操作変数法 (1)

別所俊一郎

2006 年 6 月 16 日

Today's attraction

- 操作変数法とはなにか
- 2 段階最小 2 乗推定量の標本分布
- 操作変数法の原理と前提

OLS 推定量が一致性を持たないとき

$E[\mathbf{X}u] \neq 0$ のとき

- omitted variable / errors-in-variables / simultaneous causality...

新しい推定方法

- 説明変数と誤差項が相関を持っているのが問題
- 説明変数の変動を，誤差項と相関する部分と誤差項と相関しない部分に分離
- 説明変数の変動のうち，誤差項と相関しないで変動する部分だけを用いて推定できないか
- 新たな変数（**操作変数 Instrumental Variables**）を用意して推定

説明変数と操作変数が 1 つずつの場合

説明変数を X，操作変数を Z とする．回帰モデルは

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

説明変数 X_i と誤差項 u_i が相関しているために OLS 推定量が一致性を持たないとする．

外生性と内生性

外生変数 exogenous variables 誤差項と相関していない変数（連立方程式体系では外部から与えられる変数だったことから）．

内生変数 endogenous variables 誤差項と相関している変数（連立方程式体系では内部で決定される変数だったことから）．

操作変数とは?

説明変数の変動のうち，誤差項と相関しないで変動する部分だけを抽出するために用いられる変数であり，以下の 2 つの条件を満たさなければならない

Instrument relevance （**corr**$(Z_i, X_i) \neq 0$） 操作変数は説明変数と相関を持ち，説明変数の変動を捉えることができる

Instrument exogeneity （**corr**$(Z_i, u_i) = 0$） 操作変数は誤差項と相関を持たず，説明変数の変動から誤差項と相関する部分を除去することができる．$E[\mathbf{Z}u] = 0$ と書くこともできることから，直交条件（orthogonaliry condition）とも呼ばれる

これらの条件を使って，係数 β_1 の一致推定量を得ることができる

TSLS：2段階最小2乗法

1. 説明変数 X を誤差項と相関する部分と誤差項と相関しない部分に分離する．そのために1段階目の回帰

   $$X_i = \pi_0 + \pi_1 Z_i + v_i$$

   を OLS 推定して，当てはめ値 $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ を得る

   - 操作変数 Z_i は外生変数だから，$\pi_0 + \pi_1 Z_i$ はもとの誤差項 u_i と相関しない
   - 1段階目の回帰式の誤差項 v_i はもとの誤差項 u_i と相関している
   - 1段階目の回帰式の係数 π_0, π_1 は推定されるべきパラメタ

2. 説明変数 X の誤差項と相関しない部分を使って β_0, β_1 を推定する．つまり，Y_i を当てはめ値 $\hat{X}_i$ に回帰して，推定量 $\hat{\beta}_0^{\mathrm{TSLS}}, \hat{\beta}_1^{\mathrm{TSLS}}$ を得る

なぜ操作変数が機能するのか?: 例

Wright 親子（Philip, Sewall）の問題

- 輸入関税をどうかけるか?: 関税は当時の重要な税収源
- 関税の効果の大きさを知るには，輸入財の需要関数・供給関数の形状（弾性値）を知る必要
- 対数-対数で定式化すれば係数が弾力性になるから，回帰式は

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + u_i$$

- 利用可能なデータは，1912-1922 年のアメリカのバター消費量と平均年間価格の年次データ
- 消費量と価格は需要と供給の均衡条件から決定されるから，価格 P_i と誤差項 u_i が相関しているかもしれない

なぜ操作変数が機能するのか ?: 例

- 需要曲線・供給曲線そのものがともにシフトする場合，均衡点が動く.
- 得られる (P_t, Q_t) の組み合わせは，需要曲線・供給曲線そのものがともにシフトしたときの均衡点の軌跡
- そのような (P_t, Q_t) に回帰直線を当てはめても，需要曲線にも供給曲線にも推定したことにならない
- この問題を回避するためには，需要曲線をシフトさせずに供給曲線のみをシフトさせるような第 3 の変数を見つけてくることができれば，需要曲線を推定することができる
- そのような第 3 の変数は存在するのか ?: たとえば天候．供給要因を通じて価格とは相関を持つが，需要要因とは相関しない

なぜ操作変数が機能するのか?: 例

カリフォルニアの公立小学校区のデータを用いた，児童-教師比率とテストの成績との関係

- 省略された変数があり，係数にバイアスがあるかもしれない
- 児童-教師比率と相関し，かつテストの成績に対して児童-教師比率以外の経路では相関しない変数を見つけることができれば，児童-教師比率だけの効果を抽出できる
- 外生的に児童-教師比率だけを変化させるような要因：地震によるクラス編成の変更

