2　操作変数法とハウスマン · テイラーモデルの推定

パネルデータ分析の中級編の第二回目として, 今回は操作変数法とハウスマン · テイラー法について解説します. 最初に操作変数法の中でも特に基本的な推定手法である二段階最小二乗法について説明します. 前回同様, Cameron and Trivedi (2010) による“Microeconometrics Using Stata, Revised Edition”, Stata Press の第 6 章 Linear instrumental-variables regression の内容を用いて解説を行います.

2.1　操作変数法の基本的な考え方

操作変数法の利用例を述べる前に, 操作変数法自体の考え方を最初に説明します. 変数 y を x に回帰させる単純な回帰モデルをイメージしてください. x は教育を受けた年数, y は所得とします. 変数 y の変動をうまく x だけで説明できていれば, u はホワイトノイズであり, OLS(最小二乗推定量)$\hat{\beta}$ は β の一致推定量になります.

$$y = \beta x + u \tag{6}$$

この時, 我々は次のような関係を想定していることになります.

$$\begin{array}{ccc} x & \rightarrow & y \\ & \nearrow & \\ u & & \end{array}$$

図 1

もちろん, 現実には教育年数以外の様々な要因が所得には影響を与えます. 誤差項 u はそれらの全ての情報を含んでいることになります. 例えば, 仕事に対する適性という要因を考えてみると, ある程度の基礎学力を要求される職業であれば, 誤差項に含まれる適性 u と教育年数 x の間に相関があることも考えられます. つまり,

$$\begin{array}{ccc} x & \rightarrow & y \\ \uparrow & \nearrow & \\ u & & \end{array}$$

図 2

このような変数の構造を分析の考え方に反映させたいとき, 我々は操作変数法を利用しなければなりません. そしてこのような**構造方程式**を用いてモデルを推定する場合, x には内生性があるといいます. したがって, このような構造が合理的であると思われる場面で, 単純に 6 式を OLS で推定しても, それは一致性は有していません. 一致性とは簡単に言えば, データ数 N が無限大に近くなるほど, 推定量 $\hat{\beta}$ と母数 β の差が小さくなることを指します. このような推定上の問題に対応するのが操作変数法です.

操作変数 z というものを新たに導入して, 次のような関係を考えます.

$$\begin{array}{ccccc} z & \rightarrow & x & \rightarrow & y \\ & & \uparrow & \nearrow & \\ & & u & & \end{array}$$

図 3

先と同じように x は教育を受けた年数, y は所得とした場合, z は大学への地理的な距離とします. 大学への通学に要する費用や通学時間が少なければ, 大学へ通う人も増えると考えてください. 図のような構造方程式を満たす推定量の事を操作変数推定量 $\hat{\beta}_{\mathrm{IV}}$ と呼びます. この時, z と u の間に相関は無く, z と x の間には相関があると考え, その相関が強い事が好まれます. 特に z と x の相関については後で「弱相関」というキーワードで考察します.

2.2 モデルの設定

ここでは操作変数法を説明するための設定について述べます. 我々の推定しようとするモデルは次のようなものであるとします.

$$y_{1i} = \mathbf{y}'_{2i}\beta_1 + \mathbf{x}'_{1i}\beta_2 + u_i, \quad i = 1, \ldots .N \tag{7}$$

ここで太字で表記してある $\mathbf{y}_2$ は複数の内生変数を示し, $\mathbf{x}_1$ は複数の外生変数[2]であることを示しています. $\mathbf{y}_2$ と $\mathbf{x}_1$ の間には相関はないものとします. しかし, $\mathbf{y}_2$ と u_i の間には相関があるものと考えます. 先にも述べたようにこのような**構造方程式**を考えた時, OLS 推定量に一致性はありません. そこで操作変数 $\mathbf{x}_2$ を導入します. この $\mathbf{x}_2$ は $\mathrm{E}\left(u_i|\mathbf{x}_{2i}\right) = 0$ を満たすものと仮定し, 7 式の $\mathbf{y}_2$ との間に相関を持つものとします. そして, 次に示す第一段階の回帰モデルを推定します.

$$y_{2ji} = \mathbf{x}'_{1i}\pi_{1j} + \mathbf{x}'_{2i}\pi_{2j} + v_{ji}, \quad j = 1, \ldots .m \tag{8}$$

第一段回目の回帰モデル (8 式) の特徴は右辺に外生変数だけを利用するところです. この式の右辺には 7 式の外生変数 $\mathbf{x}_1$ も含まれていることに注意してください. また, 7 式に m 個の内生変数が含まれている時は, 操作変数法を利用するためには m 個以上の操作変数が必要となります.

