Modèles non-linéaires pour données de panel

Guillaume Horny*

*Banque de France et UCLouvain

CEPII

Typologie des modèles non-linéaires en panel

- **Modèles poolés :**

$$g(\mathrm{E}[y_{it}|X_{it}]) = X_{it}\beta + u.$$

Ecriture identique à celle en coupe. Implique toutefois d'ajuster les matrices de variance pour la corrélation longitudinale

- **Modèles à effets fixes :**

$$g(\mathrm{E}[y_{it}|X_{it}, u_i]) = X_{it}\beta + u_i,$$

où u_i peut être corrélé avec X_{it}.

- **Modèles à effets aléatoires :**

$$g(\mathrm{E}[y_{it}|X_{it}, u_i]) = X_{it}\beta + u_i,$$

où u_i suit une distribution et est indépendant de X_{it}.

Aperçu des modèles non-linéaires

Différences avec les modèles linéaires:

- modèles à effets fixes sujets au problème des **paramètres incidents** lorsque N grand relativement à T
- modèles à effets aléatoires demandent généralement d'utiliser des procédures d'intégration numériques (cf transparent 45 de PS)

Modèles non-linéaires traités ici

Variable aléatoire:

- binaire: probit et logit
- continue et censurée: Tobit I et II

Généralités sur les panels sous Stata

Mise en forme des données :

- chaque observation correspond à un couple (i, t),
  pas de mise en forme "wide", sinon utiliser la commande `reshape`
- l'utilisateur doit déclarer les variables contenant les indices d'individus et de temps pour pouvoir utiliser les fonctions pour données de panel :

  ```
  xtset id year
  ```

Généralités sur les panels sous Stata (suite)

Description du panel

- `xtdescribe` nous renseigne sur les séquences d'observations pour les différents individus,
- `xtsum, xttab` nous renseignent sur les variations intra- et inter-individuelles de variables. Utile car les estimateurs à effets fixes sont peu précis lorsque les variations intra-individuelles sont faibles.

  ```
  xttab x y
  ```

- `xttrans` donne les proba de transitions d'une période à l'autre

  ```
  xttrans x
  ```

Modèles binaires et Tobit I

Commandes

VA	binaire	continue et censurée
poolé	`logit`	`tobit`
	`probit`	
effet fixe	`xtlogit,fe`	
effet aléatoire	`xtlogit, re`	`xttobit`
	`xtprobit, re`	

Modèle poolé

Commande en panel identique à celle en coupe, avec un ajustement pour la corrélation longitudinale:

```
logit y x, vce(cluster id)
probit y x, vce(cluster id)
```

Modèles à effets fixes

Problème des paramètres incidents:

- N effets fixes u_i + K régresseurs $\Rightarrow (N+K)$ paramètres,
- $(N+K) \rightarrow \infty$ lorsque $N \rightarrow \infty$,
- on doit éliminer les u_i,
  - possible pour modèles Logit, d'où le Logit conditionnel,
  - pas de solution pour le Probit,
  - pas de procédure sous Stata pour l'estimateur de Hahn et Newey (2004) du Tobit FE.

Commande:

```
xtlogit y x, fe
```

Modèles à un effet aléatoire

Contribution à la vraisemblance de l'individu i:

$$f(y_{i1},\ldots,y_{iT}|X_{i1},\ldots,X_{iT},\beta,\gamma) = \int_{\mathcal{U}} \left[\prod_{t=1}^{T} f(y_{it}|X_{it},u_i,\beta)\right] h(u_i|\gamma)du_i.$$

Pour un seul effet aléatoire:

- l'intégrale n'admet généralement pas de soution analytique,
- mais c'est une intégrale simple pour l'individu i.

