

## TD ECONOMÉTRIE - M1 ACTUARIAT

### MODÈLES À CHOIX DISCRET

#### **Exercice 1 : Étude sur la possession de biens durables**

Pour le ménage  $i$  de caractéristiques  $X_i$  (âge, revenu, ...), la possession d'un bien durable particulier procure un niveau d'utilité  $U(1, X_i)$  alors que la non-possession procure un niveau  $U(0, X_i)$ . On définit la variable  $Y_i$  telle que

$$Y_i = \begin{cases} 1 & \text{si le ménage } i \text{ possède le bien,} \\ 0 & \text{sinon} \end{cases} \quad (1)$$

et la variable  $Z_i$  par  $Z_i = U(1, X_i) - U(0, X_i)$ .

1. On suppose que le ménage choisit la situation qui lui procure le plus haut niveau d'utilité.
  - (a) Re-spécifier la variable  $Y_i$  en fonction de  $Z_i$ .
  - (b) On observe les valeurs que prend  $Y$  sur un ensemble de ménages ( $i = 1, \dots, n$ ). Le modèle postule une relation de type :  $Z = X\beta + u$ . Calculer la probabilité que le ménage possède le bien (en fonction de la fonction de répartition,  $F$ , des termes d'erreurs  $u$ ) et celle que le ménage ne le possède pas.
  - (c) Une enquête nous donne  $n$  observations indépendantes  $(Y_i, X_i)$ . Écrire la vraisemblance de ce modèle.
2. Application : Dans un article intitulé "L'achat d'un logement ne va pas sans achats d'équipements" [*Economie et Statistique*, n.161, décembre 1983], Daniel Verger étudie, à l'aide de régressions Logit, les taux d'équipement des ménages en biens durables (lave-linge, lave-vaisselle, machine à coudre, chaîne haute fidélité, appareil photo, automobile). Un extrait des résultats de cette étude est reproduit en page suivante.
  - (a) Daniel Verger ne donne pas les coefficients estimés mais indique juste leur significativité à l'aide d'un système de + et de -. Quel est le test statistique qui a permis d'assigner ces symboles ?
  - (b) Dans ce modèle, l'auteur envisage 7 dimensions explicatives (le statut d'occupation, l'âge du chef de ménage, la composition du ménage, le revenu, la profession, le lieu de résidence, le type d'habitat), toutes décrites à l'aide de plusieurs variables dichotomiques. Quel(s) test(s) peut-on envisager pour statuer sur la pertinence ou non d'une de ces dimensions explicatives ?
  - (c) Des statistiques descriptives préliminaires sur ce fichier de ménages révèlent que les ménages propriétaires possèdent en moyenne moins souvent une chaîne de haute fidélité que les ménages locataires (20% contre 27%). Cependant, à la lecture des résultats du logit, il semblerait que les propriétaires sont plus enclins à posséder une chaîne que les locataires. Ces deux résultats sont-ils contradictoires ? Faut-il remettre en cause le modèle ?
  - (d) Interpréter les résultats de la 1ère colonne du tableau.

Extrait de D. VERGER [1983], « L'achat d'un logement ne va pas sans achats d'équipements » [Economie et Statistique, n°161, décembre 1983, p.25].