ここからは 7 式を次のように簡単に書き換えて説明を行います.

$$y_i = \mathbf{x}'_i\beta + u_i \tag{9}$$

もちろん, $\mathbf{x}'_i = [\mathbf{y}'_{2i}\mathbf{x}'_{1i}]$ です. また, 操作変数についてもベクトル表記を利用して $\mathbf{z}'_i = [\mathbf{x}'_{1i}\,\mathbf{x}'_{2i}]$ とします. そして, 操作変数 $\mathbf{z}$ は次に示す条件付きモーメント制約を満たすものとします.

$$\mathrm{E}\left(u_i|\mathbf{z}_i\right) = 0 \tag{10}$$

2.3 操作変数推定量

10 式のモーメント条件は実際には $\mathrm{E}\left(\mathbf{z}_i u_i\right) = 0$ として計算します. これを具体的に変数を用いて書けば,

$$\mathrm{E}\left\{\mathbf{z}'_i\left(y_i - \mathbf{x}'_i\beta\right)\right\} = \mathbf{0} \tag{11}$$

となります. これを実際に計算する段で内生変数の個数と操作変数の個数の大小関係に注意する必要があります. 両者の数が等しい時に**丁度識別**, 内生変数の数がより多い場合を**過小識別**, 逆に, 操作変数の数が多い場合を**過剰識別**と呼んで区別します. この中で過小識別のケースでは一致推定量を求めることはできません.

[2]内生性のない変数の事を外生変数と呼びます。定数項は外生変数に含まれます。

今回は丁度識別のケースで, さらに誤差項 u_i が独立で均一分散の場合に, 有効性に優れている[3]二段階最小二乗法について解説します. これ以外の操作変数法として GMM(Generalized Method of Moment:一般化積率法) や, LIML(Limited Maximun-likelhood:制約付き情報による最尤法) といった推定手法がありますが, それについては次回以降に説明します.

2.4 推定例:処方薬に対する個人支出について

ここでは"Microeconometrics Using Stata, Revised Edition"の p184 にある処方薬に対する支出に関する実証分析例を紹介します. 利用するモデルは内生変数を一つ含む構造方程式です. データは 65 才を越える個人の医療費に関するデータで出典は Medical Expenditure Panel Survey (MEPS) です. 利用する変数は次の通りです.

ldrugexp 処方薬に対する個人支出の対数値 (被説明変数).

hi_empunion 会社または組合負担の健康保険に加入していたことを示すダミー変数.

totchr 慢性疾患の状態

age 年齢

female 性別

blhisp 人種を示すダミー変数. 黒人やヒスパニックの場合は 1.

linc 家計所得の対数値

ここでモデル推定の問題意識を説明します. 高齢者の処方薬への個人支出 (`ldrugexp`) は, 退職前に健康保険に加入 (`hi_empunion`) していたことで軽減されるのか?ということが問題意識です. 引退後の高額な医療費負担が心配な高齢者は就労中に, 余分にコストを負担して引退後の処方薬への支出を軽減できる保険に加入すると考えられます. 実際に加入していた人の負担額は, 加入していなかった人に比べどれだけ少ないのか, という事を実証分析します. データとして集めた標本はほとんどが退職 (引退) した個人のものですが, 中にはまだ就労している人のデータも含まれています.

