Ⅳ　回帰分析入門

4)　重回帰モデル

これまでは取り扱った単純回帰モデルは X と Y という 2 変量の間の因果関係を分析するものであり、Y という結果を説明する要因(原因)として X という変数 1 つが考えられるというものであった。

ここでは、Y を説明する要因が複数ある場合についての分析方法をみていく。

1. 重回帰モデル

重回帰モデルとは、Y を説明する要因が複数である場合、すなわち説明変数が 1 つだけではなく、X_1、X_2、X_3…　というように複数ある場合である。このような場合に考えられる式の 1 つに

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n$$

という線形のモデルが考えられる。

reg でとりあげた消費関数は、消費（Y）を説明する要因として所得（X）のみを考えたが、それ以外にも資産などを要因として考えることができる。資産を W という変数で表すと、

$$Y = \alpha + \beta X + \gamma W$$

というモデルを用いることが考えられる。

2. データの追加

Excelの分析ツールを用いて重回帰分析をおこなう場合、説明変数は連続した範囲におかなくてはならない。regのSheet1では、B列に説明変数である所得のデータがあり、C列に被説明変数である消費のデータがあった。

Sheet3に分析のためのデータをおくことにすると、そのためには

① 消費と所得のデータを入れ替えてコピー（**Sheet1のB列をSheet3のC列に、Sheet1のC列をSheet3のB列にそれぞれコピー**）し、

② D列に右の表の資産のデータを追加する。

作成見本は右図のようになる。

	A	B	C	D
1				
2				
3	年	消費(Y)	所得(X)	資産(W)
4	1981	168	204	447
5	1982	176	210	478
6	1983	181	216	512
7	1984	186	221	567
8	1985	194	229	620
9	1986	201	235	678
10	1987	210	240	768
11	1988	221	254	851
12	1989	231	266	965
13	1990	243	277	1091
14	1991	249	289	1048
15	1992	254	293	1071
16	1993	256	294	1069
17	1994	262	299	1122
18	1995	267	302	1179
19	1996	273	302	1234
20	1997	276	304	1278
21	1998	273	304	1279
22	1999	275	303	1304
23	2000	277	301	1396

3. 重回帰分析

Excelの分析ツールを用いて重回帰分析を行う場合、単回帰モデルの場合とほぼ同様である。相違点は入力X範囲をC4:D23とする点のみである。

分析ツールを用いて分析をおこなえば、次ページのような分析結果とグラフが作成される。

＜分析結果＞

概要

回帰統計	
重相関 R	0.99877
重決定 R2	0.997541
補正 R2	0.997251
標準誤差	2.000883
観測数	20

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	2	27604.49	13802.24	3447.515	6.64E-23
残差	17	68.06008	4.003534		
合計	19	27672.55			

	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	26.15679	9.512625	2.749692	0.013677	6.086906	46.22668	6.086906	46.22668
X 値 1	0.583752	0.060306	9.679792	2.49E-08	0.456517	0.710987	0.456517	0.710987
X 値 2	0.05438	0.007239	7.512094	8.5E-07	0.039107	0.069653	0.039107	0.069653

残差出力

観測値	予測値: Y	残差	標準残差
1	169.5499	-1.54994	-0.81893
2	174.7382	1.261777	0.666674
3	180.0896	0.910353	0.480995
4	185.9993	0.000703	0.000371
5	193.5514	0.448558	0.237
6	200.208	0.792017	0.418471
7	208.0209	1.979073	1.045665
8	220.707	0.293022	0.154822
9	233.9113	-2.9113	-1.53822
10	247.1844	-4.18443	-2.21089
11	251.8511	-2.85112	-1.50642
12	255.4369	-1.43686	-0.75918
13	255.9119	0.088148	0.046574
14	261.7127	0.287258	0.151776
15	266.5636	0.436353	0.230552
16	269.5545	3.445462	1.820447
17	273.1148	2.885246	1.524451
18	273.1691	-0.16913	-0.08936
19	273.9449	1.055122	0.557485
20	277.7803	-0.78032	-0.41229

4. 分析結果の解釈

回帰分析を行った場合にはその結果を解釈しなくてはならない。分析ツールを用いた分析結果は、非常に多くの情報を含んでいるので、多くの判断ができる反面、初学者にとっては困惑するものである。ここでは、**最初に見るべき分析結果**を挙げる。それは、

① 係数推定値

② 決定係数および自由度修正済み決定係数

① 係数推定値

係数推定値はデータの傾向をもっともよく表す回帰直線を考えるとき、その直線の方程式の係数の推定値である。この重回帰分析では、回帰直線が

$$Y = 25.157 + 0.584X + 0.0544W$$

という式で表される。ここで、もっとも注意すべき点は、係数の符号である。消費(Y)を説明する要因として所得(X)と資産(W)を考えたが、経済理論において、所得↑のとき消費↑、資産↑のとき消費↑と考えられるので、これらの係数の符号は正になるはずである。この分析結果は、係数の符号がともに正になるので、**理論と分析の結果が一致している。**

