Ⅳ　回帰分析入門

1)　2 変量データの記述

1. 散布図の描画

【 課題 17 】 下に示したものは、日本の実質家計可処分所得と実質家計最終消費支出のデータ(平成 7 年基準、単位: 兆円)[1]である。このデータを入力し、散布図を描いてみよう。

📖 散布図は次のような手順で描けばよい。

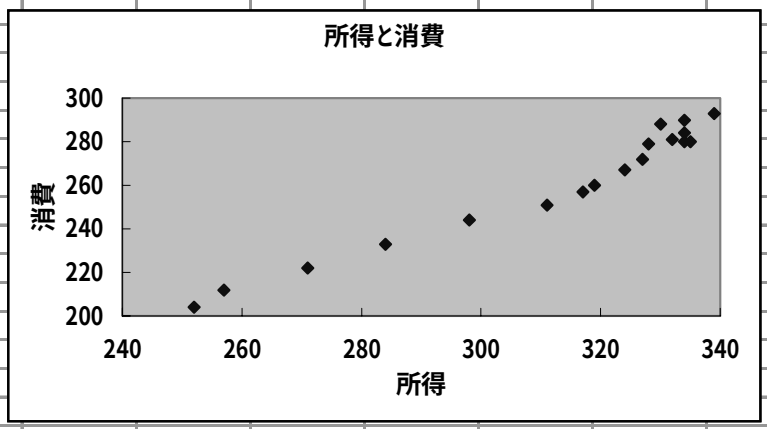
① メニューバーの「挿入」―「グラフ」でグラフウィザードを起動し、「散布図」を選択する。

② データ範囲を B2:C19 とする。

③ 下の図のようにタイトル、軸ラベルを入力し、目盛線、凡例を非表示にする。

④ グラフの作成を終えた後で、それぞれの軸の書式設定をおこない、
　縦軸　最小値：**200**　最大値：**300**　目盛間隔：**20**
　横軸　最小値：**240**　最大値：**340**　目盛間隔：**20**
とする。

	A	B	C	D	E	F	G	H
1	年	所得(X)	消費(Y)					
2	1986	252	204					
3	1987	257	212					
4	1988	271	222					
5	1989	284	233					
6	1990	298	244					
7	1991	311	251					
8	1992	317	257					
9	1993	319	260					
10	1994	324	267					
11	1995	327	272					
12	1996	328	279					
13	1997	332	281					
14	1998	335	280					
15	1999	334	280					
16	2000	334	284					
17	2001	330	288					
18	2002	334	290					
19	2003	339	293					
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								
35								
36								
37								



[1] 出典: 『平成 17 年版 国民経済計算年報』

2. 相関係数の導出

【 課題 18 】 所得と消費のデータについて相関係数を求めてみよう。

相関係数の計算式は次のような式である。

$$R = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\{n\sum X^2 - (\sum X)^2\}\{n\sum Y^2 - (\sum Y)^2\}}}$$

したがって相関係数を導出するためには、**ΣX**、**ΣY**、**ΣXY**、**ΣX²**、**ΣY²** をまず求める必要がある。ΣX、ΣYは、B 列、C 列の和を求めればよいが、ΣXY、ΣX^2、ΣY^2 を求めるためには、交差積(XY)と 2 乗(X^2, Y^2)を D 列、E 列、F 列に計算したうえで、その和を求めることになる。

手順は次のようになる。

📖 手順

① D 列に X と Y の交差積を求める。D2 セルに**=B2*C2** と入力し、これをコピーすればよい。

② E 列に X の 2 乗を、F 列に Y の 2 乗を求める。2 乗を表す演算子は '^' であり、E2 セルに**=B2^2** と入力し、これをコピーする。F 列も同様である。

③ B21 セルから F21 セルに各列の合計を求める。これらのセルがそれぞれΣX、ΣY、ΣXY、ΣX^2、ΣY^2 である。

④ C23 セルに **=(18＊D21-B21＊C21)/SQRT((18＊E21-B21^2)*(18＊F21-C21^2))** と入力する。この式と計算式とを見比べてみよ。

＜作成見本＞

	A	B	C	D	E	F
1	年	所得(X)	消費(Y)	XY	X^2	Y^2
2	1986	252	204	51408	63504	41616
3	1987	257	212	54484	66049	44944
4	1988	271	222	60162	73441	49284
5	1989	284	233	66172	80656	54289
6	1990	298	244	72712	88804	59536
7	1991	311	251	78061	96721	63001
8	1992	317	257	81469	100489	66049
9	1993	319	260	82940	101761	67600
10	1994	324	267	86508	104976	71289
11	1995	327	272	88944	106929	73984
12	1996	328	279	91512	107584	77841
13	1997	332	281	93292	110224	78961
14	1998	335	280	93800	112225	78400
15	1999	334	280	93520	111556	78400
16	2000	334	284	94856	111556	80656
17	2001	330	288	95040	108900	82944
18	2002	334	290	96860	111556	84100
19	2003	339	293	99327	114921	85849
20						
21	合計	5626	4697	1481067	1771852	1238743
22						
23		相関係数	0.980669			
24						

