第13章
内生性と操作変数

藪友良

『入門　実践する計量経済学』

(東洋経済新報社)

PPT

- 内生変数と外生変数

- 内生性の原因

  ①測定誤差、②同時方程式、③欠落変数

- 高頻度データの使用---内生性の解決①

- 2段階最小2乗法---内生性の解決②

- 2段階最小2乗法の詳細

内生変数と外生変数

- $Y_i = \alpha + \beta X_i + u_i$

  内生性　<=>　$\boldsymbol{Cov(X_i, u_i) \neq 0}$

  外生性　<=>　$\boldsymbol{Cov(X_i, u_i) = 0}$

  ---**内生変数**:誤差項と相関している変数

  外生変数:誤差項と無相関となる変数

- **説明変数に内生性があると、OLS推定量はバイアスが生じる**

- **内生性の原因**

  ①測定誤差、②同時方程式、②欠落変数

バイアス

・　説明変数が確率変数でも、説明変数と誤差項が無相関なら（外生性）一致性は満たされる。説明変数と誤差項に相関があると（内生性）、不偏性も一致性も満たされない。

[証明]

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u$$

X_1を1単位増やすと

$$Y' = \alpha + \beta_1 (X_1 + 1) + \beta_2 X_2 + \cdots + \beta_K X_K + u'$$

このとき、Yの変化分は

$$Y' - Y = [\alpha + \beta_1 (X_1 + 1) + \beta_2 X_2 + \cdots + \beta_K X_K + u']$$
$$-[\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u] = \beta_1 + \underbrace{(u' - u)}_{\boxed{\text{バイアス}}}$$

内生性の原因①
測定誤差

- 測定誤差:データの測定に伴う誤差

--- 測定が困難な変数

例) 生まれつきの能力を測る指標(IQも問題がある)

--- 記入・入力ミス

- 説明変数に測定誤差があると、内生性が生じる

[証明]

$$Y_i = \alpha + \beta X_i^* + u_i^*$$

--- X_i^*とu_i^*は無相関となる

--- X_i^* は観察できず、測定誤差e_iを含んだX_iが観察される

(測定誤差は期待値0とする)

$$X_i = X_i^* + e_i$$

--- Y_iをX_iで回帰すると、X_iと誤差項u_iに相関が生じる

$$\begin{aligned} Y_i &= \alpha + \beta X_i^* + u_i^* \\ &= \alpha + \beta(X_i - e_i) + u_i^* \\ &= \alpha + \beta X_i + \underbrace{(u_i^* - \beta e_i)}_{u_i} \end{aligned}$$

$$\boxed{X_i^* = X_i - e_i}$$

- βのOLS推定は0の方向でバイアスが生じる

--- 説明変数$X_i = X_i^* + e_i$の測定誤差が大きくなると、

X_iとY_iとの関係が弱くなる

元のモデル:　$Y_i = \alpha + \beta X_i^* + u_i^*$

回帰モデル:　$Y_i = \alpha + \beta X_i + u_i$

内生性の原因②
同時方程式

- **同時方程式：被説明変数と説明変数が同時決定の関係にある**

- **同時決定の関係にあるとき、内生性が生じる**

$$①\ Y_i = \alpha_Y + \beta_Y X_i + u_{Yi}$$

$$②\ X_i = \alpha_X + \beta_X Y_i + u_{Xi}$$

（例） 投獄率を1%増やすと犯罪率は何$\%$減るか

$$Y_i = \alpha_Y + \beta_Y X_i + u_{Yi}$$

--- 犯罪率Y_i: 10万人当たりの犯罪件数

--- 投獄率X_i: 10万人当たりの囚人数

しかし、犯罪者が増えると投獄される人数も増える

$$X_i = \alpha_X + \beta_X Y_i + u_{Xi}$$

内生性の原因③
欠落変数

- **欠落変数：含めるべき説明変数を定式化に含めない**

--- 観察できない変数がある

例)　生まれつきの能力は観察できない（IQも利用可能ではない）

- **欠落変数があると内生性が生じる**

[証明]

