

計量経済学

4. 回帰分析の基礎

矢内 勇生

経済・マネジメント学群



2019年10月17日

今日の内容



1 イントロダクション

- 線形回帰とは？
- 用語の説明

2 線形回帰の基礎

- 例：親子の身長の関係
- 回帰直線の求め方
- 説明変数が1つのモデル：単回帰
 - 説明変数が二値変数のとき
 - 説明変数が連続変数のとき

3 線形回帰モデル

- 最小二乗法

今日の目標



- 回帰分析の基本をおさえる
 - 線形回帰とは？
 - 最小二乗法 (OLS)

線形（線型）回帰とは？



線形回帰 (linear regression)

結果変数（応答変数）の平均値が、説明変数の線形関数で定義される値の変化に応じてどのように変化するかを要約する方法

説明変数の値に条件付けられた結果変数の期待値を求める

線形（線型）とは？



- 関数 $f(x)$ が線形（線型, linear）であるとは、以下の2つの性質を満たすこと
 - 加法性 $f(x+y) = f(x) + f(y), \quad \forall x, \forall y$
 - 齊次性 $f(kx) = kf(x), \quad \forall x, \forall k$
- $f(x)$ の変化の度合いが一定ということ
- 横軸を x 、縦軸を $f(x)$ とするグラフを作ると、直線になるということ

結果・応答変数と説明変数



- **結果変数・応答変数** (outcome variable, response variable) : 説明したい「結果」
 その他の呼び方：従属変数 (dependent v), 被説明変数 (explained v), 目的変数, regressand, etc.
- **説明変数** (explanatory v's) : 結果を変える要因 (原因)
 その他の呼び方：独立変数 (independent v), 予測変数 (predictor v), regressor, etc.
- 説明変数と応答変数の因果関係は回帰分析を行うための**仮定**：回帰分析では確かめられない
- 結果変数を説明変数に回帰する (regress **y on x**)

説明変数とコントロール変数



- 一般的な区別
 - 説明変数：結果を説明する主要な要因
 - コントロール [統制] 変数 (control v.)：説明変数以外で結果に影響を与える要因
- 統計学（数学）上の違い：なし
 → 説明変数とコントロール変数を区別する必要はない
- 因果推論における区別
 - 説明変数：原因と考えられる変数（**処置変数**; treatment v.f.）
 - コントロール：原因以外の変数（**交絡因子**; confounders）

ダミー変数



ダミー変数 (dummy v, indicator v)：ある属性を備えているかどうかを示す変数

- 女性ダミー：「女性」という属性を備えていれば 1, そうでなければ 0 をとる変数
- 男性ダミー：「男性」という属性を備えていれば 1, そうでなければ 0 をとる変数
- 女性が 1, 男性が 2 という値をとる変数は？

ダミー変数



ダミー変数 (dummy v, indicator v)：ある属性を備えているかどうかを示す変数

- 女性ダミー：「女性」という属性を備えていれば 1, そうでなければ 0 をとる変数
- 男性ダミー：「男性」という属性を備えていれば 1, そうでなければ 0 をとる変数
- 女性が 1, 男性が 2 という値をとる変数は？
ダミー変数とは呼ばないことが多い
 - × : 性別ダミー = 「性別」属性があるかどうか???
 - : 性別変数

単回帰と重回帰



- 単回帰 (simple regression) : 説明変数が 1 つの回帰
- 重回帰 (multiple regression) : 説明変数（コントロール変数を含む）が 2 つ以上の回帰
- 回帰 : 単回帰と重回帰を区別せずに呼ぶときを使う

