研究ノート

乱数による重回帰式の精度論

― 説明変数選択の作法 ―

平　井　孝　治

目　　次

は　じ　め　に

重回帰分析は文系，理系のいかんを問わず，いろんな学問分野で使われている。しかしながら，自由度修正済み決定係数 Q^2 がわずか 0.2 のレベルであったり，回帰式の P 値が 1% 弱であるなど，目に余るものも散見する。この程度の水準の回帰式から学問上の知見を得たとするのは如何かと思われる。この程度の精度なら，説明変数に乱数を用いても簡単に達成することが出来ることを，筆者はパソコンで何度も実証済みである。

目的変数と説明変数候補を多変量解析のソフトに丸投げすれば，とりあえずは重回帰式を提示してもらえるが，それで事たれりとするのは科学的分析の名に値しがたい。そこでこの小論では，過去幾多の乱数を用いた実験の結果から，①重回帰式のあり方と共に，②説明変数選択の作法 を精度論的に考察しようとするものである。

§1　乱数による実験の前提

この小論で想定するデータは，経済分野などの時系列データではなく，来街者調査のようなアンケート・データを仮定している。筆者の経験によれば，標本件数が数百程度のアンケート調査では，1% 程度の論外と，さらに 4% 程度の「はずれ鳥」がサンプル中に含まれるのが通常である。従って，主成分得点のプロット図や，暫定的な重回帰分析で，これらを標本空間から事前に除外しておくのが然るべきであろう。これを怠ると，見当はずれの「知見」を掴まされることになる。

私の研究室では，入手した標本のうち「5% を限度にサンプルはがし」を許容している。というのも，これ以上「はずれ鳥」がいるとすれば，むしろそれを織り込んだ知見を確立すべき

だからである。というわけで，ここでもサンプルはがしは，この 5% ルールに準拠している。検定でよく用いられる「5% の有意水準」は，まことに結構なもので，特段の事情がない限り，筆者もまた至るところでこれを重用しいている。

以下で紹介する例では，①説明変数の削り込みと並行して，②この 5% ルールによるサンプルはがしを，（擬似的な）実際値 u_i と回帰値 u_i' との残差平方 e_i^2 の大きい乱数サンプルから適用している。（以下，この論文では一貫して，実際値 u_i に対する回帰値には，表記上の都合から u_i' と記すことにする。）この 5% ルールの適用だけは，乱数標本に恣意的な手を入れたことになるが，その結果，例示の実験では，解析の対象となる標本が 215 件から途中経過を経て最終的には 205 件になることとなった。

目的変数 U には，一つの「2 択問題」V_1 と二つの「4 択問題」V_2, V_3 を想定して乱数を発生させ，もっともらしく $U = 3V_1 + V_2 + V_3$ と[1] 合成して作っている。

説明変数候補には，40 個の「2 択問題」$X_1 \sim X_{40}$ と，60 個の「4 択問題」$X_{41} \sim X_{100}$ を想定し，これもまた乱数によって擬装した。これでは「2 択問題」が多いように思われる向きもあるかもしれないが，実際のアンケートの調査票に「複数選択化」の質問が見受けられる。そのような場合には，個々の選択肢が「0，1」変数になるので，「4 択問題」との比は，この程度で妥当だと思われる。この結果，乱数による実験のデータ・サイズは合成した目的変数 U を含めて，215 件× 104 変数＝ 22,360 から出発している。

「複数選択化」の問題で，回答者が選んだ選択肢の数で数量化する場合もあるが，通常は，個々の選択肢を独立変数とみなし，「0，1」で数値数量化して分析するのが科学的な作法である。すると，A4 版表裏の調査票で調査変数の数 q が 60 個前後になるのが通常で，組織向けの調査など A3 版表裏の調査でも，120 個前後であろう。A4 版にして 5 頁以上の調査となると，調査票自体のホッチギスが外れたりなどして，サンプルのベクトル維持に苦労することになる。そればかりか，回答者にも大変な負荷をお掛けすることにもなるので，A4 版で 5 頁以上にわたる調査は，調査目的の実現のためやむを得ない場合を除いて，可及的に回避すべきであろう。

このような事情から，説明変数の候補数 q を 100 個としたが，それを 1,000 個とした過去の実験では，後に紹介する私の要求する精度にそれなりに近い回帰式が得られたことがある。それというのも，乱数変数を 1,000 個も作れば，その中には目的変数との相関がそれなりに強いものが出現する確率が高くなるからである。そのようなわけで，アンケート調査を想定して，ここでは説明変数の候補数 q を 100 個にした実験結果を例示する。

更に説明変数の削り込みに際してであるが，①目的変数との相関係数や，t 値，p 値を指標

1）目的変数としては，単変数だけではなく，主成分得点など，複数の変数を 1 次結合して合成することがしばしばある。ここでは「0，1」を 3 倍して＋ 1 すると「4 択問題」と同じ巾になるので，「2 択問題」をこのように加重して目的変数を合成した。