$$Y_i = \alpha + \beta X_i + \theta W_i + u_i^*$$

--- W_iは観察できないため、単回帰分析を行う

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

$$u_i = \theta W_i + u_i^*$$

--- X_iとW_iに相関があると、X_iとu_iにも相関が生じる

⇒　W_iを除くとβの推定にバイアスが生じる（欠落変数バイアス）

--- X_iとW_iに相関がないなら、X_iとu_iに相関は生じない

⇒　W_iを除いてもβの推定にバイアスはない

どの原因から内生性が生じるか

- ①測定誤差、②同時方程式、③欠落変数、のどれが深刻かは分析対象によって異なる

- ミクロデータなら、①②③のいずれもが重要となる
  --- 個票データは膨大で、そのチェックが不十分なことがあり、測定誤差が生じる
  --- 観察できない変数（能力、企業文化など）が多く、欠落変数が生じやすい

- マクロデータなら、②がとくに重要となる
  --- 政府や日銀はデータを慎重に記入・入力しており、測定誤差の問題は小さい傾向がある
  --- マクロ変数は同時決定しているものが多い

内生性の解決①
高頻度データの使用

- 観察頻度の高いデータ

--- 年次⇒四半期⇒月次⇒週次⇒日次⇒秒次

- 同時方程式の問題は、観察頻度を上げると解決できるかもしれない

$$Cov(X_t, u_t) = 0$$

例) 財政政策の効果

・年次データを用いる場合

政府支出(X_t) はGDP(Y_t)に影響を与えるが、GDPは政府支出に影響を与える

$$①Y_t = \alpha_Y + \beta_Y X_t + u_{Yt}$$

$$②X_t = \alpha_X + \beta_X Y_t + u_{Xt}$$

$$\Rightarrow Cov(X_t, u_{Yt}) \neq 0$$

・四半期データを用いる場合

政府が政策を決定し、実行するまで1四半期以上かかる

--- 景気の悪化を認知⇒財政支出を拡大する法案作成

⇒国会審議⇒法案可決⇒財政支出の増加

$$①Y_t = \alpha_Y + \beta_Y X_t + u_{Yt}$$

$$②X_t = \alpha_X + \beta_X Y_{t-1} + u_{Xt}$$

$$\Rightarrow Cov(X_t, u_{Yt}) = 0$$

写真の出所
https://www.irasutoya.com/2014/01/blog-post_356.html

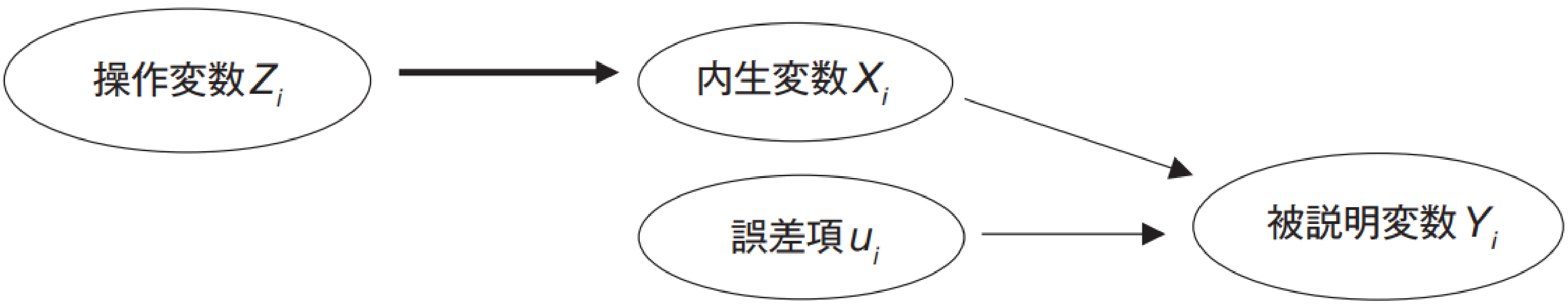
内生性の解決②
2段階最小2乗法

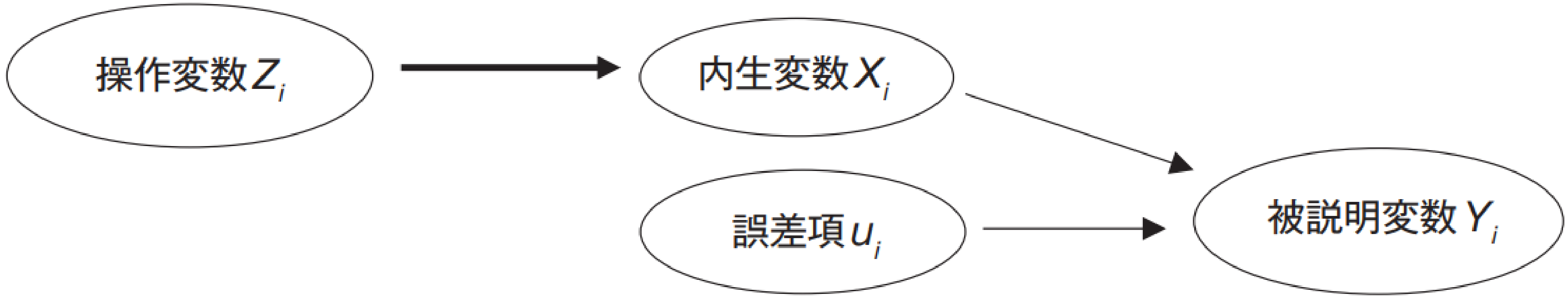
