計量経済学

12. 回帰分析の応用（続）と分析結果の提示法

矢内　勇生

2019年11月18日

高知工科大学　経済・マネジメント学群

今日の目標

- 回帰分析における応用的なテクニックを理解する
  - 変数を変換して使う
- 回帰分析結果の提示法を理解する

回帰分析の応用（続）

変数変換

線形変換 (Linear Transformation)

- 回帰式をより解釈しやすいものにするために、変数を変換する
- 1次関数を利用して変換する
  - 回帰式の実質的な意味は変わらない

測定単位の変更

- 選挙費用で得票率を説明する回帰式は、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

得票率 = 7.7 + 3.1・選挙費用 [100万円]+ 誤差

(2) 選挙費用の測定単位が1円のとき

得票率 = 7.7 + 0.0000031・選挙費用 [1円]+ 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
- しかし、実際には２つの式の意味は同じ
- 解釈の難度が違う：どちらがわかりやすい？

標準化

- 変数 x 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

$$z(x) = \frac{x - \bar{x}}{u_x} = \frac{x - x \text{の平均値}}{x \text{の不偏分散の平方根}}$$

- すべての説明変数を z 値で標準化する：

  - 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., 160cm -> 1.6m)
- その他の例：ある意見に賛成か反対かを7点尺度で尋ねる
  - 1点：強い反対，・・・，7点：強い賛成：回帰係数の解釈が難しい
  - 標準化する $\frac{\text{得点} - 4}{3}$
    - -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

スケーリングの方針

- どの単位で測ることに意味があるか？
  - 選挙費用が1円変化することの影響を議論する意味はあるか？
- 重回帰の場合：係数の値が変数ごとにあまりにも大きくばらつくことを避ける
  - １つの目安
    - 正の値しかとらない変数：0以上1以下の間に収める
    - 正負の値をとる変数：-1以上1以下の間に収める
  - ただし、結果を解釈するときに、元の測定単位が使えなくなることに注意

中心化 (1)

- 回帰式の切片の値：すべての説明変数の値が0のときの応答変数の予測値
  - 0をとらない説明変数があるとき：実質的な意味なし
  - 0が最小値または最大値のとき：データの「端」

★説明変数を中心化 (centering) する！

  - 線形変換の一種

中心化 (2)

