データ解析

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

静岡大学工学部
安藤和敏

2005.11.02

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
⋮	⋮	⋮	⋮
i	x_i	u_i	y_i
⋮	⋮	⋮	⋮
n	x_n	u_n	y_n

説明変数が2個の場合の重回帰分析

与えられたデータに「最もよくあてはまる」平面

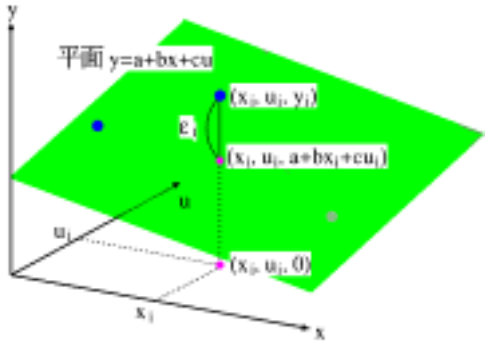
回帰方程式 $y = a + bx + cu \quad \cdots\cdots (1)$

を求めること.

目的変数　切片　偏回帰係数　説明変数

「最もよくあてはまる平面」ってどういうこと？

残差　$\varepsilon_i = y_i - (a + bx_i + cu_i)$



残差平方和 Q

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

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

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.

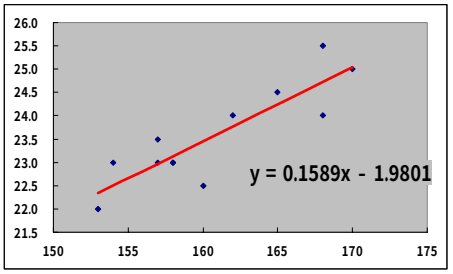
$Q(a,b,c)$を最小にする a,b,c

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

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

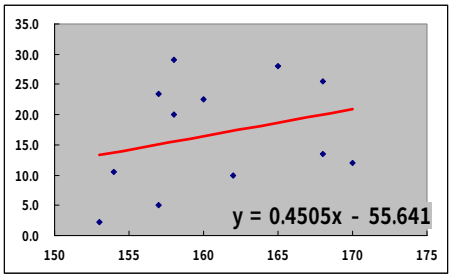
2-3 回帰分析の精度を示す決定係数

精度が良い回帰方程式



回帰方程式は，データをよく表現している．

精度が悪い回帰方程式



回帰方程式は，データを表現しているとはいえない．

決定係数

決定係数は，回帰方程式が与えられた多変数データをどれだけよく表現しているかを示す尺度である．

説明変数が2個の場合の重回帰分析

$$y = a + bx + cu$$

を回帰方程式とする．このとき，

$$\hat{y}_i = a + bx_i + cu_i \quad (i = 1, \cdots, n)$$

で定義される変数 $\hat{y}$ を予測値と呼ぶ．

残差 ε_i は以下のように書ける．

$$\varepsilon_i = y_i - (a + bx_i + cu_i) = y_i - \hat{y}_i.$$

$\hat{y}$ の平均

$$\begin{aligned}
\bar{\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i + cu_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}a + b\frac{1}{n}\sum_{i=1}^{n}x_i + c\frac{1}{n}\sum_{i=1}^{n}u_i \\
&= a + b\bar{x} + c\bar{u} \\
&= \bar{y}.
\end{aligned}$$

分散の関係

$$s_y^2 = s_{\hat{y}}^2 + s_\varepsilon^2 \cdots\cdots (3)$$

実測値の分散 ＝ 予測値の分散 ＋ 残差の分散

平方和の分解 (1)