利用するデータは `mus06data.dta` です. まずはデータを読み込んで分布を確認します.

```
. use mus06data.dta
. global x2list totchr age female blhisp linc
. summarize ldrugexp hi_empunion $x2list
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ldrugexp	10,391	6.479668	1.363395	0	10.18017
hi_empunion	10,391	.3796555	.4853245	0	1
totchr	10,391	1.860745	1.290131	0	9
age	10,391	75.04639	6.69368	65	91
female	10,391	.5797325	.4936256	0	1
blhisp	10,391	.1703397	.3759491	0	1
linc	10,089	2.743275	.9131433	-6.907755	5.744476

[3]推定量の分散が小さい事。

ここでは5つの外生変数をグローバルマクロ x2list として登録し, 簡単に利用できるようにしています. 実際にグローバルマクロを利用する時は先頭に$ 記号を付けます. その要約統計量を見ると, 約 38% の人がメディケアでカバーできない部分も支出に対応するため, 保険に加入していることが分かります.

操作変数の選択

我々の考えている構造方程式に含まれる内生変数は `hi_empunion` だけです. 他の変数については図3に示すような構造は考えづらいことが分かります. ここでは z の候補として次の4つを考えます.

ssiratio 生活保護の支給金額が全収入に示す割合. この値が大きいほど, 所得は厳しい制約があることが分かります. したがって, 保険の加入にはマイナス方向の圧力を与える事になります.

lowincome 低所得であることを示すダミー変数.

multlc 勤め先が複数の事業所を有する, 比較的規模の大きな会社であることを示すダミー変数.

firmsz 勤め先の労働力の大きさを数値化した変数.

`multlc` と `firmsz` は退職者, 自営業者, 個人で追加保険を購入した人には関係ありませんので, 変数 hi_empunion との相関はあまり期待できません. 念のため, 4つの要約統計量を確認しておきます.

```
. summarize ssiratio lowincome multlc firmsz if linc!=.
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ssiratio	10,089	.5365438	.3678175	0	9.25062
lowincome	10,089	.1874319	.3902771	0	1
multlc	10,089	.0620478	.2412543	0	1
firmsz	10,089	.1405293	2.170389	0	50

ここでは内生変数 `hi_empunion` と相関が一番強いという文脈で `ssiratio` を使うことにします. それでは実際に, 操作変数の一種である二段階最小二乗法を用いて構造方程式を推定します.

```
. ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust) first
```

コマンドは `ivregress 2sls` であり, その次に被説明変数 `ldrugexp` を入力します. その後ろに内生変数 `hi_empunion` と操作変数 `ssiratio` を等号で接続して入力します. 説明変数の最後には先にグローバルマクロとして設定した外線変数群 `x2list` を入力します. ここでは構造方程式において不均一分散の発生が考えられるので, `vce(robust)` オプションを利用しています. また, `first` を利用すると一段階目の回帰の結果も一緒に出力します.

```
.  ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust) first
First-stage regressions

                                                Number of obs     =     10,391
                                                F(   1,  10389)   =     148.60
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0385
                                                Adj R-squared     =     0.0384
                                                Root MSE          =     0.4759

------------------------------------------------------------------------------
             |               Robust
 hi_empunion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ssiratio |   -.25431    .0208617   -12.19   0.000    -.2952028   -.2134171
       _cons |  .5120564    .0119376    42.89   0.000     .4886564    .5354564
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression          Number of obs   =     10,391
                                                  Wald chi2(1)    =      34.27
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     1.4418

------------------------------------------------------------------------------
             |               Robust
    ldrugexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 hi_empunion |  -.8747104   .1494206    -5.85   0.000    -1.167569   -.5818515
       _cons |   6.811756   .0585355   116.37   0.000     6.697029    6.926484
------------------------------------------------------------------------------
Instrumented:  hi_empunion
Instruments:   ssiratio
```

一段階目の回帰の結果を見ると, ssiratio の符号が負になっており, これは先に想定した結果通りになっています. R^2 の値は小さいものの符号条件は適切で, 有意となっています. 一番重要な出力は二段目の回帰の結果です. 保険に加入していると, 処方薬に対する個人負担が軽減される結果になっています. 被説明変数の ldrugexp は対数値ですから, 保険加入により個人負担が約 90% 軽減されると解釈できます.

重要 ivregress の機能を拡張した ivreg2 という ado ファイルがあります. 操作変数法の考え方を理解できたら, ivreg2 の利用をお勧めします.

内生性の検定

二段階最小二乗法では一段階目で内生変数を操作変数と外生変数に回帰させます. しかし, 図 3 の構造方程式を考えたからといって, 本当にそこまでやる必要があるのでしょうか. 単純な重回帰分析ではだめなのでしょうか. 試しに OLS による重回帰分析を行ってみましょう.

```
.  regress ldrugexp hi_empunion $x2list, vce(robust)
```

```
Linear regression                               Number of obs     =     10,089
                                                F(6, 10082)       =     376.85
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1770
                                                Root MSE          =      1.236
```

ldrugexp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hi_empunion	.0738788	.0259848	2.84	0.004	.0229435	.1248141
totchr	.4403807	.0093633	47.03	0.000	.4220268	.4587346
age	-.0035295	.001937	-1.82	0.068	-.0073264	.0002675
female	.0578055	.0253651	2.28	0.023	.0080848	.1075262
blhisp	-.1513068	.0341264	-4.43	0.000	-.2182013	-.0844122
linc	.0104815	.0137126	0.76	0.445	-.0163979	.037361
_cons	5.861131	.1571037	37.31	0.000	5.553176	6.169085

この結果をみると, 内生変数 `hi_empunion` の値が大きく異なることが分かります.