### Les effets des caractéristiques du ménage sur les taux d'équipement

	Lave linge	Lavo vals- selle	Ma- chine à cou- dre	Plus de 7 gad- gets	Chai- ne haute- fidé- lité	Appa- reil pho- togra- phi- que	Bloc- mo- teur	Auto- mobi- le
<b>Statut d'occupation :</b>								
— locataire								
— accédant récent.....		+	+	+		+	+	+
— accédant ancien.....	+	+	+	+	+	+	+	+
— propriétaire.....	+	+	+	+	+	+	+	+
<b>Chef de ménage :</b>								
— de 30 ans ou moins.....	—	—	—	+	+	+		+
— de 30 à 45 ans.....								
— de 46 à 65 ans.....	+	—	+	—	—	—	—	—
— de 66 à 75 ans.....	—	—	+	—	—	—	—	—
— de plus de 75 ans.....	—	—	+	—	—	—	—	—
<b>Ménage composé :</b>								
— d'un individu.....	—	—	—	—	—	—	—	—
— d'un couple seul.....	—	—	—	—	—	—	+	+
— d'un couple avec un enfant.....								
— d'un couple avec deux enfants.....	+	+	—			+	+	+
— d'un couple avec trois enfants ou plus.....	+	+	+			+	+	+
<b>Marié depuis plus de deux ans</b>								
— marié depuis moins de deux ans.....	—	—					—	
<b>Revenu du ménage :</b>								
— inférieur à 35 000 F.....	—	—	—	—	—	—	—	—
— de 35 000 à 55 000 F.....	—	—	—	—	—	—	—	—
— de 55 000 à 80 000 F.....								
— de 80 000 à 110 000 F.....	+			+	+	+	+	+
— supérieur à 110 000 F.....	+			+	+	+	+	+
<b>Chef de ménage :</b>								
— inactif.....						—	—	—
— agriculteur.....	+				—	—	+	+
— patron.....	+	—			—	—	+	+
— ouvrier.....	—				—	—	—	—
— employé.....					—	—		
— cadre moyen.....	—	+	+	+	+	+	+	+
— cadre supérieur.....	—	+	+	+	+	+	+	+
<b>Lieu de résidence :</b>								
— commune rurale.....	+	—	—	—	—	—		+
— commune urbaine hors agglomération parisienne.....								
— agglomération parisienne.....					+			—
— ville de Paris.....	—			—	—	+	—	—
<b>Type d'habitat :</b>								
— habitat collectif.....								
— habitat individuel.....	+		+	+	—		+	+

\* Ces effets sont étudiés toutes choses égales par ailleurs (annexe p. 30). Pour chaque caractéristique, la situation de référence par rapport à laquelle sont étudiés les effets est indiquée en italique. L'absence de signe indique que l'effet n'est pas statistiquement significatif; le signe renforcé (+ ou —) souligne les effets les plus marqués.

## Exercice 2 : Étude sur la probabilité de rechercher un autre emploi

Dans leur article "Salaire d'efficience et théorie de la recherche d'emploi : la mobilité de l'emploi vers un autre emploi" [Economie et Statistique, n.290, pp.51-67, 1995], D. Balsan, S. Hanchane et P. Werquin s'intéressent au comportement de recherche d'emploi de personnes déjà employées (*on-the-job search*). A partir des données de l'Enquête Emploi de l'INSEE de 1988, les auteurs observent le fait que l'individu recherche ou non un autre emploi ( $Y$ ). Conformément à la théorie, ils supposent que l'individu recherche si le salaire qu'il perçoit ( $W$ ) est inférieur à un seuil (son salaire de réservation, noté  $W_r$ ) qui n'est pas observable. En revanche, ils soupçonnent que la différence  $D = W_r - W$  dépend linéairement d'une matrice  $X$  de  $k$  variables explicatives.

1. *Modélisation* : Proposer une modélisation économétrique permettant d'analyser l'influence de certaines caractéristiques de l'individu sur le comportement de recherche d'un nouvel emploi. Écrire la vraisemblance associée à votre modèle et mentionner la technique d'estimation recommandée.
2. *Lecture et interprétation des résultats trouvés par Balsan, Hanchane et Werquin* : Dans cet article, la recherche d'emploi des personnes employées fait l'objet d'une estimation logit à partir d'un échantillon qui retient le salaire perçu comme variable explicative. Afin d'évaluer l'impact global du salaire, la spécification retenue considère l'effet du salaire en tant que tel mais aussi par ses effets croisés avec certaines variables explicatives individuelles. Les caractéristiques individuelles (diplôme, âge, statut marital, région d'habitation, sexe) peuvent également intervenir indépendamment du salaire, elles sont donc introduites dans la régression. Les auteurs précisent que la variable âge est centrée par rapport à l'âge moyen de l'échantillon (36 ans). Une partie des résultats de l'article est reproduite. Le tableau 1 présente la moyenne et l'écart-type des variables tandis que le tableau 2 donne les résultats de l'estimation de la probabilité de recherche d'un autre emploi. N'apparaissent dans ce tableau que les variables qui se sont relevées significatives.
  - (a) Vrai ou Faux sur le rôle des caractéristiques individuelles : les affirmations suivantes sont-elles exactes ? (réponses à justifier)
    - i. Les femmes employées sous contrat à durée déterminée ont une probabilité de recherche d'un autre emploi plus élevée que les hommes quel que soit le statut de leur contrat.
    - ii. Les femmes mariées recherchent moins fréquemment un autre emploi que les femmes divorcées.
    - iii. L'obtention d'un diplôme plus élevé se traduit toujours par une diminution de la probabilité de recherche d'un autre emploi.
    - iv. L'effet de l'âge sur la recherche d'un autre emploi est positif puis négatif.
    - v. Une augmentation de salaire entraîne une diminution de la probabilité de rechercher un autre emploi plus importante chez les femmes que chez les hommes.
  - (b) Calculer la probabilité de recherche d'un autre emploi pour les femmes divorcées de 36 ans, possédant un BEPC, travaillant sous un contrat à durée indéterminée,

