高知工科大学 経済・マネジメント学群

計量経済学

2. 二変数の関係

矢内 勇生（やない ゆうき）

https://yukiyanai.github.io

yanai.yuki@kochi-tech.ac.jp

トピック2の目標

1. 2つの変数の関係を調べる方法を理解する

   (1) 質的変数のとき

      - クロス表（分割表）
      - 独立性の検定

   (2) 量的変数のとき

      - 散布図
      - 相関係数

2. 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

2つの変数の関係

(1) 質的変数の場合

質的変数と量的変数

- 質的変数の例：性別、支持 vs. 不支持、大学の成績（S, A, B, C, F）、好きなスポーツ
- 量的変数の例：身長、体重、年齢、年収

クロス表（分割表, contingency table）

（例）性別と内閣支持の関係

	現在の内閣を		
	支持しない	支持する	計
男性	200	300	500
女性	250	250	500
計	450	550	1000

注目するのは行か列か（1）

- 問題ごとに行（row）と列（column）のどちらに注目するか考える
- 例の場合：
  - 行：性別によって、内閣の支持・不支持が変わるか
  - 列：内閣の支持・不支持によって、男女比が異なるか

注目するのは行か列か（2）

行に注目
→行の合計を100%にする

	不支持	支持	計
男性	40%	60%	100%
女性	50%	50%	100%

列に注目
→列の合計を100%にする

	不支持	支持
男性	44%	55%
女性	56%	45%
計	100%	100%

性別によって内閣支持率は異なるか

- 標本：女性より男性のほうが内閣支持の割合が大きい

➡母集団でも男性の支持率のほうが高いといえる？

➡検定：独立性の検定

表：性別と内閣支持の関係

	不支持	支持	計
男性	200 (40%)	300 (60%)	500 (100%)
女性	250 (50%)	250 (50%)	500 (100%)
計	450 (45%)	550 (55%)	1000 (100%)

独立性の検定

- クロス表で提示される2変数に関連があるかどうか調べるための検定

✤ 内閣支持率に男女間で差がない

= 性別と内閣支持に関連がない

= 性別と内閣支持は独立

➡「独立性の検定」

▸χ^2 分布を利用するので、「χ^2 [カイ二乗] 検定」とも呼ぶ

独立性の検定の帰無仮説と対立仮説

- 帰無仮説：2変数は独立である（関連がない）
- 対立仮説：2変数は独立ではない（関連がある）

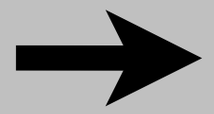
（例）

- H_0: 性別と内閣支持には関連がない
- H_1: 性別と内閣支持には関連がある

独立性の検定（χ^2検定）の考え方

帰無仮説

	不支持	支持	計
男性	45%	55%	100%
女性	45%	55%	100%
計	45%	55%	100%



実際に観測されたデータ

	不支持	支持	計
男性	200 (40%)	300 (60%)	500 (100%)
女性	250 (50%)	250 (50%)	500 (100%)
計	450 (45%)	550 (55%)	1000 (100%)

このようなサンプルはあり得ない？

帰無仮説が正しいとすれば

帰無仮説

	不支持	支持	計
男性	45%	55%	100%
女性	45%	55%	100%
計	45%	55%	100%

帰無仮説の下で
期待されるデータ

	不支持	支持	計
男性	225 (45%)	275 (55%)	500 (100%)
女性	225 (45%)	275 (55%)	500 (100%)
計	450 (45%)	550 (55%)	1000 (100%)

期待度数

帰無仮説が正しい場合のχ^2値（検定統計量）を求める

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

- i は行を表す（k は行の数）
- j は列を表す（m は列の数）
- 観測度数$_{ij}$ は i 行 j 列の観測度数

すべてのセルで $\dfrac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$ を求めて、合計すればよい

例題の場合の検定統計量を求める

観測度数

	不支持	支持
男性	200	300
女性	250	250

期待度数

	不支持	支持
男性	225	275
女性	225	275

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

$$= \frac{(200-225)^2}{225} + \frac{(300-275)^2}{275} + \frac{(250-225)^2}{225} + \frac{(250-275)^2}{275}$$

$$\approx 2.78 + 2.27 + 2.78 + 2.27$$

$$= 10.1$$

統計量を何と比較する？

- カイ二乗分布の臨界値と比較する
  - カイ二乗分布は自由度によって形が変わる
  - クロス表の場合：自由度 = (行数 - 1) x (列数 - 1)
  - 0からどれだけ離れた値を取るかを調べたいので、**棄却域を片側（右側）にとる**
- 「検定統計量 > 臨界値」なら帰無仮説を棄却する

χ^2（カイ二乗）分布 (chi-squared distribution)