親子の身長の関係



- 親の身長と子の身長の関係を調べたい：
どうする？（ヒント：2変数とも量的変数）

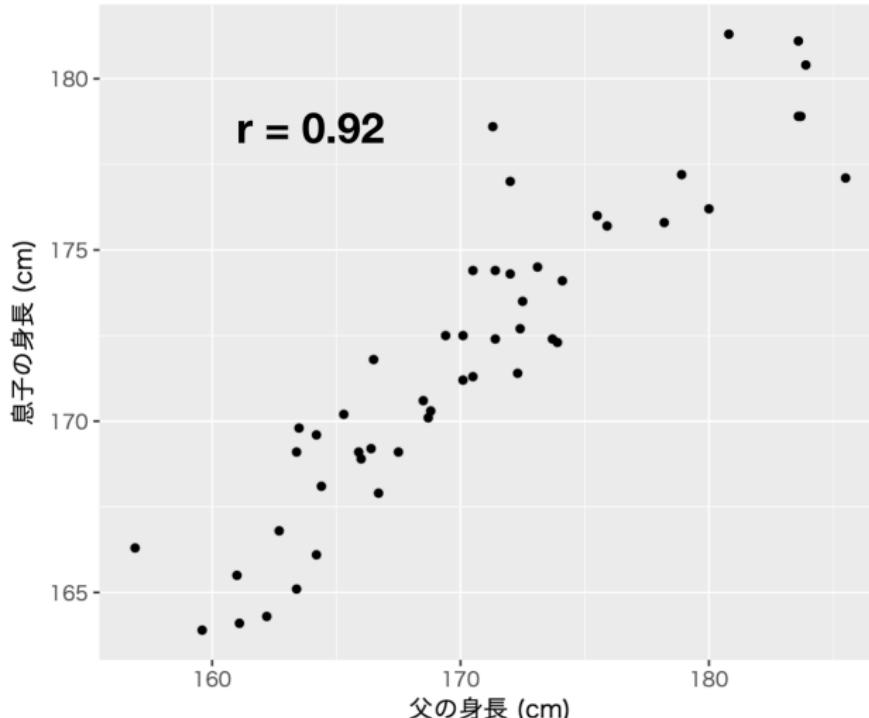
親子の身長の関係



- 親の身長と子の身長の関係を調べたい：
どうする？（ヒント：2変数とも量的変数）
- 図示する → 散布図
- 統計量を求める → 相関係数

例：親子の身長の関係

散布図と相関係数



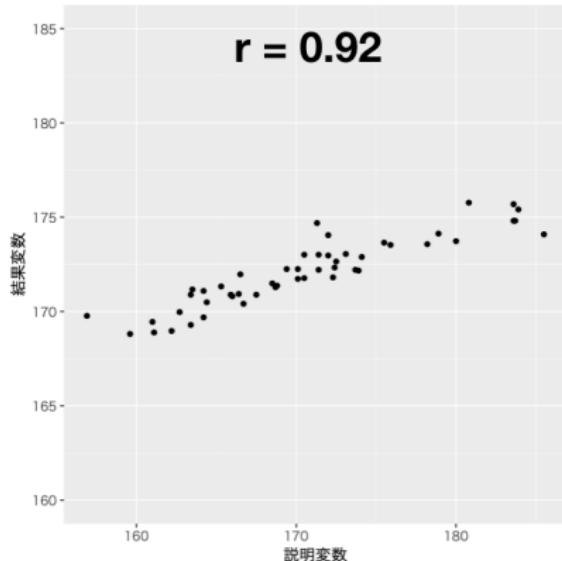
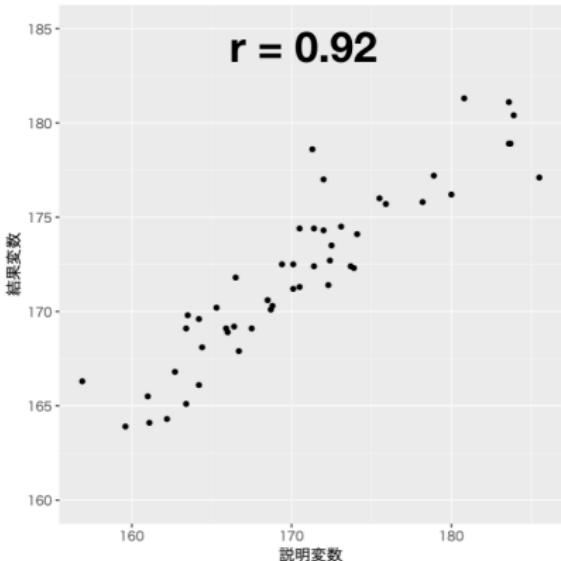
わかったことと新たな疑問



