応用科目・国際経済学 （講義資料）

Excel による回帰分析

熊　倉　正　修

はじめに

「国際経済学」の講義資料には多数の図表が含まれていますが，これらはすべてマイクロソフト社の Excel を用いて作成しています。これらの図表には簡単な回帰分析の結果をまとめたものが含まれていますが，これらの回帰分析も Excel の**分析ツール**を利用して行っています。時間の制約により授業中に回帰分析の詳細を解説することはしていませんが，これらについて質問を受けることがあります。それがこの小論を執筆した理由です。

この小論では Excel を用いて初歩的な回帰分析を行う方法を解説します。この小論の目的は受講生の皆さんに経済データの分析に関心を持ってもらうこと，私の講義資料に示した程度の回帰分析であれば皆さんでも決して難しくないことを知ってもらうことにあります。そのため，この小論は「とにかくまずやってみる」という方針で執筆しており，回帰分析の統計学的な解説は最小限に抑えています。これを読んでデータ分析に多少なりとも興味を持った人は，統計学や計量経済学の講義を聴講して本格的に勉強してもらいたいと思います。

なお，この小論の執筆直前にマイクロソフト社から Excel2010 がリリースされましたが，大半の人は現在でも Excel2007 以前のバージョンを使用しているでしょう。この小論は Excel2007 の利用を前提として執筆していますが，Excel2010 や Excel2003 を利用する場合でも回帰分析の手順はほとんど同じです。

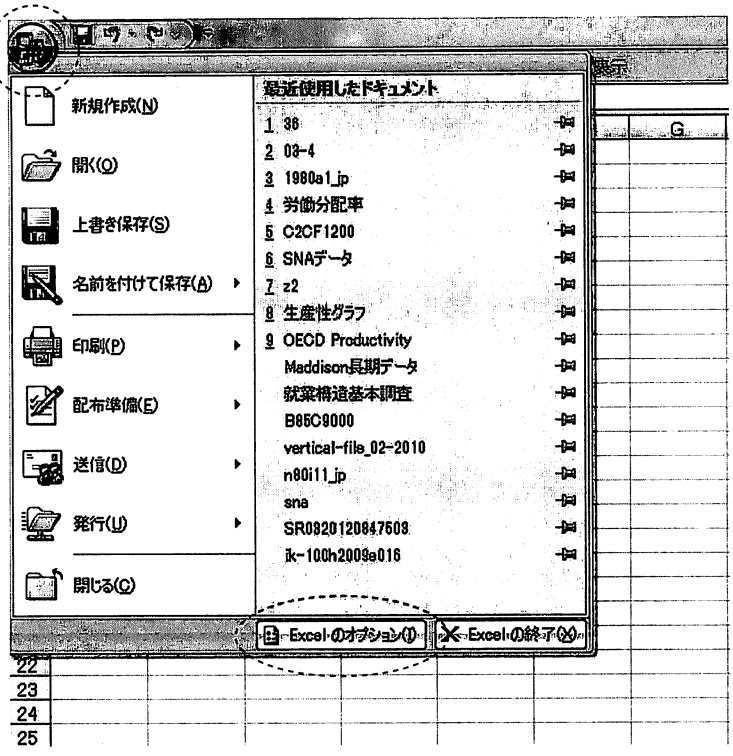
1.「分析ツール」の組み込み

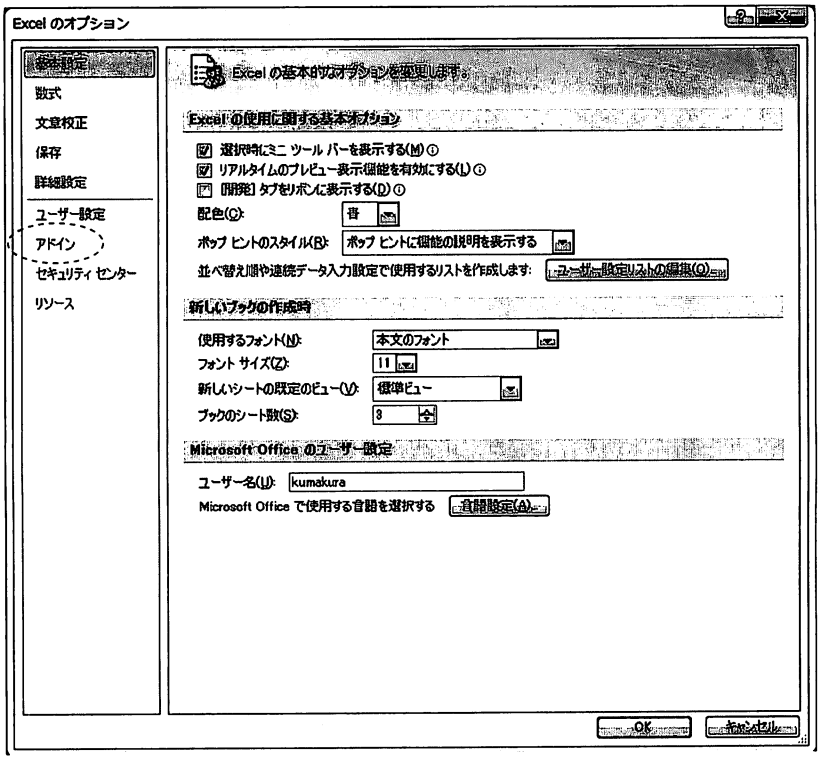
Excel で回帰分析を行うには**分析ツール**を利用できるようにしておく必要がある。まず，Excel を起動し，左上の Office ボタンをクリックする。次頁上図に示した画面が現れるので，最下段にある **Excel のオプション**をクリックする。

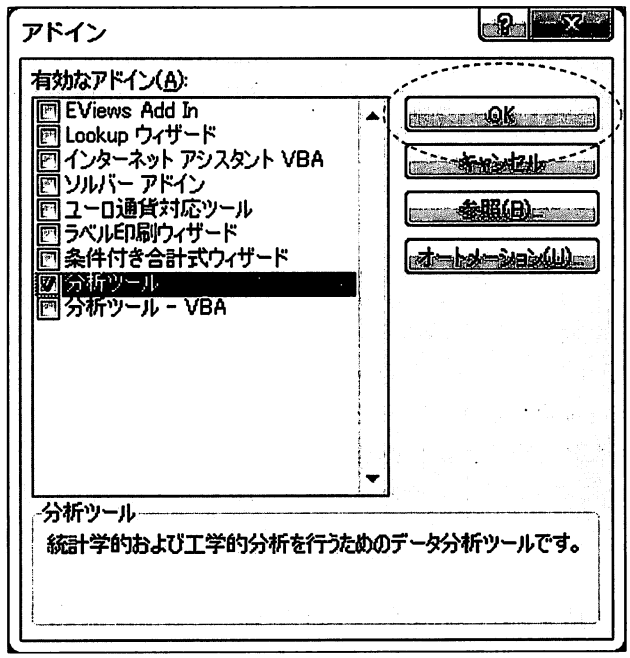
すると次頁下図のメニューが現れるので，左側の**アドインオプション**をクリックする。選択枝の中から**分析ツール**を選択した状態で，最下段の**設定**をクリックする。

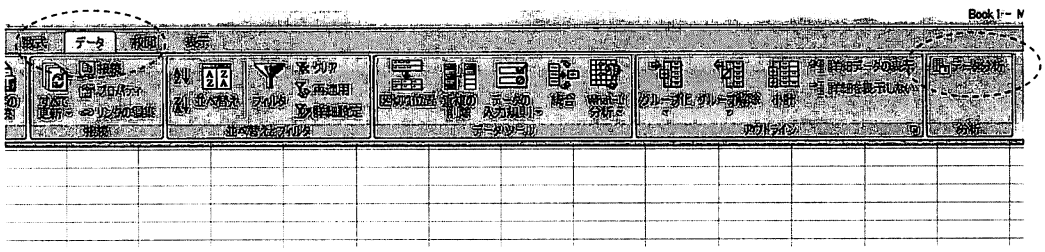
すると次の図のようなメニューが現れるので，**分析ツール**のボックスにチェックを入れて「OK」をクリックする。「インストールしますか？」と聞かれるので，「はい」を選択する。

分析ツールが組み込まれたことを確認するために，最上段の**データタブ**をクリックし，メニューを表示する。右端に**データ分析**というメニューが表示されていれば組み込み完了である。


新規作成(N)
開く(O)
上書き保存(S)
名前を付けて保存(A)
印刷(P)
配布準備(E)
送信(D)
発行(U)
閉じる(C)
最近使用したドキュメント
1 36
2 03-4
3 1980a1_jp
4 労働分配率
5 C2CF1200
6 SNAデータ
7 z2
8 生産性グラフ
9 OECD Productivity
Maddison長期データ
就業構造基本調査
B85C9000
vertical-file_02-2010
n80i11_jp
sna
SR0320120847503
ik-100h2009e016
Excel のオプション(I)
Excel の終了(X)


