Table des matières

1 La participation des femmes au marché du travail aux États-Unis. 2
1.1 Présentation descriptive de la base *mroz*. . . . 2
1.2 Les effets attendus des variables. . . . 2
1.3 Estimation des modèles Logit, Probit et le modèle de probabilité linéaire. . . . 4
1.4 L'emploi du modèle de probabilité linéaire. . . . 5
1.5 Les effets marginaux des variables sur la probabilité de travailler. . . . 7
1.6 Probabilité de travailler et expérience. . . . 7
1.7 *kidslt6*, une variable endogène ? . . . 8
1.8 Expliquer le nombre d'heures travaillées (*hours*). . . . 8
1.8.1 Le rôle de la variable *hurshrs*. . . . 9
1.8.2 Modèle MCO sur les travailleurs. . . . 9
1.8.3 Modèle MCO sur la population totale (travailleurs ou non). . . . 9
1.8.4 Modèle de sélection Tobit sur l'ensemble de la population. . . . 9
1.8.5 Modèles de comptages : Poisson, Binomial négatif, ZIP, ZINP. . . . 10

2 Education et emploi des hommes aux États-Unis. 13
2.1 L'hypothèse : *Independance of Irrelevant Alternatives* (IIA). . . . 13
2.2 Estimation d'un Logit multinomial. . . . 13
2.3 Les rapports des risques relatifs du Logit multinomial. . . . 14
2.4 Expérience et être noir : l'effet sur le statut. . . . 15
2.5 Estimation par un modèle séquentiel. . . . 16

Liste des tableaux

1 Effets marginaux des modèles logit et probit . . . 7
2 Regressions : modèles de comptages variable endogène : heures travaillées. . . . 10

Table des figures

1 Évolution de la probabilité de travailler pour une femme en fonction du nombre d'années d'expérience dans le poste actuel. . . . 3
2 Répartition des probabilités selon le modèle de probabilité linéaire. . . . 6
3 Évolution de la probabilité de travailler pour une femme en fonction du nombre d'années d'expérience dans le poste actuel. . . . 8
4 Résidu du modèle Tobit. . . . 10
5 Effet joint de la couleur de peau et de l'expérience. . . . 15
6 Arbre pour la construction du modèle Logit séquentiel. . . . 17

1 La participation des femmes au marché du travail aux États-Unis.

Présentation de la base *mroz*.

La base *mroz* est une base composée 753 individus et de 22 variables. Elle présente un échantillon de femmes mariées et leurs activités sur le marché du travail aux États-Unis. Cette base permet de couvrir des thématiques variées telle que la participation ou non au marché du travail, leurs caractéristiques personnelles telles que l'âge, leur situation familiale (nombre d'enfants) ou bien des caractéristiques relatives à leur conjoint...

Pour l'analyse qui va suivre (partie 1), nous allons nous concentrer sur dix variables : *inlf* qui étudie la participation ou non de la femme i au marché du travail. *Nwifeinc* qui présente le revenu du ménage hors salaire de la femme, cette variable comprends le revenu du patrimoine ainsi que le revenu du conjoint. En d'autres termes, c'est le revenu de transfert. Les variables éducation (*educ*) et expérience (*exper*) qui retracent le parcours de l'individu. Nous sélectionnons aussi l'âge (*age*) de l'individu. Et enfin, nous disposons de deux variables rapportant le nombre d'enfants de l'individu (la première variable (*kidslt6*) traite des enfants ayant un âge inférieur à 6 ans et la seconde traite du nombre d'enfants ayant un âge compris entre 6 et 18 ans.)

Dans la partie 8 de la première partie (question relative aux modèles de comptage), nous ajoutons les variables *hours* et *hushrs* qui quantifient respectivement le nombre d'heures travaillés par la femme et celles travaillées par son conjoint.

1.1 Présentation descriptive de la base *mroz*.

1.2 Les effets attendus des variables.

Revenu de transfert ou *nwifeinc*.

