

# Mécanismes de sélection dans les enquêtes et non-réponse non-ignorable

---

*Exposé aux Journées de Méthodologie Statistique – mercredi 16 mars 2005.*

Eric Gautier

INSEE - Unité Méthodes Statistiques

18 Boulevard Adolphe Pinard,

75675 Paris cedex 14

[eric.gautier@insee.fr](mailto:eric.gautier@insee.fr)

# PLAN

---

## I. La sélection dans les enquêtes

La sélection non-ignorable :

II. Approche paramétrique

III. Approche semi paramétrique

# I. La sélection dans les enquêtes

- Tirage initial : on tire **aléatoirement** des logements, la loi du tirage peut dépendre des variables de la base : probabilités inégales...  
⇒ La sélection est parfaitement connue.
- Non réponse totale : sélection de phase 2, la loi de répondre sachant que l'on est échantillonné est inconnue.
- Non réponse partielle : sélection de phase 3, pour des unités sélectionnées à l'issu des phases 2 et 3, certaines ne répondent pas à tous les items ⇒ autant de sélections que de variables.

Nous adoptons **une approche modèle** : on suppose que les variables que l'on mesure sont des réalisations de variables aléatoires.

- S la sélection est aléatoire, S vaut 1 si observation et 0 sinon.  
modèle : S vaut 1 si  $S^* = X\beta + u > 0$ , 0 sinon.
- Biais de sélection, par ex pour la moyenne d'une variable lorsque

$$E[Y|S = 1] \neq E[Y|S = 0]$$

et que l'on souhaite inférer sur une moyenne non conditionnelle

# Terminologie

Y partiellement observée et X parfaitement observée sont aléatoires

- **Missing Completely At Random (MAR)** : la sélection pour la variable d'intérêt est indépendante de toutes les variables de l'enquête

$$P(S = 1|Y, X) = cste$$

revient à  $P((Y, X) \in A|S = 1) = P((Y, X) \in A|S = 0) = P((Y, X) \in A)$ ,

- **Missing At Random (MAR)** (assimilé ici au cas particulier=ignorable),  
X : **toutes** les co-variables disponibles

$$P(S = 1|Y, X) = P(S = 1|X),$$

revient à  $P(Y \in A|S = 1, X = x) = P(Y \in A|S = 0, X = x) = P(Y \in A|X = x)$ ,

en quelque sorte  $L(Y|X) \amalg L(S|X)$

on peut ignorer la loi des données (approche basée sur le plan de sondage, repondération) ou la sélection (cas usuels en imputation ou économétrie)

- **NMAR** sinon

# Une sélection MAR peut poser des difficultés

Ex: hyp = loi de  $Y$  sachant  $X_1 = x_1$  et  $X_2 = x_2$  de moyenne  $a + bx_1 + cx_2$   
on s'intéresse à la loi de  $Y$  sachant  $X_1 = x_1$   
mais la sélection sur représente des modalités de  $X_2$

$$P(S = 1 | X_2 = 0) = p, P(S = 1 | X_2 = 1) = q, p \neq q$$

Pas de biais de sélection lorsque  $L(S | X_1 = x_1) \perp\!\!\!\perp L(X_2 | X_1 = x_1)$

en effet : la loi d'intérêt en population générale est de moyenne

$$a + cE[X_2 | X_1 = 0] + (b + cE[X_2 | X_1 = 1] - cE[X_2 | X_1 = 0])x_1.$$

et MCO  $\Rightarrow$  inférence sur  $E[Y | X_1 = x_1, S = 1]$

$$= a + cE[X_2 | X_1 = 0, S = 1] + (b + cE[X_2 | X_1 = 1, S = 1] - cE[X_2 | X_1 = 0, S = 1])x_1.$$

# La sélection non-ignorable

---

- NR car question sensible : revenu, patrimoine, pratiques sexuelles...
- Où est ce que cela peut se produire?
  - Au tirage? Oui mais pas pour des enquêtes ménage standard de l'INSEE, cas des plans de sondage informatifs.
  - A l'étape de la NR totale? OUI!  
ex. une lettre avis présente le sujet de l'enquête ...  
d'autant plus que peu de co-variables sont à ce stade disponibles pour obtenir de l'indépendance en conditionnant.
  - A l'étape de la NR partielle? Toujours oui, certains disent que le biais est moindre compte tenu des co-variables.

# II. Approche paramétrique

## II.1. Mélanges de lois :

II.1.1 Ex 1: nous ne disposons pas de co-variables,

modèle : la loi  $L(Y|S = 1)$  est normale de moyenne  $\mu_1$  et de covariance  $\sigma_1^2$   
 $L(Y|S = 0)$   $\mu_0$   $\sigma_0^2$

rq : MCAR  $\Leftrightarrow \mu_0 = \mu_1$  et  $\sigma_0^2 = \sigma_1^2$

la loi mélangeante (loi de sélection) = une loi de Bernoulli de paramètre  $p$ ,

alors la loi mélangée a pour moyenne  $p\mu_0 + (1-p)\mu_1$

et pour variance  $p\sigma_1^2 + (1-p)\sigma_0^2 + p(1-p)(\mu_1 - \mu_0)^2$

**PB d'identifiabilité**, impossible si pas d'observations chez les non-sélectionnés d'estimer  $\mu_0$  et  $\sigma_0^2$ .

