データ解析 (第13回)

安藤 和敏

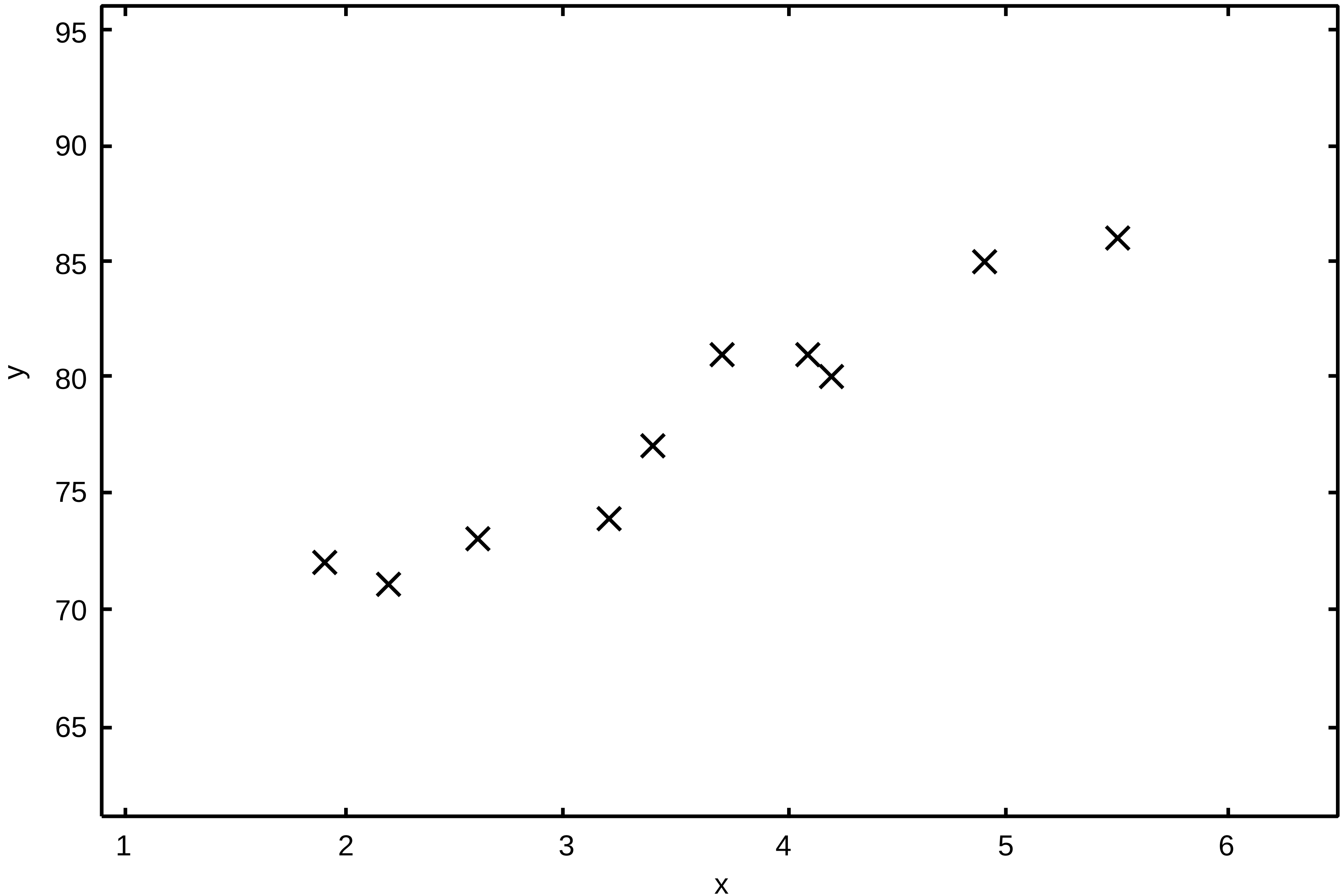
静岡大学システム工学科