とするのが通常であるが，②調査の目的や仮説と予め付しておいた変数ラベルとのすり合せを行い，意味論も加味してベストな回帰式を探るのが重回帰の作法である。しかしながら，ここで示す乱数変数に意味などあろう筈も無いので，乱数実験では，各変数の回帰係数 α_i の危うさにのみ焦点を当て，ただひたすら P 値の大きい変数から割愛することを説明変数削り込みのアルゴリズムとする。

§2　回帰式に求められる精度

回帰式の精度を高めるだけなら，質問票に目的変数をモディファイした項目を潜ませておけば，容易に目的を達成することができようが，これは元来トートロジーでしかない。即ち，「自明な知見」が得られるだけに過ぎない。主成分の創り込みのため，アンケート票にそのような仕掛けをしておくことも稀にはあるが，その際には目的変数と類似した変数を当該解析枠から予め外しておくのが作法であろう。

回帰式に求められる精度については，時々の事情にもよるが，一般には調査の目的や，個々の学問分野の定説・慣習や，そしてなによりも研究者の研究姿勢によって規定されるものではある。しかし如何なる学問分野であれ，精度としては①自由度修正済み決定係数が $Q^2 \geqq 0.4096$ でかつ，②回帰式の P 値が $P \leqq 10^{-12}$ もあれば立派な回帰式として誰しも肯うに吝さかでは無いだろう。そこで筆者はプロット図から得られた実感に基ずき，表 1 のような自由度修正済み相関係数の絶対値 r' や，Q^2 に関する解釈を施し，統計的なデータを扱う際の礎にして来た。

表 1　自由度修正済み決定係数 Q^2 の相関レベルとその解釈

自由度修正済み相関係数 r'	レベル	解　釈	自由度修正済み説明係数 Q^2
$0 \leqq r' < 0.16$	1	相関があるとは言えない	$0 \leqq Q^2 < 0.0256$
$0.16 \leqq r' < 0.25$	2	乱数でも起こりうる程度	$0.0256 \leqq Q^2 < 0.0625$
$0.25 \leqq r' < 0.36$	3	若干の相関が見られる	$0.0625 \leqq Q^2 < 0.1296$
$0.36 \leqq r' < 0.49$	4	そこそこの相関がある	$0.1296 \leqq Q^2 < 0.2401$
$0.49 \leqq r' < 0.64$	5	それなりに強い相関	$0.2401 \leqq Q^2 < 0.4096$
$0.64 \leqq r' < 0.81$	6	非常に強い相関がある	$0.4096 \leqq Q^2 < 0.6561$
$0.81 \leqq r' \leqq 1$	7	理論的な関係にある相関	$0.6561 \leqq Q^2 \leqq 1$

ちなみに筆者は「レベル 7」は理論モデルで，「レベル 6」は実用モデルとして使えるが，自由度修正済みの Q^2 がようやく 0.24 を超える程度の「レベル 5」では，せいぜい説明モデルになるに過ぎないと考えている。

ここで「自由度修正済み相関係数 r'」なるしろものを出しているが，筆者はかねてより相関係数 r はサンプル件数によって修正を掛けられて然るべきだと考えてきた。誰しも一度はそう思った筈で，あるいは誰かが既にこれを定義されているかもしれない。そもそも標本が二つしかない時には，当然相関係数は $r = \pm 1$ となる。即ち，件数が少なければ少ないほど r の絶対値は当然大きく表示される。件数が 402 件を超えているときは，さほど障害にならないが，

筆者の考えでは，自由度が 400 度以下の場合は，相関係数 r に，次のような修正を掛けるべきである。

既にどなたかが定義されているならお許しいただきたいのだが，相関係数 r は当然二つの変数間で定義されているので，r’ は，一方の変数を「目的」とし，それをもう一方の変数で「説明」すれば，自由度修正済み決定係数 $(r')^2$ が出きて，それを容易に定義することが出来る。件数を $n \geqq 3$ とすると，定義式は次のようになる。[2)]

$$(r')^2 = 1 - \frac{n-1}{n-2}(1 - r^2) \qquad \cdots (0)$$

ここに，r’ の符号は断るまでもなく Sign（r）である。

式（0）で，元の相関係数が $r^2 < \frac{1}{n-1}$ ならこのままでは定義できないが，その時は当然 r’ ＝ 0 と定義する。この小論ではこのように定義して「自由度修正済み相関係数 r’」なる概念を使用することにする。

私は 1997 年の 4 月に立命館大学に奉職して以来今日まで，院生達と共に研究や授業で年に少なくとも二回は各種の調査を設計し，実施の上，アンケート・データを解析・分析[3)] して来た。その際の重回帰分析では，表 1 でいう「レベル 6」の，自由度修正済み決定係数の到達水準を $Q^2 \geqq 0.4096$ として譲らず，特に研究室所属の院生には苦労を強いてきた。

過去に実施した数多くの調査では，概要設計の段階で①三本の主成分の創り込みと，②少なくとも三つの目的変数を想定したのであるが，実際に集めたデータを解析すると，前者は予期通りに出てくる。しかしながら，後者の予め設定しておいた目的変数に対する個々の重回帰式を同一の解析枠で結ぶ[4)] のは至難の業である。

