政治学方法論 I

第 6 回：線形回帰分析（1）

矢内 勇生

神戸大学　法学部/法学研究科

2014 年 11 月 5 日

今日の内容

線形回帰とは？

線形回帰 (linear regression)

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

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

- $f(x)$ 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

応答変数と説明変数

- 応答変数 (response variable)：説明したい「結果」

  その他の呼び方：従属変数 (dependent v), 結果変数 (outcome v), 被説明変数 (explained v), 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

- 応答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
  $\rightarrow$ 説明変数とコントロール変数を区別する必要はない

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？

  ダミー変数ではない！
  ×：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に使う

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 (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

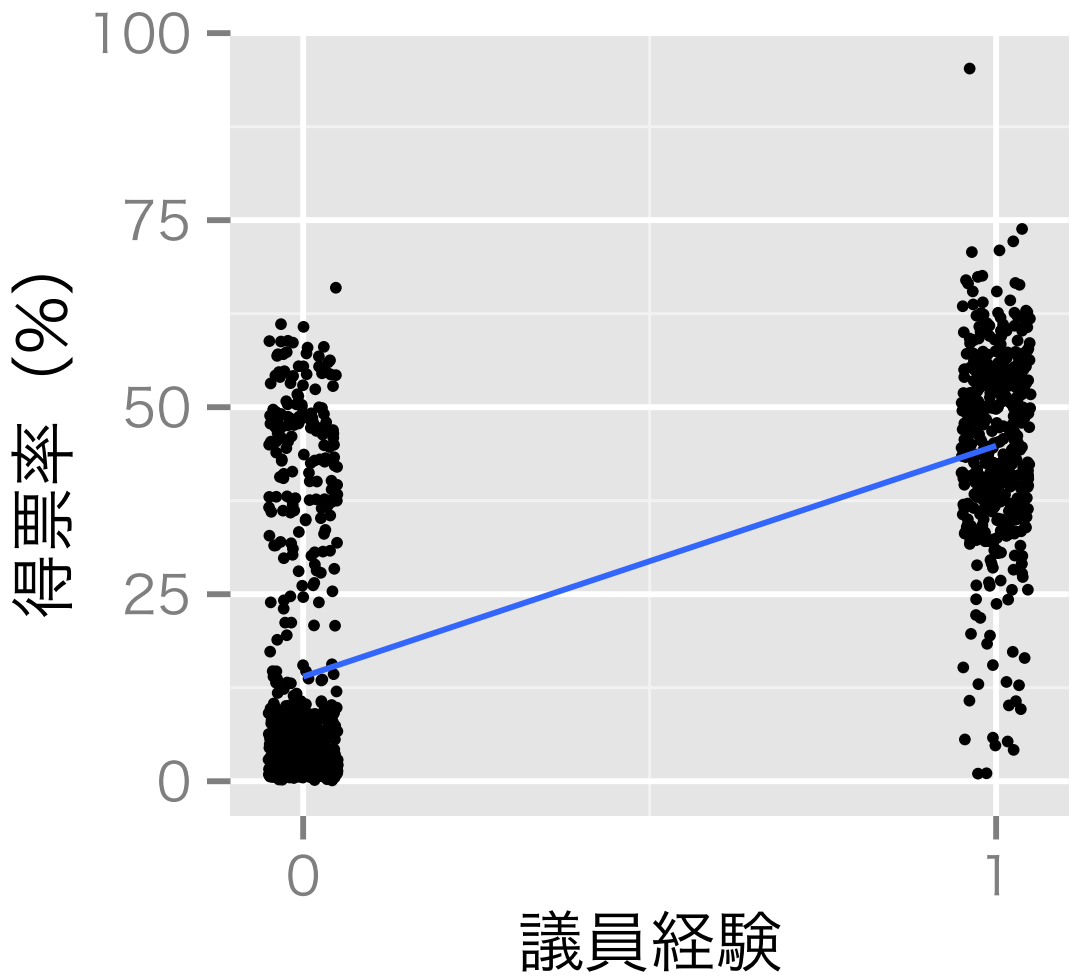


図: 議員経験の有無で得票率を説明する

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

- 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

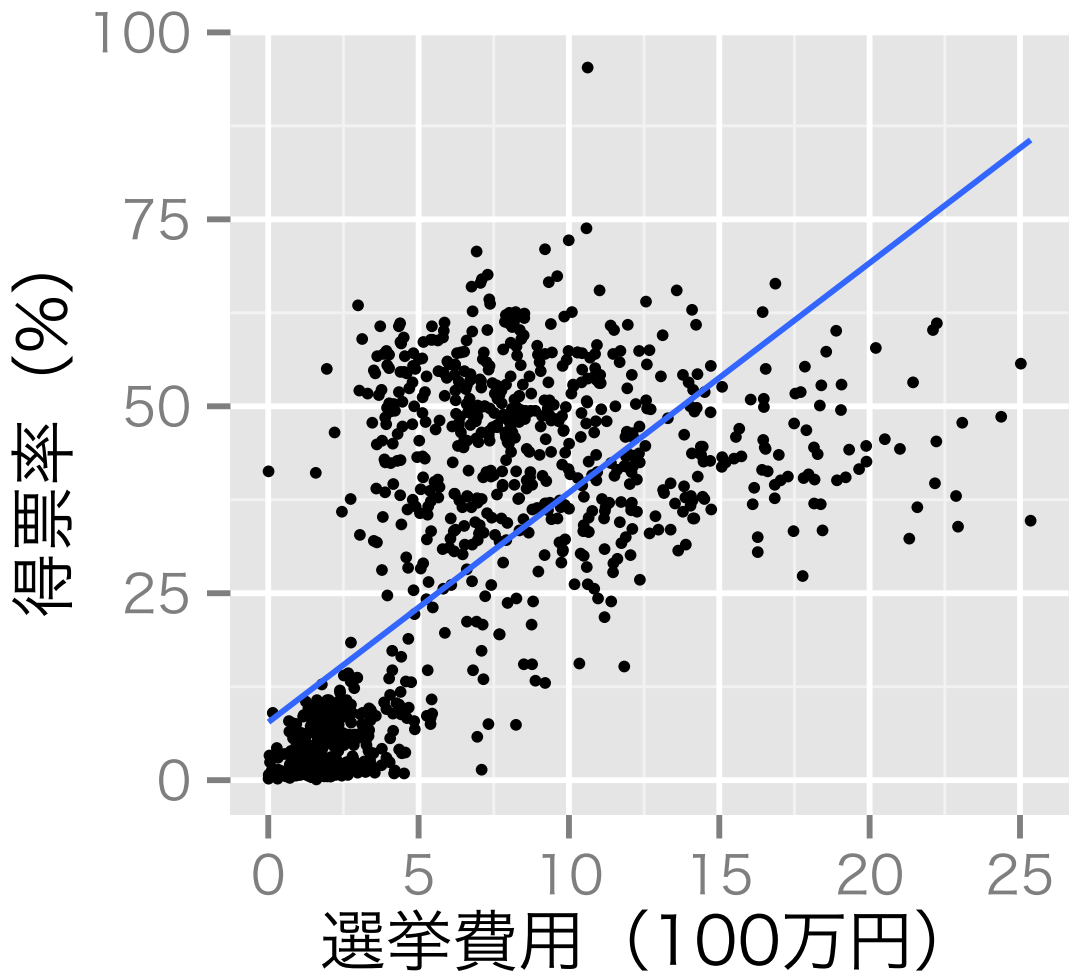


図: 選挙費用で得票率を説明する

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

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

- 応答変数：得票率 (%)
- 説明変数 1：議員経験（なし=0, あり=1）
- 説明変数 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

