Modélisation de la non réponse dans les enquêtes statistiques Cas du secteur Bâtiment, Travaux Publics et Hydraulique (BTPH)

R. Toumache et T. Akrouf. (*)

Introduction

La modélisation d'un phénomène qui prend des valeurs spécifiques demande l'utilisation des modèles spécifiques correspondants. Ainsi, la modélisation du phénomène « réponse – non réponse »dans les enquêtes statistiques fait appel à l'économétrie des variables qualitatives.

Les modèles de réponses qualitatives prennent en compte la nature discrète des variables qui dans le cas le plus simple prennent conventionnellement la valeur codée 0 ou 1. Cette forme de réponse (binaire), qualifie ceux-ci de modèles à réponses binaires (binary reponse models). Parmi ceux-ci, nous rencontrons deux types de modèles qui sont fréquemment utilisés le modèle logit et le modèle Probit.

La simplicité des formules et la facilité des calculs, conjuguées avec l'utilisation des algorithmes numériques pour ces derniers, font que le recours à ces modèles soit conseillé. La méthode d'estimation utilisée pour estimer les paramètres du modèle est la méthode du maximum de vraisemblance. La régression linéaire étant inappropriée dans ce cas.

(*) R.Toumache Maître assistant à l'INPS ;
T.Akrouf, Maître de conférences à l'INPS.

Le phénomène de non-réponse est lié à un certain nombre de variables exogènes connues à priori par expérience et/ou prédéterminés par des méthodes d'analyse des données. Dans ce qui suit nous allons présenter le modèle logit et les tests adéquats, puis nous allons appliquer ce modèle aux données de l'enquête sur le BTPH.

1– Présentation du modèle binaire logit[1] :

1.1- Définition et principe des modèles à variables qualitatives binaire[1] :

Le plus souvent la variable à expliquer est dichotomique (à deux modalités possible seulement) et les variables explicatives sont nominales (qualitatives. Les n individus caractérisés par l'ensemble des p variables sont partitionnés en deux groupes définis par les modalités de la variable Y.

La solution consiste à considérer la réalisation de la variable dépendante binaire comme provenant d'une certaine règle de décision. Cette règle est un mécanisme associant les variables explicatives x_i à l'observation de l'évènement $\{y_i = 0\}$ou $\{y_i = 1\}$. L'intuition est la suivante : supposons que la réalisation de $\{y_i = 1\}$ est plutôt associée à des valeurs élevées des x_i, celle de $\{y_i = 0\}$ à des valeurs faibles des x_i. Il existe alors une certaine valeur seuil dépendant de la combinaison linéaire des x_iB, ou B est un vecteur de paramètres, au-delà de laquelle la proportion des $\{y_i = 1\}$ l'emporte sur celle des $\{y_i = 0\}$.
Notons c cette valeur seuil. Il est également raisonnable de supposer que notre règle de décision n'est pas déterministe, c'est à dire que pour certaines observations, Yi peut être nul alors que les valeurs de x_i sont élevées. Ce caractère non déterministe peut être intégré en ajoutant à notre combinaison linéaire un terme aléatoire noté u_i. Notre règle de décision est alors :

La proportion de $\{y_i = 1\}$ est « élevé » pour $x_iB+u_i > c$, et « faible » pour $x_iB+u_i \leq c$

La règle de décision probabiliste devient alors :

$$\textbf{Prob } (y_i = 1) = \textbf{prob } (x_iB+u_i > c) = \textbf{1- prob } (u_i < c - x_iB)$$

$$\textbf{Prob } (y_i = 0) = \textbf{prob } (x_iB+u_i \leq c) = \textbf{prob } (u_i \leq c - x_iB)$$

[1] Thomas ALBAN, Econométrie des variables quantitatives, Paris, Ed DUNOD 2000 ; pages 51 à 78.

