Non réponse à l'enquête emploi et modèles probit spatiaux

Vincent Loonis : Insee

Rennes 2012

plan

1. Rappel sur le plan de sondages de l'EEC,
2. Pourquoi un nouveau modèle de non réponse de l'EEC ?
3. Le nouveau modèle,
4. la méthode d'estimation,
5. les résultats,
6. les perspectives.

Rappel sur le plan de sondages de l'EEC

1. 30 millions de logements répartis en 1,5 millions d'aires d'environ 20 logements contigus ou très proches.
2. 3000 aires enquêtées chaque trimestre,
3. $\frac{1}{6}$ des aires renouvelées chaque trimestre (panel rotatif),
4. 1 aire entièrement affectée au même enquêteur,
5. 1 enquêteur peut avoir 1 ou plusieurs aires.

On ne s'intéresse ici qu'aux aires entrantes de l'année 2010, soit

- n=24125 logements enquêtés (indicés par k), dont 21% de non répondants ,
- répartis dans 1272 aires (indicées par j),
- attribuées à 516 enquêteurs différents (indicés par i).

Des spécificités non expliquées ou non prises en compte

1. plus faible taux de non réponse de l'ensemble des enquêtes auprès des ménages de l'Insee
   - seule enquête aréolaire : effet d'entraînement *(M.Christine 2002)* dû à la proximité des logements enquêtés ?
   - Enquête par panel : effet de l'expérience des enquêteurs ?
   - Sujet consensuel dans la société contribuant à l'adhésion des enquêtés ?
2. Première enquête dont l'échantillon a été entièrement sélectionné dans les fichiers de la taxe d'habitation.
   - La pérennité des identifiants permet de décrire les enquêtés (répondants ou non) par leurs caractéristiques *au moment de l'enquête* : avec un peu de patience...
   - Test de la non ignorabilité de la non réponse avec des descripteurs liés à la mesure du chômage.

Effet d'entraînement

La propension d'un individu à répondre (ou ne pas répondre) à l'enquête dépend de ses caractéristiques mais aussi de la propension des autres enquêtés.

$$\underbrace{\begin{cases} y_k^* = \rho_1 \sum_{k' \neq k} w_{kk'}^1 y_{k'}^* + x_k^t \beta + \epsilon_k \\ y_k = 1(y_k^* > 0) \end{cases}}_{\text{Modèle Spatial AutoRégressif (SAR)}} \iff \underbrace{\begin{cases} y^* = \rho_1 w^1 y^* + x\beta + \epsilon \\ y = 1(y^* > 0) \end{cases}}_{\text{écriture matricielle}}$$

1. y_k^* propension inobservable de k à ne pas répondre,
2. y_k (observée) valant 1 si k est non répondant et 0 autrement,
3. x_k vecteur de variables explicatives,
4. β paramètre inconnu,
5. $w_{kk'}^1$ scalaire mesurant la *proximité* entre k et k',
6. ρ_1 coefficient inconnu mesurant l'effet d'entraînement.

Effet d'entraînement

w^1 est une matrice caractéristique des modèles d'économétrie spatiale (et de celle des réseaux). Elle est normalisée à 1 en ligne. Elle peut être paramétrée sous les formes suivantes :

