École nationale de la statistique et de l'analyse économique

Agence Nationale de la Statistique et de la Démographie

Cycle Ingénieur Statisticien Économiste (ISE)

Économétrie des variables qualitatives

Rédigé par :

Crépin MEDEHOUIN

Élève ingénieur statisticien économiste ISE-2

Sous l'encadrement de :

M. Mamadou Abdoulaye DIALLO

Ingénieur Statisticien Économiste (ISE)

Chercheur au CRES [1]

Résumé

Ce document se divise en trois parties, chacune examinant respectivement le modèle dichotomique simple, le modèle multinomial, et le modèle Tobit et Heckman. Dans la première partie, nous analysons les données de la neuvième vague de l'enquête Afrobaromètre pour identifier les facteurs influençant la décision des adultes Béninois de se faire vacciner contre la Covid-19. Nous utilisons également la base de données "base_estimation_enfants.dta" pour étudier l'impact de l'alphabétisation des femmes sur la santé infantile. La deuxième partie, se concentrant sur le modèle multinomial et en utilisant la base de données "base_contrat.dta", examine la nature des contrats de travailleurs. En fin la dernière partie, mettant en exergue le modèle Tobit et Heckman, s'intéresse à identifier les déterminants de la rémunération en utilisant la base de données « base_labor_market_estimation.dta».

1. Consortium pour la Recherche Économique et Sociale

Introduction

L'économétrie des variables qualitatives explore les méthodes pour modéliser des variables catégorielles dans les analyses économétriques. Trois modèles principaux sont étudiés : le modèle dichotomique simple pour les variables binaires, le modèle multinomial pour les variables présentant plusieurs catégories, ainsi que les modèles Tobit et Heckman pour les variables sujettes à une censure ou à un biais de sélection. Ces approches fournissent des outils précieux pour appréhender les relations complexes entre les variables explicatives et les variables qualitatives, principalement dans les domaines de la sociologie et de la psychologie.

Partie 1 : modèle dichotomique simple

A) Données : Afrobarometer round 9, Bénin

A-1 Construction de la variable dichotomique covid_19

Nous construisons la variable dichotomique "*vaccination*" qui prendra 1 si l'individu est vacciné et 0 sinon. Dans cette base de données, la question *"Q58A. Avez-vous reçu une vaccination contre la COVID-19, une ou deux doses ?"*, nous permet d'identifier les personnes vaccinées. Cette question donne lieu à trois modalités qui sont : *Oui, Non* et *Refuse*.
La construction de la variable dichotomique "*vaccination*" se fait donc comme suite :

$$vaccination : \begin{cases} 1\,, & si\ la\ modalite\ de\ Q58A\ est\ Oui \\ 0\,, & si\ la\ modalite\ de\ Q58A\ est\ Non \end{cases}$$

A-2 Les variables susceptibles d'étudier les raisons de la vaccination

La variable dépendante dans cette étude était le statut de vaccination et peut être expliquée par l'ensemble des variables décrites ci-dessous. En tenant compte de la littérature, plusieurs variables ont été conservées pour expliquer le statut de vaccination des adultes Béninois. L'analyse dans le **Tableau 1** montre une dépendance de quatre variables explicatives (valeur p inférieure à 5%) et trois variables explicatives (valeur p inférieure à 15%) avec le statut de vaccination, à l'exception du niveau d'éducation. Comme nous nous intéressons aux signes attendus de chaque variable, nous réalisons que certaines variables peuvent être mélangées, telles que le sexe, l'âge, le niveau d'éducation, la confiance au gouvernement, la gestion de la pandémie par le gouvernement, l'équipement des établissements de santé, la radio comme source d'information et le milieu de résidence.

TABLE 1 – Statistiques descriptives des variables explicatives

Variables	χ^2 test	**Pourcentage (%)**	
		Statut de vaccination	
		Pas vacciné	Vacciné
Sexe	p_value = 0,117		
Homme		49,2	52,1
Femme		50,8	47,9
Âge (années)	p_value = 0,078		
18-25 ans		31,5	20,6
26-35 ans		26,5	28,8
36-45 ans		17	22,3
46-55 ans		15,2	15,3
56-65 ans		6,1	8,8
Plus de 66 ans		3,8	4,2
Niveau d'éducation	p_value = 0,394		
Aucune instruction		36,2	35,5
Primaire		26,1	25,1
Secondaire		28	27,7
Supérieur		9,8	11,7
Confiance au gouvernement	p_value = 0,000		
Pas de confiance		28,7	9,9
Un peu de confiance		62,5	60,1
Beaucoup de confiance		8,9	30
Gestion de la pandémie	p_value = 0,000		
Mal		43,5	20,1
Bien		46,8	60,9
Très bien		9,7	19,1
Établissements de santé	p_value = 0,100		
Pas du tout satisfait		15,7	12,1
Assez satisfait		70,7	73
Très satisfait		13,7	14,9
Radio comme source d'info	p_value = 0,001		
Pas vraiment		19,6	16,1
Un peu		36,1	31,4
Toujours		44,3	52,6
Type de milieu	p_value = 0,002		
Urbain		50,5	44,8
Rural		49,5	55,2

Source : Calcul à partir des données d'Afrobaromètre round 9

A-3 Les coefficients des modèles logit, probit et OLS du statut de vaccination

Nous savons que, dans le modèle à variable dépendante dichotomique, les coefficients estimés ne représentent pas, comme dans le modèle linéaire, l'effet partiel des variables explicatives sur la variable explicative. Cela place le modèle OLS dans une contrainte où nous pouvons observer que la probabilité qu' individu soit vacciné ou non soit en dehors de l'intervalle [0, 1]. En effet, il révèle des coefficient négatif et même certains qui pourrons excéder 1 suit à une accroissement d'une variable explicative de plusieurs unité. Par ailleurs, dans notre étude, le modèle probit semble être meilleur que le logit car il révèle en générale des Standard errors les plus faible. ***(Tableau 2)***

