第 3 章 重回帰分析 multiple correlation analysis

浅川 伸一

2006 年 11 月 09 日

1 単回帰分析の拡張

二つの独立変数 independent variable(説明変数、予測変数とも言う) x_1, x_2 から、従属変数 dependent variable (目的変数とも言う) y を予測することを考える。

個々の独立変数 x_1, x_2 と従属変数 y_i との間に $y_i = b_0 + b_1x_1 + b_1 + e_i$ という (線形な) 関係があると仮定されるときパラメータ b_0, b_1, および b_2 を推定する問題を考える[1]。

1.1 説明変数が 2 つの場合

この場合、相関係数が 3 つある。それぞれの独立変数 x_1, x_2 と、従属変数 y との相関係数を r_{x_1y}, r_{x_2y} と表す。独立変数 x_1 と x_2 との相関係数を $r_{x_1x_2}$ とする。

x_1, x_2 それぞれ単独で y を予測するよりも、x_1 と x_2 との両者を使って従属変数 y を予測した方が予測の精度が上がると期待される。この他に x_1, x_2 の両者と y との相関係数を r_{yx1x2} あるいは r_{y12} と表記する。単回帰の場合にならって次の用語を用いる。

重回帰係数 multiple regression coefficient x_1 と x_2 の合成得点から y を予測するときの回帰係数

重相関係数 multiple correlation coefficient x_1 と x_2 の合成得点と y との相関係数

重相関係数の定義は次のとおり。

$$r_{y\cdot 12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}} \tag{1}$$

[1]独立変数が一つときは、中学数学との対応から $y = ax + b$ と記していたが、この例ではパラメータを a, b から b に変更していることに注意

この他に、3 者関係に独特な概念がある。x_1 から x_2 の影響を取り除いた成分 $x_1|x_2$ と y との相関係数 $r_{y(1|2)}$ あるいは $r_{y(x_1|x_2)}$ を

部分相関係数 semipartial correlation coefficient[2] と呼ぶ。$r_{y(x_1|x_2)}$ のように下付き添字に下付き添字がついて小さくなりすぎる場合には省略して $r_{y(1|2)}$ と表記することがある。

部分相関係数 $r_{y(1|2)}$ では x_1 から x_2 の影響を取り除いてあるが、目的変数 y については x_2 の影響を取り除いていない。y からも x_2 の影響を取り除き x_1 からも x_2 の影響を取り除いた上で相関係数を求めることができれば x_2 という変数に影響されない x_1 と y との関係が現われることになる。これを

偏相関係数 partial correlation coefficient と呼び $r_{y1|2}$ と表す。

$$r_{y1|2} = \frac{r_{y1} - r_{y1}r_{12}}{\sqrt{1-r_{y2}^2}\sqrt{1-r_{12}^2}} \tag{2}$$

偏相関係数 $r_{y1|2}$ と部分相関係数 $r_{y(1|2)}$ との間には

$$r_{y1|2} = \frac{r_{y(1|2)}}{\sqrt{1-r_{y2}^2}} \tag{3}$$

という関係がある。

次に相関係数の算出でなく、回帰について考える。単回帰の場合の回帰係数は

$$b = r\frac{s_y}{s_x} \tag{4}$$

であった。誤差 (ベクトル、あるいは残差ベクトル) と予測ベクトルとは直交し、長さの 2 乗は三平方の定理から $s_y^2 = s_e^2 + s_{\hat{y}}^2$ が成り立つことに注意すれば、$s_e^2 = s_y^2\left(1-r^2\right)$ となる。この式は $\boldsymbol{y}$ から $\boldsymbol{x}$ の影響を取り去った後の分散 (距離の 2 乗) は、元の分散に 1 から相関係数の 2 乗を引いた値の開平をかけたものである、と言い表せる。相関係数の 2 乗のことを**決定係数** coefficient of determination と呼び、従属変数の全分散のうち、独立変数で説明できる割合を表す量となる (前章のハンドアウトも参照のこと)。