- 標本平均を使った中心化

$$x_c = x - \bar{x}$$

- 基礎知識や慣習を使った中心化
  - 例1）女性ダミーの中心化：男女比が1対1だと仮定

$$\text{female}_c = \text{female} - 0.5$$

  - 例2）知能指数 (IQ) の中心化：平均は100のはず

$$\text{IQ}_c = \text{IQ} - 100$$

- 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 $z(x)$ と $z(y)$ の単回帰：

$$z(y_i) = s + tz(x_i) + e_i$$

$$z(x_i) = \frac{x_i - \bar{x}}{u_x}$$

$$z(y_i) = \frac{y_i - y}{u_y}$$

- 切片：$s = 0$

- 傾き：$t \in [-1,1] = x$ と yの相関係数

$$y_i = a + bx_i + e_i$$

$$|b| > 1 \Rightarrow u_y > u_x$$

相関係数と単回帰の回帰係数

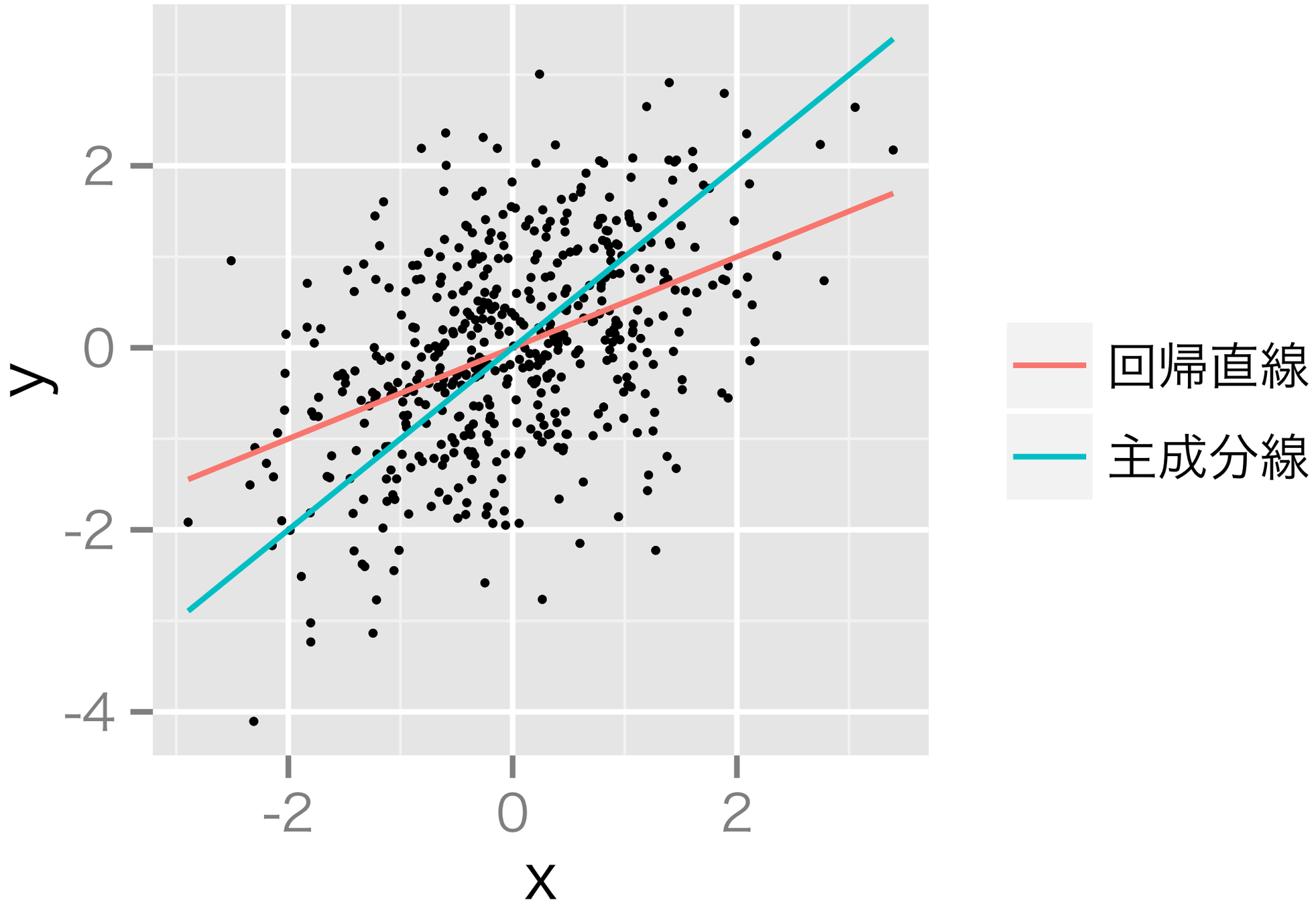
- 一般的な単回帰（標準化されていない場合）を考える
  - x と y の共分散を σ_{xy} とする
  - x と yの相関係数 ρ：