「はずれ鳥」を 5% ルールで解析枠から除く際，こちらの重回帰式の顔を立てれば，あちらの重回帰式の Q^2 が 0.4096 を下まわることなどしばしばで，一時期は要求水準を $Q^2 \geqq 0.4000$ まで切り下げようと考えたこともある。わずか $4096 \div 4000 = 2.4\%$ の差でしかないが，これが実に「厳しい壁」なのである。結局のところこの時も妥協せずに，$Q^2 \geqq 0.4096$ の水準を生涯貫徹することにして，今日に至っている。

§3　「非常に強い相関」の提示

前掲の表 1 における「各レベルの解釈」は，あくまでも筆者の唱えるものであって，もとよりそれは「然るべき筋から権威づけ」られたものではない。しかしながら筆者の豊富な経験

2）以下に定義する「自由度修正済み相関係数 r’」は，本学経営学研究科の社会人研究生　四方健雄氏との議論の結果えられたものである。

3）私の研究室では，データを解析（しグラフを出力）するところまでの工程を「解析」と称し，それから意味論を加味して知見を出す工程を「分析」と称して，Analysis という言葉を漢字で区別して使っている。

4）後に紹介する要求水準に達した回帰式を得たとき，私の研究室では「重回帰式が結んだ」とか，「目的変数が結んだ」と称している。

から割り出した数値区分に，実感した「形容詞」を附したもので，このたびの乱数による実験を経て，自信を持ってここに提示している。

重回帰式に求められる水準は，自由度修正済み決定係数のみならず，回帰式のP値などに対するもの等，当然いろいろあって然るべきあろう。ここではそれを以下に表2としてまとめて示しておく。

表2　重回帰式に求められる水準

要求の厳格度		項　　目	要求水準
①	☆	自由度修正済み決定係数	$Q^2 \geqq 0.4096$
②	☆	説明変数の数	$q \leqq 12$
③	［＊］	各説明変数のAT値	$AT \geqq 2.45$
④	☆	各説明変数のP値	$p < 5\%$
⑤	［＊＊］	回帰式のP値	$P < 10^{-12}$
⑥	☆	マルチコ変数の数	$m = 0$

自由度修正済み決定係数がいかに水準に達していようとも，説明変数の数qが多いのでは，実用はもとより説明にも窮することとなる。そこで，10個では苦しいこともままあるので，回帰式に臨席する説明変数の数qとしては12個を上限とするのが至当と思われる。

回帰式に登場する各説明変数X_iの重要性の指標は，各々t値の絶対値AT_iである。[5] 然るに，回帰係数α_iを計算するには，全説明変数から構成される分散・共分散行列V（や相関行列R）が用いられるが，ここではこれらと数学的に同値な偏差積和行列S[6] を使ってt値を説明することにする。

実際値u_iから回帰値u_i'を引いた残差e_iの平方和を残差の自由度[7] で除した「残差分散V_{ee}」に，Sの逆行列S^{-1}のそれぞれの説明変数に対応する対角成分を掛けると，各「説明変数の誤差分散」が計算される。その平方根が該説明変数の「回帰係数α_iに関する標準誤差」で，いわゆる「回帰係数のゼロ検定」に用いられることとなる。即ち，各説明変数の回帰係数α_iを，今のようにして求めた対応する標準誤差で割ると，それぞれt値が求まるというわけである。

よく知られているように，正規分布の場合は，1.96σより左右の尾部の累積確率がほぼ5%になるが，t分布でも自由度がそこそこ大きければ，それを2.00σとみなしてもさほど問題はない。なのでAT値としては通常2.00以上もあれば十分であるが，筆者としてはF値に換算して6.000ぐらいになる2.5σぐらいを想定した$AT \geqq 2.45$を要求水準としたものである。（この水準を$F \geqq 6.000$で置換してもよい。）よって，AT値に対する要求の厳格度としては，表2のごとく最も低い［＊］にとどめている。

5）通常はt値の平方である「F値」を用いて論じられるが，この小論では絶対値を用いることとし，notationとしては「AT_i」を使うことにする。

6）分散・共分散行列Vを件数倍したものと同値。

7）よく知られているように，残差平方の和でもって「残差変動」を定義し，これをその自由度「件数n－1－説明変数の数q」で除した値を「残差分散」と称する。

さらに，ことの理から説明変数の「各回帰係数 α_i がそれで誤る確率 p」は，当該回帰式の残差変動の自由度の下における t 分布で，t 倍の σ の左右尾部の累積確率によって求められる。表 2 では，この各変数の P 値に $p < 5\%$ なる厳格な基準を求めている。

多くの分析者が回帰式の精度の指標として，自由度修正済み決定係数 Q^2 の次に問題としておられるのは，おそらく回帰式の P 値と思われる。仄聞するに，これが 1% 以下なら由とする向きもあるようだが，これが 0.001% 程度でも論外で，得られた式は「怪奇式」と断ぜざるを得ない。

回帰式の P 値は,『「回帰変動[8] を説明変数の数 q で除した不偏分散」と残差分散の間に有意な差があるか否か』を問うことを背景にしたもので，従って，前者と後者に分散の差の検定に使う F 分布を適用して解することとなる。具体的には，不偏分散と残差分散の二つの自由度の下における F 分布において，前者を後者で除した値から右尾部にある累積確率が，求める回帰式の P 値である。

