慶應義塾大学ビジネス・スクール

# 統計学　ノート (4)

## 1　相関と回帰

### 1.1　線形相関

今まで1つの変数とその分布に関する統計的方法を考えてきたが，統計的な問題では複数の変数に関連するものが多い。2つまたはそれ以上の変数に関するデータ処理の2つの手法を説明する。ここでは，主に2変数について考えるが，この方法は3変数以上の多変数の処理にも拡張できる。

ある問題では，いくつかの変数の相互の関係を知るために，これらの変数を同時に調べることがある。また別の問題では，関心があるのは特定の1変数だけであるが，それをより明らかにするために他のいくつかの変数を同時に調べることもある。これらの2種類の問題は一般にそれぞれ相関および回帰とよばれる。まず，相関について述べる。

相関の問題は，関心のある1組の変数の間に何か関連があるかどうかを知りたいときに起こる。2つの変数 $x$ と $y$ の間の関連を調べるには，まずデータを $xy$ 平面上の点として図示し，関連のだいたいの形を発見することからはじめるべきである。このときのグラフは散布図とよばれる。このグラフを見れば，2つの変数の間に何かきわだった関連があるかどうか，もしあればその関連は近似的に直線的関係とみなしうるかどうかを簡単に見わけることができる。

点の散らばりの一般的な傾向の特徴をつかむには，散布図を見て，点全体が多少のうねりをもつあるなめらかな曲線の両側に散らばる傾向があるかどうか，あるいは，これらの点が1本の直線の両側に散らばっているように見えるかどうか，を見極めるべきである。どちらにも見えるような場合には，ある意味で2つの変数が直線的に関係している度合を測定することが望ましい。このような尺度をつくり出すために，それがもつべき望ましい性質を考える。

関連の尺度はそれぞれの変数の原点の選び方に無関係でなければならない。この性質を実現するには，変数をそのまま用いるより，それぞれの変数の平均からの偏差を用いたほうがよい。望ましい関連の尺度をつくるには変数 $x_i, y_i$ の代わりに $x_i - \bar{x}, y_i - \bar{y}$ を使うべきである。ここで記号 $x_i, y_i$ は第 $i$ 番目の1組の数字を表す。

関連の尺度は，$x, y$ に対して用いた測定の尺度にも無関係でなければならない。この

性質を実現するには，$x, y$ をこれらと同じ単位をもつ量で割ればよい。後で述べる理由から，この量として 2 つの標本標準偏差 $s_x, s_y$ を用いる。以上から変数 $x_i, y_i$ を $u_i = (x_i - \bar{x})/s_x$，$v_i = (y_i - \bar{y})/s_y$ に直して関連の尺度を構成することにすれば，両方の性質が満たされることになる。

（データの大部分の点は第 1 象限と第 3 象限にあって，第 2 象限と第 4 象限の点より総体的に座標の大きさが大きいような）散らばりの特性をはかる簡単な尺度として，和 $\sum_{i=1}^{n} u_i v_i$ が考えられる。この和の各項のうち，第 1 象限と第 3 象限内の各点に対応するものは正になるのに対して，第 2 象限と第 4 象限内の各点に対応するものは負となる。従って，この和が正の大きい値をとるならば，散布図は強い直線の傾向を示していると考えられる。しかし散らばりの状態をそのままにして点の数を 2 倍にしたとすれば，和の値も 2 倍になるから，上で述べたことは厳密には正しくない。よって，この和を関連の尺度として用いるには，それを点の数 $n$ で割っておく必要がある。標本標準偏差を定義した場合と同様に，ある理論的理由から $n$ ではなく，$n-1$ で割ったほうがよい。こうして得られた和 $\sum u_i v_i/(n-1)$ が求める関連の尺度である。これを相関係数といい，$r$ で表す。元の測定値で表すと，相関係数は次の式で定義される。

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{1}$$

$r$ の絶対値は直線的関係の強さを測るが，$x$ が増すと $y$ も増す傾向にあれば，$r$ の符号は正となり，$x$ が増すと $y$ が減る傾向にあれば，負となる。$r$ は 2 つの変数が直線的に関連するときのみ変数間の関係の強さの尺度として有効であり，$x$ と $y$ が密接に関連していても，その関係が直線的でないときには用いることはできない。

