第2章

データの加工法

2.1 はじめに

計量経済分析をするためにデータは不可欠である。鉄が産業のコメだといわれた時代があったが、データは計量経済分析のコメ、いわば素材である。優れた料理人が素材にこだわるように、データの吟味も必要となる。データの入手法や注意点などはほかの解説書（たとえば、筆者が書いたものとしては『経済予測入門』（日本経済研究センター（2000））に譲るとして、加工法などの技術的な話を中心に進める。

2.2 データの種類

経済統計には使用目的によってさまざまな種類のものがある。大きく分けると、国全体の集計量を扱ったマクロデータと個々のデータを扱うミクロデータに分けられる（表2.1）。マクロデータは、国内総生産（GDP）統計や消費者物価指数、失業率などである。ミクロデータは個々のサラリーマンの勤続年数や所得、個別企業の企業収益や株価の動向などである。これは、経済学がマクロ経済学とミクロ経済学に分類されることに対応したものだ。

データの性格によってデータを分類することもできる。時系列データ、クロスセクションデータ、パネルデータである。時系列データは、時間を追って記録したものである。記述されたデータの順番が重要になるGDP統計や鉱工業生産指数などが時系列データだ。クロスセクションデータは、ある時点でのさまざまな主体についてのデータである。都道府県別の失業率のデータや国ごとの成長率のデータである。パネルデータはその両者を組み合わせたものだ。パネルデータの分析は、家計の個票データの分析などに有効だ。個票データはクロスセクションデータとしては豊富にあるが、調査が大変なため時系列的に多くのデータはとれない。こうしたデータを推計する手法としてパネルデータの分析が進んでいる。

データの種類	説明
マクロデータ	国単位などの集計量
ミクロデータ	個別データ
時系列（タイム・シリーズ）データ	時間とともに変わるデータ
クロスセクション（横断面）データ	ある時間の中でのさまざまなデータ
パネルデータ	時系列データと横断面データの組み合わせ

表 2.1: データの種類

本書では、変数の表記としては x や y、z を主に使っている。たとえば、x は所得を示し、y は消費を示す。所得 x は単一のデータではない。1980 年から 1990 年までの時系列データの場合もあるし都道府県別のデータである場合もある。それを明示的に示すために、時系列の場合は x_t と表し、クロスセクションデータの場合は x_i などと表す。x が四半期データなら x_1 は 1980 年 1 − 3 月期、x_2 は 1980 年 4 − 6 月期 ... を表し、それらをまとめて x_t として表す。クロスセクションデータの場合は i は個体番号を表し、x_1 は北海道、x_2 は青森県 ... とし、一般的に x_i などと表す。

本書では時系列データを中心としているので、x_t で表記している場合が多い。また、時系列データであることが明らかな場合は省略して x_t を単に x と表す場合もある。

x_t の 1 期前の値は x_{t-1} と表す。$x(-1)$ と表す場合もあるが同じ意味である。

2.3 データの加工

経済データのうち最もよく使うのは、時間の流れに沿ってデータを持っている時系列データだ。ただ、何も加工していないデータ（原数値と呼ぶ）を見ても、データの意味していることがわかりにくいことが多い。そこで、伸び率をとったり、季節調整をかけたりして、データの特徴を捉えやすくする。

2.4 伸び率

2.4.1 前年比

データが 1 年周期の場合、前年との値を比べた前年比伸び率をとるとデータが持っている情報がわかりやすくなる場合が多い。年度ベースの実質 GDP のグラフをみてみよう（図 2.1）。原数値では傾きが緩やかになったことはわかるが、どの時点でどのくらい傾きが緩やかになったかはわかりにくいし、1 年ごとの動きもとらえにくい。これを前年比伸び率のグラフにしてみると、毎年どのくらい成長していたかがわかるうえ、90 年代に

