**2011 年度秋学期　統計学　第１５回**

# データの関係を知る (3) ー 重回帰

前回は,「緯度 $(x)$ が上がると気温 $(y)$ が下がる」というように，１つの変量 $y$ をもう１つの変量 $x$ で説明するという考えにもとづき，線形単回帰を説明しました。今回は，１つの変量 $y$ をそれ以外の複数の変量 $x_1, x_2, \ldots,$ で説明するという考えにもとづき，重回帰・重相関・偏相関という考え方を説明します。

今回の説明には「行列」が出てきます。しかし，細かい数学よりも，考え方の大枠を理解してください。一応,「画像情報処理」の講義で使った補助プリント「『行列』に慣れていない人のために」を一緒にアップロードしておきます。

---

## 重回帰

前回の例では,「緯度が上がると気温が下がる」というように,「気温」という１つの変量の変化を「緯度」というもう１つの変量の変化で表す（変量で説明する）という考え方を示しました。しかし，気温の違いは緯度だけで説明できるものではなく，他にも例えば「標高」にも関係があります。そこで,「緯度が上がり，標高が上がると気温が下がる」というように，２つ以上の変量で１つの変量を説明する方法を考えれば，データの構造をより精密に表すことができます。この方法を，前回説明した単回帰に対して**重回帰分析**といいます。

単回帰では，被説明変量 $y$ を１つの説明変量 $x$ の１次関数 $a+bx$ で表現しました。これに対して，重回帰では被説明変量 $y$ を $p$ 個の説明変量 $x_1, x_2, \ldots, x_p$ の１次関数 $y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p$ で表現します。単回帰の場合が，２次元の散布図で各データをもっともうまく代表する直線を求めるのに相当するのに対して，重回帰では，$n$ 次元の散布図で各データをもっともうまく代表する $(n-1)$ 次元の超平面（回帰超平面）を求めます。

超平面の決め方は単回帰の場合と同じで，$i$ 番のデータ $(x_{1i}, x_{2i}, \ldots, x_{pi}; y_i)$ について，実際の被説明変量 $y_i$ と超平面によって決められる被説明変量の予測値 $a_0 + a_1x_{1i} + a_2x_{2i} + \cdots + a_px_{pi}$ との差の２乗を求め，その全データについての合計が最小になるように，超平面を表すパラメータ $a_0, a_1, a_2, \ldots, a_p$ を決めます。ここでは，説明変量が２つの場合について，前回と同様に考えてみましょう。以下，$x_{1i}, x_{2i}$ はそれぞれ「$i$ 番めのデータの，説明変量１（説明変量 2）の値」を表し，データは $n$ 個あるとします。

差の２乗の合計 $L$ は,

$$\begin{aligned} L &= \sum_{i=1}^{n} \left[ y_i - (a_0 + a_1x_{1i} + a_2x_{2i}) \right]^2 \\ &= \sum_{i=1}^{n} \left[ y_i^2 - 2\left(a_0 + a_1x_{1i} + a_2x_{2i}\right) y_i + \left(a_0 + a_1x_{1i} + a_2x_{2i}\right)^2 \right] \end{aligned} \tag{1}$$

と表されます。単回帰の場合と同様に考えて，$L$ を $a_0, a_1, a_2$ で各々偏微分し，それらを 0 とおくことで，$a_0, a_1, a_2$ の値を決めます。まず，$a_0$ で偏微分して 0 とおくと，

$$\frac{\partial L}{\partial a_0} = \sum \left[ -2y_i + 2\left(a_0 + a_1x_{1i} + a_2x_{2i}\right) \right] = 0 \tag{2}$$

すなわち

$$\sum \left(a_0 + a_1x_{1i} + a_2x_{2i}\right) = \sum y_i \tag{3}$$

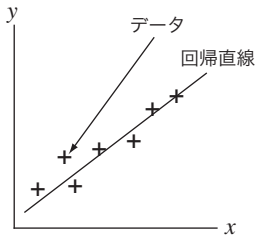


図 1: ２次元散布図と回帰直線

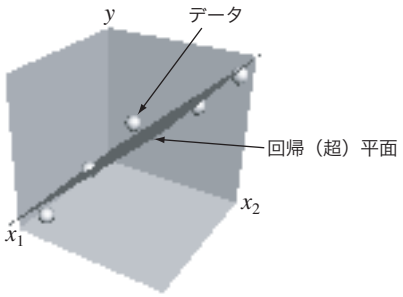


図 2: ３次元散布図と２次元回帰（超）平面

が得られます。ここで,

$$\bar{x_1} = \frac{1}{n}\sum x_{1i}, \bar{x_2} = \frac{1}{n}\sum x_{2i}, \bar{y} = \frac{1}{n}\sum y_i \tag{4}$$

