論　　文

VBAによるユーザーの判断を重視した重回帰分析ツールの開発

VBA Development of Multiple Regression Analysis Tools Emphasizing User Determination

堀　田　裕　史
HORITA Hiroshi

１．はじめに

回帰分析は、被説明変数と説明変数との関係を調べるものである。ダーウィンの従兄のゴールドウィンが親子の身長の関係を考察して以来（1889年）使われており、今日多くの教科書に登場する[1）2）3）]。

ただし回帰分析では、説明変数の候補となるデータが100種以上ある場合、何を説明変数としてよいか迷う。特に重回帰分析で、その選択は難しいことが多い。このため自動的に説明変数を選択する統計ソフトが存在する。これらでは、変数増加法や変数減少法が使用される。例えば変数増加法では、別の説明変数を追加してみてフィッテングがよくなればその変数を使用し、良くならないならばその説明変数は使用しないで、別の説明変数の候補の使用を試みる。この様にして、全ての説明変数の候補から、一定数の説明変数で、被説明変数のフィッテングがよいものを自動的に決定するのである[4）]。

回帰分析では、被説明変数のデータは結果側であり、説明変数を使った一次形式は原因の側である。被説明変数のデータの変動を、説明変数を使った一次形式で説明ができればよく、説明がどの程度できているかの尺度として決定係数の高さが求められる。

ところが、説明変数の候補が100以上ある場合については、偶然被説明変数の変動を説明するものがありうる。しかし、これでは意味がないのである。また、回帰分析では被説明変数の側は結果で、説明変数は原因となる事象であることが望まれる。被説明変数が原因で説明変数が結果と受け取れるものは好ましくない。実際は、原因と結果の判断は難しいが[5）]、ともかく、被説明変数が結果で説明変数は原因と見なせる組み合わせを探すことになる。説明変数の自動選択では、偶然説明変数の変動を与えるものや、因果関係の観点から望ましくない説明変数をも含めてしまう危険がある。

そこで、説明変数の候補が100以上もあるような重回帰分析をする際、ユーザーインターフェースに優れて使い勝手がよく、被説明変数との関係では、偶然または脈絡がな

ほりた　ひろし（食物栄養学科）

いか、因果関係が不自然なものが除外できるように、ユーザー自身が判断して説明変数を選択できるような重回帰分析ツールがあると望ましいといえる。筆者は、この様な意図のもとに重回帰分析ツールを開発したので、ここに報告する。

２．重回帰分析ツールの実装

２．１．Excel 上での実装

本重回帰分析ツールは、マイクロソフト社のExcel 2003上に実装する。

本ツールは、ユーザーインターフェースの提供、回帰分析結果の提示、視認性のよいグラフ作成など、重回帰分析に必要な処理をVBA（Visual Basic For Applications）[6）7）]で実装している。

ただし単回帰・重回帰分析それ自体は行っていない。Excelアドインで「分析ツールVBA」を有効にすると、Excelアドイン「ATPVBAEN.XLA」が組み込まれる。このアドインに単回帰・重回帰分析を実行するプロシージャーが含まれるので、シート上にデータを準備して単回帰・重回帰分析を実行する直前に、ATPVBAEN.XLA内の当該プロシージャーを起動しており、本ツールでは単回帰・重回帰実行部は作成していない。回帰分析が終了すると、本ツールの側で、結果を抽出して集積したりグラフ化等行う。

２．２．回帰分析の対象

試作する重回帰分析ツールの評価のため、使用するデータを決定した。これにより極めて具体的に、有効性・使い勝手を考慮しつつ重回帰分析ツールを実装し、評価することができることになると考える。

データは、都道府県別指標（データ）を使用した。基本的には平成11年度経済企画庁国民生活局編「新国民生活指標」[8）]を元に、人口、高齢化などに関わるデータを若干追加した[9）10）11）]。結果として、持家率など157種のデータが、全国47都道府県別に与えられている。多くの指標は人口比で与えられ、都道府県別の特徴を捉えることができる。本ソフトでは、シート名「元データ」シートに保持している。

２．３．ユーザーインターフェースの重視

具体的なユーザーインターフェースについては、次節「３．回帰分析の進め方」に本ソフトの使い方があるのでそれを参照することにして、ここでは特徴のみに触れる。

２．３．１．操作

本ソフトの操作は、Excelシートの行ボタン選択、列ボタン選択、マクロメニューの選択のみで実行可能としてある。例えば被説明変数の選択は、シート名「元データ」シートの列番号ボタンを選択する。また単回帰・単回帰グラフ化・重回帰・重回帰グラフ化・分析結果の新しいブックへの保存は、対応するマクロメニューの実行で行う。

２．３．２．説明変数の選択

Excelシートに説明変数の候補の一覧をリストアップしてある。一行で1つ説明変数の候補を示し、ユーザーの判断材料として、相関係数（絶対値のみ。重回帰の場合は重相関係数）、決定係数、回帰曲線の切片、偏回帰係数、切片のｐ値、偏回帰係数のｐ値が続いて表示される。ユーザーはこれの中から、説明変数を選択する。

２．３．３．グラフ化

グラフ化の留意事項を以下に記す。

①　グラフの種類など

単回帰では１つのグラフを、重回帰分析では使用した説明変数の数だけグラフを作成する。グラフは散布図、扱い易さから大きく広がったチャートとし、チャートオブジェクト（シート上に小さく張り付いたグラフの形式）は使用しない。

既に同一チャート名のチャートがある場合は、削除される。相関係数（重回帰の場合は重相関係数）、決定係数、回帰式（重回帰の場合は重回帰式）が表示される。

②　グラフ座標軸の最大値と最小値の自動設定

グラフの縦軸・横軸の最大値・最小値は、データの最大値と最小値に近い値で、かつきりの良い値を自動的に設定する。理由は、都道府県名の見易さのためである。軸の最大値・最小値がデータのそれと近いと、グラフのデータ間のスペースが広くなる。都道府県別指標（データ）を扱っているので、グラフ上の各データが何県のデータか表示がある方が、無い場合よりはるかに興味・共感が得られる。ただし47都道府県名の表示では県名が重なって表示され判読できないこともあり、データ間のスペースを広くして県名を見易くする必要があったのである。

２．４．例外データの除去

単回帰・重回帰分析において、例外的なデータは都道府県名のクリックで分析から除外できる。除外する都道府県の数に制限はない。また元に戻すには、都道府県名を再度クリックするだけでよい。これは、シート名「元データ」シートの第１列の都道府県名の書かれたセルで行う。セルをクリックすると、セルの色が赤になりその県は回帰分析から除外され、再度クリックするとセルが元の白になりその県は回帰分析に含められる。

２．５．分析結果の保存ツールの実装

単回帰・重回帰分析の結果のみを保存できるようにする。これは、情報量の大きい元データなどがメモリーを消費することから、回帰分析結果とグラフのみを保存するものである。結果を保存しては、別の選択肢即ち別の説明変数を使用して再度分析をすることがやりやすくなる。本ソフトは、実行中はデータを含め約1.3MBであるが、分析結果のみだと約410kBである。

２．６．コード量

VBAのコードで約700行で実装した。説明変数６個までの重回帰分析が行える。シート名「元データ」のシートに、都道府県名のセルのクリックで起動され、例外データを除外するため、当該都道府県名のセルの色を赤または白に変更するコードが約50行含まれる。その他は、標準モジュール内に約650行あり、ほぼ全ての機能を受け持っている。

３．回帰分析の進め方

３．１．単回帰分析の進め方

３．１．１．被説明変数の選択（都道府県別各種指標の選択）

シート名「元データ」シートには、都道府県別各種指標が157種含まれている。下の図３－１に例を示す。都道府県毎にデータが示されている。１・２行にデータの説明がある。第２列から第132列までは「新国民生活指標」[8]のデータであり、第133列から第158