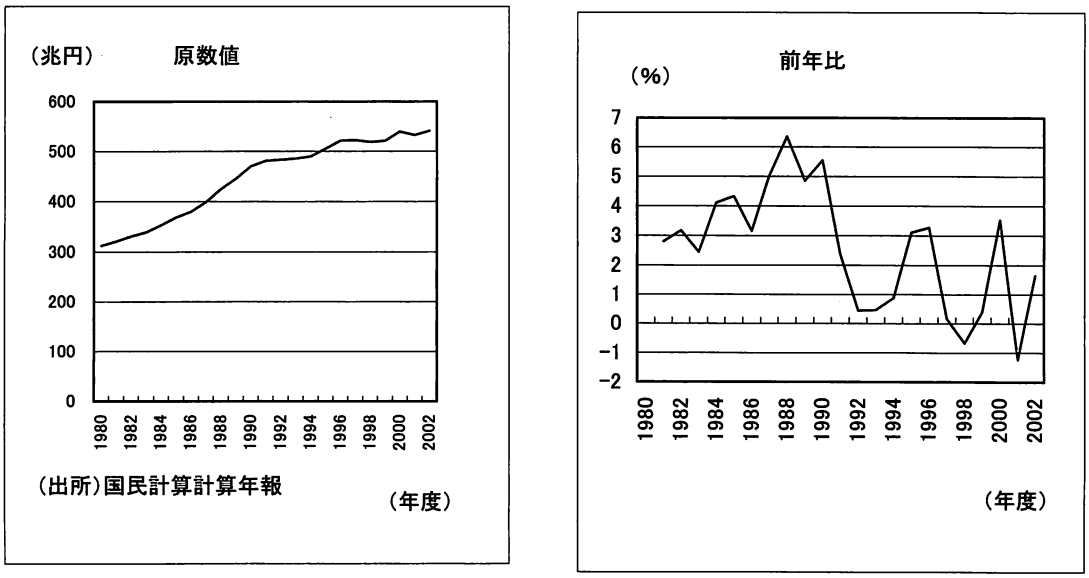

図 2.1: GDP の原数値と前年比

入って成長率が低下した様子がよくわかる。年度データの場合は前年度比とも呼ぶ。実質GDP の年度データの前期比は経済成長率と呼ばれる。前年比の計算法は前期比の計算法と同じである。

2.4.2 前期比・前年比

伸び率のなかで、最も基本的なものである。前期比は前期に比べてどの程度増えたかを比率で表わしたものである。

前年比も前期比の一種である。式で示すと次のように表わされる。

$$\frac{y_t - y_{t-1}}{y_{t-1}} \times 100 = \frac{y_t}{y_{t-1}} \times 100 - 100$$

2.4.3 前年同期比

前年同期比は、新聞紙上でもよく使われるものだ。月次データなら前年と同じ月と比べてどのくらい伸びたかを表す。季節性は毎年同じ時期に表れるものなので、前年同期比をとれば季節性は取り除かれる。月次データの場合は次の式で表される。

$$\frac{y_t - y_{t-12}}{y_{t-12}} \times 100 = \frac{y_t}{y_{t-12}} \times 100 - 100$$

2.4.4　前期比年率

「前期比年率」は四半期データでしばしば使われる。実質 GDP 成長率は年ベースでは「2 %伸びれば望ましい成長である」といったある程度の相場観がある。しかし、四半期データでは、1 − 3 月期と比べた 4 − 6 月期の伸びを表す前期比の数字を聞いてもピンと来ない場合が多い。そこで、「前期からの成長が 1 年間続くとどの程度の伸び率になるか」を計算することが有用だ。この計算は複利計算が使われる。1 四半期分の伸びが 2 %なら、2 四半期分の伸びは、1 四半期で伸びた分（1+0.02）の 2 %分伸びることになる。つまり、$(1+0.02)^2$ である。年率にするには、これを 4 回繰り返すことになる。つまり、$(1+0.02)^4$ である。式で示すと次のようになる。

$$\left(\left(\frac{y_t}{y_{t-1}}\right)^4 - 1\right) \times 100$$

2.4.5　寄与度

寄与度は、伸び率をデータの構成項目ごとに分解する手法である。設備投資のうち、情報技術（IT）関係の投資は大きく伸びている。前年同期比 20 %や 30 %で伸びている場合もある。しかし、絶対額が小さい場合、設備投資全体に対する影響は小さい。寄与度を計算すれば、IT 投資の設備投資全体の伸びに対する影響度がわかる。

x が y の構成項目の一つである時、x の y に対する寄与度は次式で表わされる。

$$\left(\frac{x_t - x_{t-1}}{y_{t-1}}\right) \times 100$$

伸び率の計算

さまざまな伸び率を GDP 統計で実際に計算してみよう（表 2.2）。2002 年 10 − 12 月期の例をとると、当期は 543 兆円、前期は 540 兆 5000 億円なので、前期比伸び率は $(543/540.5 - 1) \times 100$ で計算できる。0.5 %である。前期比年率は、$543/540.5 = 1.0046$ を 4 乗して年率の増加率を求め、それから 1 を引いて 100 をかけることによって計算でき、1.9 %になる。民間需要の寄与度は、民間需要の増加額を前期の GDP で割って求める。つまり、$(404 - 403)/540.5 \times 100$ で 0.3 となる。民間需要と公的需要と海外需要を足せば GDP になるので、それぞれの増加分の和は GDP の増加分に等しい。このことから、寄与度をすべて足したもの（$0.3 + (-0.2) + 0.4$）は GDP の前期比（0.5 %）となる。

		2002年 1－3月	4－6月	7－9月	10－12月
GDP	原数値	530.0	536.7	540.5	543.0
	前期比	0.1	1.3	0.7	0.5
	前期比年率	0.2	5.1	2.9	1.9
民間需要	原数値	392.5	397	403	404
	寄与度	−0.6	0.9	1.0	0.3
公的需要	原数値	125.6	125	125	124
	寄与度	0.2	−0.1	−0.1	−0.2
海外需要	原数値	11.9	14	13	15
	寄与度	0.5	0.4	−0.2	0.4

