政治学方法論 I

第 7 回：線形回帰分析（2）

矢内 勇生

神戸大学　法学部/法学研究科

2014 年 11 月 12 日

今日の内容

最小二乗法による回帰分析の仮定

1. 推定しているモデルの妥当性
2. 線形性：予測値は説明変数の線形関数である
3. 誤差の独立性：$\mathrm{Cor}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
4. 誤差の分散均一性：$\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i
5. 誤差の正規性：$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$

仮定そのものを確かめることはできない！

残差プロットでモデルが捉え損ねた変化を示す

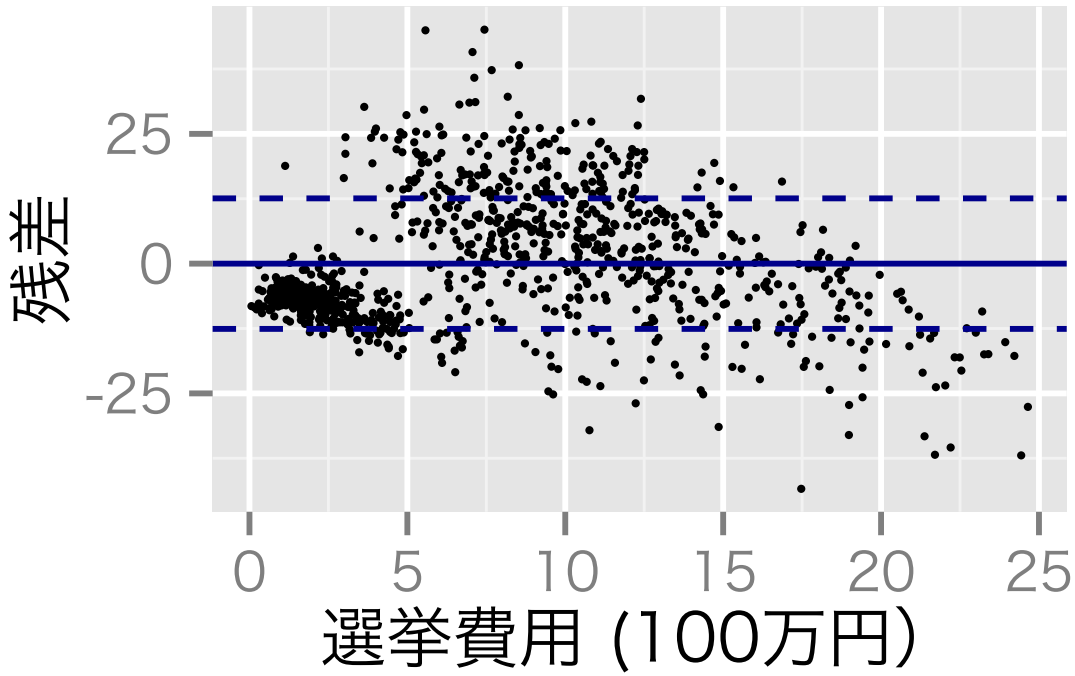

図: 得票率を選挙費用で説明するモデルの残差プロット．赤い点線は ± 1 標準偏差．残差プロットに何らかのパタンが見られるのは、モデルの欠点があるサイン

その他の回帰診断

- 外れ値の検討：スチューデント化残差 (Studentized residuals)
- テコ比 (leverage, hat values) の検討
- 観測値の影響力の検討：dfbeta, クック距離 (Cook's distance)
- etc.
- 回帰診断の詳細は、以下の本を参考に
  - Cook, R. Dennis, and Sanford Weisberg. 1999. *Applied Regression Including Computing and Graphics.* New York: John Wiley: 334–369.
  - Fox, John. 1997. *Applied Regression, Linear Models, and Related Methods.* Thousand Oaks: SAGE:267–366.

線形変換（linear transformations）

- 回帰式をより解釈しすいものにするために、変数を変換する
- 1 次関数で変換する
- 回帰式の実質的内容は変化しない

測定単位の変更（scaling）

選挙費用で得票率を求める回帰式は、以下のように表せる

1. 選挙費用の測定単位が 100 万円のとき（モデル 2）

$$\text{得票率} = 7.7 + 3.1 \cdot \text{選挙費用 (100 万円)} + \text{誤差}$$

2. 選挙費用の測定単位が 1 円のとき（モデル 2'）

$$\text{得票率} = 7.7 + 0.0000031 \cdot \text{選挙費用 (1 円)} + \text{誤差}$$

- 一見すると、1 のほうが 2 よりも選挙費用の効果が大きく見える
- 実際には、上の 2 つの回帰式はまったく同じ内容を示す
- ただし、解釈の難度が違う：どちらがわかりやすい？

z 値による標準化

- 変数の z 値（z 得点）を使って回帰分析を行うこともできる
- 変数 x の z 値は、

$$z = \frac{x - \bar{x}}{u_x} = \frac{x - x \text{の平均値}}{x \text{の不偏標準偏差}}$$

- すべての説明変数を z で標準化する：
  - 回帰係数：他の説明変数の値を一定に保ち、注目する説明変数の値を 1 標準偏差分大きくしたとき、応答変数が何単位分大きくなるか
  - 切片：すべての説明変数がそれぞれの平均値をとったときの応答変数の予測値

