KYOTO UNIVERSITY

統計的モデリング基礎②
～回帰モデリング～

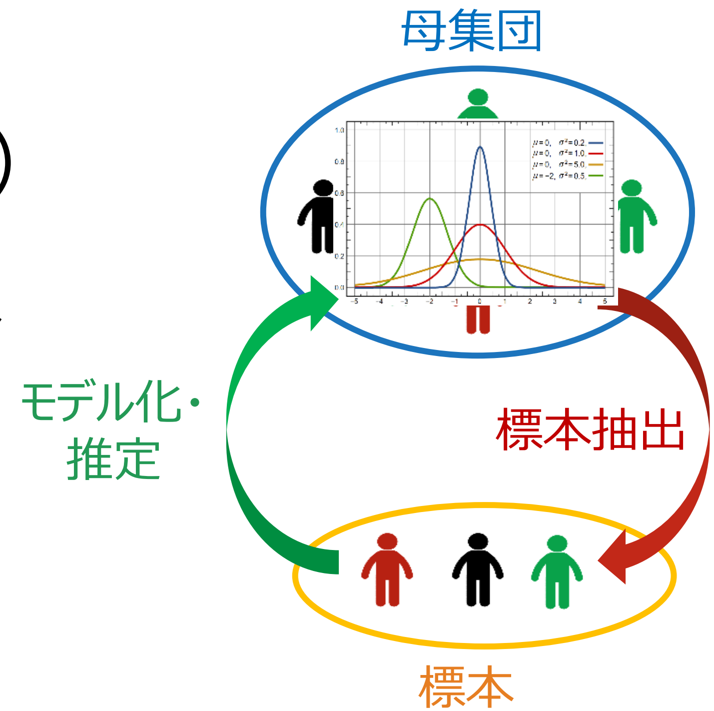
鹿島久嗣
（情報学科 計算機科学コース）

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

統計的モデリングの考えかた

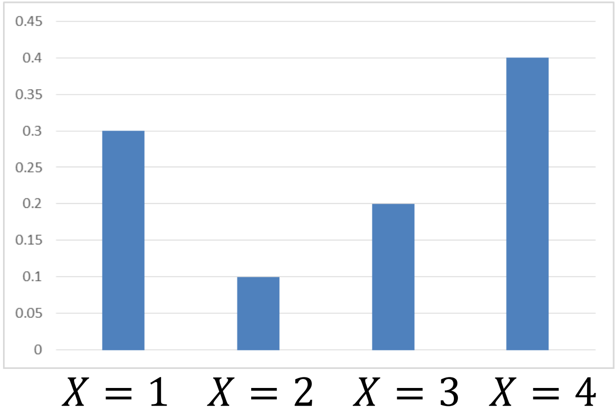
統計モデリングの考え方：
部分から全体について知る

- 母集団：
  - 興味のある集合のすべての要素
  - 確率分布
    （分布のクラスやパラメータで指定される）
- 標本：母集団からの無作為抽出あるいは確率分布に従った抽出
  - 確率変数：確率的に値が決まる変数
- 標本から母集団について推測する
  （標本抽出の逆）



離散型確率変数の代表的な確率分布：離散分布、ベルヌーイ分布と2項分布

- 離散分布 $P(X=k)=f(k)$（但し $\sum_{k\in\mathcal{X}} f(k)\ =1, f(k)\ \geq 0$）
- ベルヌーイ分布：$\mathcal{X}=\{0,1\}$上の離散分布



- 二項分布
  - ベルヌーイ試行： 1が出る確率pのベルヌーイ分布からn回 独立に抽出する
  - 二項分布：ベルヌーイ試行において1がk回出る確率を与える

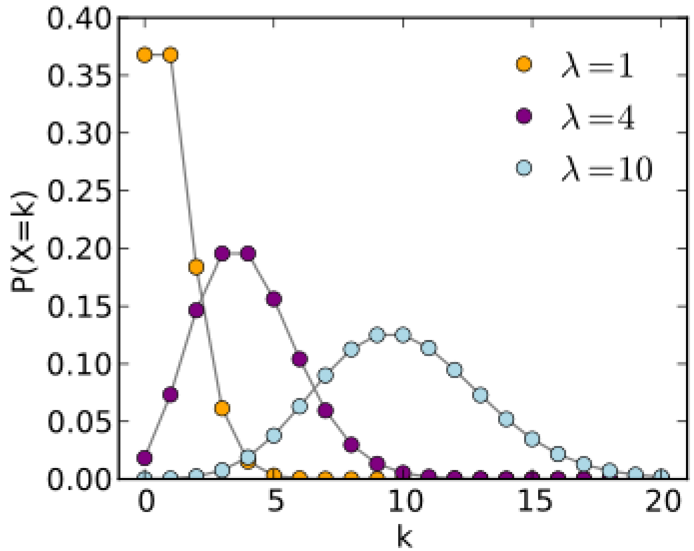
$$P(X=k\mid p)=\binom{n}{k}p^k(1-p)^{n-k}$$

