高知工科大学 経済・マネジメント学群

計量経済学

4. 回帰分析の基礎

矢内 勇生（やない ゆうき）

https://yukiyanai.github.io

yanai.yuki@kochi-tech.ac.jp

今日の目標

- 回帰分析の基本的な意味を理解する

- 線形回帰とは何か

- 最小二乗法による推定

線形回帰とは

線形（線型）回帰

- 線形回帰 (linear regression)
  - 応答変数（結果変数）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づけられた応答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：$f(x+y) = f(x) + f(y), \quad \forall x, \forall y$
  - 斉次性：$f(kx) = kf(x), \quad \forall x, \forall k$
- 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

応答変数（結果変数）と説明変数

- **応答変数** (response variable)・**結果変数** (outcome variable)：説明したい結果
  - その他の呼び方：従属変数, 被説明変数, 目的変数, regressand, etc.
- **説明変数** (explanatory variables[s])：結果に影響を与える要因
  - その他の呼び方：独立変数, 予測変数, regressor, etc.
- 説明変数と応答変数の間の因果関係は、回帰分析を行う際の**仮定**
  - 因果関係があるとは限らない
  - 回帰分析では確認できない
- 応答変数**を**説明変数**に**回帰する (regress y on x)

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)：説明変数が2つ以上の回帰
- 単に「回帰」という場合は、単回帰と重回帰の両者を指す

