データ解析 (第9回)

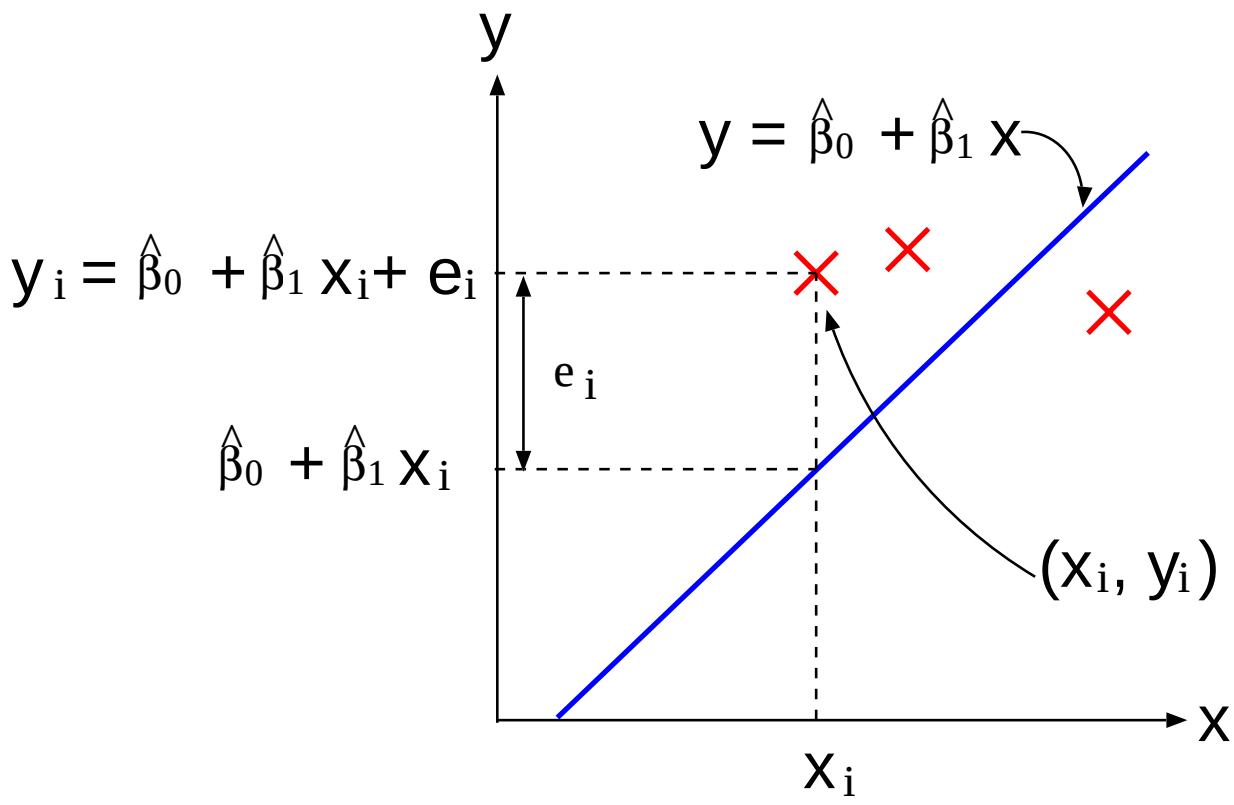
静岡大学システム工学科
安藤 和敏

第4章 (1) 最小2乗法による回帰式の推定

残差 (誤差は間違いでした!)

実際には残差 e_i が加わって,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i.$$



最小2乗法

したがって, データに「最もよくあてはまる」直線$y = \hat{\beta}_0 + \hat{\beta}_1 x$を求める問題は,

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2$$

を最小にするような, $\hat{\beta}_0$と$\hat{\beta}_1$を求める問題になった.

このようにして, $\hat{\beta}_0$, $\hat{\beta}_1$を求める方法を最小２乗法と呼んだ.

$\hat{\beta}_0$ と $\hat{\beta}_1$ の求め方

$$\text{(4.6)} \quad \frac{\partial S_e}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$\text{(4.7)} \quad \frac{\partial S_e}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

回帰直線

この連立方程式を解くと, $(\hat{\beta}_0, \hat{\beta}_1)$ は,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}.$$