表 2.2: GDP の前期比と寄与度

2.4.6 要因分解

寄与度と似た手法で、ある系列の要因分解をすることができる。回帰分析（3 章参照）を使って次の式が推計できたとする。

$$y_t = a + bx_t + cz_t + e_t \tag{2.1}$$

x が所得要因、z が価格要因のとき、y の伸び率は次のように分解できる。まず、(2.1) 式から、1 期前の同式を差し引くと、定数項が消える。

$$y_t - y_{t-1} = b(x_t - x_{t-1}) + c(z_t - z_{t-1}) + (e_t - e_{t-1})$$

両辺 y_{t-1} で割れば、伸び率の要因分解ができる。

$$\frac{y_t - y_{t-1}}{y_{t-1}} = \frac{b(x_t - x_{t-1})}{y_{t-1}} + \frac{c(z_t - z_{t-1})}{y_{t-1}} + \frac{e_t - e_{t-1}}{y_{t-1}}$$

前年同期比で要因分解するには y_{t-1} の代わりに四半期なら y_{t-4}、月次なら y_{t-12} を使う。

2.4.7 年平均成長率

長期的な趨勢を見る場合は、年単位のデータだけをみただけではわかりにくい場合がある。たとえば、日本の 80 年代と 90 年代の成長率がどの程度違っていたのかは、年ごとのデータをみるより 10 年単位の平均成長率を比べる方がわかりやすい。成長率を単純に平均しても年平均成長率にはならない。年平均成長率を計算するには、以下の計算をする。

	実質 GDP（10 億円）	平均成長率（5 年）		平均成長率（10 年）	
1980	311988				
1985	368212	1980-1985	3.4		
1990	469567	1985-1990	5.0	1980-1990	4.2
1995	504827	1990-1995	1.5		
2000	539160	1995-2000	1.3	1990-2000	1.4

（出所）国民経済計算年報

表 2.3: 実質 GDP の年平均成長率

たとえば、2000 年から 2005 年までの平均成長率は次の式で表わされる。

$$\left(\left(\frac{y_{2005}}{y_{2000}}\right)^{\frac{1}{5}}\right) - 1 \times 100$$

表 2.3 は実質 GDP の 5 年ごと、10 年ごとの年平均成長率である。たとえば、90 年代の平均成長率を調べるには、その出発点である 1990 年とその終点である 2000 年の GDP が必要だ。それらの比をとると 10 年間でどの程度成長したかがわかる。成長率は複利計算なので、この比率の 10 乗根をとることで 1 年間の成長率を求めることができる。

2.5 弾力性

弾力性（弾性値）は、ある変数が 1 %動いた時にほかの変数が何%動くかを示す。事後的な A の B に対する弾性値は、A の伸び率 ÷B の伸び率という形で計算できる。伸び率には前期比や前年同期比を用いる。ただ、この計算では、A の伸びや B の伸びがほかの要因で動いていてもその影響を排除できない。このため、経済分析で使われる弾性値は、対数線形で推計した回帰分析の係数によって求める（4 章参照）。回帰式に被説明変数のラグがある場合は、長期の弾力性と短期の弾力性が計算され、両者を区別することが重要である（8 章参照）。

2.6 階差

時系列分析では、しばしば「階差」という言葉が出てくる。これは、前期差と同じ意味で次の式で表すことができる。差分という言い方もする。階差は、Δ（または d）という記号で表す。

$$\Delta(x_t) = x_t - x_{t-1}$$

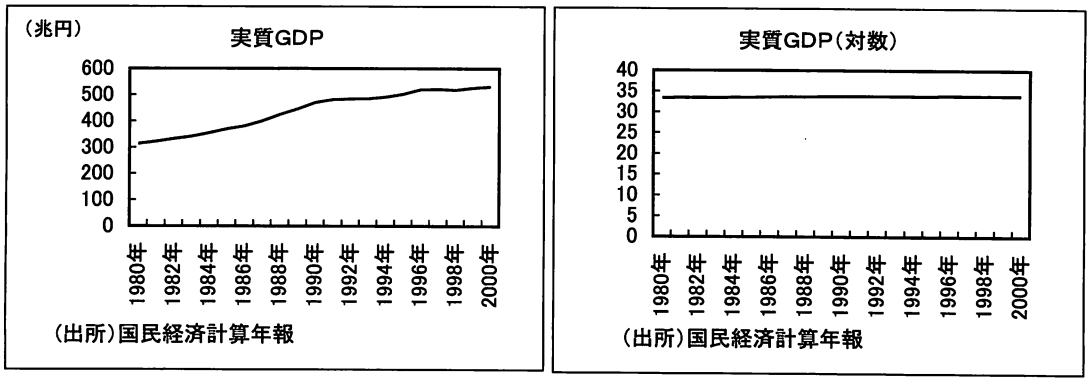

図 2.2: 対数グラフ

対数階差は、推計データの加工としてよく使われる（4 章参照)。階差は時系列モデルを分析する際の基本的な手段の一つであり、系列を定常化する一つの方法である（9 章参照)。