	1	2	3	4	5	6	7	8	9	145	146	147	148	149
1		住む.1	住む.2	住む.3	住む.4	住む.5	住む.6	住む.7	住む.8	自殺率	不慮の事故死亡率	年少人口	第1次産業	老人人口
2		危険･修理不能住宅比率(%)	最低居住水準以上住宅比率(%)	借家の1畳当たり実質家賃(円)	持家比率(%)	公害苦情受理件数(人口十万人比)	重要刑法犯罪認知件数(人口十万人比)	重要窃盗犯認知件数(人口十万人比)	交通事故発生件数(人口十万人比)	自殺死亡率(10万人当たり)平成11年	不慮の事故死亡率(10万人当たり)平成11年	2005(推計人口)15歳未満割合	第1次産業就業割合(%)	2005(推計人口)65歳以上割合
3	北海道	7.67	94.3	1510	54.0	5.0	7.6	206	451.8	26.2	30.8	13.4	9.0	20.7
4	青森	7.80	95.9	1480	71.6	32.8	4.5	103	558.2	32.5	38.6	14.3	16.9	22.1
5	岩手	6.93	95.0	1643	72.8	22.0	6.3	115	388.9	34.4	38.7	14.6	16.7	23.8
6	宮城	7.26	93.9	2257	60.7	34.3	9.1	272	471.1	24.8	31.4	14.5	8.2	18.7
7	秋田	5.92	97.1	1554	79.6	23.1	9.1	88	364.3	40.6	45.1	13.1	13.1	26.0
8	山形	5.75	96.8	1701	79.2	31.0	5.6	137	496.6	26.2	40.7	14.5	12.9	24.5
9	福島	7.11	93.7	1811	68.6	25.1	6.5	232	618.6	25.8	35.1	15.4	10.8	21.6
10	茨城	5.36	93.2	2063	70.5	41.8	12.5	245	732.6	23.5	37.7	15.1	9.4	18.0
11	栃木	5.19	93.0	2123	69.2	58.4	8.5	209	700.2	25.0	31.0	14.8	8.4	18.5
12	群馬	5.99	93.3	1992	70.4	51.0	6.7	275	854.7	24.7	35.0	15.0	7.9	19.8
13	埼玉	4.13	90.1	3085	61.9	59.2	15.0	246	596.9	22.8	22.2	15.1	2.8	15.4
14	千葉	4.40	91.2	2908	61.1	40.7	13.1	342	535.4	21.0	26.6	14.1	4.6	16.6
15	東京	6.86	78.8	4379	39.6	55.4	15.7	366	528.6	24.3	21.5	11.5	0.5	19.5
16	神奈川	4.61	87.8	3507	51.5	43.2	12.6	194	749.8	22.8	22.9	13.9	1.2	16.6
17	新潟	5.40	95.7	1847	76.9	28.9	9.8	164	542.9	33.7	44.4	14.4	9.1	22.8
18	富山	4.15	96.7	1874	79.8	17.0	6.2	281	688.2	30.9	44.6	13.8	5.6	22.5
19	石川	3.85	95.6	2041	69.9	28.3	9.3	190	736.9	22.2	39.2	14.6	5.4	20.0
20	福井	4.61	95.7	1640	76.5	31.0	6.0	184	555.4	22.7	45.7	15.3	6.5	22.0

図３－１　シート名「元データ」シート（都道府県別各種指標）の一部抜粋

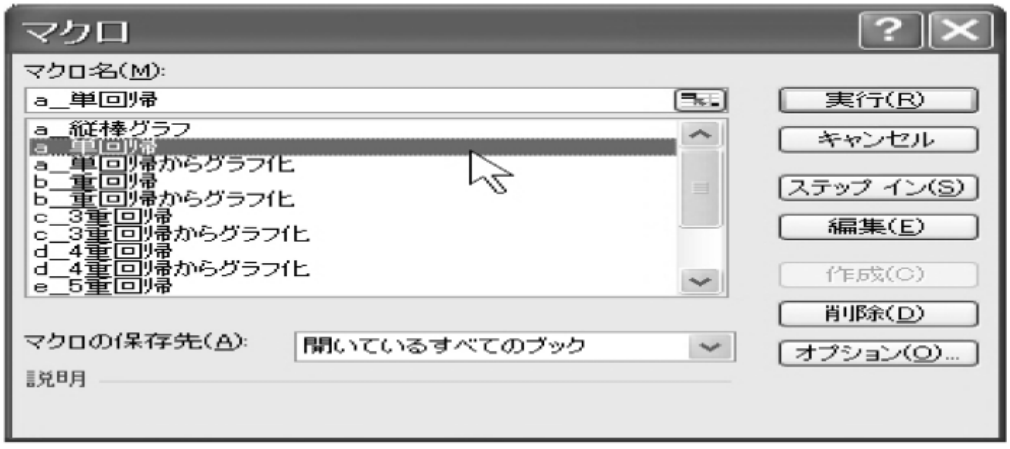

図３－２　マクロメニュー

列までは「データでみる県勢」[9)][10)]等から補ったものである。回帰分析の被説明変数の選択または縦棒グラフの対象の選択は、対応する列番号ボタンの選択で行う。

３．１．２．マクロメニュー

「元データ」シートで列が選ばれた後は、マクロメニューから選択して、縦棒グラフ作成か単回帰分析を選ぶ。マクロメニューを前頁の図３－２に示す。表示された他、「５重回帰からグラフ化」、「６重回帰」、「６重回帰からグラフ化」、「新しいブックに保存」メニューを含む。マクロメニューは、事前に列ボタン又は行ボタンの選択で適当な選択がされた後、メニューを選択して実行する。

３．１．３．縦棒グラフ

図３－３は、「元データ」シートで第５列（持家率1993年データ）の列ボタンを選択し、マクロメニューで「縦棒グラフ」を実行し作成したものである。チャート名「棒G」のチャートが作成される。同一チャート名のチャートがあると、旧い方は削除される。