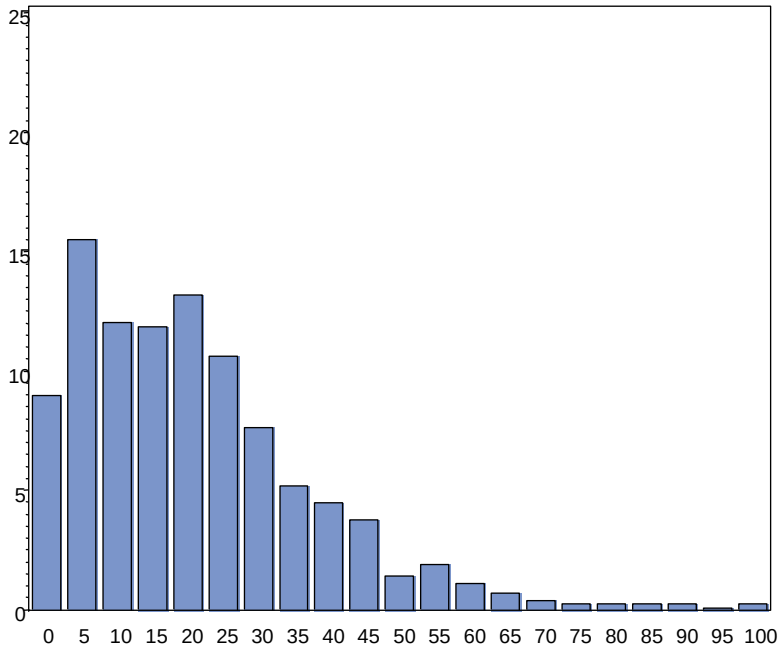
- à un coefficient de normalisation près :
  - $w^1_{kk'} = 1$ si k et k' on une frontière commune , 0 autrement,
  - $w^1_{kk'} = \exp(-\alpha d_{kk'})$ où $d_{kk'}$ est la distance entre k et k',
- Dans notre modèle :
  - $w^1_{kk'} = \frac{1}{n_j - 1}$ si $k \neq k'$ et appartiennent à la même aire : j, et 0 autrement.
  - w^1 est bloc diagonale, 1 bloc par aire. Les 1272 blocs sont de la forme :

$$w^{1,j} = \frac{1}{n_j - 1}(e_{n_j} e_{n_j}^t - I_{n_j}), 1 \leq j \leq 1272, \sum_j n_j = n$$

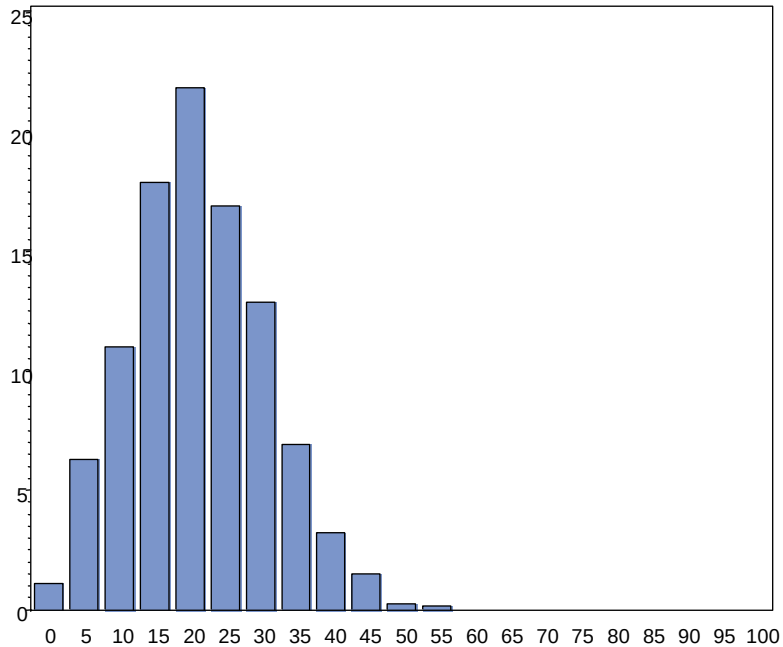
Effet d'entraînement

Répartition par aire du taux de non réponse :

Observé



Théorique sous l'hypothèse d'un mécanisme de réponse homogène



Effet enquêteur

La propension d'un individu à répondre (ou ne pas répondre) à l'enquête dépend des caractéristiques observables et inobservables de l'enquêteur.

$$\underbrace{\begin{cases} y_k^* = x_k^t \beta + \epsilon_k \\ \epsilon_k = \rho_2 \sum_{k' \neq k} w_{kk'}^2 \epsilon_{k'} + u_k \\ y_k = 1(y_k^* > 0) \end{cases}}_{\text{Spatial Error Model (SEM)}} \iff \begin{cases} y^* = x\beta + \epsilon \\ \epsilon = \rho_2 w^2 \epsilon + u \\ y = 1(y^* > 0) \end{cases}$$

1. x_k vecteur de variables explicatives, intégrant des caractéristiques observables de l'enquêteur : expérience, distance domicile-aire.
2. β paramètre inconnu,
3. ρ_2 coefficient inconnu mesurant l'effet enquêteur.