n回の試行中のどこでk回の1が現れるかの場合の数

  - モデルパラメータpによって分布の形が一意に決定される

離散型確率変数の代表的な確率分布：ポアソン分布（2項分布の極限）、その他

- ポアソン分布：$P(X = k \mid \lambda) = \dfrac{\lambda^k}{k!}\exp(-\lambda)$
  - 比較的稀な事象が何回起こるかを表現
    - 1分あたりのWebサーバアクセス数
    - ロットあたりの不良品数
  - パラメータ $\lambda > 0$
    - ２項分布のパラメータ（n, p）がない
    - ２項分布で$np = \lambda$として、$n \to \infty, p \to 0$ とするとポアソン分布になる



https://en.wikipedia.org/wiki/Poisson_distribution#/media/File:Poisson_pmf.svg

- ほか、離散型の確率分布には幾何分布、負の2項分布などがある

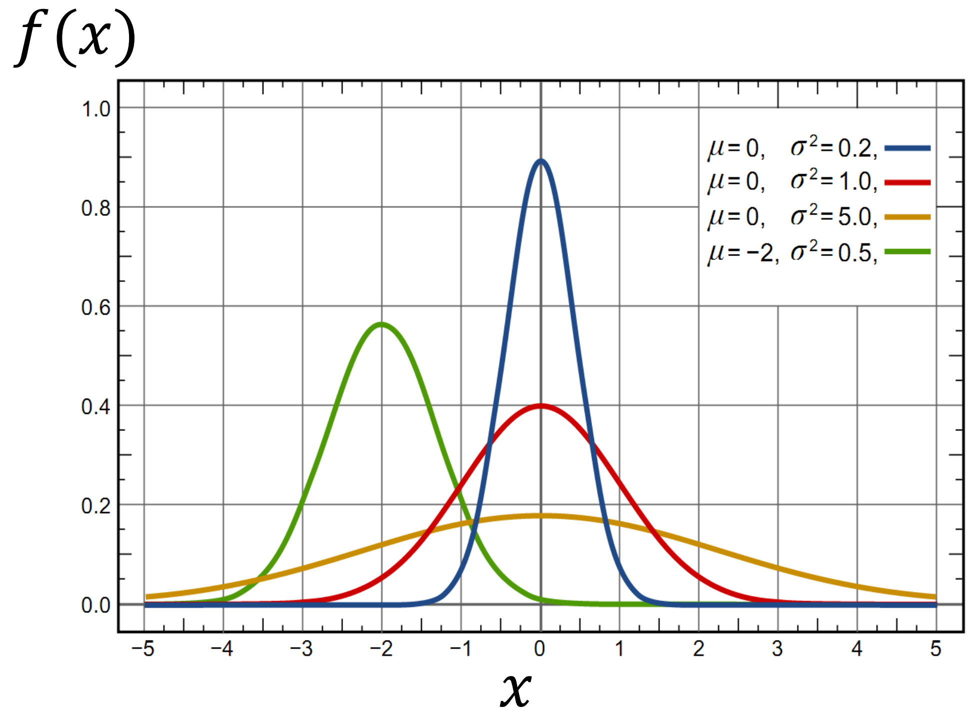
連続型確率変数の代表的な確率分布：確率密度関数で指定される

- 連続分布は確率密度関数$f(x)$で指定される
  - 確率 ＝ 確率密度の積分
    $[a,b]$内の値をとる確率：$P(a \leq X \leq b) = \int_a^b f(x)dx$
  - 連続変数がある特定の値をとる確率：$P(X = a) = 0$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
- 一様分布：閉区間$[a,b]$上の一様分布は

$$f(x) = \begin{cases} \dfrac{1}{b-a} & (a \leq x \leq b) \\ 0 & (\text{その他}) \end{cases}$$

連続型確率変数の代表的な確率分布：正規分布

- 正規分布：$f(x) = N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - パラメータ：平均μと分散σ^2