habitant la région parisienne. Le salaire moyen, par rapport au sexe, au diplôme et à l'âge, de ce groupe de personnes est de 5738 francs mensuel. De façon générale, représenter graphiquement l'évolution de cette probabilité en fonction du salaire (entre 5000 francs et 25000 francs). La probabilité de recherche d'un autre emploi est-elle une fonction convexe ou concave du salaire ?

Extrait de D. Balsan, S. Hanchane et P. Werquin, « Salaire d'efficience et théorie de la recherche d'emploi : la mobilité de l'emploi vers un autre emploi » [Economie et Statistique, n°290, 1995, p 60-63]

**Tableau 1 : Moyenne et l'écart-type des variables**

	Moyenne	Ecart-Type
Ne pas rechercher d'autre emploi	0.925	0.263
Salaire (SAL)	7145	4330
Sup. à Bac+2 (DIP1)	0.054	0.226
Bac+2 (DIP3)	0.060	0.238
Baccalauréat (DIP4)	0.106	0.308
BEP – CAP (DIP5)	0.331	0.471
BEPC (DIP6)	0.067	0.251
Aucun Diplôme (DIP7)	0.381	0.486
Age	36.2	10.801
Divorcé(e) (DIV)	0.055	0.228
Marié(e)	0.641	0.480
Veuf	0.013	0.115
Région Parisienne (PARIS)	0.219	0.414
Femme (F)	0.335	0.472
CDD (CDD)	0.045	0.207
Femme*CDD (F*CDD)	0.017	0.128
Femme* Mariée (F*MARIE)	0.187	0.390

Le salaire est exprimé en francs par mois. A l'exception de l'âge et du salaire, les variables sont

**Tableau 2 : Estimation de la probabilité de recherche d'un autre emploi**

	Coefficient estimé	t de Student
Constante	-1.7213	-11.29
Salaire (SAL)	-0.00016	-6.93
Baccalauréat ou BEP-CAP (DIP45)	0.6163	4.69
BEPC (DIP6)	0.3749	2.27
Age centré (AGEC)	-0.0416	-7.51
Age centré au carré	-0.00175	-3.79
Divorcé(e) (DIV)	0.6407	3.79
Région parisienne (PARIS)	0.3612	3.57
Femme*CDD (F*CDD)	0.5014	2.14
Femme* Mariée (F*MARIE)	-0.3853	-2.63
Femme * Salaire (F*SAL)	-0.00004	-2.10
CDD * Salaire (CDD*SAL)	0.000114	4.32
Sup. à Bac+2 * Salaire (DIP1*SAL)	0.000106	5.31
Bac+2*Salaire (DIP3*SAL)	0.000093	4.24
BEP-CAP*Salaire (DIP5*SAL)	-0.00009	-4.14