2) 単回帰モデル(その 1)

1. 回帰直線の導出

【 課題 19 】 所得と消費のデータについて Y = α + βX という 1 次式をあてはめ、回帰係数 α，β の推定値を求めよ。(C24 セルに β の推定値 b、C25 セルに α の推定値 a を求めよ。)

回帰係数の推定値を求める式は次のようなものである。

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

この式に相関係数の導出の際に求めた ΣX、ΣY、ΣXY、ΣX^2、ΣY^2 を代入すれば回帰係数の推定値がそれぞれ計算できる。

2. 予測値と残差の計算

【 課題 20 】所得と消費のデータについて各年の X のデータに対する予測値 $\hat{Y}$ と残差を求めよ。

📖 手順

① G 列に予測値を求める。予測値 $\hat{Y}$ は各 X_i について a + b X_i を計算すればよいので、G2 セルに 1986 年の X(B2 セル)に対応する予測値を求めるなら**=$C$25+$C$24*B2** とし、これをコピーすればよい。ここでは、コピーの際に絶対参照をするので、'$'がついている。

② H 列に残差を求める。残差は Y から予測値 $\hat{Y}$ を引いたものなので、H2 セルに **=C2-G2** とし、これをコピーすればよい。

3. 回帰直線のグラフへの書き入れ

散布図に回帰直線を書き入れる場合、Excelでは各Xに対応する予測値をグラフに書き入れ、それを直線でつなぐという手順をとる。

【 課題 21 】 所得と消費のデータについて散布図に回帰直線を書き入れよ。

📖 手順

① **グラフをアクティブ**（グラフエリアの枠の四隅および４辺に ■ という印が現れて入る状態。散布図のグラフエリアの白い部分をクリックすればこの状態になる。）にし、メニューバーから「グラフ」―「データの追加」を選ぶ。すると、「グラフの追加」というウインドウが開くので、G2 セルから G19 セルまでをドラッグし、OK ボタンを押す。

② ①の操作で散布図上にピンク色の点が現れたはずである。これを直線で結ぶ。ピンク色の点のひとつをダブルクリックすると、「データ系列の書式設定」ウインドウが開く。[2]そこで「パターン」のタグにおいて、**線を「指定」**をチェックし、**色を黒**に変え、**マーカーを「なし」**にする。グラフエリアの外をクリックすると回帰直線が引けたことがわかるはずである。

---

[2] この操作はうまくいかないことが多々ある。その場合にはグラフエリアの外を一回クリックした後で、もう一度この操作を繰りかえしてみると良い。

4. 決定係数の導出

決定係数は回帰における当てはまりの尺度であり、全変動のうち回帰モデルによって説明される変動の割合を示すものである。決定係数は**0**と**1**の間の値をとるが、決定係数が**0.2**や**0.3**などの小さい値であるということは、あまり関係のない**X**と**Y**の間に因果関係を想定し、分析を行っているということを意味し、モデルの再検討が必要となる。

決定係数は、相関係数との間に、

$$\text{(決定係数 } R^2\text{)} = \text{(相関係数)}^2$$

という関係がある。

5. 残差の表示

残差 e_i は従属変数の個々の観測データと回帰直線との間のズレの大きさをあらわすものであった。この残差の状態を調べることで、回帰直線のあてはまり具合いなど、さまざまな情報を入手することができる。残差を出発点としてモデルの設定やデータ間の関係を検討する分析を、残差分析(residual analysis)という。ここでは、残差を求めてそれをグラフに表示してみよう。

グラフを描くには、残差 e_i を縦軸にとり、横軸には、

① i　　②X_i　　③ $\hat{Y}_i$

などを用いる。

⌧演習問題３：消費と所得のデータについて、【 課題 21 】までに加え、決定係数を計算し、残差プロット(横軸は年とする)を描いたものを、A4 用紙縦 1 枚に納まるようにレイアウトして、印刷してみよ。