TABLE 2 – Les coefficients des modèles logit, probit et OLS du statut de vaccination

Variables	Modalités de références	Coeff Logit	Coeff Probit	Coeff OLS
Sexe	Homme			
Femme		-0.040	-0.024	-0.009
		(0.131)	(0.079)	(0.028)
Âge (années)	18 - 25 ans			
26-35 ans		0.483***	0.292***	0.105***
		(0.174)	(0.105)	(0.038)
36-45 ans		0.555***	0.335***	0.120***
		(0.192)	(0.116)	(0.041)
46-55 ans		0.418**	0.256**	0.092**
		(0.211)	(0.128)	(0.044)
56-65 ans		0.365	0.229	0.083
		(0.265)	(0.163)	(0.058)
Plus de 66 ans		0.623*	0.375*	0.131*
		(0.356)	(0.213)	(0.075)
Niveau d'éducation	Aucune instruction			
Primaire		0.045	0.026	0.009
		(0.169)	(0.102)	(0.036)
Secondaire		0.245	0.152	0.054
		(0.165)	(0.101)	(0.036)
Supérieur		0.467**	0.286**	0.096*
		(0.236)	(0.143)	(0.050)
Confiance au gouvernement	Pas de confiance			
Un peu de confiance		0.991***	0.600***	0.218***
		(0.179)	(0.107)	(0.036)
Beaucoup de confiance		2.153***	1.294***	0.458***
		(0.246)	(0.144)	(0.046)
Gestion de la pandemie	Mal			
Bien		0.498***	0.308***	0.113***
		(0.148)	(0.091)	(0.034)
Très bien		0.631***	0.392***	0.138***
		(0.230)	(0.138)	(0.049)
Établissement de santé	Pas du tout satisfait			
Assez satisfait		0.166	0.089	0.032
		(0.183)	(0.111)	(0.039)
Très satisfait		-0.137	-0.091	-0.034
		(0.247)	(0.149)	(0.053)

Variables	Modalités de références	Coeff Logit	Coeff Probit	Coeff OLS
Information par la radio	Pas vraiment			
Un peu		0.096	0.051	0.022
		(0.183)	(0.111)	(0.040)
Toujours		0.475***	0.281***	0.102***
		(0.177)	(0.106)	(0.038)
Type de milieu	Urbain			
Rural		0.345***	0.210***	0.072**
		(0.130)	(0.079)	(0.028)
Observations		1200	1200	1200
R^2 *McFadden*		0.117	0.116	0.150
Robust sd in parentheses				
$***p < 0.01, **p < 0.05, *p < 0.1$				

Source : *Calcul* à *partir des données d'Afrobaromètre round* 9

A-4 Les Odds ratios et les effets marginaux de la variable « vaccination»

Les résultats du modèle de régression logistique sont présentés dans **le Tableau 3**. Le modèle est globalement significatif ($Prob > chi2 = 0.000$), et le test de Hosmer-Lemeshow accepte l'hypothèse nulle avec une bonne spécification et une probabilité de 0.3671. La zone sous la courbe des caractéristiques opérationnelles du récepteur (ROC) est de 0.7269, et le modèle a un taux de classification correct de 66.58 %.

TABLE 3 – Odds ratios et les effets marginaux

Variables	Modalités de références	Modèle logit	
		Odds Ratios	Effets marginaux
Sexe	Homme		
Femme		0.961	-0.00852
		(0.125)	(0.0277)
Âge (années)	18 - 25 ans		
26-35 ans		1.621***	0.103***
		(0.282)	(0.0367)
36-45 ans		1.742***	0.118***
		(0.334)	(0.0404)
46-55 ans		1.519**	0.0891**
		(0.321)	(0.0447)
56-65 ans		1.441	0.0778
		(0.382)	(0.0565)
Plus de 66 ans		1.865*	0.133*
		(0.663)	(0.0743)
Niveau d'éducation	Aucune instruction		
Primaire		1.046	0.00951
		(0.177)	(0.0359)
Secondaire		1.278	0.0520
		(0.210)	(0.0349)
Supérieur		1.596**	0.0985**
		(0.376)	(0.0489)

Variables	Modalités de références	Modèle logit	
		Odds ratios	Effets marginaux
Confiance au gouvernement	Pas de confiance		
Un peu de confiance		2.693***	0.223***
		(0.481)	(0.0370)
Beaucoup de confiance		8.614***	0.471***
		(2.122)	(0.0463)
Gestion de la pandémie	Mal		
Bien		1.645***	0.109***
		(0.243)	(0.0327)
Très bien		1.879***	0.138***
		(0.432)	(0.0505)
Établissement de santé	Pas du tout satisfait		
Assez satisfait		1.181	0.0352
		(0.216)	(0.0388)
Très satisfait		0.872	-0.0289
		(0.216)	(0.0522)
Information par la radio	Pas vraiment		
Un peu		1.101	0.0205
		(0.202)	(0.0392)
Toujours		1.608***	0.101***
		(0.285)	(0.0377)
Type de milieu	Urbain		
Rural		1.412***	0.0735***
		(0.183)	(0.0276)
$Observations$		$1,200$	
$Cragg-Uhler/Nagelkerke$		0.199	
$Correct\ prediction\ rate$		66.58%	
$Area\ under\ ROC\ curve$		0.7269	
$Hosmer-Lemeshow\ test$		0.3671	
$R^2\ McFadden\ (adjusted)$		0.094	
$Wald\ \chi^2$		149.01	
$prob > \chi^2$		0.000	
$Robust\ see\ form\ in\ parentheses$			
*** $p < 0.01, ** p < 0.05, *p < 0.1$			