その他の標準化

- 単位を変えるのも標準化の 1 種（e.g. 1 円 $\rightarrow$100 万円）
- その他の例：7 点尺度である問題に賛成か反対か尋ねる
  - 1 点：強い反対 $\cdots$ 7 点：強い賛成 $\rightarrow$ 回帰係数の解釈が難しい
  - 標準化する：

$$\frac{\text{得点} - 4}{3}$$

$\rightarrow -1$ 点：強い反対、0 点：中立、1 点：強い賛成

  - 回帰係数：強い反対と中立の差、中立と強い賛成の差

回帰式の切片の解釈

- 切片の値：すべての説明変数が 0 のときの応答変数の予測値
- 0 を取らない説明変数があるとき $\rightarrow$ 実質的な意味なし
- 0 が最小値または最大値のとき $\rightarrow$ データの「端」に注目してしまう

$\downarrow$

- 説明変数を中心化する （線形変換の一種なので、回帰式の実質的な内容は変化しない）

中心化 (centering)

1. 標本平均を使って中心化する

$$c.x = x - \bar{x}$$

2. 基礎知識や慣習を使って中心化する
   - 例 1. 女性ダミーの中心化：男女比が 1 対 1 だとする

$$\mathrm{c.female} = \mathrm{female} - 0.5$$

   - 例 2. 知能指数 (IQ) の中心化：平均は 100 になるはず

$$\mathrm{c.IQ} = \mathrm{IQ} - 100$$

すべての説明変数が中心化された回帰式の切片：

すべての説明変数が平均（またはその他の中心）の値をとったときの応答変数の予測値（平均値）

標準化した変数による単回帰

標準化された変数 x と y の単回帰 $y = a + bx + \epsilon$

$$
\begin{aligned}
x &= \frac{x_{\text{raw}} - \bar{x}_{\text{raw}}}{u(x_{\text{raw}})} \\
y &= \frac{y_{\text{raw}} - \bar{y}_{\text{raw}}}{u(y_{\text{raw}})}
\end{aligned}
$$

- 切片 $a = 0$
- 傾き $b \in [-1, 1]$：x と y の相関係数：

$$|b| > 1 \Rightarrow \sigma_y > \sigma_x$$

相関係数と単回帰の回帰係数

一般的な単回帰（標準化されていない場合も含む）を考える

- x と y の共分散を σ_{xy} とする
- x と y の相関係数 ρ：

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- 回帰式の傾き b：

$$b = \rho \frac{\sigma_y}{\sigma_x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