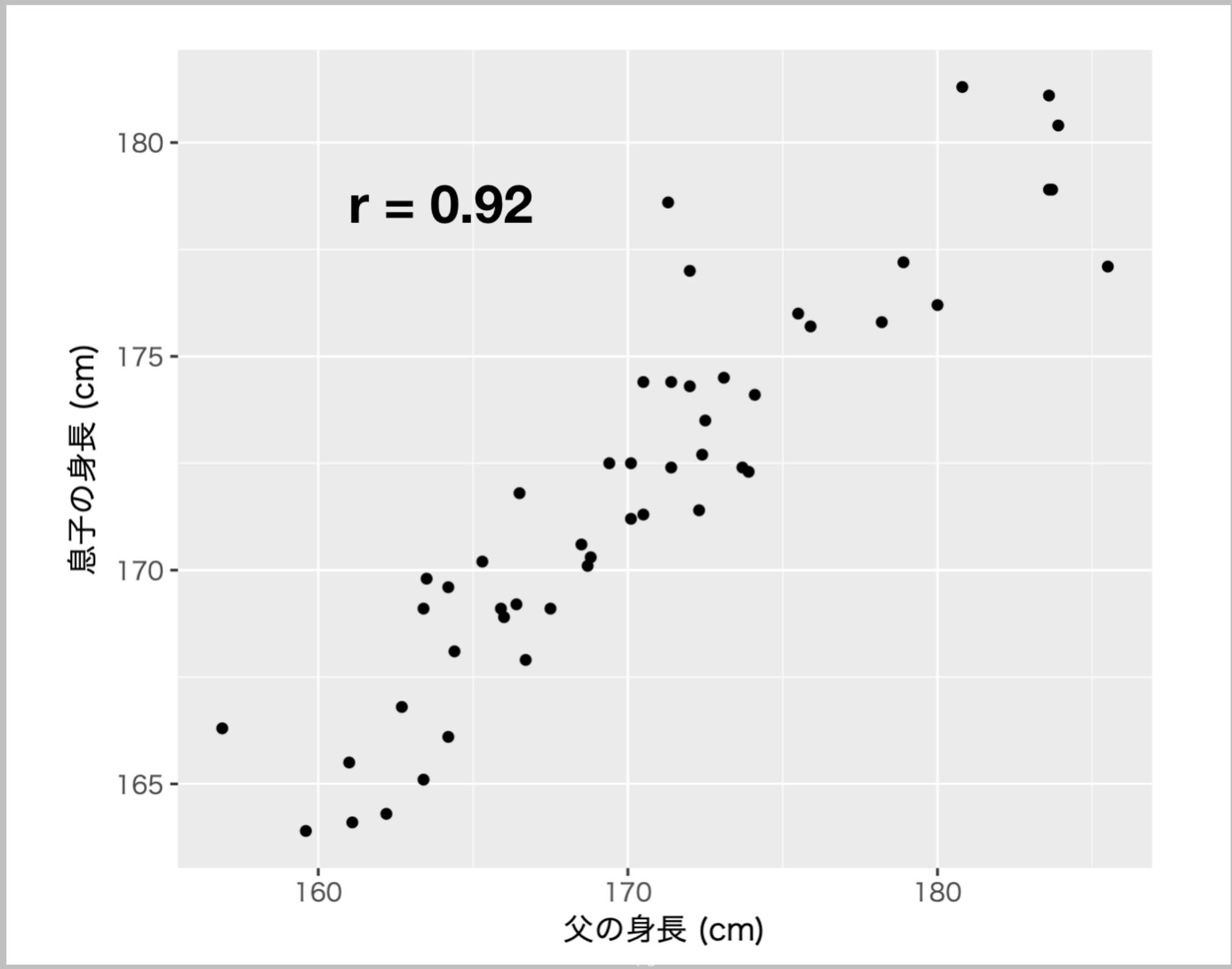
線形回帰の基礎

例：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つとも量的変数）

- 図示する：**散布図**
- 統計量を求める：**相関係数**

散布図と相関係数

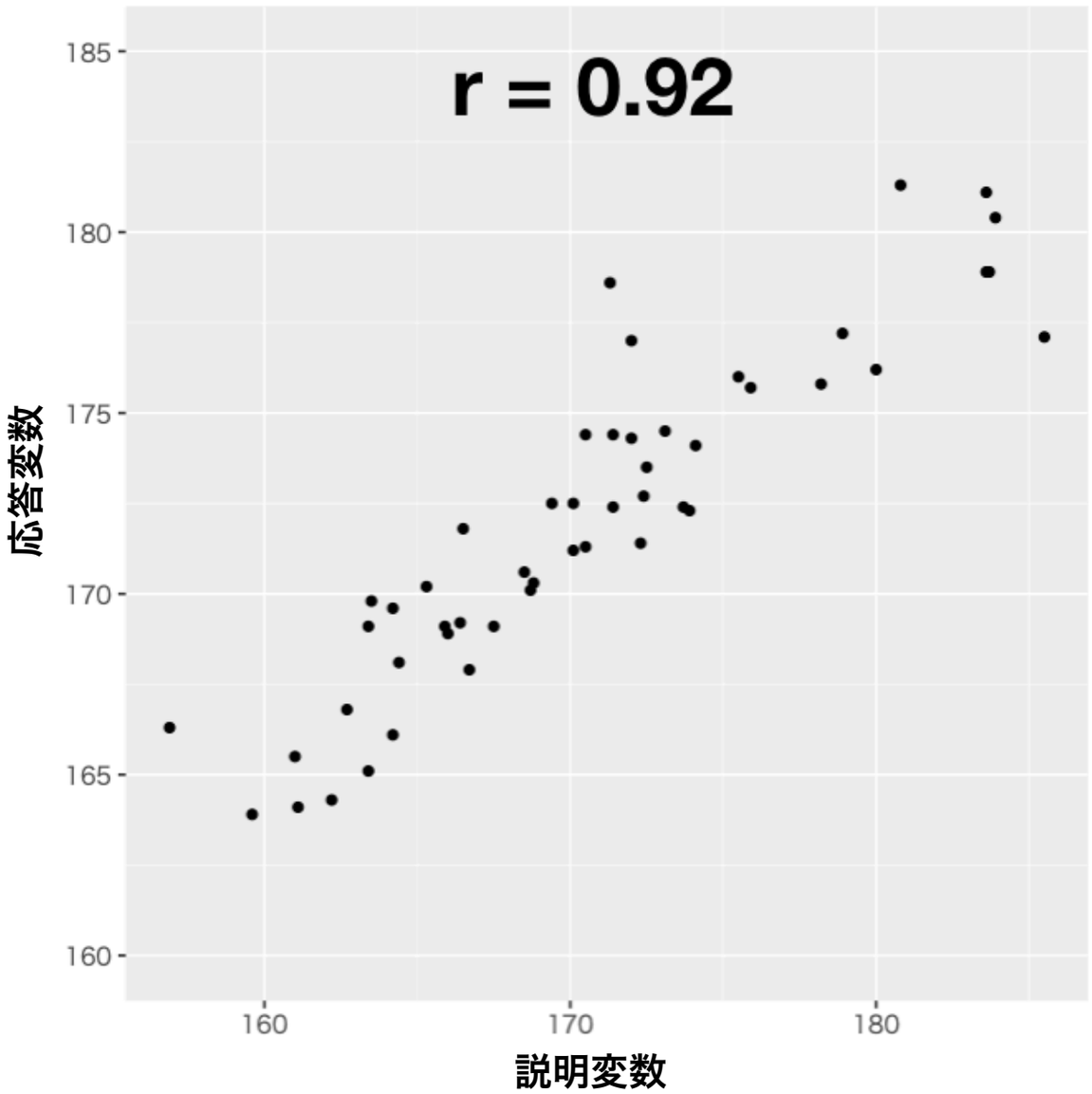