$Source$: $Calcul$ à $partir\ des\ données\ d'Afrobaromètre\ round\ 9$

L'analyse des résultats révèle que l'âge a un effet significatif sur la probabilité d'être vaccinée. Les personnes âgées de 26 à 35 ans et de 36 à 45 ans respectivement sont 1,6 fois et 1,7 fois plus susceptibles de se faire vacciner que les individus âgés de 18 à 25 ans. En ce qui concerne l'effet d'avoir confiance au gouvernement pour assurer l'efficacité d'un vaccin du COVID-19, il semble que la probabilité de se faire vacciner soit plus faible pour les personnes n'ayant aucune confiance. Par exemple, ceux ayant beaucoup confiance ont 8,8 fois plus de chance de se faire vacciner par rapport à ceux n'ayant aucune confiance. Ce degré diminue avec la confiance, les personnes peu confiantes ont 2,7 fois plus de chance d'être vaccinées par rapport à ceux n'ayant aucune confiance. Pareillement aux personnes n'ayant pas de confiance au gouvernement, les personnes pensant que la gestion de la réponse à pandémie est mal gérée sont également moins susceptibles d'être vacciner. Par ailleurs, l'écoute régulière de la radio donne une chance de 1,6 fois d'être vaccinée par rapport ceux n'écoutant pas la radio.

En fin, il ressort que le milieu rural révèle 1,4 fois plus de chance d'être vaccinée que le milieu urbain.
En ce qui concerne les effets marginaux, on sait que l'effet marginal d'une variable explicative sur la probabilité de l'évènement être vaccinée, est la variation de la probabilité suite à l'accroissement de la variable explicative d'une unité. Par exemple, nous constatons qu'au passage de la tranche des plus jeunes à celle de 26 à 35 ans, la probabilité d'être vacciné augmente de 10 %, et cette probabilité augmente de 12 % en passant des toujours des plus jeunes à la tranche 36 à 45 ans.

Par ailleurs, les résultats de prédiction varient selon les seuils choisis : à un seuil de 50%, le taux de bonne prédiction est de 66.58%, tandis qu'à des seuils plus élevés de 70% et 80%, les taux de bonne prédiction sont respectivement de 59.58% et 55.25% *(voir do-file)*.

A-5 Tirage aléatoire stratifié (suivant le milieu de résidence)

Dans notre cas, les personnes vaccinées représentent une proportion de 52%. Nous allons donc faire le tirage aléatoire parmi les non vaccinées de telle sorte que notre nouvel échantillon soit composé de 60% des personnes vaccinées. Etant donné que le nombre des personnes vaccinées est 624 et représentent 60% de notre échantillon, on obtient donc 416 pour ceux non vaccinées. Notre nouvelle d'échantillon est donc $n = 1040$ et composé de 60% des personnes vaccinées et 40% des personnes non vaccinées.
La stratification selon le milieu se fera comme suit : on observe 51.91% des personnes non vaccinées en milieu urbain et 48.09% en milieu rural. On reparti les 416 des personnes non vaccinées à tirer proportionnellement aux 51.91% et 48.09% selon le milieu urbain et rural respectivement.
On obtient au final 216 personnes non vaccinées en milieu urbain et 200 personnes non vaccinées en milieu rural. Soit un total de 416 personnes non vaccinées qui représentent 40% de nouvelle échantillon et stratifiée en milieu de résidence.

A-6 Nouveau vs ancien modèle : comparaison

TABLE 4 – Nouveau vs ancien modèle modèle logit

Variables	**Modalités de références**	**Logit**	
		Coefficient	
		Nouveau modèle	Ancien modèle
Sexe	Homme		
Femme		-0.039	-0.040
		(0.143)	(0.131)
Âge (années)	18 - 25 ans		
26-35 ans		0.467***	0.483***
		(0.188)	(0.174)
36-45 ans		0.573***	0.555***
		(0.209)	(0.192)
46-55 ans		0.448*	0.418**
		(0.230)	(0.211)
56-65 ans		0.524*	0.365
		(0.296)	(0.265)
Plus de 66 ans		0.759*	0.623*
		(0.395)	(0.356)

Variables	**Modalités de références**	**Logit**	
		Coefficient	
		Nouveau modèle	Ancien modèle
Niveau d'instruction	Aucune instruction		
Primaire		0.153	0.045
		(0.186)	(0.169)
Secondaire		0.244	0.245
		(0.180)	(0.165)
Supérieur		0.518**	0.467**
		(0.263)	(0.236)
Confiance au gouvernement	Pas de confiance		
Un peu de confiance		1.065***	0.991***
		(0.189)	(0.179)
Beaucoup de confiance		2.145***	2.153***
		(0.267)	(0.246)
Gestion de la pandemie	Mal		
Bien		0.467***	0.498***
		(0.161)	(0.148)
Très bien		0.594**	0.631***
		(0.252)	(0.230)
Etablissement de santé	Pas du tout satisfait		
Assez satisfait		0.151	0.166
		(0.201)	(0.183)
Très satisfait		-0.110	-0.137
		(0.269)	(0.247)
Information par la radio	Pas vraiment		
Un peu		0.018	0.096
		(0.200)	(0.183)
Toujours		0.438**	0.475***
		(0.194)	(0.177)
Type de milieu	Urbain		
Rural		0.344**	0.345***
		(0.142)	(0.130)
$Observations$		$1,040$	$1,200$
$Cragg-Uhler/Nagelkerke$		0.201	0.199
$Correct\ prediction\ rate$		68.37%	66.58%
$Area\ under\ ROC\ curve$		0.7293	0.7269
$Hosmer-Lemeshow\ test$		0.3346	0.3671
$R^2\ McFadden\ (adjusted)$		0.074	0.094
$wald\ \chi^2$		144.2	149.01
$prob > \chi^2$		0.000	0.000
$Robust\ seeform\ in\ parentheses$			
$***p < 0.01, **p < 0.05, *p < 0.1$			

