

MODÈLES À CHOIX DISCRET  
CORRECTION EXERCICES 1 ET 2

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

1. a. On peut définir  $Y_i$  tel que :

$$Y_i = \begin{cases} 1 & \text{si } Z_i > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $Z_i = U(1, X_i) - U(0, X_i)$

b. On note  $Z_i = X_i\beta + u_i$

On veut  $P(Y_i = 1)$ .

$$\begin{aligned} P(Y_i = 1) &= P(Z_i > 0) \\ &= P(X_i\beta + u > 0) \\ &= P(u_i > -X_i\beta) \\ &= 1 - P(u_i < -X_i\beta) \\ &= 1 - F(-X_i\beta) \\ &= F(X_i\beta) \end{aligned}$$

Avec  $F$  la fonction de répartition de  $u$ .

Donc on a :  $P(Y_i = 1) = F(X_i\beta)$  et  $P(Y_i = 0) = 1 - F(X_i\beta)$ .

*Notes :*

Dans le cas d'un modèle probit on a :  $F(X_i\beta) = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ .

Dans le cas d'un modèle Logit on a :  $F(X_i\beta) = \Lambda(X_i\beta) = \frac{1}{1+e^{-X_i\beta}} = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$ .

c. Fonction de vraisemblance :

$$L(Y, X, \beta) = \prod_i^n [F(X_i\beta)]^{Y_i} \cdot [1 - F(X_i\beta)]^{(1-Y_i)}$$

2. a. Pour tester la significativité sur un coefficient :

— Test de Wald

- Test de Student (**Attention uniquement dans le cas où on teste  $\beta = 0$ , on a test de Wald =  $t^2$** )  
En plus, on peut également mettre en oeuvre :
- Test du score
- Test du rapport de vraisemblance

*Note sur le test de Wald :*

On veut tester :

$$H_0 : \beta = \beta_0 \text{ contre } H_1 : \beta \neq \beta_0$$

Dans le test de Wald, l'estimateur du maximum de vraisemblance  $\hat{\beta}$  du paramètre  $\beta$  est comparé à la valeur  $\beta_0$ , sous l'hypothèse que la différence est distribuée approximativement selon la loi de Gauss. En pratique le carré de la différence est comparé à un seuil de la loi du Chi2. Dans le cas univarié, la statistique de Wald est :

$$\frac{(\hat{\beta} - \beta_0)^2}{\text{var}(\hat{\beta})}$$

Dans le cas particulier où la valeur testée est  $\beta_0 = 0$ , la statitisque de test d'un test de Wald correspond donc simplement au  $t$  de Student au carré. Ainsi, dans la pratique, pour tester la significativité d'un coefficient on utilisera la même méthodologie que le test de Student (On compare  $|t|$  à 1,96 au seuil de 5%).

b. Test de significativité sur un ensemble de coefficients :

- Test de Wald
- Test du score
- Test du rapport de vraisemblance

avec  $H_0 = \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$ .

c. Les statistiques descriptives permettent seulement de décrire quelles sont les caractéristiques des détenteurs d'une chaîne de haute fidélité. En particulier, elles indiquent ici qu'on trouve plus de détenteurs parmi les locataires que parmi les propriétaires. En revanche cela ne signifie pas que le fait d'être locataire influence la détention d'une chaîne haute fidélité. L'interprétation n'est donc pas contradictoire mais complémentaire. Les résultats économétriques permettent en effet de compléter les statitisques descriptives et montrent notamment que les propriétaires sont plus enclins que les locataires à avoir une chaîne haute fidélité, **toutes choses étant égales par ailleurs**.

d. Toutes choses étant égales par ailleurs :

- Les accédants anciens, et les propriétaires ont une probabilité plus importante d'avoir un lave linge que les autres (locataires et accédants récents) ;