### Exercice 3 : Disposition à payer pour améliorer la qualité de l'air

Vous êtes en charge d'analyser les dispositions à payer individuelles pour améliorer la qualité de l'air afin d'aider le gouvernement à prendre des décisions en termes de politiques environnementales. On suppose que la disposition à payer potentielle  $y_i^*$  dépend d'un ensemble  $Z_i$  de caractéristiques individuelles parmi lesquelles le revenu ( $rev_i$ ), le nombre d'enfants ( $enf_i$ ), le niveau d'éducation ( $educ_i$ ), l'âge de l'individu ( $age_i$ ), son sexe ( $sexe_i$ ), le fait que l'individu souffre ou non d'une maladie liée à la pollution telle que l'asthme ( $maladie_i = 1$  si oui et 0 sinon) et le fait qu'il soit membre ou non d'une organisation environnementale ( $environ_i = 1$  si membre et 0 sinon) selon la relation :

$$y_i^* = Z_i\beta + \varepsilon_i.$$

avec  $\varepsilon_i \sim N(0, \sigma^2)$ , i.i.d. et indépendants de toutes les variables explicatives.

Les individus qui ont une disposition à payer potentielle négative ou nulle déclarent un montant de disposition à payer nulle.

Dans un premier temps, on vous demande de déterminer quel type d'individus seraient les plus susceptibles de payer pour améliorer la qualité de l'air et cela sans analyser les montants qu'ils seraient prêts à verser.

1. Calculer la probabilité  $p_i$  que l'individu  $i$  soit prêt à payer pour améliorer la qualité de l'air en fonction de ses caractéristiques ( $Z_i$ ) et la probabilité qu'il ne veuille pas payer.
2. Écrire la fonction de vraisemblance de ce modèle associée à un échantillon de  $N$  individus.
3. Déterminer l'effet marginal sur la probabilité  $p_i$  d'une augmentation d'une unité de revenu  $rev_i$  de l'individu  $i$ .
4. Les résultats d'estimation du modèle approprié par maximum de vraisemblance donnent :

	Coefficient	p-value
Revenu	0.002	0.003
Nombre d'enfants	-0.013	0.29
Sexe (=1 si homme)	-0.054	0.02
Education	0.005	0.11
Maladie (=1 si malade)	-0.027	0.48
Age	-0.0092	0.000
Organisation envir. (=1 si membre)	0.081	0.03
Obs.	2 120	

- (a) Quelle est la probabilité estimée,  $\hat{p}_i$ , qu'un homme de 40 ans, avec 2 enfants, non malade et non membre d'une organisation environnementale, avec 12 années d'éducation et 1200 euros de revenu déclare être prêt à payer pour améliorer la qualité de l'air.

- (b) Calculer l'élasticité de la probabilité  $\hat{p}_i$  par rapport à une augmentation de 1% du revenu.
- (c) Interpréter littérairement les résultats du tableau précédent.

## Exercice 4 : Scoring et défaillances d'entreprises

Une société souhaite construire un modèle de scoring sur les défaillances d'entreprises. On cherche à savoir si la faillite d'une entreprise peut être reliée à son ratio dette sur actif total (RDAT), son taux de croissance annuel des effectifs (CROI), son excédent brut d'exploitation sur actif total (EBE), son ratio Stock sur ventes (SV), son logarithme des ventes (LV) et son secteur d'activité (SEC=manufacturier, services, autres).

1. Proposer une modélisation économétrique permettant de répondre à cette question.  
Justifier votre réponse.
2. Écrire la vraisemblance associée à ce modèle
3. Comment en pratique réaliseriez-vous cette étude ? (données utilisées, construction/codage des variables expliquée et explicatives nécessaires à l'application empirique, signes attendus, etc)
4. Expliquer la démarche permettant de savoir dans quel secteur il y a en moyenne le plus de défaillances d'entreprises, toutes choses étant égales par ailleurs ?
5. Donner la formule qui permettrait de calculer l'élasticité de la probabilité de défaillance de l'entreprise par rapport au ratio dette sur actif total.