$\Rightarrow$ les procédures d'intégration numériques "marchent bien"

Commandes pour des u_i gaussiens:

```
xtlogit y x, re
xtprobit y x, re
xttobit y x, ll(0) intpoints(20)
```

Extensions à plusieurs effets aléatoires

Pour capturer des corrélations multiniveaux, on peut vouloir spécifier plusieurs effets aléatoires
Exemple: des pays peuvent être regroupés (OCDE, zones géographiques...)
$\Rightarrow$ la contribution à la vraisemblance de chaque groupe contient une intégrale multiple

Module GLLAMM

- add-on gratuit pour Stata (www.gllamm.org),
- intégration numérique par une procédure de quadrature gaussienne adaptative.
  - \+ flexible, permet de spécifier de nombreuses variations autour des modèles usuels,
  - \- syntaxe peu intuitive, méthode générique d'intégration.

Exemple

- données issues de la Rand Health Insurance Experiment
  différentes couvertures maladies (“coinsurances”) sont assignés pour plusieurs années à des ménages
- panel non-cylindré
- données préparées pour Deb et Trivedi (2002) disponibles à:
  http://www.stata-press.com/data/mus.html

Exemple: code complet

```
use mus18data.dta, clear
list id year in 1/10
describe dmdu med mdu lcoins ndisease female age lfam child id year

xtset id year
xtdescribe
xtsum age lfam child
xttrans dmdu

global xlist lcoins ndisease female age lfam child
logit dmdu $xlist, vce(cluster id)
estimates store POOLED
xtlogit dmdu $xlist, fe nolog
estimates store FE
xtlogit dmdu $xlist, re nolog
estimates store RE
estimates table POOLED FE RE, equations(1) se b(%8.4f) stats(N ll)
```

Exemple

```
. list id year in 1/10

     +---------------+
     |     id   year |
     |---------------|
  1. | 125024      1 |
  2. | 125024      2 |
  3. | 125024      3 |
  4. | 125024      4 |
  5. | 125024      5 |
     |---------------|
  6. | 125025      1 |
  7. | 125025      2 |
  8. | 125025      3 |
  9. | 125025      4 |
 10. | 125025      5 |
     +---------------+
```

Exemple

```
describe dmdu med mdu lcoins ndisease female age lfam child id year

              storage  display
variable name   type   format    variable label
-------------------------------------------------------------------------------
dmdu            float  %9.0g     any MD visit = 1 if mdu > 0
med             float  %9.0g     medical exp excl outpatient men
mdu             float  %9.0g     number face-to-fact md visits
lcoins          float  %9.0g     log(coinsurance+1)
ndisease        float  %9.0g     count of chronic diseases -- ba
female          float  %9.0g     female
age             float  %9.0g     age that year
lfam            float  %9.0g     log of family size
child           float  %9.0g     child
id              float  %9.0g     person id, leading digit is sit
year            float  %9.0g     study year
```

Exemple

```
. xtset id year
      panel variable:  id (unbalanced)
       time variable:  year, 1 to 5, but with gaps
               delta:  1 unit
. xtdescribe

     id:  125024, 125025, ..., 632167
   year:  1, 2, ..., 5
          Delta(year) = 1 unit
          Span(year)  = 5 periods
          (id*year uniquely identifies each observation)
```

Exemple

```
Distribution of T_i:   min      5%     25%      50%      75%      95%
                         1       2       3        3        5        5

     Freq.  Percent     Cum. |  Pattern
 ---------------------------+---------
     3710    62.80    62.80 |  111..
     1584    26.81    89.61 |  11111
      156     2.64    92.25 |  1....
      147     2.49    94.74 |  11...
       79     1.34    96.07 |  ..1..
       66     1.12    97.19 |  .11..
       33     0.56    97.75 |  ..111
       33     0.56    98.31 |  .1111
       29     0.49    98.80 |  ...11
       71     1.20   100.00 | (other patterns)
 ---------------------------+---------
     5908   100.00          |  XXXXX
```

Exemple

```
. xtsum age lfam child

Variable         |      Mean   Std. Dev.       Min        Max |    Obs
-----------------+--------------------------------------------+-------
age      overall |  25.71844   16.76759         0   64.27515 |      N
         between |             16.97265         0   63.27515 |      n
         within  |             1.086687   23.46844   27.96844 | T-bar
                 |                                          |
lfam     overall |  1.248404   .5390681         0   2.639057 |      N
         between |             .5372082         0   2.639057 |      n
         within  |             .0730824   .3242075    2.44291 | T-bar
                 |                                          |
child    overall |  .4014168   .4901972         0          1 |      N
         between |             .4820984         0          1 |      n
         within  |             .1096116  -.3985832   1.201417 | T-bar
```