もし、分析結果においてどちらか1つの係数の符号が負になっていれば、理論と分析の結果が一致しないことになる。このような場合、

a) 理論が誤っている。

b) 分析の手法もしくは分析に用いたデータが誤っている。

のいずれかの原因が考えられる。

② 決定係数および自由度修正済み決定係数

決定係数は回帰モデルのあてはまり具合を示す尺度であり、0と1の間の値をとる。決定係数が1に近いほど回帰直線のあてはまりはよく、決定係数の値が小さい場合(0.5とか0.6以下の場合)には、分析の妥当性を検討する必要がある。

重回帰分析において、説明変数を追加することによって決定係数の値が大きくなるということは、説明変数の追加によってモデルの説明力が増したということである。

しかし、重回帰モデルにおける決定係数は、**被説明変数と関係のない説明変数**を追加した場合にも値が大きくなってしまうという欠点を持つ。

たとえば、次のようなデータをE列に追加し、重回帰分析をおこなってみよう。

1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
3	3	4	4	1	3	6	6	5	6	6	2	4	5	6	6	5	6	6	6

このデータは阪神タイガースのセ・リーグでの順位であり、日本の家計消費とはまったく関係ない。このようなデータを加えても、決定係数の値は高くなる。

そこで、その欠点を修正したものが自由度修正済み（または調整済み）決定係数である。

説明変数の追加によってモデルの説明力が増したかどうかはこの指標によってみる。

5) 実証分析の手順

経済学などで提唱されるさまざまな理論や、理論の上に知識や経験などを用いて考えられるさまざまな仮説は、統計データを用いた分析によって検証される。この分析が**実証分析**と呼ばれるものであり、分析の手法としては回帰分析がよく用いられる。実証分析をおこなう場合には次のような手順をとる。

1．モデルの定式化
2．モデルに含まれる変数と実際のデータの対応
3．パラメータの推定と統計量の算出
4．モデルの検討

ここでは「都道府県別死亡率」について、実証分析をおこなった１つの例を実際に実習してみよう。

1. モデルの定式化

実証分析の最初の手順はモデルの定式化である。理論や自分の知識や経験基づく仮説を数式の形であらわす。

「都道府県別死亡率」のデータを見ると、都道府県によって死亡率の高いところと低いところがある。これはなぜだろうか？これを分析するために、自分の知識や経験に基づいて 3 つのおもな原因をあげてみた。

1．寿命の問題　－　高齢者が多い場合には死亡率が高くなるではないだろうか？
2．医療機関の問題　－　医療従事者や施設などが充実していない場合には死亡率が高くなるのではないだろうか？
3．衛生状態の問題　－　衛生状態が悪い場合には死亡率が高くなるのではないだろうか？

原因は他にもあるかもしれないが、この 3 つを考えれば十分だと判断した。人によっては、この問題について 2 つの原因を考えたり、4 つ以上の原因を考えたりするかもしれない。それらは各個人の考えた仮説であり、どの仮説が適切かは実証分析の結果明らかになるものである。

この場合、Y（死亡率）、X（高齢者）、Z（医療機関）、W（衛生状態）とすると、

$$Y = a + b X + c Z + d W + u$$

というモデルが定式化できる。

2. モデルに含まれる変数と実際のデータの対応

定式化されたモデルを実際に分析するためには、データ入手し、変数と対応することが必要となる。このステップは一見簡単なようであるが、もっとも難しいものである。

ここでは、「都道府県別死亡率」の実証分析をおこなうので、原因となる変数（**説明変数**）は都道府県別データである必要がある。

次に、個々の説明変数に対応するデータを選んでいく。たとえば、**医療機関**の充実度合いをあらわすデータとして、①医師数　②病院数　③病床数　…　と、いくつかの候補が考えられる。その中から、**最適かつ入手可能なデータ**を用いる。最適なデータを考えついても、入手可能でないことが多い。

これらのデータは統計資料集から得ることができる。代表的な統計資料集としては
・日本統計年鑑
・民力
などがある。「日本統計年鑑」は時系列データが多く収録されており、「民力」には都道府県別のクロスセクションデータが多く収録されている。

統計資料集以外にも、「交通白書」「運輸白書」などの白書類には多くのデータが掲載されている。これらは図書館の参考図書コーナーに所蔵されている。

インターネット上にデータが収録されていることも多い。とくに、総務省統計局のページ(www.stat.go.jp)には政府作成の各種統計資料が収録されている上、最新の日本統計年鑑のすべての表がExcel 形式で収録されている(http://www.stat.go.jp/data/nenkan/index.htm)。また、「統計でみる都道府県・市区町村」の中にある「社会生活統計指標」の基礎データは、さまざまな都道府県別のデータを収録している(http://www.stat.go.jp/data/ssds/5.htm)。

