ERM（拡張回帰モデル）

はじめに

- ERM とは拡張回帰モデル(Extended Regression Model)を意味する言葉です。
- 線形回帰、区間回帰、プロビット、オーダードプロビットモデルなどを推定する際に生じる、次の問題に対応する場合に使用します。
    1) 内生性
    2) サンプルセレクション
    3) 非無作為割り付け
    4) パネルデータにおける個体内の相関
- これらの問題は単独で生じることもあれば、複数が同時に発生することもあります。ERM を利用すれば、これらの問題に対応し、推定上の問題を解消できます。

ERM で推定可能なモデル

- 連続変数を被説明変数とするモデルに内生性が存在する
- 二値変数、順序変数などを被説明変数とするモデルに内生性が存在する
- 内生変数の多項式モデル
- 内生性のある共変量の交差項が存在するモデル
- 内生変数と外生変数の交差項が存在するモデル
- 内生性のあるサンプルセレクションモデル
- 無作為割り付けになっていない処置モデル
- パネルデータの非線形モデルに内生性が存在する

✧ 本文中のコマンドをコピーし、Stata のコマンドウィンドウに貼り付けて実行できます。全ての操作のコマンドは、do ファイル erm.do にまとめられています。

クロスセクションにおける内生性の問題

連続変数y_iを被説明変数として、これを外生変数x_iと内生性のある共変量w_{ci}に回帰させることを考えます。

$$y_i = x_i\beta + w_{ci}\beta_c + \epsilon_i$$
$$w_{ci} = z_{ci}A_c + \epsilon_{ci}$$

ここでz_{ci}はx_iと、w_{ci}に影響を与える変数を指します。また、モデルが識別されるためにはz_{ci}はx_i以外の外生変数を 1 つ以上、利用する必要があります。その個数は内生説明変数w_{ci}の個数に対応するものとします。ここで観測できない誤差ϵ_iとϵ_{ci}は平均ゼロ、共分散は次式のようになります。

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_{1c}' \\ \sigma_{1c} & \Sigma_c \end{bmatrix}$$

そして尤度関数は

$$lnL = \sum_{i=1}^{N} w_i ln\phi_{C+1}(r_i, \Sigma)$$

ここで、

$$r_i = [y_i - x_i \quad w_{ci} - z_{ci}A_c]$$

y_iの条件付き期待値は次のようになります。

$$E(y_i|x_i, w_{ci}, z_{ci}) = x_i\beta + w_{ci}\beta_c + \sigma_{1c}'\Sigma_c^{-1}(w_{ci} - z_{ci}A_c)'$$

操作変数法の一種である二段階最小二乗法とは異なり、最尤推定により目的のパラメータを推定するというアプローチを採用します。この方法をユーザが実行しようとすると、モデルの定式化が非線形式など複雑なものになるほど、尤度関数の導出とプログラミングに時間がかかってしまいます。先に紹介したように、ある程度の非線形モデルのバリエーションを用意したところに ERM を利用するメリットがあります。

例題 1

ここで利用するサンプルデータは親の所得 **income** が子供の成績 **gpa**(大学時代の GPA)に与える影響を考察するために用意したものです。内生性のある説明変数 **hsgpa**(高校時代の GPA)を含む線形モデルを推定します。

- サンプルデータをインポートし、変数 **gpa** の内容を確認します。Stata のコマンドウィンドウで次のコマンドを実行します。

```
webuse class10, clear

codebook gpa
```

結果画面に次のように表示されます。

```
-------------------------------------------------------------------------------
gpa                                                                 College GPA
-------------------------------------------------------------------------------

                  type:  numeric (double)

                 range:  [.52,4]                      units:  .01
         unique values:  270                      missing .:  972/2,500

                  mean:   2.95351
             std. dev:   .635897

           percentiles:        10%       25%       50%       75%       90%
                              2.01      2.55      3.13      3.38      3.69
```