$Source$: $Calcul$ à $partir\ des\ données\ d'Afrobaromètre\ round\ 9$

Les deux modèles de régression logistique sont globalement significatif ($Prob > chi2 = 0.000$), et les tests de Hosmer-Lemeshow des modèles acceptent l'hypothèse nulle avec une bonne spécification et une probabilité presque identique. La ressemblance s'observent au niveau de la zone sous la courbe des caractéristiques opérationnelles du récepteur (ROC) qui est de 0.7293 pour nouveau modèle ce qui est légèrement supérieur à celui de l'ancien modèle 0.7269.

Les différences s'observent également au niveau du taux de classification correct qui semble s'améliorer pour le nouveau modèle : 68.37% contre de 66.58%.
Il en ressort dans la globalité que le nouveau modèle semble plus prédictif voire significatif que l'ancien modèle. Néanmoins, cette amélioration est légère car la variable d'intérêt était initialement presque bien repartie dans notre échantillon, soit 52% contre 48%.
On peut donc conclure que notre échantillon était en amont équilibré car la différence observée après rééchantillonnage par stratification n'est pas grande.

A-7 Estimation séparée (ensemble, urbain, rural)

Dans **le tableau 5**, il est présenté les odds ratios de l'estimation du milieu de résidence, du milieu urbain et du milieu rural.

TABLE 5 – Estimation séparée (ensemble, urbain, rural)

Variables	**Modalités de références**	**Modèle logit**		
		Ensemble Odds Ratios	Urbain Odds Ratios	Rural Odds Ratios
Sexe	Homme			
Femme		0.961	1.090	0.833
		(0.125)	(0.207)	(0.152)
Âge (années)	18 - 25 ans			
26-35 ans		1.621***	1.320	1.999***
		(0.282)	(0.344)	(0.473)
36-45 ans		1.742***	1.493	2.124***
		(0.334)	(0.419)	(0.570)
46-55 ans		1.519**	1.393	1.555
		(0.321)	(0.412)	(0.469)
56-65 ans		1.441	1.095	1.715
		(0.382)	(0.445)	(0.615)
Plus de 66 ans		1.865*	1.708	1.867
		(0.663)	(0.856)	(0.954)
Niveau d'education	Aucune instruction			
Primaire		1.046	0.984	1.041
		(0.177)	(0.246)	(0.247)
Sécondaire		1.278	1.152	1.404
		(0.210)	(0.286)	(0.315)
Supérieur		1.596**	1.482	1.698
		(0.376)	(0.449)	(0.705)
Confiance au gouvernement	Pas de confiance			
Un peu de confiance		2.693***	2.222***	3.169***
		(0.481)	(0.576)	(0.812)
Beaucoup de confiance		8.614***	6.429***	11.354***
		(2.122)	(2.312)	(3.899)
Gestion de la pandémie	Mal			
Bien		1.645***	2.322***	1.241
		(0.242)	(0.499)	(0.256)
Très bien		1.879***	2.514**	1.529
		(0.432)	(0.897)	(0.467)

Variables	**Modalités de références**	**Modèle logit**		
		Ensemble Odds Ratios	Urbain Odds Ratios	Rural Odds Ratios
Établissement de santé	Pas du tout satisfait			
Assez satisfait		1.181	1.223	1.146
		(0.216)	(0.327)	(0.299)
Très satisfait		0.872	1.155	0.676
		(0.216)	(0.412)	(0.236)
Information par la radio	Pas vraiment			
Un peu		1.101	1.183	1.001
		(0.202)	(0.304)	(0.266)
Toujours		1.608***	1.923**	1.354
		(0.285)	(0.486)	(0.345)
$Observations$		$1,200$	568	632
$Cragg-Uhler/Nagelkerke$		0.199	0.198	0.203
$Correct\ prediction\ rate$		66.58%	66.20%	68.35%
$Area\ under\ ROC\ curve$		0.7269	0.7270	0.7317
$Hosmer-Lemeshow\ test$		0.3671	0.2071	0.3923
$R^2\ McFadden\ (adjusted)$		0.094	0.071	0.078
$wald\ \chi^2$		149	72.43	80.92
$prob > \chi^2$		0.000	$1.13e-08$	$3.71e-10$
$Robust\ seeform\ in\ parentheses$				
$***p<0.01, **p<0.05, *p<0.1$				

$Source$: $Calcul$ à $partir\ des\ données\ d'Afrobaromètre\ round\ 9$

Il en ressort que la significativité de la variable âge dans le modèle global est influencée par le milieu rural. En effet, l'âge n'a pas un effet significatif sur la probabilité d'être vacciné en milieu urbain. En revanche, l'utilisation quotidienne de la ratio comme source d'information est plutôt signification en milieu urbain qu'en milieu rural.

A-8 Variable susceptible d'être endogène au modèle ?