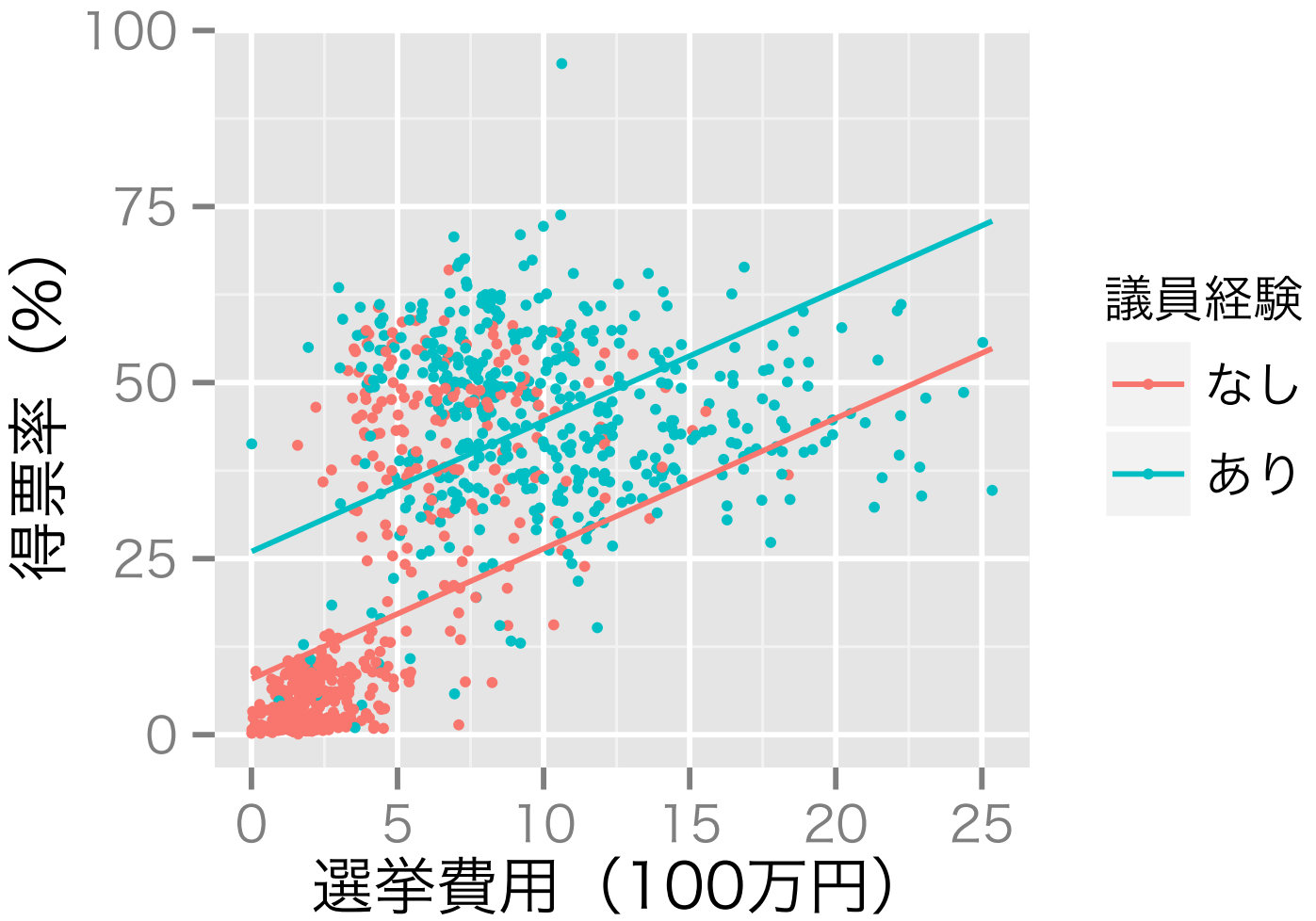


図: 議員経験の有無と選挙費用で得票率を説明する

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）
- 「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
  - 応答変数：得票率 (%)
  - 説明変数 1：議員経験（なし=0, あり=1）
  - 説明変数 2：選挙費用（測定単位：100 万円）
  - 説明変数 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