- 父親の身長が高いほど、子の身長が高い
- 新たな疑問
 - ① 父親の身長は息子の身長にどの程度影響するのか？
 - ② 父親の身長 x cm のとき、息子の身長は何 cm になりそうか？

例：親子の身長の関係

相関係数だけでは疑問に答えられない！



相関係数だけでは不十分な理由



- 相関係数が同じでも、関係の「傾き (slope)」は異なるかもしれない
 - 傾き：ある変数がもう一つの変数に与える（と想定される）影響の大きさ (**effect size**)
- 相関係数がわかつても、「**予測 (prediction)**」ができるない

直線を当てはめる

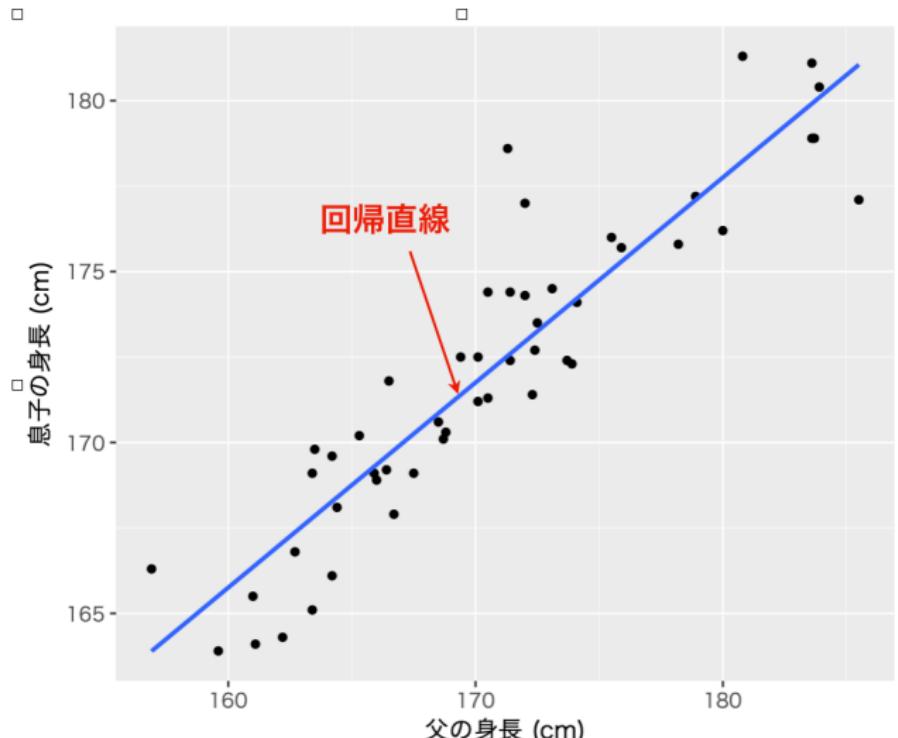


- 相関係数は、2変数の直線的な関係の強さを示す
→ 直線を引けばいいのでは？

- 直線：1次関数
→ x の値（父親の身長）から y の値（息子の身長）が予測できる！

例：親子の身長の関係

線形回帰：直線の当てはめ



回帰直線 (regression line)



- 結果（応答）変数と説明変数の関係を表す直線
 - 傾き（結果変数に対する説明変数の影響の大きさ）がわかる
 - 説明変数の値から結果変数の値を予測できる
- 結果変数の値が決まる原因を説明変数に帰する：「結果変数を説明変数に回帰する」
- 回帰分析には：
 - **1つの結果変数と、1つ以上の説明変数が必要**
 - **結果変数を縦軸、説明変数を横軸に**