Ainsi le modèle ne détermine pas exactement la réalisation de l'événement (y_i =1) ou (y_i =0),mais fournit plutôt une mesure théorique de la proportion d'observation pour lesquelles cet évènement s'est réalisé

Ensuite, puisque cette mesure théorique (la probabilité) est croissante dans son argument, la probabilité que (y_i =1) sera croissante pour les composantes de x_i dont les paramètres associés sont positifs, et décroissante pour celles dont les paramètres sont négatifs. On note également que la valeur seuil **c** dans ce modèle est identique pour toutes les observations. On peut alors fixer arbitrairement la valeur seuil à 0, ainsi que l'espérance de u_i.

Enfin, le plus important est de bien noter que l'écriture probabiliste de notre règle de décision dépend exclusivement de la distribution statistique de la seule variable aléatoire du système, c'est-à-dire de u_i. Par conséquent, dès que l'on impose une loi particulière à ce terme aléatoire, les probabilités de {y_i= 1} et {y_i = 0} pourront être calculées en faisant référence à cette loi.

Les deux lois statistiques les plus couramment utilisées dans la pratique sont la loi logistique et la loi de Laplace Gauss (distribution normale).

De l'utilisation de ces deux distributions découle alors les modèles qualitatifs binaires appelés respectivement logit et probit.

Notons **F(.)** la fonction de répartition issue de la distribution statistique du terme d'erreur u_i, et **f** la fonction de densité associée. Comme la valeur seuil peut être normalisée à 0, le modèle s'écrit de façon générale :

- **Prob (y_i = 1)= prob (u_i >- x_iB) = 1-F(-x_iB)**
- **Prob (y_i = 0)= prob (u_i ≤- x_iB) = F(-x_iB)**

L'hypothèse de symétrie de la densité de u_i autour de 0 fait que:

$$f(x_iB) = f(-x_iB)$$

et donc :

$$F(x_iB) = 1-F(x_iB)$$

Ce qui permet de considérer les valeurs observées de y_i comme les réalisations d'un processus binomial avec les probabilités F(x_i B) pour (y_i = 1), et (1- F(x_i B)) pour (y_i = 0) .A la différence des processus binomiaux usuels, les probabilités varient à chaque « expérience », dans la mesure ou elles dépendent de x_i.

On aura donc :

Prob (y_i = 1) = 1-F (- x_i B) = F (x_i B) et Prob (y_i = 0) = F (- x_i B) = 1- F (x_i B).

Pour l'analyse nous utiliserons le modèle logit parce que la taille de notre échantillon est grande, et que ce modèle donne des résultats plus significatifs.

1-2 Le modèle logistique(logit)

- **Définition :**

La régression logistique, comme l'analyse discriminante, cherche à décrire la liaison entre une variable nominale Y(variable à expliquer) et un ensemble de p variables (X_1, X_2,..., X_p). On veut également connaître l'effet d'une variable sur la variable à expliquer en tenant compte des liaisons qu'elle entretient avec les autres variables du modèle.
On fait ici l'hypothèse que l'erreur ui est une variable suivant une loi logistique qui admet comme fonction de répartition et de densité les expressions suivantes :

$$F(x_i B) = \frac{\exp(xiB)}{1+\exp(xiB)} \qquad f(xi\ B) = \frac{\exp(xiB)}{[1+\exp(xiB)]^2}$$

Remarquons que la probabilité associée à la loi logistique peut être inversée.
Si l'on note pi la probabilité que (y_i =1), on a alors la représentation suivante :

$$p_i = \text{prob}\ (y_i = 1) = F(x_i B) = \frac{\exp(xiB)}{1+\exp(xiB)}$$

D'où :

$$\frac{Pi}{1 - Pi} = \exp(xiB)$$

Ou encore:

$$\text{Log}\ \frac{Pi}{1 - Pi} = x_i B$$

- **Estimation et tests des coefficients :**

Pour estimer les coefficients B du modèle, on utilise la méthode du maximum de vraisemblance.

Les **n** observations (y_i, x_i) [ou $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$] sont indépendantes et les y_i sont des variables de Bernoulli.