La variable *Nwifeinc* décrit le revenu du ménage une fois enlevé le revenu de la femme. Il comprend alors les revenus du patrimoine et les revenus du travail du conjoint. Nous attendons un effet négatif de cette variable : plus le revenu du ménage augmente, moins la probabilité de travailler de la femme est important. Nous pouvons imaginer que si le revenu du ménage est faible alors, la femme doit contribuer par son travail à la richesse du ménage pour répondre aux besoins du ménage. Le coût d'opportunité de travail est de moins en moins important avec l'augmentation du revenu du ménage hors salaire de la femme. Nous pouvons illustrer avec les statistiques descriptives : le revenu du ménage hors salaire de la femme chez les individus qui travaillent est inférieur au revenu moyen des ménages des individus qui travaillent (21,7 contre 18,9). Nous pouvons vérifier que ces des moyennes sont significativement différentes en appliquant un test d'égalité des moyennes. La probabilité critique associée à ce test ($p_value = 0.0012$) est inférieur à 5%, nous rejetons alors hypothèse nulle d'égalité des moyennes, nous pouvons alors déduire que les deux moyennes sont significativement différentes.

Education (*educ*).

Nous attendons un effet positif de l'éducation sur la probabilité de travailler pour deux raisons principalement. Les femmes éduquées sont exposées à des périodes de chômage plus courte. Le taux de chômage parmi les non-diplômés est supérieur à celui des individus éduqués. Une deuxième explication qui va dans le même sens : après avoir supporté les coûts de l'éducation c'est-à-dire le coût financier des études ainsi que le coût d'opportunité (le coût de ne pas travailler), l'individu cherche à valoriser les enseignements de son parcours éducations en cherchant un emploi. Nous pouvons alors penser qu'il existe chez l'individu qui travaille, une volonté grande de trouver un emploi à la suite de ces études.

Expérience (*exper*).

Nous anticipons un effet positif de l'expérience sur la probabilité de travailler. Un individu ayant de l'expérience a acquis par la pratique des compétences qu'il peut valoriser sur le marché du travail le rendant alors moins sensible aux phases de chômage (pour les employeurs, son expérience est un actif intéressant). De plus, nous pouvons imaginer qu'une personne ayant de l'expérience dans son activité, a une connaissance fine du marché (relation tissée par exemple) et compétences propres au secteur.

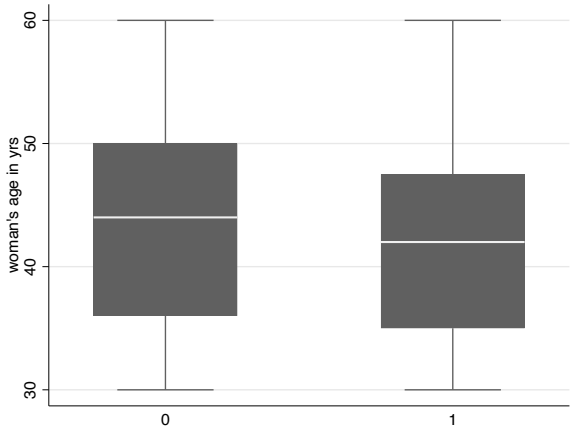
Expérience au carré ou *expersq*.

Nous attendons un effet négatif sur l'expérience mis au carré. La variable expersq permet de mettre en évidence une relation non linéaire entre l'expérience et la probabilité de travailler. Le signe attendu est négatif puisque que nous imaginons une fonction est concave (admet un maximum). En d'autres termes, une année d'expérience supplémentaire permet d'augmenter la probabilité de travailler, de plus en plus jusqu'à un certain point. A ce point, une année supplémentaire d'expérience se traduit par une augmentation de la probabilité de travailler moins grande qu'à la période précédente. Nous pouvons interpréter cela par l'obsolescence des compétences. Une personne ayant un grand nombre d'année dans le secteur peut manquer de perspective et d'un certain recul. En d'autres termes, la productivité de l'individu décroche, le rendant un peu moins attractif.

Âge.