$r$ の 2 つの性質がある。$r$ の値は不等式

$$-1 \le r \le 1$$

を満たさなければならないこと，そして $r$ の値はすべての点が一直線上にあるときのみ+1 または-1 に等しくなることである。これらの性質は数学的にも証明できる。

(1) の分子は整理すると $\sum xy - n\bar{x}\bar{y}$ となり，$r$ を計算するとき，これを使うと便利である。

## 1.2　$r$ の意味

相関係数を 2 変数間の直線的関係の強さの尺度と解釈することは，純粋に数学的な解釈であって，これら変数の間に何らかの因果関係があるという意味ではない。2 つの変数がともに増加または減少の傾向にあるということは，一方が他方に直接または間接に何らかの影響を及ぼしているということを意味するものではない。2 変数がともに別の変数によ

って影響を受け，そのために 2 変数間に強い数学的関係が生じたのかもしれない。もし相関係数が 2 変数間の関係について何らかの情報を与えるようなときには，その取り扱いに慎重を期さなければならない。相関係数を有効に利用するには，その数学的性質のみならず，適用分野の知識にも精通している必要がある。

### 1.3　$r$ の信頼性

$r$ の値は繰り返し実験で同様な標本値の系列 $r_1, r_2, r_3, \ldots$ における最初の標本値と考えられる。これら標本値の集まりは母集団から大きさ $n$ の無作為標本を抽出して得られたものと考えられる。繰り返し実験で得られた多数の $r$ の値を度数分布表に分類し，ヒストグラムを描いたとすると，$r$ の極限度数分布（または理論度数分布）のよい近似が得られるであろう。$r$ の極限分布は変数 $x, y$ に関して適当な仮定をすれば，数学的に導くことができる。例えば，$x, y$ が独立な正規変数である（このとき，$x, y$ は必ず無相関になる，つまり $x, y$ の相関係数は 0 になる）と仮定すれば，$r$ の標本分布は $n$ だけに関係することが証明される。この $r$ の分布から，$r$ の標本値がどのくらいならば，$x, y$ が無相関であるとの主張，すなわち $x, y$ の理論相関係数 $\rho$ が 0 であるとの仮説，を棄却できるかが決められる。$r$ の分布の棄却限界値，つまり $r$ がこの値を超える確率が $\alpha$ となる確率は数表で与えられている。$\rho = 0$ の検定では，$\rho$ が 0 でないならば，$\rho > 0$（または $\rho < 0$）という確たる理由がない限り，両側検定を使わなければならない。したがって，$\alpha = 0.025$ に対する限界値有意水準 0.05 の両側検定を与える。

$r$ の値が大きくないときに，2 つの変数間に相関がないという主張を否定できるためには，大標本でなければならない。相関係数は変数間の相互関係を調べる道具として有効なことも多いが，これらの変数を解析する量的な道具としては，その信頼性と解釈の面で問題があることにも注意する必要がある。

### 1.4　直線回帰

2 つまたはそれ以上の多くの変数間の関係を調べるのは，この中の特定のある変数の推定なり予測を行うとき，それらの関係を利用したいからである。予測の問題を取り扱うために考え出された方法として，回帰法がある。

データをグラフにプロットしたとき，データに与えられた範囲の $x$ の値に対して $x$ と $y$ がほぼ直線的に関係していることがわかったとする。このとき，$x$ の値から $y$ の値を予測するには，点の集まりに直線を当てはめるのがよいと考えられる。実験の繰り返し数が多ければ多いほど，直線に対して期待される精度はますます大きくなる。したがって，関係が直線的であるという仮定が妥当であれば，標本直線値は理論直線値に近いと期待される。ただし，$x$ と $y$ の間の関係が直線的であるということはデータの範囲の $x$ に対してのみ仮定されていることであるから，この範囲外の $x$ の値に対する $y$ の値の予測に標本直線値を

用いるのは正しくない。

### 1.5　最小2乗法

これまでの議論から，直線による予測の問題は，点の集まりに1本の直線を当てはめる問題に帰着する。一般に，直線の方程式は次のように書ける。

$$y = a + bx \tag{2}$$

ここで$a$と$b$は直線を定める係数，$a$は直線が$y$軸を切る点を定める係数であり，$b$は直線の傾きを定める係数である。

問題は点の集まりにうまく直線が当てはまるように，係数$a$と$b$の値を決めることであるから，この問題は本質的には何らかの有効な方法で係数$a$と$b$を推定する問題であるといえる。これらの係数を推定する方法はいろいろあるが，回帰の問題で最もよく知られている方法は最小2乗法である。