Effet enquêteur

Dans notre modèle :

- $w^2_{kk'} = \frac{1}{n_i - 1}$ si $k \neq k'$ sont interviewés par le même enquêteur i ayant en tout n_i enquêtes, et 0 autrement.
- w^2 est bloc diagonale, constituée de 516 blocs. Chacun des blocs est de la forme :

$$w^{2,i} = \frac{1}{n_i - 1}(e_{n_i} e_{n_i}^t - I_{n_i}), 1 \leq i \leq 516, \sum_i n_i = n$$

Le modèle final

Au final, le modèle est :

$$\begin{cases} y^* = \rho_1 w^1 y^* + x\beta + \epsilon \\ \epsilon = \rho_2 w^2 \epsilon + u \\ u \longrightarrow \mathcal{N}(0,1) \\ y = 1(y^* > 0) \end{cases}$$

ρ_1 et ρ_2 sont identifiables dès lors que $W^1 \neq W^2$. C'est le cas si au moins 1 enquêteur a plusieurs aires.

Estimation dans le cas général

On peut réécrire le modèle sous la forme :

$$\begin{cases} y^* = (I_n - \rho_1 w^1)^{-1} x\beta + (I_n - \rho_1 w^1)^{-1}(I_n - \rho_2 w^2)^{-1} u \\ u \longrightarrow \mathcal{N}(0,1) \\ y = 1(y^* > 0) \\ V(y^*|x) = [(I_n - \rho_2 w^2)^t (I_n - \rho_1 w^1)^t (I_n - \rho_1 w^1)(I_n - \rho_2 w^2)]^{-1} \end{cases}$$

1. modèle probit non linéaire avec matrice de variance non scalaire,
2. estimation par le maximum de vraisemblance simulé,
3. estimation par des méthodes bayésiennes,
4. méthode rendue plus délicate à utiliser à cause de calculs liés à la matrice $V(y^*|x)$.

Estimation dans notre modèle

On peut réécrire le modèle sous la forme Durbin Spatial étendue :

$$\begin{cases} y^* = \underbrace{(I_n - \rho_1 w^1)^{-1}(I_n - \rho_2 w^2)^{-1}}_{=\Omega^{-1}}(x\beta + w^2 x \underbrace{\gamma}_{=-\rho_2\beta} + u) \\ u \longrightarrow \mathcal{N}(0, \sigma^2) \\ y = 1(y^* > 0) \\ V(y^*|x) = [\Omega^t \Omega]^{-1} \end{cases}$$

1. $\mathbb{E}(y_k) = \Phi(\frac{\Omega^{-1}(x\beta + w^2 x\gamma)}{(\Omega^t\Omega)^{\frac{1}{2}}_{k,k}}) = \mu_k(\rho_1, \rho_2, \beta, \gamma|x)$,
2. où Φ est la fonction de répartition d'une loi normale.
3. On peut obtenir une estimation convergente des paramètres inclus dans μ_k par le pseudo-maximum de vraisemblance :

On dispose de données telles que $\mathbb{E}(y_k) = \mu_k(\theta, x)$ où θ est un paramètre inconnu et μ_k une fonction connue. On peut obtenir une estimation convergente de θ en maximisant n'importe quelle vraisemblance, appelée pseudo vraisemblance, se mettant sous la forme :

$$\log g = A(\mu) + B(y) + c(\mu)y.$$

Pour un enquêteur i, ayant m_i aires de taille respective n_j on prend *(Lollivier 2009)* :

$$\begin{cases} A(\mu) = \sum_{j=1}^{m_i} \sum_{k=1}^{n_j} \log(1 - \mu_{j,k}) & B(y) = 0 \\ c(\mu) = \begin{bmatrix} \log(\frac{\mu_{1,1}}{1-\mu_{1,1}}) \\ \vdots \\ \log(\frac{\mu_{j,k}}{1-\mu_{j,k}}) \\ \vdots \\ \log(\frac{\mu_{m_i,n_{m_i}}}{1-\mu_{m_i,n_{m_i}}}) \end{bmatrix} \end{cases}$$

la pseudo vraisemblance est alors :