拡張線形モデル

- 線形モデルを推定します。拡張線形回帰のコマンド **eregress** を使用します。ここでは説明変数 **hsgpa** に内生性を仮定しています。**hscomp** は学力面で 3 段階に分けた高校のランク(high/moderate/low)を示すダミー変数です。**i.**は指標子を意味する Stata の因子変数演算子です。 ⇒PDF マニュアル[U] User's Guide - 11. Language syntax

```
eregress gpa income, endogenous(hsgpa = income i.hscomp)
```

```
Iteration 0:   log likelihood = -638.58598
Iteration 1:   log likelihood = -638.58194
Iteration 2:   log likelihood = -638.58194

Extended linear regression                      Number of obs     =      1,528
                                                Wald chi2(2)      =    1167.79
Log likelihood = -638.58194                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa						
income	.0575145	.0055174	10.42	0.000	.0467007	.0683284
hsgpa	1.235868	.133686	9.24	0.000	.9738484	1.497888
_cons	-1.217141	.3828614	-3.18	0.001	-1.967535	-.4667464
hsgpa						
income	.0356403	.0019553	18.23	0.000	.0318079	.0394726
hscomp						
moderate	-.1310549	.0136503	-9.60	0.000	-.1578091	-.1043008
high	-.2331173	.0232712	-10.02	0.000	-.278728	-.1875067
_cons	2.951233	.0164548	179.35	0.000	2.918982	2.983483
var(e.gpa)	.1436991	.0083339			.1282592	.1609977
var(e.hsgpa)	.0591597	.0021403			.05511	.063507
corr(e.hsgpa,e.gpa)	.2642138	.0832669	3.17	0.002	.0948986	.4186724

- 操作変数法(二段階最小二乗法)との比較をします。一推定式の操作変数回帰のコマンド **ivregress** を使用します。両者の係数はほぼ等しいことが分かります。

```
ivregress 2sls gpa income (hsgpa = income i.hscomp)
```

```
Instrumental variables (2SLS) regression        Number of obs     =      1,528
                                                Wald chi2(2)      =    1168.01
                                                Prob > chi2       =     0.0000
                                                R-squared         =     0.6444
                                                Root MSE          =     .37906
```

gpa	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hsgpa	1.236168	.1336307	9.25	0.000	.9742563	1.498079
income	.0575042	.0055157	10.43	0.000	.0466937	.0683148
_cons	-1.217997	.3827032	-3.18	0.001	-1.968081	-.4679123

```
Instrumented:  hsgpa
Instruments:   income 2.hscomp 3.hscomp
```

拡張区間回帰モデル

被説明変数が連続変数ではなく、一定間隔の範囲の上限値や下限値で構成されるような場合に、最小二乗法ではなく、最尤法による区間回帰という推定手法を利用します。

- 内生性のある hsgpa を含む区間回帰モデルを推定します。
- gpal は gpa をいくつかのカテゴリに分けた時の下限値です。gpa には 2 より小さな値が存在しますが、カテゴリ変数 gpal は切り上げにより 2 以上の値のみになっています。
- 変数 gpau は連続変数 gpa の上限値による変数で、同じく hsgpa に内生性を仮定しています。

- 変数 gpal および gpau の内容を確認します。

```
codebook gpal gpau
```

```
-------------------------------------------------------------------------------
gpal                                                  College GPA, lower endpoint
-------------------------------------------------------------------------------

                  type:  numeric (double)

                 range:  [2,3.5]                      units:  .1
        unique values:  4                         missing .:  1,122/2,500

           tabulation:  Freq.  Value
                          197  2
                          302  2.5
                          612  3
                          267  3.5
                        1,122  .

