政治学方法論 I

5. 線形回帰 (1)

矢内 勇生

法学部・法学研究科

2016 年 5 月 18 日

神戸大学

今日の内容

1. はじめに
   - 線形回帰とは何か
   - 専門用語の定義
2. 線形回帰の基礎
   - 予測変数が 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
   - 説明変数が複数のモデル：重回帰
   - 相互作用 (interaction) を考慮に入れる

線形回帰とは？

線形回帰 (Linear Regression)

予測変数（説明変数）の線形関数で定義される値によって、結果変数（応答変数）の平均値がどのように変化するかを要約する方法。

「線形」とは何か

- 関数 $f(.)$ が以下の条件を満たすとき、$f(.)$ を線形 **(linear)** 関数と呼ぶ

  加法性 (additivity) $f(x+y) = f(x) + f(y), \quad \forall x, \forall y$

  斉次性 (homogeneity) $f(kx) = kf(x), \quad \forall x, \forall k$

- 変化率が一定
- 横軸を x、縦軸を $f(x)$ とするグラフにおいて、線形関数は直線で表される

結果変数と予測変数

- 結果変数 **(outcome variable)**：説明される変数 その他の呼び名: 応答変数 (response v), 従属変数 (dependent v), 被説明変数 (explained v), regressand, etc.
- 予測変数 **(predictor v's)**: 結果変数の原因と見做されるもの

  その他の呼び名: 説明変数 (explanatory v), 独立変数 (independent v), regressor, etc.
- 予測変数と結果変数の間の因果関係は、線形回帰を行うときの仮定： 仮定はデータによって確かめられない
- 「y を x に回帰する (we regress the outcome on the predictor(s))」

予測変数：説明変数と統制変数

- 一般的な区別
  - 説明変数：主な原因
  - 統制（コントロール）変数：主な原因以外の変数
- 統計学（数学）上の違い：なし
- 説明変数と統制変数を区別する必要はない

ダミー変数

ダミー変数： 一般的には、ある特徴の存在を表す

- 女性ダミー：「女性」という特徴を持っていれば 1、そうでなければ 0
- 男性ダミー：「男性」という特徴を持っていれば 1、そうでなければ 0
- 女は 1、男は 2 という性別変数は？

ダミー変数

ダミー変数： 一般的には、ある特徴の存在を表す

- 女性ダミー：「女性」という特徴を持っていれば 1、そうでなければ 0
- 男性ダミー：「男性」という特徴を持っていれば 1、そうでなければ 0
- 女は 1、男は 2 という性別変数は？ 変数としては問題ない

  が、通常はダミー変数とは呼ばない：「性別」という特徴は誰にでもあり、ある特徴の存在を示すのに役立たない

単回帰と重回帰

- 単回帰 (simple regression)：予測変数が 1 つ（統制変数なし）
- 重回帰 (multiple regression)： 予測変数が（統制変数を含めて）2 つ以上
- 回帰：単回帰と重回帰の両者を指す

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

- 結果変数：得票率 (%)
- 予測変数：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 (2013), hr96-09.dta（以下、特にことわりのない限りこのデータを使う. 詳しくはウェブで）

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^（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：$31 = 45 - 14 =$ 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