求める直線は予測のために使うのであるから，予測の誤差は小さくするような直線を望のが合理的である。予測の誤差とは，$y$の観測値をこれに対応する$y$の直線値との差のことである。直線より上にある点は正の誤差を与えるが，下にある点は負の誤差を与える。したがって，単に誤差の和をできるだけ小さくするという要求は意味がないので，誤差の絶対値，すなわち誤差の大きさの和をできるだけ小さくすることを要求するとよいだろう。ただ，絶対値の和は数学的な取り扱いが不便なので，絶対値の代わりに誤差の2乗の和をできるだけ小さくする方法を考える。誤差の2乗和を最小にするような係数の値が，最小2乗の意味で最も当てはまりのよい直線を決定する。

最小2乗法によって$a$と$b$の値を決定する問題についてここでは取り上げず，結果のみ示しておく。変数$x$が標本平均$\bar{x}$から測られるならば，結果の導出がいくらか簡単になる。このときの直線の方程式は次の形に書ける。

$$y' = a + b(x - \bar{x}) \tag{3}$$

$y$に(′)をつけたのは，任意の$x$に対する$y$の直線値と$y$の観測値とを区別するためである。直線の方程式が(3)の形のときでも$b$は直線の傾きを表す係数であるが，$a$は$x = 0$のときの$y$の値ではなく，$x = \bar{x}$のときの$y$の値である。直線を(3)の形で表したとき，最小2乗法による$a$と$b$の推定値は次のようになることが数学的に証明できる。

$$a = \bar{y},\ \ b = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

この公式における$x_i, y_i$は$n$組の$x, y$の値があるときの$i$番目の組の値を表す。公式から求めた$a, b$の値を(3)に代入すれば，求める最小2乗直線が得られる。この直線のことを慣例上$x$に対する$y$の回帰直線とよんでいる。以上の結果を次のようにまとめる。

$$y' = \bar{y} + b(x - \bar{x}),\ \text{ただし},\ b = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \tag{4}$$

これは $b = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ のようにも書ける。

計算のためには，$b$ に対する式を次の形に少し変形すると便利である。

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \tag{5}$$

この問題は 2 つの変数に対する相関法と回帰法の基本的な相違を非常によく説明している。相関の問題では，データは $x$ と $y$ は両方とも確率変数で，その値は標本が得られたあとでのみ定まることを意味する。しかし回帰の問題では，$x$ の値は前もって選ばれた値で，$y$ の値のみが標本によって決まる。点の集まりに対して 1 つの直線を当てはめる手法としての最小 2 乗法は，$x$ の値が前もって固定されたときも，無作為標本から得られたときもどちらの場合にも適用できる。したがって回帰法は，相関を考えるときのデータにも適用できる。他方，$x$ の値が意識的に選ばれているとき，$r$ の値は一般に $x$ の選択に強く依存するので，2 変数間の直線的関係の強さの尺度をしての意味をもたなくなる。

相関係数は 2 変数間の直線関係がどの程度強いかを記述するのには役立つが，直線関係以外の他の関係の記述にはほとんど役立たない。相関係数は回帰と連合しない限り，単独で量的な結論を下すには適さないものである。このように相関は一般に 2 変数間の関係を調べる際の 1 つの局面にすぎないのに対して，回帰はこれらの研究における基本的な手法である。

回帰という名前は一般に最小 2 乗直線に対して与えたものであるが，異なる 2 つの時点でとられた観測値を解釈するときに，誤って用いられることがしばしばある。この誤りを回帰の錯誤という。