この値に対する要求水準は $P < 10^{-12}$ と厳しそうに映るが，今までに述べた三つの基準①，②，④に加え，次の「マルチコ無し」ともなれば，ほぼ自動的に充足される基準である。そこで表 2 では，これに対する厳格度は幾分ゆるい［＊＊］に設定しておいた。事実，筆者の過去の解析では問題にしたこともないが，基準①，②，④，⑥を同時に満たす場合には，回帰式の P 値が $P \leqq 10^{-12}$ になることなどごく普通である。

さて,最後は「多重共線形性 (Multi-Co-Linearity)」についてであるが,日本ではこれを俗に「マルチコ」と称している。この現象はいくつかの説明変数の間に近似的な線形関係がある場合に生起し，その結果，近似的な線形関係を有するある説明変数と目的変数の相関係数 r が，当該変数の回帰係数 α_i の（±の）符号と矛盾してしまうことがある。この現象が残存する限り，求めた式を正当な回帰式として受け入れるには抵抗がある。このマルチコが消えない限り，まがいものの「怪奇式」なので，表 2 の克服すべき必須事項として⑥に挙げている。

以下で紹介する乱数による実験は，表 2 の水準を求めて苦戦する中で，回帰式に求められる精度を詳しく検討しようとしたものである。

§4　乱数による重回帰の実験結果

以下に示す乱数による重回帰の実験例は，趣味的に数年にわたって，十数回行ったそのうちの一つである。そこで，実験の手段を維持するため，Windows は XP，Excel は 2003，使用した解析ソフトはいずれもエスミ社のアドインソフト「Excel 多変量解析 Ver.5.0」であ

8）各回帰値 u_i' からその平均値（実際値の平均と同じ値）を引いた偏差の平方和を「回帰変動」と称するが，それをその自由度,即ち説明変数の数 q で除した値を「回帰値の不偏分散」という。なお,云うまでもないが,回帰変動を件数 n や n − 1 で除した場合は，前者は「回帰値の母分散」になり，後者は「回帰値の標本分散」になる。

る。この原稿を書いている現在では，研究や授業などで，それぞれ Windows は 7，Excel は 2010，解析ソフトは同社の「Excel 多変量解析 Ver.6.0」にバージョンアップしているが，乱数による回帰実験については，比較可能性を維持するため，従前通り Ver.5.0 を使用している。バージョンをアップして「乱数による回帰実験」をしても，おそらく従前と同じような結果がえられるものと思われる。

一回の実験につき説明変数に想定した乱数変数を 100 個から 1 個まで，時々の P 値を観察しながら削り込んでいくのであるが，いつも難しく思うのは 5% ルールではずれ鳥を解析枠からはがすタイミングである。はずれ鳥をはがす指標は残差平方の大きい標本ではあるが，はがすタイミングについては，説明変数削除のようなアルゴリズムがないので，試行錯誤と経験と勘に頼らざるを得ない。その結果，一回の実験で 300 回ぐらいの重回帰分析を繰り返すこととなる。

例示する実験では，一変数につきスタート時点では 215 件の乱数がぶら下がっており，これが 100 変数ある解析枠から実験を開始する。表 3 の「ステージ 0」がそれであるが，この時は，説明係数 R^2 は 0.381 もあるが，自由度で修正したそれは当然 $Q^2 = 0.000$ で，マルチコ変数も丁度四分の一の 25 変数存在していた。

この時点で，先ずは極端なはずれ鳥を 1 件はがし，さらに重回帰してもう 1 件，つごう 1% の標本を解析枠から剥がして，ステージ 1 に移行する。すると表 3 のごとく，自由度で修正した説明係数は $Q^2 = 0.000$ と変わりがないが，回帰式の P 値が約 30% も下がってくる結果となった。

更に 2 件のはずれ鳥をはがすのと並行して，乱数変数を削り込んでいくと，ステージ 4 に至るが，この時点で，Q^2 はわずか 0.191 に過ぎないにも拘わらず，回帰式の P 値は 1% を切り，解析ソフトでは早くも［＊＊］が点灯する段階に至る。しかし，説明変数の数 $q = 81$ のうちマルチコ変数が $m = 14$ 個もあれば，誰しもこれを線形モデルとして掲げるわけには行かない。いずれにしてもこのステージの結果は，乱数変数ですら回帰式の P 値が容易に［＊＊］に達することを意味している。

実験例では，続いて回帰のレベルが表 1 の $L = 5$，即ち自由度修正済み決定係数が $Q^2 \geqq 0.2401$ になるまで，5 番目のはずれ鳥を解析枠から外すことと並行して，乱数変数を削り込んでいくと，表 3 のステージ 5 に至る。この事例では，この時点で早くも回帰式の P 値が千分の 1 を割り込んでくるので，表 4 の［分散分析表］の判定欄に相当するところには，さしずめ［＊＊＊］を刻印すべきであろう。ここまで来れば，説明変数の数を $q = 77$ のまま，解析枠からはずれ鳥を一気に 2 件外して，ステージ 7 に至り，マルチコの数も $m = 13$ と出発時から比べるとおよそ半減する。