La vraisemblance L (y, x, B) pour l'ensemble des observations :

$$L(y,x,B) = \prod_{i=1}^{n} L(y_i, x_i, B) = \prod_{i=1}^{n} F(x_i, B)^{yi} [1 - F(x_i, B)]^{1-yi}$$

La procédure d'estimation revient à chercher la valeur de B qui maximise le logarithme de la vraisemblance :

$$\text{Log } L = \sum_{i=1}^{n} yixiB - \sum_{i=1}^{n} \log[1 + \exp(xiB)]$$

Pour apprécier l'éventuelle non-influence d'une variable ou d'une modalité x_i sur la variable Y, on teste l'hypothèse nulle H_0 :

$$(H_0) : B_i = 0 \qquad (H1) : Bi \neq 0$$

On considère alors la statistique de student :

$$t = \frac{\hat{\beta}_i}{\sqrt{\text{var}(\hat{\beta}_i)}} \rightarrow St_{(n\ -\ q)}$$ Où $\hat{\beta}_i$ est la i_{eme} composante de l'estimateur $\hat{\beta}$ de B et Var ($\hat{\beta}_i$) est la variance estimée associée à cette composante.

Pour tester l'influence d'une variable nominale à q modalités, on procède à un test de nullité des q coefficients B_j affectés à ces modalités.

D'une manière générale, l'hypothèse H_0 stipulant une éventuelle non-influence d'un ensemble de p variables (X_1, X_2, ..., X_p) sur Y, s'exprime par la nullité des q coefficients associés :

$$H_0 : B_1=B_2= \ldots = B_q = 0$$

Notons B l'estimateur des Bi sous l'hypothèse H_0 et B0 l'estimateur des coefficients du modèle alternatif.
On teste l'hypothèse nulle en calculant la statistique du rapport de vraissemblance :

$$\mathbf{LRTc} = 2\left[L(\beta) - L(\beta o)\right] \rightarrow \chi_q^{\,2}$$

On démontre qu'elle suit une distribution du khi-deux à q degrés de liberté (tel que q désigne le nombre de modalités de la variable testée), sous des hypothèses de travail convenables. Si l'hypothèse nulle est rejetée, on en déduit qu'au moins une des p variables (ou une modalité de la variable nominale) influe sur la variable Y.

2 - Application du modèle logit aux données de l'enquête menée dans le secteur BTPH et estimation des taux de non-réponse

L'application du modèle logit aux données de l'enquête menée dans le secteur BTPH nous permet de tester la signification des variables explicatives du phénomène de non-réponse. Le coefficient d'une variable sera retenu comme significativement positif au seuil de 5% lorsque la probabilité qu'il soit supérieur au fractile de la loi de khi-deux, soit inférieur à 5%. Par conséquent, c'est cette « probabilité » qui détermine le choix des variables explicatives.

Les données de l'enquête ne sont pas très fiables,donc les résultats indiqueront des tendances que l'on corrigera plus tard (dans une autre enquête sur le BTPH).En effet, le nombre des répondants dans le fichier utilisé est sous estimé, alors que le nombre de non –répondants est surestimé. La raison de ces écarts est due au fait que le fichier sur lequel figure l'ensemble de l'échantillon n'est pas complètement à jour. Par conséquent lorsque le fichier de l'échantillon a été apparié avec le fichier des répondants, afin de créer la variable (OBS) (répondants/non-répondants), 681 répondants n'ont pas apparié correctement. La

mise à jour des informations relatives aux 681 répondants aurait été trop lourde, alors ces 681 questionnaires ont été redressés.

2-1 Présentation des données

Les données utilisées sont issues de l'enquête sur le BTPH mené en 1997par le CECP pour le compte de l'ONS

Ces données sont enregistrées sur un tableau rectangulaire, ses lignes sont de 2202 entreprises et représentent la taille de l'échantillon. Ses colonnes contiennent les variables ou les caractéristiques de chaque entreprise (l'unité statistique) qui sont :

- **La wilaya** : le lieu de localisation de l'entreprise (48 wilayas)
- **NAPR** : Nomenclature des activités et des produits résumés (NAPR41, NAPR42, NAPR43)voir annexe
- **La région** : dans quelle région se situe l'entreprise ?
  - Région 1 : Centre
  - Région 2 : Sud
  - Région 3 : Est
  - Région 4 : Ouest
- **Catégorie** de l'entreprise en fonction du secteur juridique et de l'effectif de salariés (voir le point 2-2-2)
- **La fiche A** : fiche d'identification de l'entreprise.
- **La fiche B** : le questionnaire (B) concernant l'emploi, Salaires et heures travaillées.
- **La fiche c** : la feuille comptable.
- **Observation** : représente l'état d'activité de l'entreprise au moment de l'enquête, elle peut être active, ou dissoute.

