回帰分析３：回帰分析における仮定の議論

講師：遠山祐太

最終更新：2024-11-16

はじめに

最小二乗法における重要な仮定

- 仮定２： ϵ_i の条件付き期待値は0である。 $E[\epsilon_i | X_{i1}, \ldots, X_{iK}] = 0$
  - 任意の k について $Cov(X_{ik}, \epsilon_i) = 0$ （あるいは $E[\epsilon_i X_{ik}] = 0$ ）が成立。
  - **誤差項と説明変数の間に相関がない**
- 仮定４：完全な多重共線性がない。
- 実証分析の設定・問に応じて、これらの仮定の妥当性を検討しなければならない。
- 今日の問：
  - どのような状況でこれらの仮定は成立する・成立しないか？
  - 実際の分析においてどのように仮定の妥当性を吟味・サポートするか？

今日の内容

- 内生性の問題
- 多重共線性の問題
- 感度分析
- （補足資料）実証研究：テレビ露出と得票率

キーポイント：因果効果推定からみた線形回帰

- 以下の回帰式を用いて、 D の Y に対する因果効果を推定しよう。

$$y_i = \alpha_0 + \alpha_1 D_i + \beta' x_i + \epsilon_i$$

- **興味ある変数 D の変動（variation）** が以下の２点の意味で重要となる。

- 1: x_i を 統制した後に、 **D が外生的な変動**を持つこと。
  - 仮定３（条件付き期待値が0）と関連する。（バイアスなし）

- 2: x_i を 統制した後に、 **D が十分な変動**を持つこと。
  - 精緻に推定する上で重要（小さい標準誤差）
  - 仮定４（多重共線性）と関連する。

内生性

内生性問題

- $Cov(x_k, \epsilon) = 0$ が満たされないケースを **内生性問題（endogeniety problem）**とよぶ。
  - また、そのような変数 x_k を **内生変数（endogenous variable）** とよぶ。
- 例１：**欠落変数バイアス（omitted variable bias）**
  - 今日のフォーカス
- 例２：同時性（simultaneity）
  - 需要・供給推定の文脈など。
- 例３：観測誤差（measurement error）
  - 説明変数の観測誤差
- 例４：サンプルセレクション (sample selection)
  - 例：賃金は働いている人についてのみ観察される。

欠落変数バイアス（omitted variable bias; OVB）

- 以下の賃金回帰モデルを考える（真のモデル）

$$\begin{aligned} \log W_i &= \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i \\ E[u_i | S_i, A_i] &= 0 \end{aligned}$$

  W_i は賃金、 S_i は教育年数、 A_i は能力を表す。

- β_1 ：**他の要因を固定した上での** 教育年数が賃金に与える効果
- 問題点：能力 A_i を直接観察できるとは限らない。

能力 A_i を除いた回帰モデル

- 以下の回帰モデルを考える。

$$\log W_i = \alpha_0 + \alpha_1 S_i + v_i$$

  - $v_i = \beta_2 A_i + u_i$ であるから、教育年数 S_i と v_i はおそらく相関している。

- このモデルで推定した $\hat{\alpha}_1$ は β_1 について一致性がなくなる。

$$\hat{\alpha}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov(S_i, A_i)}{Var(S_i)}$$

- 一致性のみならず、不偏性についても同様の公式が成立

欠落変数バイアスの公式

- 欠落変数バイアスは以下の２つに依存する。
  1. 欠落変数（ここでは A_i ）が 左辺に与える影響： β_2
  2. 欠落変数と興味ある説明変数の関係

- 以下の表でまとめられる。
  - x_1 ：興味ある変
  - x_2 欠落変数
  - β_2 は x_2 の係数

	$Cov(x_1, x_2) > 0$	$Cov(x_1, x_2) < 0$
$\beta_2 > 0$	正のバイアス	負のバイアス
$\beta_2 < 0$	負のバイアス	正のバイアス

欠落変数バイアスの直観

- OLS推定量は(他の変数を統制した上で)被説明変数と説明変数の相関である。
- この「相関」を「因果」として解釈するには、「交絡因子がない」ことが重要。
- 初回授業における図を思い出そう（図追加）

