計量経済学応用

9. 回帰分析 (5)

矢内　勇生

2018年5月14日

高知工科大学　経済・マネジメント学群

今日の目標

- 回帰分析で変数変換を使う！
  - 線形変換
  - 中心化
  - 平均への回帰
  - 対数変換

変数の中心化（1）

- 切片 = すべての説明変数が０のときの結果変数の予測値

➡実質的な意味が分からないことがある

例：親の身長で子の身長を説明

- 切片 = 親の身長が0のときの子の身長

➡説明変数を中心化して、切片を意味あるものにする

測定単位の変更 (scaling)

- 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. 得票率 = 7.7 + 3.1 選挙費用（100万円）+ 誤差

2. 得票率 = 7.7 + 0.0000031 選挙費用（円）+ 誤差

- 一見すると、1の方が2よりも選挙費用の効果が大きく見える
- 実際には、2つの式の内容はまったく同じ
- 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

- 変数 x の z値は、

$$z_x = \frac{x - \bar{x}}{u_x} = \frac{x - x\text{の平均値}}{\sqrt{x\text{の不偏分散}}}$$

- 全ての説明変数を z で標準化すると：

  - 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、結果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結果変数の予測値

その他の標準化

- 単位を変えるのも標準化の一種 (e.g., 1円 -> 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　・・・　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

回帰式の切片の解釈

- 切片の値：すべての説明変数の値が0のときの結果変数の予測値

- 0を取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」に注目してしまう

- 説明変数を中心化しよう！

中心化 (centering)

- 標本平均による中心化：　　$x_c = x - \bar{x}$
- 基礎知識や習慣による中心化
  - 女性ダミーの中心化：男女比が1:1だと想定すると
    - c_female = female - 0.5
  - 知能指数の中心化：平均は100になるはず
    - c_IQ = 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結果変数の予測値（平均値）

標準化した変数による単回帰

- yを標準化したものをxを標準化したものに回帰する

$$z_y = a + bz_x + \epsilon$$

$$z_x = \frac{x - \bar{x}}{u_x}, \quad z_y = \frac{y - \bar{y}}{u_y}$$

- 切片　$a = 0$
- 傾き　$b = \mathrm{Cor}(x, y) \in [-1, 1]$
- 標準化していない回帰：$|b| > 1 \Rightarrow \sigma_y > \sigma_x$

相関係数と単回帰の回帰係数

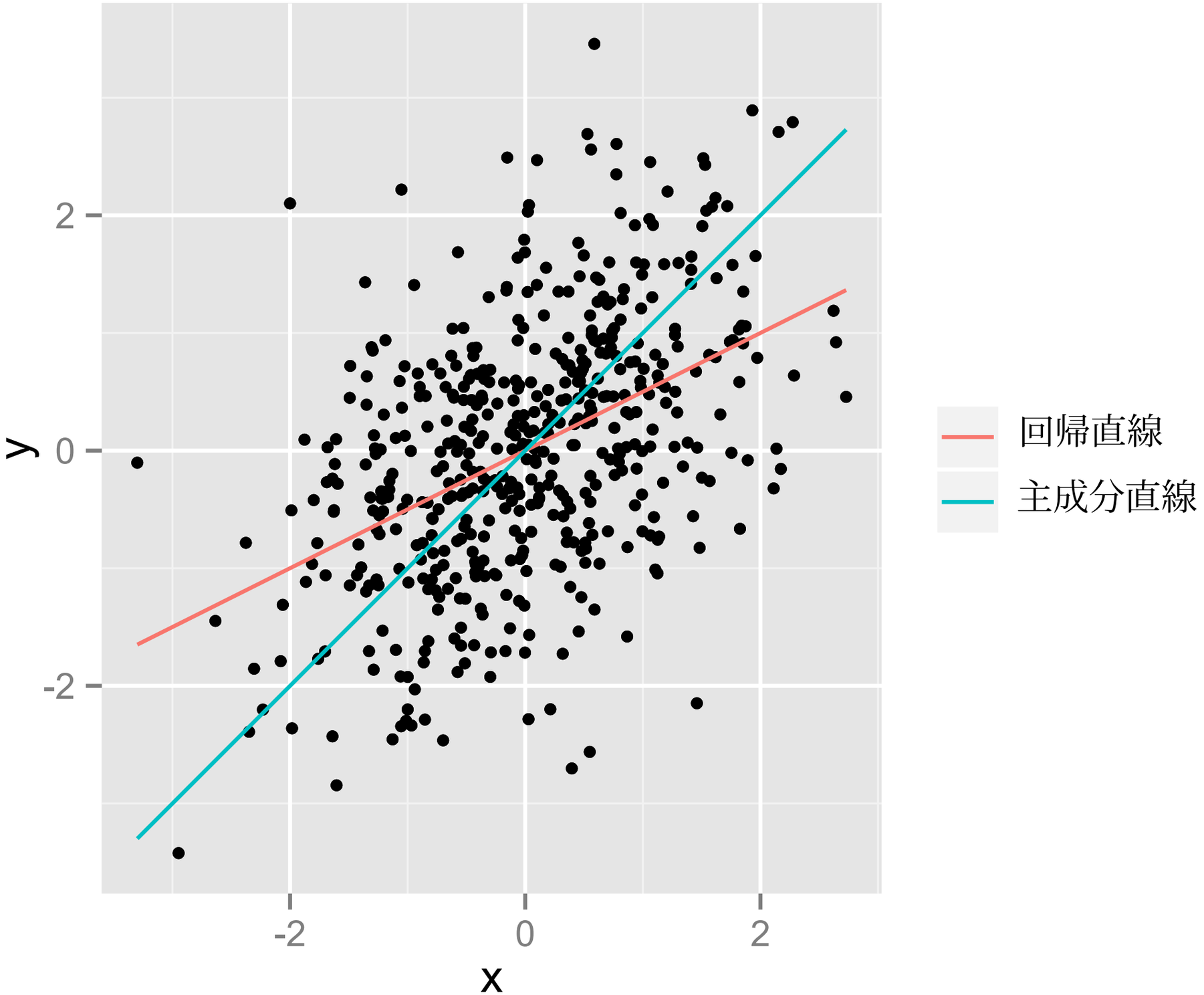
- 一般的な単回帰（標準化されていない場合も含む）を考える
- x とyの共分散を Cov(x, y) とする
- x と y の相関係数

