計量経済学応用

8. 重回帰分析 (3)

矢内　勇生

2019年5月16日

高知工科大学　経済・マネジメント学群

今日の目標

- 回帰分析を用いた統計的推定法を理解する
- 分析結果の提示法を理解する

回帰分析による推定

- データから作った散布図への直線（平面）の当てはめは、標本データの要約
- 興味があるのは母集団の特徴
- どうやって推定する？

単回帰モデル

- 単回帰モデル

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

- α、β ：母数（推定の対象）

- ε ：誤差 (error) 。説明変数以外で結果変数に影響を与えるもの。平均すると0

最小二乗法による母数の推定 単回帰の場合

- 最小二乗法によって求めた回帰係数a, bは、α, β の点推定値である

- 最小二乗推定量は以下の望ましい性質をもつ
  - 不偏性：E(a) = α, E(b) = β
  - 一致性：標本サイズを無限大にすると、推定値は母数に一致する

重回帰モデル

- 重回帰モデル

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- β_m ：母数（推定の対象）、m=0,1,2,...,k
- ε ：誤差

最小二乗法による母数の推定
重回帰の場合

- 最小二乗法によって求めた回帰係数b_0, b_1, ..., b_kは、β_0, β_1,..., β_kの推定値である
  - 不偏性：$E(b_j) = \beta_j$ ($j = 1, 2, \cdots k$)
  - 一致性

信頼区間

- 回帰分析による点推定値は、1つの標本（データ）から得られたもの

➡母数に一致するとは限らない（実際の標本サイズは有限なので）

- 統計量はばらつく（シミュレーションで確認した）
  - 標準誤差：統計量のばらつき

➡信頼区間を求める!

信頼区間の意味（1）

- 95%信頼区間とは何か？
  - よくある**誤解**：「得られた信頼区間に、真の値が入っている確率が95%」
  - 「真の値」があるなら、「得られた信頼区間に、真の値が入っている確率」は、
    - 100%（実際に入っている）

      または

    - 0%（入っていない）

    しかあり得ない

信頼区間の意味（2）

- では、95％信頼区間とは何なのか？

  1. データを生成する（新たに観測する）

  2. データを分析する

  3. 95%信頼区間を求める

- 95%信頼区間：上の1〜3までを何度も何度も繰り返し行うと、そのうち95%くらいは「真の値を含む信頼区間」が得られるだろう

信頼区間の信頼度（1）

- 信頼区間の長さ
    - 信頼度が高いほど区間が長くなる
    - 信頼度が低いほど区間が短くなる
- なぜ？
    - 区間を長くすれば、取りこぼしの確率が小さくなる
    - 区間を短くすれば、取りこぼしの確率は大きくなる

信頼区間の信頼度（2）

- では、信頼区間は長い方がいいのか？
  - No!
  - 同じ信頼度で、より信頼区間が短いほうが推定の不確実性が小さい
  - 信頼区間の長さ：標準誤差に依存
    - 標準誤差が大きい：信頼区間が長い
    - 標準誤差が小さい：信頼区間が短い

Rで回帰分析

- `lm()` 関数を使う
  - 例、myd という名前のデータセットに含まれる変 を使い、yを x1 とx2 に回帰する

```
fit <- lm(y ~ x1 + x2, data = myd)
```

summary() で結果を確認する

- lm() で推定した後、summary() で結果を確認する
- 例：summary(fit)
  - Estimate: パラメタの点推定値
  - Std. Error：標準誤差（推定の不確実性）
  - t value: t 検定で使う検定統計量
  - Pr(>|t|)：p 値

`broom::tidy()` で結果を確認する

- broom パッケージの `tidy()` 関数でも結果を確認できる

Rで信頼区間を求める

- `lm()` を実行した後、`confint()` 関数を使うと、係数の信頼区間を求めることができる。
  - 例
    - 95%信頼区間：`confint(fit)`
    - 50%信頼区間：`confint(fit, level = 0.5)`
    - 67%信頼区間：`confint(fit, level = 0.67)`
  - 上のコマンドを実行すると、信頼区間の下限値と上限値が表示される

信頼区間の図示

- ggplot2 を使えば、以下のものが図示できる
  - 回帰直線 + 95%信頼区間
    - geom_smooth(method = “lm”)
  - 回帰直線 + 89%信頼区間
    - geom_smooth(method = “lm”, level = 0.89)
  - 回帰直線のみ
    - geom_smooth(method = “lm”, se = FALSE)

回帰分析における仮説検定

- 回帰分析では、説明変数が結果変数に影響を与えているかどうかに関心がある
  - 帰無仮説：説明変数の影響はない（影響が0である）
  - 対立仮説：説明変数の影響がある（影響が0ではない）