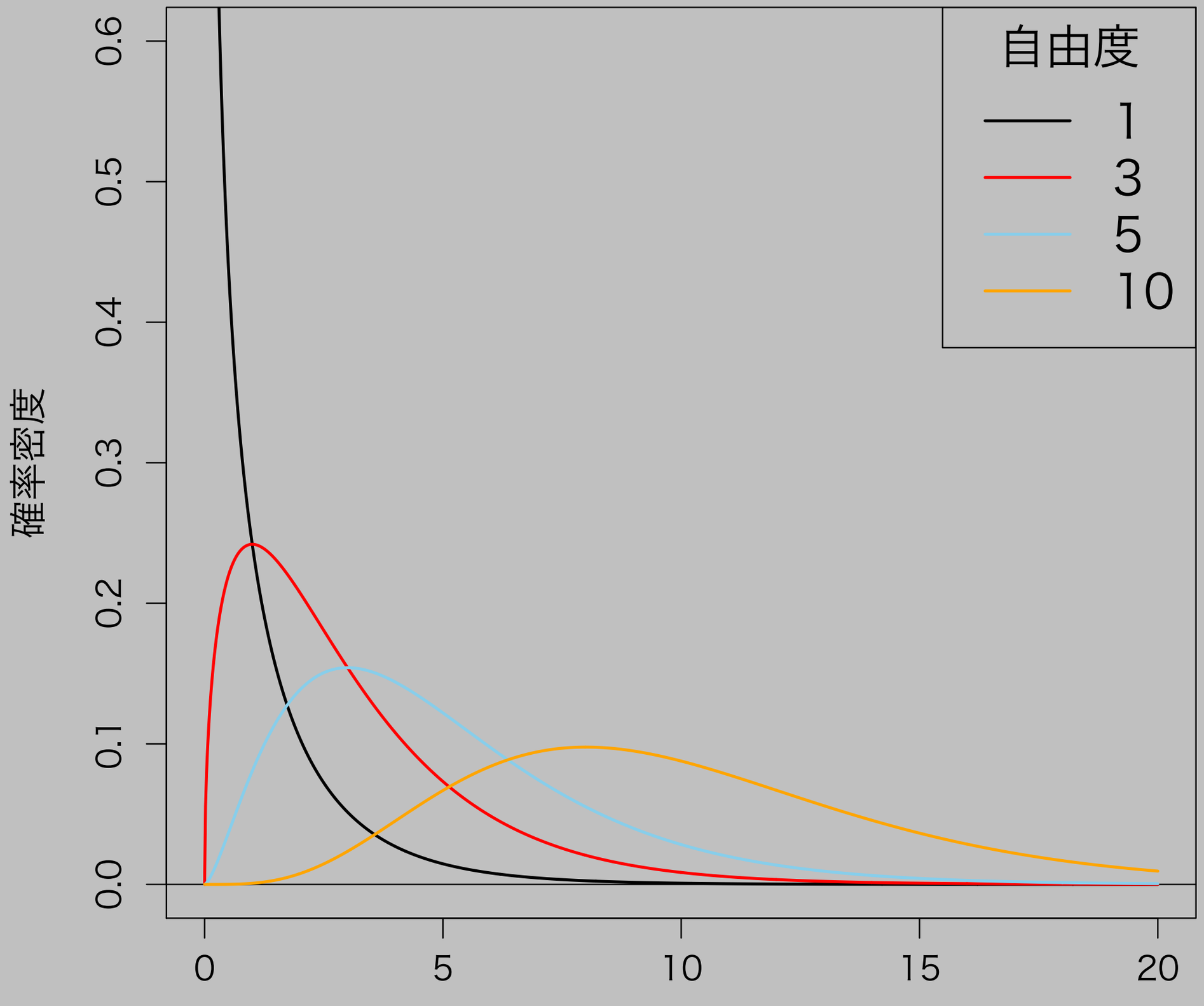
- 確率密度関数 $f(x)$ は、

$$f(x \mid k) = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \qquad (x > 0 \text{ のとき})$$

$$f(x \mid k) = 0 \qquad (x \leq 0 \text{ のとき})$$

ただし、k は x の自由度, $\Gamma(\,.\,)$ はガンマ関数

χ^2分布



自由度（degree of freedom; df）

- 自由（独立）に動かせる値の数
- 各統計量に対して自由度が定められる
  - 例：サイズ N の標本の場合
    - 標本平均の自由度は N
    - 標本（不偏）分散の自由度は $N-1$

例：有意水準5%で検定する

- 検定統計量：10.1
- 2行2列の表 → 自由度 = (2 - 1)(2 - 1) = 1

➡ 有意水準5%の臨界値 = 3.84

```
qchisq(p = 0.05, df = 1, lower.tail = FALSE)
```

➡ 検定統計量 = 10.1 > 3.84 = 臨界値

➡ 帰無仮説を棄却する

➡ 性別によって内閣支持率が異なる！

＊フィッシャーの正確確率検定 (Fisher’s exact test)

- **期待度数が**5を下回るセルがあるとき

➡ 検定統計量が大きめに出てしまうので、独立性の検定が使えない

➡ フィッシャーの正確確率検定（直接確率法）を使う

（この授業では扱わない）

2つの変数の関係

(2) 量的変数の場合

量的変数をクロス表にする

架空の例	年収 500未満	年収 500～1000	年収 1000以上
身長170cm未満	100	80	60
170cm以上	50	75	80

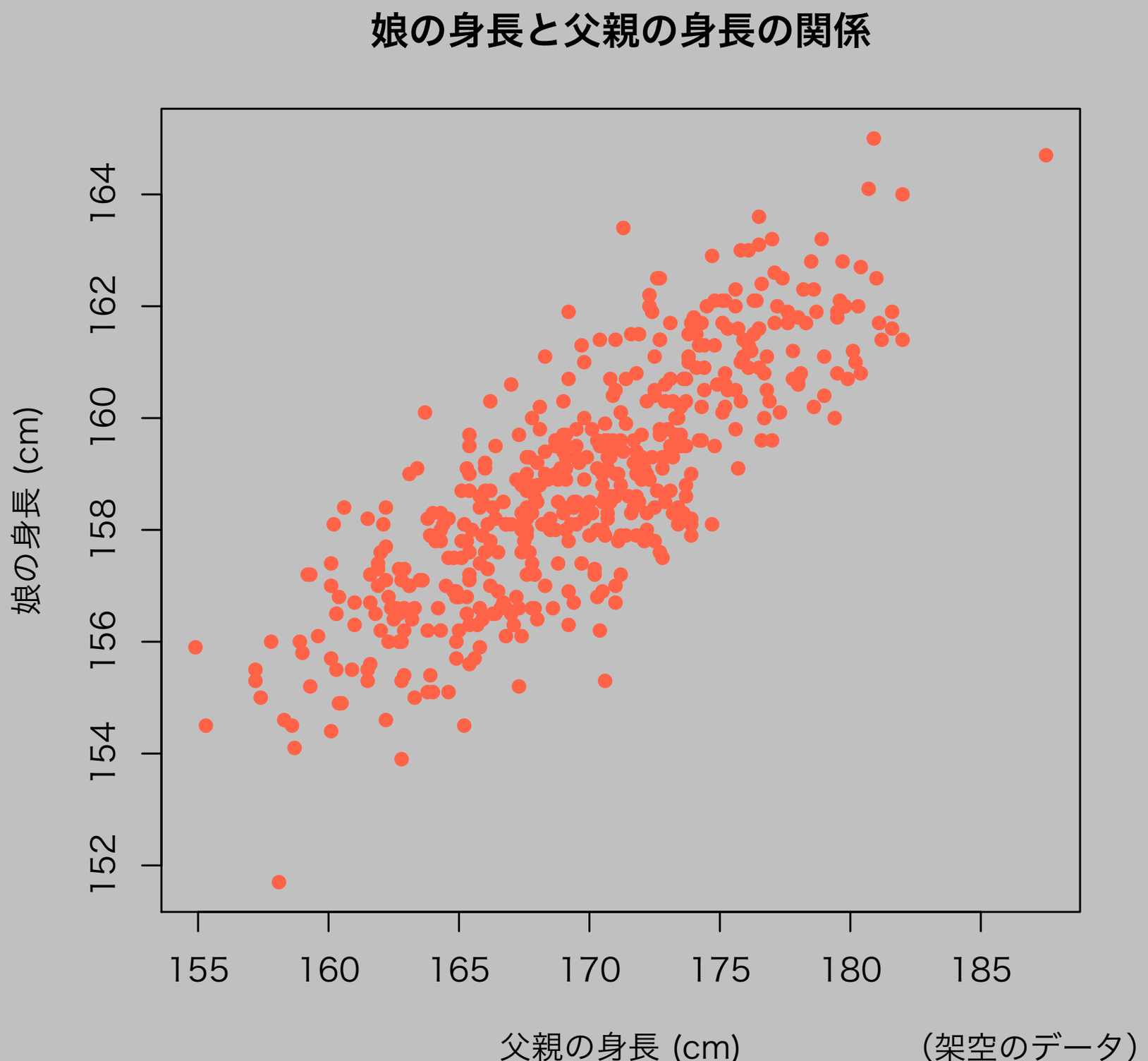
- 情報が失われる

➡ 表にせずに関係を表す

1.図示する：散布図

2.統計量を求める：相関係数

2変数の関係を図示する: 散布図（scatter plot）



相関関係

- 相関関係 (correlation)：
  - 2つの物事（変数）AとBの間の直線的な関係
  - Aの変化に合わせてBも変化する
  - 統計量：相関係数 r $(-1 \leq r \leq 1)$
  - Aが増える（減る）とき、Bも増える（減る）：正の相関 $(r > 0)$
  - Aが増える（減る）とき、Bが減る（増える）：負の相関 $(r < 0)$
  - $|r|$ が1に近いほど関係が強い