説明変数 X の外生性をどのように担保するか？

- 平均独立（mean independence）の仮定はバイアスのない推定に不可欠である。
- しかしながら、**未観測（unobserved）**要素に関する仮定であるため、その議論は難しい。
- しかも、**外生性の仮定に関する正式な検定は存在しない。**
  - 問：残差 $\hat{\epsilon}_i$ と説明変数 X_{ik} の相関を調べるのは、検定として機能するだろうか？
- どのようにの回避するか？
  - 1：統制変数を加える
  - 2：その他の識別戦略（自然実験、操作変数法、パネルデータ）

統制変数の追加

- 係数 α_1 に主眼をおく、次のモデルを考えよう。

$$y_i = \alpha_0 + \alpha_1 D_i + \beta' x_i + \epsilon_i,\ E[\epsilon_i | D_i, x_i] = 0$$

- アイデア： x に統制変数をより多く加えると、
  - 処置変数 D_i と相関する要素を統制できる。
  - 変数の欠落が避けられる。
  - D_i と ϵ_i の平均独立の仮定は蓋然性が高くなる。

- 可能な限り変数を追加すべきだろうか？そうとは限らない。
  - 問題点１：統制変数を増やすと不正確な推定になる。
  - 問題点２：悪い統制の問題（bad control problem）

悪い統制の問題

- 次のモデルを考えよう。

$$wage_i = \alpha_0 + \alpha_1 college_i + \alpha_2 occupation_i + \epsilon_i,\ E[\epsilon_i | D_i, x_i] = 0$$

- 興味のある係数 α_i は、**職業を統制した上での**大学進学が賃金に与える影響である。
- しかし、職業は明らかに大学進学に影響されている。
- 推定された α_1 には、大学進学が職業選択を通じて賃金に与える効果を捉えることができない。
- 変数 $occupation_i$ は**悪い統制（bad control）**とよばれ、変数としていれるべきではない。

変数選択の手引き

	y_i に影響する	y_i に影響しない
X_i に影響する・ X_i と同時に決定される	欠落変数バイアスを防ぐために必ず入れる	分散を増やすだけなので入れてはならない（バイアスは減らない）
X_i に影響される	入れない（悪い統制の問題）	上に同じ
X_i と相関がない	分散を小さくするので入れる（入っていなくてもバイアスはない）	上に同じ

自然実験(Natural Experiment)

- 自然実験：処置の有無があたかも「実験」されているかのようにランダムに決まる。
  - 政策変更、「たまたま」のイベント、政策適用の閾値、などなど。

- 例：
  - 天気
  - くじ引きによる政策の割り当て（徴兵制など）
  - 出生関連の出来事（双子・日付）
  - 他にもいろいろ！

- 自然実験があるとRCTの状況に近づき、分析がしやすくなる。

- 近年のミクロ実証分析では自然実験 (Natural Experiment)を活用することが重視されている。
  - 常にあるとは限らないが、思った以上に多い。
  - どの程度「ランダムか」もケース・バイ・ケース

多重共線性の問題

完全な多重共線性

- 完全な多重共線性：ある説明変数が、他の説明変数の線形結合によって表されること。
- この場合、全ての係数を推定することはできない。
- 例えば

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 \cdot x_2 + \epsilon_i$$

  かつ $x_2 = 2x_1$ の場合

- β_1 と β_2 を同時に推定することはできない。

より詳しく：

- 上のモデルは

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 \cdot 2x_1 + \epsilon_i$$

と書き表せる。これは

$$y_i = \beta_0 + (\beta_1 + 2\beta_2) x_1 + \epsilon_i$$

と同じである。
- 合成項 $\beta_1 + 2\beta_2$ を x_1 の係数として推定することはできるが、 β_1 と β_2 別々に推定することはできない。

直感

- 直感的に言えば、説明変数 x の変動が被説明変数 y の変動にどのように影響するかを捉えることで、回帰係数が推定される。
- x_1 と x_2 が完全に連動していると、 y の変動がどれだけ x_1 あるいは x_2 の変動に基づくかわからないので、 β_1 と β_2 が区別できない。