$$Y_i = \alpha + \beta X_i + u_i$$
$$Cov(X_i, u_i) \neq 0$$

- 操作変数Z_i は以下の条件を満たす変数である

① $\boldsymbol{Cov(Z_i, X_i) \neq 0}$ (関連性)
② $\boldsymbol{Cov(Z_i, u_i) = 0}$ (外生性)

図13 - 1　操作変数と内生変数との関係――単回帰分析の場合





例(勉強時間とGPA):勉強時間を増やすとGPAは改善するか？

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

--- Y_i :GPA（大学1年生1学期）、X_i :勉強時間（1日平均）

--- 能力W_iが含まれていないので、欠落変数の問題が生じる

--- 操作変数$\boldsymbol{Z_i}$:ルームメートがゲーム機を持ち込んだら1となるダミー変数（大学は全員が寮生活し、部屋の割り振りはランダムに）

--- 部屋の割り振りはランダムなので、Z_iとu_iは無相関

（$Cov(Z_i, u_i) = 0$）

--- ゲーム機が持ち込まれると勉強時間は減る（$Cov(Z_i, X_i) < 0$）

--- 2段階最小2乗法をすると、1時間勉強の時間を増やすとGPAは0.36改善した

2段階最小2乗法（$Y_i = \alpha + \beta X_i + u_i$）

- **1段階：X_iをZ_iで回帰して、予測値をもとめる**

$$\widehat{X}_i = \widehat{\gamma_0} + \widehat{\gamma}_1 Z_i \text{ 、また } X_i = \widehat{X}_i + \widehat{e}_i$$

--- $H_0: \gamma_1 = 0$が採択されたら、$Cov(Z_i, X_i) \neq 0$が満たされない

--- 予測値$\hat{X}_i$と誤差項u_iは無相関(Z_iとu_iが無相関であるため)

--- 予測値$\hat{X}_i$と残差$\hat{e}_i$は無相関

[証明] 2章より、残差の性質は

①残差の和は0（$\sum_{i=1}^{n} \hat{e}_i = 0$）

②残差と説明変数の積和は0（$\sum_{i=1}^{n} Z_i \hat{e}_i = 0$）

$$\sum_{i=1}^{n} \hat{X}_i \hat{e}_i = \sum_{i=1}^{n} (\hat{\gamma}_0 + \hat{\gamma}_1 Z_i) \hat{e}_i = \hat{\gamma}_0 \underbrace{\sum_{i=1}^{n} \hat{e}_i}_{=0} + \hat{\gamma}_1 \underbrace{\sum_{i=1}^{n} Z_i \hat{e}_i}_{=0} = 0$$