2.7 対数

経済分析ではデータに対数を使うことが多い。(自然) 対数とは、変数を x とすると、それが e の何乗であるかを表す。e のゼロ乗は 1、e の 1 乗は約 2.7、e の 10 乗は 2 万 2026 である。日本の実質 GDP は 500 兆円程度だが、これが e の何乗に当たるかを計算すると約 34 乗であることがわかる。500 兆円を数字で書くとゼロが 14 個も並ぶが、それを 34 という数字で表すことができる（図 2.2)。大きな数字をなじみやすい程度の数字にしてくれるわけだ。対数を使った推計は 4 章、対数の演算については 15 章参照。

2.8 平均

時系列データやクロスセクションデータは数値の羅列であり、数値だけをみてほかのデータと比較したり、解釈を導き出すのは難しい。そこで、データの情報量を集約して、さまざまな判断を下す方法が考えられる。平均や分散はその方法の一つである。

2.8.1 平均

通常平均という時は、データをすべて足したものをデータ数で割ったものである。A 国、B 国の GDP のデータがそれぞれ 100 サンプルあった場合、2 つのデータセットの数字を見ただけでは何の判断もできないが、A 国、B 国それぞれの GDP の平均値をとると、ある程度の判断が可能になる。データの情報を縮約した「代表値」としては最も重要なものである。n 個のサンプルがある場合は次式で表わされる。

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \cdots + x_n)}{n}$$

2.8.2 加重平均

加重平均はデータの重要度に応じてウエートをつけて平均するものである。重要度を示すウェートを $w_1 \ldots w_n$ で表わすと次式となる。

$$\bar{x} = \frac{(w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n)}{w_1 + w_2 + w_3 + \ldots w_n}$$

2.8.3 移動平均

移動平均は時系列データで用いる手法だ。変数の動きを滑らかにする効果がある。時点ごとに平均する対象を移動させるので移動平均と呼ぶ。ある期のデータを中心にして平均をとる場合は「中心移動平均」、ある期よりも過去の系列について移動平均する場合は「後方移動平均」という。系列の動きを見るには中心移動平均が望ましいが、最新期のデータについては中心移動平均は作れない。後方移動平均をとると最新期まで計算できるが、山や谷の位置がずれる。

t 期を中心に前後合わせて 3 期の移動平均をとることを 3 期中心移動平均と呼び、t 期を含めて 3 期分の過去のデータを平均する場合を 3 期後方移動平均と呼ぶ。

3 期中心移動平均 $\bar{x} = \frac{(x_{t-1} + x_t + x_{t+1})}{3}$

3 期後方移動平均 $\bar{x} = \frac{(x_{t-2} + x_{t-1} + x_t)}{3}$

図 2.3 は、対ドル円レートに、5 ヵ月中心移動平均、5 ヵ月後方移動平均をとった例である。

2.8.4 幾何平均

算術平均は、データを「足して割る」ものだが、幾何平均はデータを「掛けて累乗根をとる」ものだ。成長率の平均などに用いる。成長率のデータしか手元にない場合、成長率のデータを比率に変換して（2 %なら 1.02)、幾何平均をとると成長率の平均が計算できる。

$$\bar{x} = (x_1 \times x_2 \times x_3 \times \cdots \times x_n)^{1/n}$$

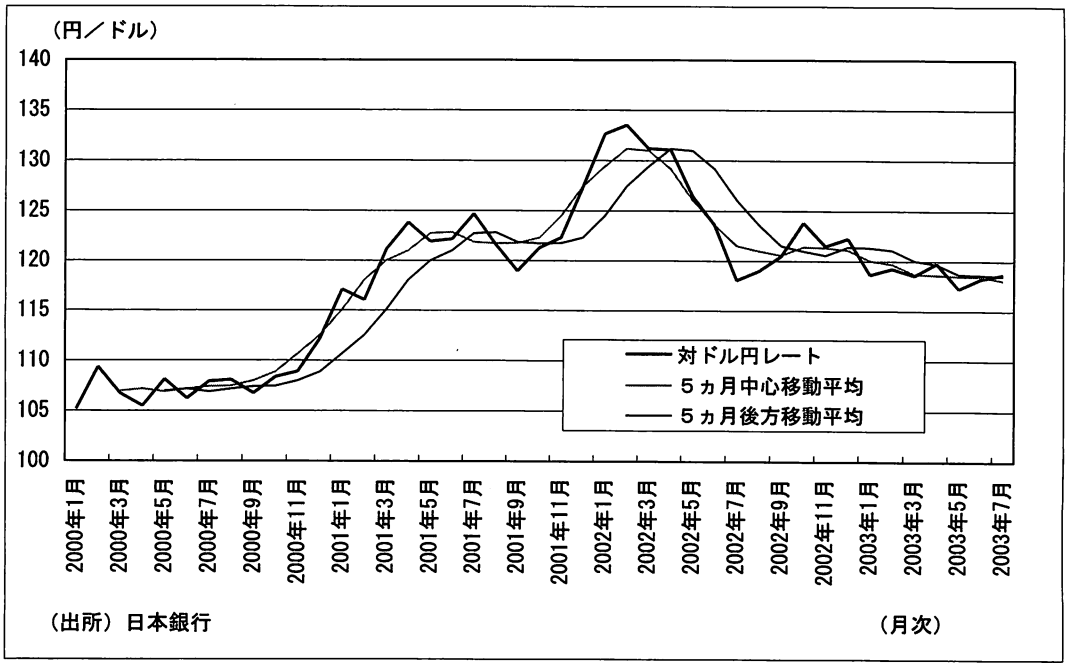

図 2.3: 対ドル円レートのグラフ

2.9 分散

2.9.1 分散

分散はその名のとおり、データの散らばり具合を表す。平均が同じでも、大きく変動するデータと小さく動くデータでは分散の大きさが変わる。データの平均値（$\bar{x}$ とする）からの差（偏差）を二乗したものの平均値だ。

$$v = \sigma^2 = \frac{((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)}{n}$$

