統計学 2

14. 回帰分析入門

矢内　勇生

2018年5月31日

高知工科大学　経済・マネジメント学群

今日の目標

- 2変数の関係を「直線」として要約する方法を理解する
  - 1つの変数がもう1つの変数に与える「影響の大きさ」を推定する
  - 1つの変数の値から、もう1つの変数の値を「予測」する

例題

- 父親の身長と息子の身長の間にはどんな関係がある？

原因と結果の関係？

- 原因：父親の身長
  - 説明変数 (explanatory variable)
  - 他の呼び名：独立変数、予測変数、入力、特徴量
- 結果：子の身長
  - 結果変数 (outcome variable)
  - 他の呼び名：従属変数、応答変数、出力

親子の身長の関係

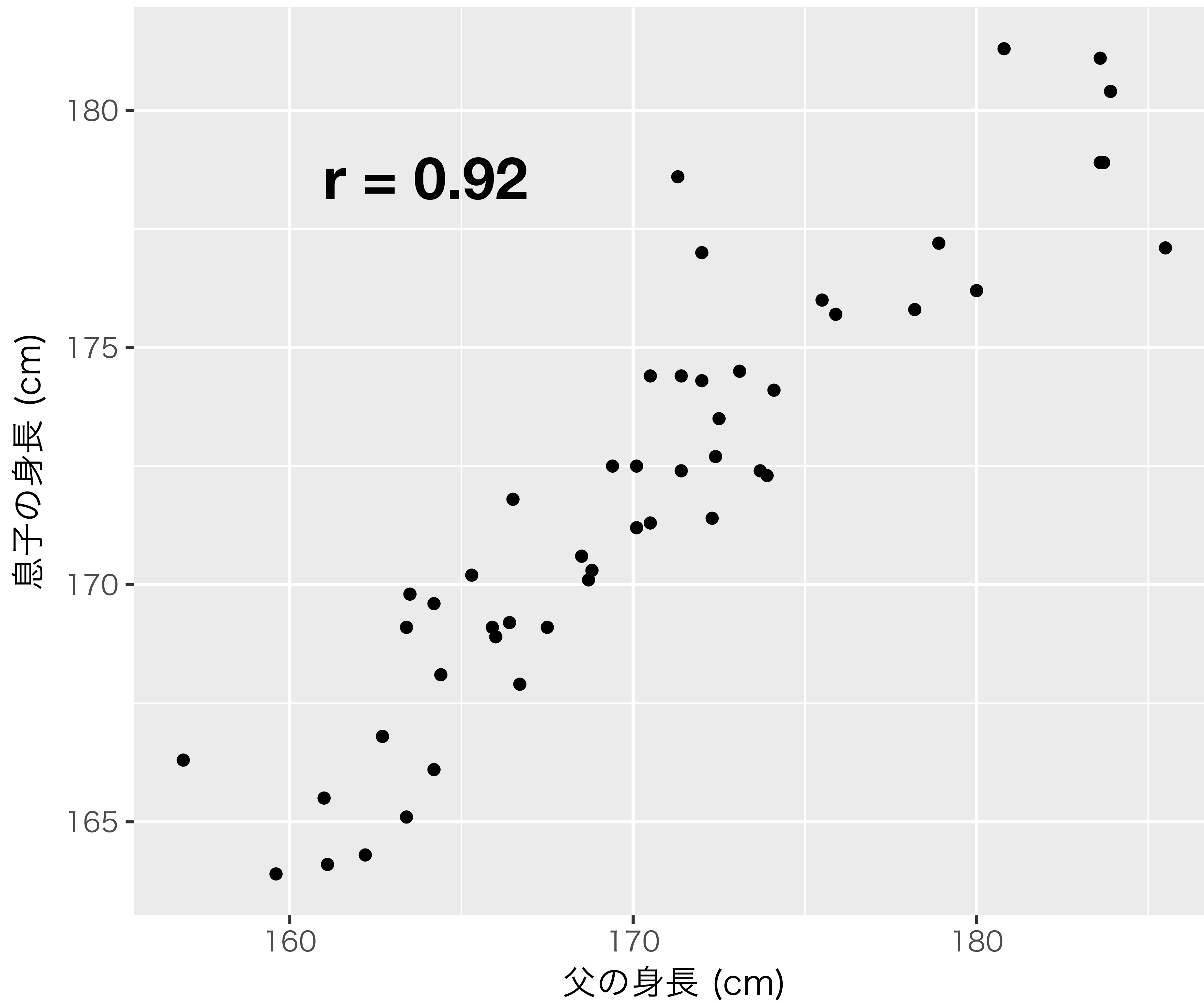
- 2変数の関係を調べたい

どうする？

（ヒント：2変数とも数量変数）

- 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、息子の身長が高い

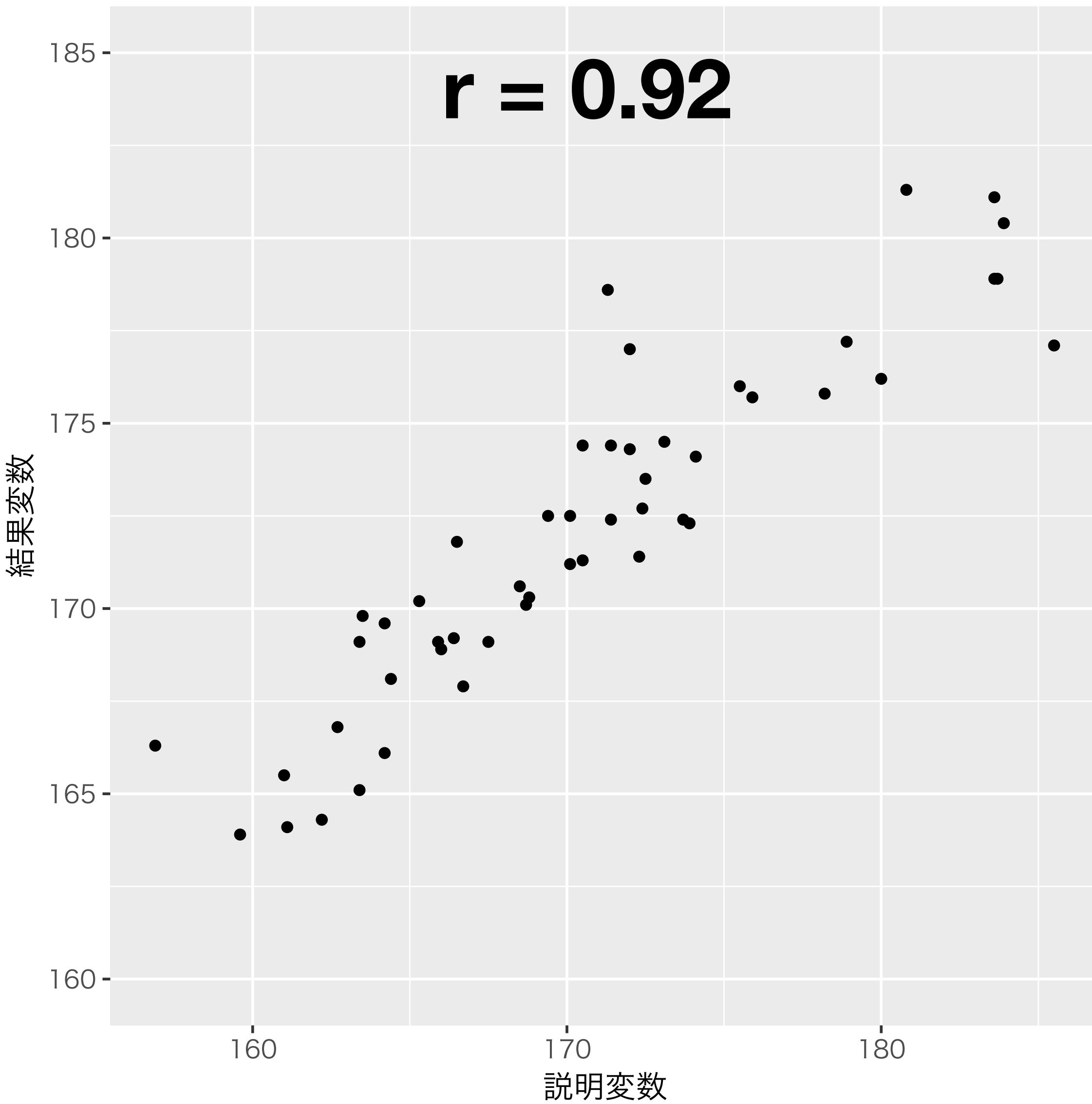