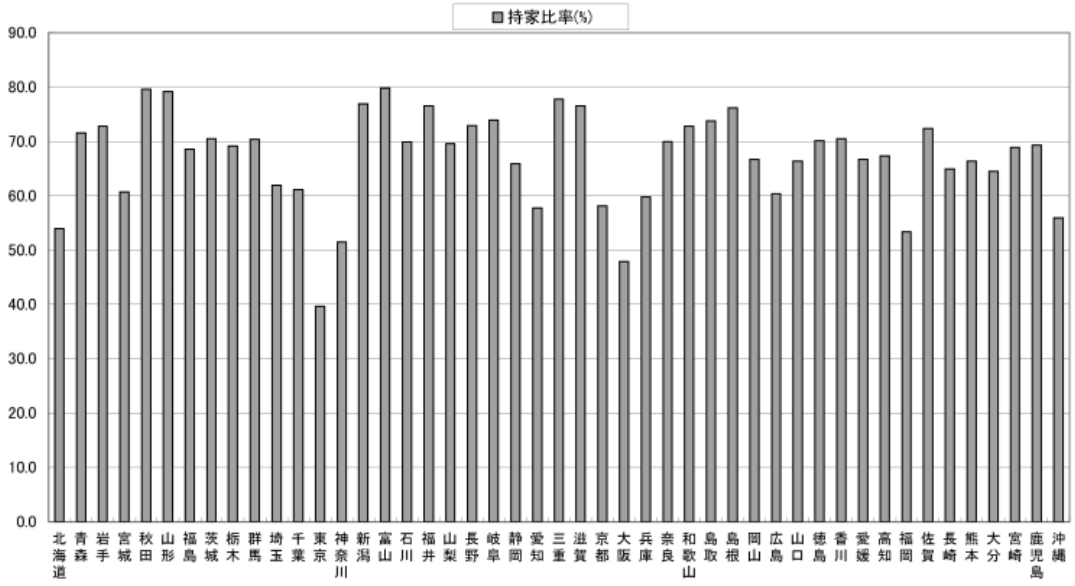

図３－３　都道府県別の持家率（1993年）

３．１．４．単回帰分析の説明変数候補一覧の作成

次頁の図３－４は、「元データ」シートで第５列（「持家率」の列）の列ボタンを選択し、マクロメニューで「単回帰」を実行した後のシート名「結果」シートの一部である。被説明変数は「持家率」と固定したまま、説明変数の候補の一覧が表示される。シート名「元データ」シートにある157種の都道府県別各種指標（データ）から「持家率」を除く全ての指標が決定係数で降順ソートされて表示される。各行が説明変数の候補を示し、156種の指標の一部が表示されている。第４列が説明変数候補となる指標のタイトルであ

り、他に相関係数（絶対値）、決定係数、回帰式、回帰式係数のｐ値も表示される。ユーザーはこの中から、説明変数として適当なものをユーザー自身の判断で選択する。

	1	2	3	4	5	6	7	8	9	10
1	列番号	被説明変数	列番号1	説明変数1	相関	決定係数	切片	係数1	切片p値	係数1p値
2	5	持家比率(%)	3	最低居住水準以上住宅比率(%)	0.859	0.7379	-139.7	2.21897	1.32E-09	1.13E-14
3	5	持家比率(%)	142	世帯数(千)	0.8266	0.6833	73.96	-0.0069	1.76E-48	8.22E-13
4	5	持家比率(%)	122	未婚率(%)	0.8261	0.6824	127.75	-2.3494	1.68E-24	8.76E-13
5	5	持家比率(%)	13	最寄りの医療機関までの距離500m未満住	0.8119	0.6592	96.089	-0.5759	2.16E-31	4.36E-12
6	5	持家比率(%)	141	人口(千人)	0.8076	0.6522	74.693	-0.0029	1.59E-46	6.93E-12
7	5	持家比率(%)	120	離婚率(人口千人比)	0.7868	0.6191	106.8	-24.126	3.11E-26	5.5E-11
8	5	持家比率(%)	143	1世帯あたり人員(人)	0.7729	0.5974	-8.351	26.8193	0.371981	1.94E-10
9	5	持家比率(%)	151	自動車(台数/千人)1998	0.7636	0.5831	20.489	0.07676	0.001201	4.31E-10
10	5	持家比率(%)	14	誘導居住水準以上住宅比率(%)	0.7494	0.5616	33.615	0.73196	1.87E-09	1.36E-09
11	5	持家比率(%)	135	人口密度(可住地面積k㎡当り)	0.7453	0.5555	72.613	-0.0038	7.68E-46	1.87E-09
12	5	持家比率(%)	146	不慮の事故死亡率(10万人当たり)平成11年	0.7421	0.5507	34.562	0.89422	7.82E-10	2.38E-09
13	5	持家比率(%)	121	婚姻率(人口千人比)	0.7395	0.5468	127.21	-10.421	1.18E-19	2.9E-09
14	5	持家比率(%)	157	自動車(台数/可住地面積k㎡当り)1998	0.734	0.5387	74.95	-0.0102	2E-42	4.36E-09
15	5	持家比率(%)	16	1人当たり畳数(畳)	0.7321	0.536	9.4878	5.21836	0.243591	4.99E-09
16	5	持家比率(%)	20	下水道等普及率(%)	0.727	0.5285	84.446	-0.3594	7.74E-33	7.2E-09
17	5	持家比率(%)	110	上級学校学生数(人口総数に対する%)	0.7163	0.5131	78.345	-4.4796	1.1E-37	1.5E-08
18	5	持家比率(%)	155	人口密度常用対数(可住地面積k㎡当り)	0.7068	0.4995	127.11	-19.879	3.93E-18	2.82E-08
19	5	持家比率(%)	117	社会教育関係職員数(人口1万人比)	0.7062	0.4987	47.631	5.00337	7.58E-20	2.93E-08
20	5	持家比率(%)	4	借家の1畳当たり実質家賃(円)	0.7022	0.4931	88.557	-0.0108	6.82E-29	3.78E-08
21	5	持家比率(%)	34	サービス支出割合(%)	0.689	0.4747	154.1	-2.2019	1.11E-14	8.59E-08
22	5	持家比率(%)	126	老人クラブ加入率(60歳以上人口比%)	0.6886	0.4742	41.32	0.69881	5.63E-13	8.79E-08
23	5	持家比率(%)	42	転職率(%)1992	0.6843	0.4683	112.68	-10.648	1.43E-19	1.14E-07
24	5	持家比率(%)	136	可住地面積半径(一世帯当り)	0.6733	0.4534	48.378	0.60308	2.91E-19	2.15E-07
25	5	持家比率(%)	28	負債年収比(倍)	0.6729	0.4528	97.949	-55.311	3.98E-23	2.2E-07
26	5	持家比率(%)	108	留学者数(15歳以上人口1万人比)	0.6541	0.4279	77.49	-0.8861	9.19E-36	6.15E-07

図３－４　単回帰分析の結果の決定係数降順ソート済み説明変数候補一覧の抜粋

３．１．５．単回帰分析のグラフ化

単回帰分析を実行し、ソート済み説明変数候補一覧がえられたら、その中から説明変数をユーザーが判断して決定する。被説明変数が「持家率」の場合、決定係数最大の「最低居住水準以上の住宅比率」は、持家率が高いと当然この変数も高くなると思われ、持家率が高い原因よりもその結果と判断し、説明変数には採用しない。２番目に決定係数の大きい「世帯数」であるが、各都道府県世帯数の絶対値で面積比率などを指しておらず説明変数に採用しない。３番目に決定係数の大きい「未婚率」であるが、単身者が多いと「持家率」が低い原因となりうるので、説明変数として採用する。

そこで、図３－４にあるように、ソート済み説明変数候補一覧のあるシート名「単回帰結果」シートで、第４行の行ボタンを選択することで、これを説明変数とすることをソフトに知らせておく。マクロメニューから「単回帰からグラフ化」を選択し、次頁の図３－５を得る。図３－５で「◆」は各都道府県の持家率の観測値、「○」は予測値を示す。予測式（回帰直線）は、図の上に「ｙ＝－2.349X1+127.746」が表示される。予測値は、この式の説明変数X1に、各都道府県の「未婚率」の観測値を代入して求められる。決定係数の値0.682、相関係数の絶対値0.826が、併せてグラフの上に示される。

回帰分析の他の各種統計情報は、シート名「作業」シートにそのまま残っているので、適宜参照できる。回帰分析自体はExcelのアドインの分析ツールの回帰分析をそのまま使用しているので、「作業」シートへの出力はExcelの出力そのままのものである。