2.9.2 標準偏差

分散はデータの二乗の平均をとっているので、もとのデータとは単位が変わる。標準偏差は分散の平方根で、もとのデータと単位が揃う。散らばり具合を表すという意味では分散と同じ概念である。

$$\sigma = \sqrt{\frac{((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)}{n}}$$

計量経済学では標準偏差を σ で表し、分散を σ^2 で表すことが多い。

2.9.3　変動（全変動）

分散の分子の部分を変動（全変動）と呼ぶ。データの平均からの差（偏差）の2乗を加えたものである。決定係数の算出で使う。

$$z = ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)$$

2.9.4　変動係数

変動係数は、標準偏差を平均で割ったものである。標準偏差は同種のデータどうし（たとえばテストの得点）を比べる場合には有効だが、平均値が大きく異なるデータの間では比較できない。大きな数字ばかりのデータの標準偏差は大きくなるし、小さな数字ばかりのデータの標準偏差は小さくなる。変動係数を見れば平均値を基準として散らばり具合を比べることができる。

$$\text{変動係数} = \frac{\text{標準偏差}}{\text{平均}}$$

2.9.5　共分散

共分散という言葉からはどういう統計なのかイメージしにくいが、「2変数の相関の度合い」を示していると考えるのがわかりやすい。ある変数が大きく動いた時にほかの変数も大きく動き、それほど動かない時には同じようにあまり動かない場合に共分散は大きくなる。式では次のように、各変数の平均からの偏差を掛け合わせたものの和をサンプル数で割ったものになる。

$$cov = ((x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y}))/n$$

2.9.6　相関係数

相関係数は、共分散をそれぞれのデータの標準偏差の積で割ったものである。共分散はデータの大きさによって変わるが相関係数はどんなデータをつかっても、マイナス1から1までの間に収まる。相関係数が1の場合は完全に正の相関があり、マイナス1の場合は負の相関がある。ゼロの場合は無相関である。詳しくは2.12節参照。平均、分散、共分散、相関係数について、表2.4にまとめた。

		説明
平均		ならした値
分散		散らばり具合をあらわす
	標準偏差	分散を変数と同じ単位に直したもの
共分散		2 つの変数の相関を表す
	相関係数	2 つの変数の相関を表す（マイナス 1 から 1 に基準化）

表 2.4: 平均と分散のまとめ

$$\text{相関係数} = \frac{x \text{ と } y \text{ の共分散}}{x \text{ の標準偏差} \times y \text{ の標準偏差}}$$

2.9.7 具体例

鉱工業生産指数と実質 GDP の前年比について統計値を比較してみよう（表 2.5、グラフは図 2.4)。1980 年度から 2002 年度までのデータを用いる。両者の平均値をとると、鉱工業生産指数の平均は 1.58 %で実質 GDP の平均の 2.56 %に比べて低いことがわかる。また、鉱工業生産指数の標準偏差は 4.53 で、実質 GDP の標準偏差の 2.03 に比べてばらつきが大きい。グラフでもその様子がわかる。標準偏差を 2 乗した分散で測っても同じ結論である。標準偏差を平均で割った変動係数は鉱工業生産指数が 2.88 に対して実質 GDP は 0.79 であり、基準を合わせた変動の大きさでも鉱工業生産の方が変動が大きいことがわかる。共分散は 7.23 で、共分散と両者の標準偏差を使って計算した相関係数は、$7.23/(4.53 \times 2.03) = 0.79$ となる。両者にはある程度正の相関があることがわかる。

	鉱工業生産（前年比）	実質 GDP（前年比）
平均	1.58	2.56
標準偏差	4.53	2.03
分散	20.55	4.12
変動係数	2.88	0.79
共分散	7.23	
相関係数	0.79	