この結果から、$\hat{X}_i$と$\hat{e}_i$の標本共分散の分子は0になる

$$\sum_{i=1}^{n} (\hat{X}_i - \bar{\hat{X}})(\hat{e}_i - 0) = \sum_{i=1}^{n} \hat{X}_i \hat{e}_i - \bar{\hat{X}} \sum_{i=1}^{n} \hat{e}_i = 0$$

- **2段階: $\boldsymbol{Y_i}$を$\boldsymbol{\widehat{X_i}}$で回帰する**

$$Y_i = \alpha + \beta \widehat{X}_i + u_i^*$$

--- $\widehat{X}_i$とu_i^*は無相関になる（推定量$\hat{\beta}_{2SLS}$は、βの一致推定量）

[証明]

$$Y_i = \alpha + \beta X_i + u_i = \alpha + \beta\left(\hat{X}_i + \hat{e}_i\right) + u_i$$

$$= \alpha + \beta \hat{X}_i + (\underbrace{u_i + \beta \hat{e}_i}_{u_i^*})$$

既に示したとおり、

--- 予測値$\hat{X}_i$と誤差項u_iは無相関

--- 予測値$\hat{X}_i$と残差$\hat{e}_i$は無相関

（例）出産と労働時間との関係

- 追加的に子供を1人出産することで、出産後の労働時間が何時間減少するか
- 2人以上の子供がいる既婚女性のデータ（1980年のU.S. census）
- 変数の定義

  Y :母親の労働時間（1979年に何週働いたか）

  X : 子供が2人より多い母親なら1、子供が2人だけなら0

- **OLS推定**

$$\hat{Y} = 21.068 - 5.387X$$

$$(0.056) \quad (0.087)$$

--- 労働時間が少ない人が子供を産む可能性？ （同時方程式）

- **OLS推定**

$$\hat{Y} = 21.068 - 5.387X$$

- **操作変数**

Z: 同性ダミー(最初の2人の子供が同性なら1となる)

--- 最初の2人が同性かはランダム（$Cov(Z_i, u_i) = 0$）

--- 最初の2人が同性だと子供をさらに欲しがる傾向あり（$Cov(Z_i, X_i) > 0$）

- **2段階最小2乗法**

第1段階: $\hat{X} = 0.346 + 0.068Z$

(0.001)　(0.002)

第2段階: $\hat{Y} = 21.421 - 6.314\hat{X}$

(0.487)　(1.275)