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

  - 回帰式の傾き b：

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

主成分直線と回帰直線



図：標準化された x と y の関係：相関係数 = 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、

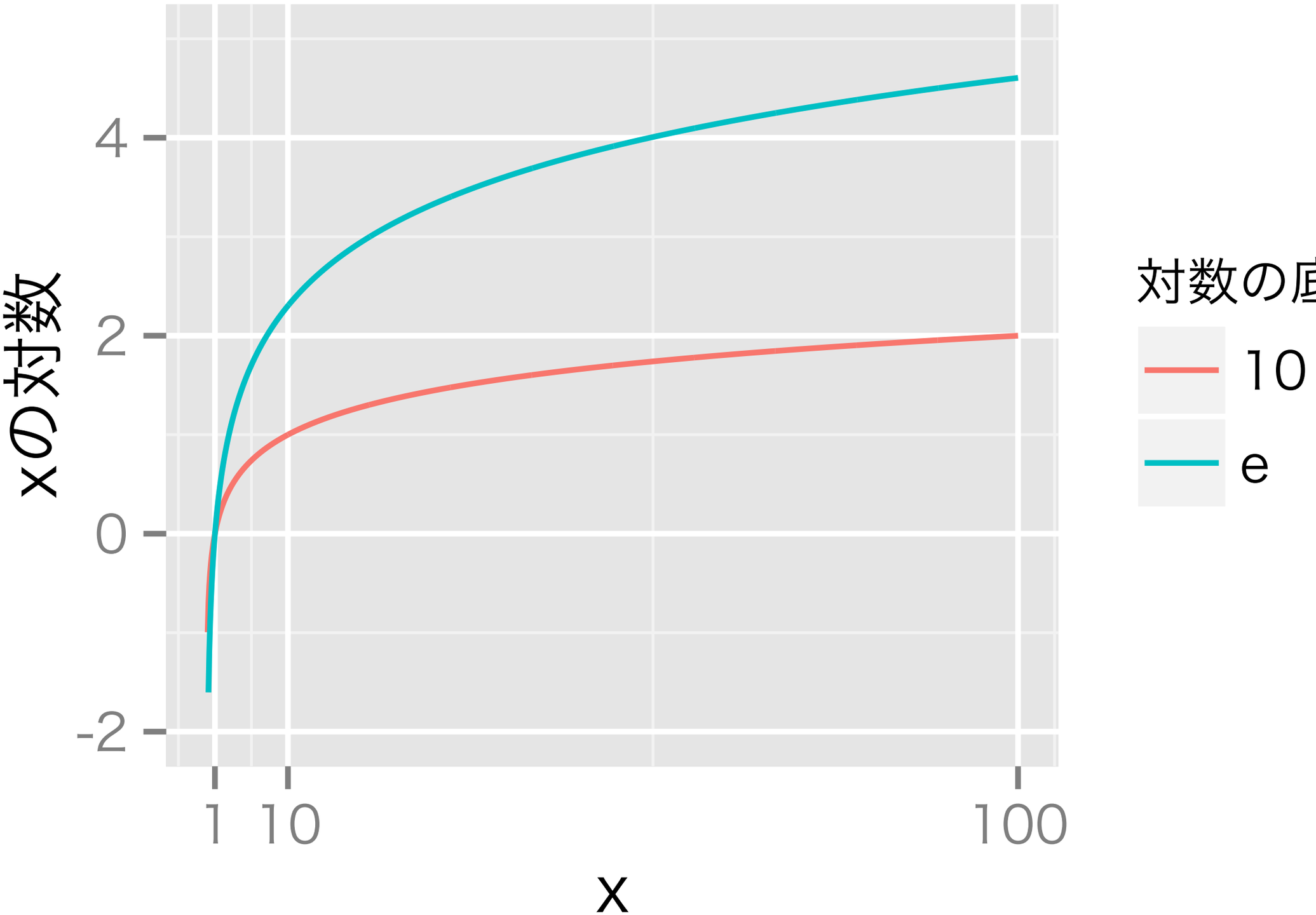
$$\hat{y}と\bar{y}の距離 < x\ と\bar{x}の距離$$