-------------------------------------------------------------------------------
gpau                                                  College GPA, upper endpoint
-------------------------------------------------------------------------------

                  type:  numeric (double)

                 range:  [2,4]                        units:  .1
        unique values:  5                         missing .:  972/2,500

           tabulation:  Freq.  Value
                          150  2
                          197  2.5
                          302  3
                          612  3.5
                          267  4
                          972  .
```

- 区間回帰モデルを推定します。拡張区間回帰のコマンド **eintreg** を使用します。

```
eintreg gpal gpau income, endogenous(hsgpa = income i.hscomp)
```

```
Iteration 0:   log likelihood = -1716.9969
Iteration 1:   log likelihood = -1716.9968

Extended interval regression                    Number of obs     =      1,528
                                                  Uncensored      =          0
                                                  Left-censored   =        150
                                                  Right-censored  =          0
                                                  Interval-cens.  =      1,378

                                                Wald chi2(2)      =     912.68
Log likelihood = -1716.9968                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
income	.0551638	.0057859	9.53	0.000	.0438236	.066504
hsgpa	1.111672	.1407083	7.90	0.000	.8358891	1.387456
_cons	-.8180699	.4032468	-2.03	0.042	-1.608419	-.0277207
hsgpa						
income	.0356351	.0019553	18.22	0.000	.0318027	.0394675
hscomp						
moderate	-.1317151	.0136277	-9.67	0.000	-.1584249	-.1050052
high	-.2320803	.0233633	-9.93	0.000	-.2778715	-.186289
_cons	2.951568	.0164465	179.46	0.000	2.919333	2.983802
var(e.gpal)	.1354248	.0090267			.1188397	.1543245
var(e.hsgpa)	.0591594	.0021403			.0551097	.0635066
corr(e.hsgpa,e.gpal)	.2700108	.0897936	3.01	0.003	.0868241	.4355353

拡張プロビットモデル

アウトカムが 0 または 1 の値をとるような回帰分析ではプロビットモデルを利用します。

- ここではプロビットモデルの説明変数 **hsgpa** に内生性が存在するものとします。
- 変数 **graduate** は大学を卒業したか(yes/no)を示すダミー変数です。

- 変数 **graduate** の内容を確認します。

```
codebook graduate
```

```
-------------------------------------------------------------------------------
graduate                                                Graduated from college
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  yesno

                 range:  [0,1]                        units:  1
         unique values:  2                          missing .:  0/2,500

            tabulation:  Freq.   Numeric  Label
                           972         0  no
                         1,528         1  yes
```

- プロビットモデルを推定します。拡張プロビット回帰のコマンド **eprobit** を使用します。

```
eprobit graduate income i.roommate, endogenous(hsgpa = income i.hscomp)
```

```
Iteration 0:   log likelihood = -1418.5008
Iteration 1:   log likelihood = -1418.4414
Iteration 2:   log likelihood = -1418.4414

Extended probit regression                      Number of obs     =      2,500
                                                Wald chi2(3)      =     330.35
Log likelihood = -1418.4414                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
graduate						
income	.1597677	.0161903	9.87	0.000	.1280353	.1915001
roommate						
yes	.2636312	.0560907	4.70	0.000	.1536954	.373567
hsgpa	1.01877	.4255124	2.39	0.017	.1847812	1.852759
_cons	-3.647166	1.18251	-3.08	0.002	-5.964843	-1.329489
hsgpa						
income	.047859	.0016982	28.18	0.000	.0445307	.0511874
hscomp						
moderate	-.135734	.0114796	-11.82	0.000	-.1582336	-.1132343
high	-.225314	.0191903	-11.74	0.000	-.2629263	-.1877017
_cons	2.794711	.0131836	211.98	0.000	2.768871	2.82055
var(e.hsgpa)	.0685893	.0019401			.0648903	.0724992
corr(e.hsgpa, e.graduate)	.3687006	.0911327	4.05	0.000	.1782749	.5325059

例題 2

順序プロビットモデル

被説明変数が大小関係のある順序変数である場合、順序プロビットモデルを用いてモデルを推定します。ここでは内生性のある説明変数を利用します。

ここで利用するサンプルデータは、女性の健康状態と、運動習慣や健康保険への加入状況の関係を考察するために用意したものです。

- 変数 **health** は 5 段階で健康状態を示しています(poor/not good/fair/good/excellent)。
- 変数 **insured** は健康保険に加入しているか(yes/no)を示すダミー変数です。これに内生性を仮定します。

- サンプルデータをインポートし、変数 **health** の内容を確認します。