TSLS 推定量の公式，標本分布

- 小標本での分布（exact な分布）を求めるのは複雑で困難
- 大標本での分布（漸近分布）を求めるのは比較的容易：一致性を持ち正規分布に従う
- 2 段階最小 2 乗推定量の公式も一般にはむずかしい
- 説明変数と操作変数が 1 つずつの場合

$$\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{zy}}{s_{zx}} = \frac{Z \text{ と } Y \text{ の標本共分散}}{Z \text{ と } X \text{ の標本共分散}}$$

$$\left(\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{\hat{x}y}}{s_{\hat{x}}^2} = \frac{\hat{\pi}_1 s_{zy}}{\hat{\pi}_1^2 s_z^2} = \frac{s_{zy}}{\hat{\pi}_1 s_z^2} = \frac{s_{zy}}{\frac{s_{zx}}{s_z^2} s_z^2} = \frac{s_{zy}}{s_{zx}} \right)$$

TSLS 推定量の一致性

- 操作変数の妥当性・外生性と，標本共分散の一致性
- 一致性の説明

$$\mathrm{cov}(Z_i, Y_i) = \mathrm{cov}(Z_i, \beta_0 + \beta_1 X_i + u_i) = \beta_1 \mathrm{cov}(Z_i, X_i) + \mathrm{cov}(Z_i, u_i)$$

操作変数が外生であれば $\mathrm{cov}(Z_i, u_i) = 0$ だから変形すると

$$\beta_1 = \frac{\mathrm{cov}(Z_i, Y_i)}{\mathrm{cov}(Z_i, X_i)}$$

いま，標本共分散が一致性を持ち，

$$s_{zy} \xrightarrow{p} \mathrm{cov}(Z_i, Y_i), \quad s_{zx} \xrightarrow{p} \mathrm{cov}(Z_i, X_i)$$

であれば,

$$\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{zy}}{s_{zx}} \xrightarrow{p} \frac{\mathrm{cov}(Z_i, Y_i)}{\mathrm{cov}(Z_i, X_i)} = \beta_1$$

TSLS 推定量の漸近正規性

中心極限定理を用いると，4 次モーメントについての適切な仮定のもとで,

$$\hat{\beta}_1^{\mathrm{TSLS}} \xrightarrow{d} N\left(\beta_1, \hat{\sigma}^2_{\hat{\beta}_1^{\mathrm{TSLS}}}\right)$$

ここで，$\sqrt{\hat{\sigma}^2_{\hat{\beta}_1^{\mathrm{TSLS}}}}$ が標準誤差であり,

$$\hat{\sigma}^2_{\hat{\beta}_1^{\mathrm{TSLS}}} = \frac{1}{n}\frac{\mathrm{var}([Z_i - \mu_z]u_i)}{(\mathrm{cov}(Z_i, X_i))^2}$$

ここから求められる標準誤差を用いて，t 検定や信頼区間の形成が可能.

2段階最小2乗法の例：タバコ消費の価格弾力性

- タバコの費用（外部性）を抑える方法のひとつがタバコ税
- タバコ税の効果の大きさは，タバコ消費の価格弾力性に依存
- 価格弾力性を推定するには，$\ln Q$ を $\ln P$ に回帰して OLS 推定すればいいというものではない
- データは 1985-95 年のアメリカ 48 州の年次データ
- $\ln P$ への適切な操作変数は？
  - 一般的な売り上げ税収に占めるタバコ税収の比率
  - 税率が高いほど税収が多くなるので価格と相関する
  - 供給要因とは相関しない

一般的なIV回帰モデル

変数が4種類

- **被説明変数** （Y_i）
- **内生性のある説明変数** （X_i）k 個.
- **説明変数に含まれる外生変数** （W_i）誤差項と相関しない．r 個.
- **操作変数** （Z_i）m 個.

Z_i は少なくとも X_i と同じだけの数が必要.

- $m = k$ **のとき** 丁度識別（exactly identified）
- $m > k$ **のとき** 過剰識別（overidentified）
- $m < k$ **のとき** 識別できない（underidentified）

識別条件とは？

直交条件がすべて成り立つとき，代数の基本定理から，$m \geq k$ でないと推定できない.

$$E\begin{bmatrix} Z_1u \\ \vdots \\ Z_mu \\ W_1u \\ \vdots \\ W_ru \end{bmatrix} = E\begin{bmatrix} Z_1(y-\mathbf{X}\beta_{\mathbf{x}}-\mathbf{W}\beta_{\mathbf{w}}) \\ \vdots \\ Z_m(y-\mathbf{X}\beta_{\mathbf{x}}-\mathbf{W}\beta_{\mathbf{w}}) \\ W_1(y-\mathbf{X}\beta_{\mathbf{x}}-\mathbf{W}\beta_{\mathbf{w}}) \\ \vdots \\ W_r(y-\mathbf{X}\beta_{\mathbf{x}}-\mathbf{W}\beta_{\mathbf{w}}) \end{bmatrix} = E\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