L'effet de l'âge est indéterminé. Nous pouvons imaginer deux relations évoluant en sens inverse. La première relation est positive. Nous pouvons imaginer une discrimination à l'embauche de la part des employeurs envers des jeunes femmes (entre 20 et 30 ans), anticipant une potentielle maternité dans les années à venir ou bien une autre situation : l'employeur s'appuie le stéréotype suivant : les femmes sont première donneuse de soin à leur enfant en cas de maladie, préfèrera un homme (réduisant alors l'employabilité de la femme) (Petit 2003). L'âge avançant, cette contrainte potentielle se lève. Un accroissement de l'âge aura un effet positif sur la probabilité de travailler. Nous pouvoir imaginer aussi un effet négatif, une discrimination négative contre les séniors. Les employeurs anticipant le départ à la retraite des individus âgés préfèrent sélectionner des individus plus jeunes. La question est de savoir quel effet l'emporte sur l'autre.

FIGURE 1 – Évolution de la probabilité de travailler pour une femme en fonction du nombre d'années d'expérience dans le poste actuel.



Les deux boîtes à moustaches montrent que les femmes qui travaillent sont plus jeunes (en s'appuyant sur le premier quartile, la médiane et le troisième quartile).

Nombre d'enfants sous 6 ans ou *kidslt6*.

La variable *Kidslt6* indique le nombre d'enfants âgés de moins de 6 ans. Nous imaginons un effet négatif de cette variable sur la probabilité de travailler. Deux raisons : la première est sociologique, souhaitant s'occuper de ses jeunes enfants (pour des raisons économiques ou personnelles), la femme peut choisir d'arrêter de travailler. Une seconde raison est économique : des inégalités salariales existe entre homme et femme, une femme renonçant à travailler, elle renonce à un salaire moins élevé que si l'homme s'arrêtait. Le choix devient économiquement pertinent pour le ménage.

Nombre d'enfants entre 6 ans et 18 ans ou *kidsge6*.

La variable *Kidsge6* décrit le nombre d'enfant ayant plus de 6 ans tout en ayant moins de 18 ans. Nous attendons un effet interdéterminé sur la probabilité de travailler de la femme. En effet, nous anticipons un effet positif du nombre d'enfants sur la probabilité de travailler : avec le nombre d'enfants augmentant, les charges associées augmentent. De plus, cette tranche d'âge correspond à l'âge de l'école, la mère peut alors travailler. Néanmoins, nous pouvons aussi imaginer un effet négatif, la mère peut toujours envie de s'occuper de ses enfants (raisons personnelle et économique). Nous pouvons faire la remarque que cette catégorie est hétérogène : les besoins ne sont les mêmes entre un enfant de 6 ans et un enfant de 18 ans. Cela interroge sur l'interprétation donné au test Nous pouvons montrer cela en réalisant un simple test du χ^2 (indépendance entre les deux variables). Nous trouvons une p_value supérieur à 5% ($Pr = 0.763$), nous ne pouvons donc pas rejeter l'hypothèse nulle d'indépendance.

1.3 Estimation des modèles Logit, Probit et le modèle de probabilité linéaire.

Le modèle *Logit*.

Présentation théorique du modèle.

Nous allons estimer le modèle suivant :

$$p_i = \frac{e^{\beta_0+\beta_1\cdot nwifeinc_i+\beta_2\cdot educ_i+\beta_3\cdot exper_i+\beta_4\cdot exper_i^2+\beta_5\cdot age_i+\beta_6\cdot kidslt6_i+\beta_7\cdot kidsab6_i}}{1+e^{\beta_0+\beta_1\cdot nwifeinc_i+\beta_2\cdot educ_i+\beta_3\cdot exper_i+\beta_4\cdot exper_i^2+\beta_5\cdot age_i+\beta_6\cdot kidslt6_i+\beta_7\cdot kidsab6_i}} \quad (1)$$

Ici, nous pouvons voir qu'une probabilité sera associée à chaque individu. Le modèle est par nature hétéroscédastique, pour les analyses puisque la variance est dépendante de l'individu i.

Estimation du modèle Logit.