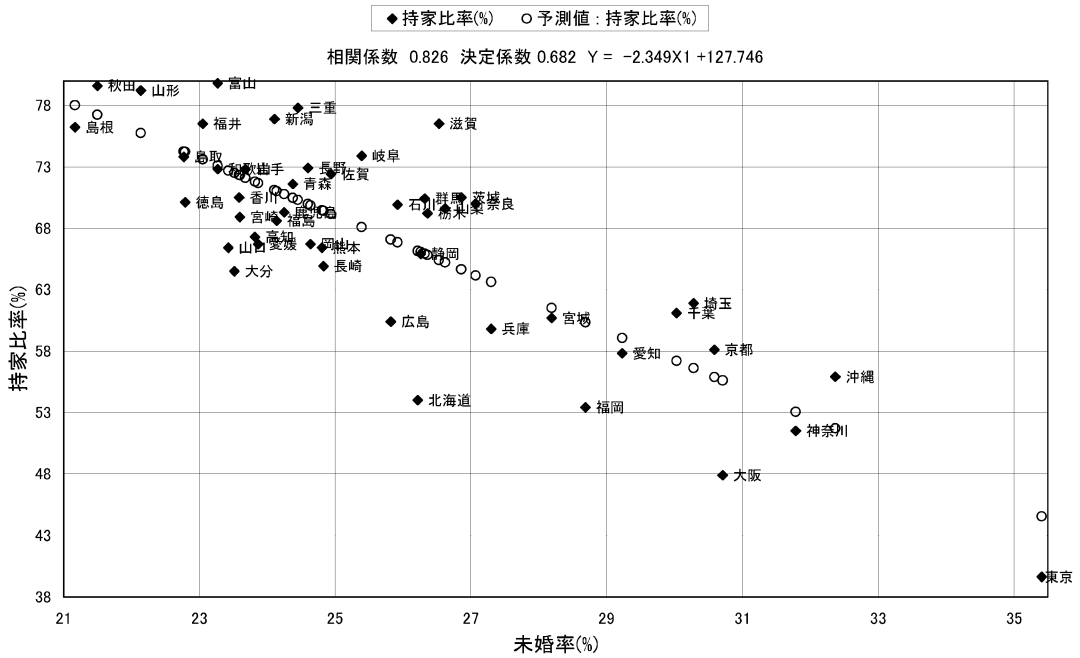

図３－５　単回帰の場合の持家比率（1993年）と未婚率（1995年）の散布図

３．２．重回帰分析の進め方

３．２．１．重回帰分析の説明変数候補一覧の作成

まず、単回帰分析後に重回帰分析を実行する手順の説明をする。

単回帰分析を実行し、ソート済み説明変数候補一覧のシートから説明変数をユーザーが判断して決定し、対応する行の行ボタンを選択する。被説明変数が「持家率」で説明変数を「未婚率」とすると、第４行行ボタンを選択する。次にマクロメニューから「重回帰分析」を選択し、次頁の図３－６を得る。シート名「重回帰結果」シートに、被説明変数は「持家率」、第１説明変数を「未婚率」と固定したまま、第２説明変数の候補の一覧が表示される。「元データ」シートの157種の都道府県別各種指標（データ）から「持家率」と「未婚率」を除く155種の指標が決定係数で降順ソートされ表示される。ユーザーはこの中から、第2説明変数として適当なものをユーザー自身の判断で選択する。

次に説明変数がｎ個（３≦ｎ≦６）の重回帰分析の説明をする。ｎ個の説明変数の重回帰は、「ｎ重回帰」とこのソフトでは呼んでおり、メニュー表示、シート名、チャート名で使われる。説明変数が２個の重回帰は、２重回帰ともいえることになる。ｎ重回帰は、（ｎ－１）重回帰のシート名「（ｎ－１）重回帰結果」シートのソート済み第（ｎ－１）説明変数候補一覧から、第（ｎ－１）説明変数をユーザーが判断して決定し対応する行の行ボタンを選択する。次に、マクロメニューから「ｎ重回帰」を選択する。すると、被説明変数、第１～（ｎ－１）説明変数を固定し、元データシートにある157種の

都道府県別各種指標（データ）から、被説明変数、第１～（ｎ－１）説明変数を除く全ての指標が、第ｎ説明変数候補として、決定係数で降順ソートして表示される。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	列番号	被説明変数	列番号1	説明変数1	列番号2	説明変数2	相関	決定係数	切片	係数1	係数2	切片p値	係数1p値	係数2p値
2	5	持家比率(%)	122	未婚率(%)	143	1世帯あたり人員(人)	0.938	0.88	62.23	-1.699	17.33	6E-09	4E-13	8E-11
3	5	持家比率(%)	122	未婚率(%)	120	離婚率(人口千人比)	0.9	0.81	130.3	-1.568	-13.8	8E-29	4E-08	2E-06
4	5	持家比率(%)	122	未婚率(%)	79	看護婦数(人口十万人比)	0.897	0.805	159.9	-3.015	-0.02	5E-24	6E-17	4E-06
5	5	持家比率(%)	122	未婚率(%)	31	生活保護世帯割合(総世帯数に占める	0.895	0.802	130.9	-2.24	-4.43	2E-28	5E-15	6E-06
6	5	持家比率(%)	122	未婚率(%)	127	献血者数(15～64歳人口比%)	0.893	0.798	161.3	-2.945	-2.44	4E-23	1E-16	9E-06
7	5	持家比率(%)	122	未婚率(%)	29	個人破産件数(人口1万人比)	0.89	0.791	136.9	-2.401	-1.35	9E-28	9E-16	2E-05
8	5	持家比率(%)	122	未婚率(%)	13	最寄りの医療機関までの距離500m未満	0.888	0.788	120.5	-1.435	-0.32	1E-25	6E-06	3E-05
9	5	持家比率(%)	122	未婚率(%)	126	老人クラブ加入率(60歳以上人口比%)	0.887	0.787	100.4	-1.828	0.377	2E-16	4E-10	3E-05
10	5	持家比率(%)	122	未婚率(%)	150	人口増加率(1985～90)	0.887	0.787	144.8	-3.071	6.712	5E-26	1E-15	3E-05
11	5	持家比率(%)	122	未婚率(%)	76	一般病院病床数(人口十万人比)	0.883	0.78	148.7	-2.712	-0.01	1E-24	5E-16	6E-05
12	5	持家比率(%)	122	未婚率(%)	144	1世帯当たり家計所得(万円:参考値)	0.883	0.779	114.6	-2.573	0.026	9E-23	6E-16	7E-05
13	5	持家比率(%)	122	未婚率(%)	73	入院患者率(一般病院在院患者数/人	0.882	0.779	147.2	-2.699	-0.01	7E-25	5E-16	7E-05
14	5	持家比率(%)	122	未婚率(%)	78	医師数(人口十万人比)	0.88	0.775	144.5	-2.438	-0.07	3E-25	2E-15	0.0001
15	5	持家比率(%)	122	未婚率(%)	80	養護・軽費老人ホーム定員数(65歳以	0.877	0.768	149.3	-2.837	-0.16	2E-23	3E-15	0.0002
16	5	持家比率(%)	122	未婚率(%)	3	最低居住水準以上住宅比率(%)	0.877	0.768	-44.9	-0.963	1.469	0.3026	0.0204	0.0002
17	5	持家比率(%)	122	未婚率(%)	142	世帯数(千)	0.876	0.767	105	-1.317	-0	4E-17	0.0003	0.0002
18	5	持家比率(%)	122	未婚率(%)	149	2005(推計人口)65歳以上割合	0.873	0.762	192.1	-3.537	-1.6	5E-14	4E-12	0.0004
19	5	持家比率(%)	122	未婚率(%)	151	自動車(台数/千人)1998	0.871	0.759	86.07	-1.623	0.038	2E-08	1E-06	0.0006
20	5	持家比率(%)	122	未婚率(%)	17	1住宅当たり敷地面積(㎡)	0.87	0.756	103.7	-1.9	0.042	2E-15	9E-10	0.0007
21	5	持家比率(%)	122	未婚率(%)	21	リサイクル率(%)	0.868	0.753	121.1	-2.329	0.592	3E-24	4E-14	0.0009
22	5	持家比率(%)	122	未婚率(%)	141	人口(千人)	0.865	0.748	107.8	-1.424	-0	5E-17	0.0002	0.0015
23	5	持家比率(%)	122	未婚率(%)	117	社会教育関係職員数(人口1万人比)	0.865	0.747	104.5	-1.791	2.28	4E-15	5E-08	0.0016
24	5	持家比率(%)	122	未婚率(%)	65	0-1歳児保育在所者数(対象世帯100	0.862	0.743	127.6	-2.122	-0.65	9E-26	7E-12	0.0025
25	5	持家比率(%)	122	未婚率(%)	18	最寄りの交通機関1km未満住宅比率(%	0.862	0.742	149.6	-2.115	-0.35	8E-21	8E-12	0.0025
26	5	持家比率(%)	122	未婚率(%)	147	2005(推計人口)15歳未満割合	0.861	0.742	96.66	-2.358	2.147	6E-11	5E-14	0.0026