- Les ménages dans lesquels le chef de familles a moins de 30 ou plus de 75 ans ont une probabilité plus faible d'avoir un lave linge que les autres. A l'inverse, ceux dont le chef de famille à entre 46 et 56 ans ont une probabilité plus importante d'en avoir un que tous les autres (on a : moins de 30 ans / plus de 75 ans < de 30 à 45 ans / de 66 ans à 75 ans < de 46 ans à 65 ans) ;
- Les ménages composés d'un seul individu ou d'un couple seul ont une probabilité plus faible que tous les autres d'avoir un lave linge. A l'inverse, les couples avec deux enfants ou plus ont une probabilité plus grande que les autres d'en avoir un ;
- La durée du mariage dans un couple n'a pas d'impact que la probabilité d'avoir un lave linge ;
- Les ménage dont le revenu est inférieur à 55 000 F ont une probabilité inférieure à celle des autres de posséder un lave linge ;
- Les ménages pour lesquels le chef de famille est un cadre moyen ont une probabilité plus faible que les autres d'avoir un lave linge ;
- Les ménages vivants à Paris ont une probabilité plus faible que les autres d'avoir un lave linge. A l'inverse, ceux qui vivent dans une commune rurale ont une probabilité plus élevée que les autres d'en avoir un ;
- Enfin, les ménages vivants dans une habitation individuelle ont une probabilité plus importante que les autres d'avoir un lave linge.

**Notes : Dans le cas des modèles à choix discrets on n'interprète jamais la valeur du coefficient, seul son signe peut être interpréter. Pour donner une interprétation quantitative des résultats il faudra calculer les effets marginaux.**

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

1. On sait que l'individu recherche un emploi si  $W_r > W$  ou encore si  $D = W_r - W > 0$ .  
 Donc on peut écrire :

$$Y_i = \begin{cases} 1 & \text{si } D_i > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $D_i = X_i\beta + u_i$ .

Fonction de vraisemblance :

$$L(Y, X, \beta) = \prod_i^n [F(X_i\beta)]^{Y_i} \cdot [1 - F(X_i\beta)]^{(1-Y_i)}$$

Ce modèle pourra être estimé par **maximum de vraisemblance**.

2. a. i. Vrai. Il faut regarder la variable *femme \* CDD*. La variable est significative et positive, on peut donc dire que c'est vrai, une femme en CDD a une probabilité plus élevée que les hommes de rechercher un emploi (car les hommes quelque soit le statut en référence de la variable croisée).
- ii. Vrai. Il faut comparer *femme \* mariée* et *divorcée*. On pourrait réaliser un test de Wald pour tester si l'écart entre les deux est significatif, or ici ce n'est pas la peine car le signe des coefficients est opposé et ils sont tous les deux significatifs. C'est donc vrai.
- iii. Faux. Car "aucun diplôme" est également en référence. On peut voir ça en regardant dans le tableau 1, dans lequel l'ensemble des catégories figure.
- iv. Vrai. Attention c'est **l'âge centré** (*age-36*). Donc l'effet est positif pour les individus de moins de 36 ans et devient négatif ensuite.
- v. Vrai. Car les coefficients de la variable *salaire* et *femme \* salaire* sont négatifs.

3. On veut calculer  $P(\widehat{Y}_i = 1)$ .

$P(\widehat{Y}_i = 1) = \frac{\exp(X_i\hat{\beta})}{(1+\exp(X_i\hat{\beta}))}$ , car c'est un modèle logit qui est estimé ici.

Etape 1 : on calcul  $X_i\hat{\beta}$  :

$$X_i\hat{\beta} = -1,7213 + (-0,00016 \times 5738) + 0,3749 + (-0,0416 \times 0) + (-0,00175 \times 0) + 0,6407 + 0,3612 + (-0,00004 \times 5738) = -1,4921$$

(Note : Attention à l'âge centré = 0).

Donc  $P(\widehat{Y_i} = 1) = 0,1836 \rightarrow 18,36\%$ .

Par ailleurs, on trouve également :

Pour 5000 francs :

$X_i \hat{\beta} = -1.3445$ , donc  $P(\hat{Y_i} = 1) = 0,2068 \rightarrow 20,68\%$ .