    - 「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

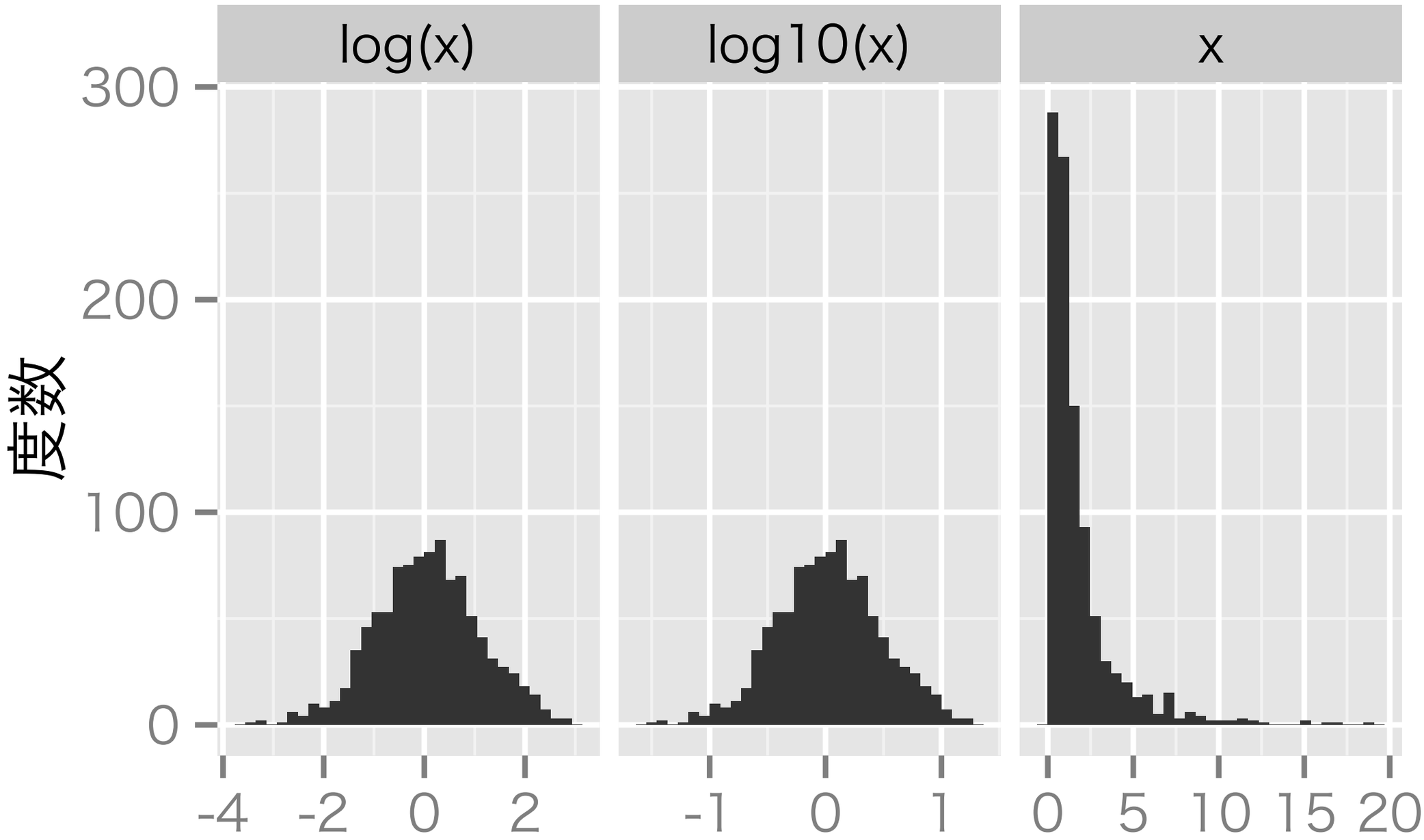
対数 (logarithm)

- 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 $1,10,100,\ldots = 10^0,10^1,10^2,\ldots$ と増えるとき、対数は $0,1,2,\ldots$ と増える
    - スケールを変更して考えられる！：大きな数を扱う（桁の違いに興味がある）ときに有効
- よく使われる対数の底：e（ネイピア数）
  - 結果がわかりやすいから
  - e^p を $\exp(p)$ と書く