```
webuse womenhlth, clear


codebook health
```

```
-------------------------------------------------------------------------------
health                                                          Health status
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  status

                 range:  [1,5]                        units:  1
         unique values:  5                    missing .:  0/6,000

            tabulation:  Freq.   Numeric  Label
                           354         1  poor
                           635         2  not good
                           999         3  fair
                         1,622         4  good
                         2,390         5  excellent
```

- 順序プロビットモデルを推定します。拡張順序プロビット回帰のコマンド **eoprobit** を使用します。

```
eoprobit health i.exercise grade, entreat(insured =grade i.workschool)
```

```
Iteration 0:   log likelihood = -9110.6053
Iteration 1:   log likelihood = -9107.4138
Iteration 2:   log likelihood = -9105.5221
Iteration 3:   log likelihood = -9105.4378
Iteration 4:   log likelihood = -9105.4376

Extended ordered probit regression              Number of obs     =      6,000
                                                Wald chi2(4)      =     544.06
Log likelihood = -9105.4376                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
health						
exercise#insured						
yes#no	.5296149	.0614054	8.62	0.000	.4092626	.6499672
yes#yes	.5190249	.033697	15.40	0.000	.45298	.5850697
insured#c.grade						
no	.1079014	.0254855	4.23	0.000	.0579507	.1578522
yes	.1296456	.0106352	12.19	0.000	.108801	.1504901
insured						
grade	.3060024	.0101482	30.15	0.000	.2861122	.3258925
workschool						
yes	.5387767	.0448199	12.02	0.000	.4509313	.6266221
_cons	-3.592452	.1373294	-26.16	0.000	-3.861613	-3.323292
/health						
insured#c.cut1						
no	.6282326	.2465266			.1450493	1.111416
yes	-.7255086	.239525			-1.194969	-.2560482
insured#c.cut2						
no	1.594089	.2365528			1.130454	2.057724
yes	.4404531	.1956483			.0569894	.8239168
insured#c.cut3						
no	2.526424	.2308273			2.074011	2.978837
yes	1.332514	.1822525			.9753057	1.689722
insured#c.cut4						
no	3.41748	.2373824			2.952219	3.882741
yes	2.292828	.1734913			1.952792	2.632865
corr(e.insured, e.health)	.3414241	.0920708	3.71	0.000	.1502896	.5079557

例題 3

- ERM はパネルデータにも対応しています。
- パネルデータの場合、同一個体内のデータには相関が生じます。この相関を考慮してランダム効果モデルを推定します。
- 内生性のある変数を説明変数として利用する場合は、前述のクロスセクションにおける例題と同じ要領で、内生性を考慮した推定を実行します。

ここで利用するサンプルデータは例題 1 のデータと似ていますが、大学 ID でグループ化されています。時点の情報はありません。クロスセクションのデータで、大学 ID だけを使ったパネルデータとして分析を行います。

- サンプルデータをインポートし、パネルデータとして宣言します。

```
webuse class10re, clear

xtset
```

次のように表示されればパネルデータとして設定が完了しています。

```
. xtset

Panel variable: collegeid (balanced)
```

拡張ランダム効果線形モデル

- 最初に個体内の相関を考慮したランダム効果モデルの推定を実行します。推定結果に model1 という名前を付けて保存します。

```
xteregress gpa income

estimates store model1
```

```
(setting technique to bhhh)
Iteration 0:   log likelihood = -599.34987
Iteration 1:   log likelihood = -599.34987

Extended linear regression                      Number of obs     =      1,372
Group variable: collegeid                       Number of groups  =        100

                                                Obs per group:
                                                              min =          3
                                                              avg =       13.7
                                                              max =         20

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(1)      =    1062.86
Log likelihood = -599.34987                     Prob > chi2       =     0.0000
```

gpa	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
income	.0946589	.0029035	32.60	0.000	.0889681	.1003497
_cons	2.25479	.0325187	69.34	0.000	2.191055	2.318525
var(e.gpa)	.1210967	.0048074			.1120316	.1308954
var(gpa[col~d])	.0605453	.0101286			.0436198	.0840383