$$\mathrm{Cor}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$$

- 回帰式の傾き：
-

$$b = \mathrm{Cor}(x, y)\frac{\sqrt{\mathrm{Var}(y)}}{\sqrt{\mathrm{Var}(x)}} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

主成分直線と回帰直線



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 の予測が過少
    - x が大きいときの y の予測が過大
  - 回帰直線：xのどの値の周辺でも、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

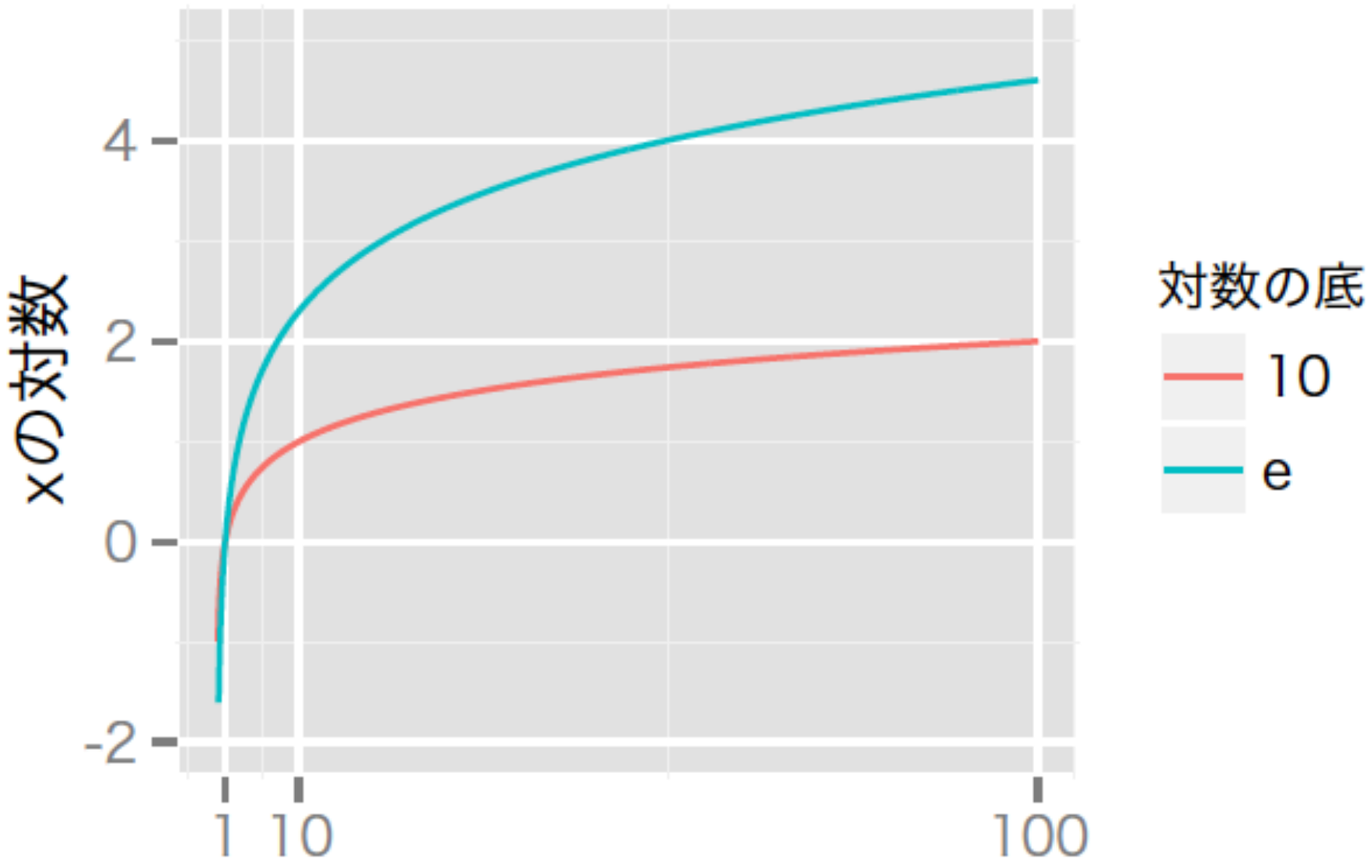


図: $\log_e x$ と $\log_{10} x$

対数変換の効果

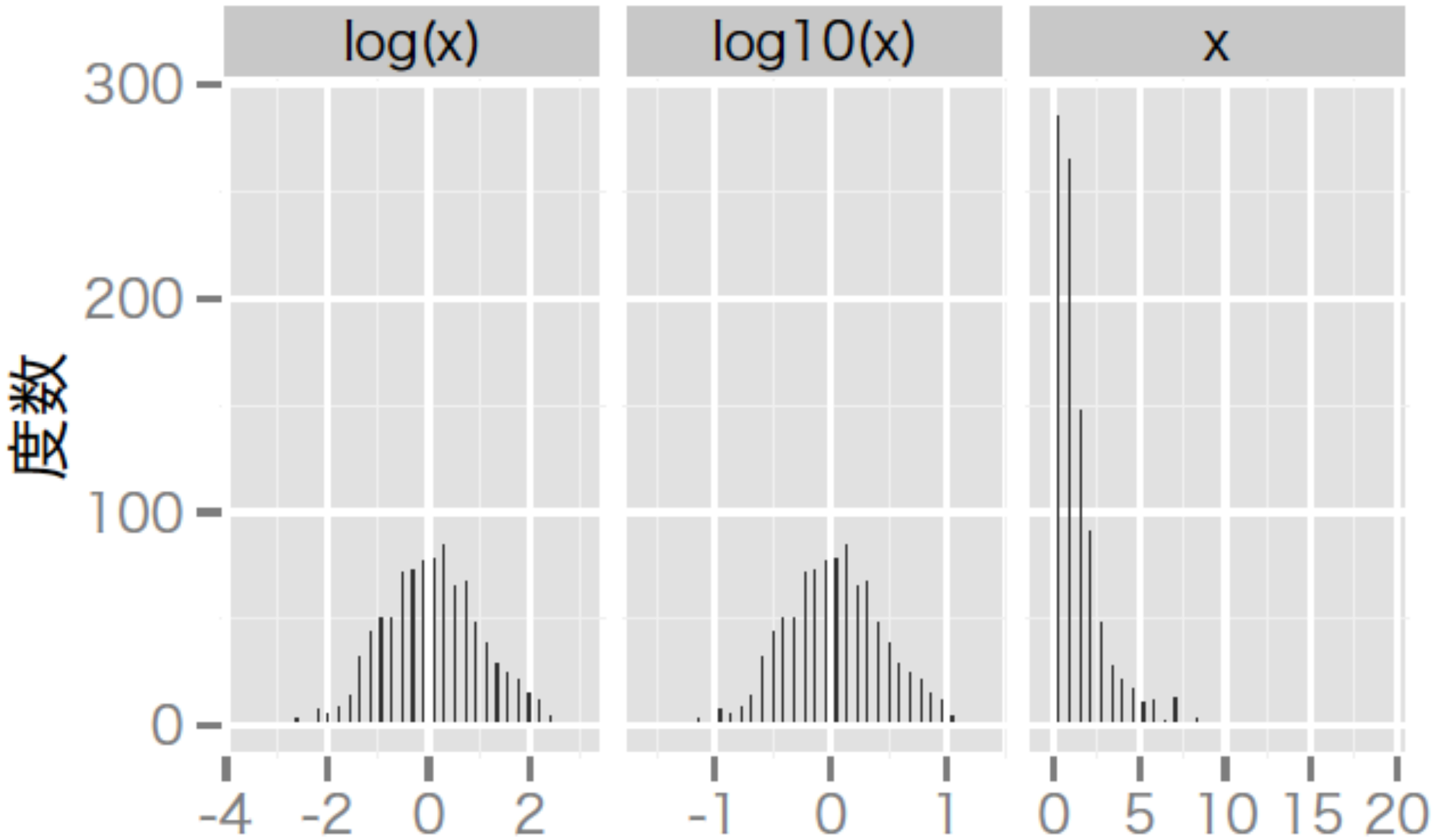


図: log(x) ($= \log_e x$), log10(x) ($= \log_{10} x$), x の分布

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：結果変数が自然対数のとき

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

- 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：