Excel のオプション
基本設定
数式
文章校正
保存
詳細設定
ユーザー設定
アドイン
セキュリティ センター
リソース
Excel の基本的なオプションを変更します。
Excel の使用に関する基本オプション
選択時にミニ ツール バーを表示する(M)
リアルタイムのプレビュー表示機能を有効にする(L)
[開発] タブをリボンに表示する(D)
配色(C):
ポップ ヒントのスタイル(R):
ポップ ヒントに機能の説明を表示する
並べ替え順や連続データ入力設定で使用するリストを作成します:
ユーザー設定リストの編集(O)...
新しいブックの作成時
使用するフォント(N):
本文のフォント
フォント サイズ(Z):
11
新しいシートの既定のビュー(V):
標準ビュー
ブックのシート数(S):
3
Microsoft Office のユーザー設定
ユーザー名(U):
kumakura
Microsoft Office で使用する言語を選択する
言語設定(A)...
OK
キャンセル





2. 回帰分析とは何か

経済学では複数の経済変数の関係をしばしば関数を用いて表現する。たとえば，y という変数が x_1 と x_2 という変数から影響を受けているとすると，これらの変数の関係を

$$y=f(x_1, x_2) \tag{1}$$

という抽象的な関数を用いて表現する。ただし経済現象は物理現象などに比べて複雑であり，一時的な要因やその他の雑多な要因も影響するため，(1)式の左辺と右辺の値は完全に一致しない場合が多い。そのような相対的に重要でない要因をまとめて e という変数で表すとすると，(1)式を

$$y=f(x_1, x_2)+e \tag{2}$$

と書き直すことができる。

経済理論から「y と x_1，x_2 の間に(2)式のような関係がある」という結論が得られても，それが本当に正しいかどうかは調べてみないと分らない。また，それが正しいとしても，(2)式の抽象的な関数のままではあまり役に立たない場合が多い。たとえば，y が家計消費，x_1 が家計の可処分所

得，x_2 が家計の負債だとして，$x_1 \to y$ の影響が正，$x_2 \to y$ の影響が負であるとする。その場合，政府が所得税減税を行えば家計の可処分所得が増加し，家計消費の増加を通じて景気が刺激される可能性が考えられる。しかし，税率を１％下げるごとにどれだけ消費が増加するか分からなければ，どれだけ減税したらよいか分らないだろう。このような問題に答えるためにしばしば利用されるのが**回帰分析**である。

たとえば，（2）式が

$$y=b_0+b_1x_1+b_2x_1^2+b_3x_2+b_4x_2^2+e \qquad (3)$$

という形になっていると仮定したとする。（3）式は（2）式に比べるとずっと具体的だが，係数の b_0 から b_4 までの値を調整することによって色々な関数を表現することができる。回帰分析とは，**できるだけ簡単な関数形を利用して現実のデータに適合するようその係数の値を選択すること，そして推計された式がどれだけ信頼できるかを統計学的に考察すること**である。たとえば，いろいろ検討した上で b_2 と b_4 はほとんど０だという結論が得られれば，（3）式は

$$y=b_0+b_1x_1+b_3x_2+e \qquad (4)$$

という簡潔な形になる。回帰分析では x_1 や x_2 を**説明変数**，y を**被説明変数**と呼ぶ。

回帰分析を行うにはデータが必要である。上記の例の場合，日本のすべての家計の消費総額と所得総額を過去に遡って調べ，n 年間の年次データをもとに推計することも可能だし，ある年における n 個の家計の消費と所得の金額を調べ，それを用いて推計することも可能である。前者のタイプのデータを**時系列データ**，後者のタイプのデータを**クロスセクション・データ**（横断面データ）と呼ぶ。（3）式に対応する推計式は

$$y_i=b_0+b_1x_{1i}+b_2x_{2i}^2+b_3x_{2i}+b_4x_{2i}^2+e_i, \qquad i=1,2,\ldots,n \qquad (5)$$

であり，i は年次データを用いる場合には年を，個別家計のデータを用いる場合には家計の識別番号を意味するが，基本的な推計方法は同じである。

なお，ここで

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad x=\begin{bmatrix} 1 & x_{11} & x_{11}^2 & x_{21} & x_{21}^2 \\ 1 & x_{12} & x_{12}^2 & x_{22} & x_{22}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{1n}^2 & x_{2n} & x_{2n}^2 \end{bmatrix}, \quad e=\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \quad b=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \qquad (6)$$

というベクトルや行列を定義すると，（5）式を

$$\boldsymbol{y}=\boldsymbol{x}\boldsymbol{b}+\boldsymbol{e} \qquad (7)$$

と簡潔に表現することができる。（5）式の右辺がもっと複雑な形になっていても，（6）式の $\boldsymbol{x}$ と $\boldsymbol{b}$ の内容を調整すれば，やはり（7）式の形に表現することができる。以下ではベクトルや行列の使用を避けるために（7）式の最も単純なケースである

$$y_i=b_0+b_1x_i+e_i, \qquad i=1,2,\ldots,n \qquad (8)$$

という**単回帰式**を例として話を進めるが，そこで得られた結果の多くは(5)式のような**重回帰式**においても成立する。2つの変数の関係はグラフを描けばある程度見当がつくため，実際に回帰分析が威力を発揮するのは3つ以上の変数が関与する重回帰式の場合である。

3. 最小二乗法

第2節において，回帰分析とは「できるだけ簡単な関数形を利用して現実のデータに適合するようにその係数の値を選択すること，そして推計された式がどれだけ信頼できるかを統計学的に考察すること」だと述べた。これらは実際には別々の作業ではなく，相互に深く関係しているが，ここでは(8)式という関数の形が選択済みだと仮定して，その係数 b_0 と b_1 をどのように決めたらよいかを考えてみよう。係数の選択基準はいろいろと考えられるが，**最小二乗法では回帰式の推計誤差の二乗和がもっとも小さくなるよう係数を選択する**。その意味は以下の通りである。

いま，(8)式においてある特定の b_0 と b_1 を選択したとする。その場合，(8)式がある x_i に関して予想する y_i の値は $y_i^*=b_0+b_1x_1$ だが，y_i は e_i からも影響を受けているため，y_i と y_i^* は一致しない。e_i は観察できない概念的な変数だが，ひとたび b_0 と b_1 が選択されれば，y_i と x_i のデータからそれに対応する値を計算することが可能である。その値を e_i^* と書くとすると，

$$y_i=(b_0+b_1x_i)+e_i^*=\underset{\text{予想値}}{y_i^*}+\underset{\text{推計誤差}}{e_i^*} \tag{9}$$

である。すべてのデータに関する推計誤差の二乗和

$$(e_1^*)^2+(e_2^*)^2+\cdots+(e_n^*)^2=\sum_{i=1}^{n}(e_i^*)^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2 \tag{10}$$

ができるだけ小さくなるように b_0 と b_1 の値を決めようというのが最小二乗法の考え方である。

上記の基準にもとづいて b_0 と b_1 を選択するとは，(10)式を b_0 と b_1 に関して微分した値が0になるように b_0 と b_1 を調節するということである。これらの条件を整理すると以下のようになる。

$$\text{(10)式を } b_0 \text{ について微分}\rightarrow\sum_{i=1}^{n}e_i^*=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)=0 \tag{11}$$

$$\text{(11)式を } b_1 \text{ について微分}\rightarrow\sum_{i=1}^{n}x_ie_i^*=\sum_{i=1}^{n}x_i(y_i-b_0-b_1x_i)=0 \tag{12}$$

(11)式と(12)式にはシグマ記号が入っているので難しそうに見えるが，これらは b_0 と b_1 に関する単純な二元一次方程式であり，中学校で習った数学で解くことができる。ただし実際の計算はそれなりに面倒なので，Excel に任せればよい。

なお，(11)式と(12)式を整理すると，

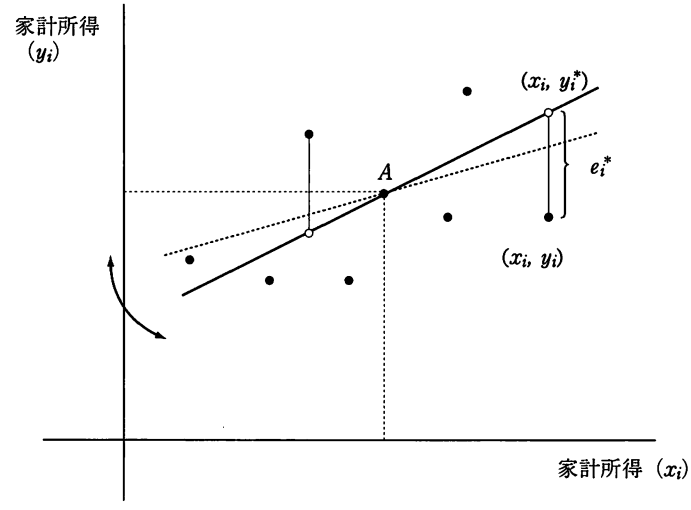
$$\bar{y}=b_0+b_1\bar{x} \tag{13}$$

という関係が得られることも知っておくとよい。ここで $\bar{x}$ と $\bar{y}$ はそれぞれ以下の算式にもとづく $x_1, x_2, \ldots, x_n$ と $y_1, y_2, \ldots, y_n$ の平均値を表している。