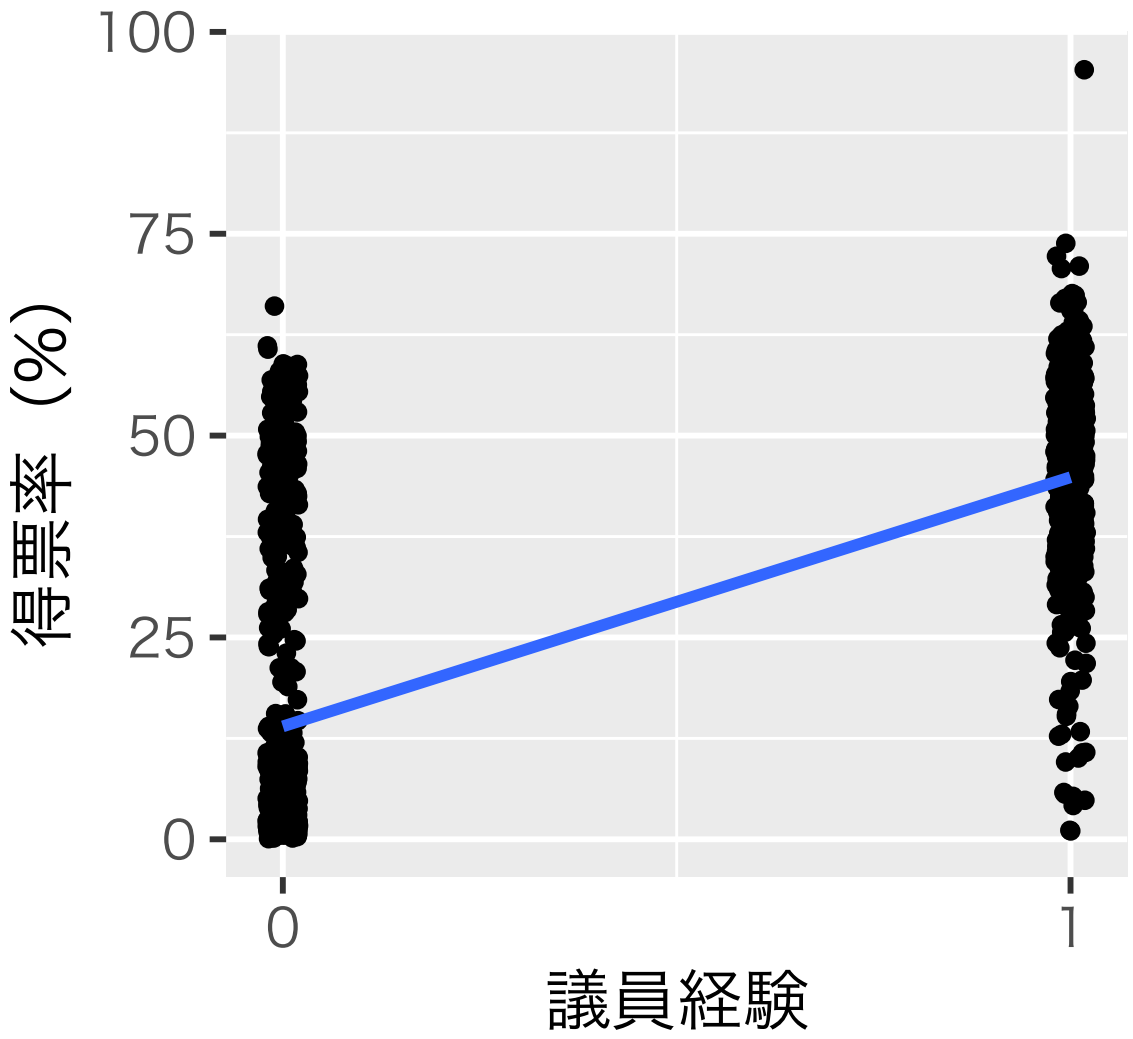

図: 議員経験の有無で得票率を説明する

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

- 結果変数：得票率 (%)
- 予測変数：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

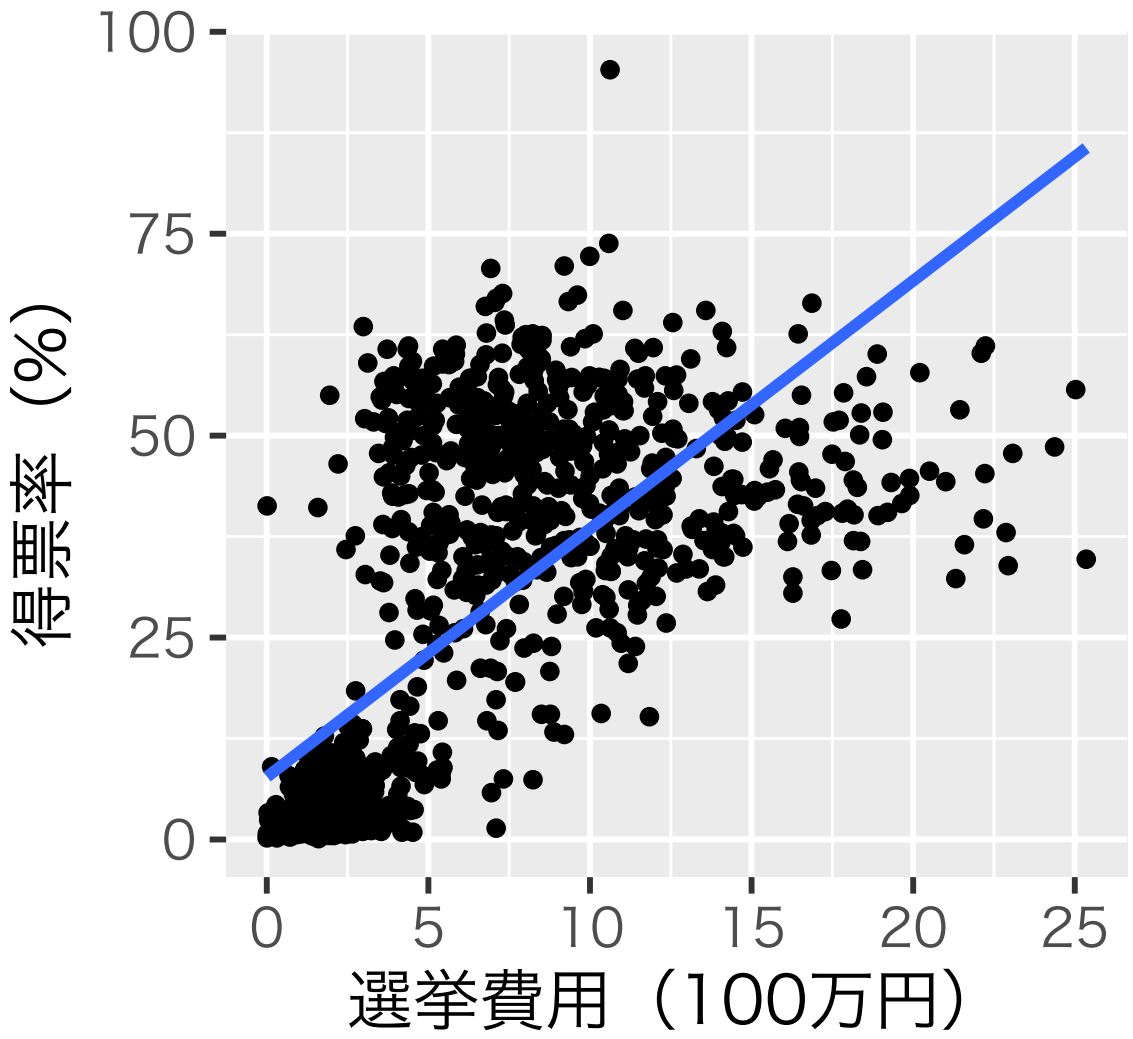

図: 選挙費用で得票率を説明する

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

モデル 3

衆議院議員総選挙での得票率を議員経験の有無と選挙費用の大きさで説明する

- 結果変数：得票率 (%)
- 予測変数 1：議員経験（なし=0, あり=1）
- 予測変数 2：選挙費用（測定単位：100 万円）
- 推定結果：

$$\text{得票率} = 7.9 + 18.1 \cdot \text{議員経験} + 1.9 \cdot \text{選挙費用} + \text{誤差}$$