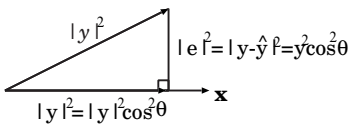

図 1: Pythagorean theorem

[2]part correlation coefficient とも

このことから、

$$s_{1|2} = s_1\sqrt{1-r_{12}^2} \tag{5}$$

である。この値を回帰式に代入すると

$$\begin{aligned} b_{y(1|2)} &= r_{y(1|2)}\frac{s_y}{s_{1|2}} && (6)\\ &= \frac{r_{y1}-r_{y2}r_{12}}{\sqrt{1-r_{12}^2}}\frac{s_y}{s_2\sqrt{1-r_{12}^2}} && (7)\\ &= \frac{r_{y1}-r_{y2}r_{12}}{1-r_{12}^2}\frac{s_y}{s_1} && (8) \end{aligned}$$

となる。これを**偏回帰係数** partial correlation coefficient と呼ぶ。

単回帰係数の場合は問題にならないことが多いが、回帰係数 $r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$ (前回のハンドアウト (14) 式) は標準偏差の単位に依存して大きくなったり小さくなったりするので、あらかじめ各変数を分散 1 となるように変換しておいてから回帰係数を求めることがある。この場合回帰係数は $\frac{s_{xy}}{s_x^2}$ で相関係数に一致する。同様に重回帰において、あらかじめ各変数を基準化しておけば偏回帰係数は

$$\begin{aligned} b_{y(1|2)^*} &= \frac{r_{y1}-r_{y2}r_{12}}{1-r_{12}^2}\frac{s_y}{s_1} && (9)\\ &= \frac{r_{y1}-r_{y2}r_{12}}{1-r_{12}^2} && (10) \end{aligned}$$

となる。これを**標準化偏回帰係数** standard partial regression coefficient と呼ぶ。

1.2 解析的な解

y_i の値の予測値 $\hat{y}_i$ を x_1 と x_2 を用いて $\hat{y}_i = b_0 + b_1x_1 + b_1$ としたとき、i 番目のデータとその予測値との差 (誤差) を

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x_{1i} + b_2x_{2i}) \tag{11}$$

と表記する。n 組のデータが得られたとき全ての誤差の自乗和を最小にするパラメータの値を推定することを考える。誤差の自乗和を $Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ とし、これを各パラメータについて微分して 0 とおくことで以下の式を得る。

$$\begin{pmatrix} \frac{\partial Q}{\partial b_0} \\ \frac{\partial Q}{\partial b_1} \\ \frac{\partial Q}{\partial b_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y-\hat{y}_i)^2 \\ \frac{\partial}{\partial b_1}\sum_{i=1}^{n}(y-\hat{y}_i)^2 \\ \frac{\partial}{\partial b_2}\sum_{i=1}^{n}(y-\hat{y}_i)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{12}$$

推定すべき未知パラメータの数と方程式の数が同じ 3 つであるので、この連立方程式は特別な場合を除いて解くことができる。これを解くと以下の解が得られる。

$$b_0 = \bar{y} - (b_1\bar{x_1} + b_2\bar{x_2}) \tag{13}$$

$$b_1 = \frac{r_{x_1y} - r_{x_2y}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_1}} \tag{14}$$

$$b_2 = \frac{r_{x_2y} - r_{x_1y}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_2}} \tag{15}$$

ここで r_{x_1y} は独立変数 x_1 と従属変数 y との相関係数を表す。同様に $r_{x_1x_2}$ は 2 つの独立変数間の相関係数を、s_{x_1} は x_1 の標準偏差を表すものとする。

2　グラフによる表現

各変数からそれぞれの平均を引いた値を要素とするベクトル (平均偏差ベクトル) を考えれば、2 つの変数間の相関係数は平均偏差ベクトル間のなす角の余弦 cos であった。今 3 つのベクトル y, x_1, x_2 を考える

2 つの独立変数と説明変数とからそれぞれの平均を引いた平均偏差ベクトルをそれぞれ $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{y}$ とする。$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって一つの空間が定義でき