第 5 章 重回帰分析

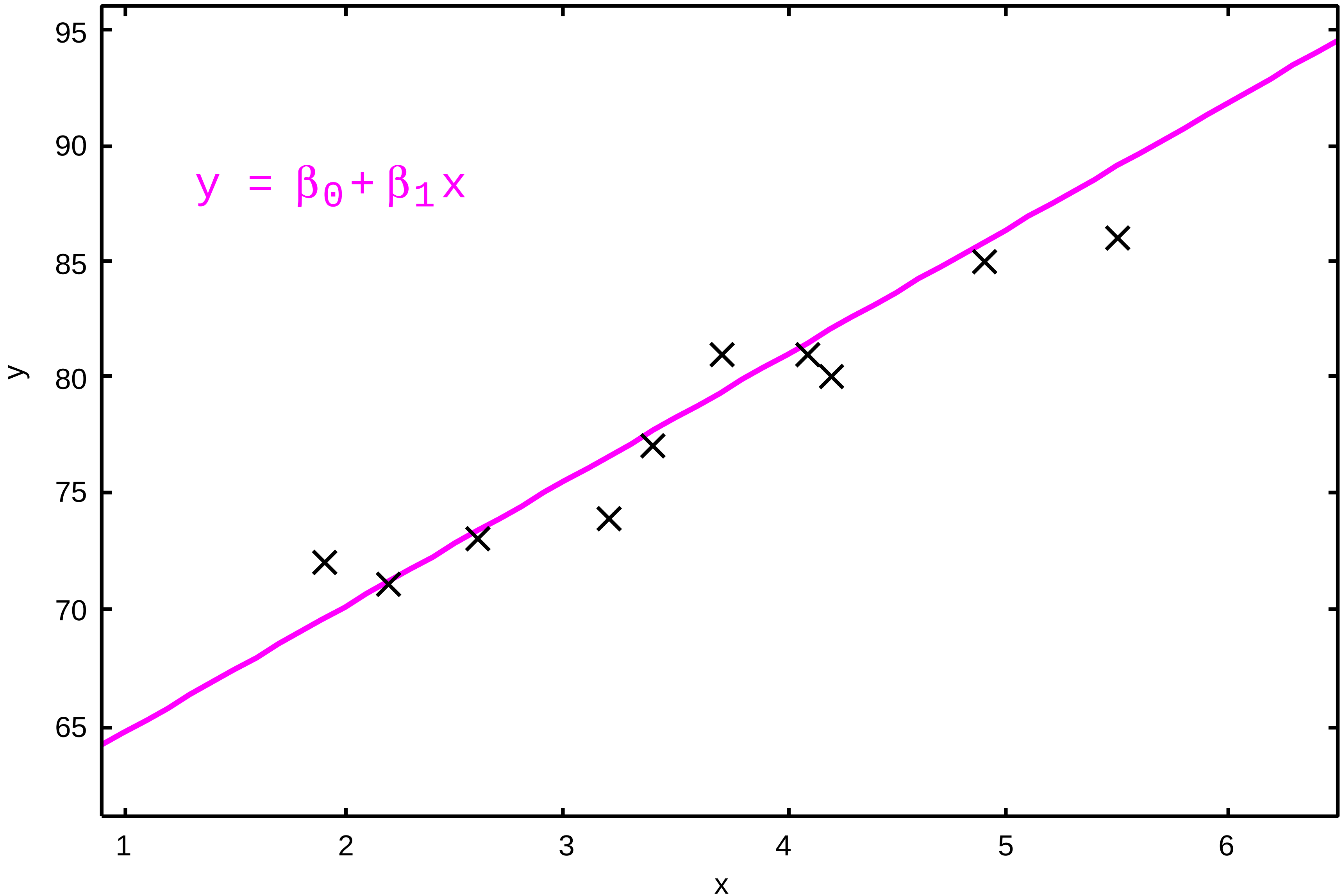
データ (x_i, y_i)