【 課題 25 】 都道府県別死亡率の分析に用いるデータを、「日本統計年鑑」からコピーし、1 枚のワークシートにまとめよ。

モデルに含まれる変数に、次のようなデータを対応させることにする。
Y … 人口千人当たりの死亡率
X … 高齢者の割合（65 歳以上人口を各県の人口で割り、%であらわしたの）
Z … 人口 10 万対医師数（医療機関の要因として）
W … し尿処理水洗化率（衛生状態の要因として）

📖 手順
① 新しいブックを開き、A 列(A2:A48)に都道府県名を入力する。連続データの作成を使って入力できる。

② 被説明変数 Y に対応するデータを B 列にコピーする。「日本統計年鑑」の「第 2 章　人口・世帯」の 2-23 表「都道府県別出生，死亡，死産，婚姻及び離婚数」（エクセル形式）をクリックし、人口 1000 人あたり死亡率のデータ（K 列）をコピーして、このワークシートの B 列に貼り付ける。

③ 説明変数 X に対応するデータを C 列にコピーする。このデータは「日本統計年鑑」の「第 2 章　人口・世帯」の 2-9 表「都道府県，年齢 3 区分別人口」（エクセル形式）の、平成 17 年の 65 歳以上人口(N 列)を総数(K 列)で割ったものである。N 列をワークシートの G 列、K 列をワークシートの H 列にそれぞれコピーし、ワークシートの C2 セルに **=G2/H2*100** と入力したものを C 列全体にコピーする。

④ 説明変数 Z は「第 21 章　保健衛生」の 21-21 表「都道府県別医療関係者数」のD列、説明変数 W は「第 26 章　環境・災害・事故」の 26-5 表「都道府県別一般廃棄物（し尿及びごみ）処理状況」のAの「し尿処理」のF列、を用いる。それぞれ、ワークシートのD列、E列にコピーする。このブックは **death** という名前で保存しておこう。

3. パラメータの推定と統計量の算出

モデルが定式化でき、対応するデータが入手できたら、分析をするのみである。分析にはExcelの分析ツールが利用可能である。

【 課題 26 】 【 課題 25 】 で作成したワークシートを、分析ツールを使って分析せよ。

4. モデルの検討

分析を行った場合にはその結果を解釈し、その結果と理論が整合的でない場合には、モデルの定式化、分析手法、データのどこが間違っているのか、またはそもそも理論が間違っているのか検討しなくてはならない。分析結果が妥当であると判断された場合、そのモデルは政策や予測などに用いられる。

① 係数推定値
係数推定値では特にその符号に注目する。モデルの定式化の際に想定して符号と分析の結果が一致するかどうかを検討する。

② 決定係数および自由度修正済み決定係数
決定係数や自由度修正済み決定係数が1に近い場合には、この分析をおこなう意味があったと判断できる。
もし、決定係数の値が低い(0.5とか0.6とか)場合には、回帰分析のあてはまりはあまりよくない。ここでは説明変数が不足していることが考えられよう。そこで、被説明変数の原因となるような他の要因を考えることになる。

【 課題 27 】 【 課題 26 】 でおこなった分析結果を検討せよ.

＜分析結果の一部＞

概要

回帰統計	
重相関 R	0.980668
重決定 R2	0.96171
補正 R2	0.959038
標準誤差	0.272696
観測数	47

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	3	80.31175	26.77058	359.998	1.8E-30
残差	43	3.197615	0.074363		
合計	46	83.50936			

	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	0.326943	0.513461	0.636744	0.527667	-0.70855	1.362435	-0.70855	1.362435
X 値 1	0.4468	0.019927	22.4222	2.36E-25	0.406614	0.486986	0.406614	0.486986
X 値 2	-0.00059	0.001135	-0.52099	0.605047	-0.00288	0.001698	-0.00288	0.001698
X 値 3	-0.00584	0.00271	-2.15413	0.036877	-0.0113	-0.00037	-0.0113	-0.00037

① 係数推定値
想定した仮説では、
- 高齢者の割合（X）が多いほど、死亡率（Y）は高い→X(Excelの表記ではX値1)の符号は＋ ⇒ 想定どおり
- 医師数（Z）が多いほど、死亡率（Y）は低い→Z(Excelの表記ではX値2)の符号は－ ⇒ 想定どおり
- 水洗化率（W）が高いほど、死亡率（Y）は低い→W (Excelの表記ではX値3)の符号は－ ⇒ 想定どおり

② 決定係数および自由度修正済み決定係数
決定係数は0.962、自由度修正済み決定係数は0.959と1に近い。