Le tableau suivant représente un exemple du tableau global des données de l'enquête.

Tableau 1. Exemple du type des données de l'enquête sur le BTPH réalisée par le CECP en 1997 :

Entreprise	Wil	CAT	NAP	NAPR	Région	A	B	C	Observation
N°25	1	4	332	41	2	A	B	C	Active
N°60	5	1	058	42	3	A	B	C	Active
N°210	16	3	346	43	1	-	-	-	Active

CECP : centre d'étude de la concurrence et des prix

2-2-Présentation des variables

L'analyse par tableaux croisés est une méthode traditionnelle pour traiter les enquêtes. Si on considère les grandes enquêtes sociologiques, ou autres, où de nombreuses variables interagissent entre elles, il serait nécessaire de faire appel à des méthodes statistiques appropriées, ces pourquoi nous avons opté pour l'application du modèle logit .

Les variables utilisées dans notre travail sont des variables qualitatives, ce sont des variables qui présentent deux valeurs, on dira aussi, qu'elles ont deux modalités, souvent notées 0 et 1.

Dans ce qui suit, la variable dichotomique est « réponse et non réponse ». (notée OBS),et représente la variable endogène

2-2-1-La variable endogène

On considère comme variable endogène dichotomique, qualitative la variable réponse (**OBS**) qui se présente comme suit :

OBS=
- 1 Si l'entreprise (i) est active et A,B,C sont présents
- 0 Si l'entreprise (i) est active, dissoute et A ou B, ou, C sont absents

Tout autre cas est considéré comme unité hors champ.

Sous ces hypothèses, nous avons les données de l'enquête, résumées dans le tableau suivant :

Tableau 2 . Proportion des entreprises ayant répondu.

Entreprises	Nombre
Entreprises ayant répondu (OBS=1)	681
Entreprises n'ayant pas répondu (OBS=0)	1511
Totale des entreprises	2192
Le taux de réponse	31%

Source : Fichier de suivi de l'enquête, exercice 97, CECP

2-2-2-Les variables exogènes

1) CAT : catégorie de l'entreprise, peut prendre.

CAT1= { 1 si l'entreprise (i) est une entreprise publique nationale
0 si non

CAT2 = { 1 si l'entreprise (i) est une entreprise locale nationale
0 si non

CAT3 = { 1 si l'entreprise (i) est une entreprise privée dont l'effectif >20
0 si non

CAT4 = { 1 si l'entreprise (i) est une entreprise privée d'effectif]10,20]
0 si non

CAT5 = { 1 si l'entreprise (i) est privée d'effectif]5,10]
0 si non

$$\textbf{CAT6} = \begin{cases} 1 \text{ si l'entreprise (i) est privée d'effectif} \leq 5 \\ 0 \text{ si non} \end{cases}$$

2) NAPR : nomenclature des activités et des produits résumés :

$$\textbf{NAPR41} = \begin{cases} 1 \text{ si NAP (i)} \in \{321\ 330\ 331\ 332\ 333, 334, 335, 336, 337, 338, 339, 340, 341\} \\ 0 \text{ si non} \end{cases}$$

NAPR41 : construction Bâtiment, pose charpente, couverture, plomberie, serrurerie,... Electricité.

$$\textbf{NAPR42} = \begin{cases} 1 \text{ si NAP (i)} \in \{020, 050, 054, 058, 343\} \\ 0 \text{ si non.} \end{cases}$$