この事例では 5% ルールに従い，解析枠より 215 件 × 5% ≧ 10 件の標本が除外できるが，

このような時には，筆者は，はじめの 4 件と，次の 3 件と，さらに次の 2 件という具合に，標本はがしに段階を設け，第一段階では最も重要な“目的変数を結び”（前述の注 4 を参照），次の段階で二番目の目的変数を，三番目の段階で三番目に結びたい目的変数を結ぶために備え，最後の 1 件で残りの目的変数を一挙に結ぶことを慣例としている。

ステージ 7 からは Q^2 を改善しつつ，重回帰式の P 値が Mega 分の 1 以下になるよう，変数をさらに削り込んでいく。そしてその目標を達成したのがステージ 8 であるが，この時点で，自由度修正済み決定係数は $Q^2 = 0.338$ となった。

ここからは「はがしの第三段階」に入り，乱数変数をおよそ 20 個ほど削り込むのと並行して，途中で 2 件の標本を解析枠から外して，最とも重要な指標である自由度修正済み決定係数

表 3　乱数回帰の各ステージにおける緒元

ステージ No.	前工程から剥がした標本件数と説明変数の削除による成果	自由度	説明変数の数	マルチコ変数の数	説明係数	自由度修正済み説明係数	回帰のレベル	回帰式の P 値	有力変数 X92 の P 値	有力変数 X92 の AT 値
	記号	d	q	m	R^2	Q^2	L	P	P_{92}	AT_{92}
0	Start	214	100	25	0.381	0.000	1	96.4%	8.43%	1.741
1	2 件	212	100	24	0.450	0.000	1	66.8%	8.00%	1.767
2	1 件，レベル 2 に到達	211	97	22	0.476	0.031	2	36.4%	5.66%	1.926
3	1 件，レベル 4 に到達	210	86	19	0.503	0.158	4	2.70%	2.32%	2.298
4	回帰式の P 値が［＊＊］に到達	210	81	14	0.503	0.191	4	0.78%	1.76%	2.405
5	1 件，レベル 5 に到達	209	77	17	0.520	0.240	5	9E-04	1.38%	2.497
6	1 件	208	77	15	0.536	0.264	5	3E-04	1.84%	2.387
7	1 件	207	77	13	0.547	0.279	5	2E-04	1.68%	2.423
8	回帰式 P 値 Mega 分の 1 に到達	207	64	9	0.543	0.338	5	7E-07	0.75%	2.713
9	1 件	206	61	6	0.551	0.362	5	8E-08	0.18%	3.187
10	1 件	205	56	7	0.560	0.395	5	2E-09	1E-03	3.360
11	レベル 6 に到達	205	45	5	0.540	0.410	6	1E-11	1E-04	3.979
12	1 件	204	45	6	0.549	0.422	6	5E-12	8E-05	4.054
13	修正済み Q^2 の Peak & Tera	204	41	5	0.540	0.425	6	8E-13	2E-05	4.403
14	レベル 6 維持の下限 & Tera	204	29	1	0.495	0.411	6	2E-14	3E-05	4.323
15	全変数の P 値が 5% 以下と回帰式 P 値が最大 & Tera	204	15	1	0.380	0.331	5	2E-13	9E-06	4.561
16	説明変数数 q の上限充足とマルチコ皆無を同時達成	204	12	0	0.331	0.289	5	6E-12	9E-05	3.999
17	レベル 5 維持の下限	204	9	0	0.280	0.247	5	1E-10	2E-04	3.852
18	レベル 4 に後退しつつも全変数の p 値 1% 以下に到達	204	7	0	0.239	0.212	4	2E-09	2E-04	3.789
19	レベル 4 維持の下限でかつ回帰式 P 値 Mega 分の 1 の下限	204	4	0	0.157	0.140	4	7E-07	3E-04	3.697
20	単回帰の Best	204	1	0	0.053	0.049	2	9E-04	9E-04	3.385

注）「& Tera」は回帰式の P 値が 1 兆分の 1 以下の意

が $Q^2 \geqq 0.4096$ に達することを目指す。この事例でそれをなしえたのはステージ 11 で，マルチコ変数も $m = 5$ 個と出発時点の五分の一に減じてはいる。しかしながら，さすがに乱数変数を用いているので，説明変数の数は基準の 12 変数には遠く及ばない $q = 45$ 変数だし，表 3 には掲載していないが，説明変数の P 値に至っては最大で実に $p = 35.55\%$ もあった。即ち，表 2 の①を満たすだけでは不十分だということである。

この次には「はがしの最終段階」として，本来なら最後のはずれ鳥を削ることによって，「他の目的変数を結びたい」のであるが，この事例ではそれを仮想していない。そこでこの事例では以降の解説に便宜なはずれ鳥を選ぶのであるが，そのために幾通りもの重回帰式を作るはめになった。その結果が次に示すステージ 13 である。

この乱数による重回帰実験で，当該ステージは自由度修正済み決定係数が $Q^2 = 0.425$ とピークに達し，回帰式の P 値も $P = 7.7 \times 10^{-13}$ と Tera 分の一を下回ることとなった。「☆」一つを 10^{-6} とすると，表 4 のような［分散分析表］の判定欄には［☆☆］と記すことになるが，いまだに説明変数が 41 個と数多くあり，さらにはマルチコ変数も依然として 5 個も残存している。従って表 2 に示す要求水準が如何に厳しいものであるかが理解されようというものである。