Q1：父親の身長は息子の身長にどの程度影響するの？

- 影響の大きさを知りたい！

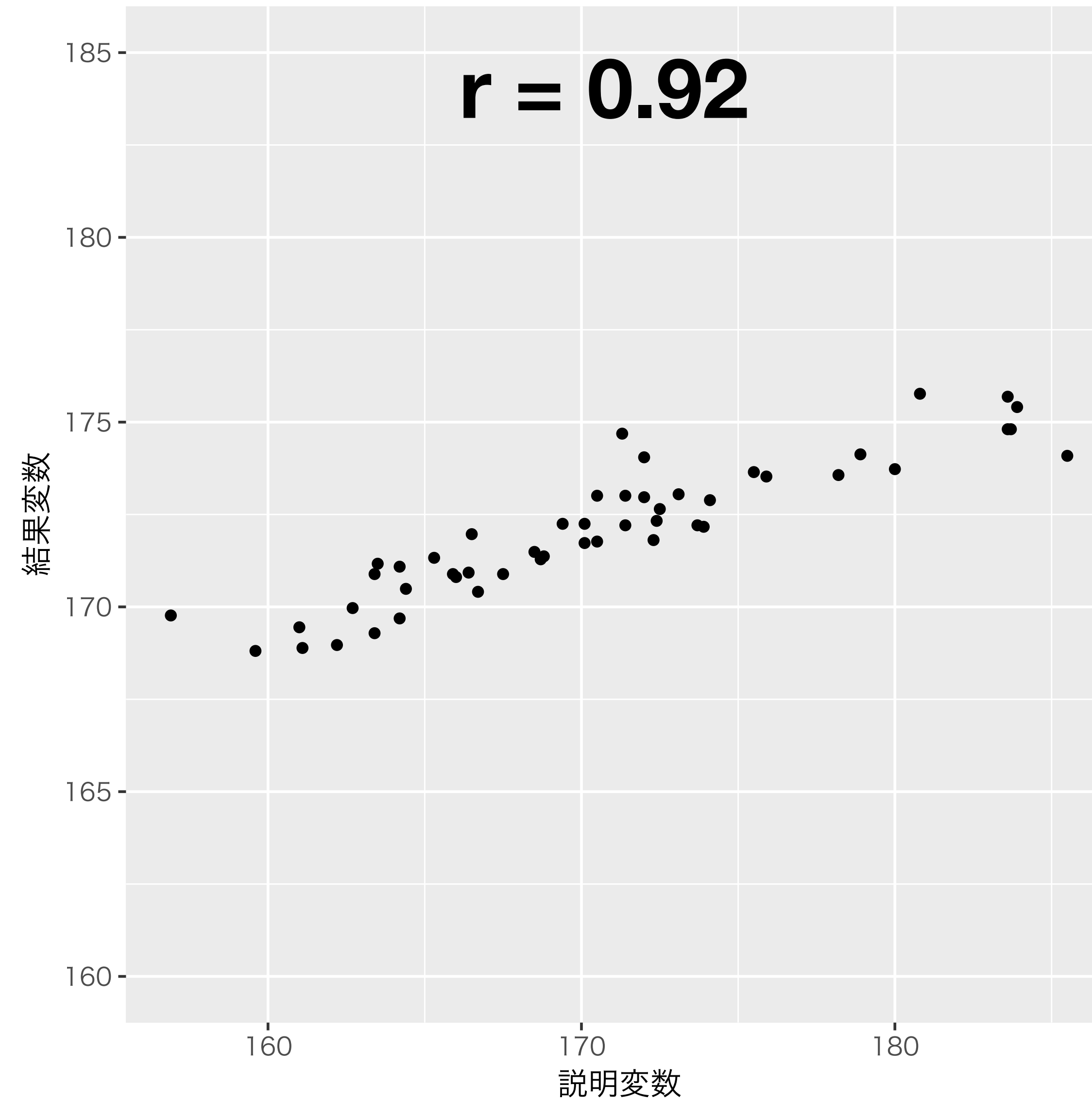
Q2：父親の身長が x cm のとき、息子の身長は何cmになりそう？

- 原因から結果を予測したい！

相関係数だけでは新たな疑問に答えられない！



説明変数の影響が大きい



説明変数の影響が小さい

相関係数だけでは不十分な理由

- 相関係数が似ていても「傾き（ある変数がもう１つの変数に与える影響の大きさ）」が異なる
- 相関係数がわかっても、「予測」ができない