結果変数が自然対数のとき

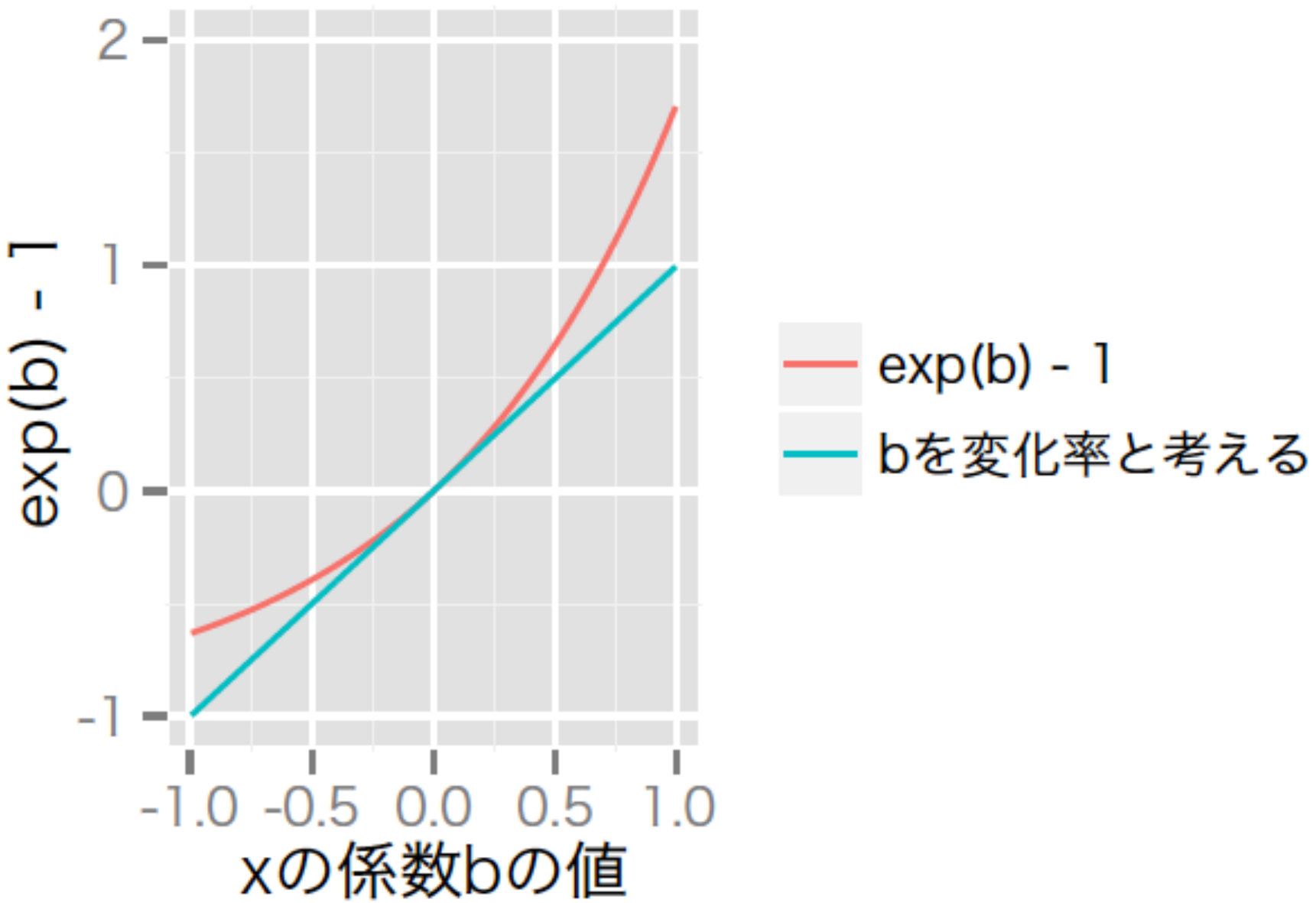


図: 係数を変化率と考える

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

結果変数	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、結果変数は b だけ増える
無変換	自然対数	説明変数が 1%増えると、結果変数が 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、結果変数が $100b$%増える
自然対数	自然対数	説明変数が 1%増えると、結果変数が $100b$%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