Pour 10000 francs :

$X_i \hat{\beta} = -2.3445$ , donc  $P(\hat{Y_i} = 1) = 0,0875 \rightarrow 8.75\%$ .

Pour 15000 francs :

$X_i \hat{\beta} = -3.3445$ , donc  $P(\hat{Y_i} = 1) = 0,0341 \rightarrow 3.41\%$ .

Pour 20000 francs :

$X_i \hat{\beta} = -4.3445$ , donc  $P(\hat{Y_i} = 1) = 0,0128 \rightarrow 1.28\%$ .

Pour 25000 francs :

$X_i \hat{\beta} = -5.3445$ , donc  $P(\hat{Y_i} = 1) = 0,0047 \rightarrow 0.47\%$ .

On trace le graphique  $\rightarrow$  fonction convexe.

TD ECONOMÉTRIE - MODÈLES À CHOIX DISCRET  
CORRECTION EXERCICES 3 ET 4

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

1. On propose le modèle suivant :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $Y_i^* = Z_i\beta + \epsilon_i$  et  $\epsilon_i \sim N(0, \sigma^2)$ , iid.

On veut  $P(Y_i = 1)$ .

$$\begin{aligned} P(Y_i = 1) &= P(Y_i^* > 0) \\ &= P(Z_i\beta + \epsilon_i > 0) \\ &= P(\epsilon_i > -Z_i\beta) \\ &= P\left(\frac{\epsilon_i}{\sigma} > \frac{-Z_i\beta}{\sigma}\right) \\ &= 1 - P\left(\frac{\epsilon_i}{\sigma} < \frac{-Z_i\beta}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{-Z_i\beta}{\sigma}\right) \\ &= \Phi\left(\frac{Z_i\beta}{\sigma}\right) \end{aligned}$$

Avec  $\Phi$  la fonction de répartition de la loi normale.

2. Fonction de vraisemblance :

$$L(Y, X, \beta) = \prod_{i=1}^n [\Phi\left(\frac{Z_i\beta}{\sigma}\right)]^{Y_i} [1 - \Phi\left(\frac{Z_i\beta}{\sigma}\right)]^{(1-Y_i)}$$

3. On veut l'effet marginal d'une augmentation d'une unité de revenu de l'individu sur sa probabilité  $p_i$  qu'il soit prêt à payer pour améliorer la qualité de l'air :

$$\frac{\delta P(Y_i = 1)}{\delta Rev_i} = \beta_{rev} \times f\left(\frac{Z_i \beta}{\sigma}\right) = \beta_{rev} \times \phi\left(\frac{Z_i \beta}{\sigma}\right)$$

Avec  $\phi\left(\frac{Z_i \beta}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{Z_i \beta}{\sigma}\right)^2/2}$ .

*Notes :* On remarque que l'effet marginal dépend de  $i$ . En effet, l'effet marginal n'est pas constant et est calculé le plus souvent au point moyen.

Attention, la formule du calcul de l'effet marginal n'est valable que pour les variables quantitatives. Pour une variable dichotomique,  $x_k$ , il faut calculer  $P(Y_i = 1|x_k = 1) - P(Y_i = 1|x_k = 0)$ .

#### 4. Prévision pour l'individu donné :

Etape 1 : Calcul de  $Z_i \hat{\beta}$

$$Z_i \hat{\beta} = 1200 \times 0,002 + (-0,013 \times 2) + (-0,054 \times 1) + (0,005 \times 12) + (-0,027 \times 0) + (-0,0092 \times 40) + (0,081 \times 0) = 2,012$$

Etape 2 : Calcul de  $P(\widehat{Y_i = 1})$

$$\begin{aligned} P(\widehat{Y_i = 1}) &= \Phi\left(\frac{Z_i \beta}{\sigma}\right) \\ &= \Phi\left(\frac{Z_i \beta}{1}\right) \\ &= \Phi(2,012) \\ &= 0,9778 \\ &= 97,78\% \end{aligned}$$

L'individu  $i$  décrit dans la question a une probabilité d'être disposé à payer pour améliorer la qualité de l'air estimée à 97,78 %.

