

ÉCONOMÉTRIE CONTRÔLE CONTINU

Note : Pour tous les exercices, vous effectuerez les tests à un seuil de risque de 5%.

Exercice 1 :

1. On veut expliquer la variable Y_i telle que :

$$Y_i = \begin{cases} 1 & \text{si l'enfant souffre d'anémie,} \\ 0 & \text{sinon} \end{cases}$$

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'enfant et de son foyer X_i : $Y_i^* = X_i\beta + \epsilon_i$. X_i correspond aux caractéristiques individuelles suivantes : l'âge, le sexe, le fait de bénéficier du programme, le revenu familial, le nombre de personnes dans le ménage, le nombre de pièces dans le logement, et si le logement est relié à l'électricité et à l'eau.

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

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

Et $P(Y_i = 0) = 1 - F(X_i\beta)$

Avec $F(\cdot)$ la fonction de répartition de la loi normale.

2. Fonction de vraisemblance :

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

3. Toutes choses étant égales par ailleurs :

- Le fait de bénéficier du programme permet de diminuer la probabilité pour un enfant de souffrir d'anémie ;
- L'âge diminue également la probabilité de souffrir d'anémie ;
- Les garçons ont une probabilité plus importante que les filles de souffrir d'anémie ;
- Enfin, ni le revenu familial, le nombre de personnes, le nombre de pièces, le fait d'avoir l'électricité et l'eau n'ont d'impact sur la survue d'un problème d'anémie chez les enfants.

4. Calcul de \hat{p}_i :

Etape 1 : Calcul de $X_i\beta$

$$X_i\beta = -0,084 + 1 \times (-0,199) + 15 \times (-0,159) + 0 \times 0,095 + 565 \times 0,000004 + 3 \times 0,012 + 4 \times (-0,039) + 0 \times 0,274 + 0 \times 0,025 = -2,7857$$

Etape 2 : Calcul de $P(Y_i = 1) = F(-2,7857) = 0,0026$

Donc la probabilité pour cette jeune fille de souffrir d'anémie est estimée à 0,26 %.

5. Non car variable de revenu est **non significative**. (Note : il ne fallait pas calculer l'élasticité ici !!).

6. (a) Ce tableau correspond à la matrice de confusion et est construit comme suit :

- A partir de l'estimation du modèle, on calcule pour chaque individu sa probabilité prédictive de souffrir d'anémie $\widehat{P}(Y_i = 1)$. Si cette probabilité prédictive est supérieure à 0,5, l'enfant est classé comme ayant souffert d'anémie durant les 12 derniers mois ($Y = 1$). Inversement, si la probabilité prédictive est inférieure à 0,5, on classe

l'enfant comme n'ayant pas souffert d'anémie ($Y = 0$). On obtient ainsi un nombre d'individus prédis en bonne et en mauvaise santé par le modèle.

- 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.

- (b) On calcule le taux de prédictions fausses :

$$\frac{1090+310}{3756} = 0,37, \text{ soit un taux de prédition fausse de } 37\% < \text{à } 50\% \text{ donc on peut garder le modèle.}$$

Exercice 2 :

1. Test de significativité globale du modèle :

$$H_0 : \beta_1 = \dots = \beta_k = 0 \text{ contre } H_1 : \exists \beta_i \neq 0$$

Statistique de test :

$$F^* = \frac{R^2}{1 - R^2} \frac{T - k}{k - 1} \sim F(k - 1, T - k)$$

Règle de décision :

Rejet de H_0 si $F^* > F(k - 1, T - k)$.

Dans notre cas, on trouve $F^* = 352,07$. On rejette H_0 , le modèle est globalement significatif.

2. Calcul de la variance résiduelle : $\frac{499234,74}{35255} = 14,16$

3. IC associé à $\beta_{prestation}$:

$$IC = [\hat{\beta} - \hat{\sigma} \times t_\alpha; \hat{\beta} + \hat{\sigma} \times t_\alpha] = [0.04646; 0.07946].$$

95% de chances que le coefficient associé à la variable *prestation* soit compris dans cet intervalle.

4. Les variables d'âges et de région sont des variables qualitatives à plusieurs modalités. Ce type de variables ne peut pas être inclus dans un modèle économétrique directement. Ainsi, pour ces deux variables, il a été nécessaire de créer autant de variables dichotomiques que de modalités, pour les intégrer ensuite dans le modèles et en laissant une catégorie

en référence. En particulier, pour la variable d'âge, on crée les 4 variables dichotomiques telle que : $age_1 = 1$ si $age = 1$, 0 sinon ; $age_2 = 1$ si $age = 2$, 0 sinon ; $age_3 = 1$ si $age = 3$, 0 sinon ; $age_4 = 1$ si $age = 4$, 0 sinon. On s'assure au préalable qu'il y a assez de monde dans chaque catégorie (au moins 5% de l'échantillon). On intègre ensuite toutes les variables dichotomique - 1 qui sera la référence. Les interprétations qui en découlent seront faites en fonction de la catégorie de référence. On fait la même chose pour la variable de région.