図３－６　重回帰分析の結果の決定係数降順ソート済み説明変数候補一覧

（第１説明変数は「未婚率」として、第２説明変数の候補である）

３．２．２．重回帰分析のグラフ化

まず、説明変数が２個の場合の重回帰分析のグラフ化の説明をする。

重回帰分析を実行し、ソート済み第２説明変数候補一覧が得られたら、その中から第２説明変数をユーザーが判断して決定する。被説明変数が「持家率」、第１説明変数「未婚率」の場合、決定係数最大の「一世帯当り人員」をここでは選択する。「一世帯当り人員」が多ければ広い家を必要とし「持家率」を高めると思われ、「持家率」の値の原因となりうるので、説明変数としての採用は妥当と思われる。第２行の行ボタンを選択し、マクロメニューから「重回帰からグラフ化」を実行する。すると、チャート名「重回帰Ｇ１」、「重回帰Ｇ２」の２つのチャートが得られる。

次頁の図３－７と図３－８は、各チャート名「重回帰Ｇ１」と「重回帰Ｇ２」である。「◆」は持家率の観測値で、「○」は予測値を示す。予測値は、単回帰の場合と異なり直線上には並ばない。図の上に、「相関係数　0.938　決定係数 0.88　Ｙ＝－1.699Ｘ1＋17.334Ｘ2＋62.227」と表示されているが、式の説明変数X1に「未婚率」、説明変数X2に「一世帯当り人員」の各都道府県の観測値が代入され予測値が求められる。図のグラフ横軸に示された「未婚率」または「一世帯当り人員」以外の観測値が代入されて予測値が得られているので、直線上に並ばないのである。相関係数は0.93、即ち93.8%と非常に高い相関が見られる。なお、相関係数とあるが、重回帰なので重相関係数を意味している。

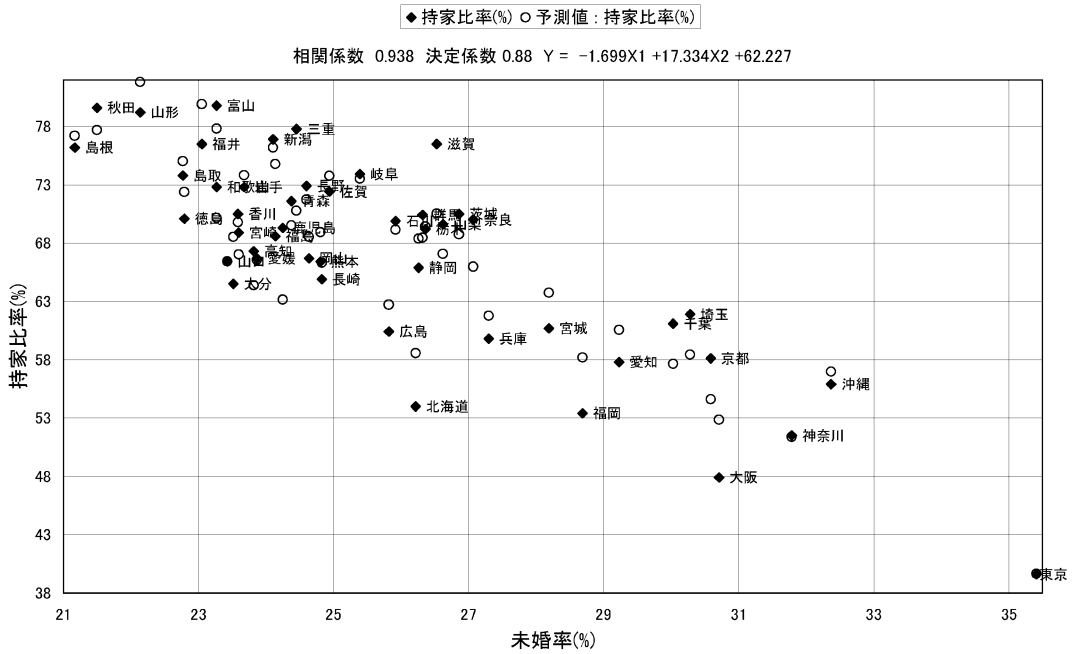

図３－７　重回帰（２説明変数）の持家比率（1993年）と未婚率（1995年）の散布図

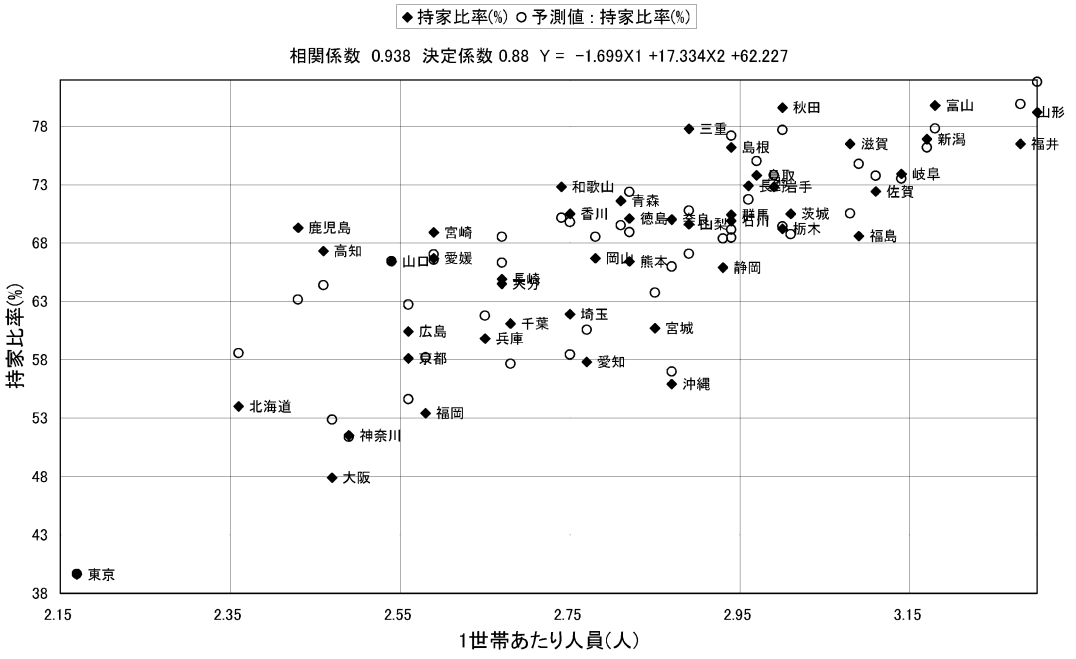

図３－８　重回帰（２説明変数）の持家比率（1993年）と１世帯あたり人員（2000年）の散布図

重回帰分析の結果の他の各種統計情報は、シート名「重回帰作業」シートにExcelの出力がそのまま残っている。例えば、自由度調整済み決定係数は、「決定係数Ｒ２」と表示されて、その値が示されている。

