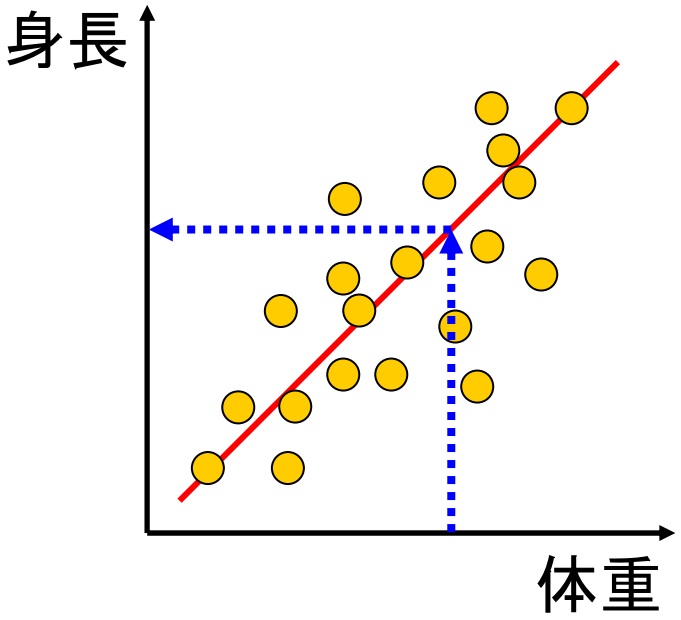
03. 回帰分析

京都大学　加納 学

Division of Process Control & Process Systems Engineering
Department of Chemical Engineering, Kyoto University

manabu@cheme.kyoto-u.ac.jp
http://www-pse.cheme.kyoto-u.ac.jp/~kano/

身近な例



体重から身長を推定できる？

身長 = 定数 × 体重 + 定数 + 誤差
y　　b_1　　x　　b_0　　e

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2$$ を最小にする係数 b を求める.