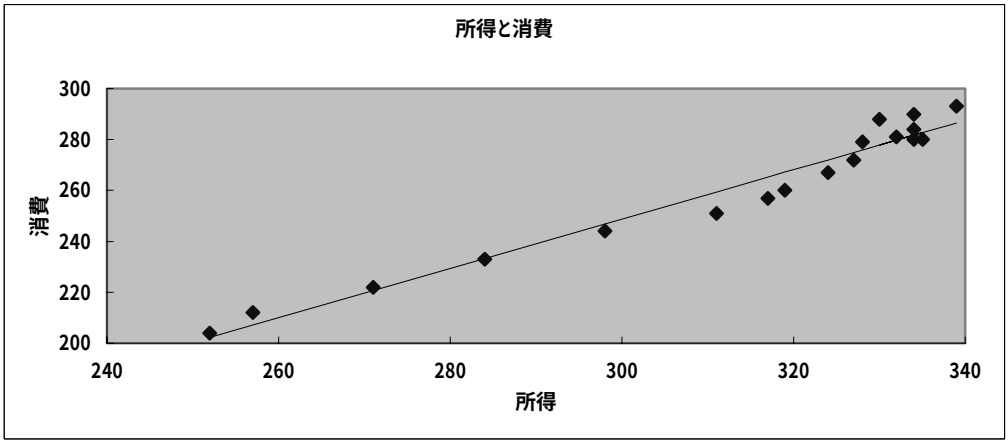
次ページに見本を示す。

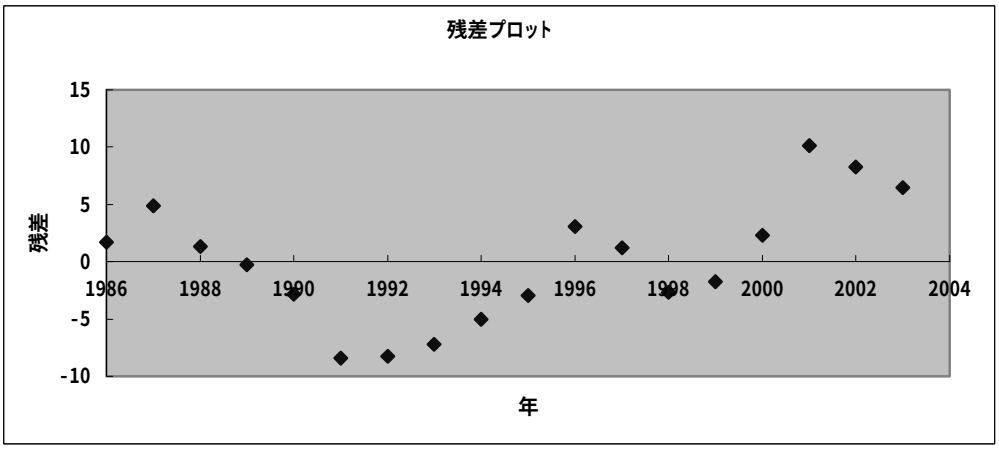
＜作成見本＞

わが国の実質家計可処分所得と実質家計消費支出　E36-000　徳山 太郎

年	所得(X)	消費(Y)	XY	X^2	Y^2	予測値	残差
1986	252	204	51408	63504	41616	202.2889	1.711132
1987	257	212	54484	66049	44944	207.132	4.868011
1988	271	222	60162	73441	49284	220.6927	1.307272
1989	284	233	66172	80656	54289	233.2848	-0.28484
1990	298	244	72712	88804	59536	246.8456	-2.84558
1991	311	251	78061	96721	63001	259.4377	-8.4377
1992	317	257	81469	100489	66049	265.2494	-8.24944
1993	319	260	82940	101761	67600	267.1867	-7.18669
1994	324	267	86508	104976	71289	272.0298	-5.02981
1995	327	272	88944	106929	73984	274.9357	-2.93568
1996	328	279	91512	107584	77841	275.9043	3.095693
1997	332	281	93292	110224	78961	279.7788	1.221196
1998	335	280	93800	112225	78400	282.6847	-2.68468
1999	334	280	93520	111556	78400	281.7161	-1.71605
2000	334	284	94856	111556	80656	281.7161	2.283948
2001	330	288	95040	108900	82944	277.8416	10.15844
2002	334	290	96860	111556	84100	281.7161	8.283948
2003	339	293	99327	114921	85849	286.5592	6.440827
合計	5626	4697	1481067	1771852	1238743		

相関係数	0.980669
b	0.968624
a	-41.8044
決定係数	0.961712





3) 単回帰モデル(その 2)

前章では、相関係数と回帰係数の推定値を交差積和(ΣXY)、2乗和(ΣX, ΣY)を求め、それを計算式に代入することによって求めた。

しかしExcelによって相関係数と回帰直線を求めるには、以下に説明するような Excelが備えている関数を用いることもできる。ここでは**reg**の例について、統計関数を用いた方法についても行ってみることにする。その際には、データを入力したセルの範囲に名前をつけておくと便利である。まず、実習の準備として、**reg**のSheet1から、年次、所得X、消費Yの部分(A3:C21)をSheet2のA3:C21にコピーし、所得のデータに _X 、消費のデータに _Y という名前を定義しておく。

