第 11 回：操作変数法
【教科書第 8 章】

北村　友宏

2025 年 12 月 9 日

本日の内容

1. 説明変数と誤差項の相関
2. 操作変数の定義
3. 2 段階最小二乗法

OLS 推定量の性質（大標本の場合）

大標本の場合，

- 誤差項の（条件なし）期待値は 0

$$E(\boldsymbol{u}) = \mathbf{0}.$$

- すべての説明変数と誤差項は無相関

$$\mathrm{Cov}(x_{ji}, u_i) = 0, \quad j = 1, 2, \cdots, k.$$

の仮定の下で，重回帰モデルの OLS 推定量 $\hat{\boldsymbol{\beta}}$ は一致性をもつ．

$$\operatorname*{plim}_{n \to \infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

⇒ 標本サイズが十分に大きい（観測値数が十分に多い）とき，重回帰モデルの偏回帰係数の OLS 推定量は真の偏回帰係数に確率収束する．（証明は省略）

ところが,

- 少なくとも 1 つの説明変数と誤差項が相関する

$$\mathrm{Cov}(x_{li}, u_i) \neq 0$$

という場合，モデルの OLS 推定量 $\hat{\beta}$ は一致性をもたない.

$$\operatorname*{plim}_{n\to\infty} \hat{\beta} \neq \beta.$$

⇒ 標本サイズが十分に大きくても（観測値数が十分に多くても），重回帰モデルの係数の OLS 推定量は真の係数と異なる値（偏った値）に確率収束する．（証明は省略）
⇒ 説明変数と誤差項に相関がある場合，係数を OLS で推定すると偏り（バイアス）が生じ，正しく推定できない.

- モデルの誤差項と相関しない変数を外生変数（exogenous variable）という．
  - e.g., ミンサー方程式における就業可能年数など
- モデルの誤差項と相関する変数を内生変数（endogenous variable）という．
  - e.g., ミンサー方程式における年収や修学年数
- 説明変数に内生変数が含まれることによって生じる OLS 推定量の偏りを内生性バイアス（endogeneity bias）という．
  - 内生性バイアスには，欠落変数バイアス，観測誤差バイアス，同時方程式バイアスなどがある（詳細な説明は省略）．

操作変数

内生性バイアスを緩和しつつ，説明変数に内生変数が含まれている式を推定する方法を考える．

- 内生変数である説明変数と相関し，かつ誤差項と相関しない変数を操作変数（Instrumental Variable, IV）という．
  - 操作変数を z_i, 誤差項を u_i とすると，$\mathrm{Cov}(z_i, u_i) = 0$ かつ $\mathrm{Cov}(z_i, x_i) \neq 0$.
- 外生変数であれば，「推定式の誤差項と相関しない」と考えることができる．
  ➡ 「内生説明変数と相関する外生変数」が，操作変数になりうる．
- 操作変数は 1 個とは限らず，複数個存在する場合もある．

操作変数法と 2 段階最小二乗法

- モデルの説明変数に内生変数が含まれている場合に操作変数を用いて係数の一致推定を試みる方法を操作変数法（instrumental variable method) という．
- 内生説明変数を，システムに登場する全ての外生変数に回帰して内生説明変数の予測値を求め，それを含む，モデルの全ての説明変数に被説明変数を回帰する方法を 2 段階最小二乗法（2-Stage Least Squares, 2SLS）という．
  - 操作変数法の 1 つと考えることができる．
- 2SLS などの操作変数法は，観測値数が十分大きいときに使われる．

2 段階最小二乗法

（y_i と）x_i が内生変数で，c_i と z_i が外生変数で，z_i が x_i と相関し，かつ u_i と相関していない場合，

$$y_i = \beta_0 + \beta_X x_i + \beta_C c_i + u_i$$

を推定する際，2SLS を用いれば，内生性バイアスを緩和できる．

- z_i を操作変数として用いる．

第 1 段階

- 内生説明変数を，システムに登場する全ての外生変数に回帰．つまり，

$$x_i = \underbrace{\delta_0 + \delta_Z z_i + \delta_C c_i}_{u_i \text{と無相関}} + \underbrace{v_i}_{u_i \text{と相関}}$$

  を OLS で推定．
  - z_i と c_i は外生変数なので定義上，u_i と無相関．
  - δ_0 は定数項なので変動せず，u_i と無相関．

  $\Rightarrow$ x_i の変動を，u_i と無相関な部分と相関する部分に分割．
- x_i の予測値 $\hat{x}_i = \hat{\delta}_0 + \hat{\delta}_Z z_i + \hat{\delta}_C c_i$ を求める．

