

ivregress - 操作変数法による線形回帰 【 評価版 】

説明変数が誤差項と相関を持つ場合、OLS 推定量には不偏性も一致性も期待できなくなります。そのような場合には操作変数を用いた推定が必要となります。

- | | |
|------------------|-----------|
| 1. OLS の前提条件 | |
| 2. OLS 推定量の性質 | |
| 3. 確率的説明変数 | |
| 4. 操作変数法 | |
| 5. 2 段階最小 2 乗法 | |
| 6. ivregress の用例 | Example 1 |
| | Example 2 |
| | Example 3 |
| | Example 4 |

1. OLS の前提条件

ここでは煩雑さを避ける意味で、単回帰モデル

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \dots, n \quad (1)$$

を対象に議論を進めて行くことにします。通常、 β_0 と β_1 の推定には最小 2 乗法 (OLS: ordinary least squares) が使用されますが、その推定が適切に行われるためには次の 4 条件が前提となる点に注意する必要があります。

前提 1 : 説明変数 x は確率変数ではなく固定変数 (fixed variable) である。

前提 2 : 誤差項 u は確率変数で、期待値は 0 である。

$$E(u_i) = 0, \quad i = 1, 2, \dots, n \quad (2)$$

前提 3 : 誤差項の分散は均一である。

$$V(u_i) = E(u_i^2) = \sigma^2, \quad i = 1, 2, \dots, n \quad (3)$$

前提 4 : 異なった誤差項は無相関である。

$$\text{Cov}(u_i, u_j) = E(u_i u_j) = 0, \quad i \neq j, i, j = 1, 2, \dots, n \quad (4)$$

これらの前提のもとで OLS 推定法は残差 2 乗和 (RSS: residual sum of squares)、すなわち

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (5)$$

を最小化する推定量 $\hat{\beta}_0, \hat{\beta}_1$ を求めます。具体的には

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (6a)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \quad (6b)$$

を連立させて解けばよく、 $\hat{\beta}_0, \hat{\beta}_1$ は次式で与えられることになります。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(= \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \quad (7a)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (7b)$$

2. OLS 推定量の性質

評価版では割愛しています。

3. 確率的説明変数

評価版では割愛しています。

4. 操作変数法

評価版では割愛しています。

5. 2 段階最小 2 乗法

評価版では割愛しています。

6. ivregress の用例

ivregress で前提となるモデル式は

$$y_i = \mathbf{y}_i \boldsymbol{\beta}_1 + \mathbf{x}_{1i} \boldsymbol{\beta}_2 + u_i \quad (8a)$$

$$\mathbf{y}_i = \mathbf{x}_{1i} \boldsymbol{\Pi}_1 + \mathbf{x}_{2i} \boldsymbol{\Pi}_2 + \mathbf{v}_i \quad (8b)$$

のように表現されます。ただし

y_i は i 番目の観測データにおける従属変数値

\mathbf{y}_i は内生の回帰変数 (ベクトル)

\mathbf{x}_{1i} は外生の回帰変数 (ベクトル)

\mathbf{x}_{2i} はモデル式 (8a) に含まれていない外生の回帰変数 (ベクトル)

を表します。 \mathbf{x}_{1i} と \mathbf{x}_{2i} は総称した形で操作変数と呼ばれます。 u_i と \mathbf{v}_i は平均値が 0 の誤差項ですが、 u_i と \mathbf{v}_i の要素間には相関があっても構いません。

▷ Example 1: 2SLS 推定法

Example データセット hsng.dta 中には米国における 1980 年の国勢調査データが州別に集約された形で記録されています。

```
. use https://www.stata-press.com/data/r18/hsng.dta *1
(1980 Census housing data)
```

ここで想定するモデルは住宅に関する月額レンタル価格 (中央値) rent を、住宅価格 (中央値) hsngval と都市部在住人口比率 pcturban とによって説明しようとするもので、モデル式は次のように記述できます。

$$\text{rent}_i = \beta_0 + \beta_1 \text{hsngval}_i + \beta_2 \text{pcturban}_i + u_i$$

この場合、モデル式中の u の変化は rent に影響を及ぼすばかりでなく、hsngval にも影響を与えるものと考えられます。すなわち hsngval は内生変数として扱う必要があるわけです。そのため hsngval と相関を持つ u とは無相関な操作変数が最低 1 つ必要となります*2。そこでデータセット中の変数のうち、世帯所得を表す faminc と地域区分を表す region*3 を操作変数として選択します。これらは共に hsngval と相関を持つものの、 u とは無相関であると考えられるからです。モデル式 (8a)、(8b) との対応で言うなら次のような設定となります。

*1 メニュー操作 : File ▷ Example Datasets ▷ Stata 18 manual datasets と操作、Base Reference Manual [R] の ivregress の項よりダウンロードする。

*2 然るべき操作変数は rent に直接影響するものではないはずですが、仮にそのような変数があったとすると、それは最初からモデル式に組み込まれていなくてはならないからです。

*3 1: North East, 2: North Central, 3: South, 4: West を表すカテゴリ変数。

変数名	モデル式変数
rent	y 従属変数
hsngval	y 内生変数
pcturban	x_1 外生変数 (操作変数)
faminc	x_2 操作変数
region	x_2 操作変数