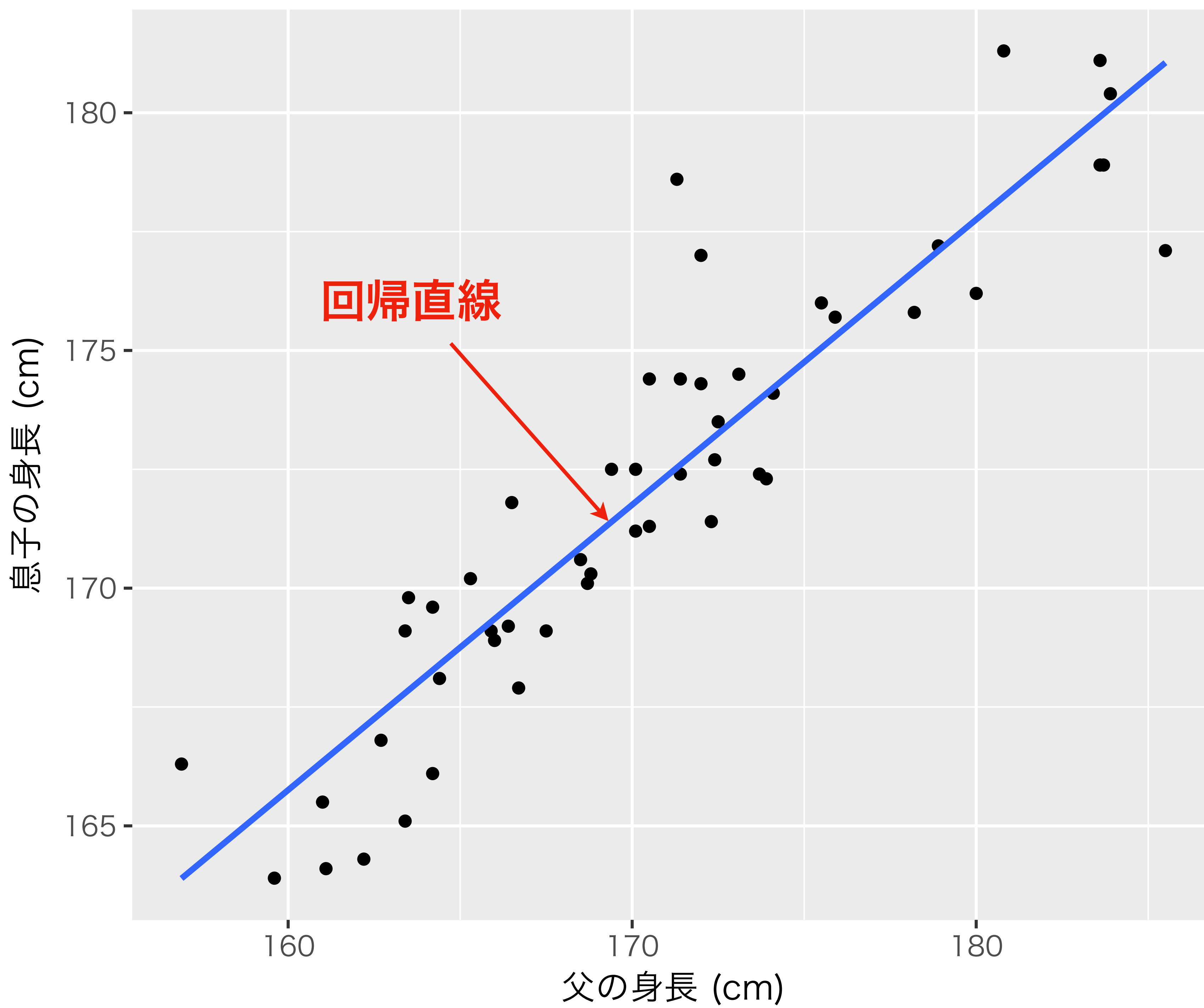
直線を当てはめる

- 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！

線形回帰分析：直線を当てはめる



回帰直線 (regression line)

- 結果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
  - 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」