直線



説明変数を x 、結果変数を y とすると、直線は 1 次関数

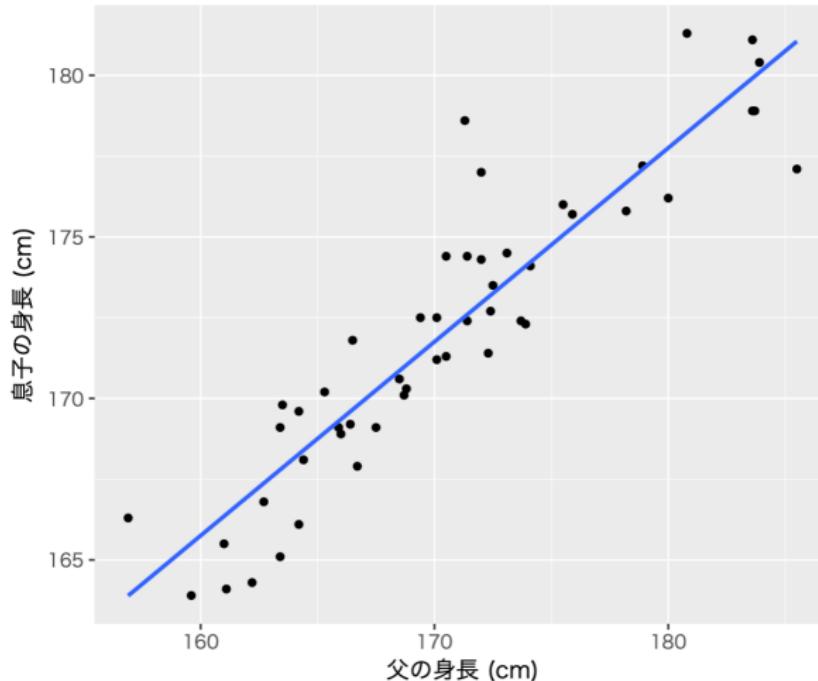
$$y = a + bx$$

で表すことができる

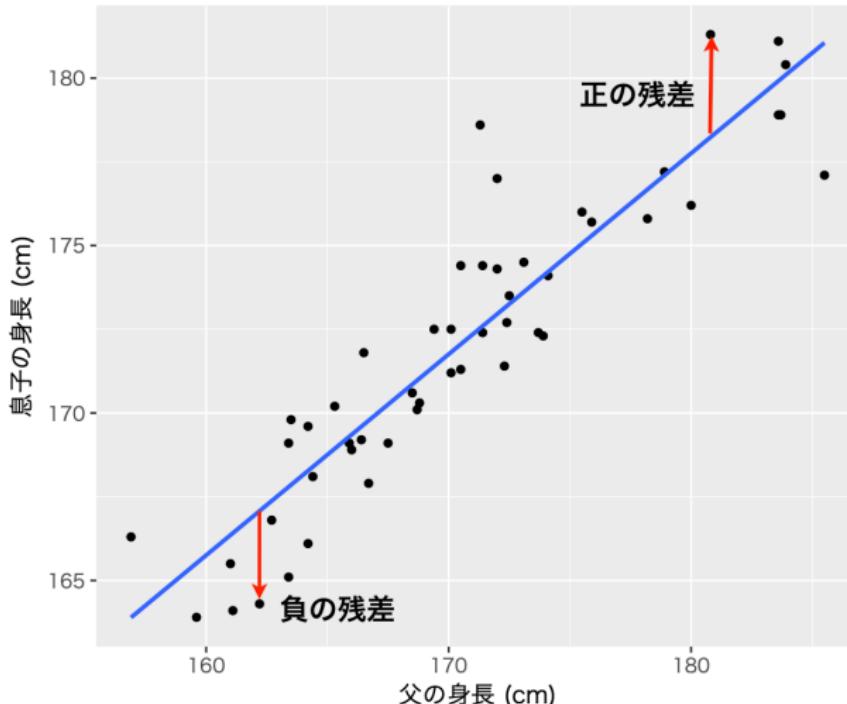
- a : y 切片 (x が 0 のときの y の値)
- b : 傾き (x が 1 単位増加したときの y の変化量)