$$\log g = \sum_{i=1}^{516}\sum_{j=1}^{m_i}\sum_{k=1}^{n_j} \log(1-\mu_{j,k}) + \sum_{i=1}^{516}\sum_{j=1}^{m_i}\sum_{k=1}^{n_j} \log\left(\frac{\mu_{j,k}}{1-\mu_{j,k}}\right) y_k =$$

$$\underbrace{\sum_{i=1}^{516}\sum_{j=1}^{m_i}\sum_{k=1}^{n_j} y_k \log(\mu_{j,k}) + \sum_{i=1}^{516}\sum_{j=1}^{m_i}\sum_{k=1}^{n_j} (1-y_k)\log(1-\mu_{j,k})}_{\text{Vraisemblance d'un probit non linéaire empilé}}$$

il suffit donc

1. d'estimer les paramètres d'un probit non linéaire empilé, dont l'espérance est $\frac{\Omega^{-1}(x\beta + w^2 x\gamma)}{\Omega_{k,k}^{\frac{1}{2}}}$ avec Ω^{-1} et $\Omega_{k,k}^{\frac{1}{2}}$ faciles à calculer car Ω est bloc-diagonale et que chaque bloc est facile à inverser.
2. d'estimer la précision de ces estimations par bootstrap. On bootstrap au niveau enquêteur, qui est le plus fin niveau *iid*.

Les résultats

TABLE: Non réponse à l'EEC : Modèle probit *spatial*

	Modèle probit			
	spatial		classique	
	$\hat{\rho}, \hat{\beta}$	$H_0 : \beta, \rho = 0$	$\hat{\beta}$	$H_0 : \beta = 0$
Effet d'entraînement ρ_1	0.283466		-	
Effet enquêteur ρ_2	0.352939		-	
constante	-0.625505		-1.1156	
Caractéristiques observables de l'enquêteur				
moins de 10 ans d'expériences	-0.002796	ns	-0.0146	
plus de 10 ans d'expériences	-0.006304		-0.0157	
temps d'accès à l'aire depuis le domicile	-0.028388	ns	0.0230	ns
Caractéristiques observables du logement				
Date de l'enquête				
1er trimestre	-0.111704	ns	-0.1197	
deuxième trimestre	0.042707	ns	0.1195	
troisième trimestre	0.056964		0.1476	
4 ème trimestre	référence			
taille du ménage				
0	0.395800		0.3989	
1 personne	0.226062		0.2367	
2 personnes	0.096392		0.1033	
3 personnes et +	référence			
âge du chef de ménage				
Moins de 30 ans	0.158959		0.1488	
30 à 40 ans	0.152271		0.1455	
40 à 50 ans	0.179626		0.1780	
50 à 60 ans	0.134102		0.1292	
60 ans et +	référence			
autres caractéristiques				
Indemnités chômage	-0.101227		-0.0905	
salarié	-0.030822	ns	-0.0417	ns
agriculteurs	-0.189820		-0.2250	

Les résultats

Selon notre modèle, on a $\mathbb{P}_k = \mathbb{P}(\text{k soit non répondant}) = \Phi\left(\frac{\Omega^{-1}(x\beta + w^2 x\gamma)}{\Omega_{k,k}^{\frac{1}{2}}}\right) = \mu_k(\rho_1, \rho_2, \beta, \gamma | x)$

1. on retrouve bien : $\frac{\sum_{k=1}^{n} \hat{\mathbb{P}}_k}{n} = \frac{\sum_{k=1}^{n} \mu_k(\hat{\rho}_1, \hat{\rho}_2, \hat{\beta}, \hat{\gamma} | x)}{n} \approx 21\%$
2. On peut calculer aussi $\frac{\sum_{k=1}^{n} \mu_k(0, \hat{\rho}_2, \hat{\beta}, \hat{\gamma} | x)}{n} \approx 27.8\%$

Selon notre modèle, la dimension aréolaire du plan de sondage de l'EEC conduit à réduire le taux de non réponse de 7 points.

Perspectives

1. Liens avec les données de panel,
2. quelques questions d'identifiabilité,
3. Tester des matrices W^1 plus *réalistes* en tenant compte de la disponibilité des (X,Y) (Triangulation de Delaunay...),
4. Tester la méthode sur d'autres enquêtes non aréolaires (Patrimoine 2010), PIAAC
5. Passer à matlab...

Merci de votre attention ! ! !