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i$$
$$\bar{y}=\frac{y_1+y_2+\cdots+y_n}{n}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{14}$$

図1においてA点が$(\bar{x}, \bar{y})$だとすると，最小二乗法はA点を支点とする直線を回転させてもっとも当てはまりのよいものを選択していることになる。このことは(4)式のように説明変数が二つ以上ある場合も同じである。(4)式の場合，図1に相当する図は三次元になり，$(\bar{x}_1, \bar{x}_2, \bar{y})$を通過する直線の中で最適なものを選択することになる。

図1　最小二乗法による回帰分析



4. Excel による推計

先に進む前に，ここで実際に Excel を用いて推計を行ってみよう。**マクロ経済学**では「消費は所得の正の関数だが，**限界消費性向**は1より小さく，所得に依存しない**独立支出**もある」として，

$$C=a+bY, \quad \text{ただし} \quad a>0, \quad 0<b<1 \tag{15}$$

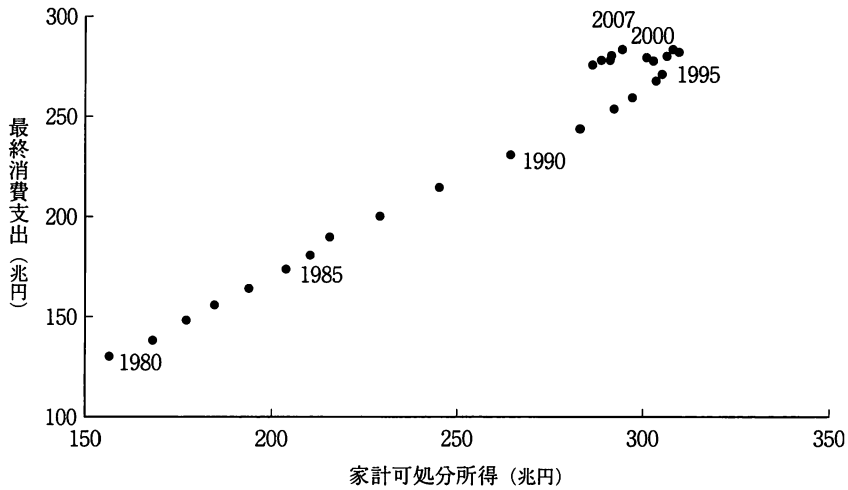
という消費関数を仮定することが多い。ここでCは消費，Yは所得を表している。aが独立支出，bが限界消費性向であり，これらが推計する係数である。

私のホームページ[1]にある「回帰分析」データセットの **Data1** ワークシートには1980－2007年の家計の最終消費支出と可処分所得（いずれも日本全体の合計値），その他のいくつかのデータが収録されている。回帰分析を行う際，いきなり推計しようとせず，まずデータをグラフに描くなどして大まかな特徴をつかむことが大切である。ここで最終消費を縦軸，可処分所得を横軸とするグラフを描いてみたところ，図2のようになった。これを見る限り，1980年代から1990年代半ばまでは(13)式で想定した直線の関係がほぼ成立しているが，1990年代後半からその関係が崩れてしまったようである。したがって1980－2007年のすべてのデータを利用して(13)式を推計した場合，十分な当てはまりは期待できそうもない。しかしそのことは後で検討することにして，ここではとりあえず最小二乗法による回帰を行ってみよう。

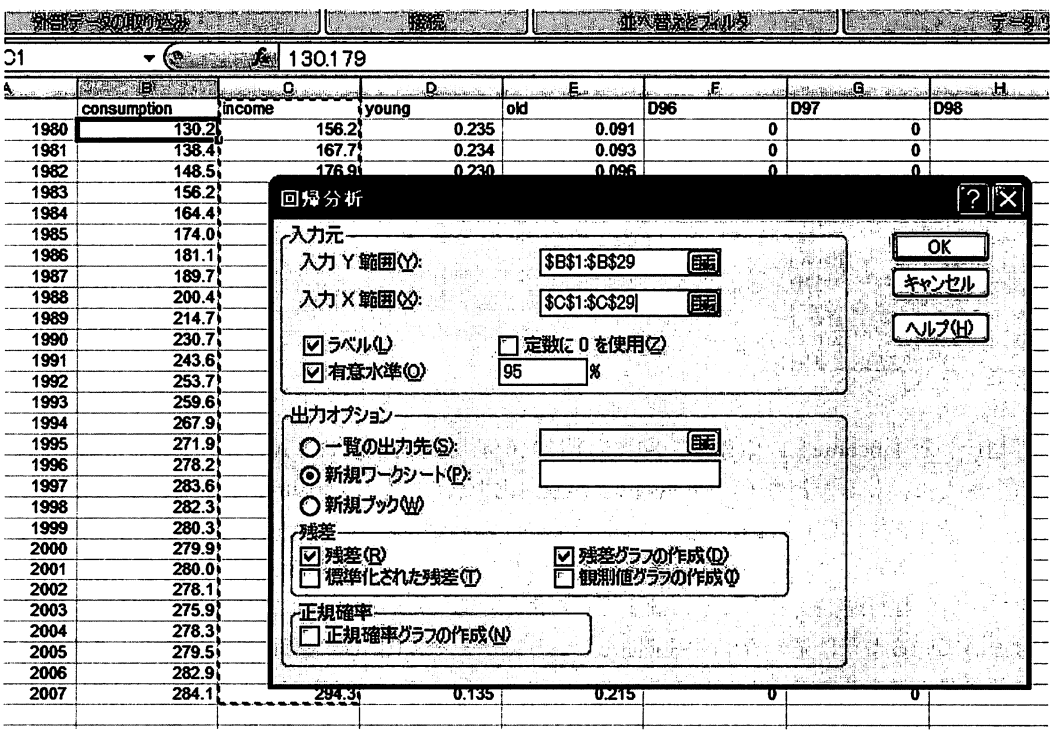
Data1 シートが開いている状態で Excel のデータリボンをクリックし，タブの右端にあるデータ

1)　http://www.econ.osaka-cu.ac.jp/~Kumakura/teaching/teaching.html

図 2　家計の可処分所得と最終消費支出



分析を選択する。現れたメニューの中から回帰分析を選択して OK をクリックすると，以下の入力画面が表示される。「入力元」の入力 Y 範囲に被説明変数の名称とデータが記載された B1 から B29 までの領域をドラッグして指定する。入力 X 範囲には説明変数の名称とデータが記載された C1 から C29 を指定する。各領域の最初の行のセルが変数の名称であることを認識させるためにラベルにチェックを入れる。その他は以下のようにチェックを入れ，OK をクリックする。

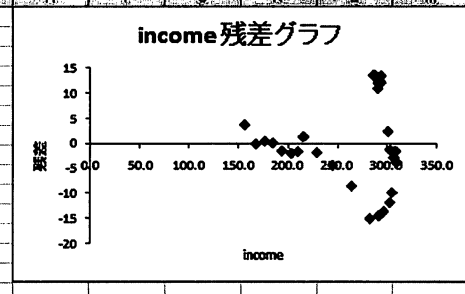


すると新しいワークシートが作成され，以下のような推計結果が出力されるはずである。

図3　家計消費関数の推計結果（1）

O3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	概要												
2													
3	回帰統計												
4	重相関 R	0.98609											
5	重決定 R2	0.972374											
6	補正 R2	0.971312											
7	標準誤差	9.070303											
8	観測数	28											
9													
10	分散分析表												
11		自由度	変動	分散	測された分散	有意 F							
12	回帰	1	75289.61	75289.61	915.1482	8.57E-22							
13	残差	26	2139.03	82.2704									
14	合計	27	77428.64										
15													
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%				
17	切片	-36.6069	9.149916	-4.00079	0.000467	-55.4148	-17.799	-55.4148	-17.799				
18	income	1.0441	0.034514	30.25142	8.57E-22	0.973155	1.115044	0.973155	1.115044				
19													
20													
21													
22	残差出力												
23													
24	観測値	値: consum	残差										
25	1	126.4957	3.683333										
26	2	138.514	-0.13518										
27	3	148.072	0.417014										
28	4	156.1931	0.040503										
29	5	165.9786	-1.54731										
30	6	176.0546	-2.04979										
31	7	182.7734	-1.58497										
32	8	188.3903	1.266389										
33	9	202.2968	-1.90528										
34	10	219.1805	-4.4803										
35	11	239.3282	-8.65617										
36	12	258.8233	-15.1979										
37	13	268.2974	-14.5864										
38	14	273.361	-13.7456										
39	15	279.8925	-11.946										
40	16	281.8113	-9.94673										
41	17	279.4765	-1.24602										
42	18	285.2619	-1.65617										
43	19	286.1784	-3.87648										
44	20	283.2741	-2.93702										
45	21	277.5452	2.361153										
46	22	267.6834	12.26829										
47	23	267.1441	10.94916										
48	24	262.364	13.55046										
49	25	264.8307	13.47968										
50	26	267.4495	12.02977										
51	27	270.782	12.10231										
52	28	270.6805	13.4393										
53													