Après avoir soumis notre modèle à un test d'endogénéité, il ressort que nous n'avons pas identifié de variable susceptible d'être endogène *(voir do-file)*. Cette constatation est essentielle pour garantir la validité de notre analyse et la robustesse de nos résultats. Toutefois, dans le cas où une variable aurait été jugée endogène, nous aurions exploré différentes solutions pour remédier à ce problème, telles que l'utilisation d'instruments supplémentaires.

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

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 »

B-1 Estimer séparément l'effet de la variable « literacy » sur les indicateurs de santé infantile. Proposer des variables de contrôle pour affiner vos résultats.

TABLE 6 – Estimation séparée de l'effet de la variable « literacy » sur les indicateurs de santé

Variables	Modalités de références	Modèle logit (Echantillon équilibré)					
		stunted_growth	Underweight	emaciation	stunted_growth	Underweight	emaciation
Literacy	No						
Yes		-0.429***	-0.485***	-0.212*	-0.157	-0.290**	-0.197
		(0.086)	(0.095)	(0.115)	(0.105)	(0.116)	(0.153)
Health insurance	No						
Yes					-0.289**	-0.225	-0.054
					(0.143)	(0.159)	(0.202)
Current marital status	Not maried						
Married/Cohabiting					-0.055	-0.338	-0.309
					(0.214)	(0.225)	(0.294)
Getting medical help for self	No problem						
Big problem					-0.002	0.092	-0.009
					(0.088)	(0.100)	(0.140)
Number of living children					-0.014	-0.021	-0.046
					(0.020)	(0.021)	(0.030)
wealth index combined	Poorest						
Poorer					-0.227**	-0.493***	-0.588***
					(0.100)	(0.114)	(0.163)
Middle					-0.826***	-0.818***	-0.700***
					(0.129)	(0.144)	(0.199)
Richer					-0.597***	-0.555***	-0.097
					(0.156)	(0.165)	(0.228)
Richest					-1.250***	-1.484***	-0.471
					(0.225)	(0.254)	(0.296)
source of drinking water							
Yes	No				-0.248***	-0.123	0.160
					(0.087)	(0.099)	(0.137)
Milieu	Urbain						
Rural					0.105	-0.014	0.066
					(0.119)	(0.129)	(0.175)
Observations		5,045	5,045	5,045	2,888	2,362	1,172
Cragg-Uhler/Nagelkerke		0.008	0.009	0.002	0.076	0.062	0.034
Correct prediction rate		79.96%	83.61%	90.70%	65.27%	64.94%	60.49%
Area under ROC curve		0.5387	0.5429	0.5429	0.6394	0.6254	0.5908
Hosmer-Lemeshow test		.	.	.	0.3030	0.4979	0.1349
R2 McFadden		0.00517	0.00620	0.00112	0.0366	0.0377	0.0186
wald chi2		24.77	26.02	3.387	121.9	103	28.31
prob>chi2		6.46e-07	3.39e-07	0.0657	0	0	0.00290
Robust sd in parentheses							
*** p<0.01, ** p<0.05, * p<0.1					Échantillon	équilibré	(Voir do-file)

Source : Calcul des auteurs à partir des données DHS-2019

Le Tableau 6 présente une analyse de l'association entre le niveau d'alphabétisation des femmes, ainsi que d'autres variables de contrôle, avec les indicateurs de santé infantile. Selon les résultats obtenus, il est observé que l'alphabétisation des femmes est associée à une réduction du risque de retard de croissance et de sous-poids chez les enfants. De plus, une corrélation significative est mise en évidence entre l'indice de richesse combiné et les indicateurs de santé infantile, suggérant que plus le niveau de richesse est élevé, moins les risques de retard de croissance et de sous-poids

sont présents chez les enfants. Par ailleurs, l'analyse révèle que seules la source d'eau potable et la possession d'une assurance santé ont un impact significatif sur le retard de croissance. Ainsi, la disponibilité d'eau potable et le fait d'avoir une assurance santé sont associés à une réduction du risque de retard de croissance chez les enfants.

B-2 Existe -t-il un lien entre les variables dépendantes ?

Variables	$\chi^2 test$	
	emaciation	
stunted_growth	$Pearson\ \chi^2(1) = 31,3973$	$Pr = 0,000$
Underweight	$Pearson\ \chi^2(1) = 399,3187$	$Pr = 0,000$

On conclut donc qu'il y a une liaison entre les variables car la p-value est inférieure au seul de 5%. Le modèle qui permet de prendre en compte cela est la regression multiple multivariée. Car il permet de modéliser avec plusieurs variables dépendantes.

Partie 2 : modèle multinomial

Cette partie s'appuie sur la base de données « base_contrat.dta ». On s'intéresse à la nature du contrat des travailleurs captée par la variable « type_contrat ».

2.1 Sous quelles conditions, il est possible de modéliser cette variable.

Pour modéliser cette variable, l'hypothèse de l'Indépendance des Irrelevant Alternatives (IIA) doit être confirmée, ce qui implique que le choix du contrat ne soit pas influencé par des caractéristiques autres que celles propres à chaque individu.

2.2 Tableau synthétique des différentes variables pertinentes

TABLE 7 – Variables explicatives

Variables	**Description**	$\chi^2 test$
Sexe	1. Masculin ; 2. Feminin	0.000
Âge (années)	De 6 à 99 ans	0.000
Situation matrimoniale	1. Marié ; 2. Célibataire ; 3. Divorcé/ Veuf	0.000
Milieu de residence	1. Urbain ; 2. Rural	0.000
Niveau d'instruction	1. Sans instruction ; 2. Primaire ; 3. Moyen ; 4.Secondire ; 5.Supérieur	0.000
Malade	1. Oui ; 2.Non	0.028

Dans cette étude, la variable principale étudiée était le type de contrat, qui était influencé par les variables décrites précédemment. Toutes ces variables ont été retenues pour expliquer les variations observées dans le type de contrat. L'analyse présentée dans le **tableau 7** indique une corrélation significative (p < 5%) entre les variables explicatives et le type de contrat.