  $\Rightarrow$ x_i の変動のうち，u_i と無相関な部分を抽出．

※ 内生説明変数が複数個あれば，各内生説明変数に対しこの作業を行う．

第 2 段階

- 推定したい式の内生説明変数 x_i を，第 1 段階で求めた予測値 $\hat{x}_i$ に変更した式

  $$y_i = \beta_0 + \beta_X \hat{x}_i + \beta_C c_i + u_i$$

  を OLS で推定.
  - $\hat{x}_i$ は u_i と無相関.
- $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_X \\ \beta_C \end{bmatrix}$ の 2SLS 推定量は,

  $$\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}.$$

識別の次数条件

「推定したい式に含まれる内生説明変数の個数」を G とし，「推定したい式に含まれない外生変数の個数」を K とする．

- $G > K$ の場合を過少識別（under-identified）という．
  - 識別不能ともいう，
  - 推定できない（操作変数を使わず無理に OLS 推定をするとバイアス発生）．
- $G = K$ の場合をちょうど識別（just-identified）という．
  - 2SLS などで推定できる．
- $G < K$ の場合を過剰識別（over-identified）という．
  - 2SLS などで推定できる．

$\Rightarrow G \leq K$ なら 2SLS などで推定できる．

$$\Downarrow$$

- 識別の次数条件（order condition）は，

$$G \leq K.$$

操作変数の例

被説明変数	内生説明変数	操作変数
所得	修学年数	最寄り大学までの距離
労働時間	子ども数	最初の 2 人の子どもの性別
修学年数	婚外子数	双子の有無
健康状態	検診回数	病院までの距離
子どもの出生児体重	母親の喫煙数	タバコの値段
財の需要量	財の価格	原材料価格
消費	GDP	政府支出

2SLS における決定係数

- モデルの説明変数と誤差項に相関がある場合，決定係数も自由度修正済み決定係数も適切に定義できない（証明は省略）.
- 2SLS を用いる目的
  ➡ モデルの説明変数と誤差項に相関がある場合に，より厳密な係数推定値を得るため（モデルの当てはまりの良さを高めるためではない.）

$$\Downarrow$$

2SLS の第 2 段階推定における R^2 や $\bar{R}^2$ は，解釈ができない.

- 参考：Wooldridge, J.M., 2019. *Introductory Econometrics: A Modern Approach*. Seventh ed., South-Western, Mason, OH, USA, p.505.

ミンサー方程式の 2SLS 推定

既婚男性 734 人のデータを用い，ミンサー方程式

$$\ln income_i = \beta_0 + \beta_Y yeduc_i + \beta_E exper_i + \beta_{EE} exper_i^2 + u_i$$

- $income_i$: 年収（万円）
- $yeduc_i$: 修学年数（年）
- $exper_i$: 就業可能年数（年）
- i : 個人番号

の説明変数のうち，$exper_i$ と $exper_i^2$ は u_i と相関しない外生変数であるが，$yeduc_i$ は u_i に含まれる観測できない要因（個人の能力など）と相関している可能性があり，u_i と相関する内生変数であるとして，2SLS で推定することを考える．

以下の変数を操作変数として用いる.

- $payeduc_i$: 父親の修学年数（年）
- $sibs_i$: 兄弟姉妹数（人）

どちらも，本人の修学年数と相関するが，本人の能力（が動かす本人の年収）とは無相関のため，外生変数であると考えられる.

$$\Downarrow$$

推定したい式に含まれる内生説明変数は $yeduc_i$ の 1 個，推定したい式に含まれない外生変数は $payeduc_i$ と $sibs_i$ の 2 個なので，過剰識別であり，識別の次数条件を満たす（2SLS で推定できる).

gretl での 2 段階最小二乗法

メニューバーから「モデル」→「操作変数法」→「2 段階最小二乗法」と操作し，

- 「従属変数」には推定したい式の被説明変数を，
- 「説明変数（回帰変数）」には推定したい式の右辺に含まれる説明変数（内生変数と外生変数両）を，
- 「操作変数」にはシステムに登場する全ての外生変数を，

それぞれ選び，標準誤差を設定して「OK」をクリックすればよい．

ただし，第 1 段階推定（内生説明変数を，システムに登場する全ての外生変数に回帰）の結果は表示されない．

$$\Downarrow$$

第 1 段階推定は，メニューバーから「モデル」→「通常の最小二乗法」と操作し，

- 「従属変数」には推定したい式の内生説明変数を，
- 「説明変数（回帰変数）」にはシステムに登場する全ての外生変数を，

それぞれ選んで結果を出力すればよい．

例えば，ミンサー方程式の場合は，

- 推定したい式の被説明変数
  - lincome
- 推定したい式の右辺に含まれる説明変数
  - yeduc（内生）
  - exper（外生）
  - exper2（外生）
- システムに登場する全ての外生変数
  - payeduc
  - sibs
  - exper
  - exper2