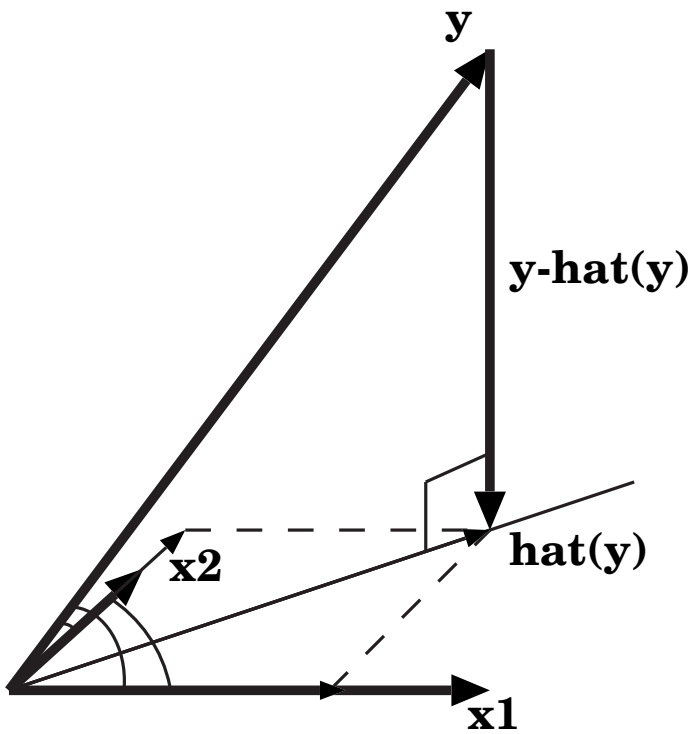

図 2: 2 つのベクトル x_1 と x_2 とによって表現される 2 次元空間へベクトル y を射影したときの模式図

る。$\boldsymbol{y}$ がこの空間内にあれば任意の定数 a と b とによって $\boldsymbol{y} = a\boldsymbol{x}_1 + b\boldsymbol{x}_2$ と

表現できる。3 つのベクトルが同一空間 (この場合同一平面) 上にあるときにはこのことが成り立つが、一般に 3 つのベクトルは同一空間内にあるとは限らない。

そこで、ベクトル $\boldsymbol{y}$ をベクトル $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって定義される空間へ射影し、この射影してできたベクトルを $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって記述することを考える。射影ベクトルを $\hat{\boldsymbol{y}}$ とすれば、$\boldsymbol{y}$ から射影ベクトルを引いた残差ベクトル $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ と残差ベクトルは直交する $\hat{\boldsymbol{y}} \perp \boldsymbol{r}$。

説明変数 $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とから目的変数 $\boldsymbol{y}$ を予測する重回帰分析とは、$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ とによって張られる空間へ $\boldsymbol{y}$ を射影し、できたベクトル $\boldsymbol{y}$ を $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ に分解すること、およびそのときの $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ にかかる係数 b_1, b_2 を求めることである。

3　ベクトルと行列による表現

従属変数 y_i $(i = 1, 2, \cdots, n)$ データをベクトル $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ とし、1 を n 個並べたベクトル $\mathbf{1}'$ と 2 つの独立変数ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2$ を合わせて n 行 3 列の行列を $\boldsymbol{X} = (\mathbf{1}', \mathbf{1}, \boldsymbol{x}_1', \boldsymbol{x}_2')$ とする。このようにすると (11) は次のように書くことができる。ただし $\mathbf{1} = \underbrace{(1, 1, \cdots, 1)}_{n \text{ 個}}{}'$ である。

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} \tag{16}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1, & x_{11}, & x_{12} \\ 1, & x_{21}, & x_{22} \\ \vdots & \vdots & \vdots \\ 1, & x_{n1}, & x_{n2} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \tag{17}$$

代数の操作からの類推を使ってこれを形式的に解くと、

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} \tag{18}$$

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} \tag{19}$$

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} \tag{20}$$

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{b} \tag{21}$$

求めるべき回帰係数と定数でできたベクトルは、従属変数ベクトルと定数からできた行列と目的変数ベクトルとを使って代数の演算のように解くことができる。(21) の左辺が 3 行 1 列のベクトルとなることを確認せよ。