2.3 Estimer et présenter le modèle logit multinomial

TABLE 8 – Estimation du modèle logit multinomial : coefficients

Variables	**Modalités de références**	**Modèle miltimonial : coefficient**		
		CDD	Contrat de prestation de service	Sans contrat
	CDI (base outcome)			
Sexe	Homme			
Fiminin		0.212	-0.006	0.510***
		(0.158)	(0.165)	(0.133)
Age (Années)		-0.029***	-0.034***	-0.034***
		(0.007)	(0.007)	(0.006)
Situation matrimoniale	Marié			
Célibataire		0.644***	0.695***	1.138***
		(0.185)	(0.193)	(0.159)
Divorcé/veuf		0.156	0.255	0.461*
		(0.332)	(0.328)	(0.261)
Milieu de résidence	Urbain			
Rural		-0.055	0.799***	0.512***
		(0.207)	(0.185)	(0.163)
Niveau d'instruction	Sans instruction			
Primaire		-0.649**	-0.877***	-1.053***
		(0.256)	(0.227)	(0.198)
Moyen		-1.116***	-2.180***	-2.945***
		(0.233)	(0.227)	(0.187)
Secondaire		-1.783***	-3.390***	-4.636***
		(0.232)	(0.254)	(0.206)
Supirieur		-2.120***	-4.162***	-5.935***
		(0.251)	(0.334)	(0.295)
Malade	Oui			
Non		-0.100	-0.483***	-0.476***
		(0.154)	(0.154)	(0.127)
Observations		3994	3994	3994
Pseudo R2		0.230	0.230	0.230
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

L'analyse des résultats met en lumière l'impact significatif du sexe sur le type de contrat. Contrairement aux hommes, les femmes ont plus de chances d'être sans contrat que d'avoir un CDI.

Concernant l'âge, il est observé que plus celui-ci augmente, moins les individus ont de chances d'avoir un contrat à durée déterminée (CCD), un contrat de prestation de service, ou d'être sans contrat, par rapport à un CDI. Cependant, les célibataires, contrairement aux personnes mariées, ont plus de chances d'avoir un CCD, un contrat de prestation de service, ou d'être sans contrat, plutôt qu'un CDI.

Par ailleurs, il apparaît que les habitants des zones rurales, contrairement à ceux des zones urbaines, ont plus de chances d'avoir un contrat de prestation de service ou d'être sans contrat, que d'avoir un CDI. En revanche, les individus en bonne santé ont moins de chances d'avoir un contrat de prestation de service ou d'être sans contrat, par rapport à ceux qui sont malades.

Enfin, contrairement aux personnes non instruites, celles qui ont reçu une instruction ont moins de chances d'avoir un CCD, un contrat de prestation de service, ou d'être sans contrat, que d'avoir un CDI.

2.4 Estimer et présenter dans un tableau les risques relatifs

TABLE 9 – Estimation du modèle logit multinomial : risques relatifs

Variables	**Modalités de références**	**Modèle miltimonial : risques relatifs**		
		CDD	Contrat de prestation de service	Sans contrat
	CDI (base outcome)			
Sexe	Homme			
Fiminin		1.236	0.994	1.664***
		(0.195)	(0.163)	(0.220)
Age (Années)		0.972***	0.967***	0.967***
		(0.006)	(0.006)	(0.005)
Situation matrimoniale	Marié			
Célibataire		1.904***	2.004***	3.121***
		(0.370)	(0.385)	(0.515)
Divorcé/veuf		1.168	1.291	1.586*
		(0.377)	(0.423)	(0.397)
Milieu de résidence	Urbain			
Rural		0.947	2.223***	1.668***
		(0.191)	(0.390)	(0.252)
Niveau d'instruction	Sans instruction	0.522**	0.416***	0.349***
Primaire		(0.136)	(0.097)	(0.071)
		0.328***	0.113***	0.053***
Moyen		(0.078)	(0.027)	(0.010)
		0.168***	0.034***	0.010***
Secondaire		(0.040)	(0.009)	(0.002)
		0.120***	0.016***	0.003***
Supirieur		(0.030)	(0.005)	(0.001)
Malade	Oui	0.905	0.617***	0.621***
Non		(0.137)	(0.094)	(0.077)
		(0.151)	(0.152)	(0.124)
Observations		3994	3994	3994
Pseudo R2		0.230	0.230	0.230
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

L'analyse des résultats indique que le sexe influence considérablement le type de contrat. Les femmes ont 1,66 fois plus de chances d'être sans contrat que d'avoir un CDI, contrairement aux hommes. De plus, l'âge est corrélé à une augmentation de la probabilité d'avoir un CDI.

En d'autres termes, avec l'âge, la probabilité de ne pas avoir de contrat ou d'avoir un CDD diminue d'environ 3 % par rapport à celle d'avoir un CDI.

De plus, les célibataires, contrairement aux personnes mariées, présentent respectivement des probabilités plus élevées de 1,9 %, 2 % et 3% d'obtenir un contrat CDD, un contrat de prestation de service ou de ne pas avoir de contrat par rapport à un CDI. De même, les individus résidant en milieu rural, contrairement à ceux habitant en milieu urbain, ont des chances plus élevées respectivement de 2,2 % et 1,6 % d'obtenir un contrat de prestation de service ou de ne pas avoir de contrat par rapport à un CDI. Cependant, par opposition aux personnes malades, celles en bonne santé ont 40 % de chances en moins d'obtenir un contrat de prestation de service ou de ne pas avoir de contrat par rapport à un CDI.