2SLS では第 2 段階において，内生説明変数そのものの代わりに，第 1 段階で推定された結果を使った内生説明変数の予測値を用いている.

$$\Downarrow$$

予測値には測定誤差が含まれる.

$$\Downarrow$$

それを考慮したうえで係数の標準誤差を求める必要がある（詳細な説明は省略）.

$$\Downarrow$$

gretl でメニューバーから「モデル」→「操作変数法」→「2 段階最小二乗法」と操作して 2SLS を実行すると，係数の標準誤差の調整が自動的に行われる.

第 1 段階推定結果

gretl: モデル1

ファイル　編集(E)　検定(T)　保存(S)　グラフ(G)　分析(A)　LaTeX

```
モデル 1: 最小二乗法(OLS), 観測: 1-734
従属変数: yeduc
不均一分散頑健標準誤差, バリアント HC1

             係数          標準誤差        t値          p値
  ---------------------------------------------------------
  const       12.3626       0.615730      20.08      1.03e-071 ***
  payeduc      0.204426     0.0267576      7.640     6.85e-014 ***
  sibs        -0.234284     0.0854614     -2.741     0.0063    ***
  exper        0.149594     0.0724683      2.064     0.0393    **
  exper2      -0.0108400    0.00241815    -4.483     8.55e-06  ***

Mean dependent var   14.13896   S.D. dependent var   2.047800
Sum squared resid    2252.604   回帰の標準誤差             1.757837
R-squared            0.267166   Adjusted R-squared   0.263145
F(4, 729)            104.0050   P-value(F)           4.49e-70
Log-likelihood      -1453.030   Akaike criterion     2916.060
Schwarz criterion    2939.053   Hannan-Quinn         2924.929
```

第 2 段階推定結果

gretl: モデル2

ファイル　編集(E)　検定(T)　保存(S)　グラフ(G)　分析(A)　LaTeX

```
モデル 2: 二段階最小二乗法(2SLS), 観測: 1-734
従属変数: lincome
内生変数(instrumented): yeduc
操作変数: const payeduc sibs exper exper2
不均一分散頑健標準誤差, バリアント HC1

               係数          標準誤差        t値          p値
  ------------------------------------------------------------
  const        4.52414       0.325281       13.91      3.51e-039 ***
  exper        0.0609592     0.0156393       3.898     0.0001    ***
  exper2      −0.00104174    0.000590447    −1.764     0.0781    *
  yeduc        0.0699093     0.0218722       3.196     0.0015    ***

Mean dependent var     6.170857   S.D. dependent var    0.356020
Sum squared resid      70.30899   回帰の標準誤差              0.310344
R-squared              0.246920   Adjusted R-squared    0.243825
F(3, 730)              25.21227   P-value(F)            1.58e-15

ハウスマン(Hausman)検定 -
  帰無仮説: OLS推定値は一致性を持つ
  漸近的検定統計量: カイ二乗(1) = 0.4871
  なお、p値(p-value) = 0.485224

Sarganの過剰識別検定 -
  帰無仮説: 全ての操作変数は有効(valid)である
  検定統計量: LM = 0.403198
  なお、p値(p-value) = P(カイ二乗(1) > 0.403198) = 0.525442

弱操作変数(weak instrument)の検定 -
  第1段階のF統計量 (2, 729) = 34.692
  10未満の値は、弱操作変数を示していると思われます
```

第 1 段階推定結果

- 父親の修学年数の係数
  - 0.204426
  - t 値は 7.64, p 値は 6.85×10^{-14}.
    ➡ 仮に「payeduc の係数が 0」だとすると，7.64 という t 値は 6.85×10^{-14}, つまりほぼ 0%の確率（1%を下回る確率）でしか出てこない.
    ➡ 有意水準 1%で，「係数は 0」の H_0 が棄却される（5%や 10%でも棄却される）.
    ➡ 父親の修学年数は（本人の）修学年数と統計的に有意に相関している.