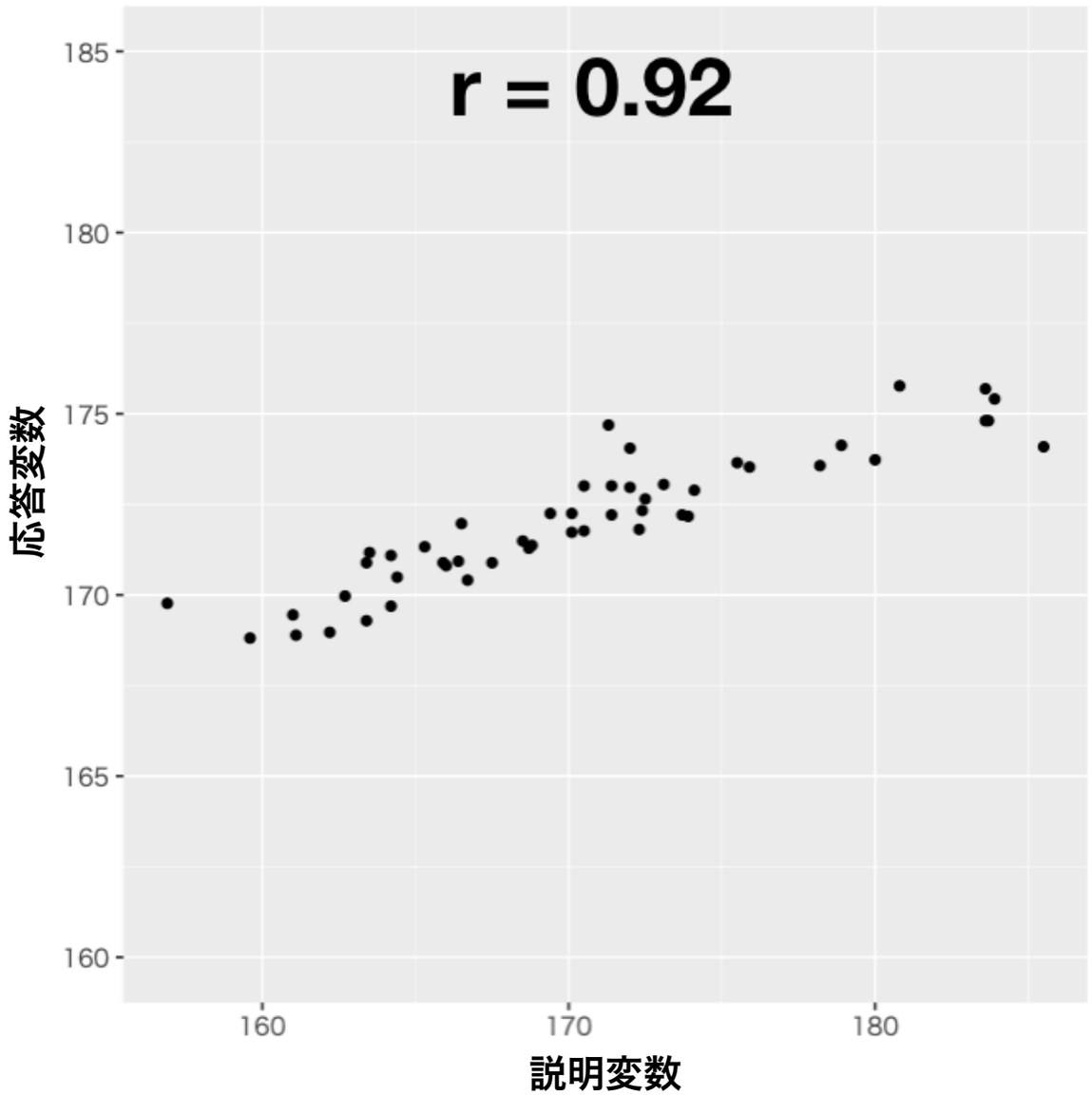
わかったことと新たな疑問

- 父親の身長が高いほど、子の身長が高い
- 新たな疑問

- 父親の身長は、息子の身長にどの程度影響するのか？
- 父親の身長が x cm のとき、息子の身長は何cm になりそうか？

相関係数だけでは、疑問に答えられない





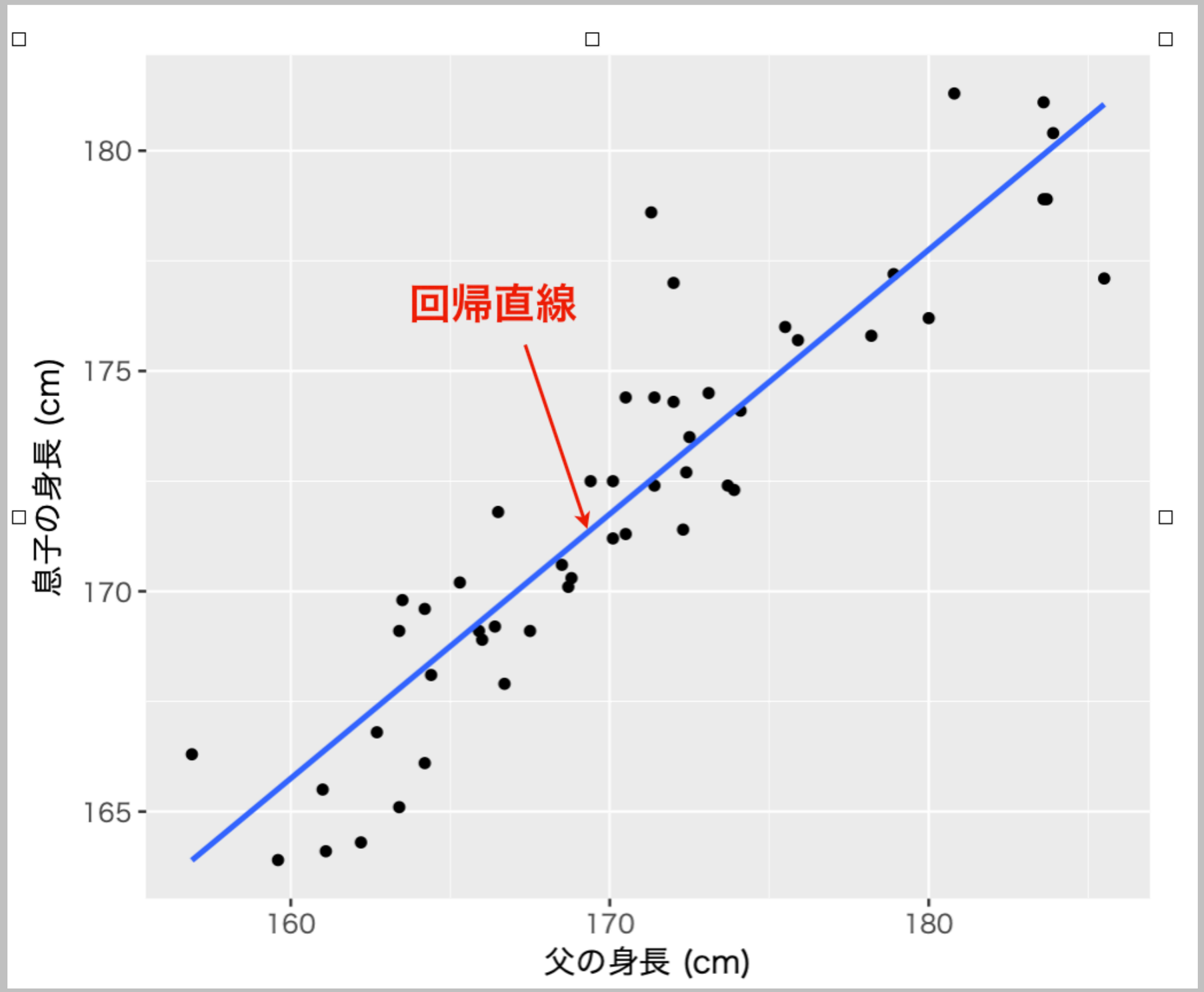
相関係数だけでは不十分

- 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：説明変数が応答変数に与える（と想定される）**影響の大きさ**
- 相関係数が判っても「**予測 (prediction)**」ができない

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!