## 1.6　推定値の標準誤差

点の集まりに回帰直線を当てはめれば，それを見ることで，回帰直線が $y$ の値をどの程度正確に予測するかを知ることができる。予測の精度を計算するには，まずすべての誤差 $y_i - y_i'$ の値をグラフ用紙の上で読みとるか，あるいは公式 (4) によって求めなければならない。予測の精度を測る有効な尺度として，予測の誤差の 2 乗の平均が使われる。これは次の式で与えられる。

$$\frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n}$$

母数 $a$ と $b$ に対する信頼限界を求める問題では，誤差平方和を $n-1$ で割るより $n-2$ で割ったほうがよいことが理論的に知られている。理論回帰直線の周りの分散を $\sigma^2$ とするとき，誤差平方和を $n-2$ $で割った結果は，$\sigma^2$ の不偏推定値であることが証明できる。その平方根を $s_e$ で表し，これを推定値の標準誤差とよぶ。すなわち，

$$s_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - y_i')^2}\ 。 \tag{6}$$

予測値 $y_i'$ をとくに求める必要がなければ，$s_e$ の計算には次の式を用いたほうが簡単である。この式は(6)を変形して導くことができる。

$$s_e = \sqrt{\frac{\sum y_i^2 - a\sum y_i - b\sum x_i y_i}{n-2}} \tag{7}$$

この式で必要な数値はすべて回帰直線式を求める計算で得られたものである。ただし，ここでの $a$ は回帰式を $y' = a + bx$ で表したときの $a$ であるから，$a$ を求める式は $\bar{y}$ ではなく，$a = \bar{y} - b\bar{x}$ であることに注意する。

理論回帰直線(最小 2 乗直線はこれの推定値である)の存在を仮定し，さらに，$y_i - y_i'$ (ただし $y_i'$ は $x = x_i$ に対する理論回帰直線値)が互いに独立で，平均 0，同じ標準偏差 $\sigma$ の正規分布に従うと仮定すれば，(6)の $s_e$ は $\sigma$ の推定値を与え，さらに，正規分布の仮定より予測の誤差に関する近似的な確率的表現を与えることもできる。たとえば，予測の誤差が $1.96s_e$ より小さい確率は約 95%であるといえる。ここでの近似は，$1.96\sigma$ をその標本推定値 $1.96s_e$ でおきかえたこと，また理論回帰直線でなく標本回帰直線を用いたことから生じたものである。すなわちこの手法は大標本法である。大部分の回帰の問題は小標本に関するもので，この手法の使用には多少の制約がある。

観測された $x$ の値の範囲外の $x$ に対する $y$ の値を予測する問題は，範囲内の $x$ の値に対する予測の場合に比べて，かなり難しい問題である。観測値の範囲外の予測を補外といい，一方，範囲内の予測を補間という。補外の難しさはその正しさを立証するのに必要な仮定が，現実の場面ではほとんど実感できないという点にある。$x$ の値の範囲外においても，例えばもし $x$ として十分大きい値が選ばれたとき直線的であることには疑問がある。さらに実際の回帰問題では観測値の範囲外の $x$ の値に対してまでも，予測の誤差が回帰直線の周りに一定な標準偏差の正規分布に従うとみなすことは多くの場合不合理である。

データの散布図を描いたとき，$x, y$ 間の関係が直線的でなくても，$x$ の関数や $y$ の関数を適当に考えて，これら 2 つの関数値間の関係が直線的であるようならば，直線回帰モデルが依然として使える。直接 $x, y$ の関係を考えるのではなく，それぞれの関数の関係を考えるこの方法は，直線回帰モデルの適用範囲をかなり広くしている。

## 1.7　小標本法

予測の誤差に関する確率的表現を与えるためには，最小 2 乗直線をその標本推定値とする理論回帰直線の存在を仮定し，さらに分布に関しても若干の仮定をする必要がある。この理論回帰直線を

$$y' = \alpha + \beta(x - \bar{x})$$

の形で表す。ここで，$\alpha, \beta$ は標本回帰直線における $a, b$ と同じ役割をする。$a, b$ の値は

$\alpha, \beta$ の推定値である。母数 $\alpha$ に関する推定や仮説の検定の問題では，$\alpha$ の推定値は $\bar{y}$ なので，正規分布の平均値に関する小標本法が適用できる。しかし，傾きを与える母数 $\beta$ の推測の問題では新しい公式が必要になる。しかし，$\beta$ に関する問題ではスチューデントの $t$ 分布が適用できることが証明されている。もし $y_i - y_i'(i = 1,2,\ldots,n)$ が互いに独立で，かつ平均 0，同一の標準偏差 $\sigma$ の正規分布に従うならば，変数

$$t = \frac{b - \beta}{s_e}\sqrt{\sum (x_i - \bar{x})^2} \tag{8}$$

は自由度 $\nu = n - 2$ のスチューデントの $t$ 分布に従うことが証明できる。

## 2　重回帰

### 2.1　重線形回帰

予測の実際問題では，2 個以上の独立変数を用いて予測を行うことが多い。1 つの変数を他の 1 つの変数でなく，いくつかの変数によって予測する問題を取り扱う方法も，1 つの変数の場合と同様である。問題は 3 次元での点の散布図に，最小 2 乗法の意味で最も適切な平面を求めることが必要になる。