とおきます。これを用いると,

$$\begin{aligned} na_0 + a_1 n\bar{x_1} + a_2 n\bar{x_2} &= n\bar{y} \\ \text{すなわち} \quad a_0 + a_1\bar{x_1} + a_2\bar{x_2} &= \bar{y} \end{aligned} \tag{5}$$

となるので，$a_0$ は

$$a_0 = \bar{y} - (a_1\bar{x_1} + a_2\bar{x_2}) \tag{6}$$

と表されます。

一方，この (6) 式で導かれた $a_0$ を再び (3) 式に代入します。すると,

$$\begin{aligned} \sum\left[\{\bar{y} - (a_1\bar{x_1} + a_2\bar{x_2})\} + a_1 x_{1i} + a_2 x_{2i}\right] &= \sum yi \\ \text{すなわち} \quad a_1\sum(\bar{x_1} - x_{1i}) + a_2\sum(\bar{x_2} - x_{2i}) &= \sum(y_i - \bar{y}) \end{aligned} \tag{7}$$

という関係が得られます。

さて，ここで差の２乗（(1) 式）を今度は $a_1$ で偏微分して 0 とおくと,

$$\frac{\partial L}{\partial a_1} = \sum\left[-2y_i x_{1i} + 2x_{1i}\left(a_0 + a_1 x_{1i} + a_2 x_{2i}\right)\right] = 0 \tag{8}$$

すなわち

$$\sum x_{1i}\left(a_0 + a_1 x_{1i} + a_2 x_{2i}\right) = \sum x_{1i} y_i \tag{9}$$

が得られます。(9) 式に (6) 式を代入すると,

$$\begin{aligned} \sum x_{1i}\left[\{\bar{y} - (a_1\bar{x_1} + a_2\bar{x_2})\} + a_1 x_{1i} + a_2 x_{2i}\right] &= \sum x_{1i} y_i \\ \text{すなわち} \quad a_1\sum x_{1i}(x_{1i} - \bar{x_1}) + a_2\sum x_{2i}(x_{2i} - \bar{x_2}) &= \sum(y_i - \bar{y})\, x_{1i} \end{aligned} \tag{10}$$

となり，さらに，(10) 式から (7) 式 $\times\bar{x_1}$ をひくと,

$$a_1\sum(x_{1i} - \bar{x_1})(x_{1i} - \bar{x_1}) + a_2\sum(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2}) = \sum(y_i - \bar{y})(x_{1i} - \bar{x_1}) \tag{11}$$

が得られます。ここで,

$$\left\{\begin{array}{rcl} \dfrac{1}{n}\sum\left(x_{1i}-\bar{x_1}\right)\left(x_{1i}-\bar{x_1}\right) & = & s_{11} \\ \dfrac{1}{n}\sum\left(x_{1i}-\bar{x_1}\right)\left(x_{2i}-\bar{x_2}\right) & = & s_{12} \\ \dfrac{1}{n}\sum\left(y_i-\bar{y}\right)\left(x_{1i}-\bar{x_1}\right) & = & s_{y1} \end{array}\right. \tag{12}$$

とおくと，(11) 式は

$$a_1 n s_{11} + a_2 n s_{12} = n s_{y1} \text{ すなわち } a_1 s_{11} + a_2 s_{12} = s_{y1} \tag{13}$$

と表されます。

また，(1) 式を $a_2$ で偏微分して 0 とおき，(9) 式から (13) 式までと同様の操作を行なうと，

$$\left\{\begin{array}{rcl} \dfrac{1}{n}\sum\left(x_{2i}-\bar{x_2}\right)\left(x_{2i}-\bar{x_2}\right) & = & s_{22} \\ \dfrac{1}{n}\sum\left(x_{2i}-\bar{x_2}\right)\left(x_{1i}-\bar{x_1}\right) & = & s_{21} \\ \dfrac{1}{n}\sum\left(y_i-\bar{y}\right)\left(x_{2i}-\bar{x_2}\right) & = & s_{y2} \end{array}\right. \tag{14}$$

とおくことで,

$$a_1 s_{21} + a_2 s_{22} = s_{y2} \tag{15}$$

という関係が得られます。

(13) 式と (15) 式をまとめると,

$$\left(\begin{array}{cc} s_{11} & s_{12} \\ s_{21} & s_{22} \end{array}\right)\left(\begin{array}{c} a_1 \\ a_2 \end{array}\right) = \left(\begin{array}{c} s_{y1} \\ s_{y2} \end{array}\right) \tag{16}$$

