13章:内生性と操作変数

藪友良

- 内生性
- 内生性の原因
- 解決策1:高頻度データの使用
- 解決策2：操作変数を用いた2段階最小2乗法
- 2段階最小2乗法の注意点

内生性

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

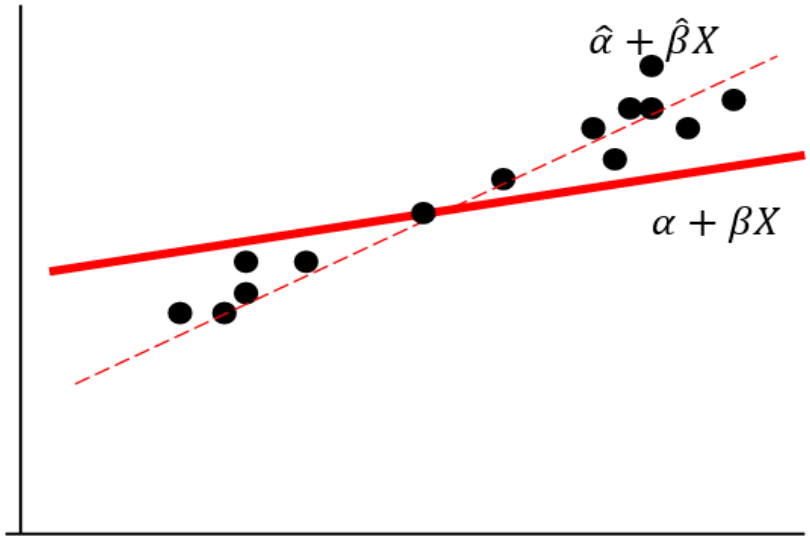
  内生性 <=> $\mathbf{cov}(X_i, u_i) \neq 0$

  外生性 <=> $\mathbf{cov}(X_i, u_i) = 0$

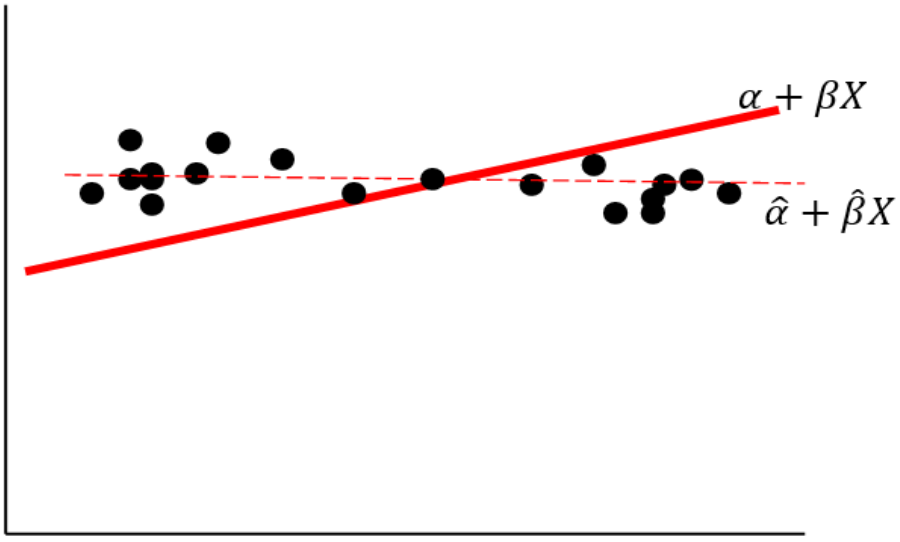
- **内生性があると、OLS推定量はバイアスが生じる**

図 3：内生性とバイアスの関係

(a) X_iとu_iに正の相関



(b) X_iとu_iに負の相関



- **内生性の原因**

  ①測定誤差、②欠落変数、③同時方程式

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) GDP（帰属家賃の推定は、不動産業の付加価値に含まれる）