x の対数



対数変換の効果



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

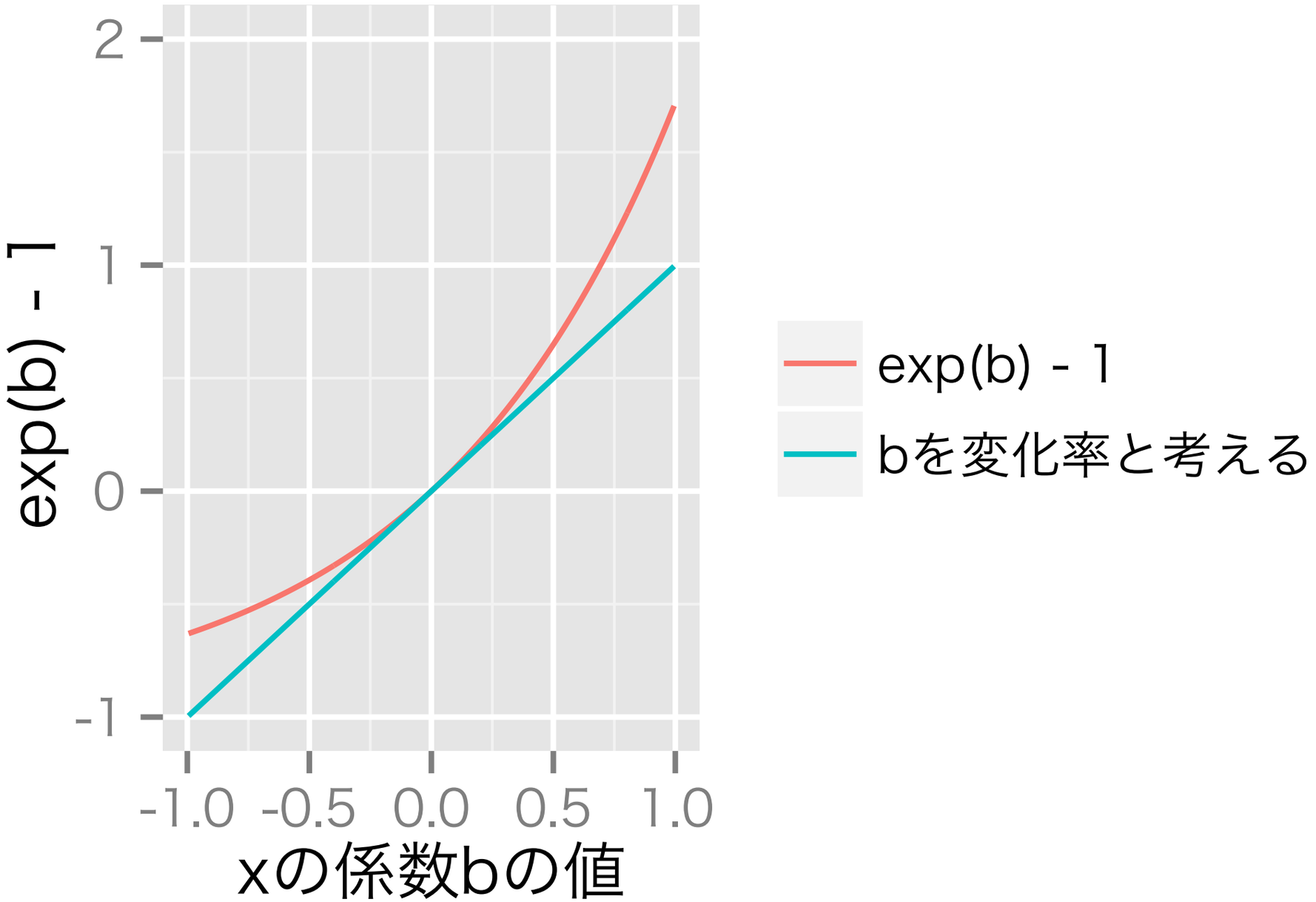
自然対数：底が e の対数

- x の自然対数：$\log_e(x)$ → 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

$$\log(y_i) = b_0 + 0.06x_i + e_i$$

  - 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 $\exp(0.06) - \exp(0) = \exp(0.06) - 1 = 0.06$ 単位増える
  - x の1単位分の増加は、y を約6%（つまり、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は、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：係数が0に近いときは、係数を変化率と考える

自然対数と10を底とする対数

$$\log_{10}(y_i) = b_0 + 0.026x_i + e_i$$

- 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 $10^{0.026} - 10^0 = 10^{0.026} - 1 = 0.06$ 単位だけ増える
- x の1単位分の増加は、y を約6%（つまり、0.06）増加させる
- 係数 0.026：このままでは、y の変化率はわからない！

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

分析結果をどうやって伝えるか？

最低限の報告内容

- 回帰モデル：式または文章
- 応答変数と説明変数（交絡を含む）の詳細な説明
- 回帰式の推定結果：（切片と）係数
  - サンプルサイズ（観測数）と R^2（R2 ではない）
- 係数の信頼区間
- 推定の不確実性を表す値（1つ以上）：標準誤差がベスト
- **結果の実質的な意味**の解釈・解説

回帰分析の結果の提示

- 図、表または式の形で表す
- 係数だけでなく、不確実性（標準誤差, t値 [検定統計量],またはp値）も一緒に示すことが必要
  - どの不確実性指標を使っているかはっきり示すこと！
- **点推定値と信頼区間を図示するのが現代の常識！**
- 観測数（標本サイズ）と決定係数（重回帰の場合は自由度調整済み決定係数）も示す
- Rのsummary() または broom::tidy() の結果をそのままコピペしない！
  - 読みやすい、綺麗な表が必要

結果提示の例：式の場合

身長=107.2 + 0.19 × 父の身長 + 0.21×母の身長

(4.93)　(0.02)　　(0.02)