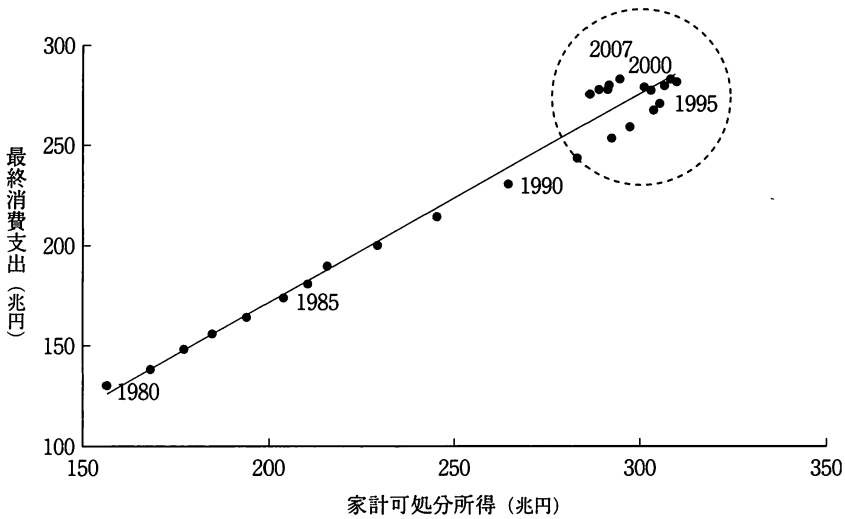


「切片」と「income」の係数の −36.6069 と 1.0441 がそれぞれ(15)式の a と b の推計値である。それをもとに計算した消費 C の予想値 C_i^* が**残差出力**表の「値：consumption」欄に，推計誤差の e_i^* が「残差」欄に出力されている。「観測値」の 1 から 28 はそれぞれ1980年から2007年に当たっている。先の散布図に回帰した式を書き加えたのが以下のグラフである。

さて，この推計結果をどのように評価したらよいだろうか。図 4 全体では回帰式はそれなりに当てはまっているように見えるが，'円で囲んだ最近のデータに関する当てはまりは悪く，このままでは減税効果を考える材料としては心もとない。また，図 3 では $a<0$，$b>1$ となっており，いずれも想定した条件を満たしていない。これらの点については第 6 節において検討する。

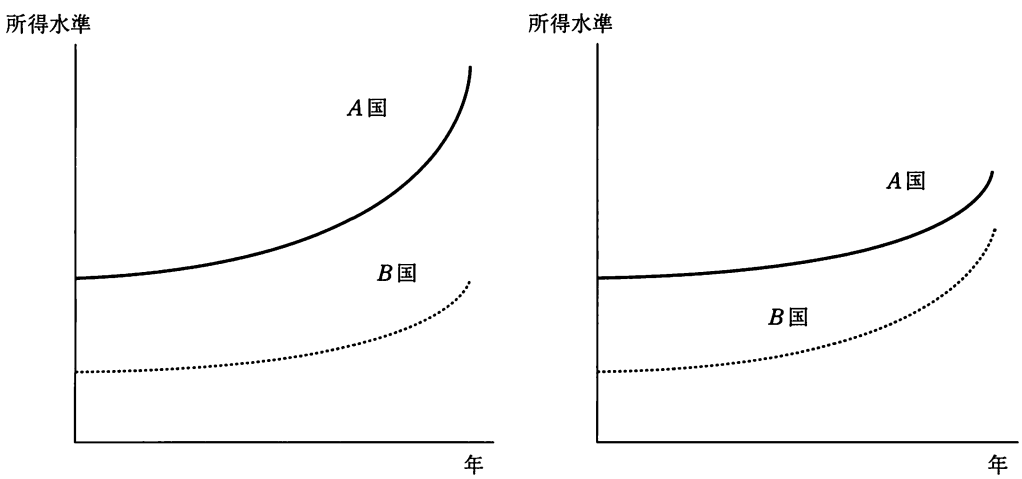
次に別の例として，回帰分析を用いて「所得水準の低い国ほどその後の所得の成長率は高い」と

図4　家計消費関数の推計結果（2）



いう**収斂仮説**を検討してみよう。図5の左パネルを見ると分かるように，初期時点で所得水準が低かった国の成長率が初期時点で所得水準が高かった国の成長率と同じか低い場合，二国の所得格差は無限に拡大してしまう。一方，右パネルのように初期の所得水準が低かった国の成長率が所得水準の高かった国の成長率をわずかでも上回っていれば，長期的には二国の所得は後者の国の水準に収斂する。左パネルの状況は貧しい国にとって絶望的なだけでなく経済合理性にも反するため[2)]，右パネルの収斂仮説を想定することが妥当である。しかし本当にそれが成立しているかどうかは調べてみないと分からない。

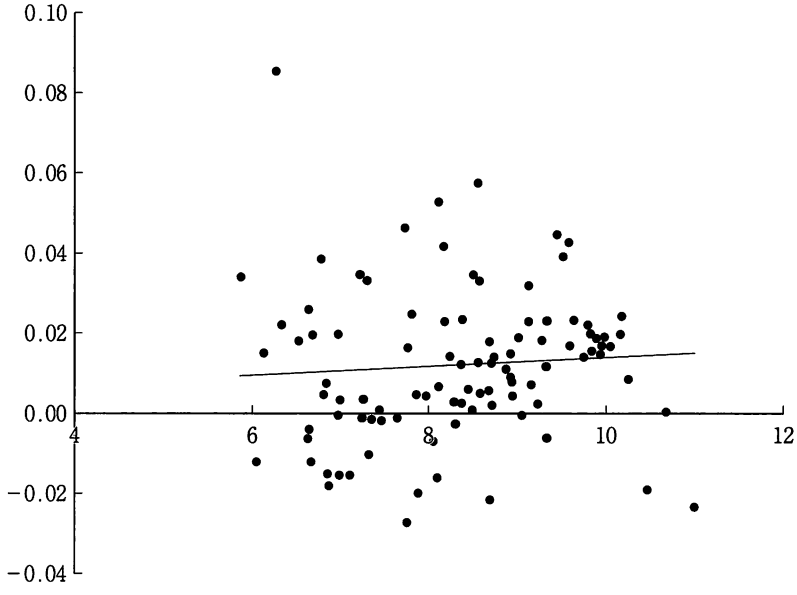
図5　収　斂　仮　説



2)　たとえば，図5の*A*国と*B*国の所得格差が非常に大きくなると，*A*国の企業は賃金水準の低い*B*国で生産活動を行おうとするだろう。その場合，*A*国の労働需要が減少し，*B*国の労働需要が増加するため，二国の賃金格差は縮小するはずである。

「回帰分析」ファイルの **Data2** シートには世界105カ国の1980年の一人当たり実質 GDP（購買力ベース）と1980－2007年の一人当たり実質 GDP の平均成長率（年率換算値），その他いくつかの変数のデータが収録されている。試しに1980－2007年の平均成長率を被説明変数とし，1980年の一人当たり実質 GDP の自然対数値を説明変数とする回帰分析を行った[3]。図6はもとのデータと推計した式を一つのグラフに描いたものである。このグラフの回帰式はわずかながら右上がりになっており，額面通り解釈すれば「豊かな国ほどその後の成長率が高い」，すなわち収斂仮説とは逆のことが生じていることになる。この推計結果は信頼できるだろうか。

