9941
35	0.9933	0.9938	0.9942	0.9946	0.9950
40	0.9939	0.9944	0.9947	0.9951	0.9954
45	0.9941	0.9946	0.9949	0.9952	0.9956
50	0.9939	0.9944	0.9947	0.9951	0.9954
55	0.9934	0.9938	0.9942	0.9946	0.9950
60	0.9922	0.9928	0.9932	0.9936	0.9941
65	0.9900	0.9908	0.9913	0.9919	0.9925
70	0.9859	0.9869	0.9877	0.9885	0.9893
75	0.9769	0.9786	0.9799	0.9812	0.9825
80	0.9520	0.9554	0.9580	0.9606	0.9633
85	0.8368	0.8470	0.8547	0.8629	0.8715

が大きく変化してみえることはなく，下で述べる新たな決定係数M^2を併用して参照する必要がないことも明らかとなった。

以上でわれわれは，回帰式が各座標軸となす角度が与えられた場合に，決定係数の大きさを，楕円体状の分布で幾何的イメージとして把握することが可能となった。

しかし，上の計測結果でも示されたとおり，その角度次第で相当大きく決定係数が変化することを知った。したがって，決定係数だけを見て説明変数を取捨選択するようなことは，今述べたような意味でも適切ではないことがわかった。

Ⅳ．新たな決定係数について

それでは，次に，各点から回帰超平面への垂線の距離で決定係数を測ることを検討してみよう。それには，通常「直交回帰」といわれている手法を用いることが考えられる。この手法で求められる決定係数の値はデータ分布の形状にのみ依存し，座標軸との位置関係には依存しないからである。(ただ，直交回帰は従来から回帰分析ないし主成分分析との関連では議論されてきたが，通常の回帰分析における決定係数の問題点を改善するという観点はなかったように思われる。実際にも，以下で述べるような新たな決定係数は実用に供されてはこなかったのである。)

まず，説明変数が１個の場合で考える（図８参照)。すなわち，

$$Y = a + bX + \varepsilon \quad (\varepsilon \text{は誤差項})$$

の場合を考えよう。

記号の簡単化のために，回帰直線の推定パラメ一タ一に∧を付けず，

$$Y = a + bX$$

で表わす。

点（X_i，Y_i）からの垂線の式は，

$$Y = (Y_i + X_i / b) - X / b$$

であるから，回帰直線とそれに垂らした垂線との交点は，

$$\left(\frac{-ab + X_i + bY_i}{1 + b^2},\ \frac{a + bX_i + b^2 Y_i}{1 + b^2} \right)$$

である。

したがって，残差平方和，すなわち点（X_i，Y_i）から交点への距離の２乗和は，

$$E = \Sigma \left\{ \left(X_i - \frac{-ab + X_i + bY_i}{1 + b^2} \right)^2 + \left(Y_i - \frac{a + bX_i + b^2 Y_i}{1 + b^2} \right)^2 \right\} = \frac{\Sigma (a + bX_i - Y_i)^2}{1 + b^2}$$

となる。

そこで，座標軸との関係に依存せず，分布の形状だけで決まるような新たな決定係数は

$$M^2 = 1 - \frac{E}{\Sigma \{ (X_i - \bar{X})^2 + (Y_i - \bar{Y})^2 \}} \quad ①$$

$$= 1 - \frac{\Sigma (a + bX_i - Y_i)^2}{(1 + b^2) \Sigma \{ (X_i - \bar{X})^2 + (Y_i - \bar{Y})^2 \}} \quad \cdots\cdots ②$$

として求められる。

①式の分母はデータX_i，Y_iから求められるから，分子のEをX_i，Y_iで表せられればよい。次にこれを求めよう。

まず，

$$\frac{\partial E}{\partial a} = 0$$

から，

$$a = -\frac{\Sigma (bX_i - Y_i)}{n} \quad \cdots\cdots ③$$

となるが，これをEに代入して，

$$E = \frac{b^2 \Sigma (X_i - \bar{X})^2 - 2b \Sigma (X_i - \bar{X})(Y_i - \bar{Y}) + \Sigma (Y_i - \bar{Y})^2}{1 + b^2} \quad \cdots\cdots ④$$