（出所）経済産業省、内閣府、1980 年度から 2002 年度を使用。

表 2.5: 鉱工業生産指数と GDP の前年比の比較

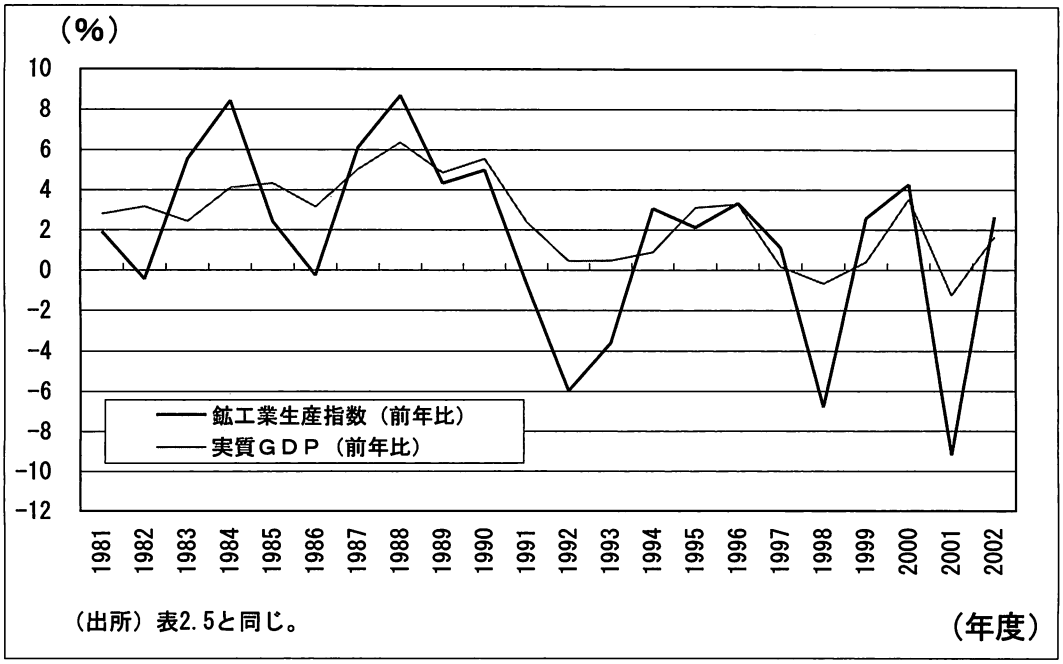

図 2.4: 鉱工業生産指数と実質 GDP（前年比）

2.10 標準化

データにはさまざまな単位があり、ばらつきも大小さまざまだ。そのまま 2 つのデータを並べてグラフを描いても何を意味するかがわかりにくいことがある。そこで、標準化して比べるという手法をとることがある。標準化とは、データを平均ゼロ、標準偏差 1 のデータに変換することである。次の式によって標準化することができ、z 値と呼ばれる。個別のデータを $x_i(i = 1, 2, \ldots)$、平均を $\bar{x}$、標準偏差を σ とすると、次の式で表される。

$$z\text{ 値} = \frac{x_i - \bar{x}}{\sigma}$$

z 値はゼロを中心とした数値になるが、これを平均 50、標準偏差 10 に換算したものが偏差値である。

$$\text{偏差値} = (\frac{x_i - \bar{x}}{\sigma}) \times \quad 10 \quad + \quad 50 = z \times 10 + 50$$

2.11 指数

指数とはある一時点を 100（または 1）として表すものだ。単位の違うデータを比べたり、ある時点を出発点とし、その後の推移を複数データで比べたい時には有効である。す

べてのデータを基準時点のデータで割れば計算できる。

$$95\text{年を}100\text{とした指数} = \frac{y_t}{y_{95}} \times 100$$

また、さまざまな品目のデータを一つの指数にする場合、ラスパイレス指数、パーシェ指数、フィッシャー指数という3種類の方法がある。

指数名	計算式	統計
ラスパイレス	$\frac{\Sigma p_t q_0}{\Sigma p_0 q_0}$	消費者物価指数、企業物価指数
パーシェ	$\frac{\Sigma p_t q_t}{\Sigma p_0 q_t}$	GDP デフレーター
フィッシャー	$\sqrt{\frac{\Sigma p_t q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_t q_t}{\Sigma p_0 q_t}}$	輸出入価格指数

ラスパイレス指数は、基準年のウエートが固定されているので変数の加工が容易である。たとえば、コンピューターと携帯電話の価格を合成した指数を「IT価格指数」として作ろうとする場合を考える。消費者物価指数にしても企業物価指数にしても年報やホームページから各品目のウエートが入手できる。たとえば、パソコンのウエートが w_a、携帯電話のウエートが w_b、パソコンの価格指数が p_a、携帯電話の価格指数が p_b とすると、両者を合成した「IT価格指数」は次の式で表される。

$$IT\text{価格指数} = \frac{w_a p_a + w_b p_b}{w_a + w_b}$$

2.12 相関係数

相関係数とは、2つの変数の関係がどのくらい強いかを表している。同じ動きをすれば1となり、正反対に動けばマイナス1となる。相関係数は単に2変数の関係をつかむのにも有効だが、さまざまな応用もできる。

$$\text{相関係数} = \frac{x\text{と}y\text{の共分散}}{x\text{の標準偏差} \times y\text{の標準偏差}}$$