例：ダミー変数

- 男女を表すダミー変数を考えよう。

$$male_i = \begin{cases} 1 & if\ male \\ 0 & if\ female \end{cases}$$

$$female_i = \begin{cases} 1 & if\ female \\ 0 & if\ male \end{cases}$$

- もし回帰に男女ダミーを両方入れると

$$y_i = \beta_0 + \beta_1 famale_i + \beta_2 male_i + \epsilon_i$$

- 任意の i について $male_i + famale_i = 1$ なので、完全な多重共線性がある。

- 常にいずれかの群のダミー変数を省略する必要がある。
- 例えば

$$y_i = \beta_0 + \beta_1 famale_i + \epsilon_i$$

- この場合、 β_1 は**男性と比較したときの**女性であることの効果として解釈される。
  - 省略された群は比較の基準となる。

練習問題

- 以下の回帰式で処置 D_i の効果を推定しよう。

$$y_i = \beta_0 + \beta_1 D_i + \epsilon_i, E[\epsilon_i | D_i] = 0$$

  D_i は 処置を示すダミー変数である。

- 問：変数 D_i がどのようなものである場合に、完全な多重共線性に持つ（つまり係数が識別できない）か？ またその直観は？

複数のダミー変数

- 複数の群に対応するときも同様にする。

$$freshman_i = \begin{cases} 1 & if\ freshman \\ 0 & otherwise \end{cases}$$

$$sophomore_i = \begin{cases} 1 & if\ sophomore \\ 0 & otherwise \end{cases}$$

$$junior_i = \begin{cases} 1 & if\ junior \\ 0 & otherwise \end{cases}$$

$$senior_i = \begin{cases} 1 & if\ senior \\ 0 & otherwise \end{cases}$$

かつ

$$y_i = \beta_0 + \beta_1 freshman_i + \beta_2 sophomore_i + \beta_3 junior_i + \epsilon_i$$

不完全な多重共線性

- 不完全な多重共線性：説明変数間の相関が強い。
- 最小二乗法でモデルを推定することはできるが、推定の精度（すなわち標準誤差）に影響する。
- 均一分散をもつ簡単なモデルを考えよう。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, V(\epsilon_i) = \sigma^2$$

最小二乗推定量の分散

- OLS推定量の条件付き分散は以下となる（漸近分散も似た形になる）

$$V(\hat{\beta}_1|X) = \frac{\sigma^2}{N \cdot \hat{V}(x_{1i}) \cdot (1 - R_1^2)}$$

- $\hat{V}(x_{1i})$ は標本分散

$$\hat{V}(x_{1i}) = \frac{1}{N}\sum(x_{1i} - \bar{x_1})^2$$

- R_1^2 は x_1 を x_2 に回帰するときの決定係数である。

$$x_{1i} = \pi_0 + \pi_1 x_{2i} + u_i$$

分散を小さくする４つの要素

1. N が大きいこと。
2. $\hat{V}(x_{1i})$ が大きいこと。
    - x_{1i} の変動が大きい！
3. R_1^2 が小さいこと。
    - R_1^2 は、他の変数が x_{1i} を線形的にうまく説明する程度を表している。
    - $R_1^2 = 1$ なる極端な場合では、 x_{1i} が他の変数の線型結合で表せる。
4. 誤差項の分散 σ^2 が小さい。
    - これは y_i の変動の説明される程度を反映している。
    - 統制変数を追加すればするほど、誤差項の分散は小さくなる。
    - 上の3点目を思い出そう。

まとめ： X の十分な変動

- X に変動があればあるほど、係数を正確に推定できる。
- **他の要素を統制した上での**変数の変動も重要である。
- 欠落変数バイアスを防ぐために統制変数を入れすぎると、 X の独立な変動がなくなる。

頑健性チェック

分析を正当化する方法

- 外生性（平均独立）の仮定の議論は難しい。
- 優れた実証分析では、懸念を払拭するために**頑健性チェック（robustness check）**が行われ、結果の頑健性が示される。
- Deryugina "Some Tips For Robustness Checks And Empirical Analysis In General" を参照
- 一般的な手法：
  - 統制変数に対する**感度分析（sensitivity analysis）**
  - **プラシーボ試験（placebo test）**　（実証例を見よ）

感度分析

- Step 1：統制変数を用いたモデルのうち妥当と思われるものを考えて推定する。(ベースライン)
- Step 2：さらなる統制変数を追加し、モデルを再推定する。
- Step 3：推定された興味のある係数（典型的には処置変数）の変化を見る。もしそれほど変化しないなら、結果は頑健である（内生性の懸念が小さい）。

なぜこの方法が外生性の議論において有用なのか？

- 外生性の懸念は、 D_i と誤差項の相関にある。
- もし統制変数を追加しても係数の推定値が変わらなければ、**欠落変数の効果はおそらく小さいであろう。**
- ただし、この手続きはフォーマルなものではなく、あくまで実践的な技術である。
- よりフォーマルな議論については Altonji, Elder, and Taber (2005) および Oster (2019) を参照