多変量解析

～第11回 重回帰分析～

回帰分析

- 回帰分析と最小2乗法の記述統計的な側面を概観する
- 線形モデルや最小2乗法という意味では，分散分析も回帰分析とまったく同じ手法である
- 線形代数と直交射影の観点から最小2乗法の理論を解説
- 回帰分析では，説明変数の値を用いて目的変数の値を説明，あるいは予測することを目的とする
- 正規線形回帰モデルの推測については，概要のみ

回帰分析の行列表示

- 説明変数は独立変数ともいう．目的変数は従属変数，非説明変数，基準変数などともいう
- 重回帰モデル

説明変数

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

目的変数　回帰係数　誤差

ただし，
例）$E(\epsilon) = 0,\ V(\epsilon) = \sigma^2$

$$\mathrm{Math} = \beta_0 + \beta_1 \times \mathrm{Physics} + \beta_2 \times \mathrm{English} + \epsilon$$

- 線形式（上式）でモデルを記述することを線形回帰と呼ぶ

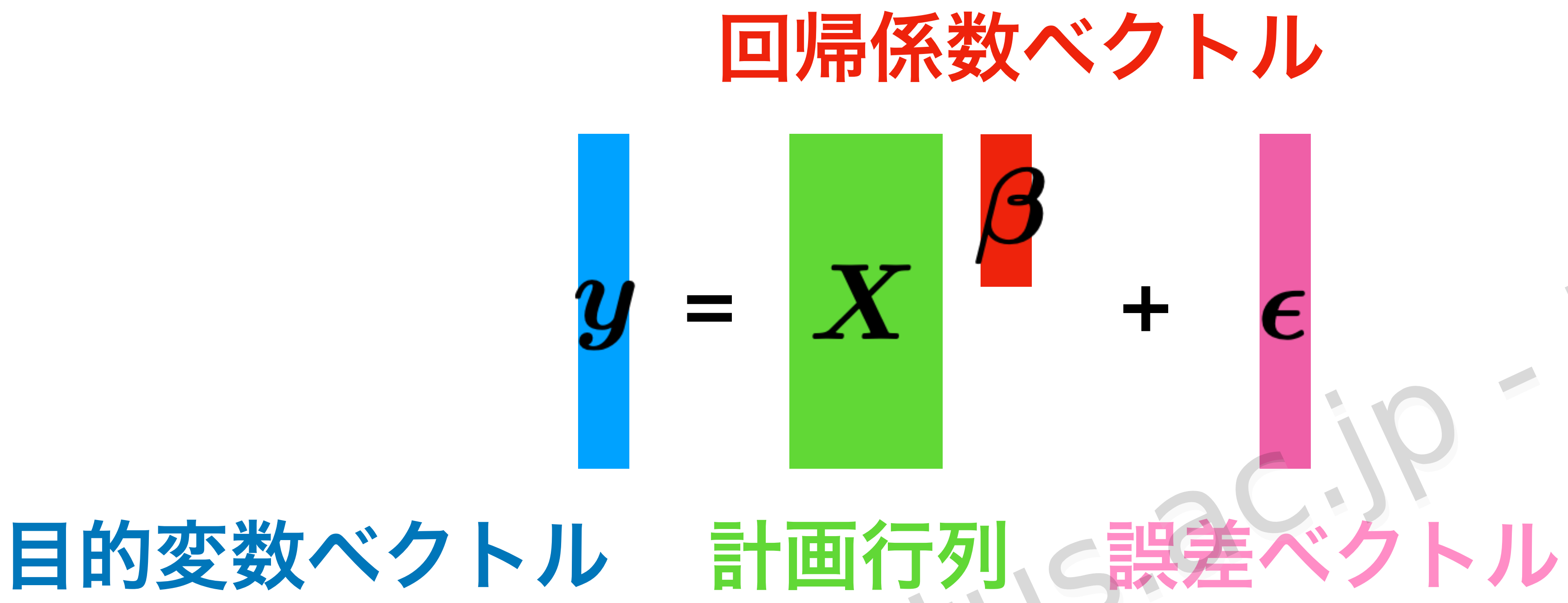
- 行列表示

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- ここでは, $n > p + 1$ とする

- イメージ



最小2乗法

- 最小2乗法の考え方
  $$\sum_{t=1}^{n}(y_t-\beta_0 x_{t0}-\cdots-\beta_p x_{tp})^2=||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||^2$$
  が最小となる$\boldsymbol{\beta}$を求めたい．ただし，
  $$||\boldsymbol{x}||=\sqrt{x_1^2+x_2^2+\cdots+x_p^2},\quad \boldsymbol{x}=(x_1,x_2,\ldots,x_p)'$$