直線を求める : a と b の値を求める

直線と点はズレる



残差 (residuals) (1)



残差 (2)



- 残差 : e
- 点（観測値, 実現値）を直線 ($a + bx$) と線からのズレに分ける

$$y_i = a + bx_i + e_i = \hat{y}_i + e_i$$

ただし、 $i = 1, 2, \dots, n$

- \hat{y}_i : 予測値 (fitted values, predicted values)
- 観測値 = 予測値 + 残差

ズレを小さくしたい



どうやってズレを小さくするか？

- 残差の平均値を小さくする？
 - プラスとマイナスが打ち消し合う：平均値のペアを通る直線は、いつも平均は 0
- 残差の二乗の総和（残差平方和）を小さくする：**最小二乗法**

最小二乗法 (Least Squares Method)



- 残差平方和を最小にすることで、散布図によく当てはまる（点とのズレが小さい）直線を求める方法
- 以下の式を最小にする a と b を求める

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

- 得られる結果：

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}}$$

- 回帰直線は、点 (\bar{x}, \bar{y}) を通る

結果の解釈



親子の身長の例: 息子の身長 = $69.8 + 0.6 \times$ 父の身長

- 父の身長が 1cm 高くなる毎に、息子の身長は平均すれば 0.6cm ずつ高くなる
- 父の身長が 0cm のとき、息子の身長は 69.8cm になると予測される
- 父の身長が x cm のとき、息子の身長は $(69.8 + 0.6x)$ cm になると予測される

モデル 1



衆議院議員総選挙での得票率を衆議院議員経験の有無で説明する

- 応答変数：得票率 (%)
- 説明変数：衆院議員経験がある（現職, 元職）候補者は 1, その他は 0
- 推定結果：

$$\text{得票率} = 14 + 31 \cdot \text{議員経験} + \text{誤差}$$

- 予測値 (predicted values) :

$$\widehat{\text{得票率}} = 14 + 31 \cdot \text{議員経験}$$

使用データ：浅野・矢内 (2018), hr-data.csv (以下、特にことわりのない限りこのデータを使う)

予測値と回帰係数



- 予測値：説明変数に具体的な数値が与えられたときの、応答変数の平均値（期待値）
- 予測値は $\hat{}$ （ハット）で表す
- モデル 1 の予測値：議員経験（0 または 1）が与えられたときの、得票率の平均値（期待値）

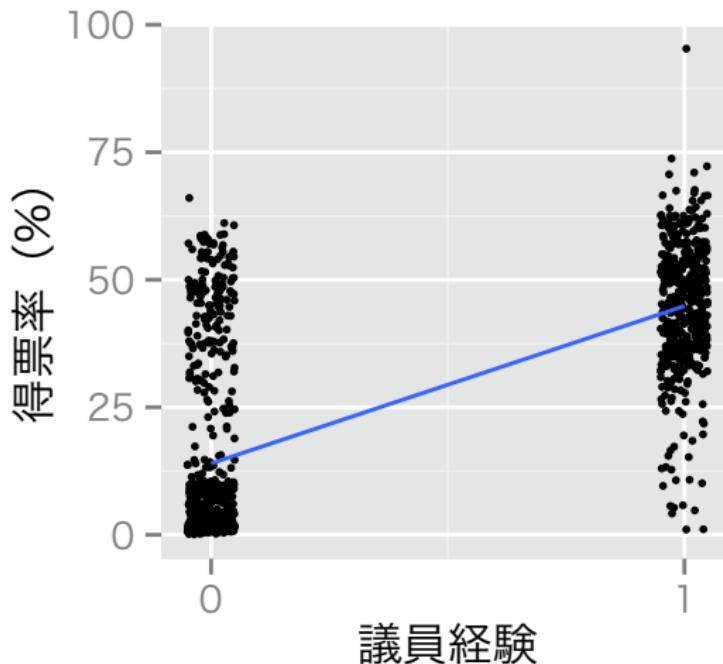
$$\widehat{\text{得票率}} = 14 + 31 \cdot \text{議員経験}$$