- 次に、単純なランダム効果モデルを推定します。推定結果に re という名前を付けて保存します。

```
xteregress gpa income


estimates store model1
```

```
Random-effects GLS regression                   Number of obs     =      1,372
Group variable: collegeid                       Number of groups  =        100

R-squared:                                      Obs per group:
     Within  = 0.4535                                         min =          3
     Between = 0.0885                                         avg =       13.7
     Overall = 0.3542                                         max =         20

                                                Wald chi2(1)      =    1065.41
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000
```

gpa	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
income	.0946657	.0029002	32.64	0.000	.0889814	.1003501
_cons	2.254745	.0331453	68.03	0.000	2.189782	2.319709
sigma_u	.25470225					
sigma_e	.34786985					
rho	.34899348	(fraction of variance due to u_i)				

- 個体内の相関を考慮したランダム効果モデルと、単純なランダム効果モデルの推定値を比較します。

```
estimates table model1 re, se equations(1)
```

Variable	model1	re
#1		
income	.09465892	.09466571
	.00290352	.00290024
_cons	2.2547901	2.2547454
	.03251866	.03314526
var(e.gpa)	.12109674	
	.0048	

- 説明変数 hsgpa(高校時代の GPA)に内生性を仮定してフィットします。推定結果に model2 という名前を付けて保存します。

```
xteregress gpa income, endogenous(hsgpa = income i.hscomp)


estimates store model2
```

```
(setting technique to bhhh)
Iteration 0:   log likelihood =  44.744857
Iteration 1:   log likelihood =  45.467002
                                 45.
(switching technique to bhhh)
Iteration 12:  log likelihood =   45.66803

Extended linear regression                      Number of obs     =      1,372
Group variable: collegeid                       Number of groups  =        100

                                                Obs per group:
                                                              min =          3
                                                              avg =       13.7
                                                              max =         20

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(2)      =    2908.41
Log likelihood = 45.66803                       Prob > chi2       =     0.0000
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
gpa						
income	.0558646	.0037641	14.84	0.000	.0484871	.063242
hsgpa	.9389094	.0783533	11.98	0.000	.7853398	1.092479
_cons	-.5592289	.2354489	-2.38	0.018	-1.0207	-.0977575
hsgpa						
income	.0427497	.001939	22.05	0.000	.0389494	.0465501
hscomp						
moderate	-.1445568	.0140903	-10.26	0.000	-.1721733	-.1169404
high	-.2338986	.0235708	-9.92	0.000	-.2800965	-.1877008
_cons	3.085305	.0170549	180.90	0.000	3.051878	3.118732
var(e.gpa)	.047076	.0024584			.0424959	.0521497
var(e.hsgpa)	.0563951	.0022393			.0521726	.0609594
corr(e.hsgpa, e.gpa)	.197656	.0867051	2.28	0.023	.0234398	.3602211
var(gpa[col~d])	.0630605	.009546			.0468711	.0848419
var(hsgpa[coll~d])	.0008785	.0007484			.0001655	.0046648
corr(hsgpa[coll~d], gpa[collegeid])	-.0991844	.2610895	-0.38	0.704	-.5485618	.3946522

- 3 つのモデルの推定値を比較します。

```
estimates table model1 model2 re, se equations(1)
```

Variable	model1	model2	re
#1			
income	.09465892	.05586457	.09466571
	.00290352	.00376406	.00290024
hsgpa		.93890944	
		.07835328	
_cons	2.2547901	-.55922891	2.2547454
	.03251866	.23544893	.03314526
gpa)	.1210	04707599	

結果から、内生性を仮定すると(model2)、推定値が大きく変化することがわかります。