- 2 本の回帰直線（次のスライド）は平行：議員経験の有無によって選挙費用の係数が変わらないようにモデル化（係数に制約をかけている）

モデル 3 の図示：散布図と回帰直線

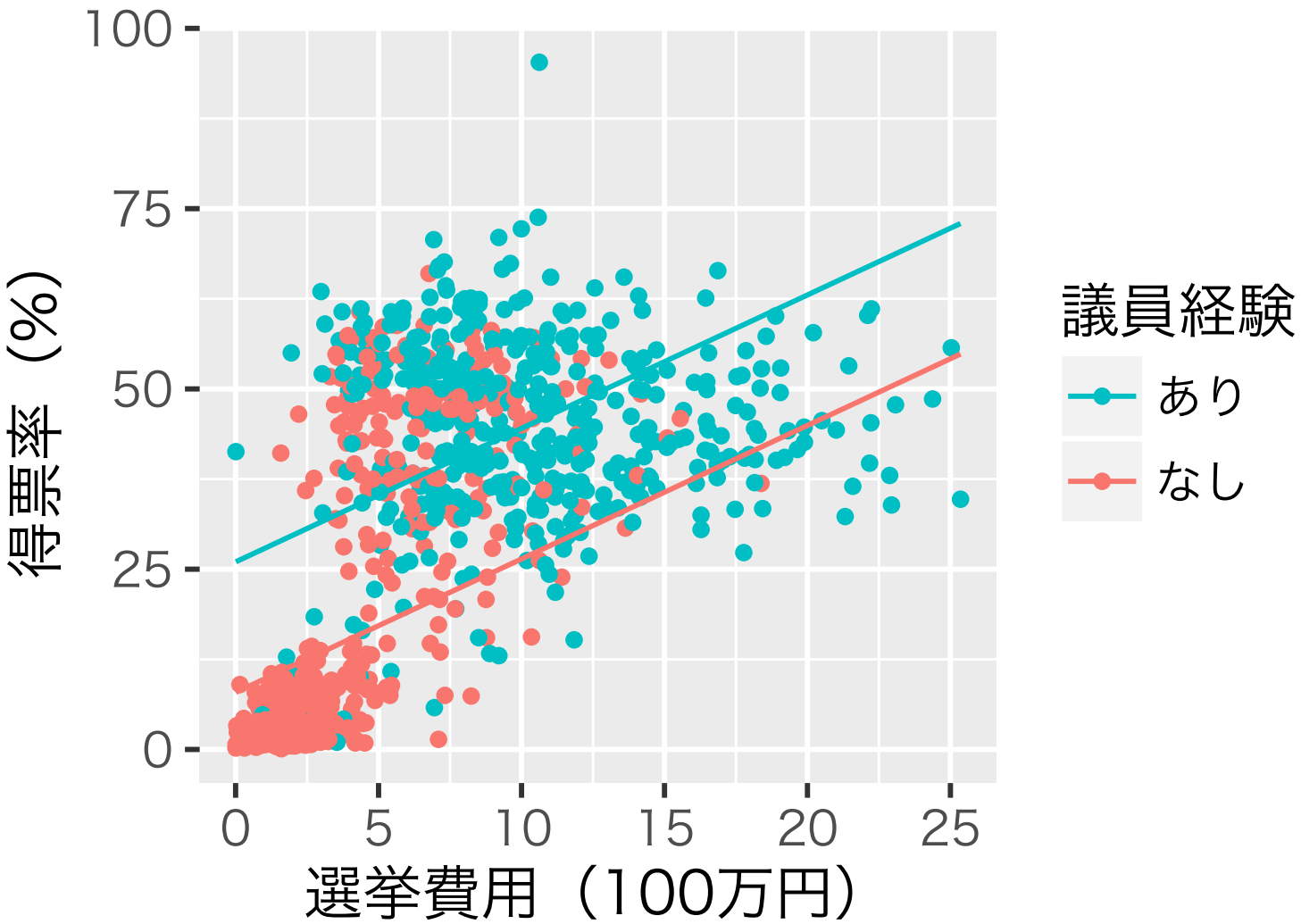

図: 議員経験の有無と選挙費用で得票率を説明する

モデル 3 が示すこと

$$\text{得票率} = 7.9 + 18.1 \cdot \text{議員経験} + 1.9 \cdot \text{選挙費用} + \text{誤差}$$