ここから後のステージでは 1 件も“はがし”が出来ないので，説明変数が一つ（単回帰）になるまで，ただひたすら説明変数の P 値の大きいものから順に，アルゴリズミカルにそれを削り込んで行く。自由度が不変の場合，説明変数を削るたびに決定係数 R^2 が落ちていく[9] のはことの理である。しかし自由度修正済みの方は，ステージ 3 → 4 のように，逆に上がることもある。それを期待しつつ，先ずは要求基準の⑤，即ち Q^2 を犠牲にして，すべての説明変数で P 値 5% 以下を目指すことになる。

それがようやく達成できたのは，先のステージ 13 から更に 26 個もの乱数変数を削り込んだ後のステージ 15 であった。このときまでには自由度修正済み決定係数は，ピークの $Q^2 = 0.452$ から 0.331 まで落ち込んで来ている。このステージでマルチコ変数はわずか 1 個になったが，しかし説明変数の数がまだ $q = 15 > 12$ なので，要求水準の②を満たすには至っていない。

変数選択のかなり初期段階でマルチコ変数を解析枠から削除してしまいたいところだが，他の説明変数を削除する過程でマルチコ符号が自然消滅することも多々あるので，ここはじっと我慢して，やはり P 値の高い変数から順に丁寧に削除して行くのが説明変数選択の作法であろう。

要求水準の④を満たすべく，説明変数数を $q = 12$ になるまで，ステージ 15 から更に 3 変

9）説明変数の数を i とすると，正確に表現すれば $R_{i+1}^2 \geqq R_i^2$ と単調減少する。

数削っていくと，次のステージ 16 に至る。このとき事例では偶然にもマルチコ変数の数が m ＝ 0 となった。即ち，q ≦ 12 と「マルチコ皆無」が同時に達成できたのではあるが，自由度修正済みの方は $Q^2 = 0.289$ と，さらに落ち込んでいる。

引き続き変数を削除して行って q ＝ 7 まで来ると，説明変数のすべての P 値が 1% 以下のステージ 18 になったが，この回帰結果を参考までに表 4 で示しておく。なおこの表は，この節の冒頭に記しておいたエスミ社のソフトが出力したものに，筆者が手を入れたものである。

表 4　標本件数 n ＝ 205，説明変数数 q ＝ 7（ステージ 18）の回帰結果

［重回帰式　目的変数　U］

説明変数	回帰係数	標準 β	F 値	P 値	判定	T 値	標準誤差	偏相関	単相関	マルチコ
78	0.299	-0.239	6.860	0.95%	［＊＊］	2.62	0.114	0.183	0.090	
61	0.309	-0.234	7.132	0.82%	［＊＊］	2.67	0.116	0.187	0.165	
28	-0.761	-0.207	8.400	0.42%	［＊＊］	-2.90	0.263	-0.202	-0.144	
66	-0.358	-0.193	8.876	0.33%	［＊＊］	-2.98	0.120	-0.208	-0.185	
4	-0.849	-0.183	10.466	0.14%	［＊＊］	-3.24	0.262	-0.225	-0.179	
79	-0.419	0.170	13.607	0.03%	［＊＊＊］	-3.69	0.114	-0.254	-0.181	
92	-0.428	0.165	14.354	0.02%	［＊＊＊］	-3.79	0.113	-0.261	-0.231	
定数項	8.548		170.02			13.04	0.656			

［精度］

決定係数	R^2 ＝	0.2395
同修正ずみ	Q^2 ＝	0.2124
重相関係数	R ＝	0.4894
同修正ずみ	Q ＝	0.4609

［分散分析表］

変動	偏差平方和	自由度	不偏分散	分散比	P 値	判定
全変動	858.08	204				
回帰変動	205.48	7	29.35	8.86	1.8E-09	[☆]
残差変動	652.60	197	3.31			

［標本の残差平方］

標本 No.	実際値	残差	回帰値	残差平方
64	10	4.67	5.33	21.79
90	2	-4.31	6.31	18.57
52	3	-4.25	7.25	18.05
12	4	-4.16	8.16	17.27
53	3	-3.91	6.91	15.31

参考までに，大きいものから順に五つだけを示しておいた。

ついでにこの表 4 に対するコメントを次に附しておく。

上段の［重回帰式］は F 値で昇順にソートしてあるので，説明変数の AT 値（t 値の絶対値）の最小は説明変数 X_{78} のそれである。然るに乱数変数 X_{78} の t 値が 2.62 なので，この重回帰は要求基準の③である $AT_i \geqq 2.45$ を満たしていることになる。なおこの変数の偏相関係数は単相関のおよそ 2 倍にもなっており，1% をわずかに下まわる 0.95% の P 値と共に，当該重回帰式にとってまことに都合のよい説明変数ではある。

次にこの回帰式の［精度］についてであるが，自由度修正済み決定係数が $Q^2 = 0.2124$ なので，回帰のレベルは L ＝ 4 である。このレベルでは，表 1 の解釈によれば，「そこそこの相関がある」程度で，説明モデルにもならない。しかし他方で［分散分析表］によれば，この回帰式は，そ