$$\widehat{\text{議員経験がない候補者の得票率}} = 14 + 31 \cdot 0 = 14$$

$$\widehat{\text{議員経験のある候補者の得票率}} = 14 + 31 \cdot 1 = 45$$

- 回帰係数： $31 = 45 - 14 = \text{議員経験がある候補者と議員経験がない候補者の平均得票率（予測値）の差}$

モデル 1 の図示：散布図と回帰直線



図：議員経験の有無で得票率を説明する

モデル 2



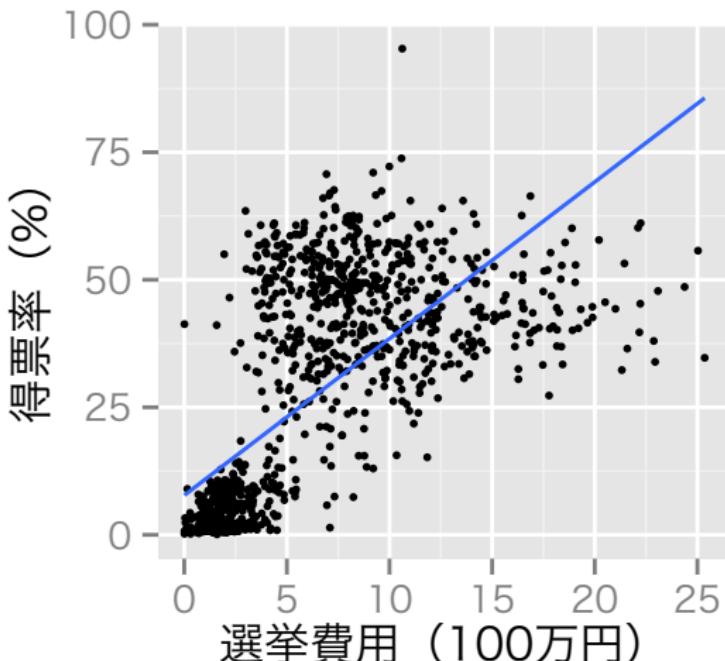
衆議院議員総選挙での得票率を選挙費用の大きさで説明する

- 応答変数：得票率 (%)
- 説明変数：選挙費用（測定単位：100 万円）
- 推定結果：

$$\text{得票率} = 7.7 + 3.1 \cdot \text{選挙費用} + \text{誤差}$$

- 回帰直線（次のスライド）上の点：
選挙費用ごとに予測される得票率：
候補者を選挙費用ごとにグループ分けしたときの、グ
ループの平均得票率

モデル 2 の図示：散布図と回帰直線



図：選挙費用で得票率を説明する

推定値の意味



$$\text{得票率} = 7.7 + 3.1 \cdot \text{選挙費用} + \text{誤差}$$

- 選挙費用の係数 3.1：選挙費用の値が 1 だけ異なる候補者を比べると、選挙費用が大きいほうが、**平均して 3.1 ポイント**高い得票率を得る
 - 選挙費用を 100 万円増やすと、得票率は 3.1 ポイント上がる
と**期待**される
 - 選挙費用を 1000 万円増やすと、得票率は 31 ポイント上がる
と**期待**される
- 切片 7.7：「選挙費用=0」の候補者の平均得票率
 - 選挙費用が 0 の候補者は存在しない！！！
 - 切片を「意味がある数字」にするには、変数変換が必要

ベクトルと行列による表現



- y_i : i 番目の個体の応答変数の値
- サンプルサイズは n : $i = 1, 2, \dots, n$
- 予測値 : $X_i\beta = \beta_1 X_{i1} + \dots + \beta_k X_{ik}$
- k : 定数項と説明変数の数の合計
- X : 説明変数の行列
- X_i : X の第 i 行
- 定数項 : $X_{i1} = 1$ for all i
- β : 係数ベクトル
- 誤差 : $\varepsilon_i \sim N(0, \sigma^2)$
- ベクトルはすべて列ベクトルとする
- 行ベクトルは、列ベクトルの転置として表す：列ベクトル a の転置 a' (a プライム) が行ベクトル