線形回帰：直線の当てはめ



回帰直線 (regression line)

- 応答変数と説明変数の関係を表す直線
  - 傾き（応答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
- 回帰分析には：
  - 1つの応答変数と、1つ以上の説明変数が必要
  - 応答変数を縦軸に、説明変数を横軸にとる

直線

- 説明変数を x , 応答変数をy とすると、直線は1次関数

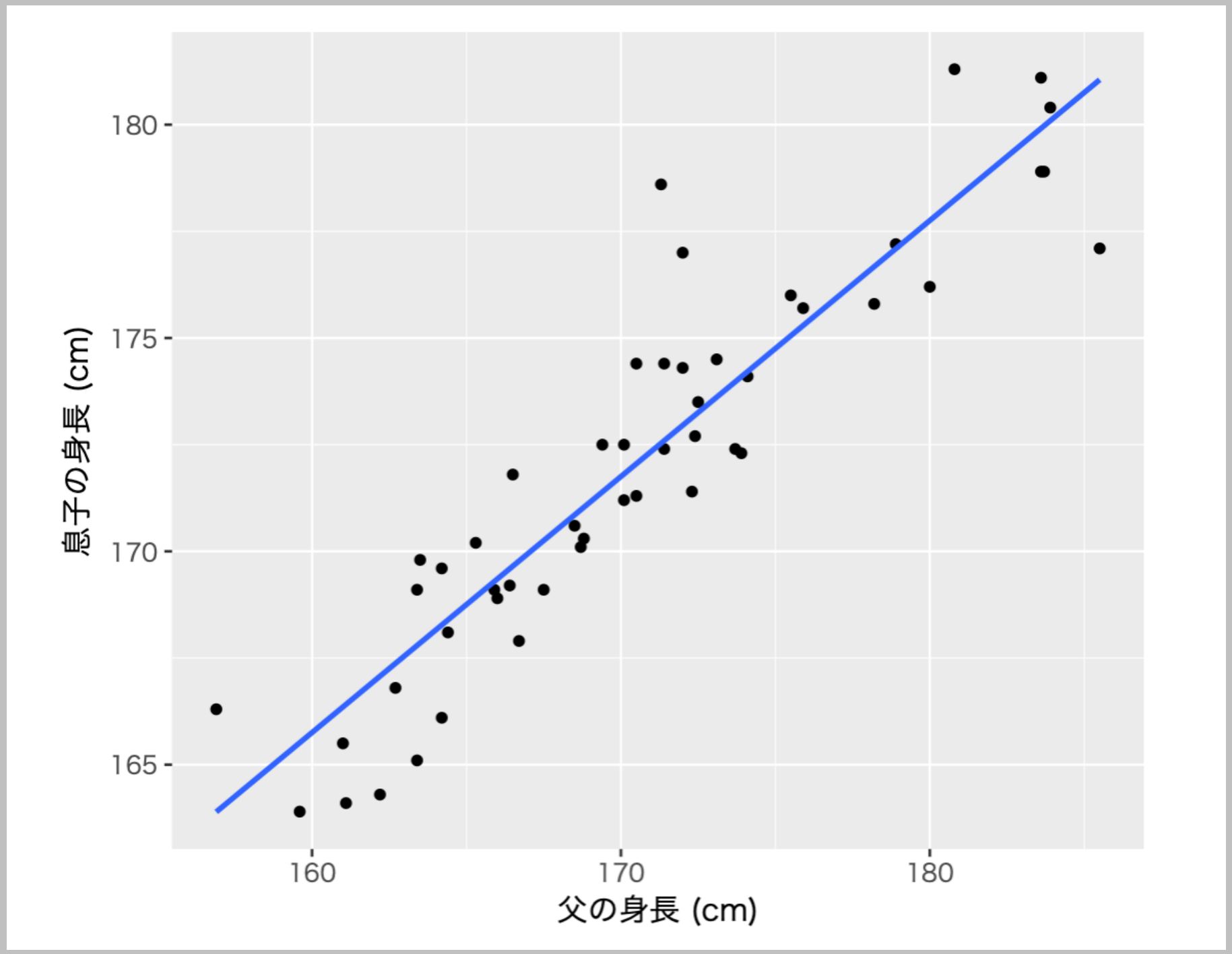
$$y = a + bx$$

で表すことができる

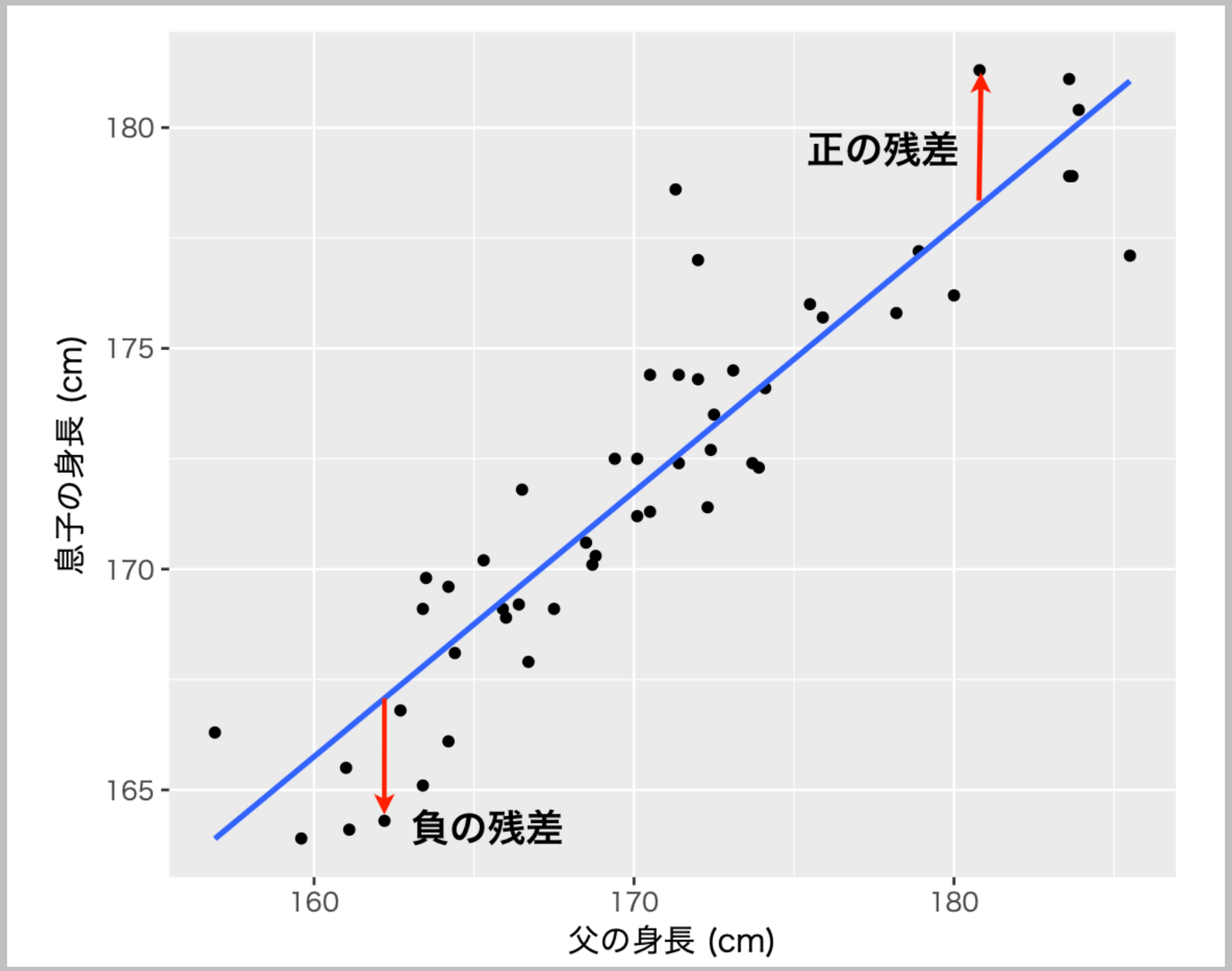
- a：y 切片（x が0のときの y の値）
- b：傾き（x が1単位増加したときの y の変化量）

- 回帰直線を求める：a と b の値を求める

直線と点はズレる



残差 (residuals)



残差とは？

- 残差：e

- 散布図上の点（観測値, 実現値）を直線（$a + bx$）とその線からのズレに分解する

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

ただし、$i = 1,2,\ldots,N$

- $\hat{y}_i$：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：平均値のペアになる座標 $(\bar{x}, \bar{y})$ を通る直線なら、残差の平均は必ず0
  - 残差の二乗の総和（残差平方和）を最小化する：**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（観測値とのズレが小さい）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：

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

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \frac{\sum x_i y_i - N\bar{x}\bar{y}}{\sum x_i^2 - N\bar{x}^2}$$

★ 回帰直線は、点 $(\bar{x}, \bar{y})$ を通る

回帰直線の意味

- 親子の身長の例

$$\text{息子の身長 (cm)} = 69.8 + 0.6 \times \text{父の身長 (cm)}$$

  - 父の身長が 1cm 高くなるごとに、息子の身長は**平均すれば** 0.6 cm ずつ高くなる
  - 父の身長が 0 cm のとき、息子の身長は 69.8 cm になる
  - 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