次に説明変数がｎ個（３≦ｎ≦６）の重回帰分析のグラフ化の説明をする。ｎ重回帰のグラフ化は、ｎ重回帰のシート名「ｎ重回帰結果」シートのソート済み説明変数候補一覧から、第ｎ説明変数をユーザーが判断して決定し、対応する行の行ボタンを選択し、マクロメニューから「ｎ重回帰からグラフ化」を実行する。すると、チャート名「ｎ重回帰G1」～「ｎ重回帰Gn」までｎ枚のチャートが作成される。ｎ重回帰分析の結果の各種統計情報は、シート名「ｎ重回帰作業」シートにExcelの出力がそのまま残っている。

４．説明変数の選択法についての考察

次に、重回帰分析で、説明変数の選択のしかたについて議論する。「自殺率」を例として、その選択方法を考察する。筆者は「自殺率」は、「持家率」と異なり、決定係数が高くかつ直感的に因果関係を認めうる説明変数の選択が難しい例であると考えている。

４．１．自殺率について

自殺率のデータは、平成11年の都道府県別自殺率である。この年、自殺率は多い順に秋田県、岩手県、青森県、新潟県、富山県、島根県、宮崎県の順である。日本の自殺率の高さは、ＯＥＣＤ諸国中最高位に位置する。世界的にみると旧共産圏の東欧諸国が高いが、大雑把にみてみると、北半球の高緯度で高く、赤道付近の低緯度では自殺率は低くなる。緯度の高低は、冬季日照時間の多寡をもたらし、冬季日照時間の少ない地方では自殺率が高いように思われる。筆者は、旧共産圏の東欧諸国の異常に高い自殺率は、長く続く経済破綻、それと高緯度地方で冬季日照時間の少なさが関係すると推察する。

日本国内の自殺率については、富山県内の新聞で平成18年頃から取り上げられるなど、関心は少しずつ高まってきている。以後、具体的に説明変数選択の仕方について記す。

４．２．物理的な因果関係が類推できる説明変数を優先

「自殺率」を単回帰分析すると、決定係数が９番目、15番目、18番目に大きい説明変数の候補として、「１月日照時間常用対数」、「年間日照時間」、「１月日照時間」が見出される。決定係数は９番目のもので約0.283であり、第１位の説明変数候補の0.435よりずっと小さい。しかしながら、日照時間が短いと人により気分に影響があるという精神保健知識は今や常識であり、物理的な量ではないが精神保健上の明白さから、説明変数として採用する。また対数は、刺激－反応理論で刺激の対数の使用はよくあるので、第１説明変数は、「１月日照時間常用対数」（日照時間は1971～2000年平均値）を採用することとする。次頁の図４－１に単回帰の結果を示す。なお日照時間のデータは理科年表[11]に拠り、主に県庁所在地のデータだが、埼玉県は熊谷市であるなど例外もある。

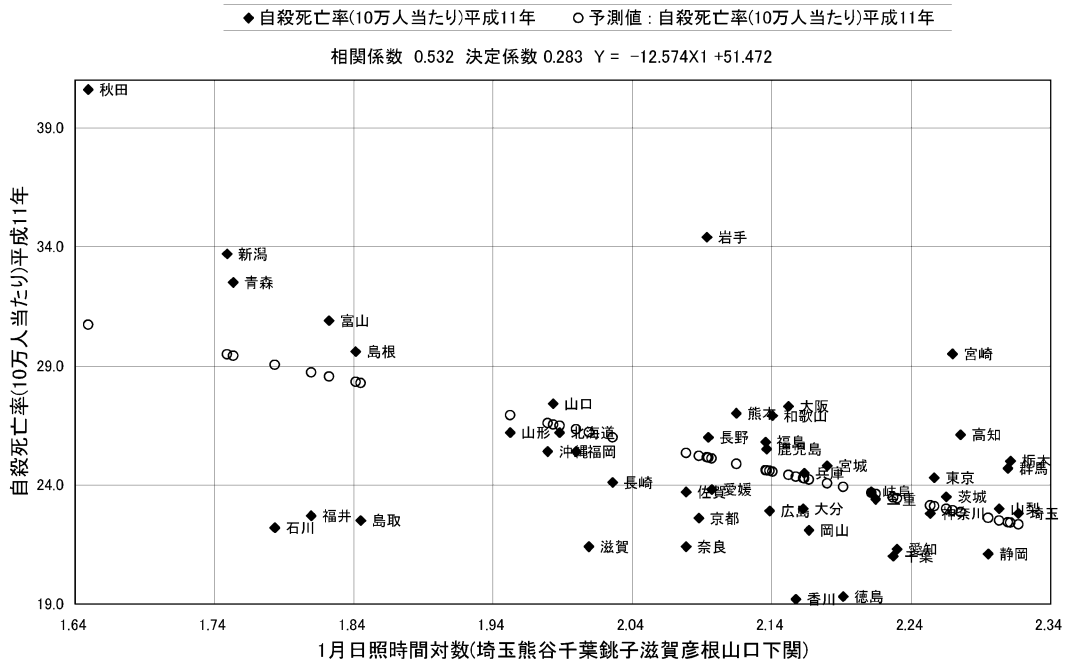

図４－１　自殺率（1999年）と１月日照時間（1971～2000年平均値）常用対数の散布図

４．３．基本的統計量でない、社会構造・文化等に関わる説明変数は単純なものを優先

被説明変数を「自殺率（平成11年）」、第1説明変数を「１月日照時間常用対数」として重回帰の説明変数候補を出すと、決定係数の降順で１番目「児童・生徒1人当たり校地面積（平方m）」、２番目「パソコン普及率（%）」、３番目「大学等進学率（%）」と説明変数の候補が出る。

これらは、社会統計の人口密度、老人人口比率などの様に意味するものがクリアーな基本的な指標ではなく、意味するものが比較的曖昧なまたは抽象的なデータといえよう。筆者は、１番目の「児童・生徒1人当たり校地面積（平方m）」は、教育環境の豊かさは意味せず、これが大きいと過疎と少子化の進行した痛ましい地方の現実を意味し、２番目「パソコン普及率（%）」の指標は、本格的普及開始の1994年のデータであり、先取の気性の強い風土を意味し、３番目「大学等進学率（%）」は、将来への投資をできるだけの豊かさを意味すると考える。

筆者は、社会統計の基本的データでない場合、直接の関係がわかりにくいのでなるべく説明変数として採用しないことにするが、それでも決定係数の上位に他に採用すべきデータが無い場合には、比較的単純に理解可能なものを説明変数とすることとする。筆者は、「大学等進学率（%）」は将来投資の可能な豊かさと現世利益重視の風土を意味し、それが高いと自殺に対し抑制的に作用すると考え、第2説明変数として採用することにとする。次頁の図4－2に重回帰の結果を、追加した第2説明変数に関する分のみ示す。