2.5 Estimer et présenter dans un tableau les effets marginaux

TABLE 10 – Estimation du modèle logit multinomial : effets marginaux

Variables	**Modalités de références**	**Modèle multimonial : effets marginaux**			
		CDI	CDD	Contrat de prestation de service	Sans contrat
Sexe					
Femme	Homme	-0.0315***	-0.0149	-0.0472***	0.0936***
		(0.00957)	(0.0105)	(0.0124)	(0.0180)
Âge (Années)		0.00270***	0.000155	-0.000453	-0.00240***
		(0.000403)	(0.000478)	(0.000517)	(0.000754)
Situation matrimoniale	Mairé				
Célibataire		-0.0742***	-0.0270**	-0.0315**	0.133***
		(0.0105)	(0.0128)	(0.0143)	(0.0212)
Divorcé/Veuf		-0.0284*	-0.0187	-0.0148	0.0619*
		(0.0147)	(0.0199)	(0.0250)	(0.0332)
Milieu de résidence	Urbain				
Rural		-0.0368***	-0.0430***	0.0475***	0.0323
		(0.0100)	(0.0114)	(0.0152)	(0.0200)
Niveau d'éducation	Sans instruction				
Primaire		0.0973***	0.0218	-0.000750	-0.118***
		(0.0227)	(0.0167)	(0.0150)	(0.0248)
Moyen		0.384***	0.102***	-0.0257*	-0.461***
		(0.0357)	(0.0234)	(0.0156)	(0.0251)
Secondaire		0.653***	0.0757***	-0.0746***	-0.654***
		(0.0312)	(0.0232)	(0.0118)	(0.0158)
Supérieur		0.764***	0.0461**	-0.0978***	-0.712***
		(0.0267)	(0.0223)	(0.00983)	(0.0113)
Malade	Oui				
Non		0.0335***	0.0271***	-0.0104	-0.0502***
		(0.00871)	(0.0104)	(0.0132)	(0.0182)
Observations		3,994	3,994	3,994	3,994
Robust standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					

Les effets marginaux révèlent que plus l'âge augmente la probabilité d'avoir un contrat CDI augmente de 0,27% et la probabilité d'être sans contrat diminue de 0,24%. Nous constatons qu'en partant du sexe masculin au sexe féminin, la probabilité d'avoir un contrat CDI diminue de 3% et de 4% pour un contrat de prestation de service. Mais cela augmente la probabilité d'être sans contrat de 9%.

En outre, quitter de marié à célibataire, diminue la probabilité d'avoir un contrat CDI, de CDD et un contrat de prestation de service respectivement de 7% , 2,7% et 3%. En revanche cela augmente la probabilité d'être sans contrat de 13,33%.

De même, passer du milieu urbain au milieu rural, diminue aussi la probabilité d'avoir un contrat CDI, de CDD et un contrat de prestation de service respectivement de 3,7% , 4% et 4,8%.

Il ressort aussi que passer d'une aucune éducation à une instruction augmente la probabilité d'avoir un contrat de CDI et cette probabilité s'accroit avec le niveau d'éducation soit 9,7% pour le niveau primaire, 38,4% pour le niveau moyen, 65,3% pour le niveau secondaire et 76,4% pour le niveau supérieur. Par ricochet, cela diminue la probabilité d'être sans contrat de 11,8% pour le niveau primaire, 46% pour le niveau moyen, 65,4% pour le niveau secondaire et 71% pour le niveau supérieur.

2.6 Effectuer le test IIA

Hausman tests of IIA assumption (N=3994)

	chi2	df	P>chi2
CDI	-64.884	22	.
CDD	1.129	21	1.000
CPS	6.753	21	0.999
Sans contrat	18.880	22	0.653

suest-based Hausman tests of IIA assumption (N=3994)

	chi2	df	P>chi2
CDI	34.716	22	0.041
CDD	30.321	22	0.111
CPS	21.220	22	0.507
Sans contrat	37.950	22	0.019

Small-Hsiao tests of IIA assumption (N=3994)

				df	
CDI	-1041.096	-1030.980	20.233	22	0.568
CDD	-1073.897	-1064.367	19.061	22	0.642
CPS	-944.493	-936.871	15.243	22	0.852
Sans contrat	-632.900	-619.522	26.756	22	0.221

Les tests indiquent que l'Indépendance des Alternatives Irrélevantes (IIA) n'a pas été violée, ce qui suggère que les choix des individus restent stables indépendamment de la présence ou de l'absence d'autres alternatives.

Partie 3 : Modèle Tobit et Heckmanl

Cette partie s'appuie sur la base de données « base_labor_market_estimation.dta» On s'intéresse à identifier les déterminants de la rémunération captée par la variable « revenu » (à prendre en logarithme) dans la base de données.

3.1 Quel problème de modélisation soulève cette variable ?

La modélisation de cette variable soulève le problème de censure. En effet, on a une absence de données pour certains individus pour la variable dépendante, elle est donc censurée.