Le modèle est estimé avec la méthode du maximum de vraisemblance. Le modèle est globalement significatif au seuil de 5% et son pouvoir explicatif est satisfaisant (un pseudoR2 a presque 22%). Tous les coefficients sont significatifs à l'exception de la variable kidsge6 qui présente le nombre d'enfants âgés entre 6 et 18 ans. Nous pouvons interpréter les signes des coefficients tout d'abord : le revenu de transfert a un effet négatif sur la probabilité de travail, l'éducation et l'expérience ont un effet positif. Le coefficient associé l'expérience au carré est négatif mettant en évidence la relation non constante de l'expérience (une relation en forme de cloche qui admet un maximum). Concernant l'âge, l'effet négatif l'emporte sur le positif. Enfin avoir un enfant ayant moins de 6 ans réduit la probabilité de travailler pour la femme. Nous retrouvons les effets identifiés a priori. Les coefficients peuvent être interprétés comme des semi-élasticité mais l'interprétation avec les odd ratio (transformation exponentielle des coefficients) est plus accessible.

Le modèle *Probit.*

Présentation théorique du modèle.

Nous allons estimer le modèle suivant :

$$p_i = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{(\beta_0+\beta_1 \cdot nwifeinc_i+\beta_2 \cdot educ_i+\beta_3 \cdot exper_i+\beta_4 \cdot exper_i^2+\beta_5 \cdot age_i+\beta_6 \cdot kidslt6_i+\beta_7 \cdot kidsab6_i)^2}{2}} \tag{2}$$

Estimation du modèle Probit.

Le probit s'estime lui aussi à l'aide de la méthode du maximum de vraisemblance. Le modèle est globalement significatif. Le pseudo R2 est égal à 22%, ce qui est satisfaisant. Tous les coefficients sont significatifs à l'exception de la variable kidsge6. Les coefficients ne sont pas directement interprétables, nous devons utiliser les effets marginaux. Néanmoins, nous pouvons interpréter les signes des coefficients. Nous retrouvons les résultats du logit présenté un peu plus haut.

Le modèle de *probabilité linéaire.*

Présentation théorique du modèle.

Nous allons estimer l'équation suivante :

$$\begin{aligned} Y =& X\beta + \varepsilon \qquad\qquad \varepsilon \backsim N(0, \sigma_\varepsilon^2) \\ y_i =& \beta_0 + \beta_1 \cdot nwifeinc_i + \beta_2 \cdot educ_i + \beta_3 \cdot exper_i + \beta_4 \cdot exper_i^2 + \\ & \beta_5 \cdot age_i + \beta_6 \cdot kidslt6_i + \beta_7 \cdot kidsab6_i + \varepsilon_i \end{aligned} \tag{3}$$

Estimation du modèle de probabilité linéaire.

Le modèle de probabilité linéaire est une adaptation du modèle MCO au cadre probabiliste. Il s'agit d'une estimation par les moindres carrés (nous cherchons à minimiser la somme des carrés des résidus). Le modèle est globalement significatif mais a une capacité d'explication limitée (le R2 ajusté n'est que de 25,73% bien inférieur au seuil de 50%). Toutes les variables sont significatives à l'exception de la variable kidsge6. Les coefficients sont directement interprétables comme des effets marginaux. Ainsi, lorsque le revenu de transfert augmente de 1, la probabilité de travailler pour la femme diminue de 0,0034 point de pourcentage. Une année d'éducation supplémentaire provoque une augmentation de la probabilité de travailler pour la femme de 0,03 point de pourcentage. L'expérience a un effet positif sur la probabilité de travailler (+0,039 points de pourcentage), mais la relation n'est pas constante et admet un maximum (le coefficient associé à la variable expersq). L'âge a un effet négatif sur la probabilité de travailler et avoir un enfant de moins de 6 ans joue négativement (-0.26 points de pourcentage). Nous obtenons des résultats cohérent entre les trois modèles (même signes).Néanmoins, le dernier modèle estimé présente des limites.

1.4 L'emploi du modèle de probabilité linéaire.

Le modèle linéaire peut être utilisé dans un premier temps pour identifier des effets au préalable des différents régresseurs sur la variable d'intérêt et permettre de sélectionner les variables notamment en s'appuyant sur la significativité de ces dernières. Le modèle MCO présente directement les effets marginaux contrairement au logit et au probit.