パネルデータ分析において固定効果モデルとランダム効果モデルの判別を行うのにハウスマン検定を利用しました. ハウスマン検定は 2 つの推定方法における推定量の違いに注目して, その違いが大きいほど, 検定統計量も大きくなり, 帰無仮説 (ランダム効果が適切である) を棄却するというものでした. ここではデータクロスセクションでパネルデータではありませんが, 次のような検定統計量を考えて, 内生性の検定 (Durbin-Wu-Hausman Test) を行います.

$$T_H = \frac{\left(\hat{\beta}_{\mathrm{IV}} - \hat{\beta}_{\mathrm{OLS}}\right)^2}{\hat{V}\left(\hat{\beta}_{\mathrm{IV}} - \hat{\beta}_{\mathrm{OLS}}\right)}$$

帰無仮説は変数は外生的である, というものです. 今一度, 二段階最小二乗法でモデルを推定し, 検定を実行します.

```
. ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust)
```
(省略)
```
. estat endogenous
```

```
Tests of endogeneity
Ho: variables are exogenous

Robust score chi2(1)            =   24.935  (p = 0.0000)
Robust regression F(1,10081)    =  26.4333  (p = 0.0000)
```

一番下に表示されるのが, 目的の検定結果です. 帰無仮説は, 目的の変数は外生変数であるというものです. この結果から, `hi_empunion` の内生性が確認できました. ただ, ここで少し気になる点があります. 一段階目の OLS 推定において被説明変数 `hi_empunion` が二値変数であるという点です. 被説明変数が二値なら, ロジットやプロビットを使うべきではないかと思われるかもしれません. 結論から先に述べると, 操作変数法のフレームワークで満たされるべきは, 直行条件 $\mathrm{E}\left(u_i|\mathbf{z}_i\right) = 0$ であり, これは内生変数が二値の場合であっても満たされますので問題はありません.

しかし, 内生変数が二値であることを考慮して操作変数推定を行う `treatreg` というコマンドも Stata には用意されています. 興味ある読者は"Microeconometrics Using Stata, Revised Edition"の 6.3.8 IV estimation with a binary endogenous regressor を参照してください.

2.5 弱相関

選択した操作変数と目的とする内生変数の間の相関は強い事が望まれます. 最初に単純な相関を調べてみましょう.

```
. correlate hi_empunion ssiratio lowincome multlc firmsz if linc!=.
```

```
(obs=10,089)

             | hi_emp~n ssiratio lowinc~e   multlc   firmsz
-------------+---------------------------------------------
 hi_empunion |   1.0000
    ssiratio |  -0.2124   1.0000
   lowincome |  -0.1164   0.2539   1.0000
      multlc |   0.1198  -0.1904  -0.0625   1.0000
      firmsz |   0.0374  -0.0446  -0.0082   0.1873   1.0000
```

内生変数 hi_empunion に対する相関の値は全体的に弱いように見えます. しかし, 操作変数法における文脈でこれを調べる場合は estat firststage コマンドを用いて検定を行います. ここでは分散不均一のオプションを用いて構造方程式を推定していますので, それに対応して forcenonrobust オプションを利用します. オプション all は標準出力以外の統計量を表示します.

```
. estat firststage, forcenonrobust all
```

```
First-stage regression summary statistics
--------------------------------------------------------------------------
             |            Adjusted      Partial       Robust
    Variable |   R-sq.       R-sq.        R-sq.    F(1,10082)   Prob > F
-------------+------------------------------------------------------------
 hi_empunion |  0.0761      0.0755       0.0179       65.7602    0.0000
--------------------------------------------------------------------------

Shea´s partial R-squared
--------------------------------------------------
             |     Shea´s             Shea´s
    Variable |  Partial R-sq.   Adj. Partial R-sq.
-------------+------------------------------------
 hi_empunion |     0.0179             0.0174
--------------------------------------------------

Minimum eigenvalue statistic = 183.98
Critical Values                      # of endogenous regressors:    1
Ho: Instruments are weak             # of excluded instruments:     1
---------------------------------------------------------------------
                                   |    5%     10%     20%     30%
2SLS relative bias                 |         (not available)
-----------------------------------+---------------------------------
                                   |   10%     15%     20%     25%
2SLS Size of nominal 5% Wald test  |  16.38    8.96    6.66    5.53
LIML Size of nominal 5% Wald test  |  16.38    8.96    6.66    5.53
---------------------------------------------------------------------
```

