データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/06

静岡大学工学部
安藤和敏

2006.10.23

前回までの復習

重回帰分析のデータ
(説明変数が2個の場合)

個体番号	変数 x	変数 u	変数 y
1	x_1	u_1	y_1
2	x_2	u_2	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	u_i	y_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	u_n	y_n

残差平方和 Q

$$Q = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} \left\{ y_i - (a + bx_i + cu_i) \right\}^2$$

Qを a,b ,cを変数にもつ3変数関数として見て，$Q(a,b,c)$を最小にする a,b,cが，データに「最もよくあてはまる」平面を与えると考える．

このようにしてa,**b,c**を求める方法を**最小2乗法**と呼ぶ．

どのようにして$Q(a,b,c)$を最小にする a,b,cをもとめるのかを見ていく．

$Q(a,b,c)$を最小にする a,b,c

$$\begin{bmatrix} s_x^2 & s_{xu} \\ s_{xu} & s_u^2 \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} s_{xy} \\ s_{uy} \end{bmatrix},$$

$$a = \overline{y} - b\overline{x} - c\overline{u}$$

説明変数が3個以上の場合

重回帰分析のデータ
(説明変数が3個の場合)

号番体個	変数 x	変数 u	変数 v	変数 y
1	x_1	u_1	v_1	y_1
2	x_2	u_2	v_2	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	u_i	v_i	y_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	u_n	v_n	y_n

残差平方和 Q

$$Q = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} \left\{ y_i - (a + bx_i + cu_i + dv_i) \right\}^2$$

*Q*を *a*,*b* ,c,dを変数にもつ3変数関数として見て，*Q*(*a*,*b*,*c*,*d*)を最小にする *a*,*b*,c,dが，データに「最もよくあてはまる」平面を与えると考える．

このようにして*a*,***b***,***c***,***d***を求める方法を**最小２乗法**と呼ぶ．

$Q(a,b,c,d)$を最小にする a,b,c,d

$$\begin{bmatrix} s_x^2 & s_{xu} & s_{xv} \\ s_{xu} & s_u^2 & s_{uv} \\ s_{xv} & s_{uv} & s_v^2 \end{bmatrix}\begin{bmatrix} b \\ c \\ d \end{bmatrix} = \begin{bmatrix} s_{xy} \\ s_{uy} \\ s_{vy} \end{bmatrix},$$

$$a = \bar{y} - b\bar{x} - c\bar{u} - d\bar{v}$$

前々回やり残したこと

$$s_y^2 = s_\varepsilon^2 + bs_x^2 + cs_u^2 + 2bcs_{xu} \qquad \cdots\cdots(3)$$

残差平方和の別表現（1）

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left\{ y_i - (a + bx_i + cu_i) \right\}^2$$

$$= \sum_{i=1}^{n} \left\{ y_i - \overline{y} + \overline{y} - (a + bx_i + cu_i) \right\}^2$$

$$= \sum_{i=1}^{n} \left\{ y_i - \overline{y} + a + b\overline{x} + c\overline{u} - (a + bx_i + cu_i) \right\}^2$$

$$= \sum_{i=1}^{n} \left\{ (y_i - \overline{y}) - b(x_i - \overline{x}) - c(u_i - \overline{u}) \right\}^2$$

残差平方和の別表現（2）

（つづき）

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 + c^2\sum_{i=1}^{n}(u_i - \bar{u})^2$$

$$-2b\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - 2c\sum_{i=1}^{n}(y_i - \bar{y})(u_i - \bar{u})$$

$$+2bc\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})$$

$$= ns_y^2 + b^2 ns_x^2 + c^2 ns_u^2 - 2nbs_{xy} - 2ncs_{uy} + 2nbcs_{xu}$$

残差平方和の別表現（3）

（つづき）

$$= ns_y^2 + b^2 ns_x^2 + c^2 ns_u^2 - 2nbs_{xy} - 2ncs_{uy} + 2nbcs_{xu}$$

$$= ns_y^2 + n\{b(bs_x^2 + cs_{xu}) + c(cs_u^2 + bs_{xu})\}$$

$$\quad - 2nbs_{xy} - 2ncs_{uy}$$

$$= ns_y^2 + n\{bs_{xy} + cs_{uy}\} - 2nbs_{xy} - 2ncs_{uy}$$

$$= ns_y^2 - n\{bs_{xy} + cs_{uy}\}$$

$$= ns_y^2 - n\{b(bs_x^2 + cs_{xu}) + c(cs_u^2 + bs_{xu})\}$$

$$= ns_y^2 - nb^2 s_x^2 - ncs_u^2 - 2nbcs_{xu}$$

残差平方和の別表現（4）

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \left\{ y_i - (a + bx_i + cu_i) \right\} = 0$$

だから，

$$\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i = 0.$$

したがって，

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(\varepsilon_i - \bar{\varepsilon} \right)^2 = n s_{\varepsilon}^2 .$$

$$s_y^2 = s_{\varepsilon}^2 + b s_x^2 + c s_u^2 + 2bc s_{xu} \qquad \cdots\cdots (3)$$

分散に関する公式

$$\boxed{s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}$$

(証明)

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\overline{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\overline{x}^2 + \overline{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2
\end{aligned}$$

共分散に関する公式

$$\boxed{s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\overline{y}}$$

(証明)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - y_i\overline{x} + \overline{x}\overline{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\frac{1}{n}\sum_{i=1}^{n} y_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}\overline{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\overline{y} - \overline{x}\overline{y} + \overline{x}\overline{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\overline{y}$$