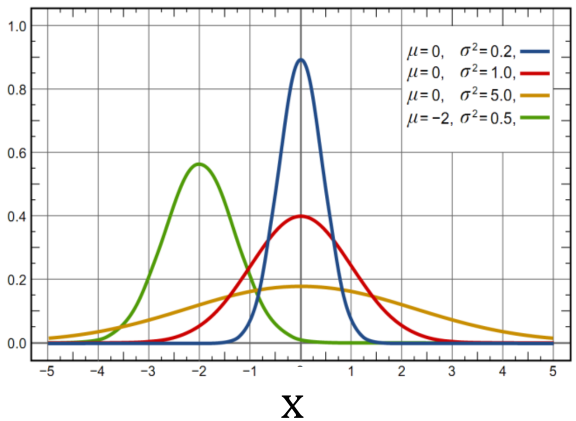


- 他、t分布、カイ2乗分布、ガンマ分布、ベータ分布、指数分布など

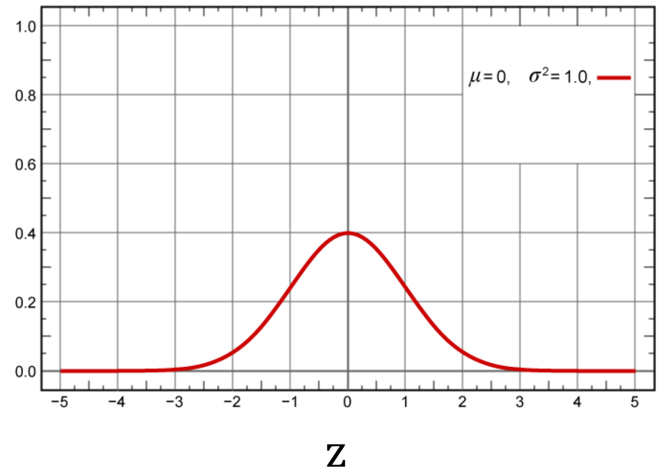
連続型確率変数の代表的な確率分布：標準正規分布

- $N(\mu, \sigma^2)$に従う確率変数Xを変数変換：$Z = \dfrac{X-\mu}{\sigma}$
- Zは平均0、標準偏差1の正規分布 $N(0,1)$に従う

確率密度関数: $f_X(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$ ➡ $f_Z(z) = \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{z^2}{2}\right)$



標準化 ➡



確率分布の特性値：
期待値は確率分布の代表値

- 確率変数Xの関数$g(X)$の期待値：確率での重みづけ平均

$$E[g(X)] = \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x & (\text{連続型確率変数}) \\ \displaystyle\sum_{x \in \mathcal{X}} g(x) f_X(x) & (\text{離散型確率変数}) \end{cases}$$

- さまざまな関数$g(X)$に対する期待値によって分布の特性を捉える
- 性質：
  - 線形性：$E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$
  - イェンセンの不等式：$E[g(X)] \geq g(E[X])$ (ただしgは凸関数)

さまざまな期待値：平均と分散

$g(X) = X$

- 平均 $\mu = E[X]$：Xの期待値　（分布の“真ん中”）

$g(X) = (X - \mu)^2$

- 分散 $\sigma^2 = \mathrm{Var}(X) = E[(X - \mu)^2]$：
  平均からの二乗偏差の期待値　（分布の“幅”）
  - $\mathrm{Var}(X) = E[X^2] - E[X]^2$
  - 標準偏差σ：分散の正の平方根
    - 正規分布なら$\mu \pm \sigma$：68%, $\pm 2\sigma$：95%, $\pm 3\sigma$：99.7%
- より一般的には（k次の）モーメント$E[X^k]$
  - 3次モーメント⇒歪度、4次モーメント⇒尖度 に関係する

- 例：厳密なサイコロ $P(X = i) = \frac{1}{6}$ の平均、分散を求めよ

平均の推定量：
標本平均

- 標本（部分）から平均（全体の性質）を知りたい
  - 標本 $S = \{x_1, s_2, \ldots, x_n\}$
- （母）平均はどのように推定できる？
- 標本平均：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ を平均 $\mu = E[X]$の推定値として使う？
  - 直感的には妥当そうだが、他にも候補は考えられるはず
    - x_nでもよいかもしれないし、適当に選んだ3つの値の中央値でもよいかもしれない…
    - 「よい」とか「よくない」は、どのように評価できるだろうか？

