

Enoncé

NB. Le travail est **individuel**. Toute ressemblance dans le dofile ou dans le document sera sévèrement sanctionnée. Toutes vos réponses doivent figurer dans un document **PDF** et un **dofile** intitulés comme suit « Prenoms_Nom_ISE2_2023 ». Toute statistique qui ne provient pas du dofile ne sera pas prise en compte. Veiller à tout consigner dans le dofile.

Partant des bases de données fournies, réaliser les tâches suivantes :
Parie

Partie 1 : modèle dichotomique simple (8 pts)

A. Veuillez télécharger les données du Round 9 de votre pays sur le site suivant <https://www.afrobarometer.org/data/data-sets/> » y compris le questionnaire

1. Veuillez construire une variable dichotomique qui permet d'identifier les personnes vaccinées contre la Covid-19. Expliquer (5 lignes)
2. Présenter un tableau descriptif des différentes variables que vous jugez pertinentes pour étudier les raisons de la vaccination. Commenter (*maximum 5 lignes*)
3. Estimer puis présenter dans un tableau synthétique les coefficients des modèles logit, probit et OLS du statut de **vaccination**. Lequel des trois semblent meilleurs ? Pourquoi ? (*maximum 10 lignes*)
4. Estimer les Odds ratio et les effets marginaux de la variable « **vaccination** » en utilisant un logit. Assurez-vous d'avoir effectué les différents tests de validation. Quel est le taux de bonne prédiction de votre modèle à 50%, 70% et 80% ? Interpréter les résultats (*maximum 20 lignes*)
5. Faire un tirage aléatoire stratifié (suivant le milieu de résidence) parmi les non vaccinés pour obtenir un nouvel échantillon dans lequel les personnes non vaccinées représenteraient 40%. (**NB.** Dans le cas où le taux de vaccination dépasse 40% choisir un seuil raisonnable à justifier et procéder au tirage)
6. Refaire l'estimation de la question 2. Comparer dans un tableau les coefficients des deux modèles logit. Commenter les résultats (*maximum 10 lignes*).
7. Faire une estimation séparée (ensemble, urbain, rural) et présenter dans un même graphique les résultats des odds ratio.
8. Y a-t-il une variable susceptible d'être endogène au modèle ? Si oui, quelle solution vous proposez ? (*maximum 5 lignes*)

B. Cette partie s'appuie sur la base de données « base_estimation_enfants.dta »

L'objectif est de modéliser l'effet de l'alphabétisation des femmes sur la santé infantile. Les indicateurs de santé infantile sont « stunted_growth ; Underweight ; emaciation »

- 1 Estimer séparément l'effet de la variable « literacy » sur les indicateurs de santé infantile. Proposer des variables de contrôle pour affiner vos résultats. Présenter les résultats dans un tableau synthétique. Commenter (*maximum 10 lignes*) [NB. *Il faudra présenter les tests de validation des modèles*]
- 2 Existe -t-il un lien entre les variables dépendantes ? Si oui, quel type de modèle permet de prendre en compte cela ? Estimer l'effet de la variable « literacy » sur la santé infantile dans le cadre d'un modèle qui prend en compte l'interdépendance potentielle entre les variables dépendantes. (*maximum 5 lignes*)

Partie 2 : modèle multinomial (6 pts)

Cette partie s'appuie sur la base de données « **base_contrat.dta** »

1. On s'intéresse à la nature du contrat des travailleurs captée par la variable « **type_contrat** ». Sous quelles conditions, il est possible de modéliser cette variable. (*maximum 5 lignes*)
2. Faire un tableau synthétique des différentes variables que vous jugez pertinentes. Commenter (*maximum 5 lignes*).
3. Estimer et présenter le modèle logit multinomial en fixant comme modalité de référence « CDI ». Commenter (*maximum 10 lignes*)
4. Estimer et présenter dans un tableau les risques relatifs. Commenter (*maximum 10 lignes*)
5. Estimer et présenter dans un tableau les effets marginaux. Commenter (*maximum 10 lignes*)
6. Effectuer le test IIA. Commenter (*maximum 5 lignes*)

Partie 3 : modèle Tobit et Heckman (7 pts)

Cette partie s'appuie sur la base de données « **base_labor_market_estimation.dta** »

1. On s'intéresse à identifier les déterminants de la rémunération captée par la variable « **revenu** » (*à prendre en logarithme*) dans la base de données. Quel problème de modélisation soulève cette variable ? (*maximum 5 lignes*)
2. Faire un tableau synthétique des différentes variables que vous jugez pertinentes pour le problème posé. Commenter (*maximum 10 lignes*).
3. Estimer un modèle linéaire sur l'ensemble de l'échantillon puis dans le sous-échantillon des travailleurs. Que constatez-vous ? Prédire les revenus pour les non travailleurs ? (*maximum 5 lignes*)
4. Estimer un modèle de sélection de Heckman avec les mêmes variables. Interpréter le ratio de Mills. Puis commenter les résultats. Prédire les revenus pour les non travailleurs ? Comparer les résultats des prédictions avec ceux de la question 3. Commenter (*maximum 10 lignes*)
5. Est-ce que le fait d'être marié et le fait d'avoir un enfant sont-elles endogènes ? Justifier ? Quelles solutions préconisez-vous s'il y en a ? (*maximum 5 lignes*)

Bon courage