SOLUTION : imposer des restrictions sur les paramètres,

rq suite :  $\mu_0 = \mu_1$  et  $\sigma_0^2 = \sigma_1^2$  revient à faire l'hypothèse MCAR

II.1.2 Ex 2: Nous disposons d'une co-variable X.

modèle : la loi  $L((Y, X)|S = 1)$  est normale de moyenne  $\begin{pmatrix} \mu_{1Y} \\ \mu_{1X} \end{pmatrix}$   
 et de matrice de covariance  $\begin{pmatrix} \sigma_{1YY}^2 & \sigma_{1XY}^2 \\ \sigma_{1XY}^2 & \sigma_{1XX}^2 \end{pmatrix}$

$L((Y, X)|S = 0)$  est de moyenne  $\begin{pmatrix} \mu_{0Y} \\ \mu_{0X} \end{pmatrix}$  et de cov  $\begin{pmatrix} \sigma_{0YY}^2 & \sigma_{0XY}^2 \\ \sigma_{0XY}^2 & \sigma_{0XX}^2 \end{pmatrix}$

8 paramètres sur 11 sont identifiables :  $p, \mu_{1X}, \mu_{0X}, \mu_{1Y}, \sigma_{1YY}, \sigma_{1XX}, \sigma_{1XY}$   
 et  $\sigma_{0XX}$ .

Exemple d'hypothèses rendant le modèle identifiable :

- $L(Y|S = 1, X = x) = L(Y|S = 0, X = x)$  = hypothèse MAR,
- $L(X|S = 1, Y = y) = L(X|S = 0, Y = y)$  = « hypothèse protectrice »,

Mais aussi :

- On peut aussi imposer que la sélection ne dépendent non plus ou de X ou de Y mais plutôt de  $X + \lambda Y$ .  $\lambda$  non identifiable mais permet d'évaluer la sensibilité des inférences.



## II.2. Modèles de sélection :

$$\begin{cases} Y &= X\beta_1 + \varepsilon_1, \\ S^* &= X\beta_2 + \varepsilon_2, \end{cases} \quad \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \text{ suit la loi } N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right]$$

Le mécanisme de sélection est bien NMAR :

$$\begin{aligned} P(S^* < 0 | Y = y, X = x) &= P\left(\varepsilon_2 < -x\beta_2 \mid \varepsilon_1 = \frac{y - x\beta_1}{\sigma}, X = x\right) \\ &= 1 - \Phi\left(\frac{x\beta_2 + \rho\sigma^{-1}(y - x\beta_1)}{\sqrt{1 - \rho^2}}\right), \end{aligned}$$

$L(Y | X = x, S^* \geq 0)$  est de moyenne  $x\beta_1 + \rho\sigma\lambda(x\beta_2)$   
et de variance  $\sigma^2 + \rho^2\sigma^2(x\beta_2\lambda(x\beta_2) - (\lambda(x\beta_2))^2)$

$L(Y | X = x, S^* < 0)$  est de moyenne  $x\beta_1 - \rho\sigma\lambda(-x\beta_2)$   
et de variance  $\sigma^2 + \rho^2\sigma^2(x\beta_2\lambda(-x\beta_2) - (\lambda(-x\beta_2))^2)$

test de sélection MAR = test de Student de nullité du coef de l'inverse du ratio de Mills  $\lambda$  dans régression augmentée via une estim. en 2 étapes.

$$L_i = P(S^{*3} \geq 0 | X = x_i, Y = y_i) \frac{1}{\sigma} \varphi\left(\frac{y_i - x_i \beta_1}{\sigma}\right)$$

■ Remarques :

- maximum de vrais

$$= \Phi\left[\frac{1}{\sqrt{1-\rho^2}}\left(x_i \beta_2 + \frac{\rho}{\sigma}(y_i - x_i \beta_1)\right)\right] \frac{1}{\sigma} \varphi\left(\frac{y_i - x_i \beta_1}{\sigma}\right),$$

si  $i$  est sélectionné et sinon  $L_i = P(S^{*3} < 0 | X = x_i) = \int_{\mathcal{R}} \Phi\left(-\frac{x_i \beta_2 + \rho u}{\sqrt{1-\rho^2}}\right) \varphi(u) du,$

- 2 étapes permet de faire de l'imputation par prédiction.
- Imputer par simulation  $\Rightarrow$  ne plus faire en 2 étapes (existe en 3) par acceptation rejet en simulant des couples  $(Y, S^*)$
- Très sensible à une mauvaise spécification
- L'hypothèse de loi conditionnelle normale en population générale n'est pas testable.