Néanmoins au-delà de facilité de l'utilisation, ce modèle présente des limites. La première limite est graphique. La variable d'intérêt ne peut prendre que deux valeurs. La droite de régression (dans la

mesure d'une représentation possible avec plus de 2 variables explicatives) ne va pas pouvoir ajuster le nuage de points de manière satisfaisante. De plus, les valeurs prises par la variable endogène n'ont pas de sens en soit, elles n'indique que des modalités. Or, l'estimateur des MCO se construit sur les valeurs de la variable endogène : $\beta = (X'X)^{-1}X'y$.

La seconde limite porte sur les résidus. La valeur ε_i peut prendre deux valeurs :

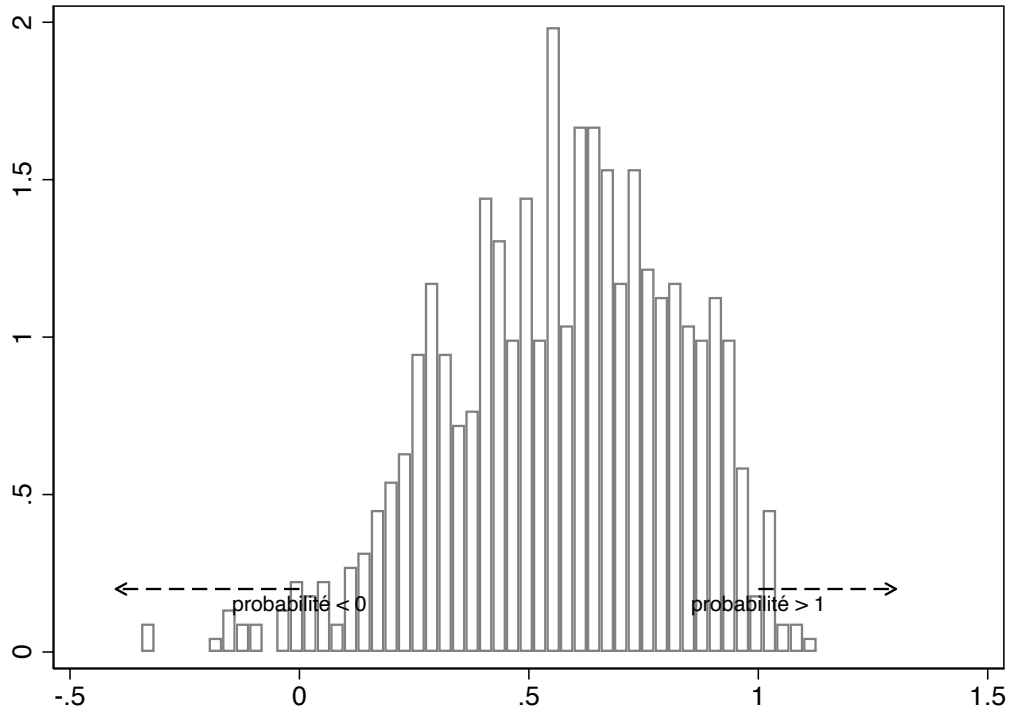
$$\begin{array}{llll} Si\ y_i = 1 & alors & \varepsilon_i = 1 - x_i\beta \\ Si\ y_i = 0 & alors & \varepsilon_i = -x_i\beta \end{array}$$

Nous calculons désormais l'espérance conditionnelle de l'erreur en rappelant l'hypothèse sur le terme d'erreur que pose le modèle des MCO $E(\varepsilon_i, X_i = x_i) = 0$.

$$\begin{aligned} E(\varepsilon_i, X_i = x_i) = &\ p_i(1 - x_i\beta) + 1 - p_i(-x_i\beta) \\ E(\varepsilon_i, X_i = x_i) = &\ p_i - x_i\beta \\ E(\varepsilon_i, X_i = x_i) &= 0 \\ p_i - x_i\beta &= 0 \\ p_i &= x_i\beta \end{aligned}$$

Une probabilité doit être compris entre 0 et 1, rien n'assure que $x_i\beta$ ne soit compris entre 0 et 1. Nous pouvons vérifier ce principe dans notre cas avec la représentation suivante :

FIGURE 2 – Répartition des probabilités selon le modèle de probabilité linéaire.