主成分直線 (principal components line) と回帰直線

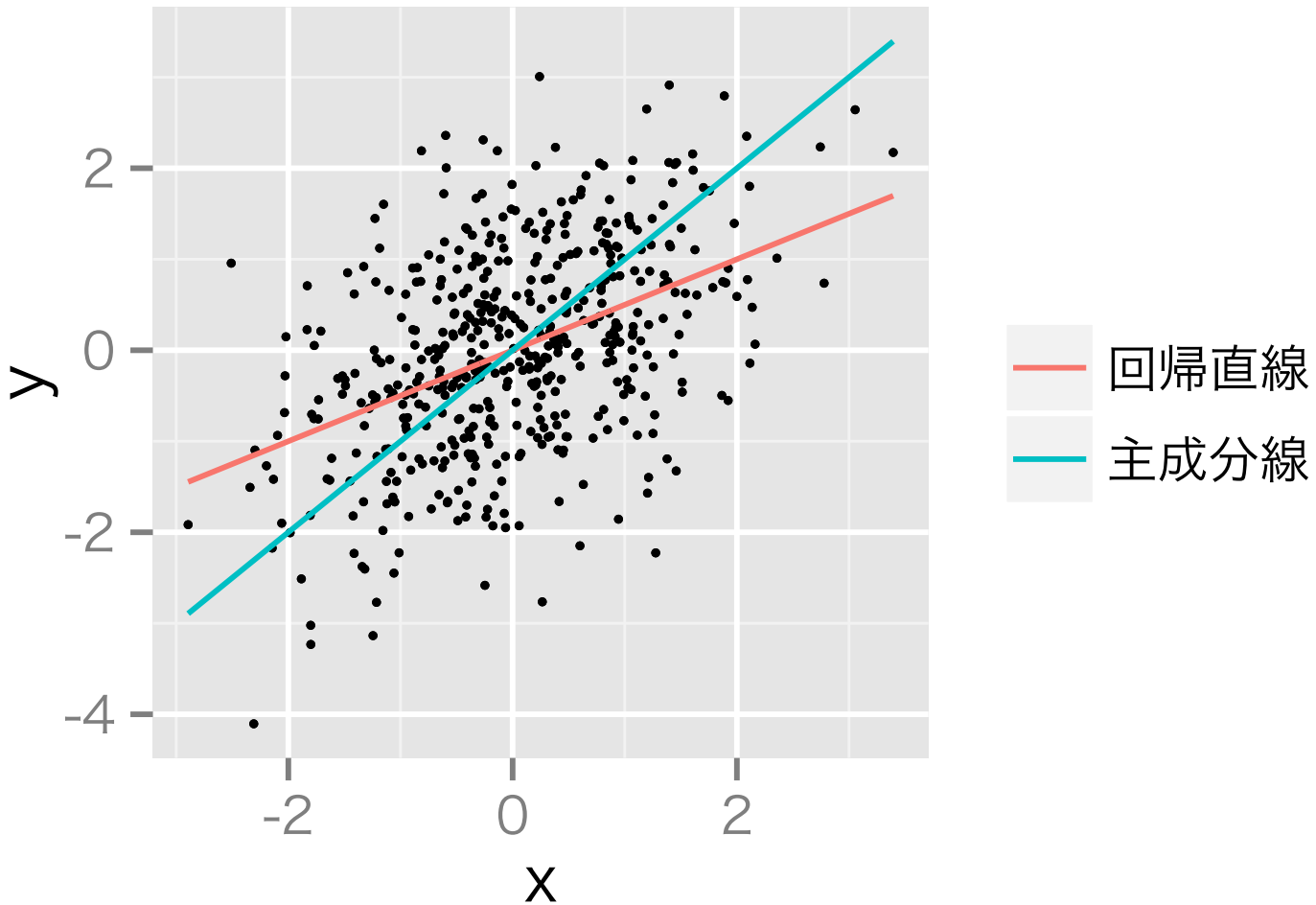

図: 標準化された x と y の関係：$\rho = 0.5$

平均への回帰 (regression to the mean)

主成分直線と回帰直線を比較する

- 主成分直線
  - x が小さいときの y の予測が過小
  - x が大きいときの y の予測が過大
- 回帰直線：どの x の周辺でも、データの中心を予測
- 平均への回帰：標準偏差で測ったとき、

  $\hat{y}$ と $\bar{y}$ の距離 $<$ x と $\bar{x}$ の距離
  - 「どんな変数も次第に平均に近づく」とは言っていない
  - 予測値の平均値からの乖離は、説明変数の平均値からの乖離より小さい（割り引いて考える）ということ

対数 (logarithm)

- 対数：指数関数の逆関数
- $x = a^p$ のとき、p を「a を底とする x の対数」と呼び、$p = \log_a x$ と書く
- 定義域：$x > 0$
- 例：底が 10 の対数
  - x が $1, 10, 100, \cdots = 10^0, 10^1, 10^2, \ldots$ と増えるとき
  - 対数は $0, 1, 2, \ldots$ と増える

  $\rightarrow$ スケールを変更して考えられる：大きな数を扱う（桁の違いに意味がある）ときに有効
- よく使われる対数の底：e（ネイピア数）– 結果が分かりやすいから（e^p を $\exp(p)$ と書く）

