データ解析 (第 11 回)

安藤 和敏

静岡大学システム工学科

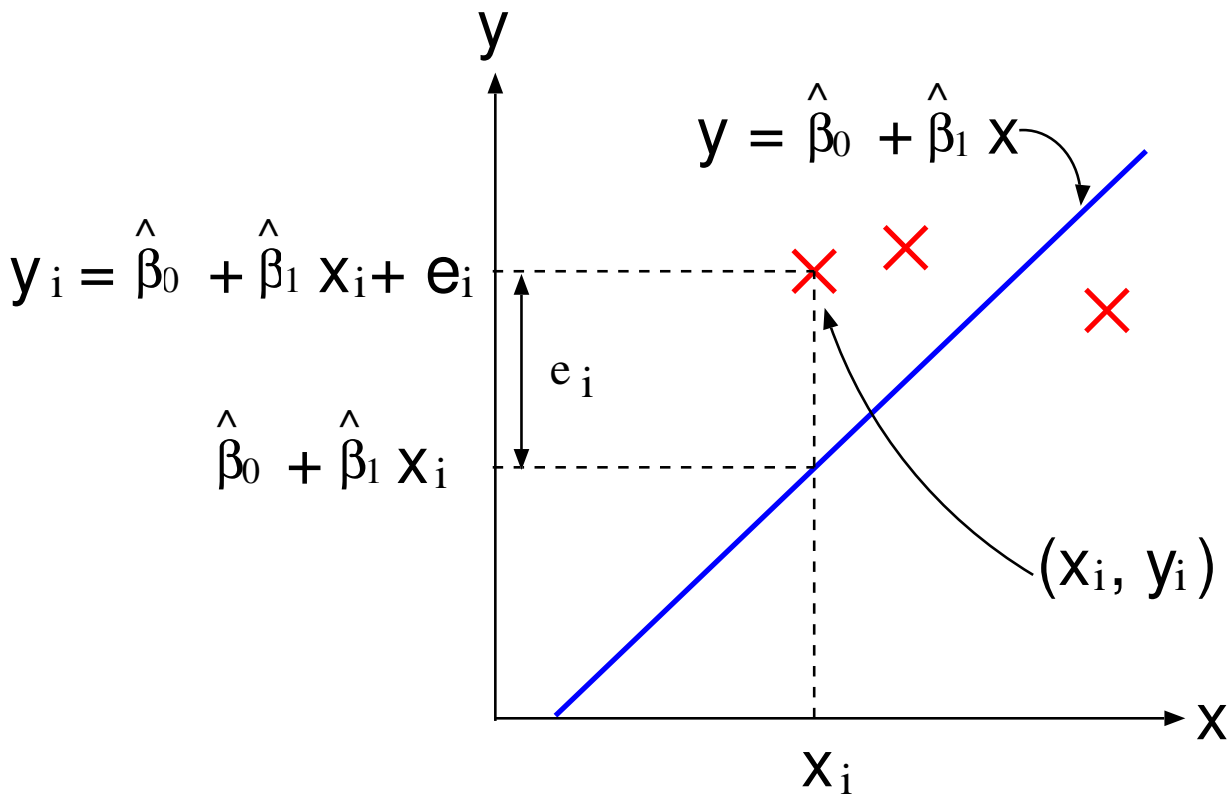
[復習] 単回帰分析のデータ

データ 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

[復習] 残差

実際には残差 e_i が加わって,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i. \quad (i = 1, \cdots, n)$$



[復習] 回帰係数

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.

[復習] 回帰直線

最小２乗法によって求まる $(\hat{\beta}_0, \hat{\beta}_1)$ は,

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

(3) 回帰係数の検定と推定 (p. 50–51)

前回までは確率の話は入ってこなかったが, 今日から確率の入った話をする.

- 単回帰モデル
- 確率変数 y_i
- 確率変数 $\hat{\beta}_1$
- $\hat{\beta}_1$ から導かれる t-分布

単回帰モデル

単回帰モデル

- 説明変数 x と目的変数の y の間には,
  $y = \beta_0 + \beta_1 x$ という関係があることを仮定.

単回帰モデル

- 説明変数 x と目的変数の y の間には,
  $y = \beta_0 + \beta_1 x$ という関係があることを仮定.
- ただし, β_0 と β_1 は未知のパラメータであるので,
  それを「推定」することしかできない.

単回帰モデル

- 説明変数 x と目的変数の y の間には,
  $y = \beta_0 + \beta_1 x$ という関係があることを仮定.
- ただし, β_0 と β_1 は未知のパラメータであるので,
  それを「推定」することしかできない.
- 前回求めた $\hat{\beta}_0$, $\hat{\beta}_1$ は, これらの推定量である.