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

$$X_i = X_i^* + e_i$$

--- 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

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

$X_i^* = X_i - e_i$

- 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

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

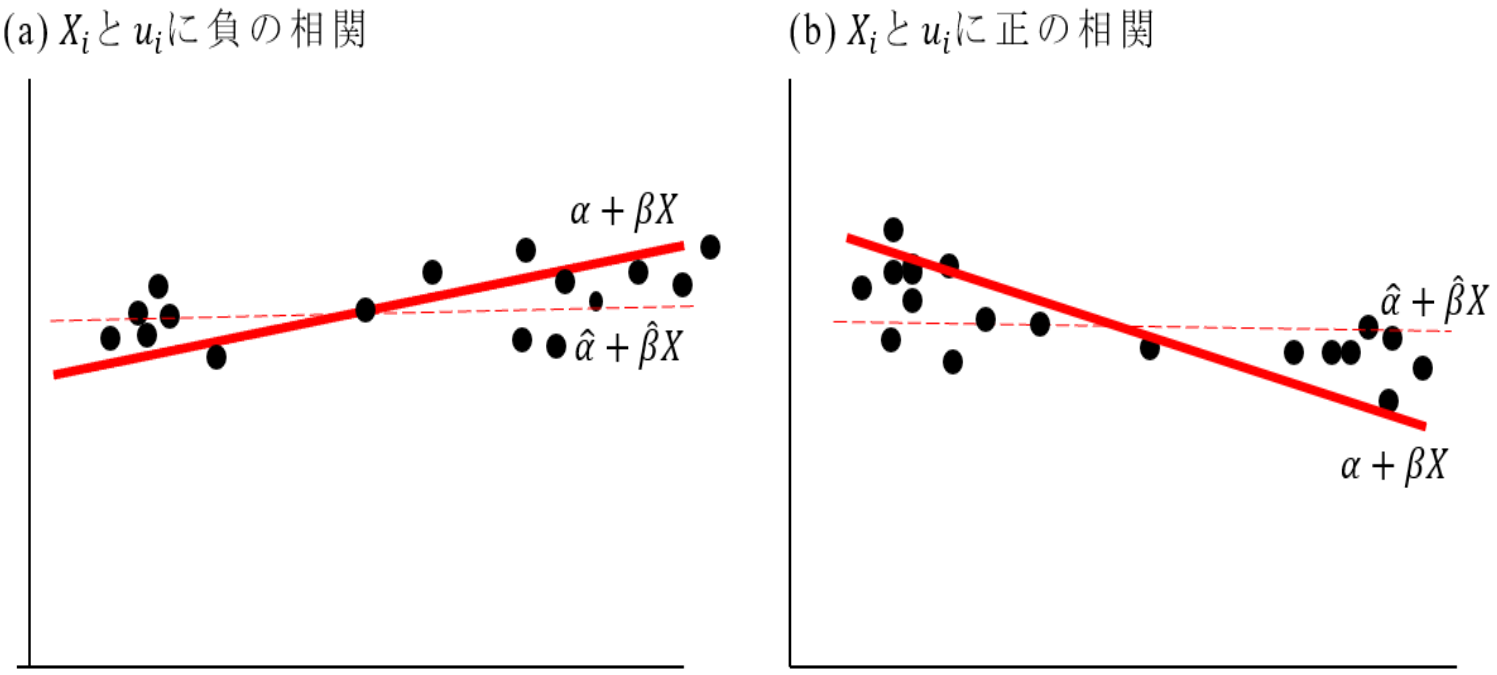
--- $\beta > 0$なら、$X_i = X_i^* + e_i$と$u_i = u_i^* - \beta e_i$に負の相関がある

$\Rightarrow \quad \beta > \hat{\beta}$

--- $\beta < 0$なら、$X_i = X_i^* + e_i$と$u_i = u_i^* - \beta e_i$に正の相関がある

$\Rightarrow \quad \beta < \hat{\beta}$

図 2：内生性とバイアスの関係



②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**

--- 変数間の関係を理論的に考察し、含めるべき変数は含めることが重要

--- 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

③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**

$$①Y_i = \alpha_Y + \beta_Y X_i + u_{Yi}$$
$$②X_i = \alpha_X + \beta_X Y_i + u_{Xi}$$

（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

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

どの要因が深刻なのか

- ①測定誤差、②欠落変数、③同時方程式、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、①測定誤差、②欠落変数、が重要となる
  --- 個票データは膨大であり、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、③同時方程式、が重要となる
  --- 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

解決策1
データの観察頻度をあげる

- 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

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

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

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

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

- 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

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

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

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

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

解決策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 $\mathbf{cov(Z_i, X_i) \neq 0}$ (操作変数の関連性)

② $\mathbf{cov(Z_i, u_i) = 0}$ (操作変数の外生性)