# Ex 1 : Mélange de lois, appariement EE-ERF, 99

---

- EE & ERF, loi du salaire ERF dans les 3 groupes:
  - Les réponses en clair à EE
  - Les réponses en tranche à EE
  - Les non-réponses
- Test d'égalité des comportements
- 4 sous groupes dans lequel nous estimons le mélange
  - Les hommes de CS commençant par 3 et 4
  - Les femmes de CS commençant par 3 et 4
  - Les hommes de CS commençant par 5 et 6
  - Les femmes de CS commençant par 5 et 6

# Tests hommes 3-4

## Test d'égalité des coefficients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	61	0.14004	1.62	0.0017
Denominateur	5118	0.08646		

⇒ on rejette l'égalité des coeffs entre les 2 groupes

## Test d'égalité des coefficients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	61	0.24423	2.82	<.0001
Denominateur	5118	0.08646		

⇒ on rejette l'égalité des coeffs entre les 2 groupes

## Test d'égalité des coefficients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	122	0.18161	2.10	<.0001
Denominateur	5118	0.08646		

⇒ on rejette l'égalité des coeffs entre les 3 groupes

# Tests femmes 3-4

## Test d'égalité des coefficients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	57	0.08527	1.26	0.0903
Denominateur	3381	0.06757		

## Test d'égalité des coefficients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	57	0.07439	1.10	0.2819
Denominateur	3381	0.06757		

## Test d'égalité des coefficients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	114	0.07959	1.18	0.0987
Denominateur	3381	0.06757		

# Tests hommes 5-6

## Test d'égalité des coefficients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	61	0.06211	1.18	0.1577
Denominateur	7225	0.05255		

## Test d'égalité des coefficients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	61	0.089	1.69	0.0006
Denominateur	7225	0.05255		

## Test d'égalité des coefficients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	122	0.07516	1.43	0.0014
Denominateur	7225	0.05255		

# Tests femmes 5-6

## Test d'égalité des coefficients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	60	0.07273	1.07	0.3277
Denominateur	6340	0.06780		

## Test d'égalité des coefficients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	60	0.08243	1.22	0.1233
Denominateur	6340	0.06780		

## Test d'égalité des coefficients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	120	0.07516	1.13	0.1547
Denominateur	6340	0.06780		

# Ex2 : modèle de sélection, la sélection = la non-réponse en clair ou en tranche.

var	coef	Student	p-value	var	coef	Student	p-value	var	coef	Student	p-value
Intercept	11,647	1,039	<,0001	cs45	0,222	0,023	<,0001	EB	-0,061	0,019	0,002
mills	-1,032	0,580	0,075	cs46	-0,021	0,123	0,864	EC	-0,073	0,021	0,001
lnheur	0,426	0,020	<,0001	cs47	0,075	0,052	0,146	ED	-0,209	0,092	0,024
tpart	-0,286	0,032	<,0001	cs48	0,026	0,096	0,786	EE	-0,126	0,055	0,021
femme	-0,069	0,019	0,000	cs52	0,039	0,016	0,012	EF	-0,172	0,072	0,017
a1	-0,111	0,019	<,0001	cs53	0,091	0,038	0,017	EG	0,003	0,025	0,901
a3	-0,094	0,074	0,201	cs54	-0,171	0,131	0,194	EH	-0,066	0,017	0,000
a4	-0,123	0,113	0,278	cs55	-0,024	0,047	0,605	EJ	-0,212	0,060	0,000
anc1	-0,096	0,042	0,022	cs56	-0,176	0,019	<,0001	EK	-0,120	0,048	0,012
anc3	0,081	0,014	<,0001	cs63	-0,085	0,039	0,027	EM	-0,185	0,038	<,0001
anc4	0,097	0,026	0,000	cs64	-0,141	0,058	0,015	EN	-0,121	0,021	<,0001
etcl	-0,002	0,030	0,944	cs65	0,165	0,094	0,080	EP	-0,343	0,109	0,002
dip1	0,269	0,034	<,0001	cs67	-0,342	0,131	0,009	EQ	-0,246	0,070	0,001
dip3	0,160	0,025	<,0001	cs68	-0,284	0,084	0,001	ER	-0,120	0,018	<,0001
dip4	0,046	0,025	0,065	cs69	-0,144	0,036	<,0001	tent1	-0,022	0,034	0,505
dip5	0,027	0,012	0,020	fonc0	0,035	0,016	0,029	tent2	0,042	0,055	0,440
dip7	-0,061	0,010	<,0001	fonc2	0,019	0,015	0,198	tent3	0,020	0,032	0,534
dipm	5,172	2,946	0,079	fonc3	0,071	0,099	0,474	tentm	-0,044	0,012	0,000
cs31	-0,131	0,274	0,633	fonc4	0,010	0,016	0,536	tsoi	-0,009	0,023	0,708
cs33	0,507	0,049	<,0001	fonc5	0,022	0,018	0,235	tnui	0,084	0,019	<,0001
cs34	0,539	0,049	<,0001	fonc6	0,100	0,035	0,004	tsam	0,050	0,030	0,101
cs35	0,149	0,135	0,269	fonc7	0,032	0,014	0,024	tdim	0,094	0,031	0,002
cs37	0,304	0,109	0,005	fonc8	0,169	0,073	0,020	idf	-0,054	0,088	0,544
cs38	0,230	0,135	0,088	fonc9	0,241	0,050	<,0001	proxy	-0,130	0,073	0,075
cs42	0,271	0,022	<,0001	foncm	-0,691	0,398	0,083	quali1	0,769	0,441	0,081
cs43	0,326	0,065	<,0001	EA	-0,154	0,043	0,000	sirmiss	-0,024	0,030	0,410



