高知工科大学 経済・マネジメント学群

計量経済学

12. 回帰分析の応用 (2)

やない ゆうき
矢内 勇生

https://yukiyanai.github.io

yanai.yuki@kochi-tech.ac.jp

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
    - 線形変換、中心化
    - 対数変換
  - さまざまな仮説を検証する
    - 「回帰係数 = 0」以外の仮説を検定する
    - 線形ではない理論について、線形回帰で考える

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

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

  - 例2）知能指数 (IQ) の中心化：平均は100のはず

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

- 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 z_x と z_y を用いた単回帰：

$$z_{yi} \sim \text{Normal}(s + r z_{xi}, \sigma)$$

ただし、$Z_{yi} = \dfrac{y_i - \bar{y}}{u_y}, \;\; Z_{xi} = \dfrac{x_i - \bar{x}}{u_x}$

- 切片 $s = 0$
- 傾き　$r \in [-1, 1]$：x と yの相関係数
- つまり、通常の回帰：$y_i = a + b x_i + e_i$ で

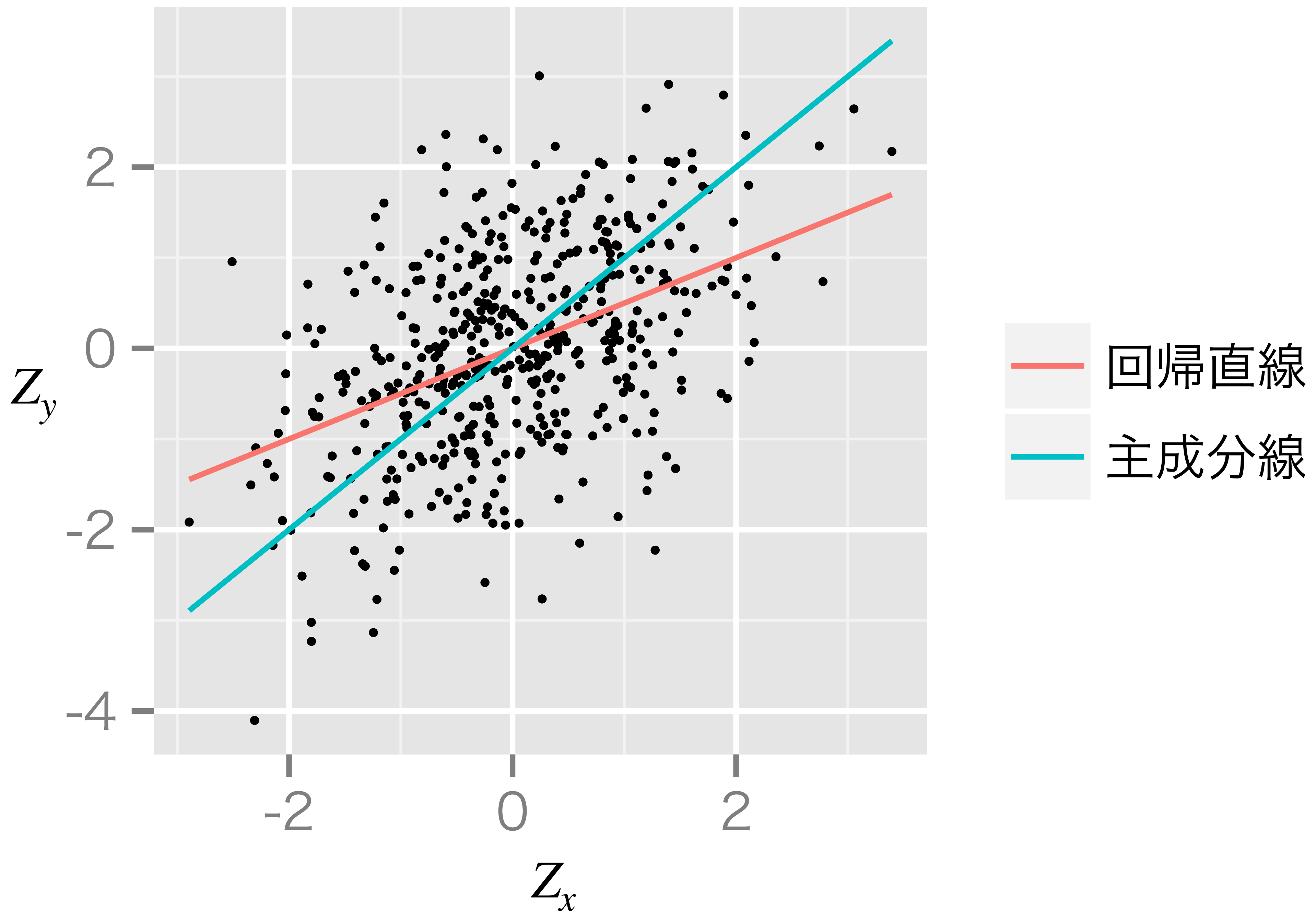
$$|b| > 1 \Rightarrow u_y > u_x$$

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 $\mathrm{Cov}(x, y)$ とする
  - x と yの相関係数 r：$r = \dfrac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$
  - 回帰式の傾き b：

$$
\begin{aligned}
b &= \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \frac{\sqrt{\mathrm{Var}(x)}}{\sqrt{\mathrm{Var}(x)}} \frac{\sqrt{\mathrm{Var}(y)}}{\sqrt{\mathrm{Var}(y)}} \\
&= r \frac{\sqrt{\mathrm{Var}(y)}}{\sqrt{\mathrm{Var}(x)}} = r \frac{\mathrm{SD}(y)}{\mathrm{SD}(x)}
\end{aligned}
$$

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、

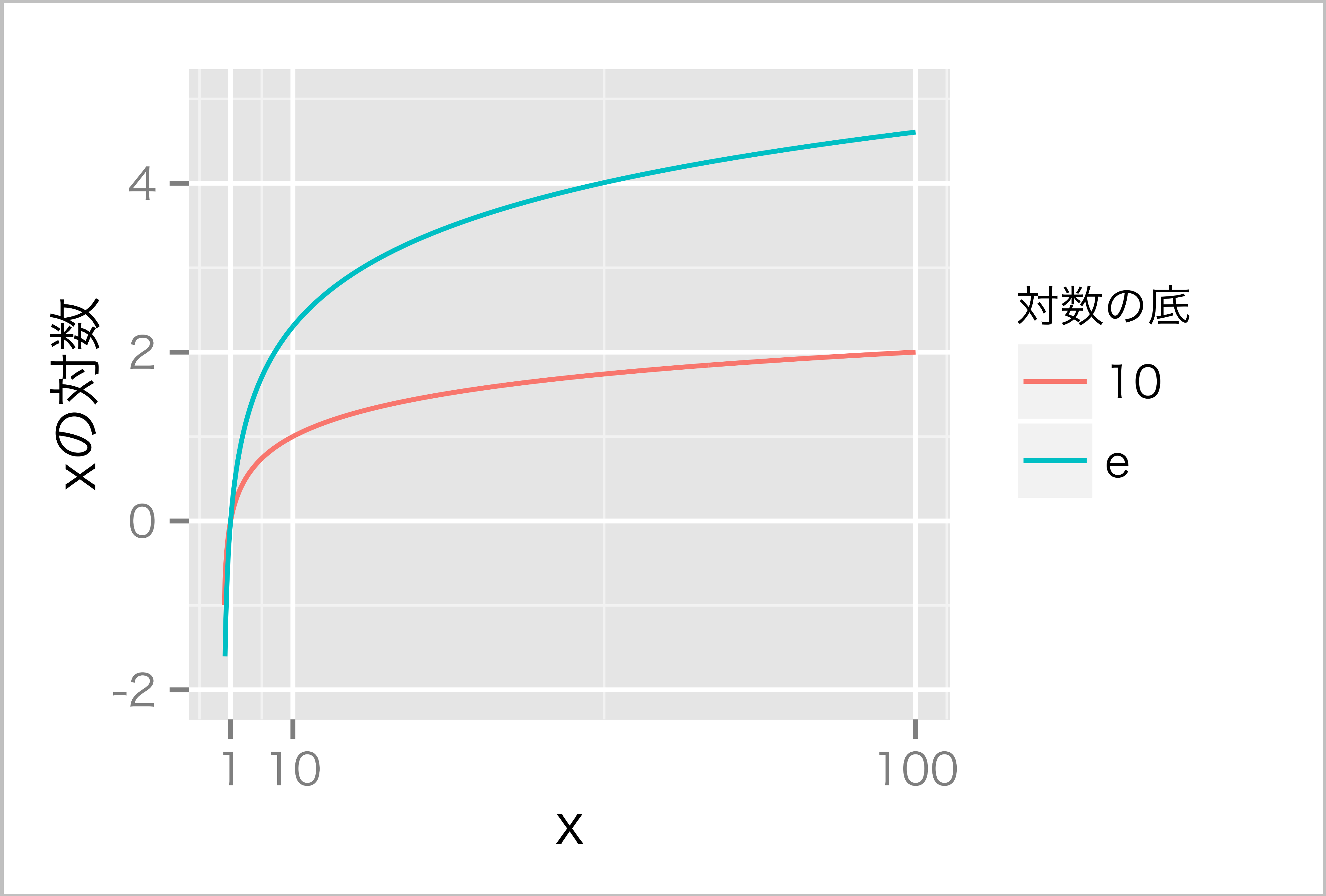
$\hat{y}$ と $\bar{y}$ **の距離** $<$ x と $\bar{x}$ **の距離**