図6　初期の所得水準とその後の経済成長率の関係



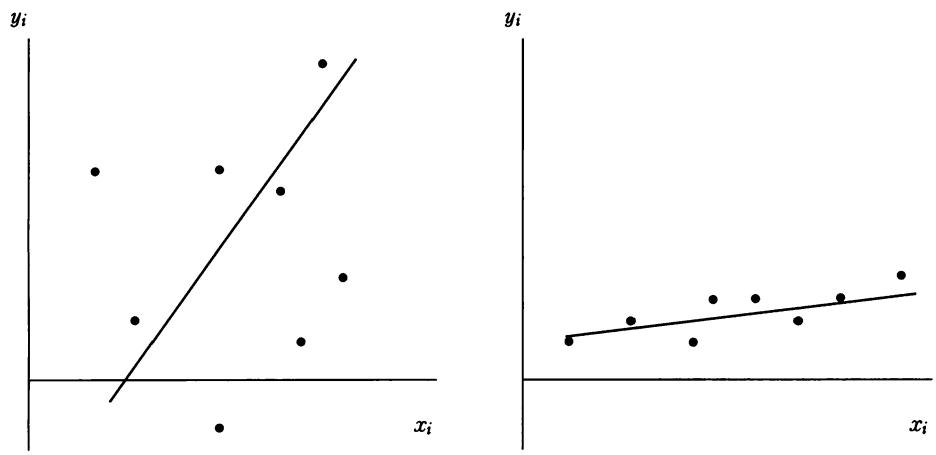
最小二乗法によって(8)式のような回帰式を推計した場合，説明変数と被説明変数の間に因果関係があってもなくても各係数の推計値が唯一つに決まり，図6のような回帰線が選択されてしまう。なお，先に「(3)式を推計して b_2 と b_4 の値がほとんど0だったら，実際の関係は(4)式だ」と述べたが，これは必ずしも「ある変数の係数の推計値が十分に大きい（小さい）場合，それはその変数が被説明変数に大きな影響を与えている」ことを意味しない。図7の左パネルでは x_i と y_i の間に何の関係もないが，無理やり最小二乗法で y_i を x_i に対して回帰して図中の直線が選択された場合，推計された x_i の係数は大きな正の値になっているはずである。一方，右パネルでは x_i と y_i の間に正の因果関係があるが，選択された回帰直線の勾配が緩やかなため，x_i の係数は0に近い値になっているはずである。しかし，「x_i の値をもとに y_i の値を予測する」上で左パネルの直線は全く役に立たず，右パネルの直線は有用である。x_i や y_i の単位を変えれば係数の推計値は変化するため[4]，係数の値が0に近いかどうかによって個々の説明変数の被説明変数に対する影響力を判断す

3)　もとの変数を自然対数値に変換するのはデータのばらつきを均すためであり，必須ではない。「世界経済の論点」追加資料参照。

4)　たとえば，図6において横軸に対数値ではなくもとの値を用いた場合，回帰線の勾配がどうなるか考えてみてもらいたい。

ることはできず，回帰式が意味をなしているかどうかも判断できない。それではどうしたらよいのか。

図7　意味のある回帰式と意味のない回帰式



5. 推計結果の解釈

ここで図3の出力表に戻り，左上の部分を見やすく整えた上で再掲してみよう（表1）。この表に含まれている数値のうち，先に観察した係数の推計値以外はいずれも回帰式や個々の説明変数の妥当性を判断するための手掛かりである。本格的に回帰分析を行うにはこれらの意味をよく理解していなければならないが，詳細な説明は森棟他（2008，第1章），滝川・前田（2009，第23－24章），縄田（2009，第13章）などに譲り，ここでは表1の大まかな意味だけを説明する。

この推計は以下のような回帰式を想定して行ったものだった。

$$y_i = b_0 + b_1 x_i + e_i, \qquad i = 1, 2, \ldots, n \tag{8}$$

（8）式において実際に x_i（所得）が y_i（消費）に規則的な影響を与えていたとしても，不規則要因の影響を表す e_i がどの程度のばらつきを持つかによって推計の精度や回帰式の有用性は左右される。図8の左パネルは e_i のばらつきが大きいケース，右パネルは e_i のばらつきが小さいケースである。左パネルの場合，e_i がどのような値をとるかによって y_i の値が大きく変化するため，回帰式の係数である b_0 や b_1 を正確に推計することが困難になる。また，かりにこれらの係数の値が分かったとしても，実際の y_i の値は予想値 y_i^* からかなり乖離してしまう可能性が高く，回帰式の予想能力も低くなる。

それでは e_i のばらつきの大きさはどのようにして調べたらよいのだろうか。ある変数のばらつき具合を測る代表的な指標が**分散**と**標準偏差**である。ある変数 z_i の真の平均値を μ，分散を σ^2，標準偏差を σ と書くとすると，分散と標準偏差の定義はそれぞれ

$$\sigma^2 = \frac{\sum_{i=1}^{n}(z_i-\mu)^2}{n}, \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(z_i-\mu)^2}{n}} \tag{16}$$

表1　家計消費関数の推計結果（再掲）

概　要

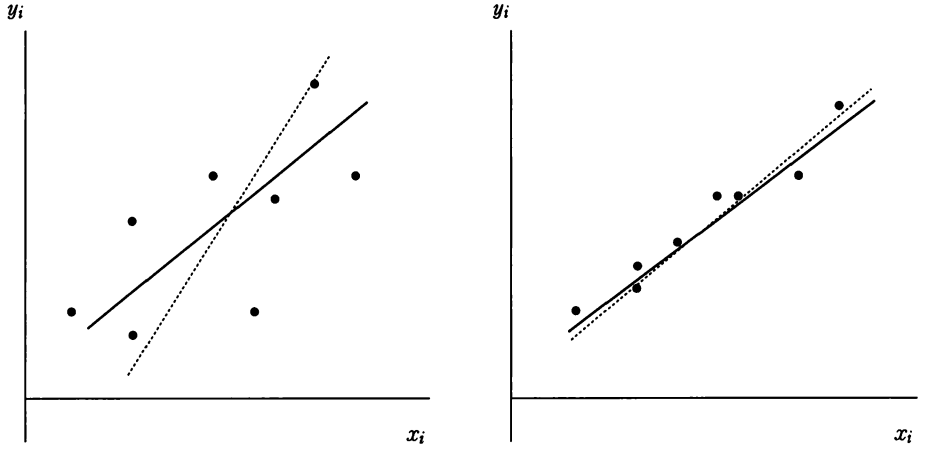
回帰統計	
重相関 R	0.986
重決定 R2	0.972
補正 R2	0.971
標準誤差	9.070
観測数	28

分散分析表

	自由度	変　動	分　散	観測された分散比	有意 F
回帰	1	75,289.61	75,289.61	915.15	0.00
残差	26	2,139.03	82.27		
合計	27	77,428.64			

	係　数	標準誤差	t	P-値	下限95%	上限95%	下限95.0%	上限95.0%
切片	−36.61	9.15	−4.00	0.00	−55.41	−17.80	−55.41	−17.80
income	1.04	0.03	30.25	0.00	0.97	1.12	0.97	1.12

図8　誤差項のばらつきと推計式の関係



であり，$\mu=0$ の場合には

$$\sigma^2=\frac{\sum_{i=1}^{n} z_i^2}{n}, \qquad \sigma=\sqrt{\frac{\sum_{i=1}^{n} z_i^2}{n}} \tag{17}$$

となる。

上記と同様に(8)式の y_i と e_i の分散と標準偏差を定義すると

$$\sigma_y^2=\frac{\sum_{i=1}^{n} (y_i-\mu_y)^2}{n}, \qquad \sigma_y=\sqrt{\frac{\sum_{i=1}^{n} (y_i-\mu_y)^2}{n}} \tag{18}$$

$$\sigma_e^2=\frac{\sum_{i=1}^{n}e_i^2}{n},\qquad \sigma_e=\sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n}} \tag{19}$$

となる。(18)式の μ_y は y_i の真の平均値を表している。(18)式の μ_y と(19)式の e_i の値は分からないので，これらをそれぞれ $\bar{y}$ と e_i^* で代用し，以下のように μ_y と e_i の分散と標準偏差を推計する。

$$\hat{\sigma}_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1},\qquad \hat{\sigma}_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \tag{20}$$

$$\hat{\sigma}_e^2=\frac{\sum_{i=1}^{n}(e_i^*)^2}{n-2}=\frac{\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2}{n-2},\qquad \hat{\sigma}_e=\sqrt{\frac{\sum_{i=1}^{n}(e_i^*)^2}{n-2}}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2}{n-2}} \tag{21}$$

(20)式では分母が n ではなく $n-1$ になっており，(21)式では分母が $n-2$ になっている。これらは表1の「分散分析表」に示されている**自由度**という概念と結びついている。自由度とは「自由に動き回れるデータの数」のことである。(18)式には $y_1, y_2, \ldots, y_n$ という n 個の自由に動き回れるデータがある。しかし(20)式では y_i の真の平均値 μ_y を $\bar{y}$ によって代用している。$\bar{y}$ は(14)式を用いて計算した値であるため，かりに $y_1, y_2, \ldots, y_{n-1}$ という $n-1$ 個のデータと $\bar{y}$ が与えられれば最後の y_n を逆算することができる。すなわち，先に $\bar{y}$ を決めてしまった場合，(20)式の分子の中で「自由に動き回れる変数」は $n-1$ 個しかない。同様に，(21)式においても真の b_0 と b_1 の代わりに自分で推計した値を利用している。b_0 と b_1 の推計値は $y_1, y_2, \ldots, y_n$ に(11)式と(12)式の条件（制約）を課して求めた値であるため，これらの推計値が先に与えられた場合，$y_1, y_2, \ldots, y_n$ のうち独立に動き回れるものの数は $n-2$ 個しかない。これだけではまだ(20)式や(21)式の説明になっていないが，「分母に自由度を用いると $\hat{\sigma}_y$ と $\hat{\sigma}_e$ が偏りのない σ_y と σ_e の推計値になる」ことが知られているものとして先に進むことにする。