Nous pouvons voir sur cette figure 2, qu'il existe des valeurs de probabilité qui n'appartiennent pas à l'intervalle $[0, 1]$. Pour faire face à ce problème, nous pourrions modifier les valeurs des probabilités pour qu'elles prennent les valeurs extrêmes (0 et 1). Cette solution n'est pas entièrement satisfaisante, les modèles Logit et Probit semblent être plus intéressant.

Enfin, l'estimateur des MCO est ici hétéroscédastique. C'est-à-dire que la variance dépend des individus. Cette contrainte peut être levée en considérant des estimateurs robustes (estimateur *Sandwitch*).

1.5 Les effets marginaux des variables sur la probabilité de travailler.

Dans cette section il s'agit d'étudier les effets marginaux décrits dans les modèles Logit et Probit.

TABLE 1 – Effets marginaux des modèles logit et probit

	(1) mfx Logit	(2) mfx Probit
=1 if in lab frce, 1975		
HouseInc-womanWage	-0.0213*	-0.0120*
	[-2.53]	[-2.48]
years of schooling	0.221***	0.131***
	[5.09]	[5.18]
Actual experience	0.206***	0.123***
	[6.42]	[6.59]
ExperienceSq	-0.00315**	-0.00189**
	[-3.10]	[-3.15]
woman's age in yrs	-0.0880***	-0.0529***
	[-6.04]	[-6.23]
kids < 6y	-1.443***	-0.868***
	[-7.09]	[-7.33]
kids 6 - 18y	0.0601	0.0360
	[0.80]	[0.83]
Constant	0.425	0.270
	[0.49]	[0.53]
Observations	753	753

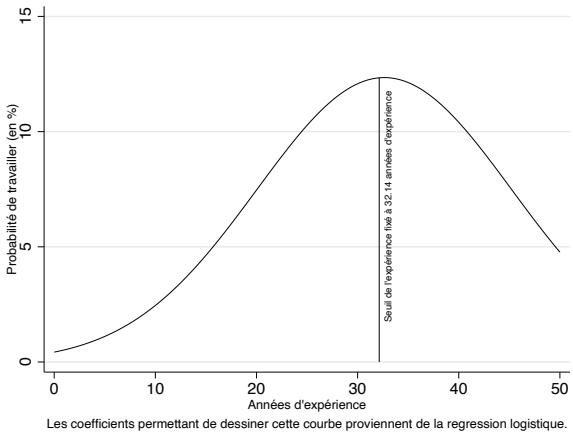
t statistics in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.6 Probabilité de travailler et expérience.

Graphiquement, nous représentons l'évolution de la probabilité de travailler pour une femme aux États-Unis (en 1975) lorsque nous faisons évoluer le nombre d'année d'expérience dans le poste actuel.

De manière évidente dans la figure 3, la relation n'est pas linéaire, elle a la forme d'une cloche (une fonction concave). Cela signifie que jusqu'à un certain seuil, une année d'expérience supplémentaire dans le poste actuel accroit la probabilité de participer au marché du travail, après, une année supplémentaire fait diminuer la probabilité supplémentaire de travailler.

FIGURE 3 – Évolution de la probabilité de travailler pour une femme en fonction du nombre d'années d'expérience dans le poste actuel.