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \{y_i - (a + bx_i + cu_i) \\ &\quad + (a + bx_i + cu_i) - \bar{y}\}^2 \\ &= \sum_{i=1}^n \{\varepsilon_i + (a + bx_i + cu_i) - \bar{y}\}^2 \\ &= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n \{(a + bx_i + cu_i) - \bar{y}\}^2 \\ &\quad + \sum_{i=1}^n 2\varepsilon_i \{(a + bx_i + cu_i) - \bar{y}\}\end{aligned}$$

$$\sum_{i=1}^n \varepsilon_i \{(a + bx_i + cu_i) - \bar{y}\} = 0$$

$$\begin{aligned}&\sum_{i=1}^n \varepsilon_i \{(a + bx_i + cu_i) - \bar{y}\} \\ &= \sum_{i=1}^n \varepsilon_i \{(a - \bar{y}) + bx_i + cu_i\} \\ &= \sum_{i=1}^n \varepsilon_i (a - \bar{y}) + \sum_{i=1}^n b\varepsilon_i x_i + \sum_{i=1}^n c\varepsilon_i u_i \\ &= (a - \bar{y}) \sum_{i=1}^n \varepsilon_i + b \sum_{i=1}^n \varepsilon_i x_i + c \sum_{i=1}^n \varepsilon_i u_i \\ &= 0.\end{aligned}$$

平方和の分解 (2)

$$\begin{aligned}&\sum_{i=1}^n \varepsilon_i^2 \\ &= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n \{(a + bx_i + cu_i) - \bar{y}\}^2 \\ &= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2\end{aligned}$$

$$\therefore s_y^2 = s_\varepsilon^2 + s_{\hat{y}}^2.$$

決定係数

$$s_y^2 = s_\varepsilon^2 + s_{\hat{y}}^2$$

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{s_y^2 - s_\varepsilon^2}{s_y^2} = 1 - \frac{s_\varepsilon^2}{s_y^2}$$

R^2は**決定係数**と呼ばれる.

$0 \leqq R^2 \leqq 1$が成り立ち, 1に近いほど回帰方程式の精度が良いと考えられる.

補正決定係数

実は説明変数の数を増やしていけば, R^2は1に近くすることができる. 説明変数の数による影響を排除するために, 決定係数のかわりに以下で定義されるR^{*2}を考えることもある.

$$R^{*2} = 1 - \frac{s_\varepsilon^2 / (n - p - 1)}{s_y^2 / (n - 1)}$$

ここで, pは説明変数の数. R^{*2}は**補正決定係数**と呼ばれる.

重相関係数

$\hat{y}$ と y の相関係数

$$r_{y\hat{y}} = \frac{s_{y\hat{y}}}{s_y s_{\hat{y}}}$$

は**重相関係数**と呼ばれる.

重相関係数^2=決定係数 (1)

$$ns_{y\hat{y}} = \sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})$$
$$= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})$$
$$= \sum_{i=1}^{n}(\varepsilon_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})$$
$$= \sum_{i=1}^{n}\varepsilon_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}\varepsilon_i(a + bx_i + cu_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = ns_y^2$$

重相関係数^2=決定係数 (2)

$$s_{y\hat{y}} = \frac{s_{y\hat{y}}}{s_y s_{\hat{y}}} = \frac{s_{\hat{y}}^2}{s_y s_{\hat{y}}} = \frac{s_{\hat{y}}}{s_y}$$

$$\therefore s_{y\hat{y}}^2 = \frac{s_{\hat{y}}^2}{s_y^2} = R^2$$

重相関係数の性質

予測値 $\hat{y}$ は, 回帰方程式の切片aと偏回帰係数 b, c によって

$$\hat{y}_i = a + bx_i + cu_i$$

で定義される.

任意の α, β, γ に対して

$$\tilde{y}_i = \alpha + \beta x_i + \gamma u_i$$

で定義される変数 $\tilde{y}$ を考えると,

$$r_{y\tilde{y}} \le r_{y\hat{y}}.$$

本日のまとめ

- 次の関係式の導出を理解した.

$$s_y^2 = s_\varepsilon^2 + s_{\hat{y}}^2.$$

- 決定係数と補正決定係数の意味を理解した.
- 決定係数と重相関係数の関係を理解した.
- Excelを用いた重回帰分析で, 決定係数, 補正決定係数などを計算する方法を理解した.