表1では，「分散分析表」の「残差」の「分散」が(21)式の $\hat{\sigma}_e^2$ を表しており，「回帰統計」の「**標準誤差**」が $\hat{\sigma}_e$ を表している。先に誤差項 e_i のばらつきが大きいと回帰式が不安定になり，係数の正確な推計が困難になると述べた。このことをもう少し正確に表現すると以下のようになる。(8)式において x_i が不確定な要因を含まない確定した変数だとしても，e_i が不確実要因を含む**確率変数**であるため，y_i もやはり確率変数である。そして(10)式から分かるように，b_0 や b_1 の計算には x_i と y_i の両方のデータが用いられているため，b_0 や b_1 もまた確率変数である。b_0 や b_1 の不確実性の根源は e_i の不確実性にあるため，e_i の標準偏差が分かればそれに対応する b_0 や b_1 の標準偏差も計算できる。

e_i の真の標準偏差の代わりに先の $\hat{\sigma}_e$ を利用し，それに対応する b_0 と b_1 の標準偏差を計算した結果が，表1下段の「標準誤差」である。たとえば，income の行の標準誤差 0.03 は「推計された b_1 の標準偏差が 0.03 である」こと，すなわち，「与えられたデータから b_1 の唯一の値を選択しろと言われれば 1.04 だが，1.04±0.03 程度の推計誤差は大いにありうる」ことを意味している。したがって，たとえば係数の推計値が 0.5 で標準偏差が 0.7 の場合，「真の係数は −0.2 と 1.2 の範囲の間のどの値をとっているか分からないため，実際には 0 かもしれない（＝本当はその説明変数は被説明変数に何の影響を与えていないかもしれない）」ことを意味している。

上記の点をシステマティックに判断するための材料が「標準誤差」の右側に出力されている「t」

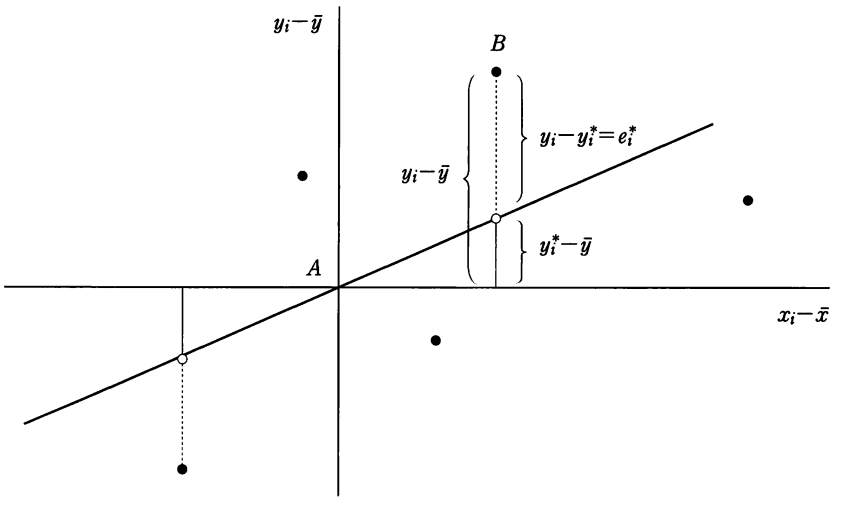
と「P-値」である。t は回帰分析では **t 値**と呼ばれ，

$$\frac{b_1}{b_1 \text{の標準誤差}} \tag{22}$$

を意味している。係数の推計値をその不確実性の大きさを表す標準誤差で除しているから，この値が 0 から十分に離れているか否かがその変数が**統計的に有意か否か**（＝その変数が被説明変数に有意味な影響を与えているか否か）の一つの判断材料になる。図 7 左パネルでは b_1 の推計値は大きくても標準誤差も大きくなり，t 値は小さくなる。右パネルでは b_1 の推計値は小さくても標準誤差がさらに小さいため，t 値は相対的に大きな値をとる。「P 値」とは「b_0 や b_1 が本当は 0 である確率」を意味しており，t 値が小さいほど大きくなる。表 1 では b_0 の P 値も b_1 の P 値も 0.00 となっており，「これらが本当は 0 である」可能性がほとんどゼロであることが分かる。私たちにとっては t 値より P 値のほうが分かりやすいが，レポートなどで回帰分析の結果を報告する場合，個々の係数の推計値の脇や下に括弧つきで t 値だけを示すことが多い。推計に用いるデータの数（**標本サイズ**）にもよるが，おおむね t 値が絶対値で 2 を超えていればその説明変数は被説明変数に意味のある影響を与えており，1 未満ならほとんど関係がないと考えてよい。

最後に，表 1 左上の「重決定 R2」と「補正 R2」の意味を説明しておこう。「回帰統計」表の値はいずれも回帰式全体の当てはまり具合を判断する材料であり，先に解説した回帰式の標準誤差もその一つだった。先述したように，最小二乗法によって係数を選択した場合，推計された回帰式は必ず $\bar{y}=b_0+b_1\bar{x}$ という関係を満たしている。図 1 を $A(\bar{x}, \bar{y})$ を原点として描き直すと，図 9 のようになる。たとえば，図の B 点に当たる y_i に関する回帰式の予想値は y_i^* だから，y_i の $\bar{y}$ からの乖離のうち回帰式によって説明できた部分は $y_i^*-\bar{y}$，説明できなかった残差は $y_i-y_i^*$ である。

図 9　回帰式の説明力



さて，ここで

$S_T=\sum_{i=1}^{n}(y_i-\bar{y})^2$：全変動（Total sum of squares）

$S_E=\sum_{i=1}^{n}(y_i^*-\bar{y})^2$：回帰式により説明された変動（Explained sum of squares）

$S_R=\sum_{i=1}^{n}(y_i-y_i^*)^2$：残差の変動 (Residual sum of squares)

という値を定義すると，常に

$$S_T=S_E+S_R \tag{23}$$

という関係が成立することが知られている。さらに

$$R^2=\frac{S_E}{S_T}=1-\frac{S_R}{S_T} \tag{24}$$

という値を定義すれば，それが回帰式のあてはまり具合（説明力）の指標になっていることを直観的に理解できるだろう。この R^2 が表1左上の「**重決定 R2**」である[5]。この指標は正式には**決定係数**と呼ばれており，その値が0に近いほど回帰式の説明力が低く，1に近いほど説明力が高いことを示している。

ただし，推計式が重回帰式の場合，若干の追加説明が必要である。(4)式や(5)式のような重回帰式において説明変数を増やしてゆくと，y_i のばらつきのうち回帰式で説明可能な部分が増えることはあっても減ることはない。極端な例として，標本サイズが n の時に $n-1$ 個の独立した説明変数を持つ（＝切片の分も含めて n 個の係数を含む）重回帰式を推計した場合，これらの説明変数と被説明変数が無関係であっても y_i のばらつきを100％説明できてしまう[6]。したがって，(5)式の決定係数が(8)式の決定係数より大きくなるのは当たり前であり，そのことを根拠に(5)式が(8)式より優れているとは言えない。先述したように，回帰分析では「できるだけ簡単な関数形を利用して」「現実のデータに上手くマッチするようその係数値を選択する」が，上記のことは**これら二つの目標の間にトレードオフがある**ことを示している。

「できるだけ簡単な関数形を利用する」という目標と「回帰式の説明力を高める」という目標のバランスをとるために利用されるのが**修正決定係数（自由度修正済決定係数）**である。ここで(18)式の決定係数を書き直すと以下のようになる。

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-y_i^*)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(e_i^*)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{25}$$

修正決定係数はこれを以下のように修正する。

$$\bar{R}^2=1-\frac{\dfrac{\sum_{i=1}^{n}(e_i^*)^2}{n-k}}{\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}=1-\frac{\bar{\sigma}_e^2}{\bar{\sigma}_y^2} \tag{26}$$

(26)式の $n-k$ は $e_1^*, e_2^*, \ldots, e_n^*$ の自由度であり，単回帰式の場合には先に説明したように $k=2$，説明変数が2つある重回帰式の場合には $k=3$ となる。

(26)式において説明変数の数を増やしてゆくと，$\sum(e_i^*)^2$ は小さくなるが $n-k$ も小さくなる。$\sum(e_i^*)^2$ が小さくなるスピードが $n-k$ が小さくなるスピードより速ければ $\bar{\sigma}_e^2$ は小さくなるが，そうでなければ大きくなる。説明変数の数を増やすことで $\bar{R}^2$ が大きくなるのは前者のケースだから，

5) 表1の「重相関 R」は重決定を0.5乗した値，すなわち $\sqrt{R^2}$ のことだが，これは実際には利用されない。

6) たとえば，図1や図9において $y_i=a+bx_i+cx_i^2$ という2次式の3つの係数を調整すれば，任意の3点を通る式にすることができる。

「そのような場合にだけ変数を増やした回帰式のほうがもとの回帰式より優れていると判断しよう」というのが修正決定係数の考え方である。実際の回帰分析では決定係数はほとんど利用されず，もっぱら修正決定係数が参照される。表1では左上の「**補正 R2**」が修正決定係数を表している。