No.	説明変数 x	目的変数 y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

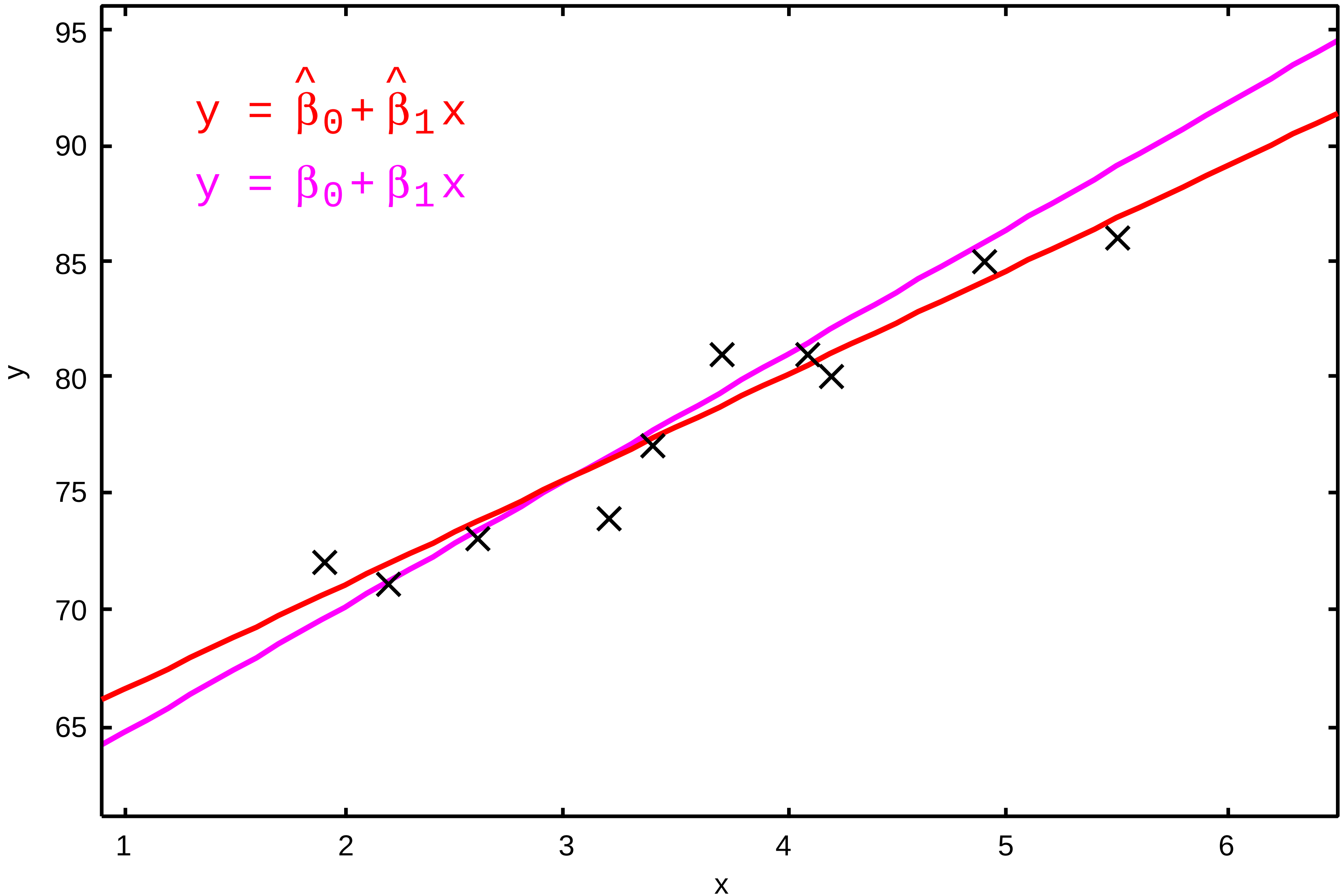
データの散布図



単回帰モデル $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



直線の推定



$y = \hat{\beta}_0 + \hat{\beta}_1 x$
$y = \beta_0 + \beta_1 x$
95
90
85
80
75
70
65
y
1
2
3
4
5
6
x

5.2 説明変数が 2 個の場合の解析方法

データ (x_{i1}, x_{i2}, y_i)

No.	説明変数		目的変数
	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$
i	x_{i1}	x_{i2}	y_i
$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$
n	x_{n1}	x_{n2}	y_n
平均	$\bar{x}_1$	$\bar{x}_2$	$\bar{y}$

重回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,\ \varepsilon_i \sim N(0,1) \qquad (5.3)$$

(ε_i は互いに独立に $N(0,\sigma^2)$ に従う.)

(1) 最小２乗法による回帰式の推定

予測値と誤差

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \tag{5.4}$$

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) \tag{5.5}$$

残差平方和

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2})\}^2 \quad (5.6)$$

単回帰分析のときと同様に, S_e を最小にするような, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ を $\beta_0, \beta_1, \beta_2$ の推定値とする.

S_e を偏微分

(ホワイトボードで)

$$\frac{\partial S_e}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \quad (5.7)$$

$$\frac{\partial S_e}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \quad (5.8)$$

$$\frac{\partial S_e}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \quad (5.9)$$

(5.7)〜(5.8) を整理(ホワイトボードで)

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i \tag{5.10}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i \tag{5.11}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i2}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i \tag{5.12}$$

$\Rightarrow$ 正規方程式

$\hat{\beta}_0$ の計算

(5.10) 式より，

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2. \end{aligned} \tag{5.13}$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 \tag{5.14}$$