推定量としての標本平均の好ましさ：標本平均は不偏性と一致性をもつ

- 標本平均は平均の推定値として好ましいか？
- 不偏性 $E_S[\bar{X}] = \mu$：標本平均の期待値は母集団の平均に一致する
  - E_Sは標本についての期待値　（何度も標本をとり直して、何度も標本平均を求めたときの、それらの平均）
- 一致性：標本サイズが大きくなるほど母集団の平均μに近づく
  - 標本平均の分散 $\mathrm{Var}_S[\bar{X}] = \dfrac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$　（大数の法則）

    $(= E_S[(\bar{X}-\mu)^2])$

    σ^2は母分散

推定量としての標本平均の好ましさ：標本平均はBLUE（最良な線形不偏推定量）

- 効率性：推定値の分散が小さいこと
  - 標本平均$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$の代わりに最初の値を使う$\tilde{x} = x_1$とする
  - 標本平均のほうが「効率的」
    - 標本平均の分散$\frac{\sigma^2}{n}$ <最初の値の分散σ^2
- BLUE（最良な線形不偏推定量）：加重平均で表されるすべての不偏推定量のなかで、最も分散が小さい（効率的）なもの
  - 加重平均による推定量 $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} a_i x_i$

分散の推定量：
不偏分散

- 標本分散：$\frac{\left(x^{(1)}-\bar{x}\right)^2+\cdots+\left(x^{(n)}-\bar{x}\right)^2}{n}=\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 不偏性をもたない：$E_S\left[\frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}-\bar{X}\right)^2\right]=\frac{n-1}{n}\sigma^2$
- 不偏分散：$\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 不偏性をもつ：期待値が母集団の分散に一致する
- どちらも一致性はもつ：
  - 標本サイズが大きくなるほど母集団の分散に近づく
  - nが大きいところではnも$n-1$も大した違いはない

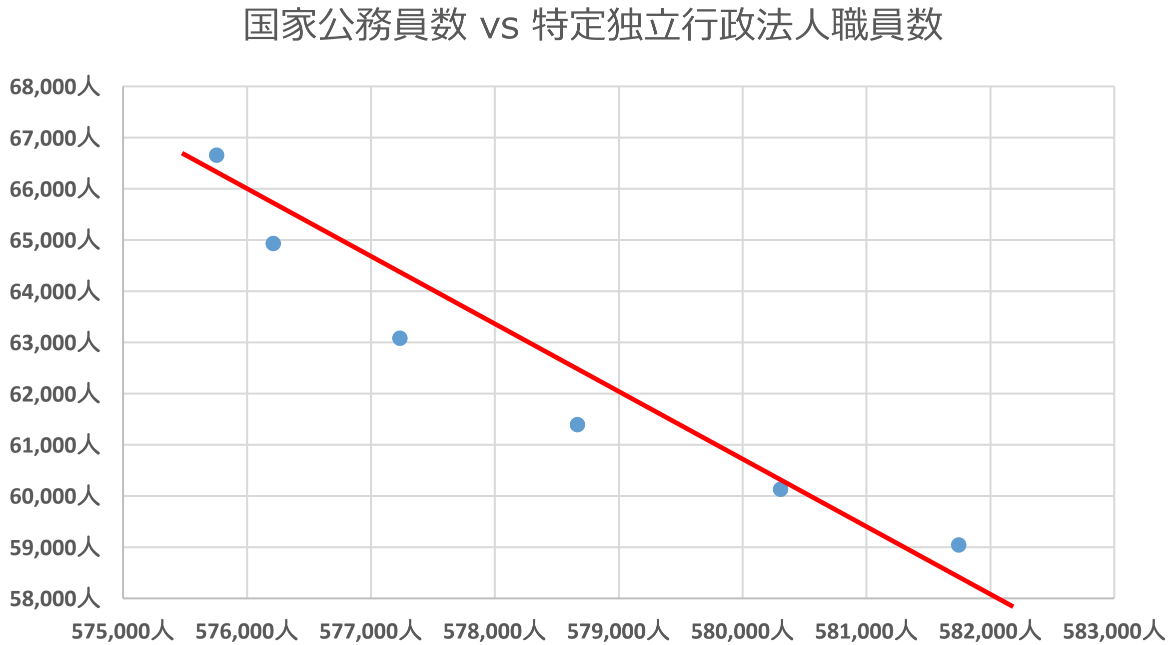
回帰

回帰：
片方の変数でもう片方の変数を説明

- 相関 (correlation) は二変数 x, y を区別せずに対等に扱う
  - 一方が増えたときに他方が増える (減る) 関係性を調べる
  - 例：身長と体重
- 回帰 (regression) は変数 x で変数 y を説明する
  - 一方から他方が決定される様子や程度を調べる
  - 例：年齢と血圧、所得と消費
  - xを独立変数・説明変数、yを従属変数・応答変数などとよぶ