注：括弧内は標準誤差

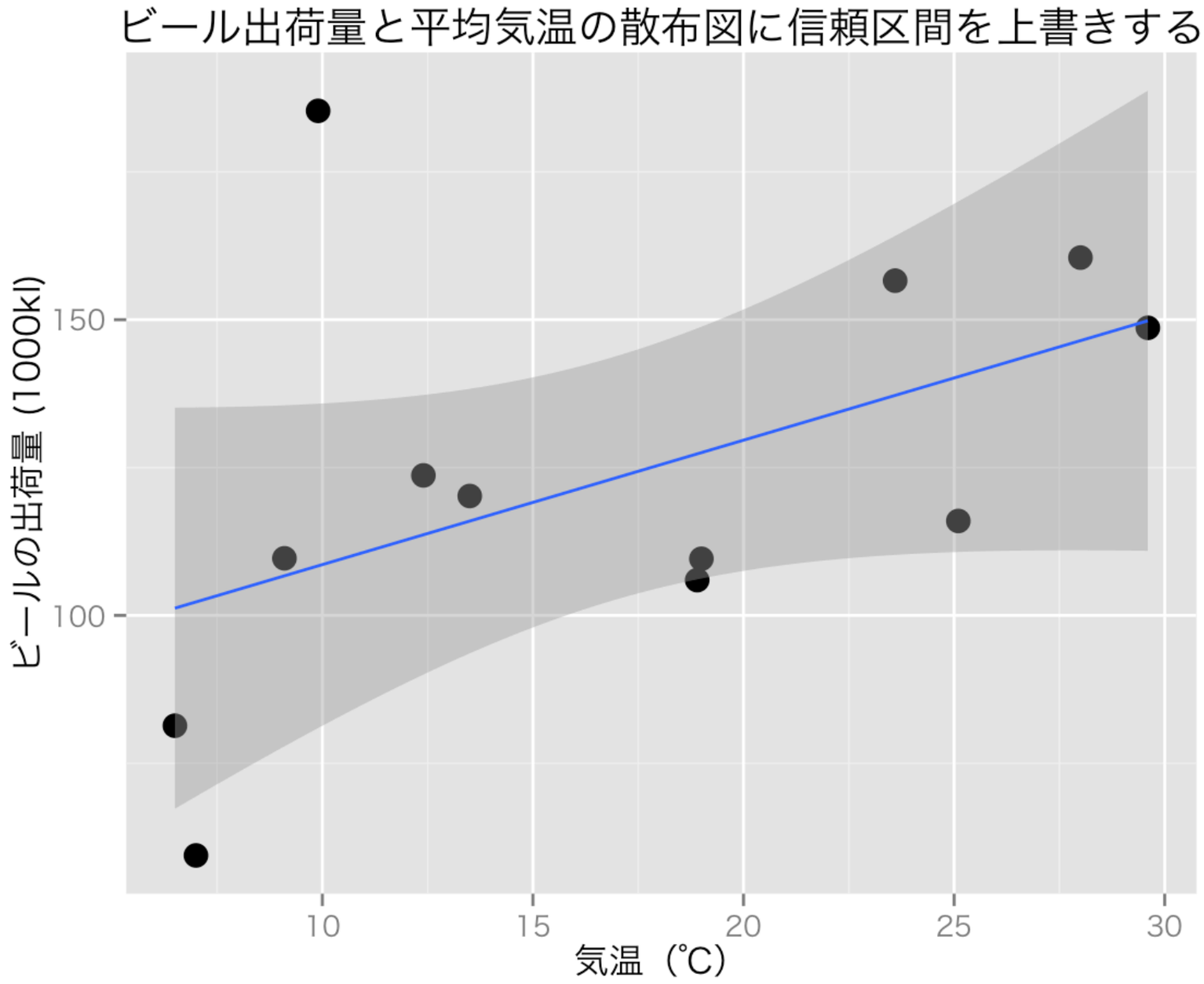
- 括弧内に*p*値を書けば、ある有意水準の下で棄却されるか受容されるかが一目でわかる：しかし、**標準誤差 (se) がオススメ**
- 標準誤差が書かれている場合の目安：有意水準5%なら、係数÷SE の値が2以上なら帰無仮説を棄却
- t値（検定統計量）を書いても理論的には問題ないが、臨界値を求めないと棄却か受容か判断できないので、あまり好まれない