内生変数 1 個に対して操作変数を 1 個利用した丁度識別の場合のモデルに対して, 弱相関の検定を実行しました. 結果画面には 3 つの表がありますので順番に説明します. 一番上の表は一段目の回帰モデルの情報です. この表で弱相関の検定結果を述べることが厳密にはできません. あくまで参考情報として利用します. R^2 と adj R^2 は一段目の回帰モデルの決定係数です. モデルの説明力は弱いこ

とが分かります. 外生変数の影響を除いた時の偏回帰係数 Partial R^2 は僅か 0.0179 しかありません. 表の一番右側にある F 統計値は前出の一段目の回帰における外生変数の t 統計量を二乗した値となっています. この F 統計値が 10 より大きければ弱相関ではないという一つの目安がありますので, 覚えておくと良いでしょう.

中央の表にある Shea の偏回帰係数は内生変数が複数個存在する場合に, 複数ある一段目の回帰式の説明力を示す統計量です. この値がある範囲にあれば良いという具体的な目安はありません.

肝心なポイントは一番下の表にある Stock and Yogo の検定結果になります. 表の一行目は丁度識別の場合に出力されません. 2 行目と 3 行目に 2SLS と LIML の推定量に関する境界値の情報を出力します. 我々は 2SLS を行っていますので, "`2SLS Size of nominal 5% Wald test`" に注目します. 帰無仮説は "操作変数は弱相関である" です. 今, 10% の境界値は 16.38 となっています. 構造方程式の推定では `vce(robust)` のオプションを利用していますので, 実際には多少のズレはありますが, 手元のデータから求めた検定統計量は 183.98 となっており, 大きく境界値を上回っています. また, 参考情報である一番上の表に出力されている `Robust F(1,10082)` の値も 10082 も 16.38 を上回っていますので, 弱相関であるという帰無仮説は棄却できます.

同書では操作変数として先に挙げた他の変数を利用して, 内生変数の推定値の符号や大きさを比較する感度分析を行っています. 操作変数を変更しても構造方程式の符号に変化が生じないという結果が得られれば, 推定結果はかなり頑健であると言えます.

操作変数法のまとめ

- 説明変数に内生性が想定できる場合, これを OLS で推定すると, 一致推定量は得られない.
- 内生変数と相関の強い操作変数を見つけて操作変数法を実行する.
- 二段階最小二乗法で丁度識別の場合でも, 内生性と弱相関をチェックする.

2.6 パネルデータにおける操作変数法

クロスセクションデータにおける操作変数法による推定のフレームワークは, パネルデータの場合にもそのまま利用できます. ここでは主に "Microeconometrics Using Stata, Revised Edition", Stata Press の第 8 章 Linear panel-data models:Basics の内容を用いて解説を行います. 最初にここで推定するモデルを確認しておきます.

$$y_{it} = \mathbf{x}'_{it}\beta + \alpha_i + \epsilon_{it} \tag{12}$$

パネルデータ分析に利用するこの基本的なモデルには α_i と ϵ_{it} という 2 つの誤差要素 (error component) が存在します. 固定効果モデルを用いると, 時間に対して不変な値を持つ変数は削除されてしまいますが, 時間とともに変動する変数については一致推定量を提供します. 固定効果モデルでは $\mathbf{x}_{it}$ は固定効果 α_i とは相関を持ちますが, ϵ_{it} とは無相関であるという仮定を置きます.