x の対数

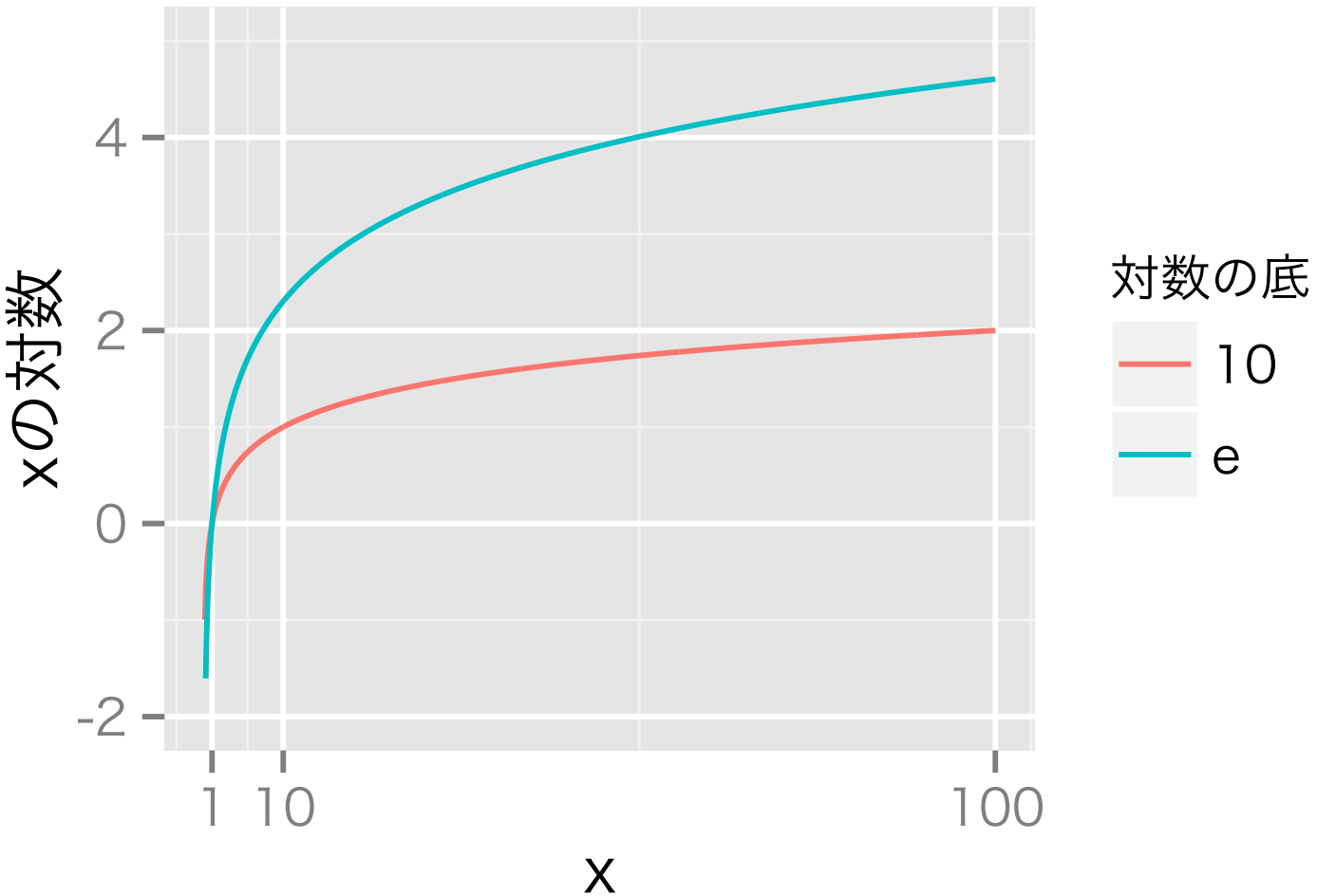

図: $\log_e x$ と $\log_{10} x$

対数変換の効果

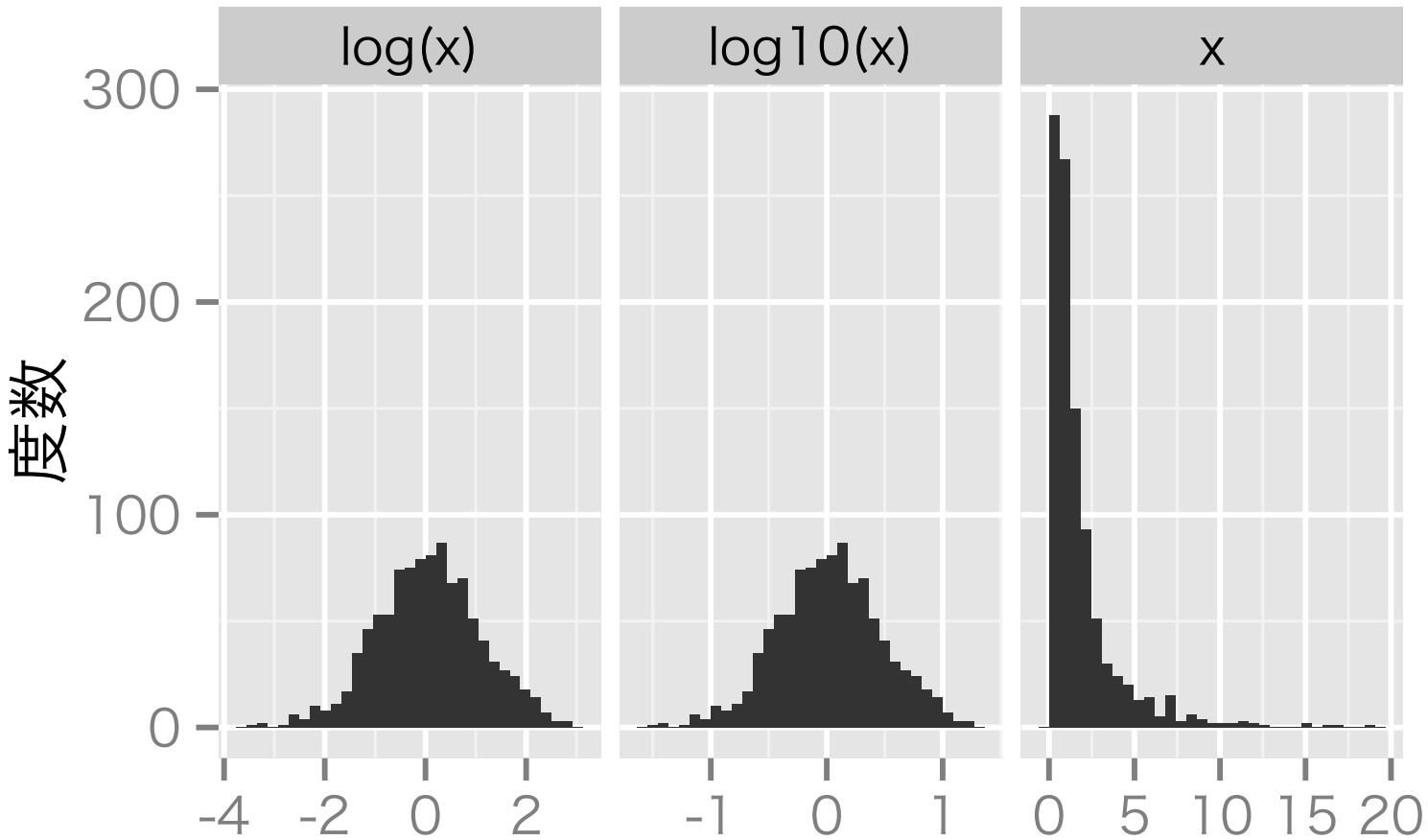

図: log(x) ($= \log_e x$), log10(x) ($= \log_{10} x$), x の分布

自然対数：底が e の対数

- x の自然対数：$\log_e(x) \rightarrow$ 単に $\log(x)$ と書く
- 自然対数を使う理由：結果がわかりやすい
- 例：応答変数が自然対数のとき

$$\log y_i = b_0 + 0.06 x_i + \epsilon_i$$

  - x が 1 単位増えると、$\log(y)$ は 0.06 単位増える
  - x が 1 単位増えると、y は $\exp(0.06) - 1 = 0.06$ 単位増える
  - $x1$ 単位の変化は y を約 6% (0.06) 増加させる
  - 係数 0.06：y の変化率（ただし、この近似が使えるの係数が 0 に近いときだけ）