回帰モデルの表現法



線形回帰モデルを式で表す

- 表現法 1

$$\begin{aligned}y_i &= X_i \beta + \varepsilon_i \\&= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \\\varepsilon_i &\sim N(0, \sigma^2) \text{ for } i = 1, 2, \dots, n\end{aligned}$$

- 表現法 2

$$y_i \sim N(X_i \beta, \sigma^2), \text{ for } i = 1, 2, \dots, n$$

or

$$y \sim N(X \beta, \sigma^2 I)$$

これらのモデルに最小二乗法を適用し、 $\hat{\beta}$ と $\hat{\sigma}$ を得る

R で線形回帰モデルに最小二乗法を当てはめる



- `lm()` 関数を使う
- 推定結果を確認するには
 - ① `summarize()` を使う
 - ② `arm::display()` を使う
 - ③ `broom::tidy()` を使う
- 詳しくは[ウェブ](#)で

回帰係数ベクトル β の最小二乗推定量 (1)



- 線形回帰モデル : $y = X\beta + \varepsilon$, $\varepsilon_i \sim N(0, \sigma^2)$
- 誤差を最小にする β を見つけたい : 誤差の平方和

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - X_i \beta)^2$$

を最小にする β を見つけたい

- β は母数であり、観測できないので、代わりに残差の平方和

$$\sum_{i=1}^n e_i^2 = e'e = (y - X\hat{\beta})'(y - X\hat{\beta})$$

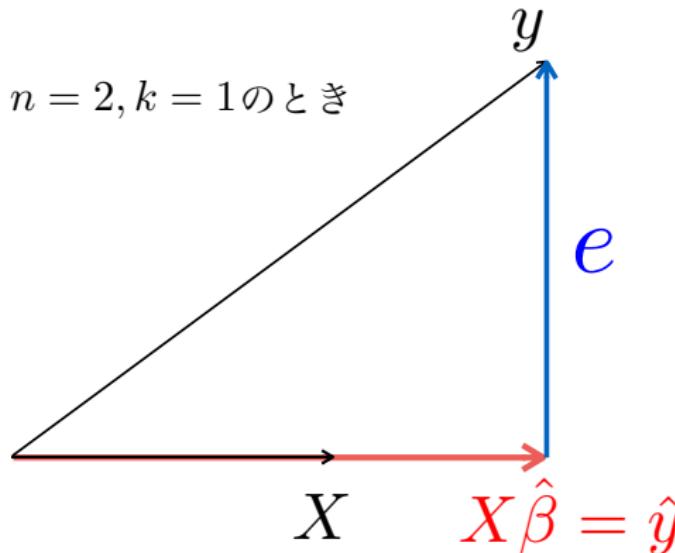
を最小にする $\hat{\beta}$ を見つける

→ y と $X\hat{\beta}$ の距離を最小化する $\hat{\beta}$ を見つければよい

回帰係数ベクトル β の最小二乗推定量 (2)



- 残差 : $e = y - X\hat{\beta} = y - \hat{y}$
- y と $X\hat{\beta}$ の距離 : X と e が直行するときに最小



回帰係数ベクトル β の最小二乗推定量 (3)



- 直行条件 : $X'_m e = 0$ for all m , $m = 1, 2, \dots, k$
よって、

$$\begin{aligned}
 X'e = \mathbf{0} &\iff X'(y - X\hat{\beta}) = \mathbf{0} \\
 &\iff X'X\hat{\beta} = X'y \quad \text{正規方程式} \\
 &\iff (X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \\
 &\iff I\hat{\beta} = (X'X)^{-1}X'y \\
 &\iff \hat{\beta} = (X'X)^{-1}X'y
 \end{aligned}$$

- $(X'X)^{-1}$ が存在すれば、これが最小二乗推定量である
- β の最小二乗推定量 : y の線形関数である！