# III. approches semi paramétrique

1. Une méthode par moindres carrés pondérés (Beaumont 00')
2. Une méthode par vraisemblance empirique (Qin, Leung et Shao 02')

La vrais est construite sur les observations conjointes de  $(y_i, x_i)_{i=1}^n$

La loi de la sélection est paramétrée :  $w(x, y, \theta) = P_\theta(S = 1 | X = x, Y = y)$

En notant  $f(x, y)$  la densité du couple on obtient

$$\left\{ \prod_{i=1}^n w(x_i, y_i, \theta) f(x_i, y_i) \right\} \prod_{i=n+1}^N \iint (1 - w(x, y, \theta)) f(x, y) dx dy = W^n (1 - W)^{N-n} \prod_{i=1}^n \frac{w(x_i, y_i, \theta) f(x_i, y_i)}{W}$$

où  $W = p(M = 0) = \iint w(x, y, \theta) f(x, y) dx dy$

La vrais observée sur tout l'échantillon s'écrirait plutôt

$$\left\{ \prod_{i=1}^n w(x_i, y_i, \theta) f(x_i, y_i) \right\} \prod_{i=n+1}^N \int (1 - w(x, y, \theta)) f(x_i, y) dy$$

Les contraintes de calage :

$$p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i (w(x_i, y_i, \theta) - W) = 0, \quad \sum_{i=1}^n p_i (x_i - \mu_x) = 0,$$

- Résolution en calculant les 0 de la fonction de plusieurs variables

$$\left\{ \begin{array}{l} \lambda_2 = \frac{N/n - 1}{1 - W} \\ \sum_{i=1}^n \frac{x_i - \bar{X}_N}{1 + \lambda_1(x_i - \bar{X}_N) + \lambda_2(w(y_i, x_i, \theta) - W)} = 0 \\ \sum_{i=1}^n \frac{w(y_i, x_i, \theta) - W}{1 + \lambda_1(x_i - \bar{X}_N) + \lambda_2(w(y_i, x_i, \theta) - W)} = 0 \\ \sum_{i=1}^n \frac{w(y_i, x_i, \theta) - W}{1 + \lambda_1(x_i - \bar{X}_N) + \lambda_2(w(y_i, x_i, \theta) - W)} = \sum_{i=1}^n \frac{\partial \log w(y_i, x_i, \theta)}{\partial \theta} \end{array} \right.$$

- On obtient  $(\hat{\theta}, \hat{W}, \hat{\lambda}_1, \hat{\lambda}_2)$  puis  $\hat{p}_i = \frac{1}{n [1 + \hat{\lambda}_1(x_i - \bar{X}_N) + \hat{\lambda}_2(w(y_i, x_i, \hat{\theta}) - \hat{W})]}$

- Un estimateur de la moyenne, approche modèle, convergent et asymptotiquement normal  $\sum_{i=1}^n \hat{p}_i y_i$
- Possibilité de faire de la pseudo vraisemblance empirique pour obtenir un estimateur de la vraisemblance semi paramétrique non biaisé vis à vis du plan de sondage, si pas de NR partielle.
- Développements : possibilités de tests de nullité du coefficient de Y dans le Logit, contraste de chi2 pour simplifier le problème d'optimisation.

# Conclusion

---

- La sélection peut engendrer des biais pour une inférence de sondage basée sur le plan de sondage ou pour l'économètre
- Importer par exemple une équation d'une enquête à l'autre nécessite d'être conscient de ces mécanismes
- Des tests des hypothèses existent
- La multiplicité des approches rend compte de la difficulté du problème et aucune n'est vraiment satisfaisante
- Un bon travail de collecte et de suivi pourrait simplifier le difficile travail statistique.
- Les approches semi-paramétriques sont peu robustes ou non identifiable. La bonne approche est peut être non paramétrique.