Exemple

```
. xttrans dmdu

    any MD |  any MD visit = 1 if
 visit = 1 |         mdu > 0
if mdu > 0 |         0          1 |     Total
-----------+----------------------+----------
         0 |     58.87      41.13 |    100.00
         1 |     19.73      80.27 |    100.00
-----------+----------------------+----------
     Total |     31.81      68.19 |    100.00
```

Exemple

```
estimates table POOLED FE RE, equations(1) se b(%8.4f) stats(N ll)
```

Variable	POOLED	FE	RE
lcoins	-0.1572		-0.2404
	0.0109		0.0163
ndisease	0.0503		0.0782
	0.0040		0.0055
female	0.3092		0.4631
	0.0446		0.0663
age	0.0043	-0.0342	0.0073
	0.0022	0.0184	0.0032
lfam	-0.2048	0.4788	-0.3022
	0.0470	0.2597	0.0645
child	0.0922	0.2705	0.1935
	0.0728	0.1685	0.1002
_cons	0.6039		0.8630
	0.1108		0.1569

Tobit II

Tobit II poolé

Commande:

```
heckman y1 x1, select(y2 = x2) vce(cluster id)
```

Tobit II à effets fixes

$$
\begin{aligned}
y_{1it} &= \begin{cases} y_{1it}^* & \text{if } y_{2it} = 1 \\ - & \text{if } y_{2it} = 0 \end{cases}, \\
y_{1it}^* &= X_{1it}\beta_1 + u_{1i} + w_{1it}, \\
y_{2it} &= \mathbb{1}[y_{2it}^* \geq 0] \\
y_{2it}^* &= X_{2it}\beta_2 + w_{2it},
\end{aligned}
$$

où w_{1it} et w_{2it} peuvent être corrélés.

Procédure:

- estimation des T modèles de sélections,
- calcul des T ratios de Mills correspondant $\widehat{\lambda}_t$,
- régression linéaire de y_1 sur $(X_1, d_1\widehat{\lambda}_1, \ldots, d_T\widehat{\lambda}_T)$.

Tobit II à effets fixes (suite)

Programme:

```
*** Etape 1: estimation des T modèles de sélection
probit y2 x2 if date == 1
predict xb1, xb
...

*** Etape 2: calcul des termes de correction
gen d1m1 = d1 * normalden(xb1) / normal(xb1)
...

*** Etape 3: régression linéaire
regress y1 x1 d1m1 d2m2...
```

Tobit II : modèle avec un facteur de charge

$$
\begin{aligned}
y_{1it} &= \begin{cases} y_{1it}^* & \text{if } y_{2it} = 1 \\ - & \text{if } y_{2it} = 0 \end{cases}, \\
y_{1it}^* &= X_{1it}\beta_1 + \lambda u_i + w_{1it}, \\
y_{2it} &= \mathbb{1}[y_{2it}^* \geq 0] \\
y_{2it}^* &= X_{2it}\beta_2 + u_i + w_{2it},
\end{aligned}
$$

où w_{1it} et w_{1it} sont indépendants. Si w_1, w_2 et u_i sont normaux centrés réduits, on a:

$$
\text{var}[(\lambda u_i + w_{1it}, u_i + w_{2it})'] = \begin{pmatrix} \lambda^2 + 1 & \lambda \\ \lambda & 2 \end{pmatrix}.
$$

Pas de routine pré-programmée, mais estimation possible avec GLLAMM.

Tobit II : réécriture du modèle avec un facteur de charge

Soient:

- $q_{it} = (y_{1it}, y_{2it})'$,
- $d_{1it} = \mathbb{1}[q_{it} = y_{1it}]$,
- $d_{2it} = \mathbb{1}[q_{it} = y_{2it}]$.