NAPR42 = Travaux publics agricoles

$$\textbf{NAPR43} = \begin{cases} 1 \text{ si NAP (i)} \in \{342, 344, 345, 346, 347, 348, 349, 350, 351, 353\} \\ 0 \text{ si non.} \end{cases}$$

NAPR43 : Autres travaux publics ; hors agricoles et pétroliers

3) la région :

R1 Centre =
- 1 si Wilaya (i) ∈ {44, 16, 06, 17, 02, 09, 10, 35, 26, 42, 15}
- 0 si non

R2 Sud =
- 1 si Wilaya (i) ∈ {01, 08, 07, 39, 47, 33, 03, 30, 30, 11, 37}
- 0 si non

R3 Est =
- 1 si Wilaya (i) ∈ {23, 05, 34, 25, 36, 24, 18, 40, 28, 04, 19}
- 0 si non

R4 ouest =
- 1 si Wilaya (i) ∈ {46, 32, 29, 27, 45, 31, 48,20 14 38 13)
- 0 si non

2-3 Estimation des modèles

L'estimation de la probabilité de réponse s'exprime en fonction de l'estimateur $\hat{\beta}$ du modèle logit

A cette étape de l'estimation nous allons prendre les variables suivantes comme témoins par exemple **NAPR41, CAT5, région1** pour que la matrice $\left[\frac{\delta^2 \log L}{\delta\beta\delta\beta'}\right]$ soit inversible.

Nous avons donc effectué une régression logistique de la réponse (OBS), en prenant pour variables explicatives, la catégorie, la région, la nomenclature des activités et des produits résumés.

L'ajustement du modèle par la méthode du maximum de vraissemblance, après trois itérations a donné le modèle prédictif suivant :

Tableau 3 Estimation des coefficients par le modèle logit

Variable	Coefficients	Student	Probabilité	Std error
Constante	-0,708796	-3,893797	0,0001	0,182032
CAT1	1,098008	4,538418	0,000	0,241936
CAT2	0,567580	3,045675	0,0023	0,185765
CAT3	-0,172564	-0,897412	0,3696	0,192290
CAT4	0,054276	0,238120	0,8118	0,227937
CAT6	0,063823	0,297068	0,7664	0,214843
N2	0,364058	1,809493	0,0705	0,201193
N3	0,132766	1,209069	0,2268	0,109809
R2	-0,495144	-3,015045	0,0026	0,164224
R3	-0,731874	-6,241455	0,0000	0,117260
R4	-0,358649	-2,7767715	0,0055	0,129163
Loglikelihood	-1301,177			
OBS =1	681			
OBS =0	1511			

Source : l'estimation a été faite par le logiciel de statistique EVIEWS

Le tableau ci-dessus présente les résultats de l'ajustement de la variable binaire réponse et non-réponse par le modèle logit, les données dont on dispose sont au nombre de 2192 avec 681 « un » pour indiquer la présence de réponse et 1511 « zéros » pour indiquer l'absence de réponse (pas de réponse, ou les cas de dissolution)

Afin de savoir si les trois variables (catégories, nomenclature des activités et produits résumés, la région, (**CAT1, CAT2, CAT3, CAT4, CAT6, R2, R3, R4, N2, N3**) ont un pouvoir explicatif sur la présence d'une réponse, nous avons ajusté et estimé différents modèles logistiques ne tenant pas compte de régresseurs, catégorie, nomenclature des activités et produits résumés et la région. Les variables catégorie et nomenclature

des activités et produit résumés n'expliquent pas de manière significative la réponse, selon la colonne de la probabilité qui nous permet de vérifier leur degré de signification.

Avec un loglikelihood (-1301, 117). Ce qui indique a priori une bonne adéquation des données au modèle logistique.

Pour confirmer cette remarque, nous avons testé la signification des variables l' une après l'autre par « **like lihood ratio test** » (LRT).