2.12.1 時差相関係数

時差相関係数は、ある変数とほかの変数の1期前、1期先など時間がずれた変数の相関係数をとったものだ。変数の先行、遅行関係を明らかにするのに役立つ。2つのグラフを

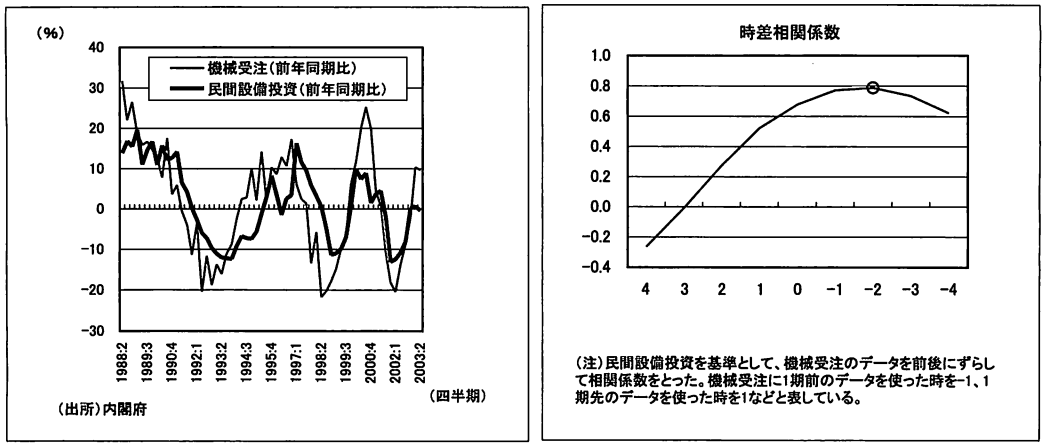

図 2.5: 機械受注と設備投資

並べて、どのくらいずれているのかを眺める作業を数字で表現したものだと考えるとわかりやすい。統計的な因果関係などを表現しようとしているのではないことに注意する必要がある。

たとえば、機械受注が設備投資に対してどのくらい先行しているかを調べようとする場合、設備投資の前年同期比のデータに対し、機械受注の前年同期比のデータを 1 四半期前とか 1 四半期後とか、さまざまにずらして相関係数をとる（図 2.5)。その結果、当期の設備投資のデータと 2 四半期前の機械受注のデータの相関係数が最も高くなった。この分析で機械受注統計は設備投資に対して 2 四半期先行していることがわかる。

2.12.2　自己相関係数

自己相関係数は、ある系列 x_t について自分自身の過去の値との相関係数をとることである。たとえば、1 階の自己相関係数は $x_t(=x_2, x_3, \ldots, x_n)$ と $x_{t-1}(=x_1, x_2, \ldots, x_{n-1})$ との相関係数である。トレンドのある変数の場合は、自己相関は強くなる。自己相関係数は、時系列モデルで、AR や MA の次数がいくつになるかを決める際に使う（9 章参照)。

2.13　主成分分析

主成分分析は、回帰分析とは異なる分析法である。回帰分析では x と y の関係を求めるのに対し、x と y から新たな変数 z を作り出すものである。簡単に表せば次のような違いがある。

回帰分析　$y_i = a + bx_i$

主成分分析　$z_i = ax_i + by_i$

回帰分析は y_i を説明できるように x_i という変数を使い a,b という係数を推計するのに対し、主成分分析は、x_i と y_i の動きのうち共通なもの z_i（主成分）を取り出すように a と b を決める。簡単に 2 変数について説明しよう。各国について x_{1i}、x_{2i} という 2 種類のデータがあるとする。主成分分析は次の式を作成することである。

$$z_{1i} = a_1 x_{1i} + a_2 x_{2i} \tag{2.2}$$

z_{1i} が x_1 と x_2 のそれぞれの特性を最大限反映されたものにするということは、z_{1i} の分散を最大化することと同じである。しかし、この条件だけではさまざまな a_1、a_2 の組み合わせが考えられるため、$a_1^2 + a_2^2 = 1$ という条件をつける。

具体的には次のように計算する。$\mathbf{A}$ を x_{1i} と x_{2i} の相関係数行列、λ を $\mathbf{A}$ の固有値、$(a_1, a_2)'$ を固有ベクトルとする。相関係数行列には対角要素が 1、それ以外の要素は個々の相関係数が入る。

$$\mathbf{A}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{2.3}$$

(2.3) 式から固有値 $\lambda_1, \lambda_2 (\lambda_1 > \lambda_2)$ が求まり、λ_1 が第 1 主成分の固有値、λ_2 が第 2 主成分の固有値となる。λ_1 に対応する固有ベクトル $(a_1, a_2)'$ の要素が (2.2) 式の a_1、a_2 となる。

以上の計算で、(2.2) 式から z_{1i} が各国について計算できる。主成分得点はそれを標準化したものである。主成分得点は次の式で表される。

$$\text{主成分得点} = \frac{z_{1i} - z_{1i}\text{の平均}}{z_{1i}\text{の標準偏差}}$$

寄与率は、各主成分がもとの変数の情報をどの程度反映しているかを表し、第 p 主成分の寄与率は、固有値合計に対する第 p 主成分の固有値の比率で計算できる。累積寄与率は、第 1 主成分から第 p 主成分までの和である。例えば第 1 主成分の寄与率、第 p 主成分までの累積寄与率はそれぞれ次の式で表される。

$$\text{第 1 主成分の寄与率} = \frac{\text{第 1 主成分の固有値}}{\text{固有値合計}}$$

$$\text{第 } p \text{ 主成分までの累積寄与率} = \text{第 1 主成分の寄与率} + \quad \cdots + \text{第 } p \text{ 主成分の寄与率}$$