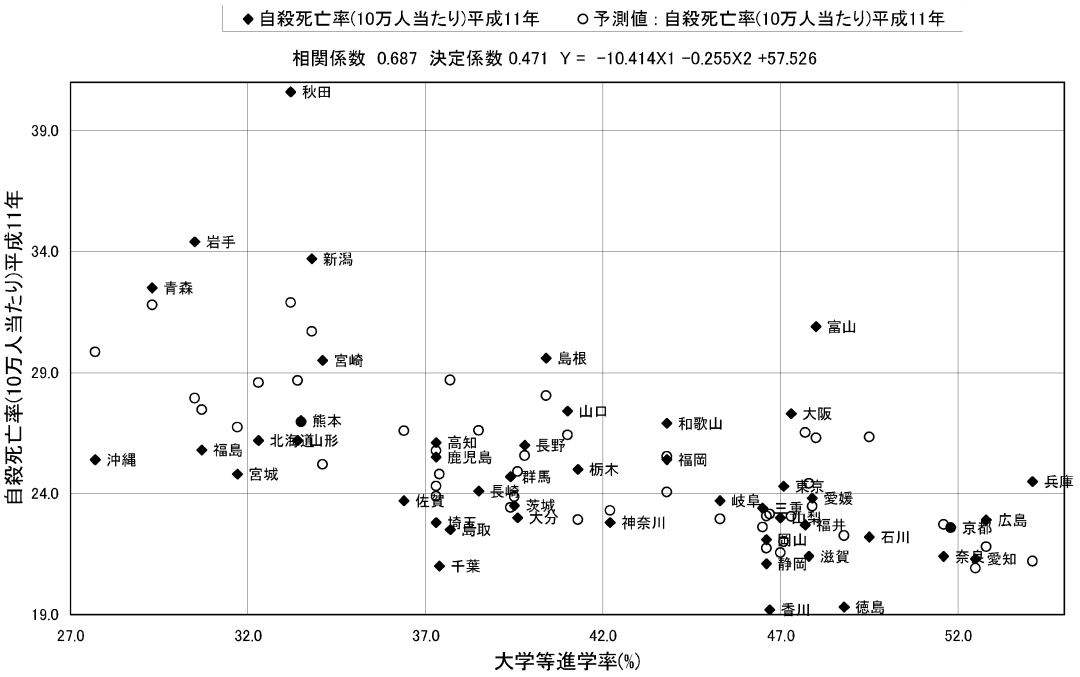

図４－２　自殺率（1999年）と大学等進学率（1997年）の散布図

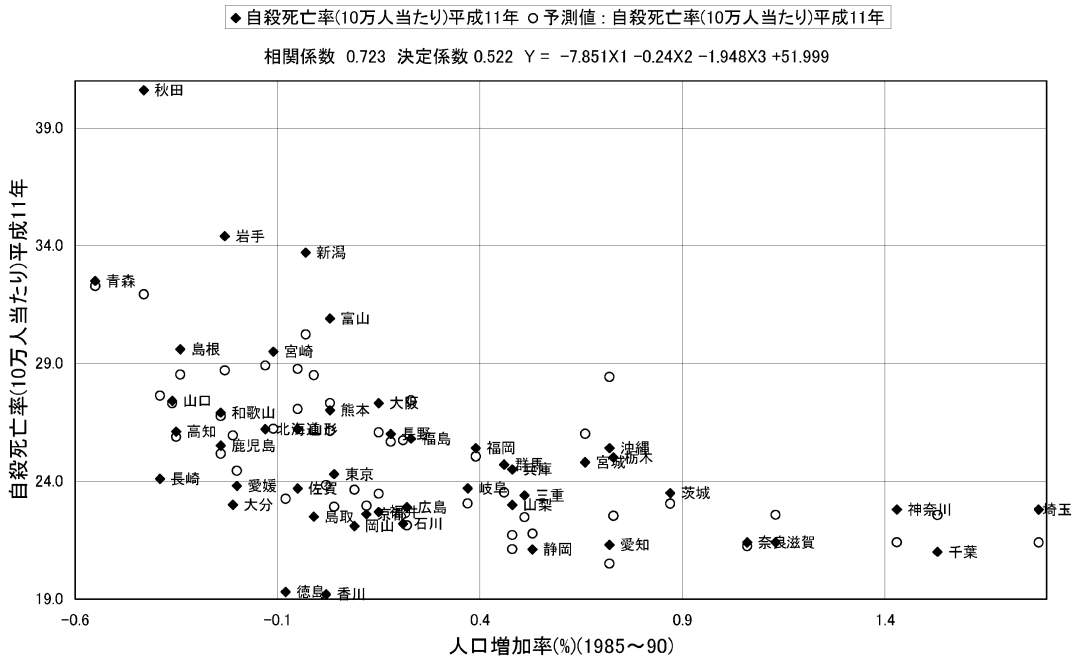

図４－３　自殺率（1999年）と人口増加率（%）（1985～90年平均）の散布図

４．４．社会的統計量は基本的で因果関係が類推しやすい説明変数を優先

被説明変数を「自殺率（平成11年）」、第１説明変数「１月日照時間常用対数」、第２説明変数「大学等進学率（%）」として、３重回帰の第３説明変数候補を出すと、決定係数の降順で、６番目「人口増加率（%）（1985～90年平均）」、８番目「2005年65歳以上人口割合（1997年推計値）」[9]が出る。人口増加率、65歳以上割合とも、基本的な社会統計量であり、自殺とは独立した統計量である。人口増加率は、負の地域は９～14年前に若者の流出、頼れる親戚等の減少等を経験し、自殺との関連性が窺える。老人割合の方は体力・気力の衰え等自殺率とより直截的に関係しそうに思われる。ここでは、決定係数を優先し、人口増加率（%）を採用する。前頁の図４－３に３重回帰の結果を、追加した第３説明変数に関する分のみ示す。自殺率と人口増加率（%）は負の相関がある。

４．５．特定の分野・業界に関わる複数の指標は１つを代表として説明変数に採用

被説明変数と第１～第３説明変数を固定し、４重回帰の第４説明変数候補を出すと、決定係数の降順で、１番目に「デイサービスセンター利用状況（65歳以上人口百人比）」が出る。福祉関係では他に決定係数20位以内に、特別養護老人ホーム定員数、ショートステイ利用状況、養護・軽費老人ホーム定員数、身体障害者ホームヘルパー派遣世帯数、身体障害者更正援護施設定員数、老人福祉施設従事者数が含まれる。福祉関係と自殺率とは関連があると判断される。福祉関係で決定係数最大の「デイサービスセンター利用状況」を、代表として第４説明変数に採用する。図４－４に４重回帰の結果を、追加した第４説明変数「デイサービスセンター利用状況」に関する分のみ示す。第４説明変数の偏回帰係数は負で負の相関があり、図ではわかりにくいが、右下がりである。