- $n>p+1$ かつ $\mathrm{rank}\boldsymbol{X}=p+1$を仮定する:
  つまり$\boldsymbol{X}$の列ベクトルの集合が線型独立である
  ベクトルの集合が線型独立または一次独立であるとは、集合のベクトルの線型結合によるゼロベクトルの表示が自明なものに限ることをいう

$$\boldsymbol{1}_n\beta_0+\boldsymbol{x}_1\beta_1+\cdots+\boldsymbol{x}_p\beta_p=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix}\begin{pmatrix}\beta_0\\ \vdots\\ \beta_p\end{pmatrix}$$

- 最小2乗解：$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

- 解の確認：任意のベクトル $\boldsymbol{c}$ に対して，

$$
\begin{aligned}
||\boldsymbol{y} - \boldsymbol{X}c||^2 &= ||\boldsymbol{y} - \boldsymbol{X}c - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 \\
&= \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{c})\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{c})\right) \\
&= ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 + ||\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{c})||^2 \\
&\geq ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2
\end{aligned}
$$

(注) $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{y} - \underline{\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0}$

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y}$$

- 予測値ベクトル：$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P_X}\boldsymbol{y}$
  ただし, $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

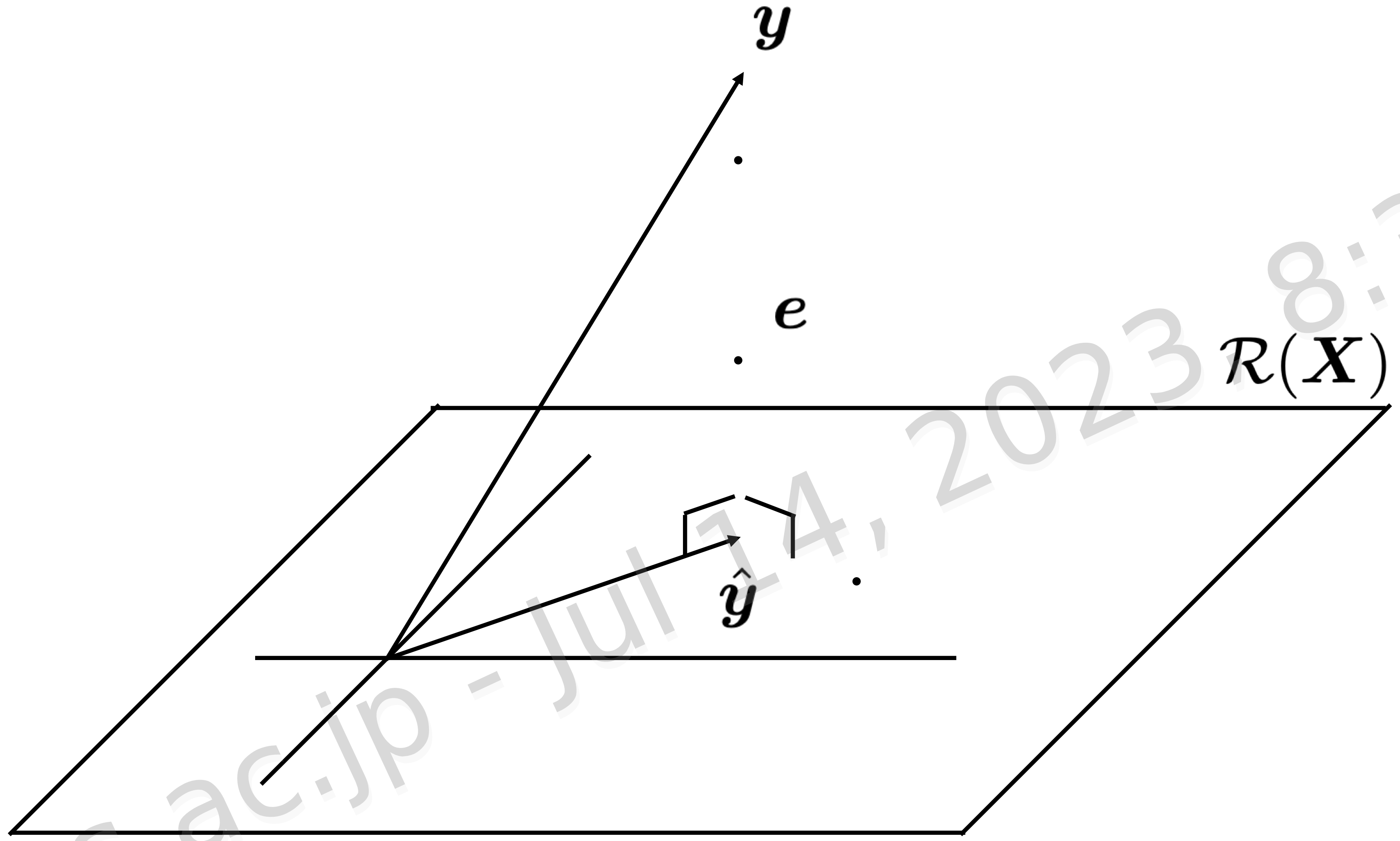
- 残差ベクトル： $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}$