のP値が実に一億分の一を下回っている線形モデルでもある。このことからも理解されるように，重回帰式のP値が一億分の一程度では，学術的には何の価値もないだけではなく，下手をすると，致命的な知見を導くリスクすらある。

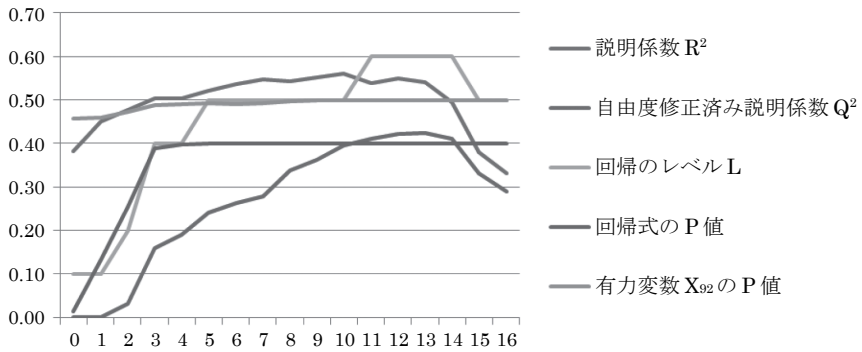
表4のような回帰式ならどの学会でも通用しそうだが，乱数でもこの程度まで来るのであるから，考えものである。

説明変数を絞り込んでいくと，最後は「一変数で重回帰」することになるが，この事例では乱数変数X_{92}が残った。その結果の緒元を表3のステージ20に示している。結果から見れば，当初100個あった乱数変数のうち最も相関の高いX_{92}[10]で目的変数Uを説明することとなったが，この場合の自由度修正済み決定係数Q^2が，Uと有力な変数X_{92}の（§2で導入した）自由度修正済み相関係数r'の平方になることは言うまでもない。その値は$Q^2 = 0.0488$だったので，回帰のレベルはL＝2となり，文字通り「乱数でも生起しうる程度」の回帰である。

単回帰の場合は，t分布に従う説明変数のP値と，F分布に従う回帰式のP値が一致する。というのも，説明変数が一つしかないので，回帰係数のリスクがそのまま回帰式のリスクということになる。この事例の場合その値はp＝0.086%で，「判定」は10^{-3}以下を意味する［＊＊＊］ということになる[11]。

なお，次の図1は，表1の緒元を折れ線グラフに表したもので，脚注2で紹介した四方健雄氏に作成してもらったものである。

図1　表3の緒元を1次式で然るべく変換した値の折れ線グラフ



§5　各ステージの緒元と有力変数

最後の§として，例示した「乱数による重回帰式」を総括するために，先ずは自由度で修正済みの決定係数Q^2と，先ほど§3の表3に掲げた緒元との相関関係を調べ，検討してみるこ

10）もっとも標本を削るたびに相関係数が変わってくるので，注意を要する。

11）ちなみに，この時のt値の絶対値はAT=3.385になっていた。

とにしよう。表 2 の要求基準から見れば，ステージ 17 以降の重回帰式は，最後のステージ 20 を除いて，面白くはあってもここでは特段の意味が無いので，この節では考察の対象からは外すことにする。

そういうわけで，ステージ 0 からステージ 16 までに限って，Q^2 と他の項目との自由度で修正済した相関係数 r’ を求めると，表 5 の左側のごとくなった。対象とするステージが 17 通りあるので，この r’ を求める際の自由度には d ＝ 16 を適用することになる。すると回帰のレベル L は，Q^2 を数値区分して定義しているので，表 1 でいう「理論的な関係にある相関」の 0.953 になっても不思議ではない。

後に「有力変数 X_{92}」については詳述するが，レベル L の他，Q^2 と各ステージの標本の数 n に依存する自由度 d とは 0.930 の，変数 X_{92} の P 値とは 0.927 の，マルチコ変数の数 m とは（負で）0.882 の，X_{92} の t 値[12] の絶対値とは 0.827 の，それぞれ「自由度修正済み相関」となった。これらの相関はいずれも表 1 でいう r’ と「理論的な関係にある相関」であることを示している。

それに対し，各ステージの回帰式の P 値や説明変数の数 q との自由度修正済み相関係数は，それぞれ 0.784 と 0.758 だったので，二つとも「理論的な関係にある相関」につぐ「強い相関関係」にあることが判る。

これに対し通常の説明係数 R^2 はというと，Q^2 がそれを使って自由度で修正されているだけなのに，当該とわずか 0.386 の相関しか見せておらず，表 1 の表現を借りていうと，ようやく「そこそこの相関が見られる」程度であった。このことは，回帰式の精度を見る際に，修正前の決定係数 R^2 の値はほとんど意味がないという厳然たる事実を含意していることになる。

なお，いつでも成り立つ $R^2 \geqq Q^2$ という関係から，R^2 と Q^2 の差があればあるほど，改善の余地があり，「優れた回帰式」ほど Q^2 が R^2 に接近していく。