#### 5. L'élasticité de la probabilité de cet individu $i$ d'être disposé à payer pour améliorer la qualité de l'air par rapport à une augmentation de 1% de son revenu est telle que :

$$\frac{\delta P(Y_i = 1)}{\delta Rev_i} \times \frac{Rev_i}{P(\widehat{Y_i = 1})}$$

On fait le calcul :

$$\begin{aligned}
 \frac{\delta P(Y_i = 1)}{\delta Rev_i} &= \beta Rev_i \times \phi\left(\frac{Z_i \beta}{1}\right) \\
 &= 0,002 \times \frac{1}{\sqrt{2\pi}} e^{-(2,012)^2/2} \\
 &= 0,00003
 \end{aligned}$$

$$\begin{aligned}
 \frac{Rev_i}{\widehat{P(Y_i = 1)}} &= \frac{1200}{0,9778} \\
 &= 1227,244
 \end{aligned}$$

→ Elasticité = 0,036

Une augmentation de 1% du revenu entraîne une augmentation de 0,036 % de la probabilité que l'individu  $i$  soit disposé à payer pour améliorer la qualité de l'air.

#### 6. Interprétation des résultats :

Toutes choses étant égales par ailleurs :

- Une augmentation du revenu entraîne une augmentation de la probabilité que l'individu soit disposé à payer pour améliorer la qualité de l'air ;
- On observe par ailleurs que les hommes ont une probabilité plus faible que celle des femmes de se déclarer disposé à payer pour améliorer la qualité de l'air ;
- On voit que l'âge des individus a un effet négatif sur leur probabilité de se déclarer disposé à payer pour améliorer la qualité de l'air ;
- On observe que le fait d'être membre d'une organisation environnementale augmente la probabilité que l'individu soit disposé à payer pour améliorer la qualité de l'air ;
- Enfin, on voit que le nombre d'enfants, le niveau d'éducation et le fait d'être malade n'ont en revanche pas d'impact sur la probabilité d'être disposé à payer pour améliorer la qualité de l'air (variables non significatives).

### Exercice 4 : Scoring et défaillance d'entreprise

#### 1. On veut expliquer la variable $Y_i$ telle que :

$$Y_i = \begin{cases} 1 & \text{si l'entreprise fait faillite,} \\ 0 & \text{sinon} \end{cases}$$

La variable que l'on souhaite expliquer est une variable dichotomique. Dans ce cadre-là, on propose le modèle suivant :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $Y_i^*$  la variable latente du modèle qui dépend linéairement des caractéristiques de l'entreprise  $X_i$  :  $Y_i^* = X_i\beta + \epsilon_i$ .

## 2. Fonction de vraisemblance :

$$L(Y, X, \beta) = \prod_{i=1}^n [F(\frac{Z_i\beta}{\sigma})]^{Y_i} \cdot [1 - F(\frac{Z_i\beta}{\sigma})]^{(1-Y_i)}$$

Avec  $F$  la fonction de répartition de  $\epsilon$  (de la loi normale ou logistique selon si on utilise un modèle Probit ou un Logit).

3. En pratique, pour estimer la probabilité de faire faillite pour une entreprise il faut tout d'abord collecter les données nécessaires. Premièrement, nous avons besoin de constituer un échantillon représentatif de plusieurs entreprises avec une proportion suffisamment importante d'entreprises qui ont effectivement fait faillite et d'autres non. A partir de cette information là, nous pourrons contruire notre variable à expliquer telle que :

$$Y_i = \begin{cases} 1 & \text{si l'entreprise fait faillite,} \\ 0 & \text{sinon} \end{cases}$$

Il faudra ensuite collecter un certain nombre de caractéristiques pour ces mêmes entreprises. Nous cherchons en particulier à récollecter des caractéristiques qui expliqueraient le fait que l'entreprise fasse faillite (c'est-à-dire les variables RDAT, CROI, EBE, SV, LV et SEC). Ces variables là seront les variables explicatives du modèles notées  $X_i$ . On peut ainsi réécrire le modèle tel que :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $Y_i^*$  la variable latente du modèle qui dépend linéairement des caractéristiques de l'entreprise  $X_i$  :  $Y_i^* = X_i\beta + \epsilon_i$ .