結果提示の例：表の場合

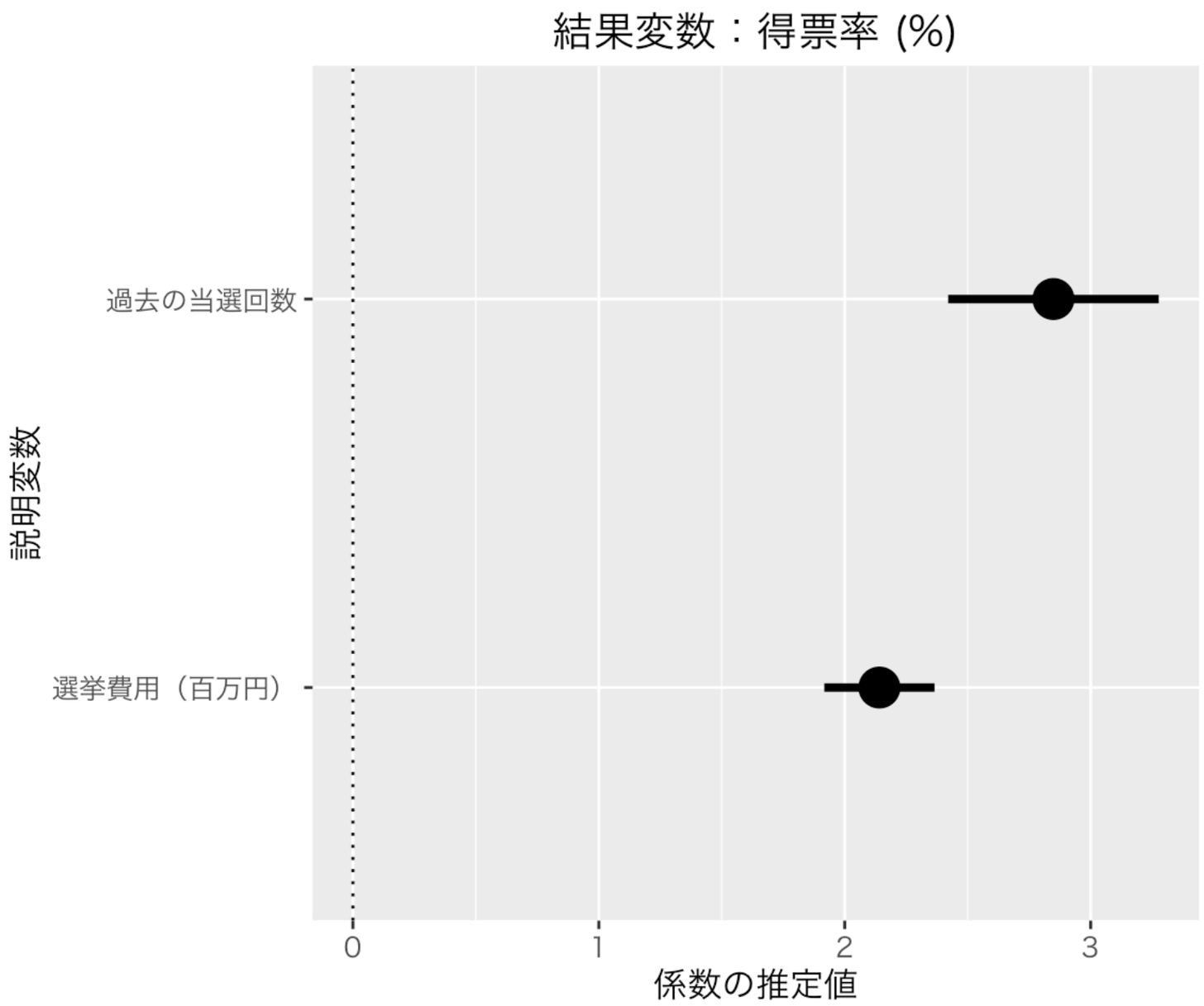
表1. 回帰分析の結果（結果変数は自民党の得票率）

説明変数	推定値	標準誤差	95%信頼区間 下限	95%信頼区間 上限	*p*値
説明変数1	−0.10	0.37	−0.85	0.65	0.79
説明変数2	0.07	0.46	−0.86	0.99	0.89
説明変数3	1.68	0.27	1.14	2.22	0.00
説明変数4	0.77	0.05	0.67	0.87	0.00
説明変数5	0.25	0.35	−0.45	0.95	0.47
説明変数6	42.15	0.33	41.48	42.83	0.00
観測数	47				
自由度調整済み決定係数	0.88				
F 統計量	66.11				
自由度 (5, 41)					

結果提示の例：単回帰の図示



結果提示の例：重回帰の図示



注：点は係数の推定値、線分は95%信頼区間を表す。