単回帰モデル

- 説明変数 x と目的変数の y の間には,
  $y = \beta_0 + \beta_1 x$ という関係があることを仮定.
- ただし, β_0 と β_1 は未知のパラメータであるので,
  それを「推定」することしかできない.
- 前回求めた $\hat{\beta}_0$, $\hat{\beta}_1$ は, これらの推定量である.
- 実際に観測されるデータ (x_i, y_i) の間には,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (4.1)$$

という関係を仮定する.

単回帰モデル

- 説明変数 x と目的変数の y の間には,
  $y = \beta_0 + \beta_1 x$ という関係があることを仮定.
- ただし, β_0 と β_1 は未知のパラメータであるので,
  それを「推定」することしかできない.
- 前回求めた $\hat{\beta}_0$, $\hat{\beta}_1$ は, これらの推定量である.
- 実際に観測されるデータ (x_i, y_i) の間には,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (4.1)$$

という関係を仮定する. ここで, ε_i は, 期待値 0
分散 σ^2 の正規分布にしたがう確率変数である.

単回帰モデル

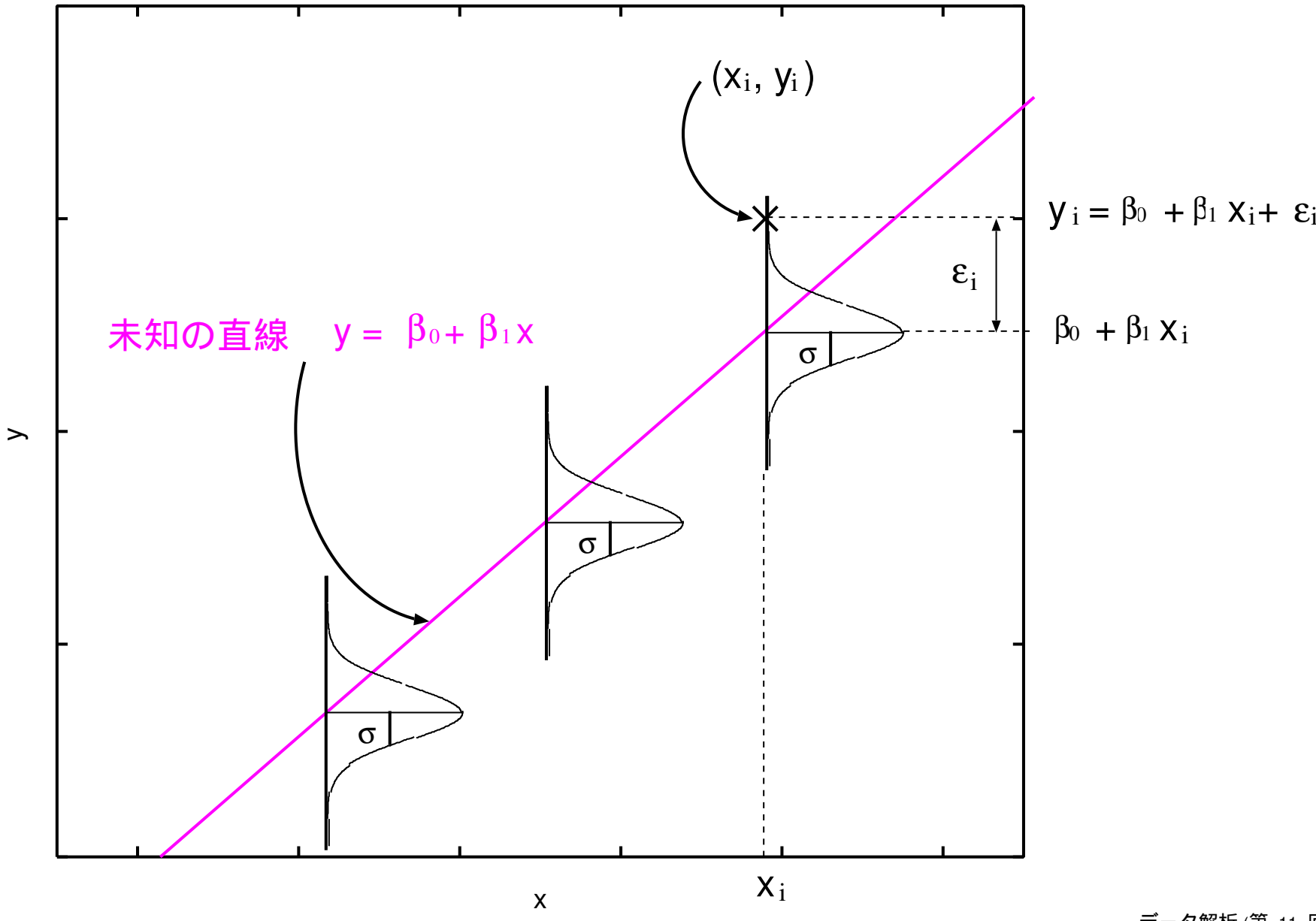
- 説明変数 x と目的変数の y の間には, $y = \beta_0 + \beta_1 x$ という関係があることを仮定.
- ただし, β_0 と β_1 は未知のパラメータであるので, それを「推定」することしかできない.
- 前回求めた $\hat{\beta}_0$, $\hat{\beta}_1$ は, これらの推定量である.
- 実際に観測されるデータ (x_i, y_i) の間には,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.1}$$