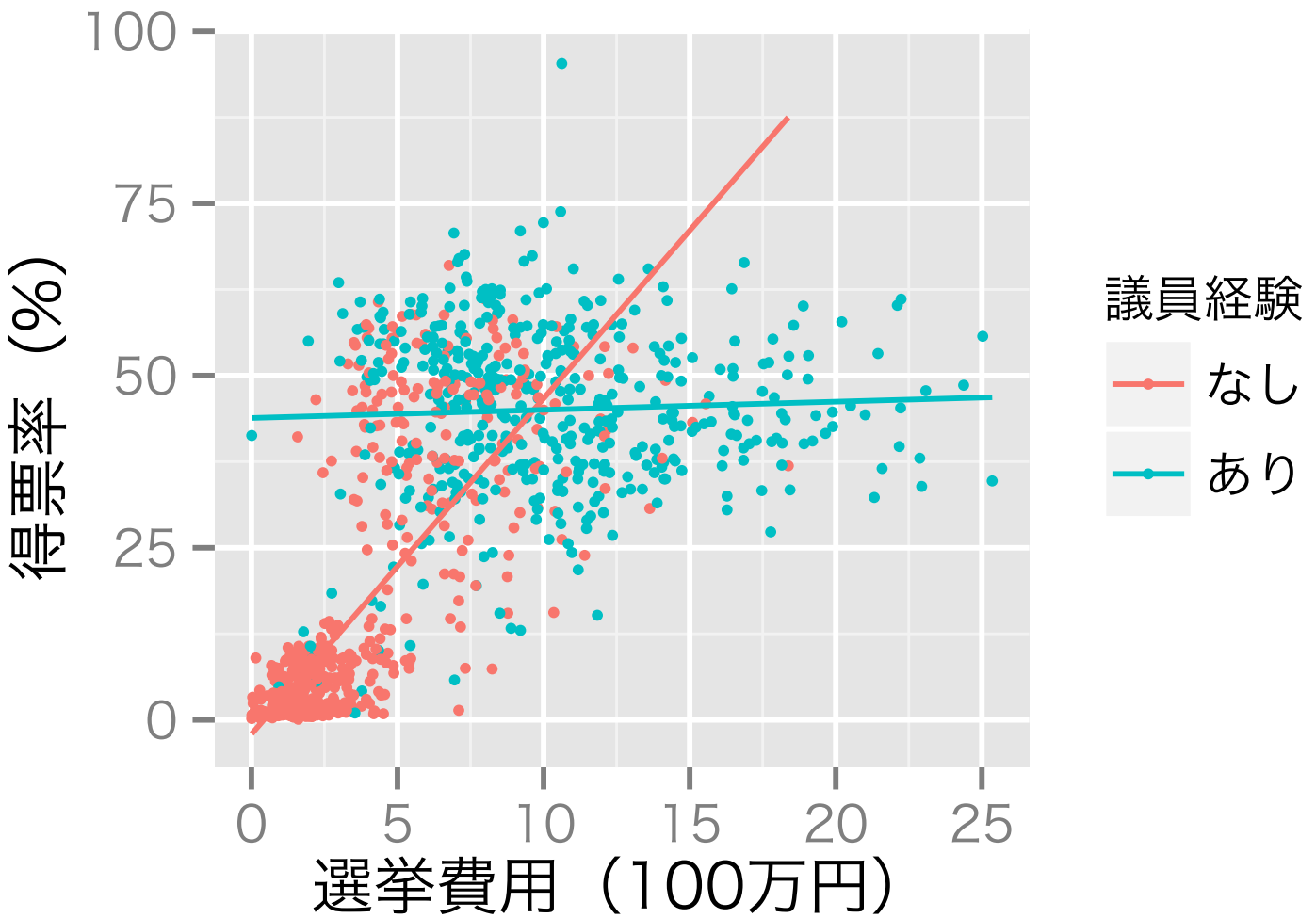


図: 議員経験の有無と選挙費用で得票率を説明する

モデル 4 の意味：各推定値の意味

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

- 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！

モデル 4 の意味：場合分けして考える

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

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：

$$\begin{aligned} \widehat{\text{得票率}} &= -2.1 + 45.9 \cdot 1 + 4.9 \cdot \text{選挙費用} - 4.8 \cdot 1 \cdot \text{選挙費用} \\ &= -2.1 + 45.9 + 4.9 \cdot \text{選挙費用} - 4.1 \cdot \text{選挙費用} \\ &= 43.8 + 0.1 \cdot \text{選挙費用} \end{aligned}$$