Pour chaque variable explicative, lorsque elle est de type qualitatif avec plusieurs modalités, comme c'est le cas de la variable de secteur d'activité, il faut créer autant de dichotomiques que de modalités, ici  $m = 3$ , pour en insérer  $m - 1$  dans le modèle, la dernière sera la référence (au préalable on s'assure qu'il y a au moins 5% de l'échantillon dans chaque modalité, si ce n'est pas le cas nous effectueons les regroupements nécessaires et pertinents). Une fois l'ensemble de nos variables explicatives recodées lorsque c'était nécessaire, nous estimons le modèle précédemment décrit par un modèle Logit

ou Probit. A l'issue de cette estimation nous effectuerons les tests nécessaires avant l'interprétation. En particulier on s'assurera de la qualité d'ajustement du modèle en construisant la matrice de contingence. Si la qualité est correcte nous testons la significativité de chaque variable avant d'en interpréter le signe (Test de Wald). Là, nous pourrons interpréter les effets des caractéristiques. On peut par exemple s'attendre à un effet négatif de la variable RDAT, etc.

4. Pour savoir dans quel secteur il y a en moyenne le plus de défaillances d'entreprise nous devons intégrer comme variables explicatives autant de dichotomiques que de secteurs à analyser, moins un (la référence). Par exemple, on insère dans le modèle les dichotomiques correspondant au secteur manufacturier et au secteur des services, la catégorie autres sera en référence. A l'issue de l'estimation, on pourra interpréter la différence de probabilité de faire faillite entre le secteur manufacturier et autres ainsi qu'entre le secteur des services et la catégories autre. En effet, si les coefficients sont significatifs on interprétera la différence selon le signe du coefficient, sinon cela implique qu'il n'y a pas de différence entre la catégorie considérée et la référence. Pour statuer de la différence entre la catégorie manufacturier et services, si les deux sont significatifs et de même signe, il faudra effectuer un test supplémentaire pour tester si l'acrt est significatif ou non (Test de Wald). Une fois cela fait nous pourrons alors ordonner les secteurs dans lesquels il y a plus ou moins de risque de défaillance.
5. Calcul de l'élastcité de la probabilité de défaillance de l'entreprise par rapport au ratio dette sur actif total :

$$\frac{\delta P(Y_i = 1)}{\delta RDAT_i} \times \frac{RDAT_i}{P(\widehat{Y_i = 1})} = \beta_{RDAT_i} \cdot f(X_i \beta) \times \frac{RDAT_i}{F(X_i \beta)}$$

**TD ECONOMÉTRIE - MODÈLES À CHOIX DISCRET**  
**CORRECTION EXERCICES 5 ET 6**

**Exercice 5 : Politique de dividendes**

- On note la variable  $Y_t$  qui vaut 1 si l'entreprise verse des dividendes à la date  $t$ , 0 sinon.  
 On propose le modèle suivant :

$$Y_t = \begin{cases} 1 & \text{si } y_t^* > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $y_t^* = \beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t + \epsilon_t$  et  $\epsilon_t \sim N(0, \sigma^2)$ , iid.

On veut  $P(Y_t = 1)$ .

$$\begin{aligned} P(Y_t = 1) &= P(y_t^* > 0) \\ &= P(\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t + \epsilon_t > 0) \\ &= P(\epsilon_t > -(\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t)) \\ &= P\left(\frac{\epsilon_t}{\sigma} > \frac{-(\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t)}{\sigma}\right) \\ &= 1 - P\left(\frac{\epsilon_t}{\sigma} < \frac{-(\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t)}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{-(\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t)}{\sigma}\right) \\ &= \Phi\left(\frac{\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t}{\sigma}\right) \end{aligned}$$

Donc  $P(Y_t = 0) = 1 - \Phi\left(\frac{\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t}{\sigma}\right)$

Avec  $\Phi$  la fonction de répartition de la loi normale.