を得る。さらに，

$$\frac{\partial E}{\partial b} = 0$$

とおくと，

$$b^2 \Sigma (X_i - \bar{X})(Y_i - \bar{Y}) + b \{ \Sigma (X_i - \bar{X})^2 + \Sigma (Y_i - \bar{Y})^2 \} - \Sigma (X_i - \bar{X})(Y_i - \bar{Y}) = 0$$

が得られる。これからもわかるとおり，このｂに関する２次方程式において，

$$D = \text{determinant} = \{ \Sigma (X_i - \bar{X})^2 + \Sigma (Y_i - \bar{Y})^2 \}^2 + 4 \{ \Sigma (X_i - \bar{X})(Y_i - \bar{Y}) \}^2 > 0$$

であるからｂは常に実根を持ち，

$$b=\frac{-\{\Sigma(X_i-\bar{X})^2+\Sigma(Y_i-\bar{Y})^2\}\pm\sqrt{D}}{2\Sigma(X_i-\bar{X})(Y_i-\bar{Y})} \quad ⑤$$

となる。

これを，②式に代入すれば，

$$a=-\frac{\Sigma(bX_i-Y_i)}{n}=-b\bar{X}+\bar{Y}$$

$$=\frac{[\{\Sigma(X_i-\bar{X})^2+\Sigma(Y_i-\bar{Y})^2\}\pm\sqrt{D}]\bar{X}}{2\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}+\bar{Y} \quad ⑥$$

となる。以上でパラメーターａ，ｂが，ともにデータX_i，Y_iにより表現できた。⑤式，⑥式を②式に代入すれば，回帰直線への垂直距離で測った新たな決定係数M^2が，データX_i，Y_iで明示的に表わされる。

さて，次に一般の場合について検討してみよう。

まず，データが，

$(Y_i,\ X_{1i},\ X_{2i},\ \cdots\cdots\cdots\cdots,\ X_{Pi})$

$i=1,\ \cdots\cdots,\ n$

として与えられるとすると，回帰平面

$$Y=a+b_1X_1+b_2X_2+\cdots\cdots\cdots+b_pX_p \quad ⑦$$

へ垂らした垂線の長さの２乗和は，

$$L=\Sigma\frac{(Y_i-a-b_iX_{1i}-b_2X_{2i}-\cdots\cdots b_pX_{pi})^2}{1+b_1^2+b_2^2+\cdots\cdots+b_p^2}$$

である。

このとき，Ｌを$(a,\ b_1,\ b_2,\ \cdots\cdots\cdots,\ b_p)$に関して最小化する。

$$\frac{\partial L}{\partial a}=0$$

から，

$$a=\bar{Y}-b_1\bar{X}_1-b_2\bar{X}_2-\cdots\cdots-b_p\bar{X}_p$$

となる。よって，Ｌは，

$$L=\Sigma\frac{(\tilde{Y}_i-b_1\tilde{X}_{1i}-b_2\tilde{X}_{2i}-\cdots\cdots-b_p\tilde{X}_{pi})^2}{1+b_1^2+b_2^2+\cdots\cdots+b_p^2}$$

$$=\frac{1}{1+b_1^2+b_2^2+\cdots\cdots+b_p^2}\times$$

$$(1-b_1-b_2\cdots\cdots-b_p)A\begin{pmatrix}1\\-b_1\\-b_2\\\vdots\\\vdots\\-b_p\end{pmatrix}$$

と書ける。

ただし，

$\tilde{Y}_i=Y_i-\bar{Y}$，$\tilde{X}_{ki}=X_{ki}-\bar{X}_k$（ｉ＝１，２，……，ｎ；ｋ＝１，２，……，ｐ）

であり，Ａは，

$$A=\Sigma\begin{pmatrix}\tilde{Y}_i\\\tilde{X}_{1i}\\\tilde{X}_{2i}\\\vdots\\\tilde{X}_{pi}\end{pmatrix}(\tilde{Y}_i,\tilde{X}_{1i},\tilde{X}_{2i},\cdots,\tilde{X}_{pi}) \quad ⑨$$

で定義される偏差積和行列である。

さらに，$x^T=(1-b_1-b_2-\cdots\cdots-b_p)$とおくと，

$$L=\frac{x^TAx}{x^Tx}$$

となる。

このＬの最小値はＡの最小固有値λmin（Ａ）であることが知られている。

一方，⑨式より，

$$\Sigma(Y_i-\bar{Y})^2+\Sigma(X_{1i}-\bar{X}_1)^2+\cdots\cdots\cdots+\Sigma(X_{pi}-\bar{X}_p)^2=trA$$