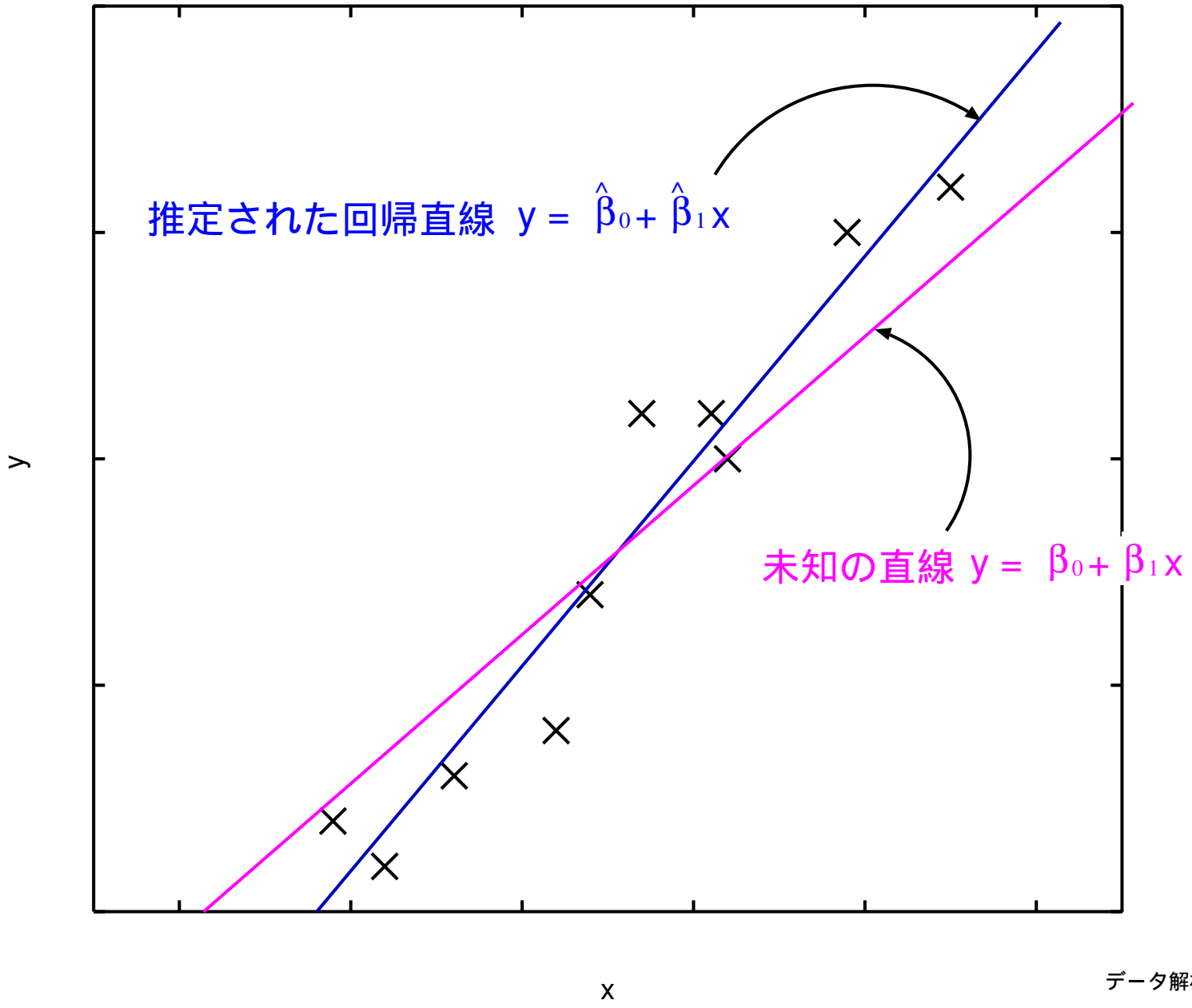
という関係を仮定する. ここで, ε_i は, 期待値 0 分散 σ^2 の正規分布にしたがう確率変数である.