それでは 2SLS 推定法を用いた形で `ivregress` を実行してみます。なお、ダイアログの設定に際しては次の点に注意してください。

- 内生変数としては y を指定する。
- 操作変数としては x_2 のみを指定する。 x_1 も操作変数ではあるが、これは独立変数として指定する。

具体的なダイアログの操作は次のように行います。

- Statistics ▸ Endogenous covariates ▸ Linear regression with endogenous covariates と操作
- Model タブ: Dependent variable: rent
Independent variables: pcturban
Endogenous variables: hsngval
Instrumental variables: faminc i.region^{*4}
Estimator: Two-stage least squares (2SLS) (デフォルト)

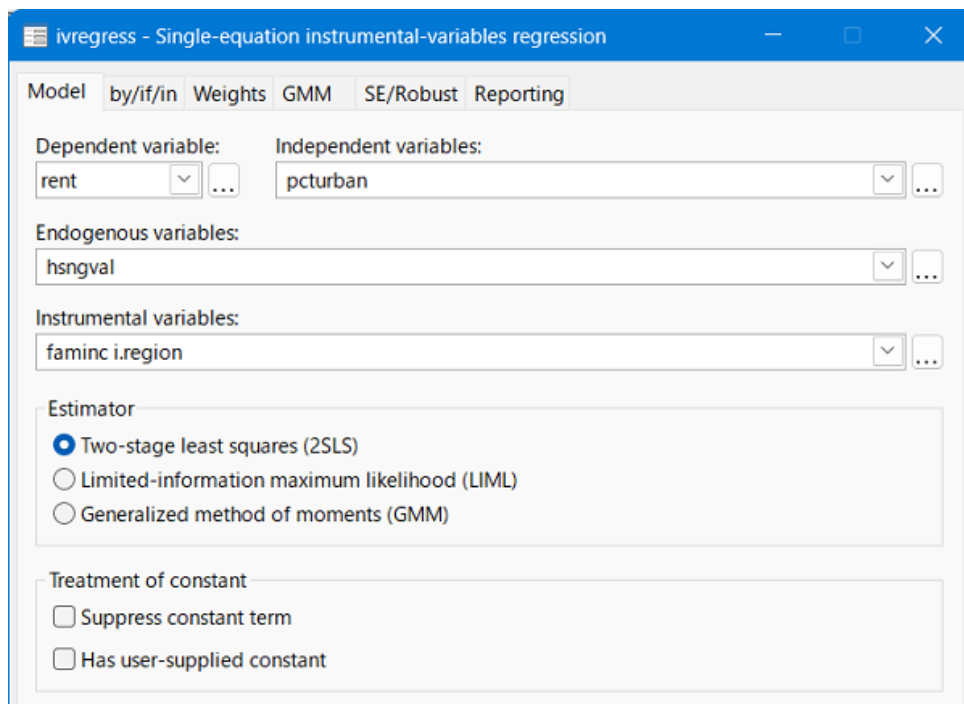


図1 ivregress ダイアログ - Model タブ

*4 因子変数演算子 i . については CS 版解説書 BR01 の因子変数の項を参照。

```
. ivregress 2sls rent pcturban (hsngval = faminc i.region)
```

Instrumental variables 2SLS regression		Number of obs	=	50
		Wald chi2(2)	=	90.76
		Prob > chi2	=	0.0000
		R-squared	=	0.5989
		Root MSE	=	22.166

rent	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
hsngval	.0022398	.0003284	6.82	0.000	.0015961	.0028836
pcturban	.081516	.2987652	0.27	0.785	-.504053	.667085
_cons	120.7065	15.22839	7.93	0.000	90.85942	150.5536

Endogenous: hsngval
Exogenous: pcturban faminc 2.region 3.region 4.region

pcturban については p 値が 0.785 であるため、有意でないことがわかります。なお、内生変数とそれに対し設定された操作変数の一覧が末尾に示されている点に注意してください。

参考までに hsngval の内生性を無視して OLS 推定を実行したときの結果を示しておきます。

```
. regress rent hsngval pcturban *5
```

```
. regress rent hsngval pcturban
```

Source	SS	df	MS	Number of obs	=	50
Model	40983.5269	2	20491.7635	F(2, 47)	=	47.54
Residual	20259.5931	47	431.055172	Prob > F	=	0.0000
				R-squared	=	0.6692
				Adj R-squared	=	0.6551
Total	61243.12	49	1249.85959	Root MSE	=	20.762

rent	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
hsngval	.0015205	.0002276	6.68	0.000	.0010627	.0019784
pcturban	.5248216	.2490782	2.11	0.040	.0237408	1.025902
_cons	125.9033	14.18537	8.88	0.000	97.36603	154.4406

hsngval に対する区間推定値には次のような違いが見られます。

OLS 推定	IV 推定
[0.0011, 0.0020]	[0.0016, 0.0029] 2SLS

◁

*5 メニュー操作：Statistics > Linear models and related > Linear regression

▷ Example 2: LIML 推定法

評価版では割愛しています。

▷ Example 3: GMM 推定法 (1)

評価版では割愛しています。

▷ Example 4: GMM 推定法 (2)

評価版では割愛しています。