# Annexe : régression mélange 1

régression hommes cadres et prof int / InWrf avec 3 types de déclaration

var	coef	Student	p-value	var	coef	Student	p-value	var	coef	Student	p-value
scont1	11,412	0,095	<,0001	anc30	0,042	0,017	0,013	dip50	-0,001	0,020	0,951
stran1	10,952	0,264	<,0001	anc31	0,087	0,052	0,093	dip51	-0,011	0,061	0,863
snrep1	11,326	0,419	<,0001	anc32	0,089	0,077	0,252	dip52	0,197	0,085	0,021
Inheur0	0,154	0,022	<,0001	anc40	0,115	0,016	<,0001	dip70	-0,083	0,024	0,001
Inheur1	0,317	0,063	<,0001	anc41	0,121	0,047	0,010	dip71	-0,059	0,068	0,385
Inheur2	0,122	0,107	0,256	anc42	0,264	0,072	0,000	dip72	-0,005	0,098	0,961
tpart0	-0,269	0,033	<,0001	ancm0	-0,712	0,210	0,001	cs310	-0,115	0,085	0,174
tpart1	-0,086	0,094	0,364	ancm1	0,086	0,228	0,706	cs311	0,278	0,231	0,228
tpart2	-0,448	0,139	0,001	ancm2	0,620	0,338	0,067	cs312	0,665	0,268	0,013
a10	-0,181	0,018	<,0001	etcl0	0,053	0,018	0,003	cs330	-0,090	0,028	0,001
a11	-0,205	0,054	0,000	etcl1	0,033	0,055	0,549	cs331	-0,241	0,079	0,002
a12	-0,219	0,091	0,017	etcl2	0,120	0,091	0,185	cs332	-0,317	0,123	0,010
a30	0,086	0,014	<,0001	dip10	0,305	0,022	<,0001	cs340	-0,128	0,031	<,0001
a31	0,068	0,039	0,081	dip11	0,334	0,065	<,0001	cs341	-0,218	0,101	0,031
a32	0,060	0,065	0,355	dip12	0,719	0,097	<,0001	cs342	-0,234	0,145	0,107
a40	0,140	0,012	<,0001	dip30	0,142	0,021	<,0001	cs350	-0,201	0,051	<,0001
a41	0,157	0,034	<,0001	dip31	0,083	0,063	0,186	cs351	-0,341	0,114	0,003
a42	0,078	0,056	0,170	dip32	0,405	0,091	<,0001	cs352	-0,126	0,220	0,568
anc10	-0,022	0,018	0,231	dip40	0,075	0,021	0,000	cs380	-0,031	0,021	0,136
anc11	-0,059	0,055	0,285	dip41	0,009	0,062	0,886	cs381	-0,063	0,055	0,252
anc12	0,121	0,092	0,190	dip42	0,199	0,086	0,021	cs382	-0,284	0,081	0,001

var	coef	Student	p-value	var	coef	Student	p-value	var	coef	Student	p-value
cs420	-0,321	0,032	<,0001	fonc22	-0,133	0,075	0,075	EB1	0,051	0,114	0,652
cs421	-0,413	0,103	<,0001	fonc40	-0,064	0,040	0,109	EB2	0,062	0,152	0,682
cs422	-0,288	0,192	0,133	fonc41	-0,050	0,089	0,576	EC0	0,020	0,028	0,470
cs430	-0,380	0,032	<,0001	fonc42	0,230	0,174	0,186	EC1	0,023	0,086	0,786
cs431	-0,367	0,133	0,006	fonc50	-0,060	0,058	0,300	EC2	-0,140	0,128	0,273
cs432	-0,431	0,183	0,018	fonc51	0,100	0,150	0,505	ED0	0,020	0,032	0,539
cs450	-0,284	0,030	<,0001	fonc52	0,305	0,331	0,356	ED1	-0,025	0,104	0,811
cs451	-0,397	0,091	<,0001	fonc60	-0,015	0,022	0,474	ED2	0,055	0,107	0,607
cs452	-0,188	0,127	0,139	fonc61	-0,050	0,062	0,419	EE0	-0,053	0,021	0,013
cs460	-0,281	0,020	<,0001	fonc62	0,108	0,096	0,261	EE1	0,019	0,061	0,750
cs461	-0,408	0,057	<,0001	fonc70	0,006	0,022	0,790	EE2	-0,122	0,078	0,120
cs462	-0,311	0,081	0,000	fonc71	0,031	0,067	0,646	EG0	0,023	0,030	0,440
cs470	-0,346	0,022	<,0001	fonc72	0,025	0,084	0,763	EG1	-0,002	0,081	0,984
cs471	-0,383	0,063	<,0001	fonc80	0,010	0,018	0,585	EG2	-0,071	0,123	0,566
cs472	-0,359	0,091	<,0001	fonc81	-0,109	0,054	0,042	EH0	-0,048	0,026	0,060
cs480	-0,309	0,024	<,0001	fonc82	-0,066	0,078	0,395	EH1	-0,164	0,070	0,020
cs481	-0,375	0,068	<,0001	fonc90	0,197	0,027	<,0001	EH2	0,033	0,110	0,764
cs482	-0,264	0,095	0,005	fonc91	0,115	0,075	0,125	EJ0	-0,054	0,022	0,014
fonc00	0,022	0,022	0,318	fonc92	0,366	0,114	0,001	EJ1	-0,132	0,060	0,027
fonc01	-0,038	0,066	0,562	fonc0	-0,173	0,107	0,105	EJ2	-0,184	0,091	0,043
fonc02	0,005	0,098	0,959	fonc1	-0,105	0,313	0,738	EK0	-0,044	0,029	0,124
fonc20	-0,036	0,019	0,062	fonc2	0,736	0,272	0,007	EK1	-0,067	0,071	0,345
fonc21	-0,131	0,060	0,029	EB0	-0,005	0,036	0,890	EK2	-0,286	0,119	0,017