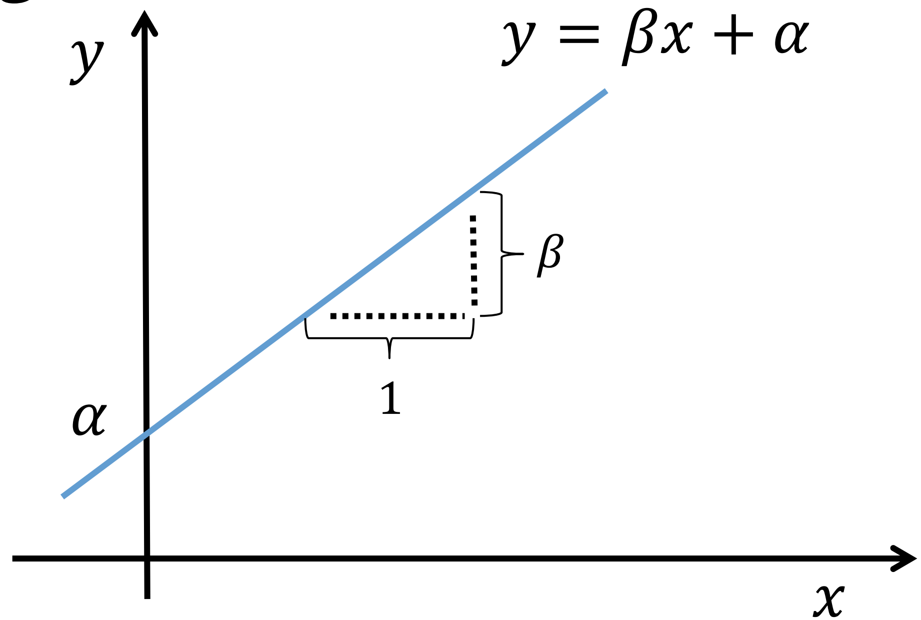
回帰の問題：
片方の変数からもう片方を説明するモデルをデータから推定

- 2つの変数 x と y の組についてN組のデータがある
  - $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ と $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right)$
- y を x で説明（予測）するモデルgがほしい
  - 概ね $y = g(x)$ となるg
  - 例えば直線をgとして仮定
- gの使い道：
  - 予測
  - 因果関係の発見（ただし注意が必要）



基本的な回帰モデル：
線形回帰モデル

- 線形モデル： $y = g(x) = \beta x + \alpha$
  - β ：傾きパラメータ（xが1増えると、yが1増える）
  - α ：切片パラメータ
- xとyの間に直線的な関係を仮定する
  - yがxの線形関数に依存



回帰モデルのパラメータ推定問題の定式化：モデルとデータの食い違いを最小化する最小二乗法

- データ：$\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(n)})$ と $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)$
- モデルの出力する予測値：$\hat{y}^{(i)} = \beta x^{(i)} + \alpha$
- モデルの予測と実際のデータとの食い違いを定義する：

$$\ell(\alpha, \beta) = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta x^{(i)} + \alpha\right)\right)^2$$

  - 食い違いを二乗誤差で測る

- 最適化問題（最小化）：

$$\left(\hat{\alpha}, \hat{\beta}\right) = \mathrm{argmin}_{\alpha,\beta}\, \ell(\alpha, \beta)$$

最小二乗法の解：二乗誤差を最小化する解

- $\ell(\alpha, \beta)$をαとβで偏微分して0とおいて、解くと：

$$\hat{\beta} = \frac{\sum_i (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_i (x^{(i)} - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \beta \bar{x}$$

- xとyの共分散：$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$
- xの不偏分散：$S_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x^{(i)} - \bar{x}\right)^2$

最小二乗法の性質：不偏性と推定精度

- いくつかの仮定の下で不偏性をもつ
  - 母集団において$\epsilon^{(i)} = y^{(i)} - \left(\beta^* + \alpha^* x^{(i)}\right)$が同一の分布に従い一定の分散$\sigma^2$、互いに無相関、$\epsilon_i$と$x_i$が無相関などの仮定
  - 不偏性：$\mathrm{E}\left[\hat{\beta}\right] = \beta^*, \mathrm{E}[\hat{\alpha}] = \alpha^*$ （標本の取り方についての期待値）