因果関係を探る

- （単）回帰分析
  結果である目的変数 y と原因である説明変数 x の関係を単回帰式を用いて表現しようとする手法

$$y = b_1 x + b_0$$

（標本）偏回帰係数

現実には，目的変数は説明変数以外の要因にも影響されるため，それらの n 番目の標本(測定値)が単回帰モデルによって表現されると考える．

$$y_n = \beta_1 x_n + \beta_0 + \varepsilon_n$$

母偏回帰係数

誤差項 ε_n は互いに独立に $\mathrm{N}(0, \sigma^2)$ に従うと仮定する．

目的変数の予測

- 目的変数の期待値
  誤差項 ε_n の期待値は 0 であるから

$$E(y_n) = E(\beta_1 x_n + \beta_0 + \varepsilon_n) = \beta_1 x_n + \beta_0$$

- 目的変数の予測値
  母偏回帰係数 β_0, β_1 の推定値 b_0, b_1 が得られれば

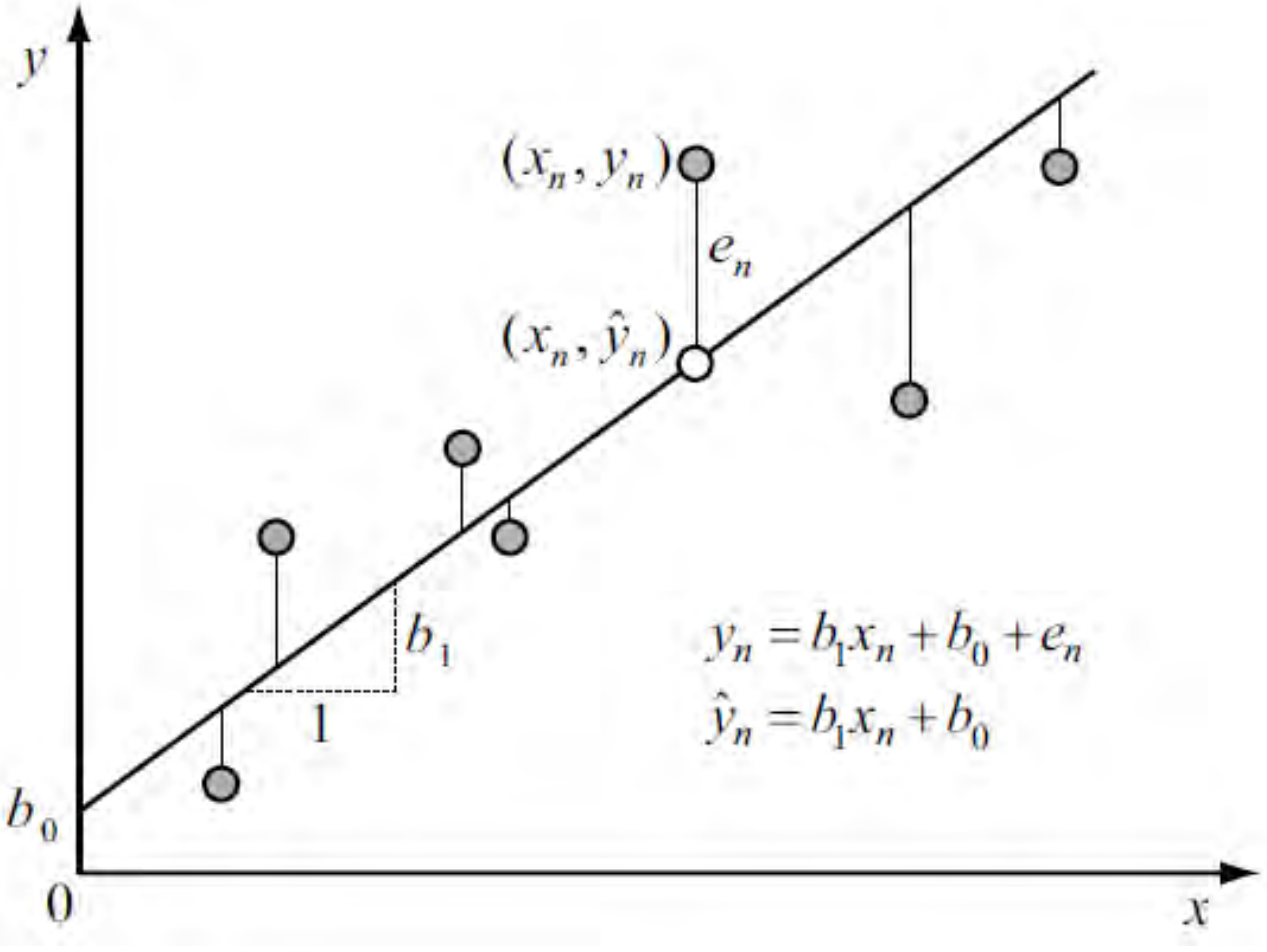
$$\hat{y}_n = b_1 x_n + b_0$$

- 残差
  目的変数の測定値と予測値の差

$$e_n = y_n - \hat{y}_n = y_n - b_1 x_n - b_0$$

回帰分析における誤差の考え方

- 目的変数 y に影響を与える説明変数 x 以外の要因をまとめて誤差とみなすため，y のみに誤差がある，つまり，x は正確に指定できると考える．



最小二乗法

- 最小二乗法

  残差平方和(目的変数の測定値と推定値の差の二乗和)が最小となるように，偏回帰係数を決定する.

予測値

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$$

残差平方和

$$Q = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

$$= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}})$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

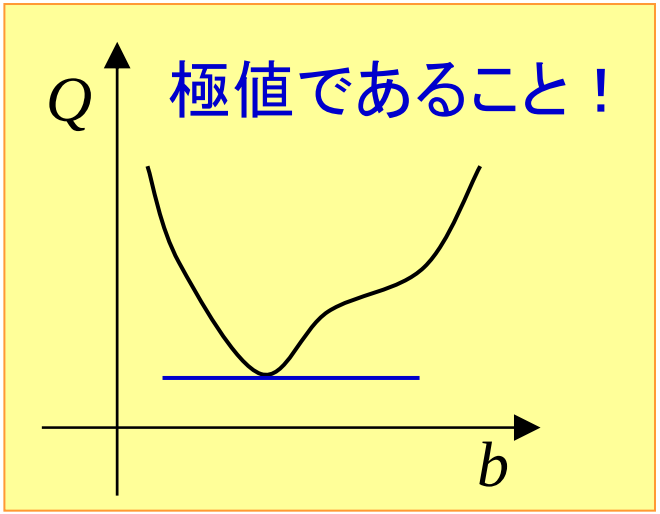
正規方程式の導出

- 残差平方和

$$Q = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}})$$
$$= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b}$$

- 必要条件

$$\frac{1}{2}\frac{\partial Q}{\partial \boldsymbol{b}} = \frac{1}{2}\left[\frac{\partial Q}{\partial b_0} \quad \frac{\partial Q}{\partial b_1}\right]^T = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b} - \boldsymbol{X}^T \boldsymbol{y} = 0$$



正規方程式

$$\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b} = \boldsymbol{X}^T \boldsymbol{y}$$

偏回帰係数の推定

- 正規方程式

$$\begin{bmatrix} N & N\bar{x} \\ N\bar{x} & \boldsymbol{x}^T\boldsymbol{x} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} N\bar{y} \\ \boldsymbol{x}^T\boldsymbol{y} \end{bmatrix}$$

- 偏回帰係数の推定値

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\boldsymbol{x}^T\boldsymbol{y} - N\bar{x}\bar{y}}{\boldsymbol{x}^T\boldsymbol{x} - N\bar{x}^2} = \frac{(\boldsymbol{x} - \bar{x}\boldsymbol{1})^T(\boldsymbol{y} - \bar{y}\boldsymbol{1})}{(\boldsymbol{x} - \bar{x}\boldsymbol{1})^T(\boldsymbol{x} - \bar{x}\boldsymbol{1})} = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}^2}{s_x^2}$$