## Exercice 5 : Politique de dividendes

Vous êtes embauché pour analyser les politiques de dividendes des entreprises. Soit une société par action susceptible de distribuer des dividendes à ses actionnaires de façon régulière à chaque date  $t = 1, \dots, T$ . On suppose que le montant des dividendes potentiels  $y_t^*$  dépend d'un ensemble de caractéristiques de l'entreprise parmi lesquelles le montant  $ben_t$  des bénéfices de l'année écoulée et le montant  $if_t$  des investissements futurs anticipés de la firme à la date  $t$  selon la relation :

$$y_t^* = \beta_0 + \beta_1 ben_t + \beta_2 if_t + \varepsilon_t.$$

avec  $\varepsilon_t \sim N(0, \sigma^2)$ , i.i.d. et indépendants de toutes les variables explicatives.

On suppose que les dividendes ne sont effectivement versés que lorsque les dividendes potentiels sont positifs.

Dans un premier temps, en tant qu'analyste financier, les actionnaires vous demandent de déterminer la probabilité qu'à une date  $t$  l'entreprise étudiée verse effectivement des dividendes et cela sans analyser la valeur de ceux-ci.

1. Calculez la probabilité que l'entreprise verse des dividendes à la date  $t$  en fonction de ses caractéristiques ( $ben_t, if_t$ ) et la probabilité que l'entreprise ne verse pas de dividendes.
2. Écrivez la fonction de vraisemblance de ce modèle associée à un échantillon de  $T$  observations.
3. Déterminez l'effet marginal sur la probabilité de distribution de dividendes d'une augmentation d'une unité des bénéfices  $ben_t$  de l'entreprise à la date  $t$  quelconque.
4. On vous communique une prévision des résultats de l'entreprise pour l'année  $T + 1$  :  $ben_{T+1} = 100$  et  $if_{T+1} = 200$ . Les résultats d'estimation du modèle approprié par maximum de vraisemblance donnent :  $\hat{\beta}_0 = 1$ ,  $\hat{\beta}_1 = 0.05$ ,  $\hat{\beta}_2 = -0.029$ . Fournissez aux actionnaires :
  - (a) la probabilité estimée que les dividendes soient effectivement versés en  $T + 1$ ,
  - (b) l'élasticité de la probabilité de versement en  $T + 1$  par rapport à une augmentation de 1% des bénéfices attendus.

## Exercice 6 : État de santé

A partir d'une enquête étasunienne auprès de 7397 individus, un économètre souhaite s'intéresser aux déterminants de l'état de santé des individus. Pour ce faire, on dispose d'un certain nombre d'informations concernant l'ensemble des individus interrogés dans le cadre de l'enquête :

- l'état de santé avec deux modalités,  $santé=1$  si l'individu se déclare en bonne ou très bonne santé et  $santé=0$  si l'individu se déclare en mauvaise ou très mauvaise santé ;
- l'âge et le carré de l'âge ( $Age$   $Age2$ , mesurée en années) ;
- le genre de l'individu,  $sex=1$  si l'individu est une femme ;
- la zone d'habitation,  $rural=1$  si l'individu habite dans une zone rurale ;
- la région d'habitation avec 5 modalités : le Nord-Est ( $Region1$ ), le Midwest ( $Region2$ ), le Sud ( $Region3$ ), l'Ouest ( $Region4$ ), et Autres (modalité en référence) ;
- la taille de l'individu ( $Height$  exprimé en cm) ;
- le poids de l'individu ( $Weight$  exprimé en kg) ;
- le nombre d'individus composant le ménage de l'individu ( $House\_size$ ) ;
- si l'individu est diabétique ( $Diabetes=1$  si l'individu est diabétique).

1. Proposer une modélisation économétrique permettant de savoir si l'état de santé d'un individu peut être relié aux autres caractéristiques individuelles présentes dans la base de données. Vous donnerez l'expression de la probabilité qu'un individu soit en bonne santé en fonction des ces différents facteurs ainsi que l'expression de la probabilité que l'individu soit en mauvaise santé.
2. Ecrire la fonction de vraisemblance de ce modèle.
3. Les résultats de l'estimation du modèle approprié par maximum de vraisemblance, en supposant que les termes d'erreur suivent une loi logistique de moyenne nulle et de variance unitaire, sont reportés dans le tableau 1.
  - (a) Que pouvez-vous dire de la significativité globale du modèle ?