2.13.1 主成分分析の計算例

主成分分析を使って、各国の IT 化度を測ってみよう。使用したデータは、ITU（電気事業連合）の 2002 年のデータで、携帯電話、インターネットユーザー、パソコンのそれ

	携帯	インターネットユーザー	パソコン
インドネシア	5.5	1.9	1.1
シンガポール	79.1	54.0	50.8
タイ	26.0	7.8	2.8
フィリピン	17.8	2.6	2.2
マレーシア	34.9	27.3	12.6
韓国	68.0	55.2	55.6
香港	93.0	43.1	38.7
台湾	106.5	38.3	39.6
中国	16.1	4.6	1.9
日本	62.1	44.9	38.3
米国	48.8	53.8	62.5

表 2.6: IT 関連統計（人口比、%）

ぞれ人口に占める比率をアジアの各国・地域についてまとめたものだ（表 2.6）。

これらの IT 関連統計から一つの成分を取り出し、それを IT 化度の指標とする。3 つの統計から一つの主成分を取り出すという主成分分析である。携帯電話普及率を x_1、インターネットユーザー比率を x_2、パソコン普及率を x_3 とすると、それぞれの相関係数行列は次の通りである。

$$A = \begin{pmatrix} 1 & x_1 と\ x_2 の相関係数 & x_1 と\ x_3 の相関係数 \\ x_1 と\ x_2 の相関係数 & 1 & x_2 と\ x_3 の相関係数 \\ x_1 と\ x_3 の相関係数 & x_2 と\ x_3 の相関係数 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1.00 & 0.79 & 0.76 \\ 0.79 & 1.00 & 0.97 \\ 0.76 & 0.97 & 1.00 \end{pmatrix}$$

この行列 A について、次式を使い固有値 (λ) と固有ベクトル $(a_1, a_2, a_3)'$ を求める。

$$A \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \tag{2.4}$$

固有値は 2.68、0.29、0.03 の 3 つで、最も大きい固有値 (2.68) に対応する固有ベクトルの要素は、$(0.54, 0.60, 0.59)'$ である。固有値合計のうち第 1 主成分の固有値は 89.4 % を占め、寄与率は 89.4 % である。つまり、第一主成分でもとの変数の約 9 割の情報を反映していることになる。

	第1主成分の主成分得点
シンガポール	1.74
韓国	1.71
米国	1.52
台湾	1.47
香港	1.36
日本	0.87
マレーシア	-0.74
タイ	-1.70
中国	-1.98
フィリピン	-2.00
インドネシア	-2.25

表 2.7: IT 化度の主成分得点

最大固有値の固有ベクトルの各要素は第一主成分を求める際の各指標にかかる係数となる。第一主成分は次の式によって求めることができる。

$$z_1 = 0.54x_1 + 0.60x_2 + 0.59x_3$$

x_1,x_2,x_3 に各国の IT 指標を代入すれば各国について主成分 z が計算できる。たとえば、日本の場合は 0.87 となる。ただ、主成分の値だけでは相対的な位置関係は把握できにくいため、第一主成分を平均をゼロ、標準偏差 1 として基準化して主成分得点を計算する。

主成分得点をまとめたのが表 2.7 である。IT 化度のもっとも高い国がシンガポール、低い国がインドネシアであり、日本はほぼ中央に位置することがわかる。

2.14 まとめ

この章では回帰分析に入る前のデータの加工や変換法に関する基本的な手法を解説した。概要は表 2.8 のようにまとめられる。

項目		内容
伸び率	前期比	前期と比べた伸び率
	前年同期比	季節性が除去できる
	前期比年率	ある月（四半期）の伸びが1年間続いた時の伸び率
	年平均成長率	5年、10年など長期成長率の年間平均
	弾力性	ある変数が1％伸びた時、別の変数が何％伸びるか
	寄与度	構成データの全体のデータへの寄与分
対数		伸び率計算などに使用
平均	平均	最も重要な「代表値」
	加重平均	平均するデータにウエートをつける
	移動平均	時間とともに平均するデータを移動する
	幾何平均	成長率の平均につかう
分散	分散	散らばり具合を表す
	標準偏差	分散の平方根
	変動	分散の分子部分
	変動係数	標準偏差の単位を揃える
	共分散	2変数の相関の度合いを表す
	相関係数	共分散を－1から1に基準化
標準化	z値	平均と標準偏差で標準化
	偏差値	50を平均として標準化
指数	ラスパイレス指数	基準年の数量を基準にする
	フィッシャー指数	比較年の数量を基準にする
	パーシェ指数	2指数の統合
相関係数	相関係数	相関の度合いを示す
	時差相関係数	どれくらいデータが遅れているかを示す
	自己相関係数	自分自身の過去のデータとどの程度相関しているかを示す
主成分分析		複数のデータの共通部分を取り出す

表 2.8: データ加工の種類