- 兄弟姉妹数の係数
  - −0.234284
  - t 値は −2.741, p 値は 0.0063.
    ➡ 仮に「sibs の係数が 0」だとすると，−2.741 という t 値は 0.63%の確率（1%を下回る確率）でしか出てこない．
    ➡ 有意水準 1%で，「係数は 0」の H_0 が棄却される（5%や 10%でも棄却される）．
    ➡ 兄弟姉妹数は修学年数と統計的に有意に相関している．

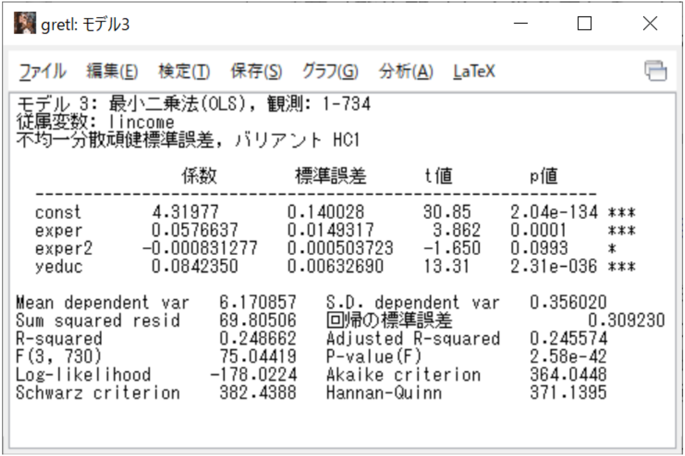
⇒「ミンサー方程式に含まれない外生変数」とした「父親の修学年数」も「兄弟姉妹数」も，「ミンサー方程式の説明変数に含まれる内生変数」の「(本人の）修学年数」と統計的に有意に相関している．
⇒「父親の修学年数」も「兄弟姉妹数」も操作変数として機能している可能性がある（より一般的な判断方法の説明は省略）．

第 2 段階（ミンサー方程式）推定結果

- 修学年数の係数
  - 0.0699093（符号は正）
  - t 値は 3.196, p 値は 0.0015.
    ➡ 仮に「yeduc の係数が 0」だとすると，3.196 という t 値は 0.15%の確率（1%を下回る確率）でしか出てこない.
    ➡ 有意水準 1%で，「係数は 0」の H_0 が棄却される（5%や 10%でも棄却される）.
    ➡ 修学年数は年収と統計的に有意に相関している.
  - 年収と修学年数についてはログ＝レベル・モデルの関係.
    ➡ 就業可能年数（とその 2 乗）を一定としたうえで，修学年数が 1 年長くなると，年収が平均して 6.99093%高くなる傾向がある.

ミンサー方程式の OLS 推定結果

このデータで，ミンサー方程式を OLS で推定した結果は以下のとおり．



gretl: モデル3

ファイル 編集(E) 検定(T) 保存(S) グラフ(G) 分析(A) LaTeX

```
モデル 3: 最小二乗法(OLS), 観測: 1-734
従属変数: lincome
不均一分散頑健標準誤差, バリアント HC1

              係数          標準誤差       t値        p値
  ---------------------------------------------------------------
  const       4.31977       0.140028      30.85     2.04e-134 ***
  exper       0.0576637     0.0149317      3.862    0.0001    ***
  exper2     -0.000831277   0.000503723   -1.650    0.0993    *
  yeduc       0.0842350     0.00632690    13.31     2.31e-036 ***

Mean dependent var   6.170857   S.D. dependent var   0.356020
Sum squared resid    69.80506   回帰の標準誤差            0.309230
R-squared            0.248662   Adjusted R-squared   0.245574
F(3, 730)            75.04419   P-value(F)           2.58e-42
Log-likelihood      -178.0224   Akaike criterion     364.0448
Schwarz criterion    382.4388   Hannan-Quinn         371.1395
```

OLS と 2SLS での，修学年数の係数推定値の違い

- OLS では，修学年数の係数は 0.084235 なので，就業可能年数（とその 2 乗）を一定としたうえで，修学年数が 1 年長くなると，年収が平均して 8.4235%高くなる傾向がある，という解釈となる．
- 2SLS では，修学年数の係数は 0.0699093 なので，就業可能年数（とその 2 乗）を一定としたうえで，修学年数が 1 年長くなると，年収が平均して 6.99093%高くなる傾向がある，という解釈となる．

$$\Downarrow$$

- OLS では修学年数が年収に与える定量的な効果がやや過大に計測されている可能性がある.
- ただし，推定結果の画面より，ハウスマン検定では「OLS 推定値は一致性をもつ」という H_0 が有意水準 10%でも棄却されないため，この場合は 2SLS 推定を行う意味がなく，OLS 推定が支持される（詳細な説明は省略）.

今日のキーワード

外生変数，内生変数，内生性バイアス，操作変数，操作変数法，2 段階最小二乗法，過少識別，ちょうど識別，過剰識別，次数条件

次回までの準備

- 今回の講義スライドを読み直す.
- 「提出課題 8」に取り組む.
- 教科書第 9 章第 1 節，第 3 節～第 4 節を読む.