変数 $y, x_1, x_2$ によって決まる 3 次元平面の任意の平面の方程式は

$$y = a_0 + a_1 x_1 + a_2 x_2 \tag{9}$$

の形に書けるから，問題は 3 つの母数 $a_0, a_1, a_2$ を最小 2 乗法で推定することである。ここで予測の誤差は平面からの点の垂直距離である。これらの誤差の 2 乗和を最小にすることは単直線回帰の場合と同じ手法で，数学的に処理できる。$n$ 個の点が利用できるならば，最小 2 乗法による $a_0, a_1, a_2$ の推定値は，次に示す 3 元連立 1 次方程式を解いて得られる。

$$\begin{aligned} a_0 n + a_1 \sum x_1 + a_2 \sum x_2 &= \sum y, \\ a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_1 x_2 &= \sum x_1 y, \\ a_0 \sum x_2 + a_1 \sum x_1 x_2 + a_2 \sum x_2^2 &= \sum x_2 y. \end{aligned} \tag{10}$$

この結果は変数の数がもっと多い場合にも一般化できる。一般に，回帰の問題で現れるこのような連立 1 次方程式を正規方程式という。

回帰問題における最小 2 乗法の魅力的な一面は，回帰式の母数の推定値を求めるための方程式が簡単に書けることである。

従属変数 $y$ に関係のある重要な変数をできるだけ多く取り入れた回帰関数の注目すべき利点の 1 つは，これらの変数の中の 1 つを除いて，他のすべての変数を固定させるとき，この 1 つの変数が $y$ にどの程度影響するかを調べることができることである。すなわち，すべての変数が同時に変化するときには分析が困難な，さまざまな関係を詳しく分析することができる。回帰関数は平均的な関係を示すもので，単線形回帰のときに与えた仮定と同じものに基づいている。

考慮する変数の数を変えたとき，係数が変化する。このとき，$y$の予測においてどちらのほうがよりよい精度で予測できるかという問題を考える。この問題の解決には，それぞれに対する推定値の標準誤差を計算しなければならない。

コンピュータによって多数の変数を取り入れることができる。しかし，多数の変数を含む回帰関数が実際に意味のある少数の変数を含む回帰関数より有利な場合はそれほど多くはない。また，複雑な回帰関数を用いる際の困難と余分な費用を考えると，簡単な回帰関数を用いたほうが有利かもしれない。したがって，従属変数$y$を予測する線形回帰関数をつくるために多数の変数の中から重要ないくつかの変数をいかにして選ぶかという問題が起こる。

このための1つの方法は逐次選択法とよばれるもので，次のような手順によるものである。まず，独立変数のそれぞれに対する$y$の線形回帰式を求め，これらの回帰式ごとに推定値の標準誤差を計算する。このとき最小の標準誤差を与える変数を，第1の変数として選ぶ。次に2つの独立変数(そのうちの1つは前の手順で選ばれた変数，他の1つは残りの独立変数の1つ)のそれぞれに対する$y$の重回帰式を求め，前と同様に，各重回帰式ごとに推定値の標準誤差を計算して，最小値を与える変数を第2の変数として選ぶ。このように，前に選んだ変数とすべての残りの独立変数の1つを加えた重回帰式をつくっていき，一定個数の変数が得られるまで，または推定値の標準誤差の値が目立った減少をしなくなるまで続ける。ある段階で推定値の標準誤差の値が目立った減少をしなくなったなら，さらに変数を追加してもこれ以上改善されることはないだろう。

多数の独立変数を含む回帰関数の取り扱いには，細心の注意を払わなければならない。なぜなら，これら変数のいくつかのものの間にはかなり強い相関があるかもしれないからである。逐次選択法は初期の段階では高度に相関のある独立変数を選ぶことにはならないが，この方法をあまり長く続けると，高度に相関のある変数さえも最良なものとして選んでしまう。多くの変数を必要以上に含めると，正しくない解釈に陥ることに注意しなければならない。