- Fonction de vraisemblance :

$$L(Y, X, \beta) = \prod_{t=1}^n [\Phi\left(\frac{\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t}{\sigma}\right)]^{Y_t} [1 - \Phi\left(\frac{\beta_0 + \beta_1 \text{ben}_t + \beta_2 \text{if}_t}{\sigma}\right)]^{(1-Y_t)}$$

3. On veut l'effet marginal d'une augmentation d'une unité des bénéfices  $ben_t$  de l'entreprise à la date  $t$  sur sa probabilité de distribution de dividendes :

$$\frac{\delta P(Y_t = 1)}{\delta ben_t} = \beta_1 \times f\left(\frac{\beta_0 + \beta_1 ben_t + \beta_2 if_t}{\sigma}\right) = \beta_1 \times \phi\left(\frac{\beta_0 + \beta_1 ben_t + \beta_2 if_t}{\sigma}\right)$$

Avec  $\phi\left(\frac{\beta_0 + \beta_1 ben_t + \beta_2 if_t}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{\beta_0 + \beta_1 ben_t + \beta_2 if_t}{\sigma}\right)^2/2}$ .

4. a) Prévision en  $T + 1$  :

Etape 1 : Calcul de  $X_{t+1}\hat{\beta}$

$$X_{t+1}\hat{\beta} = 1 + 100 \times 0,05 + 200 \times 0,029 = 0,2$$

Etape 2 : Calcul de  $P(\widehat{Y_{t+1}} = 1)$

$$\begin{aligned} P(\widehat{Y_{t+1}} = 1) &= \Phi\left(\frac{\beta_0 + \beta_1 ben_t + \beta_2 if_t}{\sigma}\right) \\ &= \Phi\left(\frac{\beta_0 + \beta_1 ben_t + \beta_2 if_t}{1}\right) \\ &= \Phi(0,2) \\ &= 0,5793 \\ &= 57,93\% \end{aligned}$$

La probabilité que l'entreprise verse des dividendes en  $T + 1$  est estimée à 57,93 %.

- b) L'élasticité de la probabilité de versement de dividendes en  $T + 1$  par rapport à une augmentation de 1% des bénéfices est telle que :

$$\frac{\delta P(Y_{t+1} = 1)}{\delta ben_{t+1}} \times \frac{ben_{t+1}}{P(\widehat{Y_{t+1}} = 1)} = 0,05 \times \frac{1}{\sqrt{2\pi}} e^{-(0,2)^2/2} \times \frac{100}{0,57926} = 3,375$$

→ Elasticité = 3,375

Une augmentation de 1% des bénéfices entraîne une augmentation de 3,38 % de la probabilité que l'entreprise verse des dividendes.

## Exercice 6 : Etat de santé

- On veut expliquer la variable  $Y_i$  telle que :

$$Y_i = \begin{cases} 1 & \text{si l'individu est en bonne santé,} \\ 0 & \text{sinon} \end{cases}$$

La variable que l'on souhaite expliquer est une variable dichotomique. Dans ce cadre-là, on propose le modèle suivant :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* > 0, \\ 0 & \text{sinon} \end{cases}$$

Avec  $Y_i^*$  la variable latente du modèle qui dépend linéairement des caractéristiques de l'individu  $X_i$  :  $Y_i^* = X_i\beta + \epsilon_i$ .  $X_i$  correspond aux caractéristiques individuelles suivantes : l'âge, l'âge au carré, le sexe, la zone d'habitation, la région, la taille, le poids, la taille du ménage et si l'individu est diabétique.

A partir de ce modèle, on peut définir  $P(Y_i = 1)$ .

$$\begin{aligned} P(Y_i = 1) &= P(Y_i^* > 0) \\ &= P(X_i\beta + \epsilon_i > 0) \\ &= P(\epsilon_i > -X_i\beta) \\ &= 1 - P(\epsilon_i > -X_i\beta) \\ &= F(X_i\beta) \end{aligned}$$

Avec  $F(\cdot)$  la fonction de répartition des termes d'erreurs.