のように行列で表現することができます。$s_{11}, s_{22}$ はそれぞれ変量 $x_1, x_2$ の分散，$s_{12} = s_{21}$ はそれぞれ変量 $x_1$ と $x_2$ の共分散です。また，(16) 式の行列は**分散共分散行列**とよばれています。(16) 式は 2 元連立 1 次方程式で，これを解くことで $a_1, a_2$ を求めることができ，さらにそれを (6) 式に代入すると $a_0$ が求められます。

ここまでは変量が 2 つの場合でしたが，変量が $p$ 個の場合は，$j, k$ を変量の番号 ($j = 1, 2, \ldots, p$ , $k = 1, 2, \ldots, p$) として,

$$\left\{\begin{array}{rcl} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(x_{ji}-\bar{x_j}\right)\left(x_{ki}-\bar{x_k}\right) & = & s_{jk} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(x_{ki}-\bar{x_k}\right) & = & s_{yk} \end{array}\right. \tag{17}$$

とおくと,

$$\left(\begin{array}{ccccc} s_{11} & \cdots & s_{1j} & \cdots & s_{1p} \\ \vdots & \ddots & & & \vdots \\ s_{j1} & & s_{jj} & & s_{jp} \\ \vdots & & & \ddots & \vdots \\ s_{p1} & & s_{pj} & & s_{pp} \end{array}\right)\left(\begin{array}{c} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_p \end{array}\right) = \left(\begin{array}{c} s_{y1} \\ \vdots \\ s_{yj} \\ \vdots \\ s_{yp} \end{array}\right) \tag{18}$$

という分散共分散行列の式が得られ，この $p$ 元 1 次連立方程式を解くことで回帰方程式が求められます。

## 多重共線関係

表 1 のデータでは，どのデータについても $x_2$ が $x_1$ の 2 倍の関係になっています。このような場合，散布図は図 3 のようになり，回帰（超）平面は $y$ 軸に平行になります。しかし，このような平面は 1 次式 $y = a_0 + a_1x_1 + a_2x_2$ の形では表現することができません。したがって，これまでに述べた方法でこの平面を求めることはできません。実際，$x_2 = \alpha x_1$ とおくと

$$\left\{ \begin{array}{rcl} s_{11} & = & \dfrac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2 \\ s_{22} & = & \dfrac{1}{n}\sum_{i=1}^{n}(x_{2i} - \bar{x_2})^2 = \dfrac{1}{n}\sum_{i=1}^{n}(\alpha x_{1i} - \alpha\bar{x_1})^2 = \alpha^2 s_{11} \\ s_{12} & = & s_{21} = \dfrac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2}) = \dfrac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x_1})(\alpha x_{1i} - \alpha\bar{x_1}) = \alpha s_{11} \end{array} \right. \tag{19}$$

ですから，分散共分散行列は

$$\begin{pmatrix} s_{11} & \alpha s_{11} \\ \alpha s_{11} & \alpha^2 s_{11} \end{pmatrix} \tag{20}$$

となり，逆行列が存在しませんから，(18) 式の連立 1 次方程式は解けません。また，この平面は直線 $x_2 = 2x_1$ の上にしか存在しませんから，この平面を求めたところで，$x_1$ と $x_2$ から $y$ を予測するという回帰分析の目的は果たせません。

このような状態は，点 $(x_1, x_2)$ が直線上に並んでいる，すなわち $x_1$ と $x_2$ の相関係数が 1 になっているとき生じます。このように，説明変量のうちの 2 つの相関係数が 1 に非常に近い値になっていることを，この 2 つの説明変量が**多重共線関係**にあるといいます。重回帰分析を行う際には，多重共線関係が生じないように説明変量を選ぶ必要があります。

| $y$ | $x_1$ | $x_2$ |
|---|---|---|
| 9 | 6 | 12 |
| 14 | 2 | 4 |
| 12 | 1 | 2 |
| 10 | 9 | 18 |
| 13 | 3 | 6 |

表 1: 多重共線関係

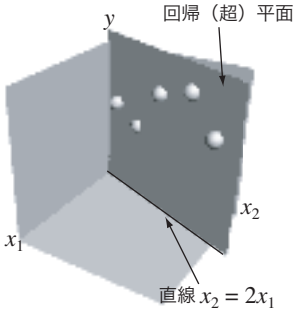


図 3: 多重共線関係

---

## 重相関係数

ここまでの方法で，回帰超平面 $y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p$ を求めたとします。この 1 次式に各データ $x_{1i}, x_{2i}, \ldots, x_{pi}$（くりかえしますが，$i$ はデータの番号，$i = 1, 2, \ldots, n$）を代入すると，$i$ 番のデータに対して回帰による予測値 $Y_i$ が得られます。一方，$i$ 番のデータに対しては，すでに $y$ の実測値 $y_i$ が得られています。そこで，予測値 $Y$ と実測値 $y$ との相関係数を考えます。この値は予測値と実測値の食い違い具合であり，また回帰超平面のまわりへの実測値の寄り集まり具合を表しています。この相関係数を，$y$ と $x_1, x_2, \ldots, x_p$ の**重相関係数**といい，$r_{y\cdot 12\ldots p}$ という記号で表します。