Pour le calcul du seuil, c'est-à-dire le niveau où une année d'expérience supplémentaire n'accroit plus la probabilité de travailler, nous cherchons le niveau d'expérience qui donne un *odd-ratio* égal à 1, alors nous posons les opérations suivantes :

$$\begin{aligned} OR_{experience} &= \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{p_1}{1-p_1} \cdot \frac{1-p_0}{p_0} = 1 \\ &= e^{\beta_1 + 2 \cdot \beta_2 \cdot exper_{seuil} + \beta_2} = 1 \\ log(OR_{experience}) &= \beta_1 + 2 \cdot \beta_2 \cdot exper_{seuil} + \beta_2 = 0 \\ exper_{seuil} &= -\frac{\beta_1 + \beta_2}{2 \cdot \beta_2} \end{aligned}$$

En remplaçant les coefficients par leur valeur trouvée dans la régression Logit, nous pouvons en déduire les résultats suivants :

$$\begin{aligned} exper_{seuil} &= 32.13 \\ p_{seuil} &= 12.34 \end{aligned}$$

1.7 *kidslt6*, une variable endogène ?

1.8 Expliquer le nombre d'heures travaillées (*hours*).

Il existe quatre raisons de l'endogénéité : hétérogénéité inobservée, causalité inverse, causalité commune, erreur de mesure sur les variables explicatives. Économétriquement, l'endogénéité se traduit par la relation suivante :

$$cov(X_i, \varepsilon_i) \neq 0$$

L'endogénéité va biaiser les résultats. Dans notre cas, nous suspectons de l'endogénéité portée par la variable kidslt6 pour plusieurs raisons. Tout d'abord, nous pouvons indiquer qu'il existe une relation entre la variable kidslt6 et les autres variables du modèle dont notamment âge, l'expérience, l'éducation,

le nombre d'enfant âgé de plus de 6 ans. Nous pouvons illustrer avec une simple régression MCO. La régression est globalement significative et l'ensemble des variables sont significativement différents de 0 au seuil de 5%.

1.8.1 Le rôle de la variable *hurshrs*.

Le nombre d'heures travaillés par le conjoint peut être utilisé pour mesurer l'engagement familial du conjoint. De cette manière, le conjoint présentant un nombre d'heure faible, peut signifier davantage disponibilité pour la femme lui permettant de travailler. Nous attendons alors un effet négatif, les deux variables évoluant en sens inverse. Une autre explication, un conjoint travaillant peu, peut signifier que la femme doit trouver un travail pour répondre aux besoins du foyer.

L'interaction entre l'éducation et l'expérience va permettre de capter un effet joint de ces deux variables et de mettre en évidence la conjugaison des deux effets (forte expérience et forte éducation). Aussi, nous pouvons imaginer deux situations que cette variable va capter : des jeunes arrivant sur le marché du travail (peu d'expérience et un niveau d'éducation élevé), tout comme des personnes ayant de l'expérience sans avoir un niveau d'éducation élevé.

1.8.2 Modèle MCO sur les travailleurs.

En estimant seulement les individus qui travaillent, nous avons une sous-population spécifique avec des caractéristiques qui lui sont propres qui sont inobservables dans notre modèle considéré. Les estimateurs sont biaisés (biais de sélection), nous ne pouvons pas en déduire des résultats globaux pour la population. En d'autres termes, nous ne pouvons utiliser le principe de l'inférence.

1.8.3 Modèle MCO sur la population totale (travailleurs ou non).

En estimant le modèle sur l'ensemble de la population, nous sommes dans une situation aussi biaisée. Le problème de générer une telle régression est que nous pouvons faire des prédictions d'heures négatives. Or cette situation n'est pas possible. Nous faisons face à une situation de censure. Nous allons construire un modèle qui permet de borner l'output à 0 (limite basse). Par l'utilisation d'un modèle de censure, un modèle Tobit, nous allons atténuer le biais. Nous pouvons aussi vérifier la qualité du modèle notamment normalité des résidus. Le test de Jarque-Berra nous indique l'hypothèse nulle de normalité des résidus est rejetée au seuil de 5%.

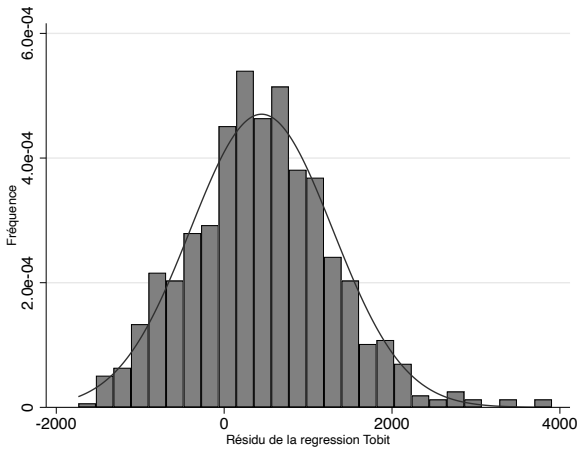
1.8.4 Modèle de sélection Tobit sur l'ensemble de la population.