2-3-1 Test des variables

Pour tester la signification des variables, nous allons calculer **LRTc** (like lihood ratio test calculé) en utilisant la formule suivante :

$$\mathbf{LRTc} = 2\left[L(\beta) - L(\beta o)\right] \rightarrow \chi_q^{\,2}$$

et nous comparons cette valeur avec celle de LRTt (de la table) au seuil de signification 5% à partir de la table de Khi-deux tel que **q** : le nombre de contraintes

2-3-1-1. Tableau 4.Test (like lihood ratio test) de la variable catégorie

Variable	L(B)	L(Bo)	LRtc	LRt_t	Q	Décision
CAt1 CAt2 CAt3 CAt4 CAt6	-1301,117	-1334,415	66,59	11,07	5	LRtc>LRt_t Alors la variable catégorie est significative

Ce tableau montre que la variable catégorie est significativement différente de Zéro car au seuil $\alpha = 0{,}05$ le **LRtc** calculé est égale à **66,59** qui est supérieur à **LRt_t** tabulé **(11,07).**

2-3-1-2 Test (like lihood ratio test) de la variable « région »

Tableau 5. Test de la variable région

Variable	L(B)	L(Bo)	LRtc	LRt_t	Q	Décision
R2 R3 R4	-1301,117	-1321,189	40,544	7,81	3	LRtc>LRt_t Alors la variable région est significative

Ce tableau montre que la variable région est significativement différente de zéro car au seuil α =5% le **LRtc** calculé est égale à **40,544** qui est supérieur à **LRt_t** tabulé **(7,81).**

2-3-1-3 Test (likelihood ratio test) de la variable « NAPR » (nomenclature des activités et des produits résumés) :

Tableau 6 :Test de la variable NAPR

Variable	L(B)	L(Bo)	LRtc	LRt_t	Q	Décision
N2 N3	-1301,117	1303,18	4,126	5,991	2	LRtc<LRt_t Alors la variable NAPR n'est pas significative

Le tableau ci-dessus montre que la variable NAPR est faible et non significativement différente de zéro puisque au risque α =0,05 le LRtc calculé est égal à **4,126** qui est inférieur à **LRtt** tabulé **(5,991).**

Il est souvent commode de résumer les ajustements des modèles emboîtés dans un tableau d'analyse de la déviance, analogue au tableau d'analyse de la variance, en effectuant une régression linéaire multiple.

Le tableau suivant donne l'analyse de la variance associée à l'ajustement de la variable réponse en fonction des régresseurs **catégorie, NAPR** et **Région.**

Tableau 7. Tableau d'analyse de la déviance.

Modèle	Constante	Catégories	Région	NAPR	Loglikelihood	Test LRt
1	-1,8120759	-1,1970734	-0,1567105	0,04709065	-1319,3708	
2	0,1118242	-0,1730623	-0,1507028	/	-1333,4730	28,1662
3	-2,1569595	0,0664674	/	0,0437848	-1326,9539	15,1662
4	-3,1569595		-0,165369	0,0664674	-1333,5803	28,419

Le tableau ci-dessus donne les résultats de l'ajustement de la variable binaire réponse et non réponse par un modèle logistique, en utilisant les régresseurs deux à deux. La colonne intitulée test **LRt** contient des différences entre chaque modèle et le modèle complet, qui prend en considération les trois variables simultanément

En regardant les coefficients des variables ajustées et leurs écart types dans le tableau 3 on s'aperçoit que c'est la modélisation de la fonction logit comme fonction de la catégorie et de la région qui rend ces variables significatives dans la régression. Une explication probable de ceci est la forte corrélation existante entre les variables **NAPR** et **catégorie**, qui masque probablement l'effet de la variable **NAPR** ou **catégorie**.

Pour cela nous considérons comme meilleur modèle qui explique le phénomène de non-réponse le modèle donné dans le tableau 8,qui prend en considération la catégorie et la région.