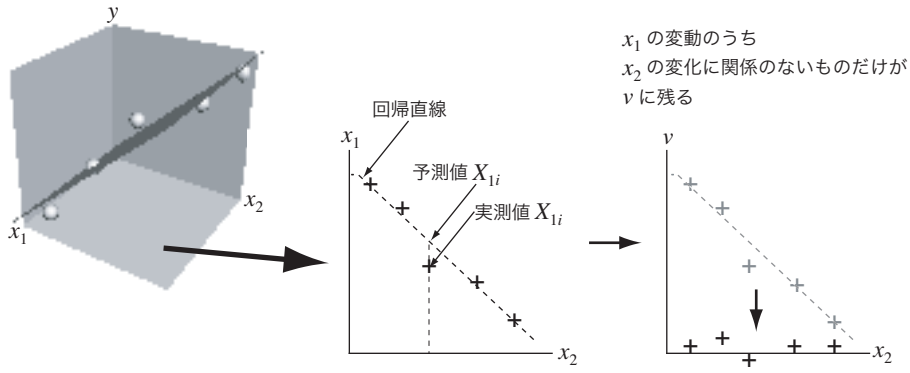


図 4: 偏相関係数

**偏相関係数**

いま，被説明変量 $y$ を $p$ 個の変量 $x_1, x_2, \ldots, x_p$ から予測するかわりに，1 番の説明変量 $x_1$ を除いた $x_2, \ldots, x_p$ から予測するとしてみます。ここまでで述べた方法で，$x_2, \ldots, x_p$ による回帰方程式を求めたとき，$i$ 番のデータの被説明変量 $y$ の予測値を $Y_i$，実測値を $y_i$ とします。さらに，説明変量 $x_1$ も同様に $x_2, \ldots, x_p$ から予測し，$i$ 番のデータの説明変量 $x_1$ の予測値を $X_{1i}$，実測値を $x_{1i}$ とします。

ここで，$y$, $x_1$ それぞれの残差を $u_i = y_i - Y_i, v_i = x_{1i} - X_{1i}$ とするとき，変量 $u$ と $v$ の相関係数を，「$x_2, \ldots, x_p$ の影響を除いた $y$ と $x_1$ の**偏相関係数**」といい，$r_{y1\cdot 2\ldots p}$ という記号で表します。

これではわかりにくいので，3 次元散布図で考えてみましょう。$y$ と $x_1$ を，それぞれ残りの $x_2$ から予測します。これは，図 4 のように，$x_2 - y$ 平面や $x_2 - x_1$ 平面で回帰超平面（この場合は回帰直線）を求めることに相当します。

この図のように，$x_1$ の残差 $v$ は，$x_2$ に依存して変化する変動を $x_1$ から除いたものになっていることがわかります。同様に，$y$ の残差 $u$ は，$x_2$ に依存して変化する変動を $y$ から除いたものになっていることがわかります。したがって，$u$ と $v$ の相関係数は，$y$ と $x_1$ から，それぞれ $x_2$ に依存して変化する変動を除いた上で，計算した相関係数になっていることになります。

偏相関係数は，講義第１３回で触れた「見かけ上の相関」と関連してよく取り扱われます。見かけ上の相関とは，変量 $y$ と $x_1$ の間には本来相関関係はないはずなのに，計算上の相関係数が 1 に近い値になることです。これは，もう 1 つの隠れた変量 $x_2$ があって，$x_2$ と $x_1$ および $x_2$ と $y$ のそれぞれに相関関係があるために，見かけ上 $y$ と $x_1$ の間にも相関関係があるように見えるためです。

第 4 回の講義プリントの例のように，小学校の全学年の児童に対して $y$：身長，$x_1$：成績というデータをとると，「身長と成績との間には相関がある」という結果が出ます。しかし，$x_2$：年齢という変量を考え，この影響を除いた偏相関係数を計算すると，身長と成績の相関は本当はないことがわかります。

「では，『成績の影響を除いた，年齢と身長の相関』もほとんどないことにならないのか？」と思った人もいるのではないでしょうか？ これは，数式の上では正しい結論です。しかし，実際には意味のない結論です。なぜならば，「みかけ上の相関」は，「身長と成績に相関があるように見えるが，実は『年齢』という隠れた量があって，年齢が成績，身長それぞれの大小に影響している」すなわち「年齢が，成績，身長それぞれを説明している」という仮定から導かれるものだからです。しかし，その仮定が正しいかどうかは，相関係数や偏相関係数を見ても不明で，別の観点からの考察が必要です。