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：$i = 1, 2, \ldots, n$
- 予測値：$X_i\beta = \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$
- 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：$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$
- 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

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n \end{aligned}$$

- 表現法 2

$$\begin{aligned} y_i &\sim \mathrm{N}(X_i\beta, \sigma^2), \quad \text{for } i = 1, 2, \ldots, n \\ &\text{or} \\ y &\sim \mathrm{N}(X\beta, \sigma^2 I) \end{aligned}$$

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 パッケージの `display()` 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：$y = X\beta + \epsilon, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$
- 誤差を最小にする β を見つけたい：誤差の平方和

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

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

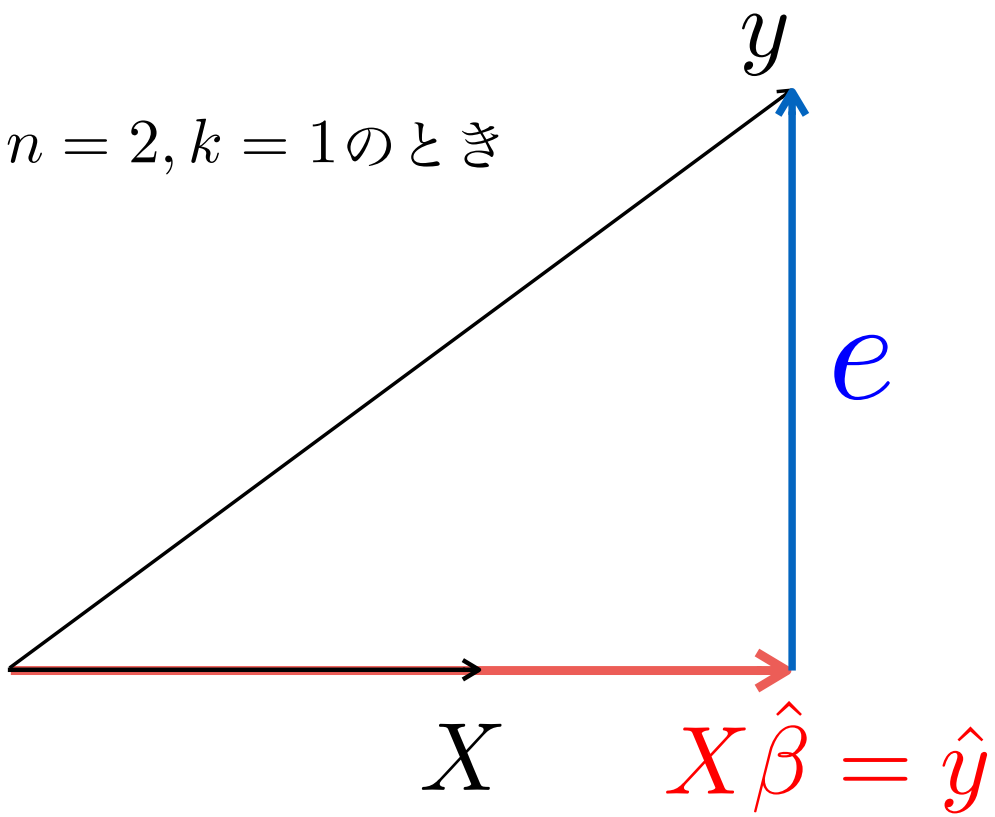
$$\sum_{i=1}^{n} e_i^2 = e'e = (y - X\hat{\beta})'(y - X\hat{\beta})$$

を最小にする $\hat{\beta}$ を見つける

$\rightarrow y$ 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, $\quad m = 1, 2, \ldots, k$
  よって、