Tableau 8. Meilleur modèle d'estimation

Variable	Coefficient	Std ERROR	T Student	Probabilité
Constante	-0,670309	0,180173	-3,720364	0,0002
CAt1	1,112061	0,241534	4,604164	0,0000
CAt2	0,587273	0,185402	3,1677567	0,0016
CAt3	-0,166324	0,191990	-0,866318	0,3864
CAt4	0,061424	0,227588	0,269890	0,7873
CAt6	0,067998	0,214573	0,316901	0,7513
R2	-0,487868	0,163644	-2,981269	0,0029
R3	-0,720782	0,116901	-6,165766	0,0000
R4	-0,351220	0,128935	-2,724017	0,0065
Loglikelihood	-1303,18			
OBS DEP=1	681			
OBS DEP=0	1511			

Un accroissement d'une variable explicative se traduit par une augmentation de la probabilité de réponse (voir tableaux n° 8 et 9). Si le coefficient de cette variable est positif comme **CAt1, CAt2, CAt4, CAt6** et une diminution de la probabilité si le coefficient est négatif comme les variables **CAt3, R2 ,R3, R4**

Pour avoir des résultats précis, nous avons calculé les différents cas possibles de probabilité. Nous avons **4x6=24** cas possibles d'où le tableau suivant :

Tableau 9 : **Probabilité de réponse par catégorie et région**

Variable	Catégories	Pij
Région 1 : Centre	Cat1	0,608676
	Cat2	0,479253
	Cat3	0,302244
	Cat4	0,352314
	Cat5	0,338428
	Cat6	0,353815
Région 2 : Sud	Cat1	0,488473
	Cat2	0,361028
	Cat3	0,210070
	Cat4	0,250349
	Cat5	0,238999
	Cat6	0,251585
Région 3 : Est	Cat1	0,430692
	Cat2	0,309209
	Cat3	0,174018
	Cat4	0,209214
	Cat5	0,199234
	Cat6	0,210304
Région 4 : Ouest	Cat1	0,522617
	Cat2	0,393110
	Cat3	0,233643
	Cat4	0,276857
	Cat5	0,264730
	Cat6	0,278175

Source : l'estimation par logiciel TSP7

CAT1 est une entreprise publique nationale ;
CAT2 est une entreprise locale nationale ;
CAT3 est une entreprise privée dont l'effectif >20 ;
CAT4 est une entreprise privée d'effectif]10,20] ;
CAT5 est une entreprise privée d'effectif]5,10[;
CAT6 est une entreprise privée d'effectif ≤5.

Les résultats récapitulés dans le tableau précédent, nous permettent de faire les remarques suivantes :

- dans toutes les régions les catégories, 1,2 donnent une grande probabilité de réponse.
- dans toutes les régions les catégories 3 et 5 donnent une petite probabilité de réponse.
- La plus grande probabilité se retrouve dans la région 1 (centre) et pour la catégorie 1 (entreprises publiques nationale).
- La plus petite probabilité se retrouve dans la région 3 (Est) et pour la catégorie 3 (entreprises privées dans l'effectif est supérieur à 20 employés).

Conclusion

Nous avons essayé à travers ce travail de proposer un modèle (logit) pour calculer les probabilités de réponses. La réalisation des tests de vraisemblances nous permettent de déterminer des variables qui expliquent le phénomène de non-réponses L'application du modèle logit et des tests de vraisemblance aux données de l'enquête sur le BTPH menée par le CECP en 1997 nous ont permis de déterminer un certain nombre de variables explicatives du phénomène de non-réponses et calculer les probabilités de réponses dans cette enquête.

L'estimation par le modèle (logit) nous a permis de dégager les résultats suivants :

L'analyse séparée des estimateurs dégagés par le maximum de vraisemblance relève que le phénomène de non-réponse est expliqué par les variables ; **catégorie de l'entreprise et la région de sa localisation**

L'analyse par le modèle logit nous a permis aussi de calculer les probabilités de réponse par région et par catégorie d'entreprise.

A la lecture des données fournies par le modèle logit, les deux variables les plus pertinentes pour expliquer la non-réponse sont :la catégorie (catégorie de l'entreprise en 6 modalités) ; La région (localisation de l'entreprise en 4 modalités)

Le modèle logit montre que les entreprises publiques nationales (catégorie 1) ont un taux de réponse plus fort que les autres catégories.