- Fonction de vraisemblance :

$$L(Y, X, \beta) = \prod_{i=1}^n [F(X_i\beta)]^{Y_i} \cdot [1 - F(X_i\beta)]^{(1-Y_i)}$$

Avec  $F$  la fonction de répartition de  $\epsilon$  (de la loi normale ou logistique selon si on utilise un modèle Probit ou un Logit).

- On regarde le test du rapport de vraisemblance → LR chi2 (10) = 1664,16 et la p-value associée (Prob > chi2 = 0,0000). Le résultat du test nous conduit à rejeter  $H_0$ , le modèle est globalement significatif.
- En moyenne et toutes choses étant égales par ailleurs :

- Plus un individu est âgé, plus sa probabilité d'être en bonne santé décroît. Autrement dit, l'âge impacte négativement la probabilité d'être en bonne santé. De plus, puisque le coefficient rattaché à la forme quadratique de l'âge n'est pas significativement différent de 0, il n'y a pas d'effet non linéaire de l'âge sur la probabilité d'être en bonne santé ;
- Les femmes ont une probabilité plus importante d'être en bonne santé par rapport aux hommes ;
- Une personne vivant dans une zone rurale a une probabilité moins importante d'être en bon état de santé par rapport à une personne vivant dans une zone urbaine ;
- Pour les régions : Une personne vivant dans le Midwest, le Sud, ou l'Ouest, a une probabilité moins importante d'être en bonne santé par rapport à une personne vivant dans le Nord-Est. Attention : on ne peut toutefois pas faire de classement entre les trois variables de régions sans faire de test complémentaire ;
- Plus une personne est grande, plus sa probabilité d'être en bonne santé est importante ;
- Plus le poids d'une personne est important, plus sa probabilité d'être en bonne santé décroît ;
- Plus le nombre de personnes dans le ménage de l'individu est important, moins sa probabilité d'être en bonne santé est importante.

c) On peut se référer au peuso- $R^2$  pour avoir une idée de la qualité du modèle. Toutefois cet indice est toujours bas et on lui préfère le taux de prédictions fausses. En particulier, nous allons construire la matrice de confusion :

- A partir de l'estimation du modèle, on calcule pour chaque individu sa probabilité prédite d'être en bonne santé  $\widehat{P(Y_i = 1)}$ . Si cette probabilité prédite est supérieure à 0.5, l'individu est classé comme étant en bonne santé ( $Y=1$ ). Inversement, si la probabilité prédite est inférieure à 0.5, on classe l'individu comme étant en mauvaise santé ( $Y=0$ ). On obtient ainsi un nombre d'individus prédis en bonne et en mauvaise santé.
- On confronte ensuite ce classement avec les vraies valeurs observées de la variable santé.
- On peut alors retrouver le nombre de personnes pour lesquelles le modèle prédit correctement/mal. Le taux de prédictions fausses correspond alors à :  $\frac{\text{Nb de mal classés}}{\text{nb d'individus}}$ . Si ce taux est supérieur à 50%, cela signifie que le modèle prédit encore plus mal que le hasard. Dans ce cas là le modèle ne pourra être utilisé.

#### 4. Taux de prédictions fausses :

$$\frac{1205+760}{7397} = 26,56\%$$

→ inférieur à 50% donc ok. Attention cependant, on remarque que le modèle prédit assez mal les individus en mauvaise santé :  $\frac{760}{1954}$ . Importante de regarder colonne par colonne !!

#### 5. a) Oui (test du rapport de vraisemblance).

- b) On voit que la variable croisée est significative. Cela signifie que l'effet négatif de l'âge est d'autant plus important pour les individus diabétiques.
6. Calcul du taux de prédictions fausses :  $\frac{1248+653}{7397} = 25,7\%$ . On préfère donc le modèle 2 dans lequel le taux de prédition est plus faible.