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

- $(X'X)^{-1}$ 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！

係数の標準誤差 (standard errors)（1）

- 標準誤差 (se)：推定の不確実性を示す
- 係数の推定値の信頼区間を求めるために利用される
- `display()` で R の `lm` オブジェクトを表示すると、coef.se (coefficient standard errors) として表示される
- $\hat{\beta} \pm 2\mathrm{se}$ の範囲にある数は、データと整合的な推定値と考えられる
- $\frac{\hat{\beta}-\beta}{\mathrm{se}}$ は、自由度 $n-k$ の t 分布に従う (k には定数項を含む)
  （n が大きいときは標準正規分布で近似できるが、R に t 分布の関数が用意されているので、近似する必要はない）

係数の標準誤差 (standard errors)（2）

- 推定値の不確実性は互いに相関
- $V_\beta = (X'X)^{-1}$ とおくと、推定値の分散・共分散行列は $V_\beta \hat{\sigma}^2$
  - 行列の対角要素：各係数の推定値の分散 $\sqrt{V_{\beta ii}}\hat{\sigma}$：$\hat{\beta}_i$ の標準誤差
  - 行列の非対角要素：対応する 2 つの係数の推定値の共分散 $\frac{V_{\beta ij}}{\sqrt{V_{\beta ii} V_{\beta jj}}}$：$\hat{\beta}_i$ と $\hat{\beta}_j$ の相関

統計的有意 (statistical significance)

有意水準を 0.05 に設定すると

- おおよそ $\hat{\beta}_i \pm 2\mathrm{se}$ の範囲に 0 が含まれないとき、その係数をもつ説明変数の効果の向き（正か負か）がはっきりする
- そのとき、その効果は「統計的に有意である」とされる
- 統計的有意は、効果の向きをはっきりさせるだけで、効果の大きさについては何も示さない
- 実際の研究では効果の大きさ（実質的に意味があるのか, substantive significance）を示すことが必要かつ重要
- 有意水準を 0.05 にしなければいけない理論的論拠はない

  $\rightarrow$ 「$p < 0.05$ だから良い結果」とは限らない！！！

p 値 (p values)

- R で `lm` オブジェクトに対して `summary()` を使うと、“`Pr(>|t|)`” の列に表示される
- 説明変数が応答変数に影響を与えていない（$\beta = 0$、つまり、帰無仮説が正しい）ときに、現在分析中のデータを得る確率
  - 「帰無仮説が正しい確率」ではない!
  - 「対立仮説が間違っている確率」ではない!
- p 値（帰無仮説が正しいと仮定して計算）が小さい
  $\rightarrow$ 分析中のデータを得る確率は低い（にも拘らず、現にデータを持っている）
  $\rightarrow$ 帰無仮説が間違っていると考えることにする（帰無仮説を棄却する）
- 「p 値 $=$ 有意水準」ではない!：p 値はデータから計算するもの、有意水準は自分で（恣意的に）決めるもの

信頼区間 (confidence intervals)

- 自由度 $n-k$ の t 分布の 100α パーセンタイルを $t_{\alpha,n-k}$ とする
- $\hat{\beta}$ の 95 パーセント信頼区間：

$$[\hat{\beta} - t_{.025,n-k} \cdot \mathrm{se},\ \hat{\beta} + t_{.975,n-k} \cdot \mathrm{se}]$$

- 95 パーセント信頼区間：同じ母集団から、同じ手続きでデータを取り、分析するという作業を繰り返し行ったとき、求めた信頼区間のうちの 95%は母数（β の真の値）を区間内に含む
  - 「この信頼区間に母数が含まれる確率が 95%」ではない!
  - 1 つの信頼区間に母数が含まれる確率は、0 か 1 のいずれか
- シミュレーションで理解しよう！（ウェブ参照）

来週の内容

線形回帰分析（2）

- 線形回帰の前提と回帰診断
- 変数変換
  - 線形変換
  - 標準化
  - 中心化
  - 対数変換
- 結果の報告法
  - 何を報告するか
  - どのように報告するか