La variable q_{it} est une VA:

- normale lorsque $d_{1it} = 1$
- de Bernouilli (donc binomiale) lorsque $d_{2it} = 1$.

Exemple

- données simulées
- échantillon de 100 individus, observés sur 5 périodes.

Exemple

```
*** simulation des données
<SNIP>
. list id date x1 u y1 y2 in 1/10, clean
```

	id	date	x1	u	y1	y2
1.	1	1	-.15982537	.65174226	-.4322912	1
2.	1	2	-1.1814841	.65174226	.5927979	1
3.	1	3	.92545607	.65174226	4.320402	1
4.	1	4	-1.5261284	.65174226	.	0
5.	1	5	.89645764	.65174226	2.787327	1
6.	2	1	.9960802	.60328083	3.428029	1
7.	2	2	.33409693	.60328083	.	0
8.	2	3	.18082472	.60328083	.	0
9.	2	4	.52460684	.60328083	1.147292	1
10.	2	5	.09943156	.60328083	1.928639	1

Exemple

```
. *** création de q:  empile y1 et y2 dans une unique variable resp
. *** créé une variable num_eq, = 1 lorsque resp=y1et =2 pour resp=y2
. clonevar resp1 = y1
(160 missing values generated)
. clonevar resp2 = y2
. gen     row     = _n
. reshape long resp, i(row) j(num_eq)
(note: j = 1 2)

Data                               wide   ->   long
-----------------------------------------------------------------
Number of obs.                      500   ->    1000
Number of variables                  13   ->      13
j variable (2 values)                     ->   num_eq
xij variables:
                              resp1 resp2   ->   resp
-----------------------------------------------------------------
```

Exemple

```
. *** créé les indicatrices d1 et d2.
. tab num_eq, gen(d)
. <SNIP>
. sort id date num_eq
. list id date num_eq row x1 u y1 y2 in 1/11, clean
```

	id	date	num_eq	row	x1	u	y1	y2
1.	1	1	1	1	-.15982537	.65174226	-.4322912	1
2.	1	1	2	1	-.15982537	.65174226	-.4322912	1
3.	1	2	1	2	-1.1814841	.65174226	.5927979	1
4.	1	2	2	2	-1.1814841	.65174226	.5927979	1
5.	1	3	1	3	.92545607	.65174226	4.320402	1
6.	1	3	2	3	.92545607	.65174226	4.320402	1
7.	1	4	1	4	-1.5261284	.65174226	.	0
8.	1	4	2	4	-1.5261284	.65174226	.	0
9.	1	5	1	5	.89645764	.65174226	2.787327	1
10.	1	5	2	5	.89645764	.65174226	2.787327	1
11.	2	1	1	6	.9960802	.60328083	3.428029	1

Exemple

```
. *** mise en forme des variables explicatives
. *** 1ère équation
. gen d1_x1 = d1*x1
. gen d1_x2 = d1*x2

.
. *** 2d équation
. foreach var in x1 x2 x3 {
  2.      gen d2_‘var’ = d2*‘var’
  3. }
```

Exemple

```
*** estimation
eq fac: d2 d1

gllamm resp d1_x1 d1_x2 d1 d2_x1 d2_x2 d2_x3 d2, nocons /*
*/ i(id) family(gauss binom) link(ident probit) /*
*/ fv(num_eq) lv(num_eq) eq(fac) adapt nip(15)
```

GLLAMM

Consulter www.gllamm.org pour:

- le manuel ($\approx$ 140 pages)
- des fichiers de données pour s'entraîner
- une liste des papiers utilisant GLLAMM

Conclusion

Points forts et faibles de Stata:

+ fonctions pour les modèles usuels pour données de panel
+ les méthodes en 2 étapes sont faciles à programmer
+ extensions possibles avec GLLAMM aux modèles multiniveaux ou à équations simultanées, grâce aux facteurs de charge
- au delà, Stata très peu flexible
  Alternatives pour les économètres créatifs: R, GAUSS, Matlab...

Merci de votre attention!

guillaume.horny@banque-france.fr