var	coef	Student	p-value	var	coef	Student	p-value	var	coef	Student	p-value
EL0	0,001	0,027	0,976	tent10	-0,082	0,014	<,0001	tsam0	-0,014	0,012	0,220
EL1	-0,109	0,071	0,125	tent11	0,002	0,042	0,959	tsam1	-0,086	0,035	0,013
EL2	-0,232	0,112	0,039	tent12	-0,044	0,066	0,510	tsam2	-0,101	0,050	0,044
EM0	-0,042	0,046	0,358	tent20	-0,045	0,015	0,003	tdim0	0,043	0,014	0,002
EM1	-0,001	0,161	0,994	tent21	0,031	0,045	0,498	tdim1	0,129	0,039	0,001
EM2	-0,680	0,210	0,001	tent30	-0,018	0,014	0,211	tdim2	0,093	0,057	0,101
EN0	-0,071	0,020	0,000	tent22	0,032	0,065	0,624	idf0	0,110	0,012	<,0001
EN1	-0,009	0,053	0,872	tent31	0,021	0,041	0,603	idf1	0,058	0,032	0,071
EN2	-0,317	0,084	0,000	tent32	0,085	0,063	0,175	idf2	0,135	0,044	0,002
EP0	-0,224	0,036	<,0001	tentm0	-0,023	0,019	0,225	proxy0	0,048	0,009	<,0001
EP1	-0,134	0,113	0,236	tentm1	0,104	0,055	0,060	proxy1	0,058	0,027	0,029
EP2	-0,151	0,137	0,271	tentm2	-0,229	0,087	0,009	proxy2	0,225	0,040	<,0001
EQ0	-0,109	0,029	0,000	tsoi0	0,043	0,012	0,000	quali10	0,051	0,032	0,119
EQ1	-0,127	0,096	0,187	tsoi1	0,093	0,035	0,008	quali11	0,032	0,053	0,551
EQ2	-0,359	0,128	0,005	tsoi2	0,133	0,048	0,005	quali12	0,051	0,051	0,318
ER0	-0,110	0,025	<,0001	tnui0	0,039	0,015	0,009	sirmiss0	0,022	0,021	0,284
ER1	-0,105	0,080	0,188	tnui1	-0,026	0,043	0,543	sirmiss1	0,053	0,056	0,343
ER2	-0,295	0,108	0,006	tnui2	-0,017	0,065	0,792	sirmiss2	0,382	0,093	<,0001

## Annexe 2 : Modèle de sélection 2

---

- La sélection correspond à la non-réponse en clair.

probit											
variable	coef	Chi2	p-value	variable	coef	Chi2	p-value	variable	coef	Chi2	p-value
Intercept	-0,52	0,23	0,02	cs53	-0,11	0,11	0,31	f6	-0,11	0,12	0,36
Inheur	0,06	0,06	0,32	cs54	0,18	0,07	0,01	g1	-0,21	0,22	0,33
tpart	0,06	0,04	0,18	cs55	-0,05	0,10	0,63	g2	-0,14	0,12	0,26
femme	0,01	0,03	0,82	cs56	-0,03	0,10	0,76	h0	-0,18	0,09	0,04
a1	0,00	0,04	0,98	cs63	0,07	0,06	0,29	j1	-0,07	0,11	0,53
a3	0,10	0,03	0,00	cs64	0,02	0,10	0,84	j2	0,02	0,09	0,87
a4	0,17	0,03	<,0001	cs65	-0,15	0,10	0,14	j3	-0,27	0,10	0,01
anc1	0,06	0,04	0,22	cs67	0,11	0,06	0,08	k0	-0,10	0,09	0,29
anc3	0,03	0,04	0,50	cs68	0,12	0,09	0,20	l0	-0,22	0,09	0,02
anc4	0,07	0,04	0,07	cs69	0,07	0,17	0,67	m0	-0,14	0,12	0,25
ancm	0,54	0,20	0,01	fonc0	-0,02	0,06	0,69	n1	-0,12	0,10	0,24
cdd	0,38	0,23	0,10	fonc2	-0,08	0,05	0,11	n2	-0,14	0,09	0,12
int	0,30	0,40	0,45	fonc3	-0,18	0,07	0,01	n3	-0,18	0,11	0,09
ast	-0,14	0,50	0,78	fonc4	-0,05	0,07	0,46	n4	-0,08	0,15	0,61
etcl	-0,01	0,04	0,80	fonc5	0,01	0,07	0,90	p1	-0,09	0,12	0,42
dip1	0,04	0,06	0,46	fonc6	0,00	0,06	1,00	p2	-0,08	0,13	0,54
dip3	-0,02	0,06	0,75	fonc7	0,04	0,06	0,51	p3	-0,29	0,13	0,03
dip4	0,06	0,05	0,26	fonc8	-0,13	0,06	0,04	q1	-0,06	0,10	0,56
dip5	0,02	0,05	0,72	fonc9	0,02	0,10	0,82	q2	-0,11	0,09	0,21
dip7	0,01	0,05	0,76	foncm	0,53	0,19	0,00	r1	-0,16	0,09	0,08
dipm	0,86	0,58	0,13	a0	-0,20	0,16	0,21	r2	-0,36	0,13	0,01
cs31	0,20	0,22	0,35	b0	-0,31	0,10	0,00	tent1	-0,04	0,03	0,24
cs33	0,15	0,10	0,15	c1	-0,55	0,17	0,00	tent2	-0,06	0,04	0,13
cs34	0,12	0,10	0,23	c2	-0,21	0,13	0,11	tent3	0,00	0,04	0,90
cs35	0,45	0,16	0,01	c3	0,04	0,13	0,77	tentm	0,02	0,05	0,69
cs37	0,32	0,08	<,0001	c4	-0,15	0,12	0,21	tsoi	-0,01	0,03	0,76
cs38	0,31	0,08	0,00	d0	-0,09	0,10	0,38	tnui	0,00	0,04	0,98
cs42	0,06	0,10	0,52	e1	0,06	0,13	0,62	tsam	-0,03	0,03	0,37
cs43	0,01	0,09	0,92	e2	-0,16	0,10	0,10	tdim	0,01	0,04	0,75
cs45	0,07	0,10	0,50	e3	-0,14	0,11	0,23	idf	0,05	0,03	0,07
cs46	0,14	0,07	0,06	f1	-0,19	0,13	0,13	proxy	0,12	0,02	<,0001
cs47	0,17	0,07	0,01	f2	-0,06	0,14	0,69	quali1	-1,01	0,05	<,0001
cs48	0,31	0,07	<,0001	f3	-0,31	0,13	0,01	sirmiss	0,19	0,05	0,00
cs52	0,08	0,07	0,27	f5	0,04	0,09	0,68				