- さらに, σ^2 も未知のパラメータ.

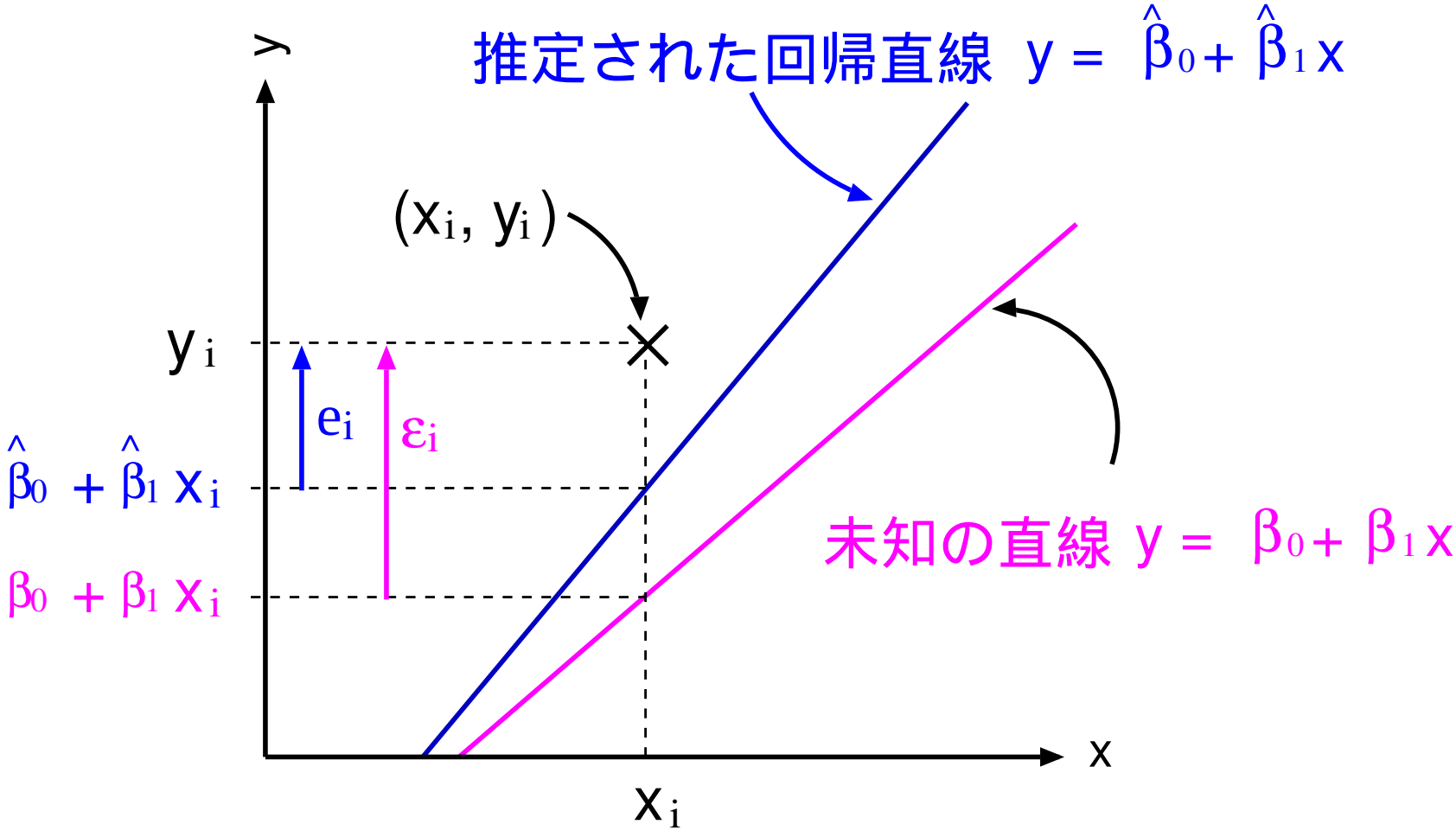
単回帰モデルのイメージ(テキスト図4.2)



未知の曲線と推定された回帰曲線



e_i と ε_i の関係



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

確率変数 y_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

(β_0, β_1, x_i は定数で, ε_i は正規分布にしたがう確率変数) であるから, y_i も正規分布にしたがう確率変数になる.