変化率としての係数：応答変数が自然対数のとき

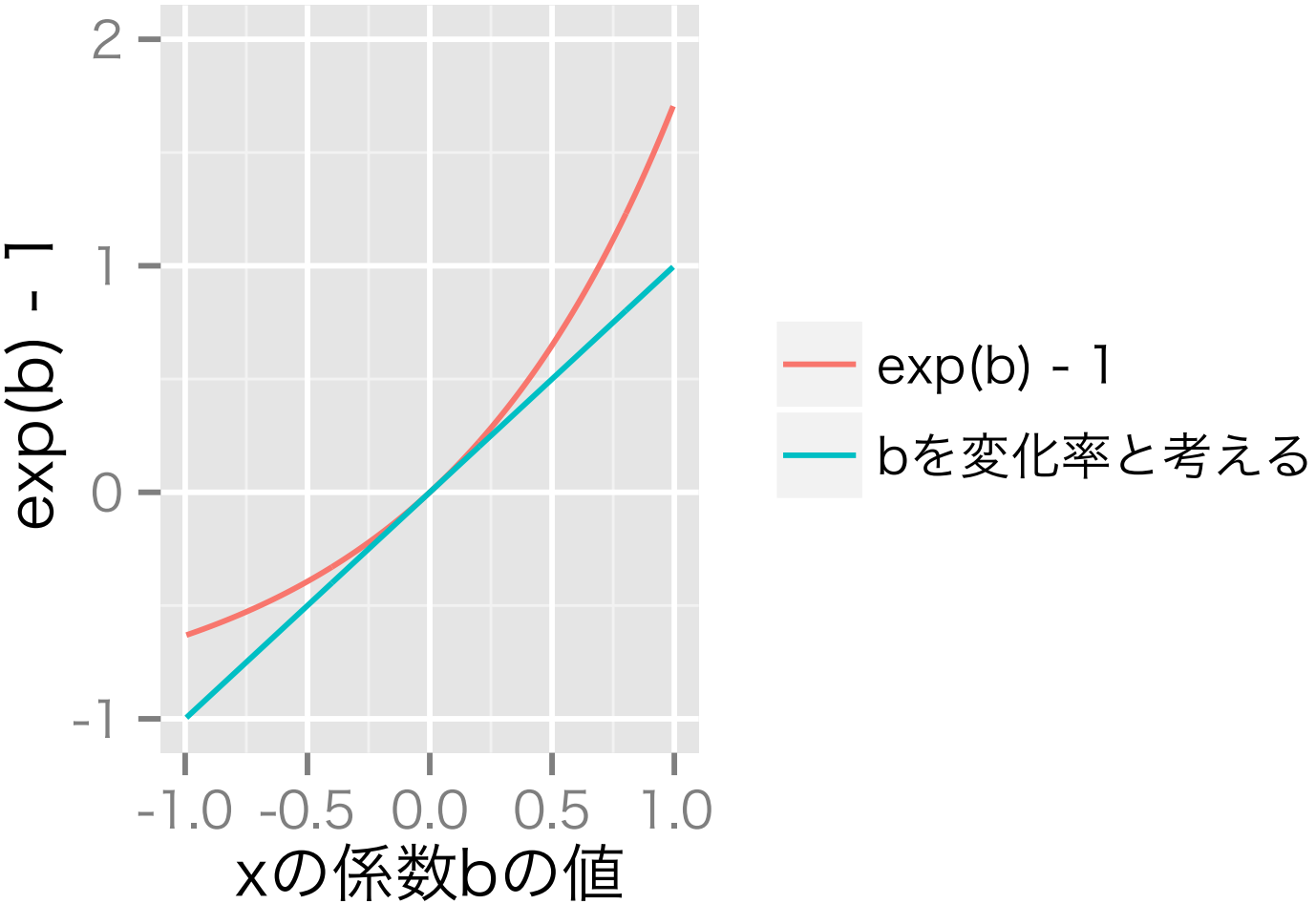

図: 係数を変化率と考える

自然対数と 10 を底とする対数

$$\log_{10} y_i = b_0 + 0.026 x_i + \epsilon_i$$

- x が 1 単位増えると、$\log_{10}(y)$ は 0.026 単位増える
- x が 1 単位増えると、y は $10^{0.026} - 1 = 0.06$ 単位だけ増える
- x1 単位の変化は y を約 6% (0.06) 増加させる
- 係数 0.026：このままでは y の変化率がわからない！

対数変換したモデルの解釈

応答変数	説明変数	係数 b の意味
無変換	無変換	説明変数が 1 単位増えると、応答変数は b だけ増える
無変換	自然対数	説明変数が 1%増えると、応答変数が b だけ増える
自然対数	無変換	説明変数が 1 単位増えると、応答変数が $100b$%増える
自然対数	自然対数	説明変数が 1%増えると、応答変数が $100b$%増える（弾力性）

注：b が 0 に近くないときは $\exp(b) - 1$ を計算する必要がある

参考：森田 (2014) 第 5 章; Tufte, E. (1974) *Data Analysis for Politics and Policy*: 108–134

最低限の報告内容

- 回帰モデル：式または文章で説明する
- 応答変数と説明変数（コントロール変数）の詳細な説明
- 回帰式の係数
- サンプルサイズと R^2 (R2 ではない！)
- 推定の不確実性を表す値（1 つ以上）
  - 係数の標準誤差 (se)
  - 係数の t 値
  - 係数の p 値
- 結果の実質的な意味の解説

実質的な意味を報告する

例：$\widehat{\text{得票率}} = 7.7 + 3.1 \cdot$ 選挙費用

- ダメな説明の例：「選挙費用の係数は 3.1 で、この効果は統計的に有意である。」
  - 「3.1」の意味は？
  - この結果をなぜ気にする必要があるの？
- 実質的な意味を文章で説明する
  - 「3.1」の意味：「選挙費用を 100 万円増やすごとに、得票率が 3.1 ポイントずつ上昇すると予測される。」
  - 実質的重要性：「選挙費用は、約 1 万円から 2500 万円の範囲に分布しており、標準偏差は約 500 万円である。標準偏差 1 つ分の選挙費用の変化（例えば、選挙費用を 500 万円から 1000 万円に増やすこと）は、得票率を約 15.5 ポイント上昇させる。得票率が 15.5 ポイント変わることは、選挙にとって重大な結果の変化である。$(50\text{-}7.7)/3.1 = 13.6$ だから、選挙費用を 1400 万円使えば、過半数票を確保し、選挙に勝つことが期待できる。」(あくまで説明のための例です)

結果を報告する方法

報告の仕方（スタイル）は状況によって様々

1. 文章 + 式
2. 文章 + 表：表に何を含める？
3. 文章 + 図
   - 散布図と回帰直線
   - キャタピラプロット

報告法を決める基準

- 誰に、何を伝えたいか
- モデルの複雑さ：1 つ（または少数）の図で内容を示せるか
- 相互作用がモデルに含まれるか
- 投稿する雑誌の規定
- etc.