- (b) Interprétez littérairement les résultats du modèle.
- (c) Comment jugeriez-vous de la qualité d'ajustement du modèle ? (Description de la démarche, calculs à effectuer, règle de décisions, etc.)
4. La matrice de confusion du modèle est reportée dans le tableau 2. Que pouvez-vous en conclure ?
  5. On souhaite maintenant savoir si l'impact de l'âge est le même pour les personnes qui ont du diabète, et celles qui n'en ont pas. Pour cela, l'économètre introduit une variable croisée dans le modèle (*age \* diabete*). La nouvelle estimation par l'estimateur du Maximum de vraisemblance donne les résultats reportés dans le tableau 3.
    - (a) Le modèle est-il globalement significatif ?
    - (b) Interpréter la variable croisée.
  6. Par ailleurs, la matrice de confusion associée à ce modèle est reportée dans le tableau 3.
    - (a) Que pouvez-vous en conclure ?
    - (b) Quel modèle est à privilégier selon vous ?

Les tableaux :

Tableau 1

```
. logit sante age age2 sex rural region2-region4 height weight house_size

Iteration 0:  log likelihood = -4660.7274
Iteration 1:  log likelihood = -3868.548
Iteration 2:  log likelihood = -3829.0863
Iteration 3:  log likelihood = -3828.6453
Iteration 4:  log likelihood = -3828.6453

Logistic regression                                         Number of obs      =     7,397
                                                               LR chi2(10)       =    1664.16
                                                               Prob > chi2      =     0.0000
                                                               Pseudo R2        =     0.1785

Log likelihood = -3828.6453
```

sante	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	-.078417	.0137653	-5.70	0.000	-.1053965 -.0514375
age2	.0002269	.0001417	1.60	0.109	-.0000509 .0005046
sex	.2730203	.0807448	3.38	0.001	.1147634 .4312772
rural	-.3002209	.0576646	-5.21	0.000	-.4132414 -.1872005
region2	-.3742795	.08739	-4.28	0.000	-.5455607 -.2029983
region3	-.8447963	.08551	-9.88	0.000	-1.012393 -.6771998
region4	-.5901148	.0878226	-6.72	0.000	-.7622439 -.4179856
height	.041999	.004617	9.10	0.000	.0329498 .0510482
weight	-.0141321	.0020583	-6.87	0.000	-.0181663 -.010098
house_size	-.038888	.0187994	-2.07	0.039	-.0757342 -.0020418
_cons	-1.380897	.8286454	-1.67	0.096	-3.005012 .2432186

Tableau 2

Logistic model for sante

Classified	True		Total
	D	~D	
+	4238	1205	5443
-	760	1194	1954
Total	4998	2399	7397

Classified + if predicted Pr(D) >= .5  
 True D defined as sante != 0

Tableau 3

Logistic regression	Number of obs	=	7,397
	LR chi2(11)	=	1809.68
	Prob > chi2	=	0.0000
Log likelihood = -3755.8896	Pseudo R2	=	0.1941

sante	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	-.0807618	.0138866	-5.82	0.000	-.107979 -.0535446
age2	.0002853	.0001433	1.99	0.046	4.52e-06 .0005661
age_diabetes	-.0257589	.0024186	-10.65	0.000	-.0304993 -.0210185
sex	.2880039	.0817124	3.52	0.000	.1278507 .4481572
rural	-.3125382	.0583881	-5.35	0.000	-.4269769 -.1980996
region2	-.3787804	.0887351	-4.27	0.000	-.5526981 -.2048628
region3	-.8447496	.0868582	-9.73	0.000	-1.014988 -.6745108
region4	-.5985655	.0891524	-6.71	0.000	-.7733011 -.42383
height	.0401895	.0046711	8.60	0.000	.0310343 .0493447
weight	-.011433	.0021009	-5.44	0.000	-.0155507 -.0073153
house_size	-.0376946	.0190909	-1.97	0.048	-.075112 -.0002772
_cons	-1.246554	.8367696	-1.49	0.136	-2.886593 .3934839

Tableau 4

Logistic model for sante

Classified	True		Total
	D	~D	
+	4345	1248	5593
-	653	1151	1804
Total	4998	2399	7397

Classified + if predicted Pr(D) >= .5  
True D defined as sante != 0