## 2.2 非線形回帰

2つの変数$x, y$間の非線形な関係も，その多くは$x, y$について適当な関数を選んで変換することにより，線形な関係に変形することができる。これは理論的にはつねに可能であるが，そのときの関係は非常に複雑になるか，または何が適切な関数であるか決めることができないものである。そのような場合には，$x, y$間の関係を非線形のまま処理した方がよいことも多い。回帰モデルとして直線では不十分なときに用いられる最も簡単な曲線は放物線である。その式は次のように表される。

$$y = a_0 + a_1 x + a_2 x^2 \tag{11}$$

最小2乗法によって$xy$平面上の点の集まりにこの種の曲線を当てはめる問題は，3次元

空間の点の集まりに重回帰式を当てはめる問題と同じである。1 組の点の集まりに放物線を当てはめるとき，係数$a_0, a_1, a_2$を与える正規方程式は次のようになる。

$$\begin{aligned} a_0 n \quad &+ a_1 \sum x + a_2 \sum x^2 = \sum y, \\ a_0 \sum x &+ a_1 \sum x^2 + a_2 \sum x^3 = \sum xy, \\ a_0 \sum x^2 &+ a_1 \sum x^3 + a_2 \sum x^4 = \sum x^2 y. \end{aligned} \tag{12}$$

点の集まりに対して直線より放物線のほうがどれほど当てはまりがいいかをみるためには，それぞれごとの予測の誤差を計算し，それを比較すればよい。ただ放物線の場合には (6) の分母の$n-2$を$n-3$として計算しなければならない。2 つの推定値の標準誤差の間にわずかの差しかないならば，放物線を用いることによる利点はほとんどないと考えられる。なお，この方法はさらに高次の多項式によって表される曲線に拡張できる。

いくつかの独立変数に基づく重線形回帰関数を解釈するときには，変数間に強い相関がありうるので，慎重に取り組まなければならない。線形性の仮定が満たされないような問題には別の困難がある。各$y$の平均値は 1 つの平面上にないと信じることができるような理論的もしくは経験的証拠があるならば，独立変数の 1 次の項の他に 2 次の項を取り入れた，より一般的な回帰関係を考えるべきである。そのようなときには，線形回帰関数 (9) は非線形回帰関数

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_1 x_2 + a_5 x_2^2 \tag{13}$$

に取り替えられる。

2 次の項を含ませることは回帰関数における項の数を急に増やすことになるので，独立変数の数が非常に少ないとき以外はあまり使われない。

2 つの変数のそれぞれについて適当な関数を選ぶことによって，2 つの変数間の非線形な関係を禁じてきまたは正確な線形関係に変形することが可能な場合がよくある。重回帰問題でも同じことがよく使われる。この手法は，それを使わないと非常に複雑になる関係を単純化するのに用いられる。

## 2.3 信頼性

重回帰関数または非線形回帰関数の信頼性を求める公式は，単線形回帰の場合とほとんど同じである。推定すべき係数の数を考慮に入れなければならないので，この場合の推定値の標準誤差は単回帰の場合をわずかに修正したものとなり，次の式で定義される。

$$s_e = \sqrt{\frac{1}{n-k} \sum_{j=1}^{n} n(y_j - y'_j)^2} \tag{14}$$

ここで，$k$は回帰式の中の係数 (定数項を含む) の数で，$y'_j$は$y_j$に対応する回帰関数値である。たとえば，(9) と (11) では$k=3$で，(13) では$k=6$である。単線形回帰と同種の仮定のもとで，推定値の標準誤差は回帰関数によって$y$を予測するとき，どのくらいの大

きさの誤差が予想されるかを示すために用いられる。

理論回帰係数の推定値としての標本回帰係数の信頼性を求める公式は，単線形回帰の場合よりかなり複雑である。これらの信頼性の問題を解くために利用できる，重回帰に対するコンピュータ・プログラムができている。しかしこれらを使うには注意が必要である。つまり，単線形回帰で用いた手法は重回帰や非線形回帰においても使えるが，理論を正当化するのに必要な仮定が実際に満たされていないならば，かなり問題であろう。

**参考文献**

[1] 初等統計学 原書第4版，P.G.ホーエル，培風館(1981)。

[2] 統計学入門，東京大学教養学部統計学教室編，東京大学出版会(1991)。