- 直交射影行列：べき等性かつ対称性を満たす行列

$$\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P_X}, \quad \boldsymbol{P_X} = \boldsymbol{P}_{\boldsymbol{X}}'$$

- $\boldsymbol{X}$の列の張る$\mathcal{R}^n$の中の
  $p+1$次元部分空間：$\mathcal{R}(\boldsymbol{X})$

- $\mathcal{R}(\boldsymbol{X})$ の直交補空間：$\mathcal{R}(\boldsymbol{X})^{\perp}$



決定係数

- 定義的な関係式：$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$
  左から $\mathbf{1}_n^{'}$ をかけて要素の和をとると，$\mathbf{1}_n^{'}\boldsymbol{e} = 0$ である．
  したがって，予測値ベクトルの標本平均は実測値ベクトルの標本平均 $\bar{y}$ に一致することがわかる

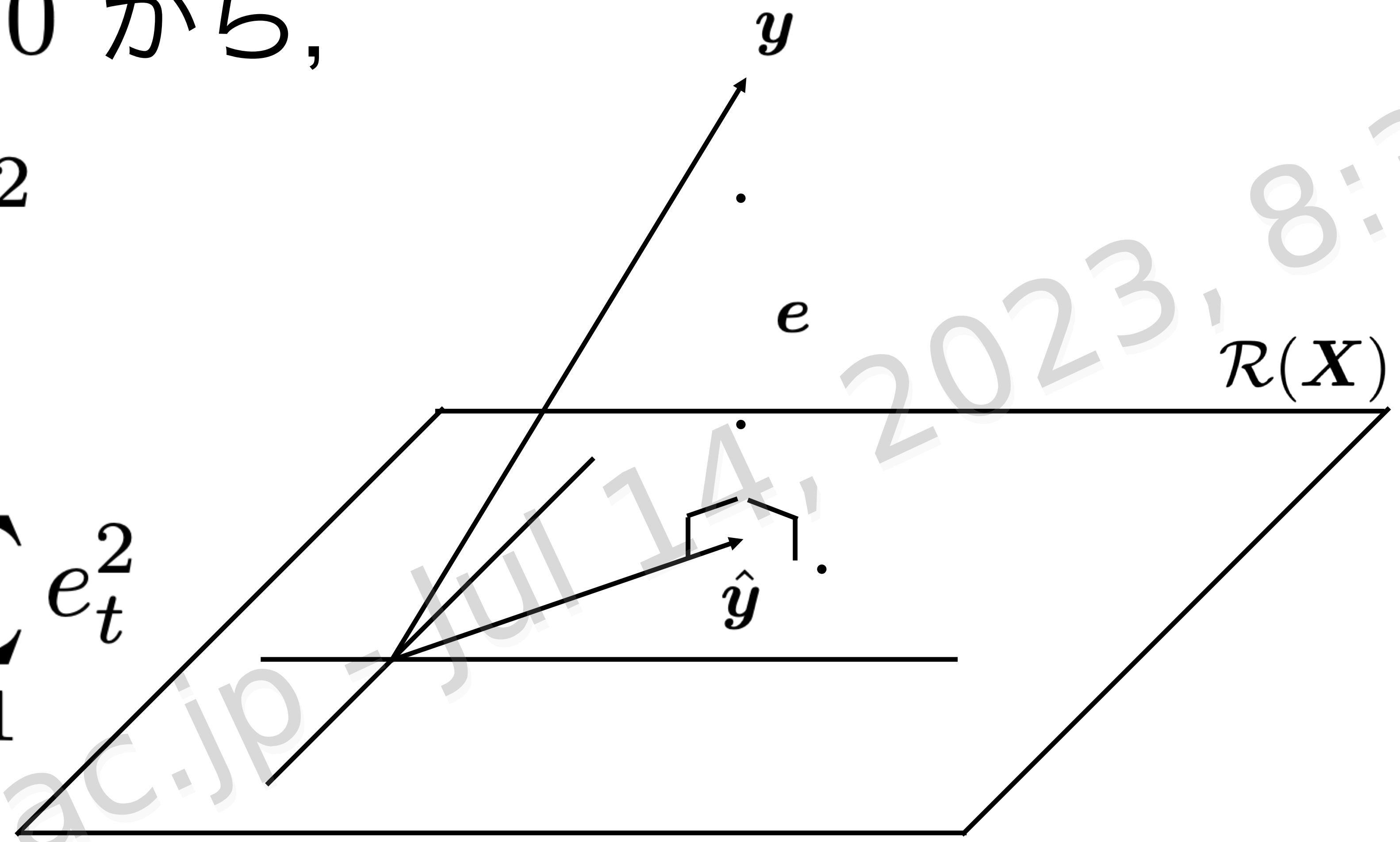
$$\underbrace{\boldsymbol{X}^{'}}_{1\text{行目が}\mathbf{1}_n^{'}}\underbrace{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}_{\text{残差ベクトル}\boldsymbol{e}} = \mathbf{0}$$