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

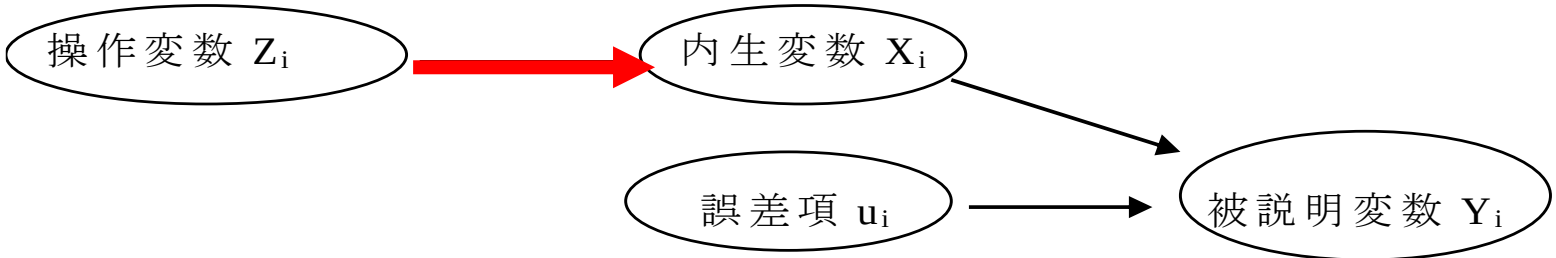
--- 操作変数$\boldsymbol{Z_i}$:ルームメートがゲーム機を持ち込んだら1となるダミー変数

(この大学は、全員が寮生活し、ルームメートの割り振りはランダム)

--- 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($cov(Z_i, X_i) < 0$)

--- 二段階最小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_{1i}で回帰して、予測値をもとめる**

$\widehat{X_i} = \widehat{\gamma_0} + \widehat{\gamma_1} Z_{1i}$ 、また $X_i = \widehat{X_i} + \widehat{e_i}$

--- $\gamma_1 = 0$なら、$\mathrm{cov}(Z_{1i}, X_i) \neq 0$が満たされない

--- 予測値$\hat{X}_i$と誤差項u_iは無相関(Z_{1i}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(証明参照)

[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($\sum_{i=1}^{n} Z_i \hat{e}_i = 0$)

よって、

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

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

- **2段階: Y_iを$\widehat{X_i}$で回帰する**

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

--- $\hat{X}_i$とu_i^*は無相関になる（推定量$\hat{\beta}^{TSLS}$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、