    - 「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

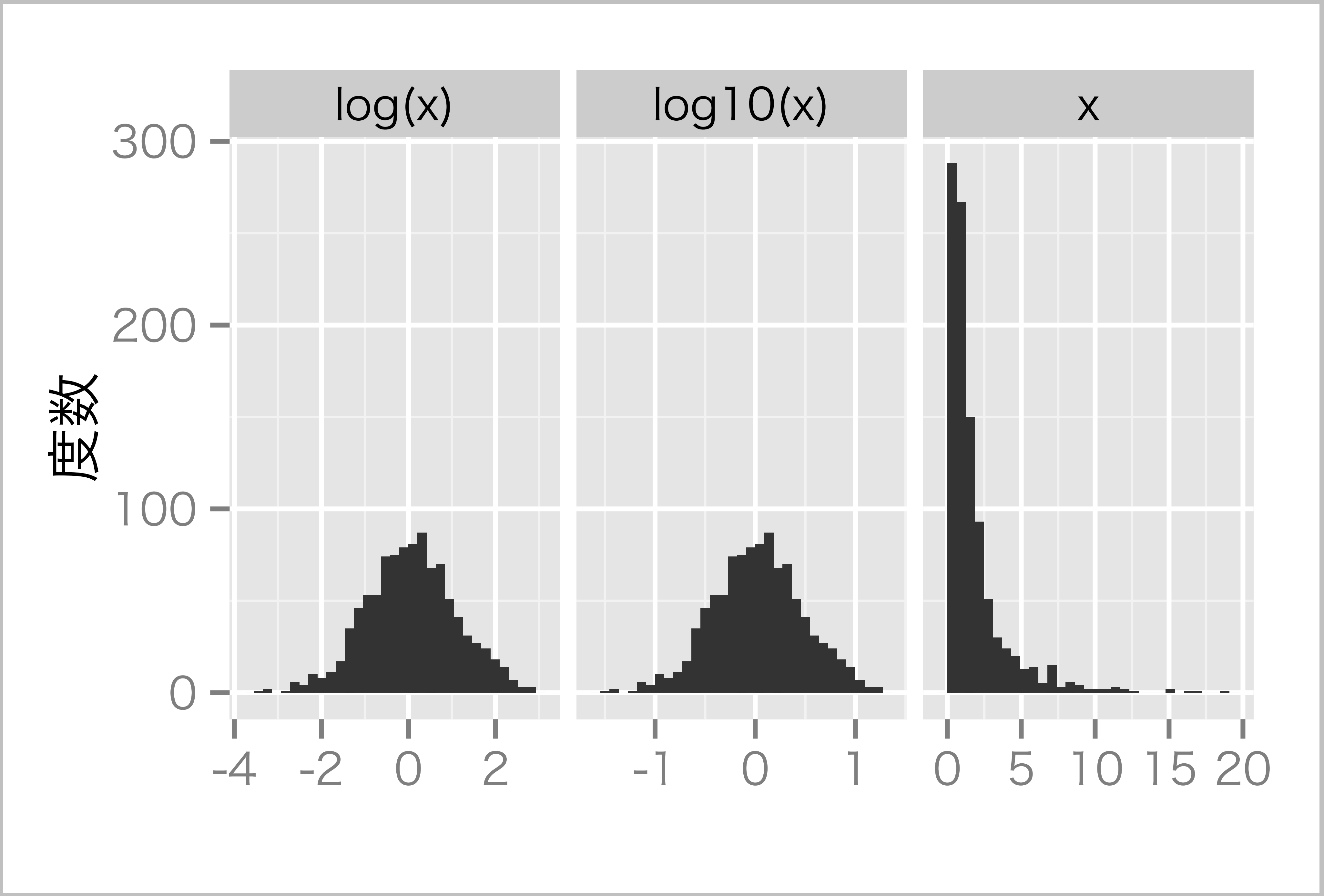
対数 (logarithm)