式（+文章）で報告する：例

$$\widehat{\text{得票率}} = \underset{(0.69)}{7.91} + \underset{(1.23)}{18.10} \cdot \text{議員経験} + \underset{(0.12)}{1.85} \cdot \text{選挙費用}$$

（括弧内は標準誤差, $n = 1124$, 自由度調整済み $R^2 = 0.56$）

表（+文章）で報告する：例1

表: 最小二乗法による推定結果：応答変数は得票率

説明変数	係数の推定値	標準誤差
切片	7.91	0.69
議員経験	18.10	1.23
選挙費用	1.85	0.12
標本サイズ（n）	1124	
自由度調整済み R^2	0.56	

表（+文章）で報告する：例2

表: 最小二乗法による推定結果：応答変数は得票率 (%)

	係数の推定値	
説明変数	モデル 3	モデル 4
切片	7.91	−2.07
	(0.69)	(0.72)
議員経験	18.10	45.91
	(1.23)	(1.58)
選挙費用	1.85	4.87
	(0.12)	(0.16)
議員経験 × 選挙費用		−4.76
		(0.21)
標本サイズ（n）	1124	1124
自由度調整済み R^2	0.56	0.70

注：括弧内は標準誤差

図（+文章）で報告する：回帰直線を示す例

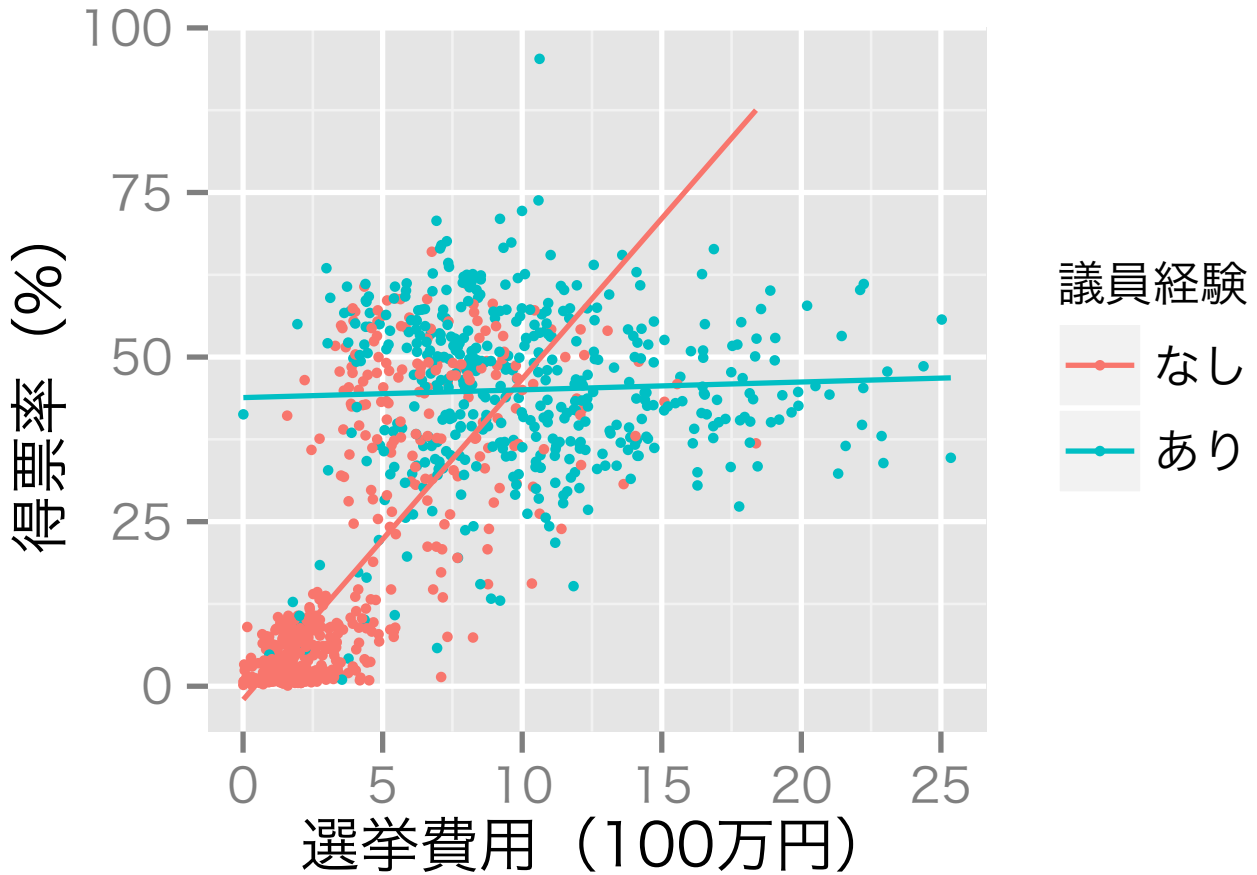

図: 選挙費用で得票率を説明する：議員経験によって回帰直線の傾きを変える

図（+文章）で報告する：キャタピラプロットを示す例

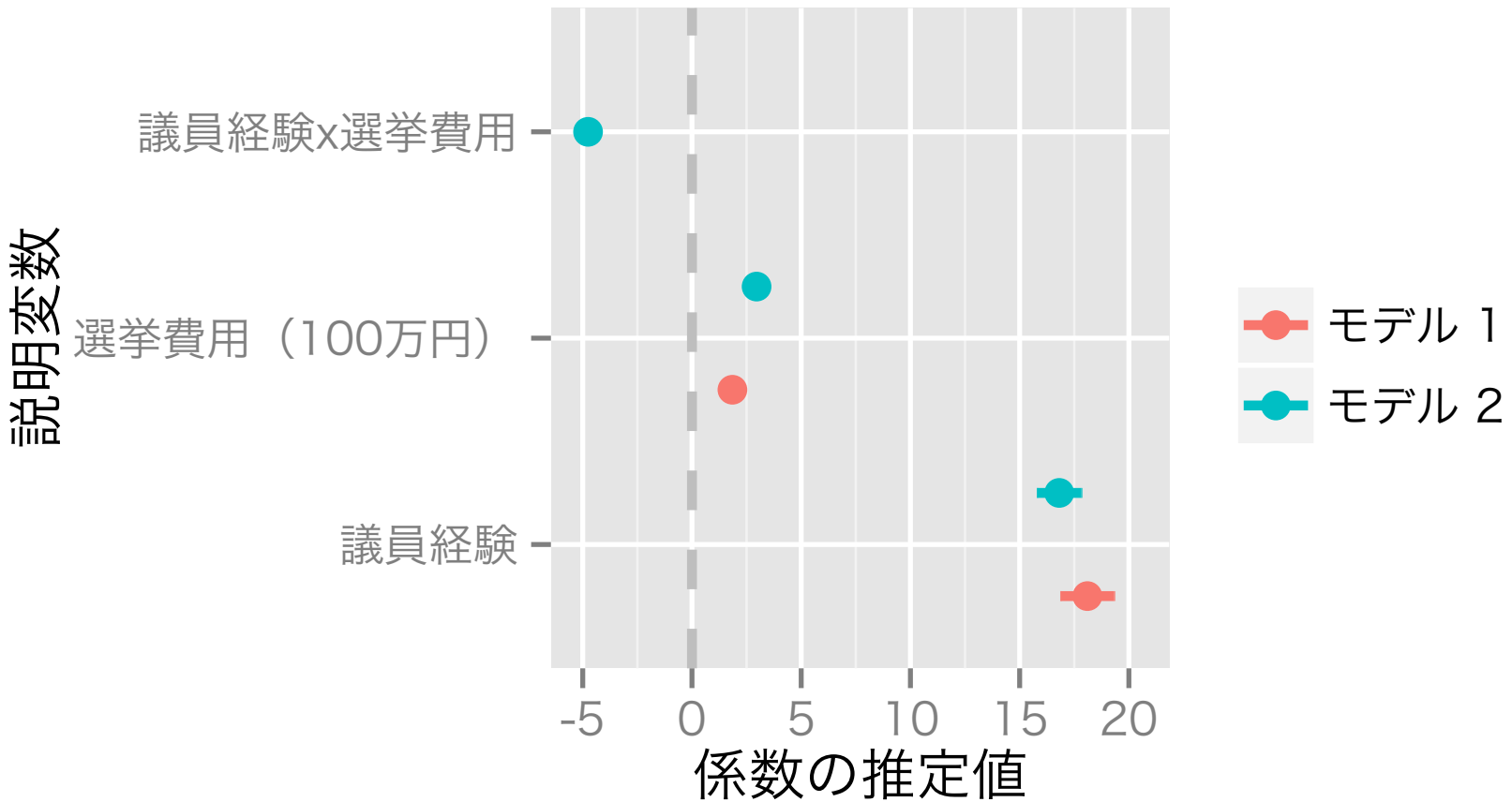

図: 係数の推定値と 95%信頼区間