--- $\hat{X}_i$と誤差項u_iは無相関

--- $\hat{X}_i$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

  weeksm1:母親の労働時間（1979年に何週働いたか）

  morekids：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**

$$\widehat{\text{weeksm1}} = 21.068 - 5.387 morekids$$

  --- 働きたくない人が子供を産む可能性？ （同時方程式）
- **2段階最小2乗法**

  samesex: 最初の2人の子供が同性なら1となるダミー変数

--- 最初の2人が同性かはランダム（誤差項とは無相関）

--- 最初の2人が同性だと子供をさらに欲しがる？（$cov(morekids_i, samesex_i) > 0$）

---第1段階：$\widehat{\text{morekids}} = 0.346 + 0.067 samesex$

(0.001)*** (0.0019)***

第2段階：$\widehat{\text{weeksm1}} = 21.421 - 6.314 \widehat{morekids}$

(0.487)*** (1.274)***

内生性を取り除くことで、子供の労働時間への負の効果は大きくなる

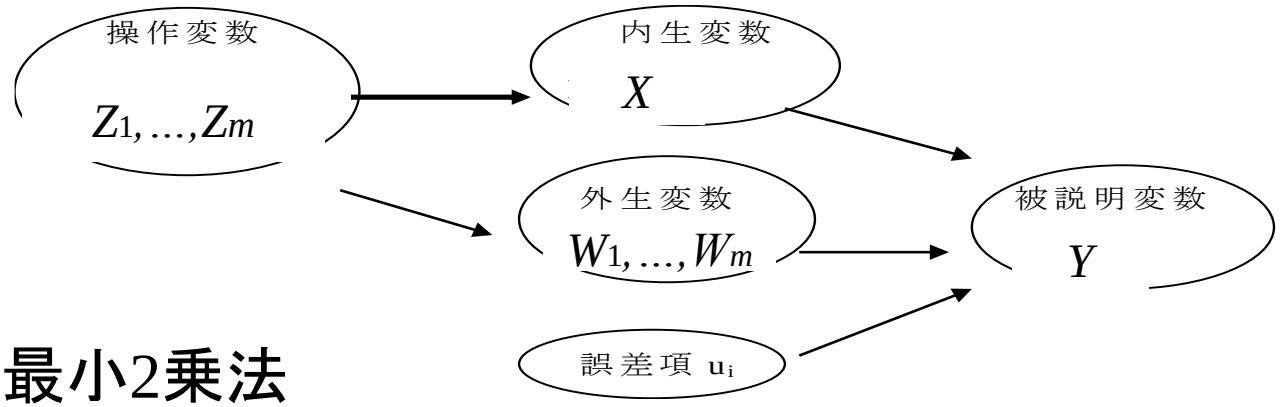
一般的ケース

- $Y_i = \alpha + \beta X_i + \theta_1 W_{1i} + \cdots + \theta_r W_{ri} + u_i$

  --- X_iは内生変数である（$cov(X_i, u_i) \neq 0$ ）

  --- 外生変数Wはr個ある（$cov(W_{ji}, u_i) = 0$ ）

  --- 操作変数がm個ある（Z_{1i} 、…、 Z_{mi} ）



- 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

  $$\widehat{X_i} = \widehat{\gamma_0} + \widehat{\gamma_1} Z_{1i} + \cdots + \widehat{\gamma_m} Z_{mi} + \widehat{\gamma_{m+1}} W_{1i} + \cdots + \widehat{\gamma_{m+r}} W_{ri}$$

  2段階:　Y_iを$\widehat{X_i}$と外生変数で回帰する

  $$Y_i = \alpha + \beta \widehat{X_i} + \theta_1 W_{1i} + \cdots + \theta_r W_{ri} + u_i^*$$

二段階最小2乗法の注意点

- **前提条件①(操作変数の関連性、$cov(Z_i, X_i) \neq 0$)**

  --- 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

  $$\widehat{X_i} = \widehat{\gamma_0} + \widehat{\gamma_1} Z_{1i} + \cdots + \widehat{\gamma_m} Z_{mi} + \widehat{\gamma_{m+1}} W_{1i} + \cdots + \widehat{\gamma_{m+r}} W_{ri}$$

  --- 経験則：$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_m = 0$（操作変数の係数が全て0）
  としたF統計量が10を超える

- **前提条件②(操作変数の外生性、$cov(Z_i, u_i) = 0$)**

  --- 両者の関係がないことを、言葉を尽くして説得的に説明する

  --- 操作変数が2個以上なら、前提条件②の妥当性を大まかに確認できる

  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を使うと$\hat{\beta}_1^{TSLS}$、Z_2を使うと$\hat{\beta}_2^{TSLS}$となる

  --- $\hat{\beta}_1^{TSLS}$と$\hat{\beta}_2^{TSLS}$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

  --- $\hat{\beta}_1^{TSLS}$と$\hat{\beta}_2^{TSLS}$の値がほぼ同じなら、操作変数はどちらも正しい

（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？**

  ---分断後の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**　$GDP = \alpha + \beta_1$ 保護指数 $+\beta_2$ 緯度 $+u$

  --- 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）

- **操作変数Z：　植民者死亡率**

  --- 植民者死亡率が高い地域は、宗主国からの移民による植民が進まなかった

  現地の法制度は搾取的制度となり、それが現在の制度に悪影響を与えた

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

  --- 100年以上前の死亡率は現在のGDPへの直接的影響はない

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

・**OLS推定**

GDP = 4.73 + 0.47保護指数 +1.58 緯度

(0.39)*** (0.06)*** (0.71)**

・**2段階最小2乗法（操作変数：植民者死亡率）**

1段階: $\widehat{\text{保護指数}}$ = 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 $\widehat{\text{保護指数}}$ − 0.65 緯度

(1.29) (0.22)*** (1.34)

--- 財産権の保護の係数は、OLSで0.47、TSLS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 $\widehat{\text{保護指数}}$ − 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には影響を与えていない（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

2) 過剰識別制約の検定 (J Test)

① 全ての操作変数を用いてTSLSを行う。そして、予測値と残差を求める

$$\widehat{Y_i} = \hat{\alpha}^{TSLS} + \hat{\beta}^{TSLS} X_i$$

$$\widehat{u_i}^{TSLS} = Y_i - \widehat{Y_i}$$

② 残差を操作変数で回帰し、操作変数の外生性が満たされているかを検定する

$$\widehat{u_i}^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi}$$

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

--- H_0（操作変数と誤差項は無相関）が棄却されたら、
操作変数は不適切と判断される

--- H_0（操作変数と誤差項は無相関）を採択したら、
操作変数は適切と判断される

＊ 統計学では、H_0の採択はH_0の正しさを意味しない
（操作変数の有効性を示す弱い結果）

まとめ

- 内生性とその原因

  ①測定誤差、②欠落変数、③同時方程式

- 高頻度データの使用

- 操作変数を用いた2段階最小2乗法

- 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