- 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

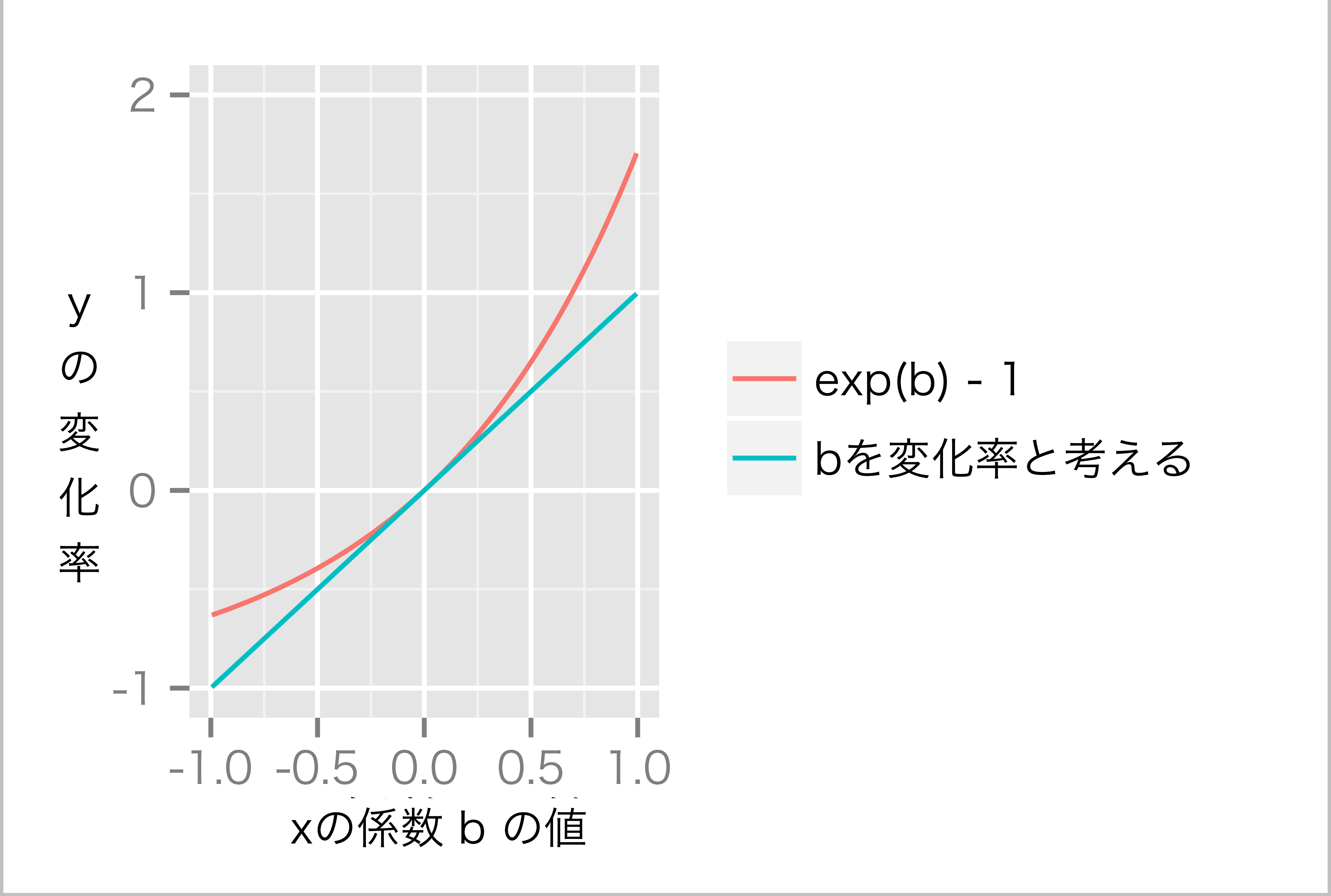
自然対数：底が e の対数

- x の自然対数：$\log_e(x)$ → 単に $\log(x)$ と書く
  - 経済学では、$\ln x$ とされることも多い
- 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 $\exp(0.06) - \exp(0) = \exp(0.06) - 1 \approx 0.06$ 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ことができる