2変数の関係を表す統計量：
相関係数（correlation coefficient）

- 変数 x と変数 y の相関係数 r

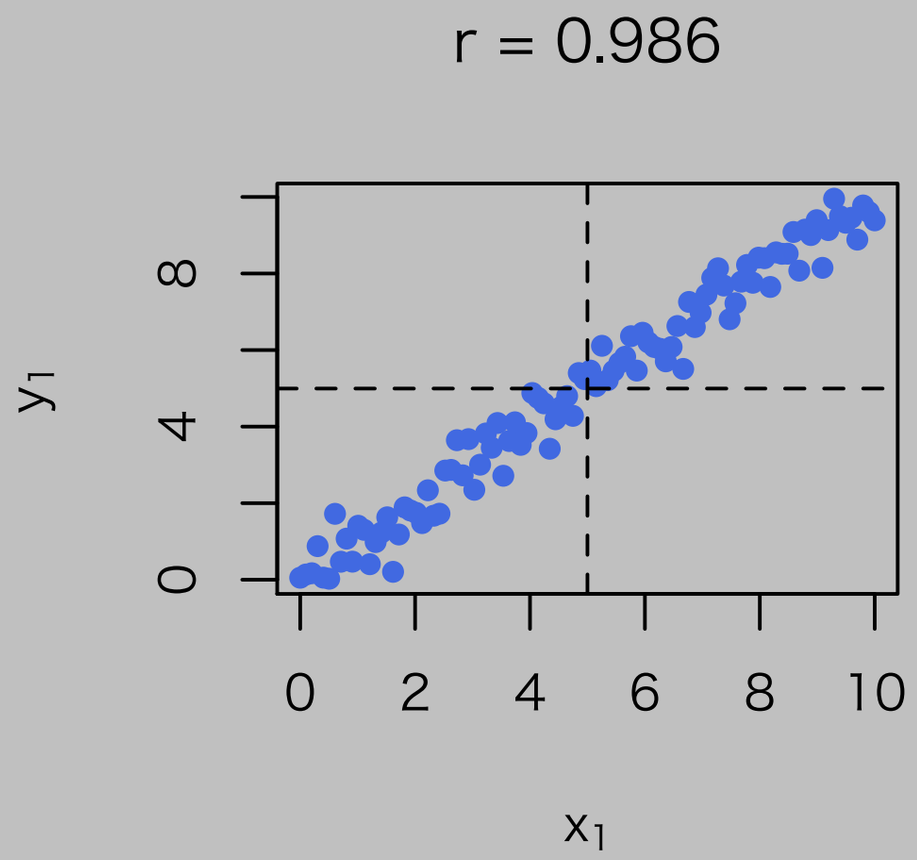
$$r = \frac{x\text{と}y\text{の共分散}}{\sqrt{x\text{の不偏分散}}\sqrt{y\text{の不偏分散}}}$$

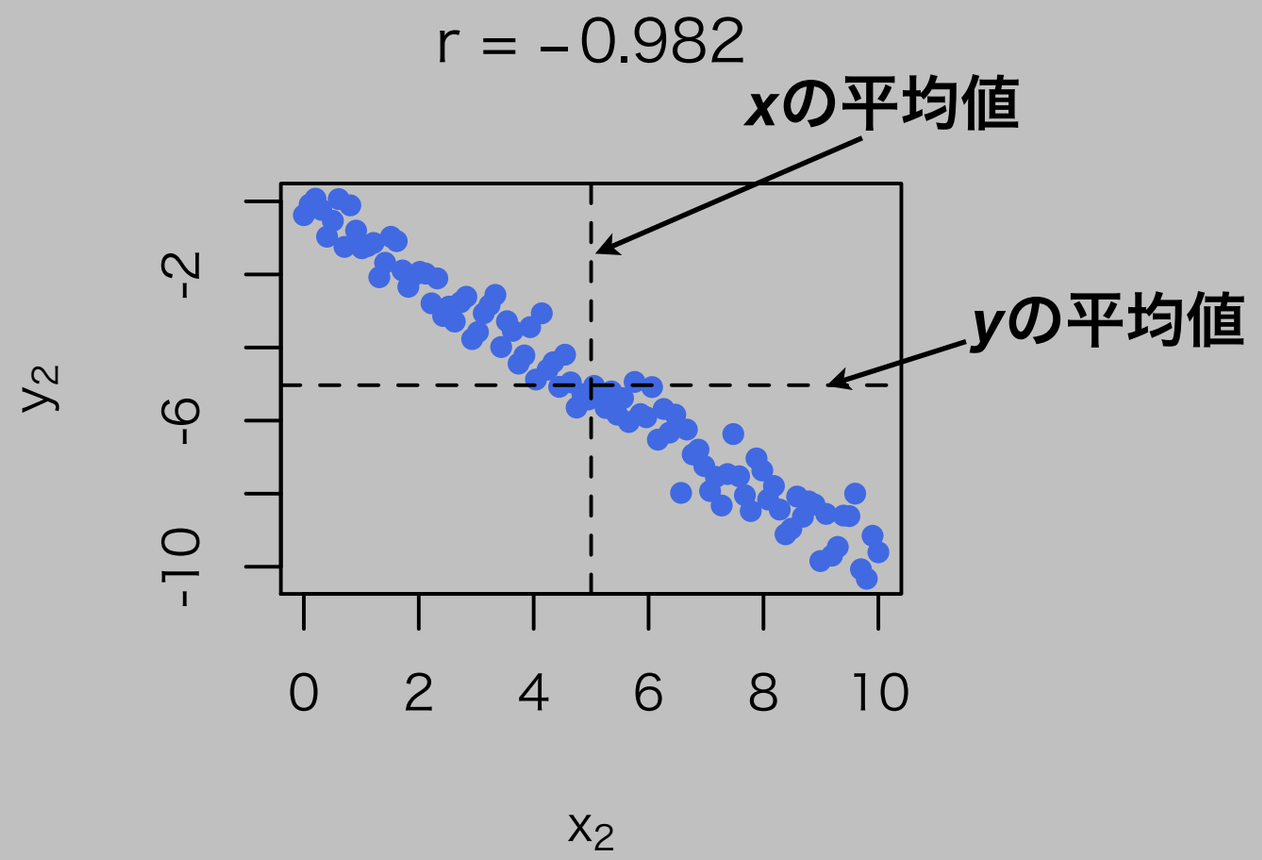
$$= \frac{\dfrac{\sum_{i=1}^{N}[(x_i-\bar{x})(y_i-\bar{y})]}{N-1}}{\sqrt{\dfrac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}\sqrt{\dfrac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N-1}}}$$