**régression augmentée InWRF ensemble**

var	coef	Student	p-value	var	coef	Student	p-value	var	coef	Student	p-value
Intercept	10,001	0,371	<,0001	cs52	0,075	0,024	0,002	f5	0,004	0,020	0,858
mills	0,437	0,307	0,154	cs53	-0,011	0,033	0,733	f6	-0,075	0,035	0,032
Inheur	0,285	0,017	<,0001	cs54	0,133	0,047	0,004	g1	0,046	0,066	0,485
tpart	-0,204	0,016	<,0001	cs55	-0,077	0,021	0,000	g2	-0,037	0,041	0,370
femme	-0,112	0,005	<,0001	cs56	-0,218	0,017	<,0001	h0	-0,133	0,049	0,006
a1	-0,077	0,006	<,0001	cs63	-0,004	0,019	0,857	j1	-0,098	0,026	0,000
a3	0,088	0,026	0,001	cs64	-0,045	0,017	0,008	j2	-0,043	0,017	0,010
a4	0,142	0,042	0,001	cs65	-0,044	0,040	0,275	j3	-0,255	0,069	0,000
anc1	-0,007	0,016	0,668	cs67	-0,053	0,030	0,079	k0	-0,075	0,029	0,010
anc3	0,062	0,009	<,0001	cs68	-0,068	0,033	0,040	l0	-0,073	0,058	0,207
anc4	0,165	0,018	<,0001	cs69	-0,110	0,034	0,001	m0	-0,118	0,042	0,005
ancm	0,173	0,136	0,203	fonc0	0,011	0,012	0,322	n1	-0,128	0,035	0,000
cdd	0,074	0,105	0,484	fonc2	-0,022	0,021	0,312	n2	-0,116	0,040	0,004
int	0,197	0,110	0,074	fonc3	-0,131	0,048	0,006	n3	-0,197	0,049	<,0001
ast	-0,186	0,080	0,020	fonc4	-0,032	0,018	0,076	n4	-0,133	0,034	<,0001
etcl	0,064	0,008	<,0001	fonc5	0,008	0,012	0,523	p1	-0,181	0,031	<,0001
dip1	0,248	0,016	<,0001	fonc6	0,060	0,010	<,0001	p2	-0,184	0,032	<,0001
dip3	0,126	0,010	<,0001	fonc7	0,042	0,015	0,005	p3	-0,226	0,075	0,003
dip4	0,100	0,017	<,0001	fonc8	-0,002	0,034	0,962	q1	-0,170	0,023	<,0001
dip5	0,019	0,009	0,033	fonc9	0,230	0,019	<,0001	q2	-0,165	0,032	<,0001
dip7	-0,053	0,009	<,0001	foncm	0,156	0,134	0,245	r1	-0,159	0,043	0,000
dipm	0,275	0,252	0,276	a0	-0,175	0,058	0,003	r2	-0,245	0,093	0,008
cs31	0,458	0,065	<,0001	b0	-0,188	0,081	0,020	tent1	-0,109	0,012	<,0001
cs33	0,478	0,042	<,0001	c1	-0,384	0,143	0,008	tent2	-0,085	0,016	<,0001
cs34	0,550	0,035	<,0001	c2	-0,060	0,056	0,285	tent3	-0,039	0,006	<,0001
cs35	0,478	0,114	<,0001	c3	0,095	0,027	0,000	tentm	-0,022	0,009	0,021
cs37	0,608	0,080	<,0001	c4	-0,153	0,042	0,000	tsoi	0,029	0,006	<,0001
cs38	0,607	0,079	<,0001	d0	-0,029	0,029	0,323	tnui	0,040	0,007	<,0001
cs42	0,289	0,024	<,0001	e1	0,035	0,029	0,226	tsam	-0,005	0,008	0,571
cs43	0,241	0,015	<,0001	e2	-0,094	0,044	0,030	tdim	0,051	0,007	<,0001
cs45	0,234	0,023	<,0001	e3	-0,106	0,040	0,008	idf	0,126	0,014	<,0001
cs46	0,249	0,038	<,0001	f1	-0,123	0,053	0,019	proxy	0,065	0,030	0,033
cs47	0,211	0,044	<,0001	f2	-0,178	0,029	<,0001	quali1	-0,316	0,234	0,178
cs48	0,312	0,078	<,0001	f3	-0,124	0,082	0,130	sirmiss	0,078	0,048	0,104