であるから，②式のM^2に対応して，一般の場合には，

$$M^2=1-\frac{\lambda_{min}(A)}{trA}$$

と決めることが，まず考えられる。しかしながら，M^2をこのように定義すると，ｔｒＡはＡの固有値の和に等しいので，一般には，

$$\frac{\lambda_{min}(A)}{trA}<1$$

である。

すなわち，言い換えると，一般には$M^2=0$とはならないのである。

そこで，②式のM^2に対応する一般の場合のM^2としては，

$$M^2=1-\frac{(p+1)\lambda_{min}(A)}{trA} \quad \cdots\cdots\cdots\cdots ⑩$$

と定義することにしよう。

このように定義されたM^2は，$0\leqq M^2\leqq 1$を満たし，Ａの固有値が全て等しいとき，つ

まり，Ａ＝λＩ（Ｉは単位行列）となってデータが球状に分布するときのみ$M^2=0$となることから，従来の決定係数R^2と類似の性質をもつ。

なお，現実のデータから求められる回帰軸（主軸）が完全に座標軸に平行になることは稀であろうが，そのような場合に，⑩がゼロになることまでも要請するならば，各データをまず基準化（normalize）してから⑩を計算すればよいことがわかる。

さて，このように定義されるM^2の性質を調べるために，３次元（説明変数２個，データ数25,000個）の場合について，モンテカルロ法による簡単なシミュレーションを行った結果が表３に示されている。

表３　M^2のシミュレーション結果

Ａ＝１	Ａ＝２	Ａ＝３	Ａ＝４	Ａ＝５
0.0115	0.5073	0.7308	0.8361	0.8887
Ａ＝６	Ａ＝７	Ａ＝８	Ａ＝９	Ａ＝10
0.9227	0.9411	0.9551	0.9644	0.9706
Ａ＝11	Ａ＝12	Ａ＝13	Ａ＝14	Ａ＝15
0.9760	0.9792	0.9824	0.9851	0.9866
Ａ＝16	Ａ＝17	Ａ＝18	Ａ＝19	Ａ＝20
0.9883	0.9897	0.9907	0.9918	0.9925
Ａ＝21	Ａ＝22	Ａ＝23	Ａ＝24	Ａ＝25
0.9933	0.9938	0.9944	0.9948	0.9953
Ａ＝26	Ａ＝27	Ａ＝28	Ａ＝29	Ａ＝30
0.9956	0.9959	0.9962	0.9965	0.9967

Ⅴ．まとめ

最後にこれまでの議論を整理してみよう。まず，ある決定係数が与えられた場合に，そのデータ分布状態を楕円ないし楕円体等で近似した場合の幾何的イメージは表１，２で明らかとなった。次に，回帰分析でより説明力のある変数を試行錯誤で探すような場合，通常使用される決定係数の大きさで評価することは危険である。回帰直線の座標軸に対して成す角度が小さい分布では，その分布の回帰直線の回りの散らばりが小さくても決定係数が極めて小さくなる場合があり，逆に，その角度が軸に対して45°の分布では，散らばりが大きくても決定係数がみかけ上大きくなる場合があるのである。

従来からの決定係数R^2がそのような性質をもつことは，その定義からみて明らかともいえるが，むしろ，問題はそのような場合に別途参照すべき代替的な指標が実用化されてこなかったことにあるといえる。

そのようなことから，従来からの決定係数R^2に加えて，本稿で述べた，⑩式で定義される別の決定係数M^2を併用して参照していくことが考えられる。それは，古くから知られている直交回帰を使用し，それにより求められる決定係数に若干の修正を加えたものであり，極めて簡単なアルゴリズムにより求められることから実用的でもある。

経済データの場合にはトレンド，系列相関

等の時系列データに関わる複雑な問題が存在するが，化学や医学や心理学等でクロス・セクションを中心とした回帰分析を行う場合には，決定係数でみた説明力がかなり重視されることが多い。そのような場合には，特に，M^2も同時に参照することが必要不可欠であると思われる。

参 考 文 献

松本和幸(1989)「最小２乗法により求められる決定係数の問題点」mimeo.