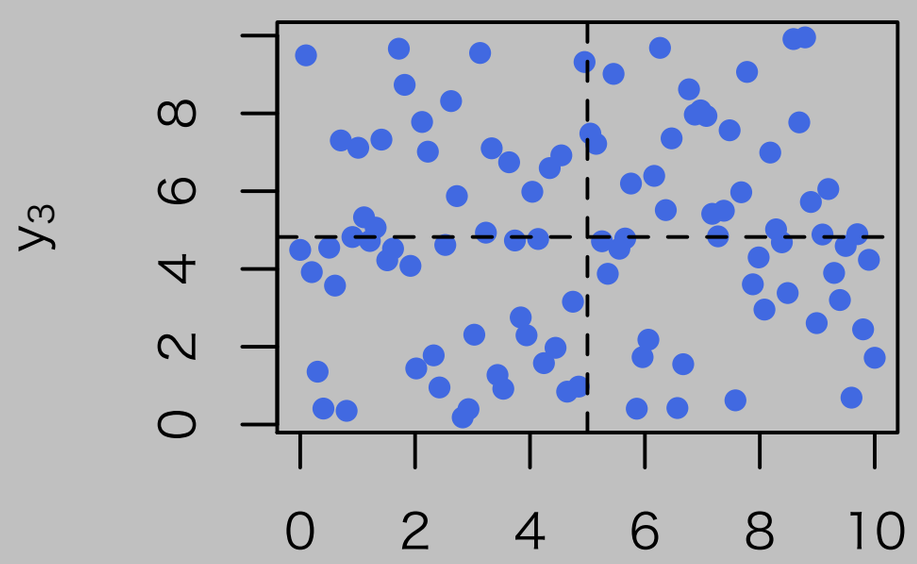
$$= \frac{\sum_{i=1}^{N}[(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}$$

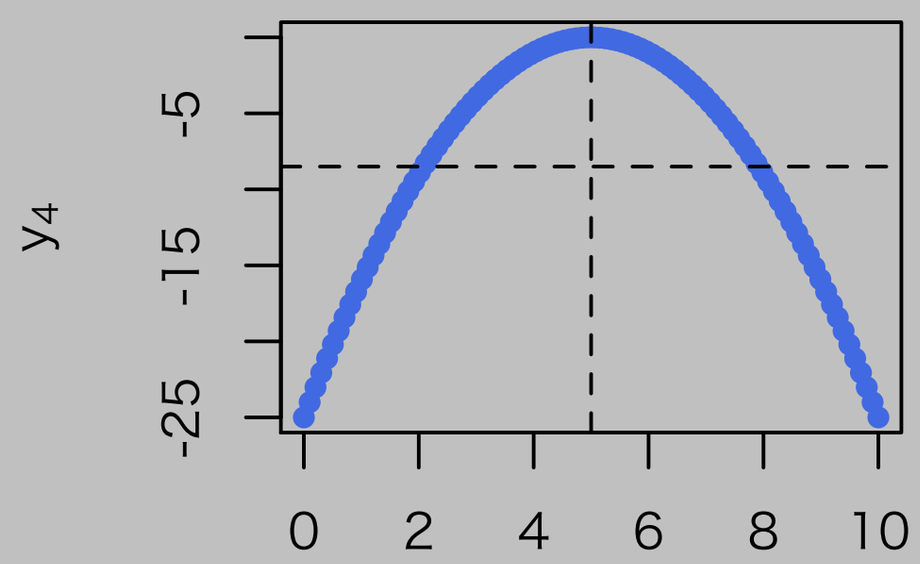
相関係数の特徴

- 2変数の**直線的な**関係の強さを表す
- 取り得る値の範囲は [-1, 1]
  - 1：正の直線的関係（一方が大きくなるとき、他方も大きくなる）が最も強い
  - −1：負の直線的関係（一方が大きくなるとき、他方が小さくなる）が最も強い
  - 0：直線的関係がない（曲線的関係は強いかもしれないことに注意）
- **因果関係はわからない**
- **因果関係を仮定するとして、原因が結果にどれだけ影響を与えるかはわからない**


r = 0.986
y1
x1
0
4
8
0
2
4
6
8
10


r = −0.982
xの平均値
yの平均値
y2
x2
-2
-6
-10
0
2
4
6
8
10


r = 0.044
y3
0 2 4 6 8
0
2
4
6
8
10


r = 0
y4
-5
-15
-25
0
2
4
6
8
10

トピック2：前半のまとめ

- 2つの変数のまとめ方：変数の種類によって異なる
  - 質的変数：クロス集計表、独立性の検定（カイ二乗検定）
  - 量的変数：散布図、相関係数

散布図と相関係数は必ずセットで使う

回帰分析の基礎

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)：研究において興味がある結果
  - 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、単回帰と重回帰の両者を指す