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！

対数をとらない場合

$$\hat{y}_i = a + bx_i$$

- $x_j = x_i + 1$　とする

$$\hat{y}_j = a + bx_j = a + b(x_i + 1) = a + bx_i + b$$

- よって、x を1単位増やしたときの y の増分は、

$$\hat{y}_j - \hat{y}_i = b$$

- これは、回帰式の導関数から明らか：$\dfrac{d\hat{y}}{dx} = b$

応答変数のみ自然対数の場合

$$\widehat{\log(y_i)} = a + bx_i \quad \Rightarrow \hat{y}_i = \exp(a + bx_i) = \exp(a)\exp(bx_i)$$

- $x_j = x_i + 1$　とする（つまり、1単位の増分を考える）

$$\widehat{\log(y_j)} = a + bx_j = a + b(x_i + 1) = a + bx_i + b$$

$$\Rightarrow \hat{y}_j = \exp(a + bx_i + b) = \exp(a)\exp(bx_i)\exp(b)$$

- ここから、x を1単位増やしたときの y の変化率は、

$$\frac{\hat{y}_j - \hat{y}_i}{\hat{y}_i} = \frac{\hat{y}_j}{\hat{y}_i} - 1 = \exp(b) - 1$$

- b が0に十分近いとき、$\exp(b) - 1 \approx b$

- よって、x が1単位増えると、y は $100b$ パーセント増加する

説明変数のみ自然対数の場合

$$\hat{y}_i = a + b\log(x_i)$$

- $x_j = 1.01x_i$　とする（つまり、1%の増分を考える）

$$\hat{y}_j = a + b\log(x_j) = a + b\log(1.01x_i) = a + b\log(x_i) + b\log(1.01)$$

- ここで、$\log(1.01) \approx 0.01$ なので、

- $\hat{y}_j - \hat{y}_i = b\log(1.01) \approx 0.01b$

- よって、x が1%増えると、y は $0.01b$ 単位増加する

応答変数と説明変数が自然対数の場合

$$\widehat{\log(y_i)} = a + b\log(x_i)$$

- $x_j = 1.01x_i$　とする（つまり、1％の増分を考える）

$$\widehat{\log(y_j)} = a + b\log(x_j) = a + b\log(x_i) + b\log(1.01)$$

- よって、
- $\widehat{\log(y_j)} - \widehat{\log(y_i)} = b\log(1.01) \approx 0.01b$
- よって、x が1%増えると、$\log(y)$ は $0.01b$ 増加する
- ここで、$\log(y) + 0.01 \approx \log(y) + \log(1.01) = \log(1.01y)$ なので、$\log(y)$ が0.01 増える

  というのは、y が1%増えることに匹敵する
- したがって、x が1%増えると、y は b % 増加する（つまり、b は弾力性）

対数変換したモデルの解釈

応答変数	説明変数	係数の推定値 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 単位増える
無変換	自然対数	説明変数が 1% 増えると、応答変数が b/100 単位増える
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 b%増える

さまざまな仮説検定

経済理論を回帰分析で実証する

例：労働生産性と実質賃金（西山ほか 2019：第4章）

- Y：生産量
- L：労働投入量
- 生産関数：$Y = AL^{\beta_1}$
    - ただし、$A > 0,\ \beta_1 > 0$
- w：労働者の賃金
- p：生産物の価格
- 企業の利潤　$\pi = pY - wL = pAL^{\beta_1} - wL$