内生変数への対応

しかし, 先に説明した `ivreg` コマンドの場合と同じく, $\mathbf{x}_{it}$ が ϵ_{it} と相関を持つようなケースも考えれらます. その例を賃金関数の推定を用いて次に示します. 利用する変数は次の通りです.

lwage 賃金の対数値 (`lwage`) とその個人の学歴 (`ed`)

exp フルタイムで過去に働いた年数

wks 就労時間 (週/年)

ed 学歴

これらの変数を利用して単純な固定効果モデルの推定から話を始めます. 個人単位のデータには級内相関があることを仮定してクラスタオプションを利用します.

```
. use mus08psidextract.dta, clear
. xtreg lwage exp exp2 wks ed, fe vce(cluster id)
```

```
note: ed omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =      4,165
Group variable: id                              Number of groups  =        595
R-sq:                                           Obs per group:
     within  = 0.6566                                         min =          7
     between = 0.0276                                         avg =        7.0
     overall = 0.0476                                         max =          7
                                                F(3,594)          =    1059.72
corr(u_i, Xb)  = -0.9107                        Prob > F          =     0.0000
                                 (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0040289    28.24   0.000     .1058753    .1217004
        exp2 |  -.0004244   .0000822    -5.16   0.000    -.0005858   -.0002629
         wks |   .0008359   .0008697     0.96   0.337    -.0008721    .0025439
          ed |          0  (omitted)
       _cons |   4.596396   .0600887    76.49   0.000     4.478384    4.714408
-------------+----------------------------------------------------------------
     sigma_u |  1.0362039
     sigma_e |  .15220316
         rho |  .97888036   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

学歴 `ed` は時間に対して不変な変数なのでここでは自動的に削除されています. 次に, 変数の内生性について考えてみます. つまり, 年間の就業時間 wks は既婚/独身であることに大きく影響されると考えます. 図 3 の z として変数 `ms`(既婚/独身のダミー) の存在を仮定します. 利用するコマンドは `xtivreg` です. もちろん, ここで時間に対して不変な変数を導入しても自動的に削除されます.

```
. xtivreg lwage exp exp2 (wks=ms) ,fe
```

```
Fixed-effects (within) IV regression          Number of obs     =      4,165
Group variable: id                            Number of groups  =        595
R-sq:                                         Obs per group:
     within  =      .                                       min =          7
     between = 0.0172                                       avg =        7.0
     overall = 0.0284                                       max =          7
                                              Wald chi2(3)      =  700142.43
corr(u_i, Xb)  = -0.8499                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wks |  -.1149742   .2316926    -0.50   0.620    -.5690832    .3391349
         exp |   .1408101   .0547014     2.57   0.010     .0335974    .2480228
        exp2 |  -.0011207   .0014052    -0.80   0.425    -.0038748    .0016334
       _cons |    9.83932   10.48955     0.94   0.348    -10.71983    30.39847
-------------+----------------------------------------------------------------
     sigma_u |  1.0980369
     sigma_e |  .51515503
         rho |  .81959748   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(594,3567) =     4.62        Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   wks
Instruments:    exp exp2 ms
------------------------------------------------------------------------------
```

パネルデータ分析の場合, クロスセクションデータの時とは違って, 内生性を検定する estat endogenous や弱相関の検定を行う estat firststage は利用できません. したがって, 両者の推定結果を慎重に考察してモデルの適切さを判定します. まず, IV 推定の結果を見ると, 1 週間余分に働くと 11.5% 賃金が少なくなるという, 合理性に欠ける結果になっています. しかし, 有意ではありません. また, 就労年数 exp による賃金カーブについて考えてみると 64 年 $[= 0.1408/(2 \times 0.0011)]$ でピークを迎えるという結果になっています.

一方, 内生性を考慮しない単純な固定効果モデルの結果を見ると, wks の符号は逆に正であり, 推定値自体が大きくなっています. また, 就労年数 exp については 142 年でピークを迎えるということですから, 実質的に賃金カーブに就労年数は影響を与えないと考えられます. 2 つの wks の推定値を比較すると, 非常に大きな差があります. 操作変数として利用した ms は二値データですので, wks との相関が非常に弱く, このような結果を招いていると考えられます. かと言って, wks の符号条件を無視することはできませんので, ms 以外のより適切な操作変数を探すことになります.

重要 xtivreg コマンドでランダム効果モデルの推定も可能です. その時, ハウスマン検定で操作変数を利用した時の固定効果モデルとランダム効果モデルの選別も行えます.

重要 xtivreg にも機能拡張した xtivreg2 という ado ファイルがあります.

2.7 ハウスマン · テイラー法

ここでは固定効果モデルを利用するときに, 関心のある変数が時間に対して不変であるために削除されてしまうという問題に対する対応方法の一つ, ハウスマン · テイラー法について考えます. 先の賃金関数の例で ed を取り除くことなく, 学歴が賃金に与える影響を考察しようというのが目的です.

パネルデータ分析のフレームワークで考えると，この場合は，観測不可能な異質性 α_i と学歴 `ed` には相関があることが想像できます．つまり，`ed` は内生変数であると考えることができます．

具体的な推定に先立ち，モデルの設定を説明します．

$$y_{it} = \mathbf{x}'_{1it}\beta_1 + \mathbf{x}'_{2it}\beta_2 + \mathbf{w}'_{1it}\gamma_1 + \mathbf{w}'_{2it}\gamma_2 + \alpha_i + \epsilon_{it} \tag{13}$$

サブスクリプト 1 のついた変数は α_i に無相関，サブスクリプト 2 のついた変数は α_i と相関を持ちます．また，変数 $\mathbf{x}$ は時変であるが，$\mathbf{w}$ は時間に対して変化しないと考えます．12 式が固定効果モデルであるケースでは $\mathbf{x}_{it}$ と α_i は相関を持つしていましたが，それに比べると，ここでは**より強い仮定**を置いていることになります．さらに全ての説明変数は ϵ_{it} と無相関であるとします．

それではモデル推定を行います．ここでは Cornwell and Rupper (1988) の提示したモデルをそのまま推定します．