ゆえに, データ (x_i, y_i) $(i = 1, \cdots, n)$ に最も良くあてはまる直線は,

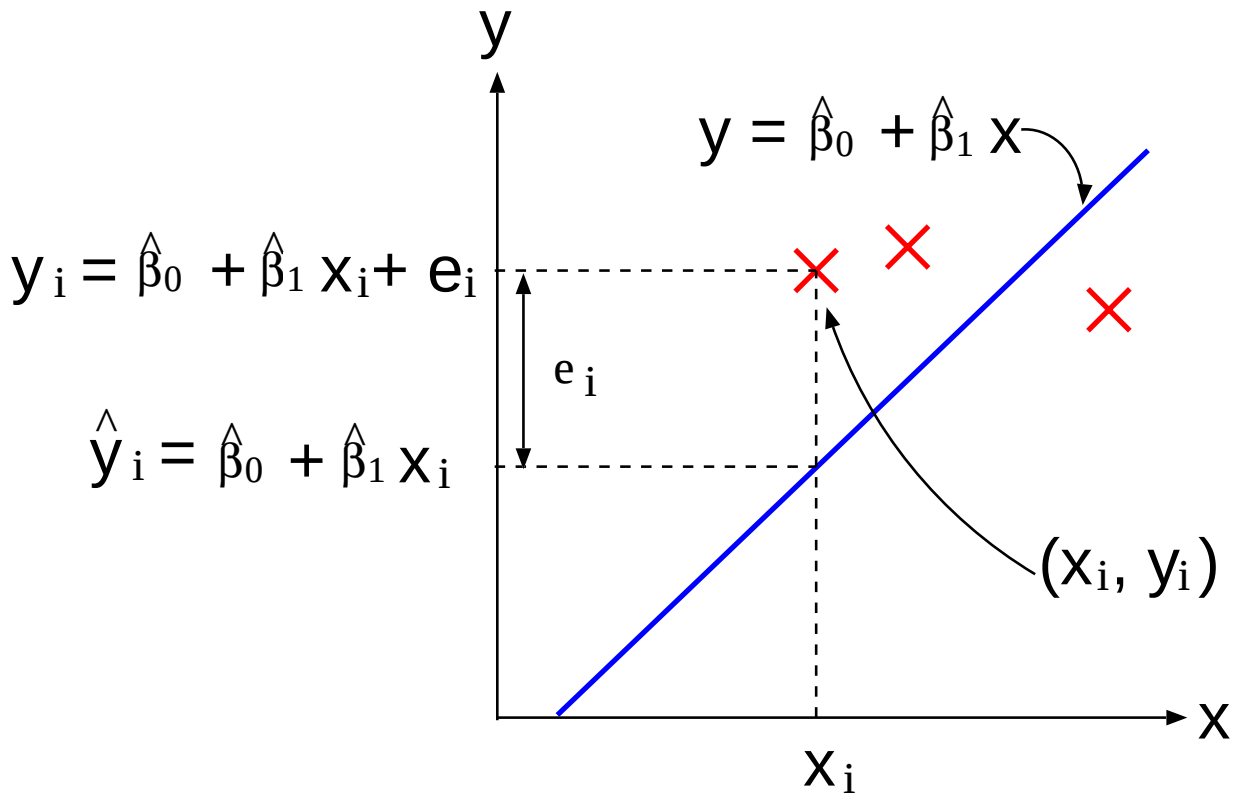
$$y = \frac{S_{xy}}{S_{xx}}(x - \bar{x}) + \bar{y}.$$

この式を y の x への回帰直線と呼び, $\hat{\beta}_1$ を回帰係数と呼ぶ.

予測値 $\hat{y}_i$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (4.3)$$

を予測値と呼ぶ.



(2) 寄与率と自由度調整済み寄与率 (p. 48–49)

平方和の分解(1)

$$
\begin{aligned}
S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}^2 \\
&= \sum_{i=1}^{n}\{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2 + \sum_{i=1}^{n}\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}^2 \\
&\quad +2\sum_{i=1}^{n}\{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\} \\
&= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}^2 + 2\sum_{i=1}^{n} e_i\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}
\end{aligned}
$$

$\sum_{i=1}^{n} e_i\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\} = 0$

$$\begin{aligned}
&\sum_{i=1}^{n} e_i\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\} \\
&= \sum_{i=1}^{n} e_i(\hat{\beta}_0 - \bar{y}) + \sum_{i=1}^{n} e_i \hat{\beta}_1 x_i \\
&= (\hat{\beta}_0 - \bar{y}) \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i .
\end{aligned}$$

ここで, (4.6) より $\sum_{i=1}^{n} e_i = 0$ であり, (4.7) より $\sum_{i=1}^{n} e_i x_i = 0$ であるから,

$$\sum_{i=1}^{n} e_i\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\} = 0.$$

平方和の分解(2)

ゆえに,

$$
\begin{aligned}
S_{yy} &= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n} \{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}^2 \\
&= S_e + \sum_{i=1}^{n} \{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}^2 \qquad (4.19)
\end{aligned}
$$

平方和の分解(3)

その一方で,

$$S_e = S_{yy} - \hat{\beta}_1 S_{xy} \qquad (4.17)$$

であるから, $S_R = \hat{\beta}_1 S_{xy} = \frac{S_{xy}^2}{S_{xx}}$ と置くと,

$$S_{yy} = S_e + S_R. \qquad (4.20)$$

(4.19) と (4.20) より,

$$S_R = \sum_{i=1}^{n} \{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\}^2$$

平方和の分解(4)

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{S_{yy}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{S_e} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{S_R}$$

y の全変動 ＝ 誤差による変動 ＋ 回帰による変動

寄与率

y の全変動のうち, 回帰による変動 S_R の占める割合

$$R^2 = S_R/S_{yy} = \frac{S_{yy} - S_e}{S_{yy}}$$

$$= 1 - \frac{S_e}{S_{yy}} \qquad (4.22)$$

は寄与率と呼ばれ, 1 に近いほどよい. 当然のことながら, $0 \leq R^2 \leq 1$ である.

寄与率と相関係数

$$R^2 = \frac{S_R}{S_{yy}} = \frac{S_{xy}^2/S_{xx}}{S_{yy}}$$

$$= \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = r_{xy}^2 \qquad (4.23)$$

このことからも,

$$-1 \leq r_{xy} \leq 1$$

であることがわかる (問題 2.11).

自由度と自由度調整済寄与率 (補正 R2)

(4.19) 式の各平方和に対して, 以下のように自由度が対応する

平方和	自由度
S_{yy}	$\phi_T = n - 1$
S_R	$\phi_R = 1$
S_e	$\phi_e = n - 2$

$$R^{*2} = 1 - \frac{S_e/\phi_e}{S_{yy}/\phi_T} \qquad (4.24)$$

は, 自由度調整済寄与率と呼ばれる.

例題1(先週の続き)

時間があれば...