推定された重回帰式は点 $[\bar{x}_1, \bar{x}_2, \bar{y}]$ を通る．

(5.13) を (5.11) に代入

$$\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \frac{1}{n}\sum_{i=1}^{n} x_{i2} \right) \sum_{i=1}^{n} x_{i1}$$
$$+\hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i \tag{5.15}$$

(5.13) を (5.12) に代入

$$\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \frac{1}{n}\sum_{i=1}^{n} x_{i2}\right)\sum_{i=1}^{n} x_{i2}$$

$$+\hat{\beta}_1 \sum_{i=1}^{n} x_{i2}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i \tag{5.16}$$

(5.15) を整理(ホワイトボード)

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_{i1}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_{i1} \right)^2 \right)$$
$$+\hat{\beta}_2 \left(\sum_{i=1}^{n} x_{i1}x_{i2} - \frac{1}{n} \sum_{i=1}^{n} x_{i2} \sum_{i=1}^{n} x_{i1} \right) \tag{5.17}$$
$$= \sum_{i=1}^{n} x_{i1}y_i - \frac{1}{n} \sum_{i=1}^{n} x_{i1} \sum_{i=1}^{n} y_i$$

(5.16) を整理 (ホワイトボード)

$$
\begin{aligned}
&\hat{\beta}_1 \left(\sum_{i=1}^{n} x_{i1} x_{i2} - \frac{1}{n} \sum_{i=1}^{n} x_{i2} \sum_{i=1}^{n} x_{i1} \right) \\
&+ \hat{\beta}_2 \left(\sum_{i=1}^{n} x_{i2}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_{i2} \right)^2 \right) \\
&= \sum_{i=1}^{n} x_{i2} y_i - \frac{1}{n} \sum_{i=1}^{n} x_{i2} \sum_{i=1}^{n} y_i
\end{aligned}
\tag{5.18}
$$

平方和 S_{11}, S_{22}

$$\begin{aligned} S_{11} &= \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \\ &= \sum_{i=1}^{n} x_{i1}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_{i1}\right)^2 \end{aligned} \tag{5.19}$$

$$\begin{aligned} S_{22} &= \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2 \\ &= \sum_{i=1}^{n} x_{i2}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_{i2}\right)^2 \end{aligned} \tag{5.20}$$

偏差積和 S_{12}

$$
\begin{aligned}
S_{12} &= \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\
&= \sum_{i=1}^{n} x_{i1} x_{i2} - \frac{1}{n} \sum_{i=1}^{n} x_{i2} \sum_{i=1}^{n} x_{i1}
\end{aligned}
\tag{5.21}
$$

$$S_{11}\hat{\beta}_1 + S_{12}\hat{\beta}_2 = S_{1y} \tag{5.25}$$

$$S_{12}\hat{\beta}_1 + S_{22}\hat{\beta}_2 = S_{2y} \tag{5.26}$$

行列を使って書きなおすと,

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix} \tag{5.27}$$

と得る.

偏回帰係数

もし, $\det \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} \neq 0$ ならば, この行列は逆行列を持ち,

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix} = \frac{1}{S_{11}S_{22}-S_{12}^2} \begin{bmatrix} S_{22} & -S_{12} \\ -S_{12} & S_{11} \end{bmatrix} \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix} = \frac{1}{S_{11}S_{22}-S_{12}^2} \begin{bmatrix} S_{22}S_{1y} - S_{12}S_{2y} \\ -S_{12}S_{1y} + S_{11}S_{2y} \end{bmatrix}. \tag{5.28}$$

多重共線性

$\det \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} = 0$ の場合, データに「多重共線性が存在する」という. このとき,

$$\begin{aligned} S_{11}S_{22} - S_{12}^2 = 0 &\Leftrightarrow \frac{S_{12}^2}{S_{11}S_{22}} = 1 \\ &\Leftrightarrow r_{x_1x_2}^2 = \left\{ \frac{S_{12}}{\sqrt{S_{11}S_{22}}} \right\}^2 = 1 \\ &\Leftrightarrow r_{x_1x_2} = \pm 1. \end{aligned} \tag{5.31}$$

多重共線性 (続き)

即ち, 点 (x_{i1}, x_{i2}) $(i = 1, \cdots, n)$ の全てが一直線上に並んでいる状態である (p. 13 を見よ).
この場合, x_1 と x_2 のどちらか一方が分かれば, もう一方が決まってしまうので, y を説明する変数は一つでいいということになる.

$r_{x_1 x_2}$ がちょうど 1 や-1 でなくとも, 1 か-1 に近い値の場合も同様のことが言えるし, (5.28) の計算が精度上まずくなることもある.

例題 1(時間があれば)