経済理論を回帰分析で実証する（続）

- 利潤最大化の1階の条件

$$\frac{\partial \pi}{\partial L} = pA\beta_1 L^{\beta_1 - 1} - w = 0$$

$$\Leftrightarrow \frac{w}{p} = A\beta_1 L^{\beta_1 - 1} = \beta_1 \frac{Y}{L}$$

- w/p：実質賃金
- Y/L：労働生産性
- よって、次のような回帰式が考えられる

$$\widehat{\text{実質賃金}} = \beta_0 + \beta_1 \text{労働生産性} \quad (1)$$

ただし、$\beta_0 = 0, \ \beta_1 > 0$

どんな仮説を検証する？

- (1) の回帰式について、「労働生産性が上がるほど実質賃金が上がる」という仮説を検証したいとき
  - 帰無仮説「$\beta_1 = 0$」で検定を行う
- 労働投入量を k (> 0) 倍したときに、生産量がk 倍より多くなる（収穫逓増）か、少なくなる（収穫逓減）か確かめたいとき
  - 帰無仮説「$\beta_1 = 1$」で検定を行う
    - 帰無仮説が棄却され、β_1 の推定値 $b_1 > 1$なら、収穫逓増
    - 帰無仮説が棄却され、β_1 の推定値 $b_1 < 1$なら、収穫逓減

「回帰係数 = 0」以外の帰無仮説を検定する

- 単回帰モデル：$Y_i = \alpha + \beta X_i + \epsilon_i, \ \ \epsilon_i \sim \text{Normal}(0, \sigma)$
  - 帰無仮説 $\beta = \beta_{\text{null}}$；対立仮説 $\beta \neq \beta_{\text{null}}$
  - β の推定値 b

$$\frac{b - \beta_{\text{null}}}{\text{SE}(b)} \sim t\,(N - 2)$$

  - t 統計量を計算する際、回帰係数から β_{null} を引いたものを標準誤差で割る
  - この値を使って t 検定を行う

式変換してから回帰する

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- $\beta = \gamma + \beta_{\text{null}}$ とおく

$$Y_i = \alpha + (\gamma + \beta_{\text{null}})X_i + \epsilon_i$$

$$\Leftrightarrow Y_i - \beta_{\text{null}}X_i = \alpha + \gamma X_i + \epsilon_i$$

- $(Y - \beta_{\text{null}}X)$ を X に回帰する
  - 帰無仮説： $\beta = \gamma + \beta_{\text{null}} = \beta_{\text{null}} \Rightarrow \gamma = 0$
  - 対立仮説： $\beta = \gamma + \beta_{\text{null}} \neq \beta_{\text{null}} \Rightarrow \gamma \neq 0$
- 検定には γ の推定値を使うが、推定したいのは β であることに注意する！

生産関数を対数変換する

- 生産関数：$Y = AL^{\beta_1}$
- 両辺の自然対数をとると、

$$\log(Y) = \log\left(AL^{\beta_1}\right) = \log(A) + \beta_1 \log(L)$$

- Y の対数をLの対数に回帰すれば、β_1 を推定できる
  - β_1：労働投入量に対する生産量の弾力性
  - 切片の推定値は、Aの対数の推定値

二次式の推定

- 例：年齢が投票率に与える影響を推定する
    - 若者の投票率は低い：年齢が投票率を上げる
    - 後期高齢者の投票率は低い：年齢が投票率を下げる？？
        - ある年齢までは投票率が上がるが、その後は投票利率が下がる

投票率を年齢に回帰する

- Y_i：個人 i $(i = 1,2,\ldots,N)$ が投票する確率
- X_i：i の年齢
- 回帰モデル：$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i + \beta_2 X_i^2, \sigma)$ (2)
    - ある年齢までは投票率が上がり、そこから下がるなら、$\beta_2 < 0$ になるはず

★ 疑問：$\beta_0 + \beta_1 X_i + \beta_2 X_i^2$ は二次関数なので、線形回帰ではないのでは？？？

線形回帰で二次関数を推定する

- $Z = X^2$ とおけば、(2) の回帰モデルは、
  - $Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i + \beta_2 Z_i, \sigma)$
  - つまり、X と Z という2つの説明変数をもつ重回帰であるとみなせる
- 注意：ただし、偏回帰係数の解釈には注意が必要
  - β_1 の推定値は、Z の値を一定にたもったときに X がY に与える影響とは解釈できない
    - 理由：X を動かすと、必ずZ も動く！

次回

分析結果を伝える方法