Par ailleurs, la proportion des entreprises privées de plus de 20 employés est importante pour les non-répondants que pour les répondants.

Bibliographie

1-Alain Stuart, OS (ECOM) « **The ideas of sampling** » Charles GRIFFIN and company LTD High Wycombe Monographic, 1984.

2-Alain Vicnot, Denis Boujet "**Traitement de l'information: Statistique et probabilités** » Viebert, Paris, 1995.

3-Anestis Antoniadis. Jaque beruyer, rené Cermana «**Régression non linéaire et applications** », Economica, Paris, 1994.

4-Christian Gourieroux, « **Théorie des sondages** », Economica, paris, 1981.

5-Christian Gourieroux, **Econométrie des variables qualitatives** », Paris, 1972.

6-Christian Chanbaz et nadine Legendre, « **Calcul des pondérations dans le panel Europeen de ménages** », INSEE méthodes N°84,85,86.

7-Ctherine Berthier et François Dupont, « **L'incidence du caractère obligatoire des enquêtes** » ; INSEE méthodes N°69,70,71

8-Cochran.W.G, « **Sampling techniques** » 3ème Edition Willy, 1977.

9-D, Blizeau; J-L Dubois, " **Connaître les conditions de vie des ménages des pays en voie de développement** », Ministère de la coopération et de développement ; France, 1989.

10-D.Grange et Lebart, « **Traitement statistique des enquêtes** », DUNA, Paris, 1978.

11-Didier Continien et daniel Verger, « **Les imputations économètriques** », l'exemple de l'enquête revenues fixaux 1992, INSEE méthodes N°59,60,61.

12-Jeau-Jaques Drosberke, Bernard Fichit, Philip Tassi « **Les sondages** », ECONOMICA, Paris, 1987.

13-IVES Tille, **Théorie des sondages**,Ed .Dunod, Paris 2001

14- Thomas ALBAN, **Econométrie des variables qualitatives** Paris, Ed DUNOD 2000 pages 51-78.

ANNEXE

Liste des nomenclatures des activités et produit s (NAP) données par la NAPR (NAP RESUMEE)	
NAPR41 : Bâtiment, construction, pose charpente, couverture, plomberie, serrurerie, climatisation, peinture, élécticité	
NAP 321	fabrication de bâtiments en bois
NAP 330	Fabrication de bâtiments
NAP 331	Maçonnerie, platerie, travaux en ciment et en béton armé pour le bâtiment, terrassement et démolition de bâtiments
NAP 332	Charpente en bois, menuiserie de bâtiments, pose
NAP 333	Couverture, plomberie, étanchéité et insonorisation
NAP 334	Serrurerie de bâtiments
NAP 335	Installation de climatisation non industrielle
NAP 336	Protection incendie
NAP 337	Fumisterie et ramonage non industriels
NAP 338	Peinture de bâtiments
NAP 339	Décoration et aménagement de locaux divers, installation de rideaux et de stores
NAP 340	Montage de constructions métalliques
NAP 341	Installation d'électricité (y compris pose d'enseignes lumineuses)
NAPR 42 : TRAVAUX PUBLICS AGRICOLES	
NAP 020	Reboisement
NAP 050	Aménagement de périmètres irrigués (non compris drainage agricole de la NAP 054)
NAP 054	Drainage agricole

NAP 058	Défence et restaurartion des sols
NAP 0343	Terassement et travaux ruraux
NAPR 43 : AUTRES TRAVAUX PUBLICS (non compris agricoles et pétroliers)	
NAP 342	Entreprises de travaux publics et souterrains
NAP 344	Travaux maritimes et fluviaux
NAP 345	Travaux de routes et d'aerodromes
NAP 346	Travaux de voies ferrées (y compris l'installation de matériels électrique)
NAP 347	Travaux urbains et travaux d'hygiene publique
NAP 348	Installation de réseaux et de et de eentrales éléctriques et téléphoniques
NAP 349	Pose de canalisation à grande distanee
NAP 350	Fumistric industrielle
NAP 351	Installations thermiques industrielles
NAP 353	Travaux liés à l'exploitation des mines