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ の期待値

$$E(y_i) = \beta_0 + \beta_1 x_i \quad (i = 1, \cdots, n).$$

(証明)

$$\begin{aligned} E(y_i) &= E(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 x_i + E(\varepsilon_i) \qquad (\because \text{(2.24) より}) \\ &= \beta_0 + \beta_1 x_i. \qquad (\because \varepsilon_i \sim N(0, \sigma^2)) \end{aligned}$$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$の分散

$$V(y_i) = \sigma^2 \quad (i = 1, \cdots, n).$$

(証明)

$$\begin{aligned} V(y_i) &= V(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= V(\varepsilon_i) \qquad (\because \text{(2.27) より}) \\ &= \sigma^2. \qquad (\because \varepsilon_i \sim N(0, \sigma^2)) \end{aligned}$$

y_i がしたがう分布

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2) \quad (i = 1, \cdots, n)$$

のとき,

y_i は期待値 $\beta_0 + \beta_1 x_i$, 分散 σ^2 の正規分布にしたがう:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad (i = 1, \cdots, n).$$

ここで, $\beta_0, \beta_1, \sigma^2$ は未知のパラメータである.

y_i と y_j の共分散 $(i \neq j)$

$i \neq j$ ならば,

$$C(y_i, y_j) = 0.$$

(証明)

$$\begin{aligned}
C(y_i, y_j) &= E((y_i - E(y_i))(y_j - E(y_j))) \qquad (\because \text{(2.43)}) \\
&= E(\varepsilon_i \varepsilon_j) \qquad (\because E(y_i) = \beta_0 + \beta_1 x_i) \\
&= E(\varepsilon_i) E(\varepsilon_j) \qquad (\because \varepsilon_i \text{ と } \varepsilon_j \text{ は独立}) \\
&= 0 \cdot 0 \qquad (\because \varepsilon_i \sim N(0, \sigma^2)) \\
&= 0.
\end{aligned}$$

確率変数 $\hat{\beta}_1$

$\hat{\beta}_1$ は正規分布にしたがう確率変数 y_i の一次結合:

$$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i. \qquad (*)$$

であるから, $\hat{\beta}_1$ も正規分布にしたがう.

式(*)の証明

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \qquad (\because \text{式(4.15)}) \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})y_i - \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\bar{y}
\end{aligned}$$

式(*)の証明(続き)

$$
\begin{aligned}
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \frac{1}{S_{xx}} \bar{y} \sum_{i=1}^{n} (x_i - \bar{x}) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i .
\end{aligned}
$$

$\hat{\beta}_1$ の期待値

$$
\begin{aligned}
E(\hat{\beta}_1) &= E\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})y_i\right) \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})E(y_i) \qquad (\because \text{式 (2.24)}) \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i) \quad (\because E(y_i)=\beta_0 - \\
&= \frac{\beta_0}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}) + \frac{\beta_1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})x_i
\end{aligned}
$$

$\hat{\beta}_1$ の期待値(続き)

$$
\begin{aligned}
&= \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) x_i \\
&= \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) x_i - \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \bar{x} \\
&= \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) \\
&= \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}
$$

$\hat{\beta}_1$ の期待値(続き)

$$
\begin{aligned}
&= \frac{\beta_1}{S_{xx}} S_{xx} \\
&= \beta_1.
\end{aligned}
$$

$E(\hat{\beta}_1) = \beta_1$. すなわち, $\hat{\beta}_1$ は β_1 の不偏推定量である.

$\hat{\beta}_1$ の分散

$$
\begin{aligned}
V(\hat{\beta}_1) &= V\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})y_i\right) \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 V(y_i) \qquad (\because \text{式 (2.51)}) \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 \qquad (\because V(y_i)=\sigma^2) \\
&= \frac{\sigma^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2
\end{aligned}
$$

$\hat{\beta}_1$ の分散 (続き)

$$
\begin{aligned}
&= \frac{\sigma^2}{S_{xx}^2} S_{xx} \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}
$$

$\hat{\beta}_1$ のしたがう分布

$\hat{\beta}_1$ は期待値 β_1, 分散 $\frac{\sigma^2}{S_{xx}}$ の正規分布にしたがう:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \qquad (4.25)$$

ここで, β_1 と σ^2 は未知のパラメータである.