偏差平方和

$$S_{xx} = \sum_{n=1}^{N}(x_n - \bar{x})^2$$

偏差積和

$$S_{xy} = \sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

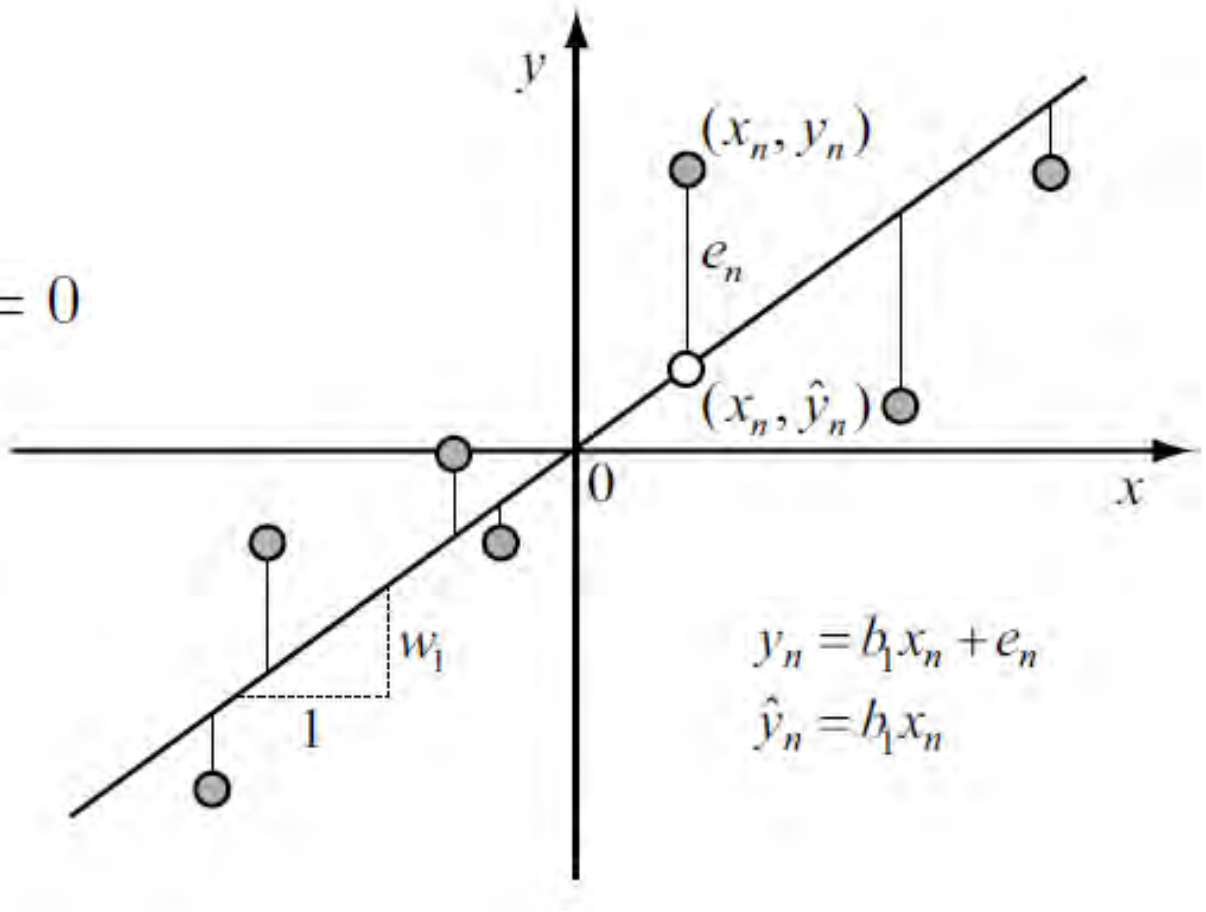
平均を0にする意味

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\bar{y} = \bar{x} = 0$$

$$b_0 = 0$$

$$\hat{y}_n = b_1 x_n$$



単回帰式（直線）が原点を通る（切片が0 となる）ように，データを移動させる．

標準化の効果

- 単回帰分析で得られる偏回帰係数の値は，説明変数と目的変数の大きさによって変化する．
  - 説明変数が質量である場合，その測定単位が kg か g かによって偏回帰係数の値は異なる．
- 測定単位の影響を排除するデータの前処理方法として，各変数の平均を0，分散を1 にする標準化がある．

$$y = \frac{y^* - \bar{y}^*}{s_{y^*}} \qquad x = \frac{x^* - \bar{x}^*}{s_{x^*}}$$

$$s_x^2 = \frac{1}{N-1}\boldsymbol{x}^T\boldsymbol{x} = 1$$

$$r_{xy} = \frac{s_{xy}^2}{s_x s_y} = \frac{1}{N-1}\boldsymbol{x}^T\boldsymbol{y}$$

標準偏回帰係数

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{x}} = r_{xy}$$

標準偏回帰係数 ＝ 相関係数

回帰分析の続き

- 母偏回帰係数を点推定し，単回帰式を求める.
- 分散分析により，回帰式が役に立つかどうかを調べる.
- 寄与率や重相関係数を求め，回帰式を評価する.
- 残差が仮定を満たしているかを確認する.
- 偏回帰係数の検定や区間推定を行う.