回帰分析の基礎

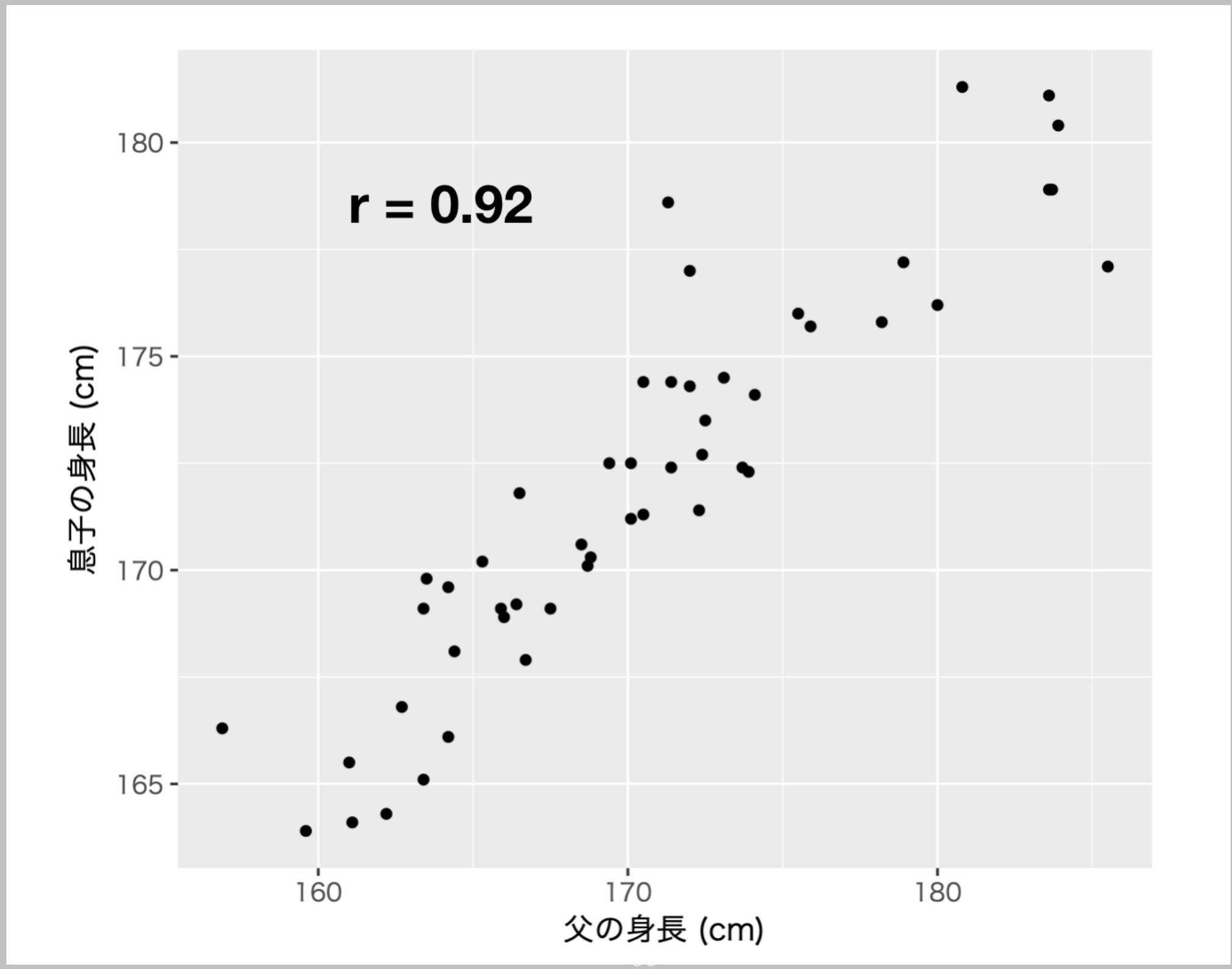
線形回帰

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

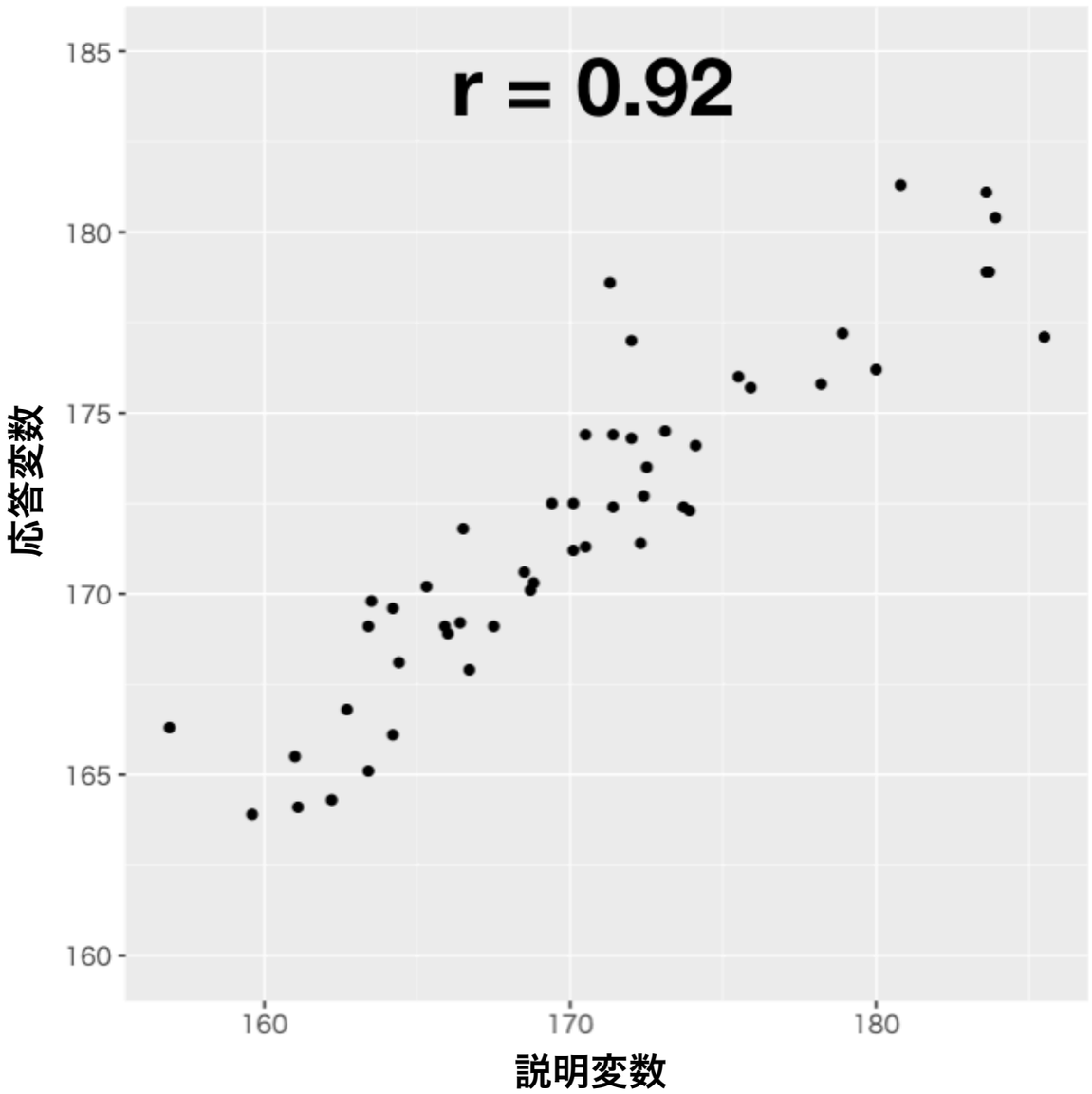


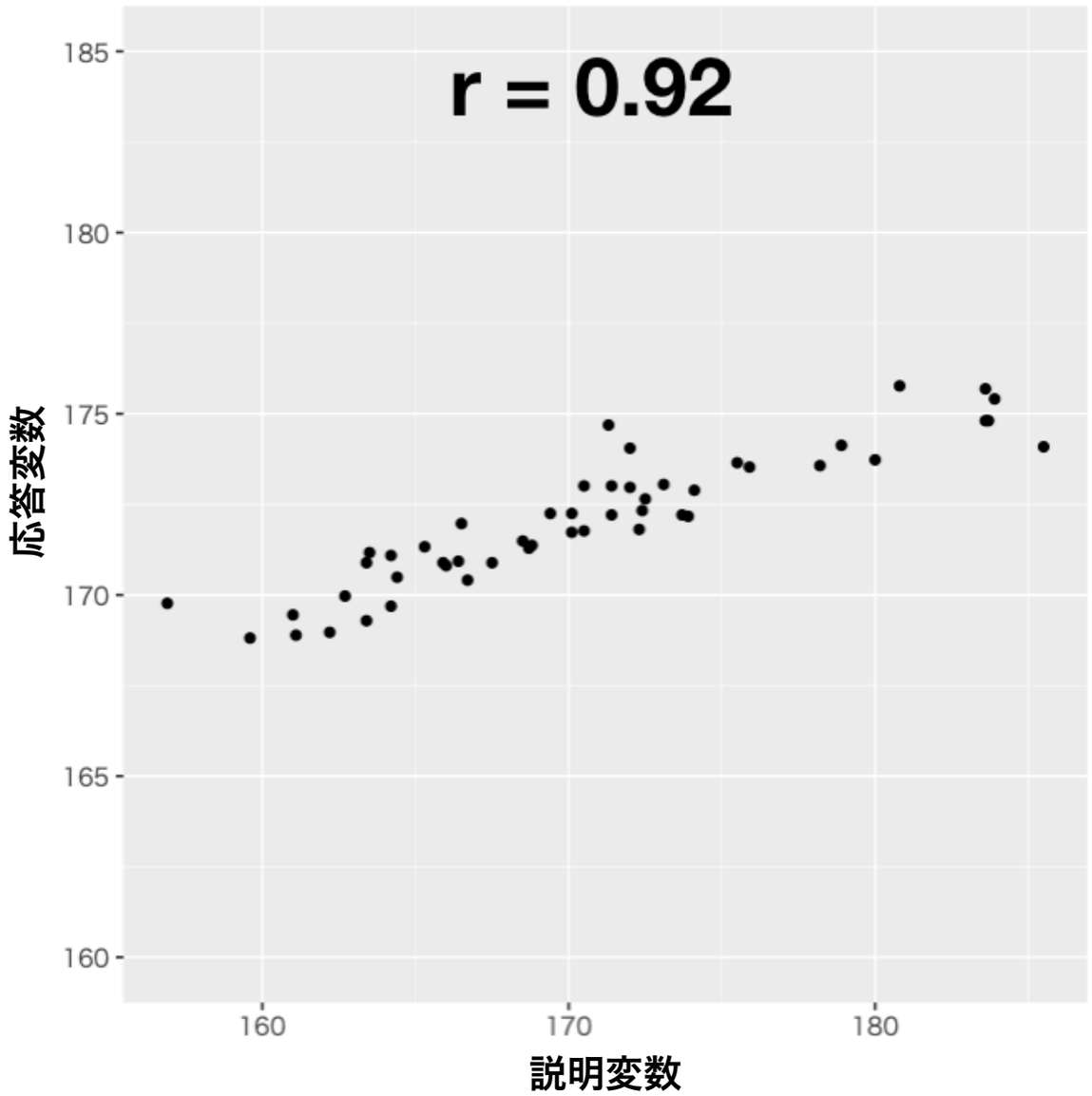
わかったことと新たな疑問