- 切片 (7.9)：候補者に議員経験がなく（議員経験=0)、選挙費用をまったく支出しない (選挙費用=0）のときに予測される得票率
- 議員経験の係数 (18.1)：選挙費用がまったく同額で、議員経験の有無が異なる候補者間の予測得票率の差
  - 選挙費用が同じなら、議員経験がある候補者のほうが平均して 18.1 ポイント高い得票率を得る
- 選挙費用の係数 (1.9)：議員経験の有無が同じで、選挙費用の額が 1 単位（100 万円）異なる候補者間の予測得票率の差
  - 議員経験の有無が同じなら、選挙費用を 100 万円増やすごとに、平均して 1.9 ポイント得票率が上がる

重回帰の回帰係数：他の要因を一定に・・・

- 各説明変数の係数：他の説明変数の値を一定に保ったとき、説明変数１単位の変化が、応答変数の予測値を何単位変化させるかを表す（単位は変数の取り方次第）

重回帰の回帰係数：他の要因を一定に・・・

- 各説明変数の係数：他の説明変数の値を一定に保ったとき、説明変数１単位の変化が、応答変数の予測値を何単位変化させるかを表す（単位は変数の取り方次第）
- 「他の変数を一定に保つ」ことはいつも可能か？

重回帰の回帰係数：他の要因を一定に・・・

- 各説明変数の係数：他の説明変数の値を一定に保ったとき、説明変数 1 単位の変化が、応答変数の予測値を何単位変化させるかを表す（単位は変数の取り方次第）
- 「他の変数を一定に保つ」ことはいつも可能か？
  → **No!!!**
- 例：「年齢」と「年齢の二乗」を説明変数に加えるとき
- 例：相互作用を考慮する（交差項を説明変数に加える）とき

モデル 4：相互作用を考える

- モデル 3：2 つの集団（議員経験なりとあり）の傾きが同じ
- モデル 4：傾きを「自由に」する
  $\rightarrow$　議員経験と選挙費用の相互作用を考慮に入れる
  - 結果変数：得票率 (%)
  - 予測変数 1：議員経験（なし=0, あり=1）
  - 予測変数 2：選挙費用（測定単位：100 万円）
  - 予測変数 3：議員経験 $\times$ 選挙費用
  - 推定結果：

$$\begin{aligned} \text{得票率} \quad = \quad & -2.1 + 45.9 \cdot \text{議員経験} + 4.9 \cdot \text{選挙費用} \\ & -4.8 \cdot \text{議員経験} \cdot \text{選挙費用} + \text{誤差} \end{aligned}$$

