

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

2006.10.19

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 $y = a + bx + cu \dots\dots (1)$

を求めること。

目的変数

切片

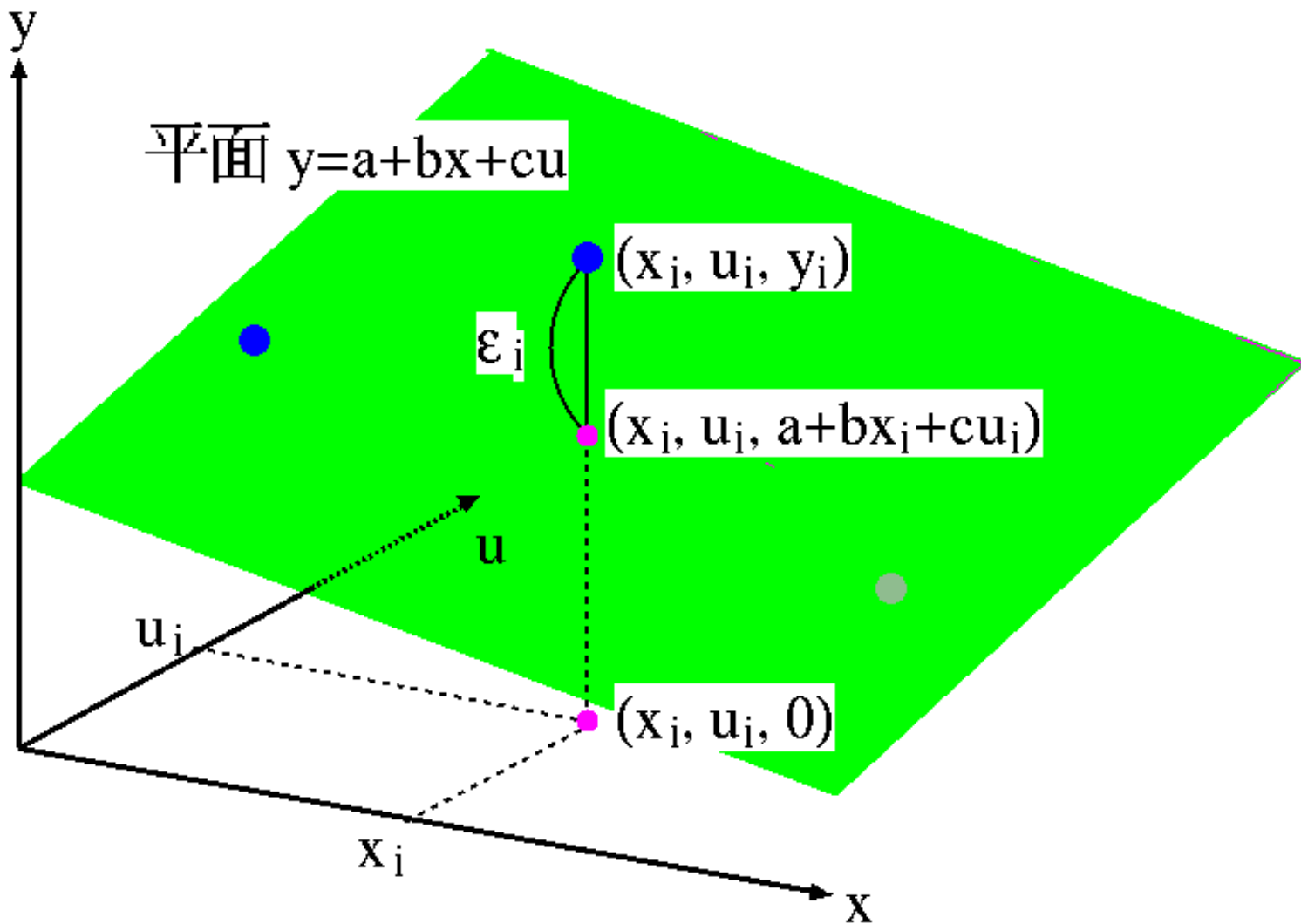
偏回帰係数

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.