- 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





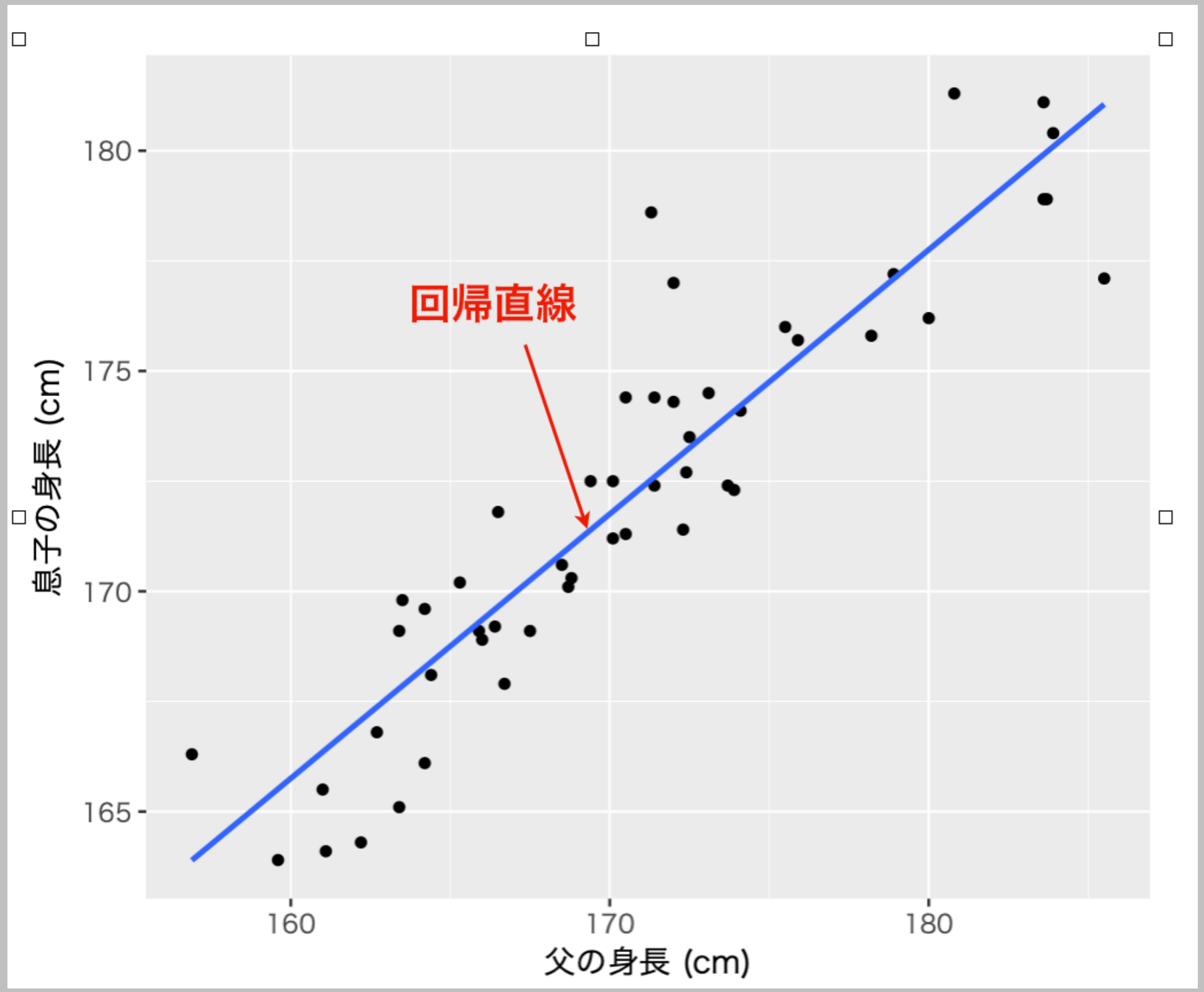
相関係数だけでは不十分

- 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（説明変数が1単位増加するのに伴って、応答変数が何単位分増加するか）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