1. 統計関数による相関係数と回帰直線の導出

Excelが備えている関数を用いた相関係数と回帰直線の導出を行ってみることにする。regのSheet2のA3:C21に、年次、所得X、消費Yのデータが入力されているものとする。

(1) 関数 PEARSON (CORREL) , RSQ

相関係数を求めるには、関数 **PEARSON**(引数1, 引数2）を用いる。PEARSON は相関係数を最初に導出した Karl Pearson (イギリス;1851-1936) にちなんでつけられた名前である。または、**CORREL** という名前の関数もあるが、どちらも全く同じものである。引数は２個あり、それぞれがデータの範囲（名前でもよい）である。戻り値は r_{xy} である。

= PEARSON(B4:B21, C4:C21)
= PEARSON(_X , _Y)

範囲B4:B21に名前_X、範囲C4:C21に名前_Yを付けてあれば、どちらの式でも結果は同じである。以下の説明では下式の書き方で示す。

関数 **RSQ**は相関係数の２乗（＝**決定係数**）を求める関数であるが、引数は PEARSON と同じである。したがって、べき乗を求める演算子 ^ を用いれば RSQ は不要となる。

= RSQ(_X, _Y)
= (PEARSON(_X, _Y)) ^ 2

どちらも全く同じ結果を与える。

(2) 関数 SLOPE とINTERCEPT[3]

SLOPE は回帰直線の傾き（回帰係数） **b** を、**INTERCEPT** は切片（回帰定数）**a** を求める関数で、どちらも引数は２個あるが、最初の引数が従属変数の範囲で、２個目の引数が独立変数の範囲をとる。引数の順序に注意しなければならない。

= SLOPE(_Y, _X)
= INTERCEPT(_Y, _X)

[3] 回帰直線の傾きと切片を求める関数には、**LINEST** という関数がある。この関数は傾きと切片以外に分析結果に関する多くの情報量を与えてくれる、非常に便利な関数である反面、使用法および結果の解釈の仕方が難しい。LINEST 関数の説明はここでは省略する。

(3) 関数 FORECAST と TREND

予測値 $\hat{Y}$ を求める関数には2種類のものが用意されている。関数**FORECAST** は引数を３個とり、FORECAST(X_i , Y範囲, X範囲) として用いる。戻り値は $a + bx_i$ として求められた数値１個である。

=FORECAST(B4, _Y, _X)　　セルB4の値を x としたときの $a + bx$ が戻り値

残りの X の値に対する予測値は、これをコピーして求めればよい。

あるいは、ｎ個の予測値を書き込む範囲を指定しておき、配列数式とすることもできる。たとえば、D4:D21の範囲を指定して、

=FORECAST(_X,_Y,_X)

を入力して、Ctrl + Shift + Enter とする。

関数 **TREND** も同じ予測値を求めるものであるが，引数の数が FORECAST より１個多く，計４個となる。一般的な型式は TREND(Y範囲, X範囲, X_i, 1)となる。最後の引数は 0 か 1 で、0 のときは、原点を通る直線 $Y = bX$ による予測値、1 のときはこれまで通りの $Y = a + bX$ による予測値を戻り値として求める。第４引数を省略した場合は、1 を指定したものとみなす。

TRENDの第３引数として、a, b の計算に用いない任意の数値を指定することもできる。たとえば、

= TREND(_Y, _X, 190 , 1)

とすれば、$\hat{Y}$ = a + b * 190 を求めることになる。regの X のデータの中には 190という数値はなく、これによって求まる $\hat{Y}$ は未知の X の値に対する**予測値**（外挿値）である。同様のことを、FORECASTを用いてもおこなうことができる。

=FORECAST(190, _Y, _X)

FORECASTでは 190 が第１引数となる点に注意されたい。また、190 という数値を直接指定するのではなく、セル番地で指定することもできる。セル N4 に190が書き込んであれば、

=TREND(_Y, _X, N4, 1)
=FORECAST(N4, _Y, _X)

とすればよい。また、N4からN13に $\hat{Y}$ = a + b x として求めたい x の値が連続して書き込まれていれば、

=TREND(_Y, _X, N4:N13, 1)
=FORECAST(N4:N13, _Y, _X)

とすればよい。範囲 N4:N13 に名前を付けて、それを使用してもよい。

3. 分析ツールの利用

Excel には統計分析を行うためのいくつかの分析ツールが付属している。これらのツールを使えば一度に詳細な分析結果を得ることができる。

分析ツールを最初に使用する場合にはメニューバーの「ツール」―「アドイン」を選び、分析ツールをチェックすることによって、分析ツールをアドイン(有効にすること)しなくてはならない。

アドインを行った後で、再びメニューバーから「ツール」を選ぶと、下のほうに「分析ツール」と表示される。ここで分析ツールを選び、回帰分析を選べばよい。