- 平均偏差：$\boldsymbol{y} - \bar{y}\mathbf{1}_n = (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n) + \boldsymbol{e}$
  $\hat{\boldsymbol{y}}^{'}\boldsymbol{e} = \hat{\boldsymbol{\beta}}^{'}\boldsymbol{X}^{'}\boldsymbol{e} = 0$ および $\mathbf{1}_n^{'}\boldsymbol{e} = 0$ から，

  $||\boldsymbol{y} - \bar{y}\mathbf{1}_n||^2 = ||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||^2 + ||\boldsymbol{e}||^2$

  要素ごとにかけば

$$\sum_{t=1}^{n}(y_t - \bar{y})^2 = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n} e_t^2$$



- 平方和の分解

$$\sum_{t=1}^{n}(y_t-\bar{y})^2=\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2+\sum_{t=1}^{n}e_t^2$$

全平方和　　回帰平方和　　残差平方和

- 決定係数：回帰式のあてはまりの良さの尺度

$$R^2=\frac{\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2}{\sum_{t=1}^{n}(y_t-\bar{y})^2}=1-\frac{\sum_{t=1}^{n}e_t^2}{\sum_{t=1}^{n}(y_t-\bar{y})^2}$$

$0\le R^2\le 1$であり，R^2が1に近いほど$\hat{\boldsymbol{y}}$が$\boldsymbol{y}$に近く，回帰式のあてはまりが良い．また，正の平方根Rを重相関係数という

Rによる実習

- treesデータ
  31本のBkack cherryの倒木の外周長（地面から4フィート6インチの円周）・樹高・容積データ

- コーヒーブレイク☕
  樹木の容積は外周長の2乗と樹高の積に比例？
  この関係式を線形モデルで表現するには対数を取れば良い？

木材博物館より引用

- treesデータの確認と準備

```
str(trees) # treesデータの構造を確認
y <- trees[,3] # treesの3列目を目的変数ベクトルとして格納
a <- rep(1:1,31) # 1ベクトルの作成
X <- cbind(a,trees[,-3]) # 計画行列の作成
X <- as.matrix(X) # 行列に変更
```

- 最小二乗推定値の計算

```
b <- solve(t(X) %*% X) %*% t(X) %*% y
esty <- X %*% b # 予測値ベクトル
c(mean(y),mean(esty)) # 平均値の確認
```

- 各平方和の計算

```
t(y-mean(y)) %*% (y-mean(y)) # 全平方和
t(esty-mean(y)) %*% (esty-mean(y)) # 回帰平方和
(e <- y - esty) # 残差ベクトル
t(e) %*% e # 残差平方和
```

- 平方和の分解

```
zen <- t(y-mean(y)) %*% (y-mean(y)) # 全平方和
kaiki <- t(esty-mean(y)) %*% (esty-mean(y)) # 回帰平方和
zansa <- t(e) %*% e # 残差平方和
(zen - (kaiki + zansa))
```

- 決定係数

```
1 - zansa/zen # 決定係数1
kaiki/zen #決定係数2（同じ計算結果）
```

- 組み込み関数lmで自動計算

```
res <- lm(Volume~1+Girth+Height,data=trees) # 組み込み関数で実施
summary(res)
```

- （演習問題）データを対数変換して回帰分析してみる