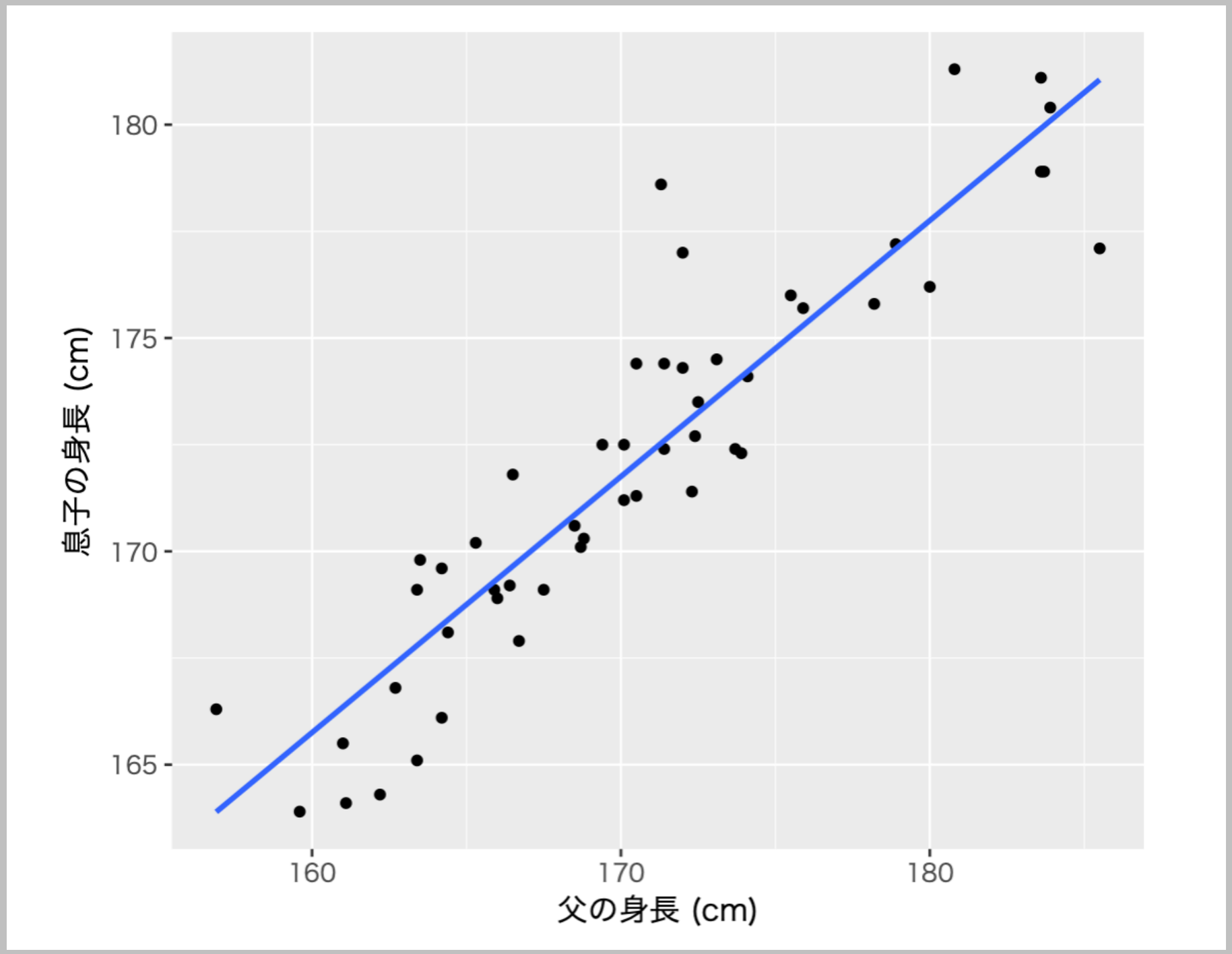
$$y = a + bx$$

で表すことができる

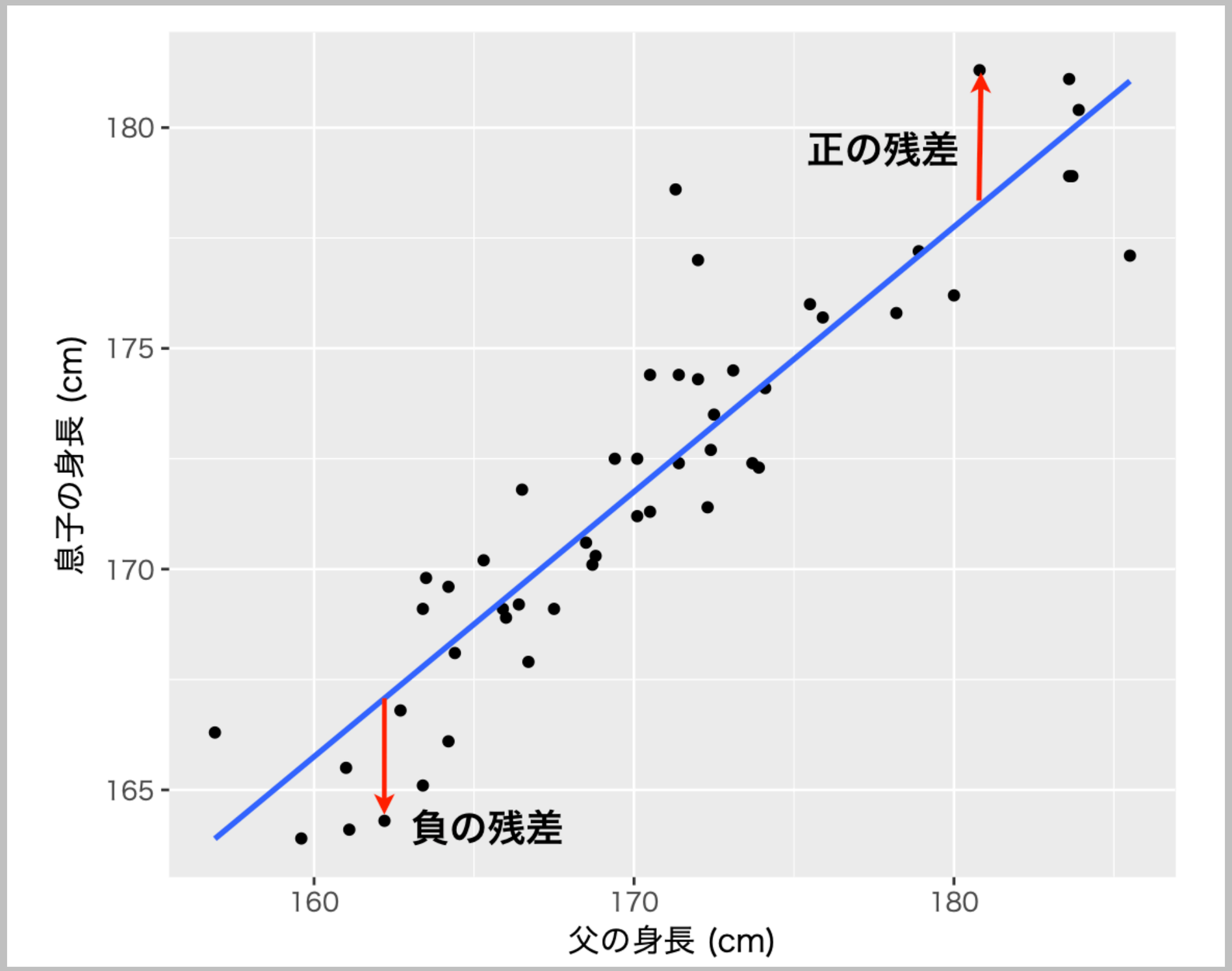
- a：y 切片（x が0のときの y の値）
- b：傾き（x が1単位増加したときの y の変化量）