▶ 回帰分析を行うときは、

- 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

$$y = a + bx$$

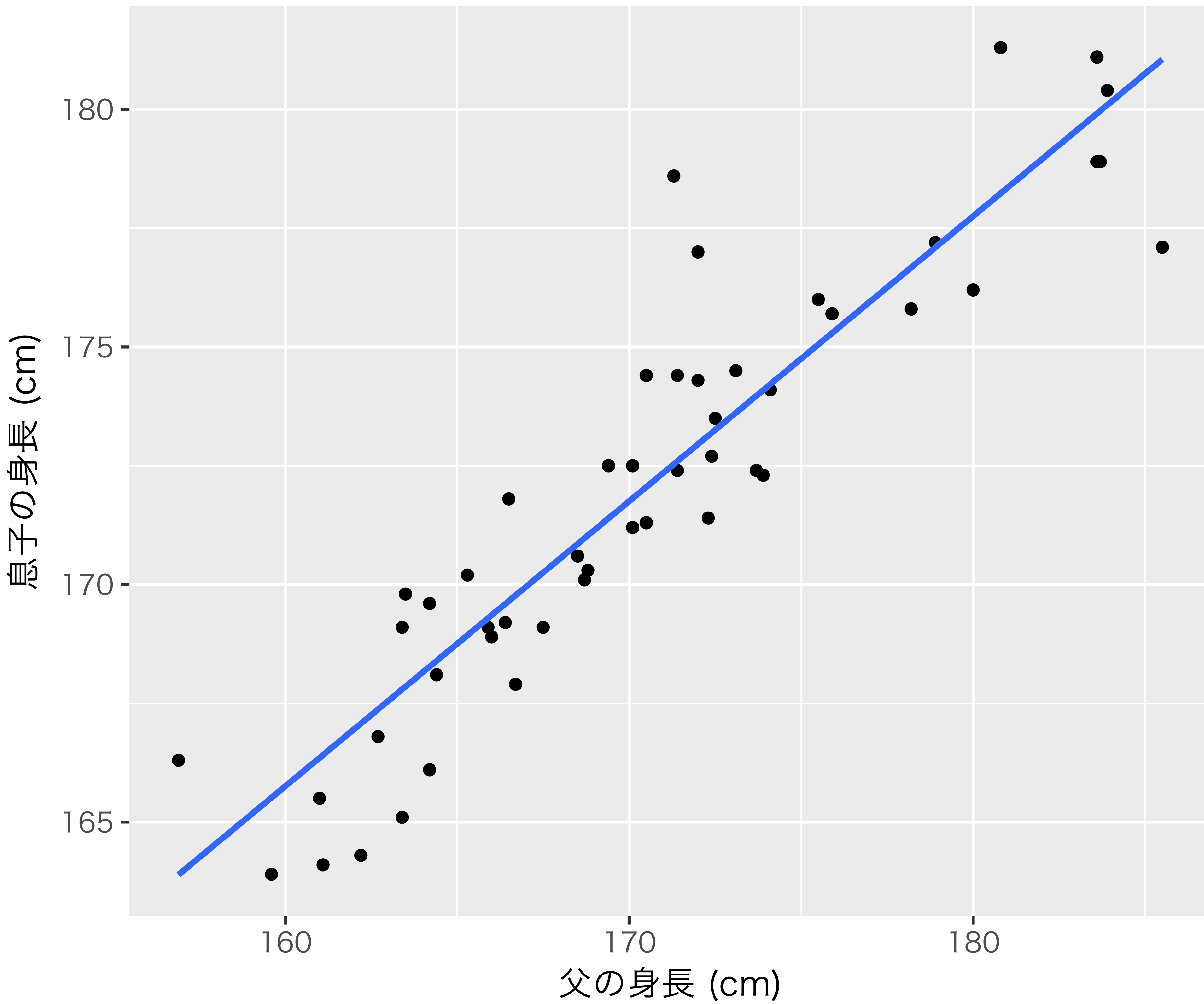
で表すことができる

- a : y切片

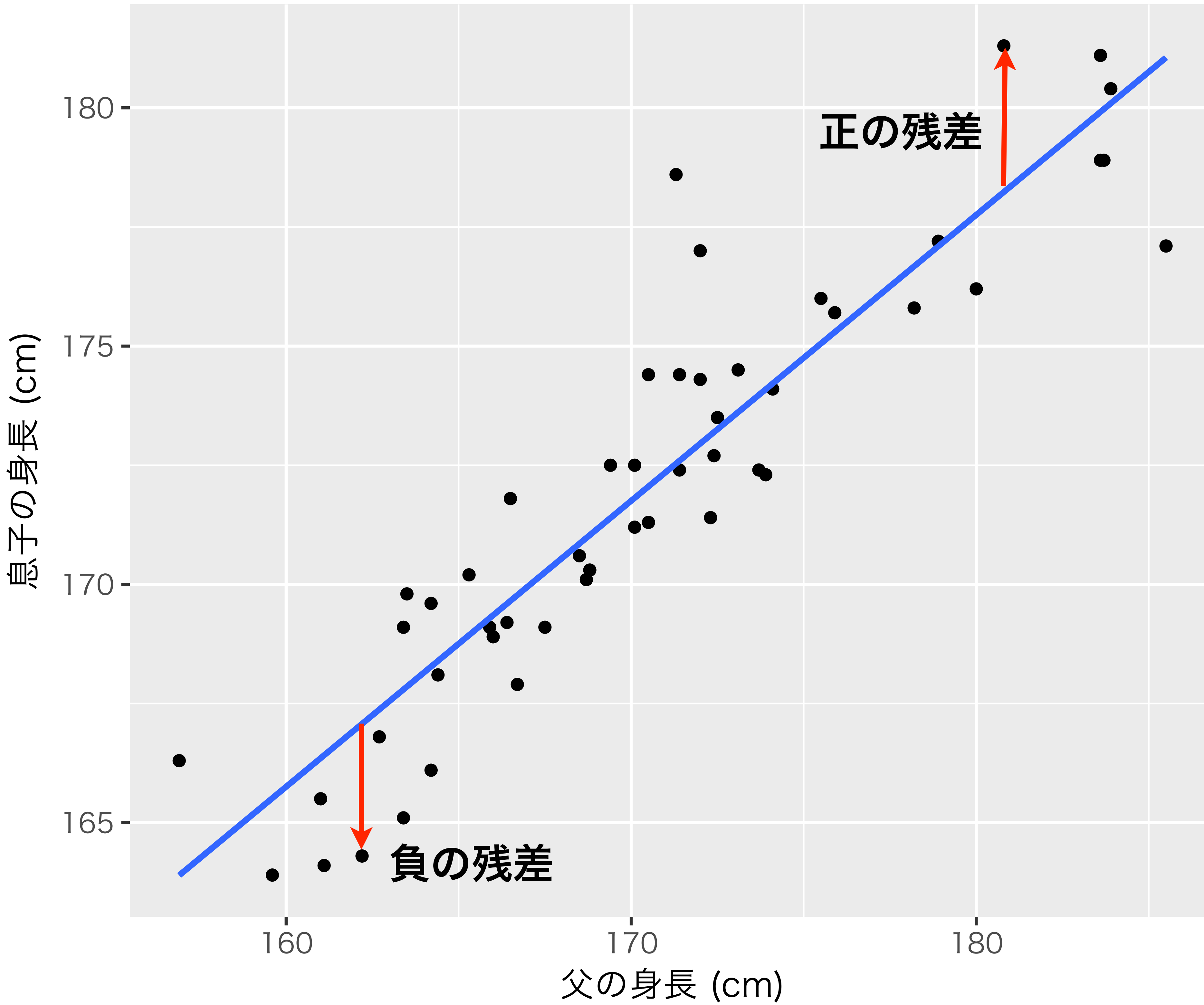
- b : 傾き

➡ 直線を求める : a と b を求める

直線と点はズレる



残差 = 直線から点までの垂直距離



残差：直線と点のずれ

- 残差 (residuals)： e
- 点を直線（a + bx） とそこからのズレに分ける

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

ただし、 $i = 1, 2, \ldots, n$

- $\hat{y}_i$: 予測値
- 観測値 = 予測値 + 残差

ズレを小さくしたい

- できるだけ「ズレ」が小さい直線を引きたい
- 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い、平均値を通ればすべて平均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

この辺りの詳細は『計量経済学』で扱う