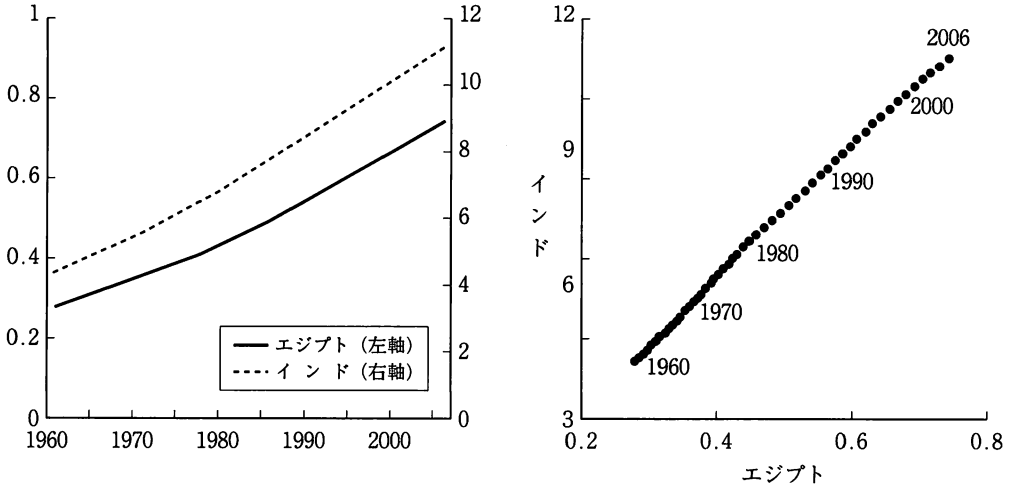
6. 有意義な回帰分析のために

6.1 時系列データと見せかけの回帰

ただし，実際の修正決定係数の解釈には注意が必要である。表1の修正決定係数は0.971という1にきわめて近い値になっており，この回帰式の説明力は非常に高いように見える。しかし，この推計式は1990年代以降の所得と消費の関係を十分に説明できていないし，他にもおかしな点があった。時系列データを用いて回帰分析を行う場合，推計した式が**見せかけの回帰**に陥っている場合が少なくないため，修正決定係数の値が高くても安心してはいけない。以下でこのことを説明しよう。

図10はエジプトとインドの人口の推移を描いたものである。左パネルでは両国の人口を縦軸に，年を横軸にとってグラフを描いており，右パネルではインドの人口を縦軸に，ブラジルの人口に横軸にとってプロットしている。

図10 エジプトとインドの人口の推移（単位：億人）

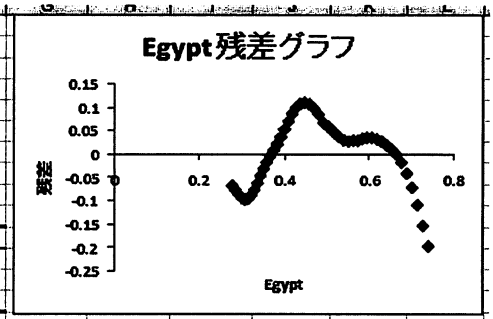


ここで(8)式においてインドの人口を被説明変数，エジプトの人口を説明変数とした回帰分析を行ったところ，図11のような結果が得られた。推計式の修正決定係数はほとんど1であり，説明変数の係数のt値もきわめて大きな値になっている。しかしこの回帰式に意味があるだろうか。図10を見る限り，インドの人口もエジプトの人口もかなり急速な増加トレンドを持っている。図10の推計式がこれらの増加トレンドを反映しているだけであり，インドとエジプトの人口の間に因果関係があるわけではないことは直観的にあきらかだろう。この例から言えることは，推計に用いる変数に上昇トレンドや下降トレンドを持つものが含まれている場合，これまで説明した方法で単純に回帰分析を行ってもうまくゆかない可能性が高いということである。

推計した回帰式が適切かどうかを判断する一つの材料になるのが推計誤差 e_t^* の動きである。Excelを用いて回帰分析を行う時に**残差（R）**と**残差グラフの作成（D）**の項にチェックを入れておく

図 11　最小二乗法によるインドとエジプトの人口の回帰結果

概要								
回帰統計								
重相関 R	0.999							
重決定 R2	0.999							
補正 R2	0.999							
標準誤差	0.074							
観測数	47							
分散分析表								
	自由度	変動	分散	観測された分散比	有意 F			
回帰	1	198.191	198.191	36,484.463	0.000			
残差	45	0.244	0.005					
合計	46	198.435						
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0.281	0.039	7.171	0.0000000057	0.202	0.360	0.202	0.360
Egypt	14.848	0.078	191.009	0.0000000000	14.691	15.004	14.691	15.004



と，各データに対応する推計誤差とそのグラフを自動的に出力してくれる。推計誤差はランダムな要因だけを反映しているはずだから，推計した回帰式が適切であれば，それが図 11 のように規則的な動きをすることはありえない。ある年の残差 e_i^* が前年の残差 e_{i-1}^* が強く相関している場合，それは回帰式に重要な問題があるサインだと考えてほしい。

図 11 のようにある年の残差 e_i^* が前年の残差 e_{i-1}^* が強く相関している場合，

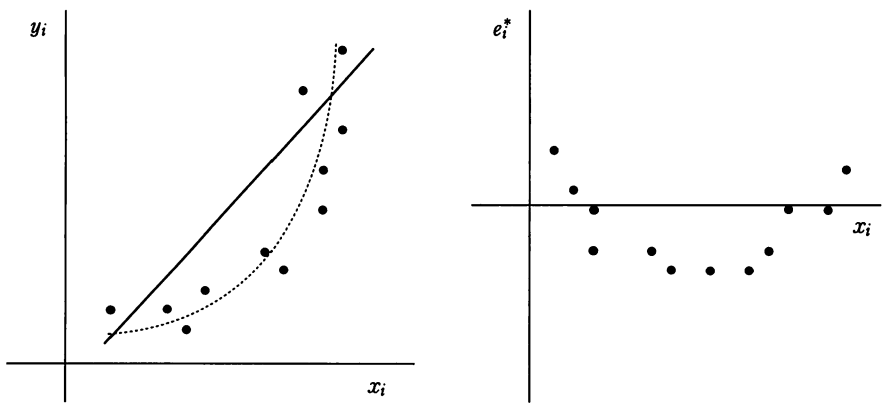
(i)　回帰式が見せかけの回帰に陥っており，実際は説明変数と被説明変数の間に関係がない

(ii)　回帰式の説明変数は被説明変数に影響を与えているが，他の重要な説明変数を見落としている

(iii)　回帰式の説明変数は被説明変数に影響を与えているが，想定した回帰式の関数の形が適切でない

といった可能性が考えられる。(iii)の例として，図 12 のように**非線形関係**にあるデータに無理やり**線形**（直線）の式を回帰するケースが挙げられる。この場合，回帰式の右辺に x_i^2 という説明変数を加え，$y_i = b_0 + b_1 x_i + b_2 x_i^2 + e_i$ という回帰式を推計してやればよい。

図 12　回帰式の選択と誤差項の系列相関

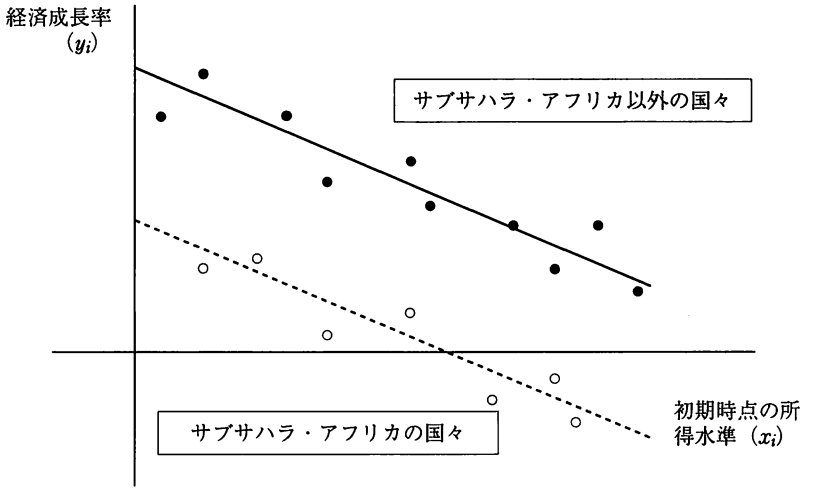


6.2　ダミー変数の利用

上記の(ii)の場合，推計結果や残差のグラフを観察しつつ，他にどのような説明変数があるか再考する必要がある。しかしこれは実際には簡単でない。先の収斂仮説の回帰分析では，予想に反して初期の所得水準がその後の成長率に与える影響が正になっていた。これは他の重要な説明変数を見過ごしていたためだろうか。