- 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

回帰分析の基礎

単回帰

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）

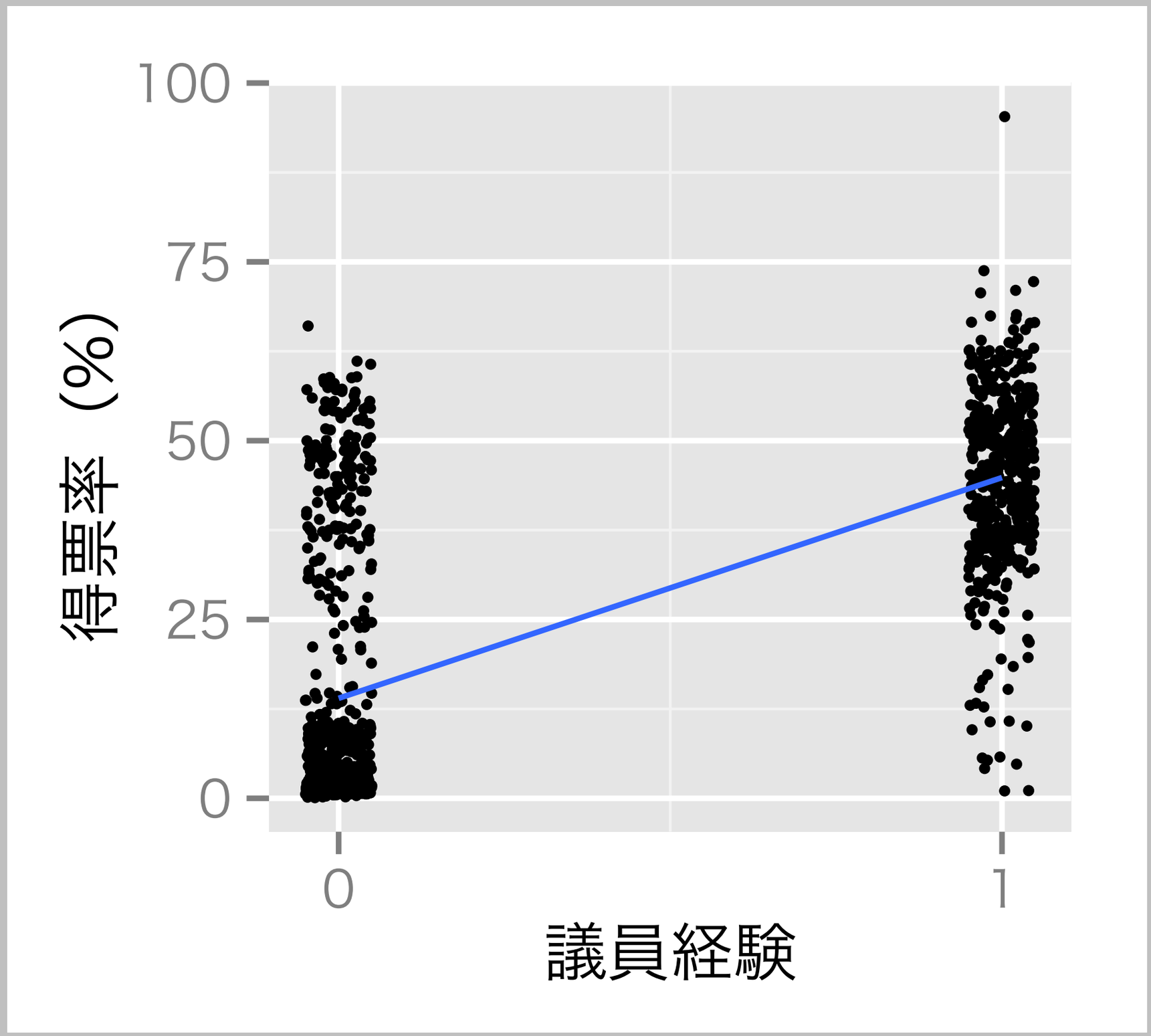
$$\widehat{\text{得票率}} = 14 + 31 \cdot \text{議員経験}$$

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

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

- 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：散布図と回帰直線



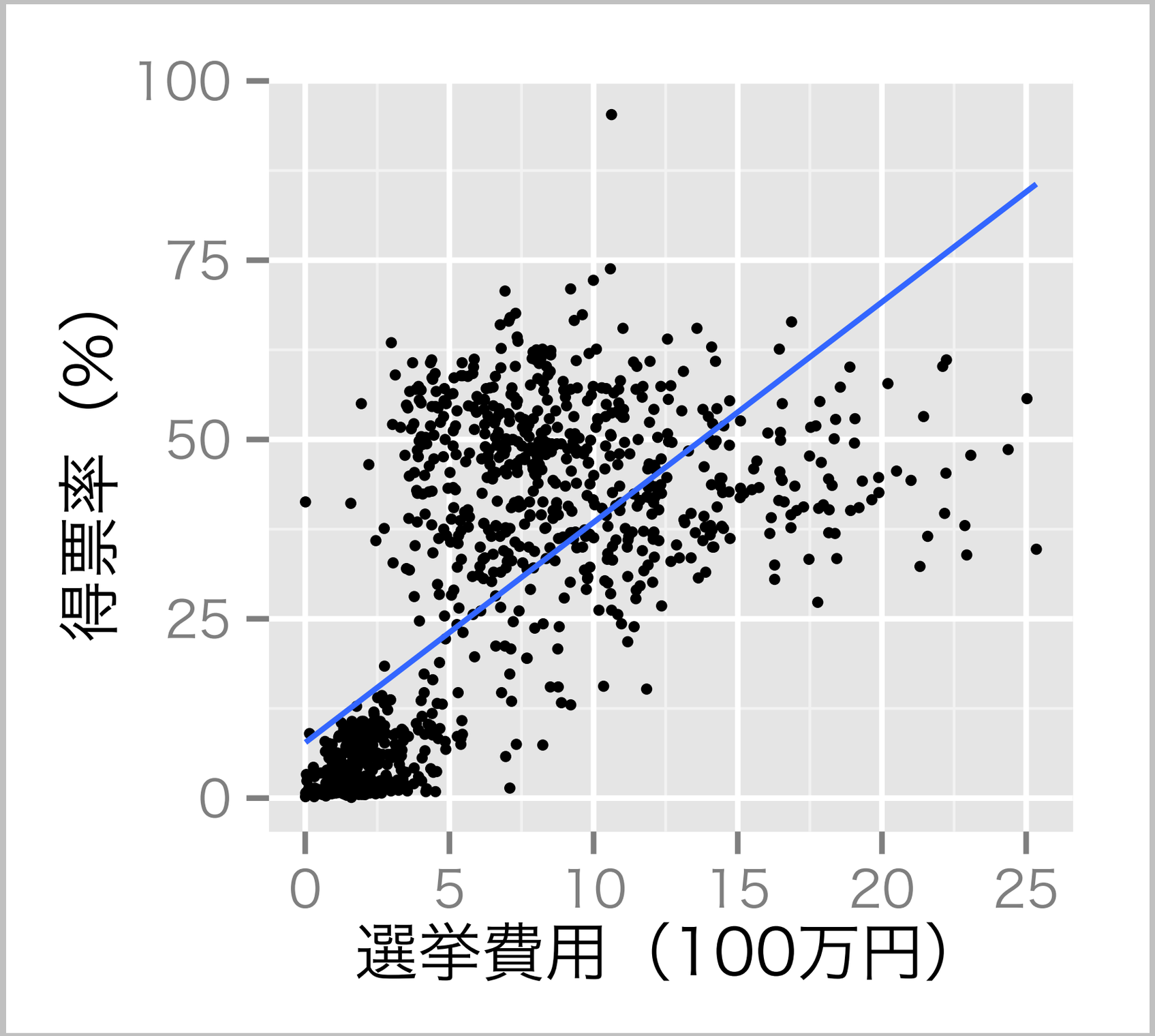
モデル2：説明変数が量的変数の場合

- 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

トピック2：後半のまとめ

- 回帰直線で2変数の「直線的」関係を要約できる
  - 直線を求める：切片と傾きを求める
  - 与えられたデータに対し、決められた方法で計算すれば求められる
    - 統計量、記述統計
  - 傾きを得ることによって、一方の値から他方の値を「予測」することができるようになる

次のトピック

3. 因果推論