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

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

この辺りの詳細は『計量経済学』で扱う

最小二乗法でa と b を求めると

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

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

- これで、回帰直線 $\hat{y} = a + bx$ が求められた
- 回帰直線は、点 $(\bar{x}, \bar{y})$ を通る

この辺りの詳細は『計量経済学』で扱う

結果の解釈：身長の例

息子の身長 = 69.8 + 0.6 × 父の身長

- 父の身長が1cm 高くなると、息子の身長は0.6cm 高くなると考えれられる
- 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

より一般的に

- 傾き b は、説明変数の値が1単位増加したとき、結果変数が何単位増加するかを表す

- 切片 a は、説明変数の値が0のときの結果変数の予測値
  - 説明変数の値が0を取り得ないとき、切片の解釈が難しい（→ 解決策は『計量経済学』で解説）

回帰分析による予測

息子の身長 = 69.8 + 0.6 × 父の身長

- 父の身長が170cm のとき、息子の身長は？
  - 予測値 = 69.8 + 0.6 × 170 = 171.8

◉ 注意

- 今回学んだのは、「点推定」
- 1つの標本から得られた結果なので、誤差がある（真実とは異なる）
- 区間推定と検定の方法は、『計量経済学』で解説する

内挿と外挿

- 内挿 (interpolation)：実際に観察された説明変数の値の範囲での予測

- 外挿 (extrapolation)：実際には観察されていない範囲での予測

    - 父親の身長：[156.9, 185.5]

        - 父の身長が180cmのときの息子の身長：内挿

        - 父の身長が200cm のときの息子の身長：外挿

外挿は危険

- 実際に観測された範囲外では、観察された関係がないかもしれない
    - 特に、直線的関係が局所的な場合
- 外挿はなるべく避ける
- 外挿を行う場合は慎重に
- 理論的におかしな外挿（例：父の身長が5cm のときの息子の身長の予測）はしない

今日のまとめ

- 2変数に原因と結果の関係がありそうなとき、回帰分析で直線的関係を推定する
  - 直線の傾きから、原因が結果に与える影響の強さを推定できる
  - 原因の値から、結果の値を予測できる