線形回帰の例

単回帰モデル

モデル1：説明変数がダミー変数の場合

- 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のダミー変数
  - 推定結果：

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

  - 予測値 (fitted values, 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) の `hr-data.csv`

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）

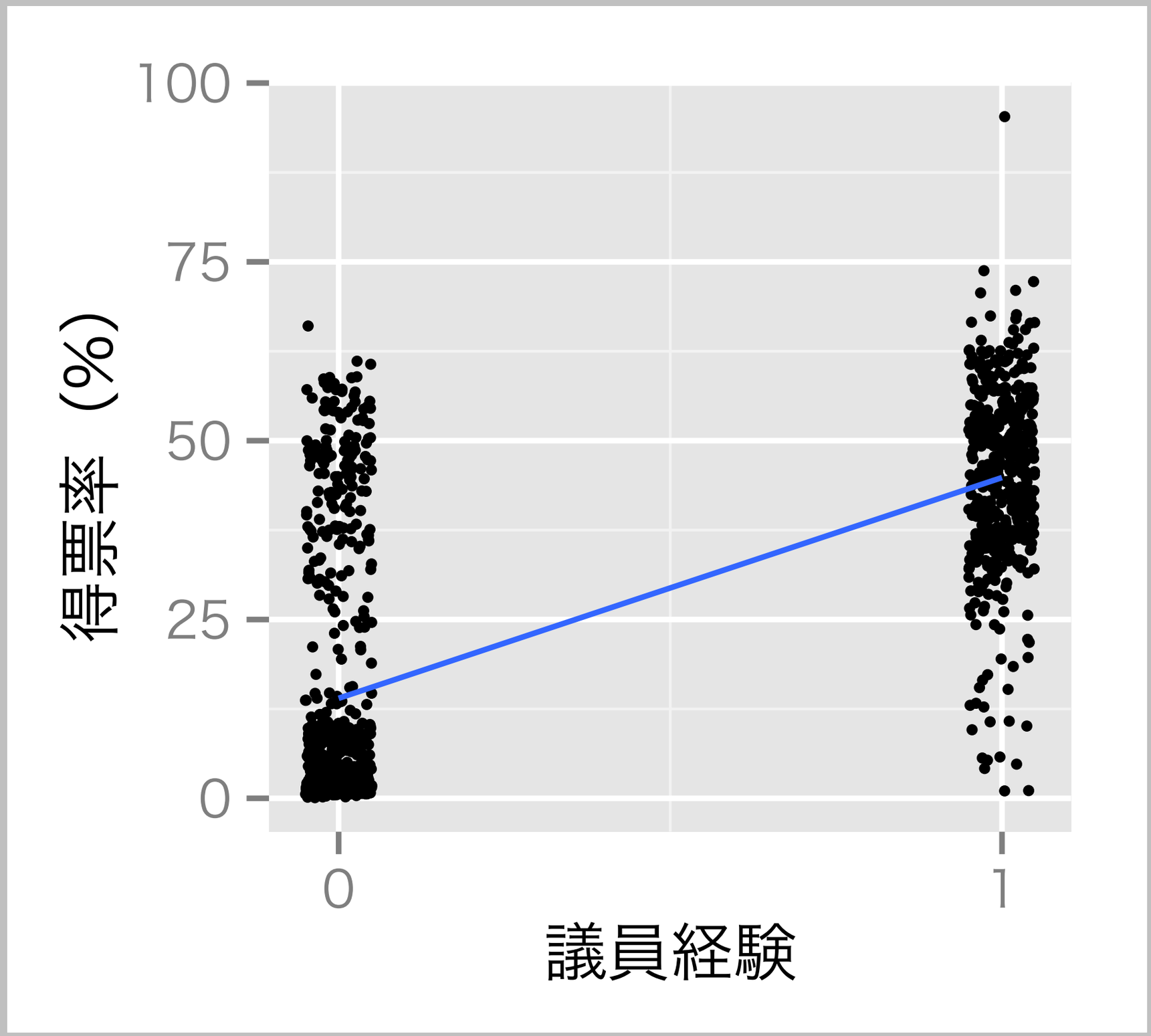
$$\widehat{\text{得票率}} = 14 + 31 \cdot \text{議員経験}$$

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

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

- 回帰係数：$31 = 45 - 14$ = 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：散布部と回帰直線