未知数は係数の数（$k+r$），方程式の数は直交条件の数（$m+r$）

一般的な IV 回帰モデル

$k = 1$ のとき,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + \cdots + +\beta_{1+r} W_{ri} + u_i$$

1 段階目は誘導型（reduced form：右辺が外生変数のみ）を OLS 推定して

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + +\pi_{m+r} W_{ri} + v_i$$

当てはめ値を求めて

$$\hat{X}_i = \pi_0 + \hat{\pi}_1 Z_{1i} + \cdots + \hat{\pi}_m Z_{mi} + \hat{\pi}_{m+1} W_{1i} + \cdots + +\hat{\pi}_{m+r} W_{ri}$$

2 段階目では当てはめ値を用いて OLS 推定を行う

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + \beta_3 W_{2i} + \cdots + +\beta_{1+r} W_{ri} + u_i$$

一般的な IV 回帰モデル

$k \geq 2$ のとき,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

1 段階目は誘導型を OLS 推定して

$$\begin{aligned}
X_{1i} &= \pi_{10} + \pi_{11} Z_{1i} + \cdots + \pi_{1m} Z_{mi} + \pi_{1,m+1} W_{1i} + \cdots + \pi_{1,m+r} W_{ri} + v_{1i} \\
\vdots \quad & \quad \vdots \\
X_{ki} &= \pi_{k0} + \pi_{k1} Z_{1i} + \cdots + \pi_{km} Z_{mi} + \pi_{k,m+1} W_{1i} + \cdots + \pi_{k,m+r} W_{ri} + v_{ki}
\end{aligned}$$

当てはめ値を求めて

$$\begin{aligned}
\hat{X}_{1i} &= \hat{\pi}_{10} + \hat{\pi}_{11} Z_{1i} + \cdots + \hat{\pi}_{1m} Z_{mi} + \hat{\pi}_{1,m+1} W_{1i} + \cdots + \hat{\pi}_{1,m+r} W_{ri} \\
\vdots \quad & \quad \vdots \\
\hat{X}_{ki} &= \hat{\pi}_{k0} + \hat{\pi}_{k1} Z_{1i} + \cdots + \hat{\pi}_{km} Z_{mi} + \hat{\pi}_{k,m+1} W_{1i} + \cdots + \hat{\pi}_{k,m+r} W_{ri}
\end{aligned}$$

2 段階目では当てはめ値を用いて OLS 推定を行う

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \cdots + \beta_k \hat{X}_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

IV 回帰モデルの仮定

Relevance と exogeneity

- 少なくとも 1 つの操作変数 Z が，W を所与としたときの内生変数 X の予測に有用
- すべての操作変数 Z が誤差項 u と無相関

操作変数法の仮定

- 説明変数に含まれる外生変数について誤差項の仮定：$E[u|W_1, ..., W_r] = 0$
- 観測値が無作為抽出：$(\mathbf{X}, \mathbf{W}, Y) \sim$ i.i.d.
- 適切な 4 次のモーメント条件：$0 < E[X^4] < \infty$...
- 操作変数の外生性：$\text{corr}(Z, u) = 0$
- $(\mathbf{X}, \mathbf{W})$ について多重共線性の不在

TSLS 推定量の性質

- 一致性
- 漸近正規性
- 検定・信頼区間の形成には t 統計量，F 統計量を用いる
- 標準誤差の推定に用いる残差は $\hat{u}_i = y_i - \mathbf{X}\hat{\beta}$

```
Sample(adjusted): 49 96 IF YEAR > 1990
Included observations: 48 after adjusting endpoints
White Heteroskedasticity-Consistent Standard Errors & Covariance
Instrument list: TAXSO

Variable           Coefficient Std. Error t-Statistic Prob.
C                   9.719877  1.528322  6.359835  0.0000
LOG(AVGPRS/CPI)   -1.083587  0.318918 -3.397693  0.0014

R-squared           0.401129     Mean dependent var  4.538837
Adjusted R-squared  0.388110     S.D. dependent var  0.243346
S.E. of regression  0.190354     Akaike info criterion   -3.276967
Sum squared resid  1.666792     Schwarz criterion -3.199000
F-statistic       11.71294     Durbin-Watson stat  1.992587
Prob(F-statistic)  0.001313
```