- $\mathrm{Var}\left[\hat{\beta}\right] = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^2}$：広範囲の$x^{(i)}$があったほうが精度がよい

- $\mathrm{Var}[\hat{\alpha}] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^2}\right)$：原点付近の$x^{(i)}$があったほうが精度がよい

決定係数：
従属変数をモデルがどの程度説明できたかを測る

- 決定係数 R^2 ：モデルの予測値 $\hat{\mathbf{y}} = \left(\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(n)}\right)$ とデータ $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)$との相関係数の2乗

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(\hat{y}^{(i)} - \bar{y}\right)\left(y^{(i)} - \bar{y}\right)\right)^2}{\left(\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{y})^2\right)\left(\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2\right)} = \frac{\sum_{i=1}^{n}\left(\hat{y}^{(i)} - \bar{y}\right)^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}$$

  - $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)$ の変動（分母）のうち、回帰式が説明できる変動（分子）の割合

- 相関係数は$-1 \leq R \leq 1$なので、決定係数 $0 \leq R^2 \leq 1$
  - 決定係数が1に近いほどデータへのモデルの当てはまりがよい

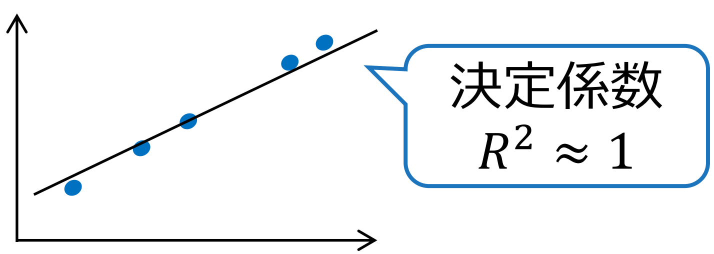
決定係数：
従属変数をモデルがどの程度説明できたかを測る

- **y** の変動の分解：

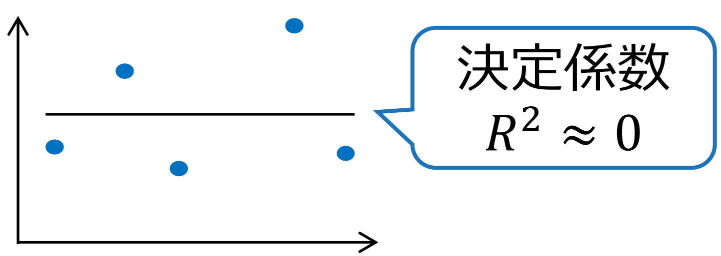
$$\sum_{i=1}^{n}\left(y^{(i)}-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}^{(i)}-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y^{(i)}-\hat{y}^{(i)}\right)^2$$

y の変動	回帰式の予測$\hat{y}^{(i)}$が説明できる変動	残差の平方和 $\sum_{i=1}^{n}{\epsilon^{(i)}}^2$

$\underbrace{\sum_{i=1}^{n}\left(\hat{y}^{(i)}-\bar{y}\right)^2}_{\text{回帰による説明}}\ \underbrace{\sum_{i=1}^{n}{\epsilon^{(i)}}^2}_{\text{回帰後に残るばらつき}}$



$\underbrace{\sum_{i=1}^{n}\left(\hat{y}^{(i)}-\bar{y}\right)^2}_{\text{回帰による説明}}\ \underbrace{\sum_{i=1}^{n}{\epsilon^{(i)}}^2}_{\text{回帰後に残るばらつき}}$



課題：
回帰モデリングを試してみよう！

- 自分でデータを見つけよう！
  - 従属変数と独立変数を決めよう！
  - データ：$\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ と $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right)$
- 回帰モデルを推定してみよう！：$\hat{y}^{(i)} = \beta x^{(i)} + \alpha$

$$\hat{\beta} = \frac{\sum_i (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_i \left(x^{(i)} - \bar{x}\right)^2}, \quad \hat{\alpha} = \bar{y} - \beta \bar{x}$$

- 決定係数を計算、データと回帰モデルをプロットしてみよう！
- 推定に使用しないデータに対しても、予測を評価してみよう

まとめ：
回帰モデリング

- 回帰では、（1個ないし複数の）独立変数から従属変数を説明・予測するモデルを作る
- 線形回帰モデル：独立変数が線形に効くモデル
- 最小二乗法によって回帰モデルのパラメータが求まる
- モデルの当てはまりは決定係数によって測る