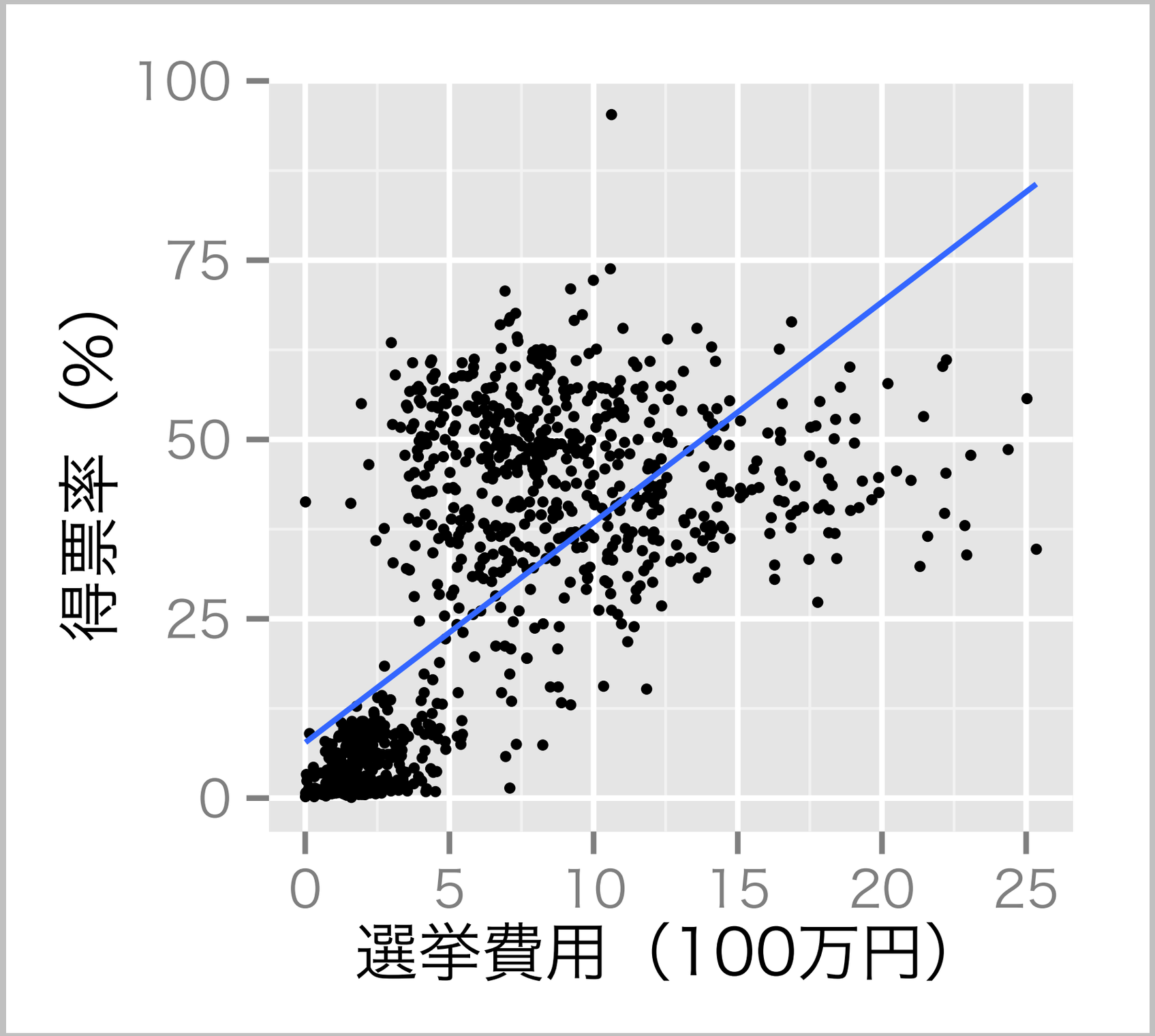
モデル2：説明変数が量的変数の場合

- 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：

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

  - 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



推定値の意味

- 得票率 $= 7.7 + 3.1 \cdot$ 選挙費用 $+$ 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、**平均すれば** 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1,000万円増やすと、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！
    - 切片は「意味がない」？？？（後で解決する）

Rで回帰直線を求める

- `lm()` 関数を使う
- 推定結果を確認するには
  - `summarize()` を使う
  - `broom::tidy()` を使う
    - 詳しくは web の実習資料で

次回

因果推論 I