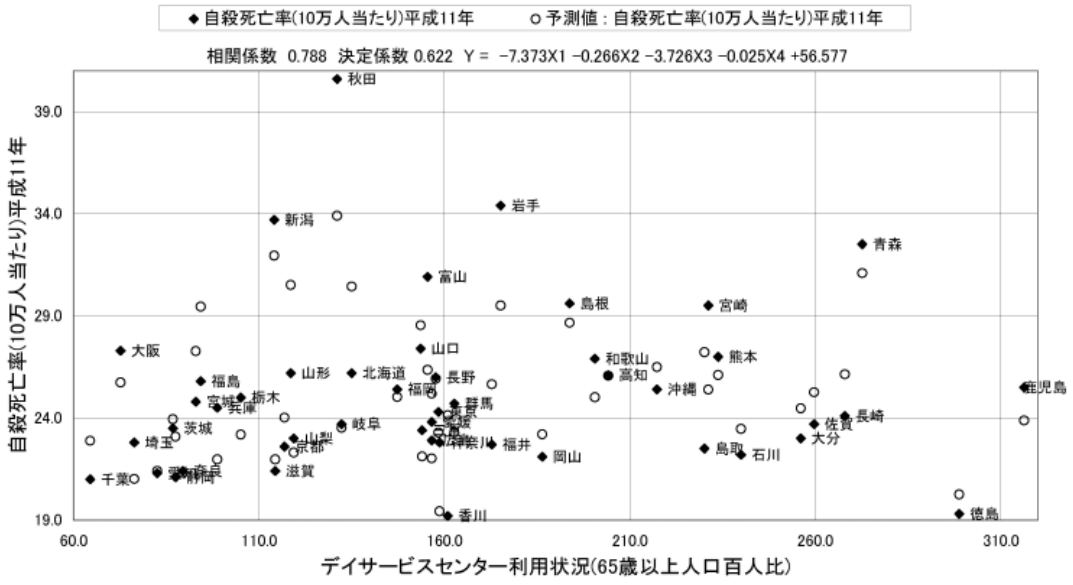

図４－４　自殺率（1999年）とデイサービスセンター利用状況

４．６．説明変数となるべきデータの不足・不備の場合の代替指標の採用

被説明変数と第１～第４説明変数を固定し、５重回帰で第５説明変数候補を出すと、決定係数の降順で、２番目に「有効求人倍率（倍）（1997）」が出る。平成11年の自殺率には平成の長期不況の影響が考えられ、自殺の原因となりうるので、説明変数として「有効求人倍率（倍）」を採用する。図4－5に追加した第５説明変数に関する分のみ示す。「有効求人倍率（倍）」の偏回帰係数は負で、図は全体としては右下がりである。

「失業率」そのものは100番目以下にやっと出る。特にダメージの強い失業を反映する「一家の主たる家計支持者の失業率」は北陸（富山、石川、福井、新潟の４県）としてしか統計になく、都道府県別指標（データ）として存在しない。このため失業率関係が上位に現れないと推量する。「有効求人倍率（倍）」は、「失業率」データそのものより就職難の実態を反映し、自殺率の回帰分析では失業関係指標を代替していると考えられる。

なお６重回帰の説明変数候補は、生活保護世帯割合（総世帯数に占める割合%）である。

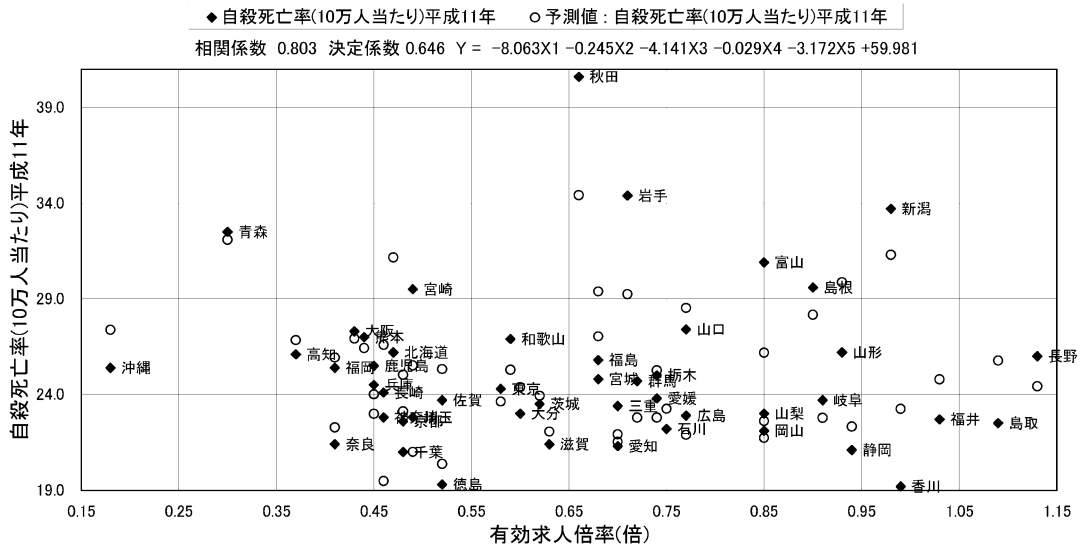

図４－５　自殺率（1999年）と有効求人倍率（倍）（1997年）の散布図

5. まとめ

既に述べたように、ユーザーの判断を重視した重回帰分析ツールを開発し、Excel 2003上にＶＢＡ約700行で実装した。本ソフトウェアは、ユーザーインターフェースを重視し、Excelの行ボタン、列ボタン選択とマクロダイアローグからマクロを実行するだけで操作が可能である。結果はビジュアルなグラフ（チャート）化し、詳細はExcelシートで表示する。印刷等Excelの全機能が使え、ユーザーインターフェースの良好なツールである。既述のとおりである。実際、筆者の重回帰分析に要する時間は短縮された。

説明変数の選択に際しては、説明変数の候補を決定係数の高い順に全て表示するので、その中からユーザーが判断して決定することになる。その説明変数の選択にあたっては、一定の方針が必要となる。「４．説明変数の選択法についての考察」で記したとおりである。

６．今後の課題

６．１．説明変数選択の方法の明確化

本ソフトは、重回帰分析を考えており、比較的大量のデータから説明変数を探している。自動的に求める変数増加・減少法と違い、説明変数の候補一覧を決定係数の大きい順にソートし、ユーザー自身がそれから判断するのである。「４．説明変数の選択法についての考察」で説明変数の選択方法を記したが、157種類の都道府県別指標の場合、被説明変数と使用中の説明変数を除く約150の候補からの選択となり、実際には難しい問題を含んでいる。自殺率の例では「人口増加率（％）（1985～90年平均）」を採用したが、「2005年65歳以上人口割合（1997年推計値）」の方が決定係数は若干低いが、自殺との関連が直感で理解し易い。既に記した以上に明快な説明変数選択基準の設定が望まれる。

６．２．パス解析の導入の検討

本ソフトでは、157種類の都道府県別指標（データ）を用い説明変数６個までの重回帰分析を行った。この重回帰分析では被説明変数と説明変数の計７個の関係は表しうるが、他の150種類程度の指標との関係は表現できない。パス解析を使えば、他の都道府県別指標と被説明変数との関係が表現できる可能性が高くなるので検討に値する[5)]。

６．３．特定都道府県別指標（データ）等の更新

用いたデータの多くは、「平成11年版　新国民生活指標」[8)]を元にしており、それ以外も使用している。当時の社会分析には適するとしても、今の社会分析には古すぎると思われデータの更新が必要である[12)]。データ量は50kB（数字５万字相当）だが、都道府県別各種指標（データ）の入手をはじめ、数値を正確に入力する作業等が必要となる。

参考文献

１）水上茂樹編：『栄養情報処理論』、講談社（2004年３月）
２）浅井長一郎：『データとデータ解析』、放送大学教育振興会（1992年３月）
３）柳井春夫、高木廣文編：『多変量解析ハンドブック』、現代数学社（1986年４月）
４）阿部圭司：『Excelで学ぶ回帰分析』、ナツメ社（2004年9月）
５）小島隆矢：『Excelで学ぶ共分散分析とｸﾞﾗﾌｨｶﾙﾓﾃﾞﾘﾝｸﾞ』、オーム社（2003年12月）
６）㈱アンク編：『Excel　2000　ＶＢＡ辞典』、翔泳社（2000年1月）
７）田中亮：『Excel　ＶＢＡ　スパテク　３５８』、翔泳社（2000年1月）
８）経済企画庁国民生活局編：『平成11年版新国民生活指標』、大蔵省印刷局（1999年７月）
９）矢野恒太記念会編：『データでみる県勢2001』、国勢社（2000年12月）
10）矢野恒太記念会編：『データでみる県勢CD-ROM2004』、矢野恒太記念会（2003年12月）

11）国立天文台編：『理科年表2005』、丸善（2004年11月）
12）総務省統計局編：『社会生活統計指標2007』、日本統計協会（2007年１月）
（平成19年９月28日受付、平成19年10月９日受理）