収斂仮説に関する研究では，「アフリカは特別だ」という結果が示されていることが多い（小峰2008，40頁）。**Data2** シートのデータをよく観察すると分かるように，サブサハラ・アフリカ（Sub-Saharan Africa，サハラ砂漠以南のアフリカ）には非常に貧しい国々が多く，その中には1980－2006年の平均経済成長率がマイナスだった国が少なくない。これらの国々には旧植民地時代から引き継いだモノカルチャー的な産業構造，複雑な民族問題による内紛や国境紛争，天候不良やエイズの蔓延といった深刻な社会的・経済的問題があり，国連や日本政府も援助に力を入れている。そこで，これらの国々では他の地域の国々と初期の所得水準が同一でもその後の成長率が低くなると予想し，図13のような関係を仮定してみよう。その場合，(8)式をどのように修正したらよいだろうか。

図13　所得水準と経済成長率の関係



この場合，サブサハラ・アフリカの国々とそれ以外の国々に関して別々に(8)式の推計を行うことも可能だが，そのような方法はあまり効率的でない。それぞれの回帰分析に用いるデータの数が少なくなり，正確な推計が困難になるからである。

上記の問題に対処する一つの方法はダミー変数を利用することである。ダミー変数とは0か1かの値をとる人工的な変数のことである。ここで

$$D_i=\begin{cases}0 & i\text{がサブサハラ・アフリカの国の場合}\\ 1 & i\text{がそれ以外の国の場合}\end{cases} \tag{27}$$

というダミー変数を定義し，

$$y_i=\alpha+\beta D_i+\gamma x_i+e_i \tag{28}$$

という回帰式を推計したとする。その場合，サブサハラ・アフリカの国々に関しては

$$y_i=\alpha+\gamma x_i+e_i \tag{29}$$

という式を，それ以外の国々に関しては

$$y_i=(\alpha+\beta)+\gamma x_i+e_i \tag{30}$$

という推計式を適用していることになり，図12の点線の切片がαに，実線の切片が$\alpha+\beta$に対応する。実際に(28)式を推計し，βのt値（の絶対値）が小さければ，二つのグループを区別する必要がないと判断すればよい。

回帰分析においてダミー変数が必要になることは非常に多い。たとえば，「所得水準の消費水準への影響」ではなく，「所得水準の被服費支出への影響」を知りたいとする。衣料品への消費性向は性別や年齢層によってまちまちである可能性が高いが，うまくダミー変数を活用すればこれらの違いを考慮しつつ簡潔な推計を行うことが可能になる。

6.3 消費関数の再検討

最後にまとめとして，先の消費関数についてもう一度考えてみよう。一国の経済が成長するにつれて家計の所得や消費は増加するから，先に利用した家計消費額と家計可処分所得のデータは上昇トレンドを持っている可能性が高い。そこで，(15)式の代わりに，

$$\Delta C_t=a+b_0\Delta Y_t+b_1\Delta Y_{t-1}+e_t \tag{31}$$

という式を考えてみよう。ここで $\Delta C_t=C_t-C_{t-1}$ と $\Delta Y_t=Y_t-Y_{t-1}$ は前年からの消費と所得の変化を表している。(31)式では，ある年に所得が1円増えると消費は同年にb_0円，翌年にb_1円増加する。説明変数に前年の所得の変化を追加したのは，家計の消費行動が所得の変化に反応するまでに時間がかかる可能性を考慮したためである。この式では1円の所得増加は長期的に消費b_0+b_1を円増加させるため，これが限界消費性向となる。C_tが上昇トレンドを持っていればΔC_tは正になっている年が多いだろうが，(31)式において平均的な消費の変化はaに反映されるため，(15)式に比べると見せかけの相関に陥る可能性は低いと思われる[7]。

次に一時的な要因の影響も考えておこう。我が国では1989年に消費税が導入され，1997年4月には消費税率が3％から5％に引き上げられた。1989年の導入は決定後すぐに実施されたが，1997年の税率改訂は1994年にすでに閣議決定されており，国民が将来の増税を考慮して消費行動を調整することが可能な状況にあった。そのため，1996年から1997年初にかけて税率引き上げ前の駆け込み消費があり，1997年後半から1998年にかけてはその反動で消費が落ち込んだと言われている。「1996年に駆け込み消費があり，その反動が1998年に現れた」という仮説を(31)式に盛り込むために，

7) 厳密にはもう少し慎重な定式化が必要だが，この小論の範囲を超えるのでここでは取り扱わない。

$$D_t^{96}=\begin{cases}0 & t\text{が1996年以外の場合}\\ 1 & t\text{が1996年の場合}\end{cases}$$
$$D_t^{98}=\begin{cases}0 & t\text{が1998年以外の場合}\\ 1 & t\text{が1998年の場合}\end{cases} \qquad (32)$$

というダミー変数を定義し，(31)式を以下のように書き換えてみよう。

$$\Delta C_t=a_0+a_1D_t^{96}+a_2D_t^{98}+b_0\Delta Y_t+b_1\Delta Y_{t-1}+e_t \qquad (33)$$

先述したように家計が消費税率上げに反応していたとすると，$a_1>0$，$a_2<0$ だと予想される。

(33)式のような重回帰式であっても，Excel の操作方法は第3節で解説した単回帰式の場合と同一である。ただし，この小論で利用したファイルには1980年以降のデータしか収録されておらず，(32)式では $\Delta Y_{t-1}=Y_{t-1}-Y_{t-2}$ という説明変数が含まれているため，推計に用いるデータの数は1982年から2007年までの26個となる。

(33)式の重回帰式を推計した結果が表2である。補正 R2＝0.894 であることから，回帰式の説明力は高いと言える。また，すべての係数の符号が予想通りになっており，回帰式が経済学的に意味のある式になっている。限界消費性向は $b_0+b_1=0.712$ であり，まずまず妥当な値と言える。a_1 と a_2 の推計値はそれぞれ5.215と −4.771であり，1996年の駆け込み需要が約5.2兆円，1998年の反動減が約4.8兆円と推計されている[8]。すべての係数の P-値が5％未満になっており，各説明変数が被説明変数に対して意味のある影響を与えていることが分かる。

表2　消費関数の再推計結果(1)

概　要

回帰統計	
重相関 R	0.954
重決定 R2	0.911
補正 R2	0.894
標準誤差	1.720
観測数	26

分散分析表

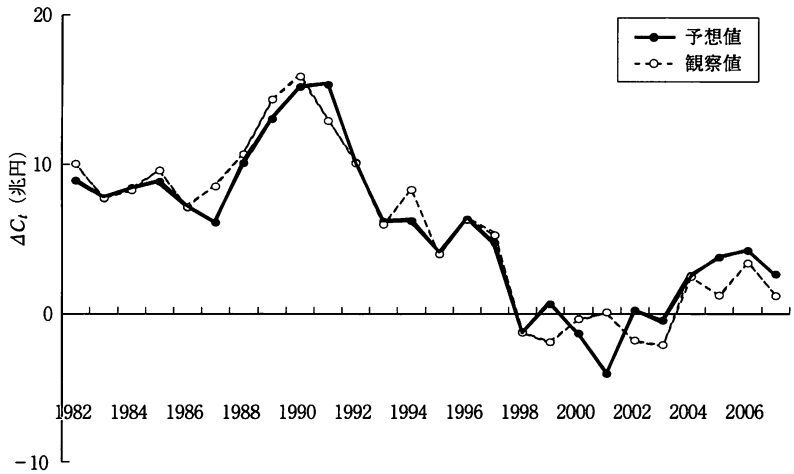
	自由度	変　動	分　散	観測された分散比	有意 F
回帰	4	633.296	158.324	53.489	0.000
残差	21	62.159	2.960		
合計	25	695.455			

	係　数	標準誤差	t	P-値	下限95％	上限95％	下限95.0％	上限95.0％
切片	2.047	0.446	4.590	0.000	1.120	2.975	1.120	2.975
income	0.541	0.083	6.492	0.000	0.368	0.714	0.368	0.714
income(-1)	0.171	0.081	2.107	0.047	0.002	0.339	0.002	0.339
D96	5.215	1.808	2.885	0.009	1.455	8.975	1.455	8.975
D98	−4.771	1.796	−2.657	0.015	−8.506	−1.036	−8.506	−1.036

8)　実際にはこれらの変化をすべて消費税に帰することは必ずしも適切ではないが，ここではこれ以上の↗

実際の消費の変動を回帰式にもとづく予想値と比較したのが図 14 である。回帰式の説明力は良好なように見受けられるが，まだ改良の余地はありそうである。どのような工夫がありうるか考えてみてもらいたい。

図 14　消費関数の再推計結果(2)



参　考　資　料

小峰隆夫（2008）『最新日本経済入門（第 3 版)』日本評論社
滝川好夫・前田洋樹（2009）『経済学のための Excel 入門（Office 2007 対応版)』日本評論社
縄田和満（2009）『Excel による統計入門（Excel 2007 対応版)』朝倉書店
御園謙吉・良永康平（2007）『よくわかる統計学Ⅱ経済統計編』ミネルヴァ書房
森棟公夫・照井伸彦・中川　満・黒住英司（2008）『統計学』有斐閣

↘分析は行われない。