単回帰の場合

- 帰無仮説：$\beta=0$
- 対立仮説：$\beta\neq0$
  - α は説明変数の影響ではないので、通常はあまり気にしない

重回帰の場合

- パタン1（複合仮説）
  - 帰無仮説：$\beta_1 = \beta_2 = \cdots = \beta_k = 0$
  - 対立仮説：少なくとも1つは0でない
- パタン2
  - 帰無仮説： $\beta_1 = 0, \beta_2 = 0, \cdots, \beta_k = 0$
  - 対立仮説： $\beta_1 \neq 0, \beta_2 \neq 0, \cdots, \beta_k \neq 0$

パタン2の仮説検定

- p値が設定した有意水準より小さいとき
  - 帰無仮説を棄却する
  - 係数は「統計的に有意 (statistically significant)」である（「優位」ではない！）
- p値が設定した有意水準以上のとき
  - 帰無仮説を*とりあえず*受容：対立仮説が正しいとはいえないという弱い結論

統計的に有意とは？

- 効果が「ゼロではない」と信じるに足る証拠がある
  - それだけ！
- 「ゼロではない」≠ 重要
- 研究においては、「重要である」ことを示すことが求めらる
  - 実質的重要性 (substantive significance) を示すことが必要（浅野・矢内 2018: pp. 165-168 を参照）
- **係数の値そのもの（効果量, effect size）を議論することが絶対に必要！！！**

効果がないことを証明できる？

- 効果がないことを証明したいとき、β=0という帰無仮説が受容されることは証拠として使える？

➡使えない！

- 統計的仮説検定の方法では、効果がない証拠を見つけることは不可能（ベイズ法でROPEというものを設定する必要）

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 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

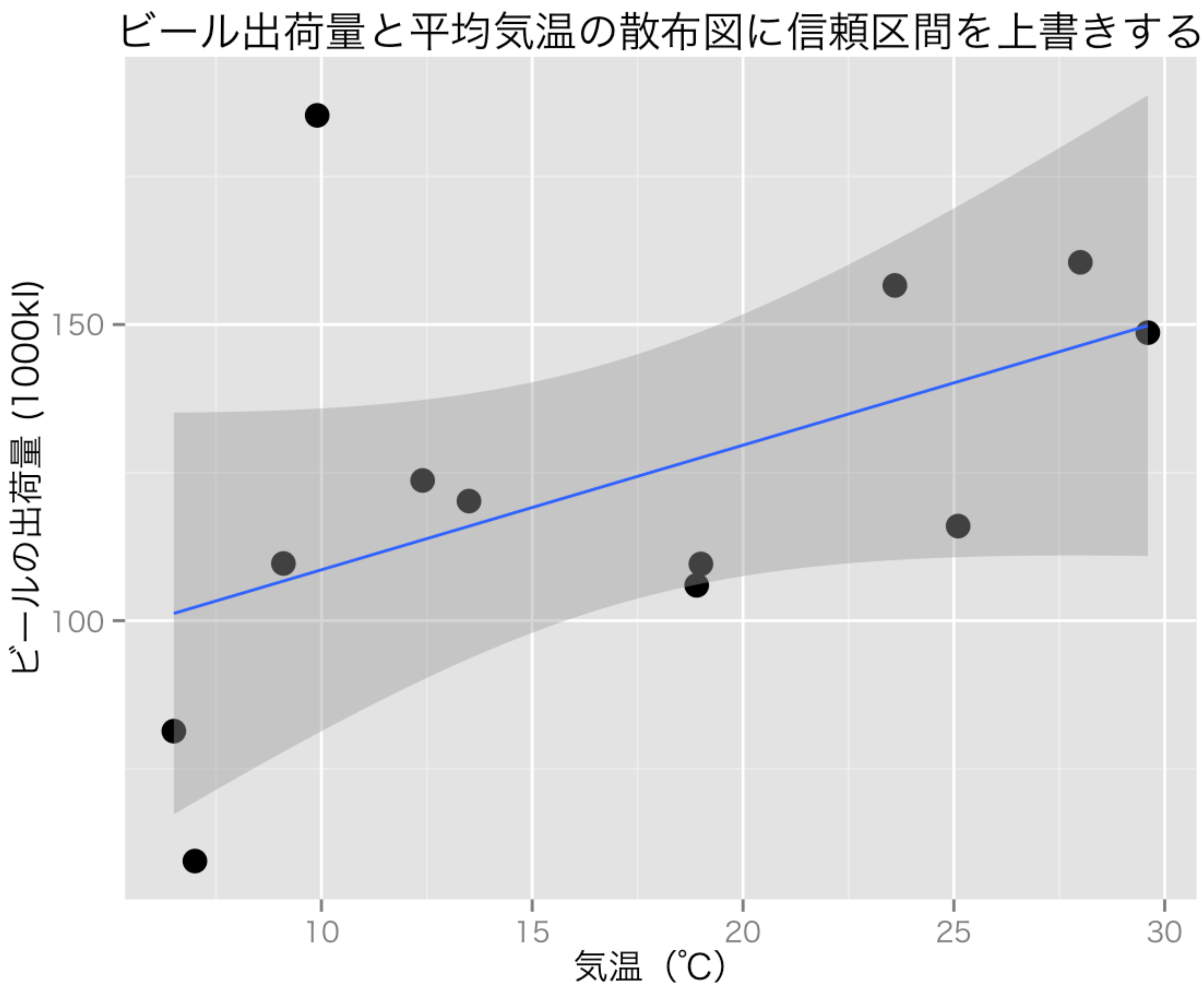
- 括弧内にp値を書けば、ある有意水準の下で棄却されるか受容されるかが一目でわかる
- seが書かれている場合の目安：有意水準5%なら、係数÷SE の値が2以上なら帰無仮説を棄却
- t値（検定統計量）を書いても理論的には問題ないが、臨界値を求めないと棄却か受容か判断できないので、あまり好まれない