話は変わるが，この小論で事例に挙げた実験ではサンプル数 n を，「はがし」によって，215 件から 205 件に減らして来た。そのため，そのつど目的変数と各乱数変数との相関係数を時々の自由度で修正を掛けることになった。その間，ステージ 20 まで残った説明変数 X_{92} は，どのステージでも一貫して，目的変数との自由度修正済み相関係数 r’（の絶対値）が，説明変数たちの中で最大の位置を保ち続けた。

元より説明変数 X_{92} には乱数が振り当てられており，目的変数 U もまた乱数変数から合成されているのではあるが，（元は負の相関の）二者間の r’ の絶対値はこの実験の間，表 5 の右側の最右列で示したように，0.188 からスタートしてステージ 9 でピークの 0.223 に達し，その後の自由度が d ＝ 204 になった各ステージでも，ほとんど減少すること無く，単回帰に達している。

12）後述するように，目的変数 U と有力変数 X_{92} はどのステージでも負の相関があったので，X_{92} に関する回帰係数 α_{92} は，一貫して負であった。

表 5　各ステージの緒元と，有力変数 X_{92} の相関関係

表 3 に表記されている自由度修正済み決定係数と，他の項目との自由度修正済み相関関係				ステージ No.	自由度 d	説明変数の数 q	X92 の AT 値	目的変数 U との r’
Q^2 と他項目との修正済み相関		r’	順位					
自由度修正済み決定係数	Q^2	1	0	0	214	100	1.741	0.188
回帰のレベル	L	0.953	1	1	212	100	1.767	0.196
自由度	d	0.930	2	2	211	97	1.926	0.205
有力変数 X_{92} の P 値	P_{92}	0.927	3	3	210	86	2.298	0.208
マルチコ変数の数	m	0.882	4	5	209	77	2.497	0.216
有力変数 X_{92} の AT 値	AT_{92}	0.827	5	6	208	77	2.387	0.209
回帰式の P 値	P_0	0.784	6	7	207	77	2.423	0.206
説明変数の数	q	0.758	7	9	206	61	3.187	0.223
決定係数	R^2	0.386	8	10	205	56	3.360	0.222
				12	204	45	4.054	0.221
「目的変数 U との X92 との相関 r’」と各項目との自由度修正済み相関係数				0.865	0.889	0.876	0.845	1

これらの値は表 1 の相関レベルでいうと，L ＝ 2 と当然低いものではあるが，文字通り「乱数でも生起しうる」レベルである。ちなみに，ステージナンバーとの相関を取ってみると，件数は 10 件であるが，修正済みで 0.865 もあった。この値は，ステージナンバーとですら，優に「理論的な関係にある相関」であることを示している。

更には，偏回帰係数 α_{92} にかかわる t 値の絶対値 AT も，ステージの進行とともに一貫して上昇し続け，ステージ 12 では AT ＝ 4.054 となり，F 値でいうと F ≧ 16.000 で，乱数としては非常に高い値を示した。AT 値と，先に求めた r’ との更なる相関も調べて見ると，表 5 にある通り，これもまた「理論的な関係にある相関」を示している。

以上の事実から，有力変数 X_{92} はこの実験の中で，合成された目的変数 U の重回帰に関し，決定的な貢献をなしていることが理解されよう。普通のアンケート調査における設計や解析においても目的変数以外に，このような有力な説明変数を ①予め埋設しておくか，もしくは，解析段階で自由度修正済みの ②相関係数を使ってそれを見出すことを是非にもお勧めしたい。ただし，それが目的変数とあまりに高い相関を持つ場合は，トートロジーになっている可能性もあるので，注意を要する。

お　わ　り　に

読者の中には，乱数ですら諸々の制約条件を満たす回帰式が普通に成立することに驚かれた方もあるかと思います。それゆえ，表 2 に掲げた「要求水準」を満たすような見事な回帰式

を創るよう，調査の設計や，データの解析段階で，妙味をプロデュースされることをお奨めします。

最後にまた敢えて再掲する六つの要求水準が「OR」で成立するだけならば，乱数であっても「重回帰式を結ぶ」ことが可能なのである。その時には乱数変数にはどんな解釈を附すことも出来るので, よって「OR」で得られたそれは価値ある回帰式とはいい難い。是非にも「AND」で成立させるべきことを，実例を用いて説いて来た所存です。

表 2　重回帰式に求められる水準

要求の厳格度		項　　目	要求水準
①	☆	自由度修正済み決定係数	$Q^2 \geqq 0.4096$
②	☆	説明変数の数	$q \leqq 12$
③	［＊］	各説明変数の AT 値	$AT \geqq 2.45$
④	☆	各説明変数の P 値	$p < 5\%$
⑤	［＊＊］	回帰式の P 値	$P < 10^{-12}$
⑥	☆	マルチコ変数の数	$m = 0$

興味や疑念のある読者のために，この事例の出発時の解析枠をメールでお送りする用意があります。筆者が「乱数による重回帰」の論文を書くためのデータを作成するのには，この事例の計算だけで約 60 時間を要していますが，追試だけなら，おそらく 10 時間ぐらいで済むと思います。宜しく。

wintel @ pl. ritsumei. ac. jp