2段階最小2乗法の詳細

一般的なケース

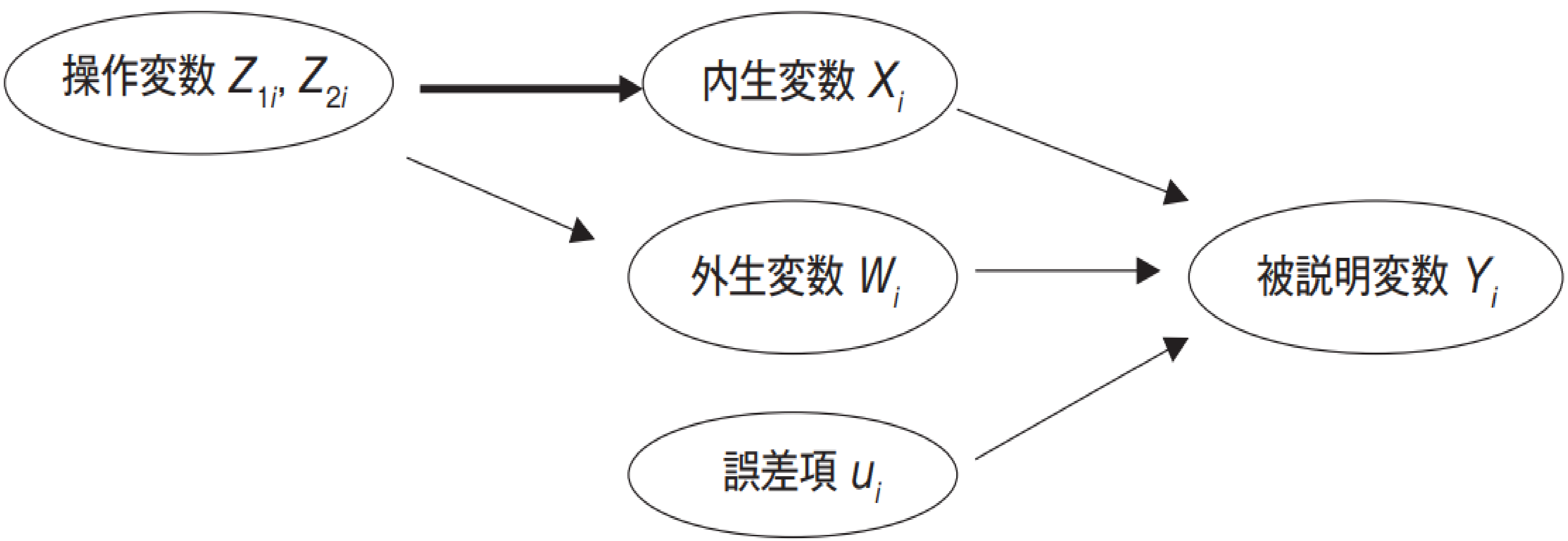
$$Y_i = \alpha + \beta X_i + \theta W_i + u_i$$

--- X_iは内生変数（$Cov(X_i, u_i) \neq 0$ ）

--- Wは外生変数（$Cov(W_i, u_i) = 0$ ）

--- Z_{1i} 、Z_{2i}は操作変数

図13 - 2　操作変数と内生変数との関係——重回帰分析の場合



- **2段階最小2乗法**

  1段階:　X_iを操作変数と外生変数で回帰する

  $$\widehat{X}_i = \widehat{\gamma_0} + \widehat{\gamma_1} Z_{1i} + \widehat{\gamma_2} Z_{2i} + \widehat{\gamma_3} W_i$$

  2段階:　Y_iを$\widehat{X}_i$と外生変数で回帰する

  $$Y_i = \alpha + \beta \widehat{X}_i + \theta W_i + u_i^*$$

  ---第1段階では、操作変数だけでなく、外生変数(W_i)を含めること（さもないと、2段階目で外生変数と誤差項u_i^*が相関してしまう）

- **条件① (関連性)の確認**

--- 1段階の結果を確認し、XとZが関係していることを示す

$$\widehat{X}_i = \widehat{\gamma_0} + \widehat{\gamma_1} Z_{1i} + \widehat{\gamma_2} Z_{2i} + \widehat{\gamma_3} W_i$$

--- $H_0 : \gamma_1 = \gamma_2 = 0$

経験則：操作変数の係数が全て0としたF統計量が10を超えるかを確認する(10を下回ると弱相関操作変数の可能性あり)

- **条件② (外生性)の確認**

--- 操作変数Zと誤差項uに関係がないことを、説得的に説明する
--- 操作変数が2個以上なら、条件②の妥当性を確認できる

方法(1): 操作変数を変えても、βの推定結果が安定しているか

--- 操作変数Z_1、Z_2があるとき、Z_1だけを使うと$\hat{\beta}_{2SLS}$、Z_2だけを使うと$\hat{\beta}'_{2SLS}$となるとしよう

--- $\hat{\beta}_{2SLS}$と$\hat{\beta}'_{2SLS}$の値が大きく異なるなら、どちらかの操作変数が仮定を満たしてないか、両方の操作変数が仮定を満たしていない

--- $\hat{\beta}_{2SLS}$と$\hat{\beta}'_{2SLS}$の値がほぼ同じなら、操作変数は適切と判断される

方法(2) 過剰識別制約の検定 (サーガン検定)