結果提示の例：表の場合

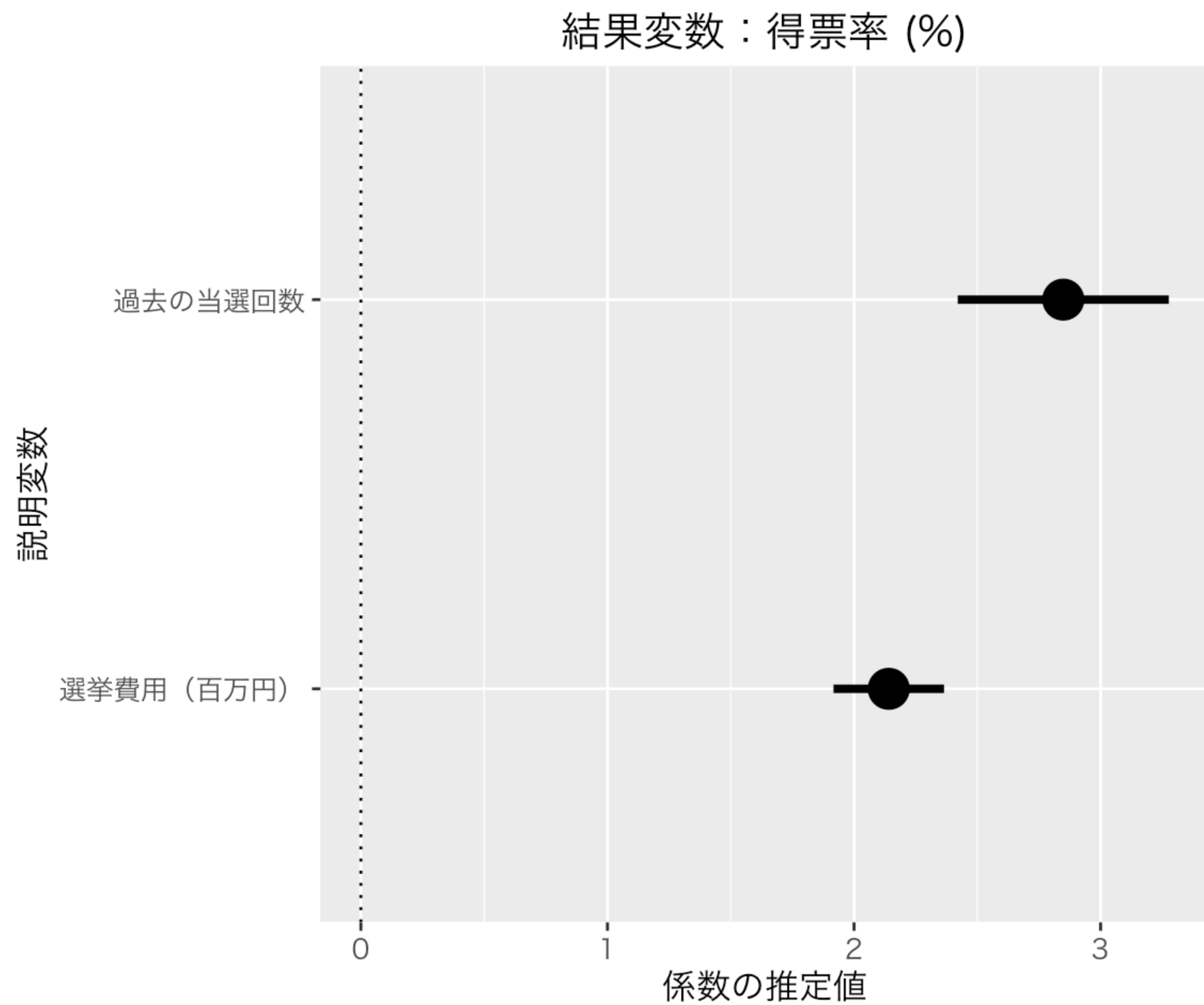
表1. 回帰分析の結果（結果変数は自民党の得票率）

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。