```
. xthtaylor lwage occ south smsa ind exp exp2 wks ms union fem blk ed, endog(exp exp2 wks ms union ed)
```

```
. xthtaylor lwage occ south smsa ind exp exp2 wks ms union fem blk ed, endog(exp exp2 wks ms union
>  ed)

Hausman-Taylor estimation                       Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

                                                Obs per group:
                                                              min =          7
                                                              avg =          7
                                                              max =          7

Random effects u_i ~ i.i.d.                     Wald chi2(12)     =    6891.87
                                                Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TVexogenous  |
         occ |  -.0207047   .0137809    -1.50   0.133    -.0477149    .0063055
       south |   .0074398    .031955     0.23   0.816    -.0551908    .0700705
        smsa |  -.0418334   .0189581    -2.21   0.027    -.0789906   -.0046761
         ind |   .0136039   .0152374     0.89   0.372    -.0162608    .0434686
TVendogenous |
         exp |   .1131328    .002471    45.79   0.000     .1082898    .1179758
        exp2 |  -.0004189   .0000546    -7.67   0.000    -.0005259   -.0003119
         wks |   .0008374   .0005997     1.40   0.163    -.0003381    .0020129
          ms |  -.0298508     .01898    -1.57   0.116    -.0670508    .0073493
       union |   .0327714   .0149084     2.20   0.028     .0035514    .0619914
TIexogenous  |
         fem |  -.1309236    .126659    -1.03   0.301    -.3791707    .1173234
         blk |  -.2857479   .1557019    -1.84   0.066    -.5909179    .0194221
TIendogenous |
          ed |    .137944   .0212485     6.49   0.000     .0962977    .1795902
             |
       _cons |   2.912726   .2836522    10.27   0.000     2.356778    3.468674
-------------+----------------------------------------------------------------
     sigma_u |  .94180304
     sigma_e |  .15180273
         rho |  .97467788   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Note: TV refers to time varying; TI refers to time invariant.
```

コマンドでは内生変数のみ，`endog()` の中に改めて入力します．時変であるか否かは Stata が自動的に判別しますので，特にユーザが指定する必要はありません．推定結果を見ると，ed の推定値が有

意な値として得られています．これを単純にランダム効果モデルで推定すると，その推定値は 0.112 となり，標準誤差は 0.0084 となります．ここにあるようにハウスマン・テイラー法で推定すると，推定値は 0.138，標準誤差は 0.0212 と大きくなります．ハウスマン・テイラー法に限らず，操作変数法を利用すると推定値の標準誤差は大きくなります．

先ほど触れたようにハウスマン・テイラー法では**より強い仮定**をおいてモデルを推定しています．この仮定が成立していることを確認するコマンドはアドファイル `xtoverid` として用意されていますので，是非，この例題で試してください．手順は `xthtaylor` コマンドで推定後に，`xtoverid` コマンドを単純に実行するだけです．帰無仮説は**より強い仮定**が成立しているというものです．

次回は GMM(一般化積率法) を用いたダイナミックパネルデータモデルの推定について解説します．

以上

2016 年 10 月

株式会社 ライトストーン