モデル 4 の図示：散布図と回帰直線

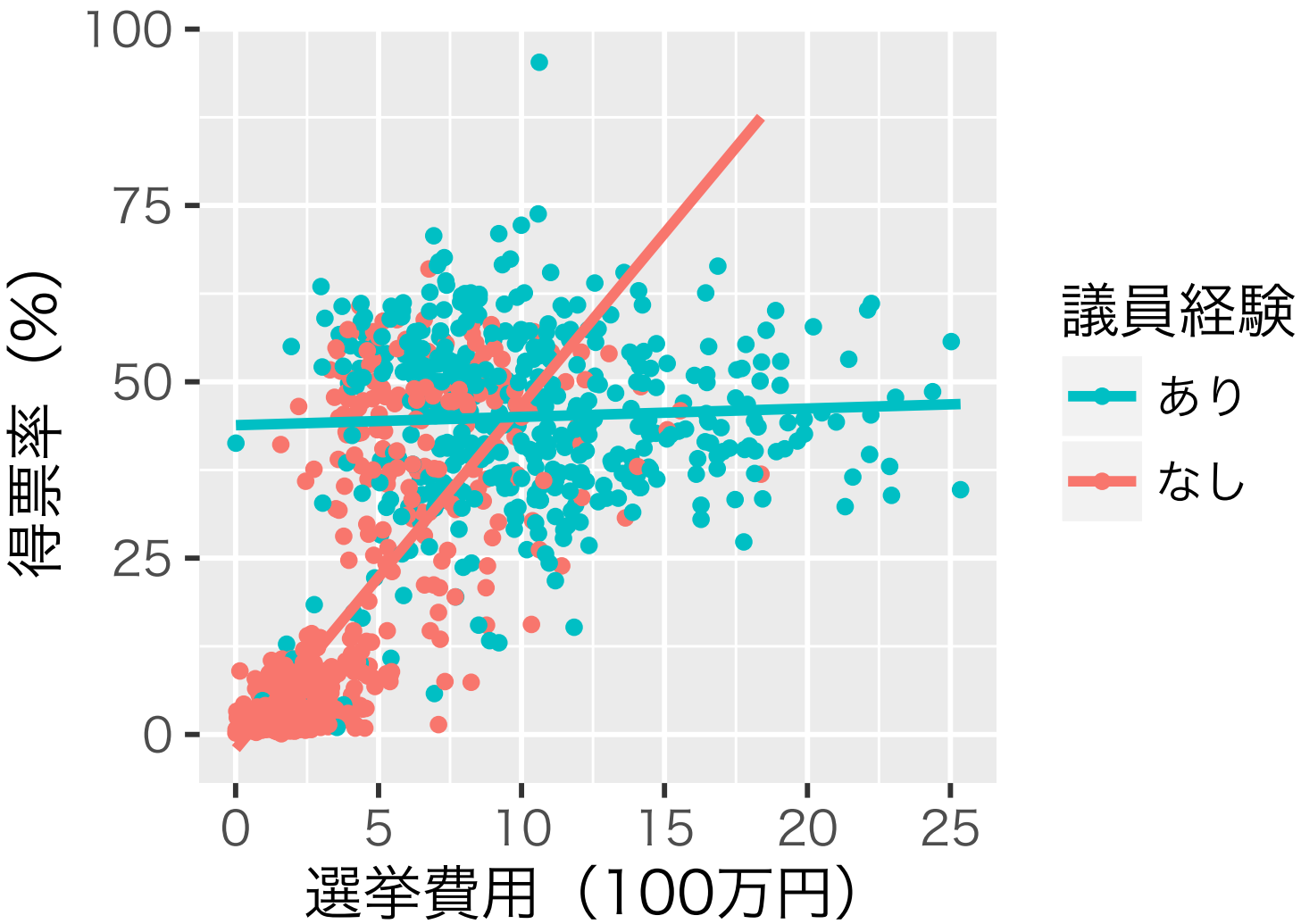

図: 議員経験の有無と選挙費用で得票率を説明する

モデル 4 の意味：各推定値の意味

- 切片：議員経験がなく、選挙費用が 0 の候補者の予測得票率（マイナス？？？）
- 議員経験の係数：選挙費用が **0** の候補者の中で、議員経験がある者と議員経験のない者の間の予測得票率の差
- 選挙費用の係数：議員経験がない者の中で、選挙費用が 1 単位だけ異なる候補者間の予測得票率の差
- 相互作用の係数：議員経験がある候補者とない候補者の間にある回帰直線の傾きの差

相互作用項を含むモデルの解釈には特に注意が必要！

相互作用 (interaction) を考慮に入れる

モデル 4 の意味：場合分けして考える

1. 議員経験がない候補者：

$$\begin{aligned}\widehat{\text{得票率}} &= -2.1 + 45.9 \cdot 0 + 4.9 \cdot \text{選挙費用} - 4.8 \cdot 0 \cdot \text{選挙費用} \\ &= -2.1 + 4.9 \cdot \text{選挙費用}\end{aligned}$$

2. 議員経験がある候補者：

$$\begin{aligned}\widehat{\text{得票率}} &= -2.1 + 45.9 \cdot 1 + 4.9 \cdot \text{選挙費用} - 4.8 \cdot 1 \cdot \text{選挙費用} \\ &= -2.1 + 45.9 + 4.9 \cdot \text{選挙費用} - 4.8 \cdot \text{選挙費用} \\ &= 43.8 + 0.1 \cdot \text{選挙費用}\end{aligned}$$