3.2 Tableau synthétique des différentes variables pertinentes pour le problème posé

TABLE 11 – Tableau synthétique des variables explicatives

Variables	Pourcentage (%)	Variables	Pourcentage (%)
Sexe		Années d'éducation	
Homme	50.69	0	25.08
Femme	49.31	6	23.31
Nombre d'enfant		10	19.62
0	60.23	13	18.38
1	14.62	16	9.92
2	12.00	18	3.15
3	6.08	21	0.54
4	3.77	Type de milieu	
5	1.85	Rurale	51.00
6	1.00	Urbaine	49.00
7	0.23	Niveau d'éducation	
8	0.08	Aucune éducation formelle	25.08
9	0.15	Primaire	23.31
Années d'expérience		Moyen	19.62
0	51.77	Secondaire	18.38
1	9.38	Supérieur	13.62
2	7.08	RÉGION	
3	7.31	DAKAR	27.15
4	5.85	DIOURBEL	14.77
5	6.69	FATICK	4.92
6	2.69	KAOLACK	9.00
7	2.54	KOLDA	4.77
8	1.54	LOUGA	5.77
9	1.00	SAINT LOUIS	11.08
10	2.46	TAMBACOUNDA	7.38
11	0.62	THIES	15.15
12	1.00		
13	0.08		

3.3 Estimer un modèle linéaire sur l'ensemble de l'échantillon puis dans le sous- écchantillon des travailleurs

TABLE 12 – Estimation du un modèle linéaire sur l'ensemble et sous échantillon

Variables	**Modalités de références**	**Ensemble**	**Sous-ensemble**
		Coeff	Coeff
Sexe	Homme		
Femme		-0.483*	-0.352***
		(0.269)	(0.074)
Âge		0.112***	0.030***
		(0.028)	(0.003)
Nombre d'enfant		0.014	-0.090***
		(0.097)	(0.030)
Dépendance		1.937***	0.200***
		(0.293)	(0.068)
Année d'expérience		0.942***	0.023**
		(0.047)	(0.011)
Type de milieu	Rural		
Urbain		0.778**	0.156*
		(0.302)	(0.089)
REGION	DAKAR		
DIOURBEL		0.590	-0.119
		(0.461)	(0.110)
FATICK		-0.923	-0.395**
		(0.606)	(0.173)
KAOLACK		-0.918**	-0.229
		(0.449)	(0.142)
KOLDA		-0.074	-0.110
		(0.577)	(0.138)
LOUGA		-0.922	-0.341**
		(0.590)	(0.162)
SAINT LOUIS		-0.764	-0.251**
		(0.473)	(0.123)
TAMBACOUNDA		-1.352**	-0.286
		(0.533)	(0.193)
THIES		-0.024	-0.239**
		(0.407)	(0.112)
Observations		1300	685
Pseaudo R2		0.389	0.180
Robust standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Les modèles montrent des variations de significativité des coefficients. Par exemple, le sexe affecte significativement la rémunération dans le sous-échantillon du modèle mais semble ne pas être significatif dans l'ensemble des travailleurs. De même, la variable nombre présente une similarité de résultats. Cette différence de significativité se retrouve également au niveau de certaine région. Pour la prédiction des revenus pour les non travailleurs : (*voir do-file*)

3.4 Estimer un modèle de sélection de Heckman avec les mêmes variables. Interpréter le ratio de Mills.

TABLE 13 – Modèle de sélection de Heckman

Variables	**Modalités de références**	**Modèle de sélection de Heckman**		
		Revenu Coeff	Travailler Coeff	Ratio de mills Coeff
Sexe	Homme			
Femme		-0.356***		
		(0.0737)		
Âge		0.0301***		
		(0.00739)		
Nombre d'enfant		-0.0973***		
		(0.0309)		
Dépendance		0.198***		
		(0.0689)		
Année d'expérience		0.0227**		
		(0.0106)		
Type de milieu	Rural			
Urbain		0.157*		
		(0.0821)		
RÉGION	DAKAR			
DIOURBEL		-0.123		
		(0.116)		
FATICK		-0.399**		
		(0.185)		
KAOLACK		-0.229*		
		(0.131)		
KOLDA		-0.110		
		(0.156)		
LOUGA		-0.336*		
		(0.176)		
SAINT-LOUIS		-0.249**		
		(0.116)		

Variables	**Modalités de références**	**Modèle de sélection de Heckman**		
		Revenu Coeff	Travailler Coeff	Ratio de mills Coeff
RÉGION	DAKAR			
TAMBACOUNDA		-0.290*		
		(0.157)		
THIES		-0.244**		
		(0.101)		
Année d'éducation			0.0931**	
			(0.0380)	
Niveau d'éducation			-0.403**	
			(0.159)	
Problème de santé			-0.135*	
			(0.0727)	
Avoir des enfants			-0.317***	
			(0.0744)	
Lambda				-0.168
				(0.337)
Observations		1,300	1,300	1,300
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

D'après les résultats, il apparaît que les femmes ont moins de probabilités que les hommes d'avoir un revenu élevé. De plus, il est observé que le nombre d'enfants est inversement proportionnel aux revenus, tandis que l'âge , la dépendance et l'expérience professionnelle sont positivement proportionnel aux revenus.

En ce qui concerne le ratio de mill, puisque la p-value est supérieure à 5%, on conclut qu'il n'est pas significatif. Cela indique ainsi que la sélection n'a pas d'effet notable sur les estimations des coefficients (*Voir do-file*).

En comparant les deux prédictions, on constate simplement que le modèle de sélection de Heckman offre de meilleures prédictions pour le revenu des non-travailleurs. Cette observation s'explique aisément car lorsque la variable dépendante est censurée, le modèle approprié est celui de Heckman.

3.5 Est-ce que le fait d'être marié et le fait d'avoir un enfant sont-elles endogènes ?

Le statut marital et le nombre d'enfants peuvent être endogènes, ce qui signifie qu'ils pourraient être liés à des facteurs non observés qui influencent également le revenu. Pour remédier à cette situation, l'utilisation de variables instrumentales pourrait être envisagée. Ces variables devraient être corrélées avec le statut marital et le nombre d'enfants, mais ne devraient pas être corrélées avec le revenu.