# Annexe 3 : Maximum de vraisemblance

- Vraisemblance paramétrique (celle qui est usuelle)
  - Maximisation de la vraisemblance par technique numérique standard  
ex. Newton-Raphson ...
  - Algorithme EM – particulièrement adapté et simple

Ex : la densité de la loi jointe  $f_\theta(y, z)$  est simple, Y tjrs observée, Z jamais.

comme  $f_\theta(y, z) = \left( \int f_\theta(y, z) dz \right) f_\theta(z|Y = y)$ ,

on a  $\log \left( \int f_\theta(y, z) dz \right) = \log(f_\theta(y, z)) - \log(f_\theta(z|Y = y))$

en multipliant les deux membres par  $f_\theta(z|Y = y)$ , en intégrant par rapport à z, en appliquant l'inégalité de Jensen à  $x \mapsto x \log(x)$

augmenter  $E_{\theta'}[\log(f_\theta(y, z))|Y = y]$ , augmente la vraisemblance.

Rq : comme  $f_\theta(y, z) \mathbb{1}_{y \in T} = \left( \iint_{T \times R} f_\theta(y, z) dy dz \right) f_\theta(z|Y \in T)$ , T est une tranche, augmenter  $E_{\theta'}[\log(f_\theta(y, z))|Y \in T]$ , augmente la vraisemblance.

■ EM :

1. Initialisation des paramètres

---

2. Succession de n cycles:

- Espérance (éventuellement conditionnelle à une information partielle ou d'observations de co-variables) pour la valeur courante du paramètre
- Maximisation et mise à jour des paramètres

Convergence si régularité vers pt stationnaire.

■ Cas des tranches ou fourchettes  $\Rightarrow$  MCEM, étape E = moyenne de simulations par acceptation-rejet.

■ Si multivarié éventuellement ECM,  
 $\Rightarrow$  on fait les maximisations une à une.

■ Vraisemblance empirique: (Owen 88')

non paramétrique ou semi paramétrique,  
inférence modèle en théorie des sondages

ex :  $Z = (Y, X)$  , on s'intéresse à une quantité sur la loi marginale de Y  
(ex. sa moyenne)

$p_i$  = probabilité que Z soit dans un rectangle élémentaire

log-vraisemblance :  $\sum_{i \in S} \log(p_i)$ , si SAS (Chen et Qin 93')

+ les contraintes  $\sum_{i \in S} p_i = 1$  que loi de proba

+ contrainte de « calage » ou de moments si on connaît par ailleurs

la moyenne de X :  $\sum_{i \in S} p_i (x_i - \bar{X}_N) = 0$

+ éventuellement d'autres contraintes

$\Rightarrow$  extrema liés

$$\hat{p}_i = \frac{1}{n \left[ 1 + \lambda (x_i - \bar{X}_N) \right]} \text{ où } \lambda \text{ est solution de } \sum_{i \in S} \frac{x_i - \bar{X}_N}{n \left[ 1 + \lambda (x_i - \bar{X}_N) \right]} = 0.$$

puis & Newton-Raphson...

---

un estimateur d'une moyenne :  $\hat{Y} = \sum_{i \in s} \hat{p}_i y_i$

Spécificités du plan de sondage (Chen et Sitter 99')

⇒ pseudo-vraisemblance empirique

$$\sum_{i \in s} w_i \log (p_i)$$

avec la contrainte de calage :  $\sum_{i \in s} p_i (x_i - \bar{X}_N) = 0$

rq 1: on peut faire donner des intervalles de confiance et faire des tests

rq 2: distance de Kullback remplacé par chi2 ⇒ formules fermées

rq 3: si MAR propriété de la vraisemblance empirique sous imputation par noyaux (Wang et Rao 02')

rq 4: cas de la sélection non-ignorable : Qin, Leung et Shao 02'