① 全ての操作変数を用いて2SLSを行う。そして、予測値と残差を求める

$$\widehat{Y_i} = \hat{\alpha}_{2SLS} + \hat{\beta}_{2SLS} X_i$$
$$\widehat{u_i}^{2SLS} = Y_i - \widehat{Y_i}$$

② 残差を操作変数と外生変数で回帰し、操作変数の外生性が満たされているかを検定する(この場合、外生変数なし)

$$\widehat{u_i}^{2SLS} = \delta_0 + \delta_1 Z_{1i} + \delta_2 Z_{mi}$$

$H_0 : \delta_1 = \delta_2 = 0$（操作変数と誤差項が無相関）

--- $H_0: \delta_1 = \delta_2 = 0$が棄却されたら、
操作変数は不適切と判断される

--- $H_0: \delta_1 = \delta_2 = 0$を採択したら、
操作変数は適切と判断される

（例）制度と経済成長との関係

・財産権が保護されている国は経済成長するのか？

---北朝鮮と韓国から明らかだが、ここでは数量的に評価したい

・ 昔植民地であった64カ国のデータ

保護指数：財産権の保護指数、1～10で評価される（アメリカは10）

GDP: 一人当たりGDPの対数（1995年）

植民者死亡率: 植民地時代の植民者死亡率の対数

緯度: 赤道を0とし、南極と北極は0.9(90度) と表記する

・ **定式化**　$\boldsymbol{GDP} = \boldsymbol{\alpha} + \boldsymbol{\beta_1}$ 保護指数 $+\boldsymbol{\beta_2}$ 緯度 $+\boldsymbol{u}$

--- 保護指数と誤差項とは相関がある

① 豊かな国ほど財産権を保護する余裕がある（同時方程式）

② 財産権の数値化が難しく、測定誤差が発生する（測定誤差）

・**OLS推定**

GDP = 4.73 + 0.47保護指数 +1.58 緯度

(0.39)　(0.06)　　　　　(0.71)

- **操作変数Z：　植民者死亡率**

---植民者死亡率が高い地域は、宗主国からの植民が進まなかった

現地の法制度は搾取的となり、現在の制度に悪影響を与えた

$$Cov(\text{植民者死亡率}_i, \text{保護指数}_i) < \mathbf{0}$$

---　100年以上前の死亡率は現在のGDPへの直接的影響はない

$$Cov(\text{植民者死亡率}_i, u_i) = \mathbf{0}$$

・2段階最小2乗法（操作変数：植民者死亡率）

1段階: $\widehat{保護指数} = 8.52 - 0.51$ 植民者死亡率 $+2.00$ 緯度

(0.81) (0.14) (1.33)

--- 植民者死亡率は1%有意、また係数0としたF値は13.1（>10）

2段階: GDP = 1.69 + 0.99 $\widehat{保護指数} - 0.65$ 緯度

(1.29) (0.22) (1.34)

--- 財産権の保護の係数は、OLSで0.47、2SLSで0.99である
内生性を考慮することで係数は倍以上になっている

・2段階最小2乗法（操作変数：移民者数）

2段階目： GDP = 2.16 + 0.92 $\widehat{保護指数} - 0.47$ 緯度

(1.17) (0.20) (1.24)

--- 移民者数(1900年)が多いと保護指数は高くなる（関連性）

--- 当時の移民者数は現在のGDPには影響を与えない（外生性）

--- 操作変数として、どちらを用いても保護指数の係数はほぼ同じ

まとめ

- 内生性とその原因

  ①測定誤差、②同時方程式、③欠落変数

- 高頻度データの使用：解決策①

- 2段階最小2乗法：解決策②

- どうやって操作変数をみつけるか

  --- 経済学、経営学、歴史、制度などを深く理解する

  --- 外生的なショックが内生変数を変化させていないか？

  例) 大地震やコロナショックなど

  --- 操作変数を用いた実証研究を読むことで、操作変数をみつけるコツがつかめる