5. On peut tout d'abord confirmer l'hypothèse selon laquelle le niveau de revenu avant le congé va diminuer la durée du congé parental. Par ailleurs, on remarque que cet effet n'est pas le même pour les hommes et pour les femmes. En particulier, on voit que pour les hommes, toutes choses étant égales par ailleurs, 1 euros de revenu supplémentaire entraîne une diminution de la durée de 0.04 semaines, alors que pour les femmes cette diminution n'est que de 0.011.

6. Toutes choses étant égales par ailleurs :
 - On voit que le montant de la prestation perçues entraîne une augmentation de la durée du congé : 1 euro de prestation supplémentaire entraîne une augmentation d'environ 0.06 pour les hommes et 0.02 pour les femmes ;
 - On voit que les femmes prennent en moyenne 3.77 semaines de plus que les hommes ;
 - On voit que les parents de 35 ans ou plus prennent en moyenne des congés plus courts que les autres. En particulier, on remarque que cet effet n'est pas le même pour les hommes et les femmes :
 - les hommes âgés de moins de 25 ans prennent en moyenne 3.25 semaines de moins que ceux de plus de 35 ans, contre un écart seulement de 0.21 pour les femmes,
 - les hommes entre 25 et 29 ans prennent en moyenne 1.95 semaines de moins que ceux de plus de 35 ans, contre un écart de seulement 0.14 pour les femmes,
 - et enfin les hommes entre 30 et 35 prennent en moyenne 0.48 semaines de moins que ceux de plus de 35 ans, contre une écart de 0.03 pour les femmes ;
 - On constate un effet de la région d'habitation : les parents qui vivent dans la région Est prennent en moyenne des congés plus long qu'ailleurs (y compris Nord et Ouest), alors que ceux vivant dans l'Ouest ou le Nord prennent des congés plus court que le Sud et l'Est, le sud étant situé entre les deux ;
 - On voit que le statut de travailleur des parents n'impacte pas la durée des congés demandée ;
 - Enfin, on voit que les parents qui prennent un congé seul dans le

couple, prennent en moyenne des congés plus longs de 1.64 semaines.

7. Prévision de la durée du congé :

$$X_i\beta = 18.11 + 1550 \times -0.039 + 437.5 \times 0.063 + 3.775 \times 1 + 1 \times -1.952 + 1 \times -0.287 + 1550 \times 0.028 + 437.5 \times -0.044 + 1 \times 1.819 = 12.72$$

La durée du congé prise par cette mère est estimée à 12.72 semaines.

8. On peut réécrire H0 sous forme matricielle : $R\beta = q$ avec :

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad q = \begin{bmatrix} 0 \end{bmatrix}$$

Sous H_0 , on a :

$$F_1 = \frac{\left(R\hat{\beta} - q\right)' \left[R(X'X)^{-1} R'\right]^{-1} \left(R\hat{\beta} - q\right) / c}{S^2} \sim F(c, N - k)$$

La valeur de la statistique F_1 est ainsi 0.34 avec une p-value associée au test de 0.5609. Donc on ne rejette pas H_0 au seuil de 5 %.

Ce résultat signifie qu'il n'y a pas de différence significative en termes de durée de congé entre les parents de la région 1 et 2.

9. Il faut tout d'abord réaliser 3 régressions distinctes : la première sur l'échantillon totale, la seconde uniquement sur le sous-échantillon des mères et la troisième uniquement sur le sous-échantillon des pères (on estime le modèle de base sans la variable de sexe et les termes d'interaction associés). A partir de ces trois régressions nous pourrons faire un test de Chow dans le but de tester la stabilité des paramètres entre les deux groupes (mères et pères).

Soit β_t , le vecteur des paramètres du modèle estimé sur l'échantillon total, β_h (resp. β_f), le vecteur des paramètres du modèle estimé sur l'échantillon des hommes (resp. femmes). L'hypothèse H_0 est $\beta_h = \beta_f = \beta$.

Sous H_0 , on a :

$$F_c = \frac{[SCR_t - (SCR_h + SCR_f)] / K_t}{(SCR_h + SCR_f) / (N_h - K_h + N_f - K_f)} \sim F(K, N - 2K)$$

$F_c > F_{0.05}(K, N - 2K)$, alors on rejette H_0 . Les coefficients ne sont pas homogènes entre les hommes et les femmes et il sera pertinent de distinguer l'analyse.

Les résultats de l'estimation reportés dans l'énoncé, et en particulier la significativité des termes d'interaction avec la variable de sexe nous permet déjà de savoir que le test conduira à rejeter H0.