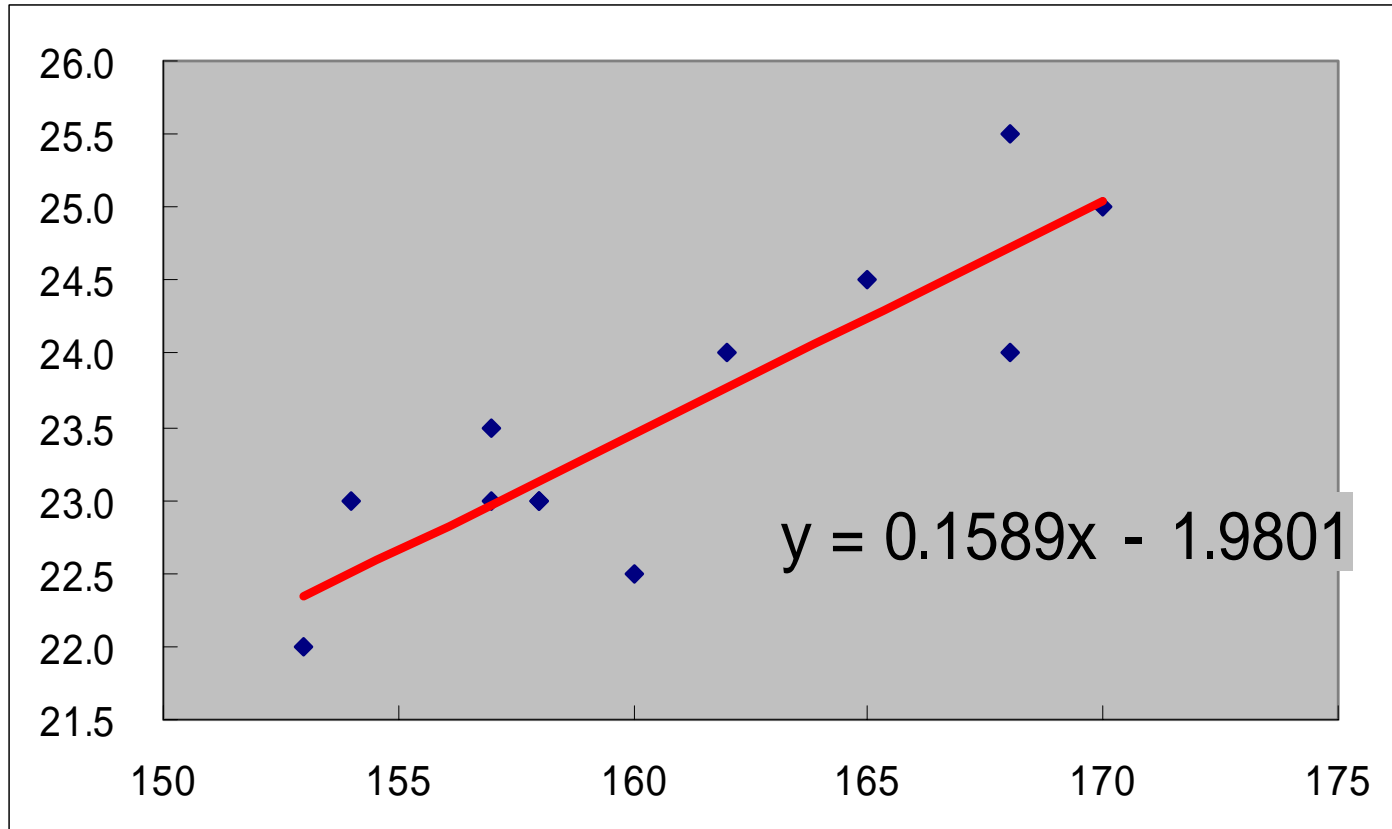
$Q(a,b,c)$ を最小にする a,b,c

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

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

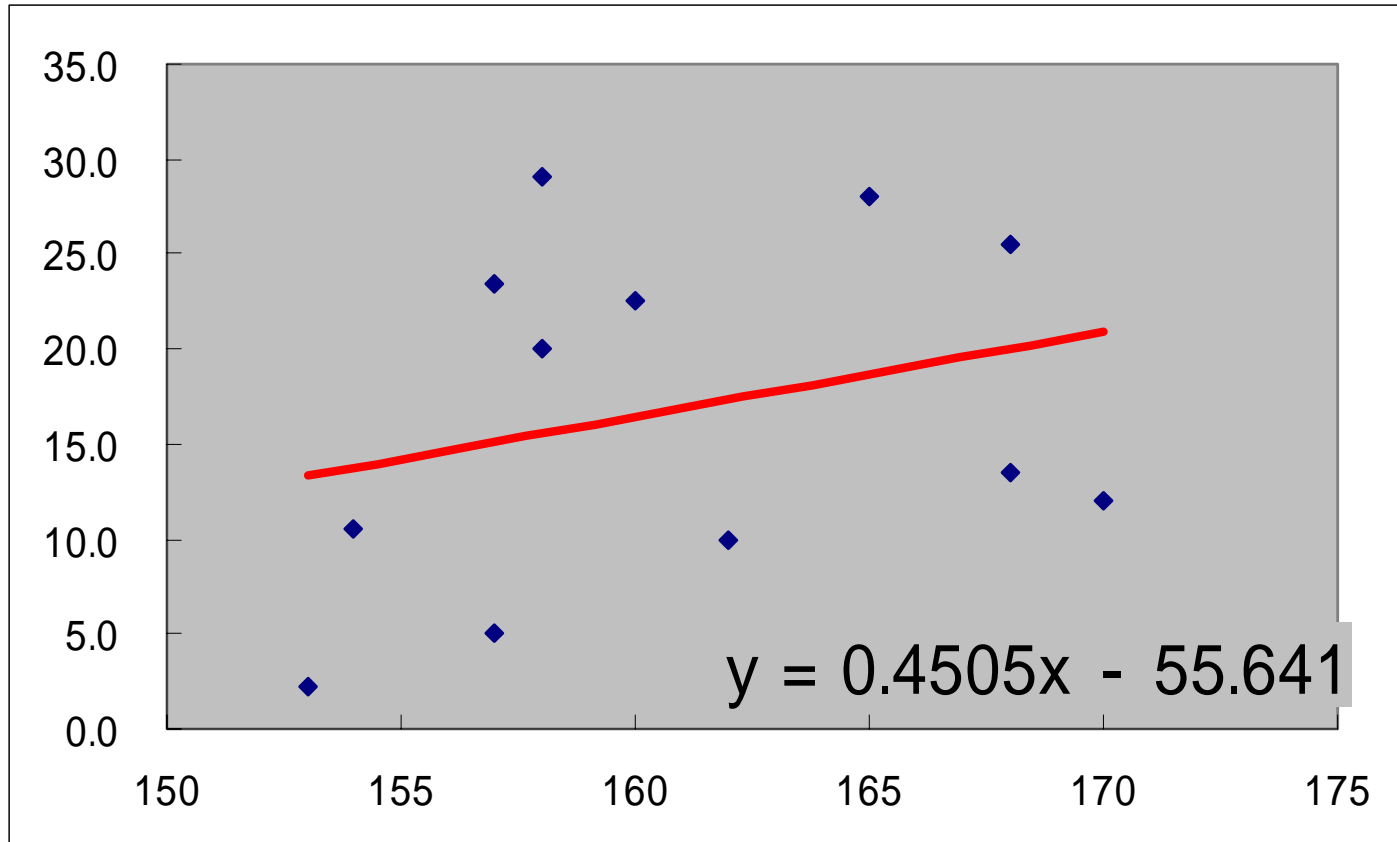
2-3 回帰分析の精度を示す決定係数

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、

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

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

$$\overline{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$$

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

$$= \frac{1}{n} \sum_{i=1}^n y_i$$

$$= \bar{y}.$$

分散の関係

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

実測値の分散 = 予測値の分散 + 残差の分散

平方和の分解 (1)

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

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

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

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

$$= \sum_{i=1}^n \varepsilon_i (a - \bar{y}) + \sum_{i=1}^n b \varepsilon_i x_i + \sum_{i=1}^n c \varepsilon_i u_i$$

$$= (a - \bar{y}) \sum_{i=1}^n \varepsilon_i + b \sum_{i=1}^n \varepsilon_i x_i + c \sum_{i=1}^n \varepsilon_i u_i$$

$$= 0.$$

平方和の分解(2)

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

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

$$= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

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

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 R^2 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)

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

重相関係数²=決定係数(2)

$$r_{y\hat{y}}^2 = \frac{s_{y\hat{y}}^2}{s_y^2 s_{\hat{y}}^2} = \frac{s_{\hat{y}}^4}{s_y^2 s_{\hat{y}}^2} = \frac{s_{\hat{y}}^2}{s_y^2} = R^2$$

$$\therefore r_{y\hat{y}}^2 = R^2$$

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,

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

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。