Dans cette régression, nous allons censurer le nombre d'heures à 0 (limite inférieur). Néanmoins, en procédant de cette manière, nous sommes aussi confrontés à des problèmes. Nous sommes incapables de différencier les vrais 0. Cela signifie que nous sommes incapables de différencier les individus qui ne souhaitent pas travailler, le salaire de réservation n'est pas suffisamment élevé (salaire minimum à partir duquel l'individu accepte de travailler), des individus qui cherchent en emploi sans succès.

Aussi, sur un plan économétrique, nous pouvons faire l'analyse des résidus. Le modèle Tobit est construit sur la normalité des résidus. Dans le cadre de ce modèle, nous faisons un test de Jarque-Berra. la normalité des résidus est rejetée au seuil de 5%, le modèle proposé n'est pas valide. L'histogramme 4 présente la répartition des résidus.

Pour faire face à ce problème (différencier les individus), nous pouvons utiliser un modèle de troncature (modèle de Heckman). Ce modèle considère que l'échantillon à disposition présente des spécificités et cherche donc à mitiger ce biais de sélection. En effet, dans la construction de cette base, les individus qui travaillent sont surement surreprésenté et donc pas représentatif de la population globale.

FIGURE 4 – Résidu du modèle Tobit.



1.8.5 Modèles de comptages : Poisson, Binomial négatif, ZIP, ZINP.

Pour étudier le nombre d'heures travaillées, nous allons présenter des modèles de comptages suivants :

— Modèle de Poisson (*poisson reg*) ;
— Modèle Binomial Négatif (*nb reg*) ;
— Modèle Zero Inflated Poisson (*zip reg*) ;
— Modèle Zero Inflated Negative Binomial (*zinb reg*).

Estimation des modèles.

Pour faciliter la comparaison entre les modèles, nous présentons ci-dessous un tableau récapitulatif présentant les différents résultats des modèles :

TABLE 2 – Regressions : modèles de comptages variable endogène : heures travaillées.

	(1) poisson reg	(2) nb reg	(3) zip reg	(4) zinb reg
hours worked, 1975				
HouseInc -womanWage	-0.00621***	-0.00932	0.0000726	-0.000237
	(-42.82)	(-0.93)	(0.50)	(-0.06)
years of schooling	0.0725***	0.107	-0.0325***	-0.0347
	(57.62)	(1.44)	(-26.37)	(-1.20)
Actual experience	0.135***	0.176*	0.0276***	0.0280
	(129.30)	(2.38)	(26.22)	(1.05)
ExperienceSq	-0.00176***	-0.00230*	-0.000548***	-0.000596
	(-108.33)	(-2.02)	(-32.96)	(-1.43)
woman's age in yrs	-0.0442***	-0.0538**	-0.0153***	-0.0150*
	(-199.36)	(-3.14)	(-66.99)	(-2.53)
kids < 6y	-0.830***	-1.083***	-0.267***	-0.289**
	(-196.91)	(-4.61)	(-62.92)	(-3.02)

kids 6 - 18y	-0.0344***	0.0522	-0.0591***	-0.0589
	(-29.29)	(0.56)	(-48.88)	(-1.85)
hours worked by husband, 1975	-0.0000548***	-0.000213	0.00000915***	0.00000272
	(-23.20)	(-1.15)	(3.84)	(0.04)
Interaction educ exper	-0.00154***	-0.00244	0.00109***	0.00120
	(-20.99)	(-0.45)	(15.24)	(0.65)
Constant	6.931***	7.020***	7.884***	7.910***
	(347.46)	(5.58)	(401.35)	(16.68)
/				
lnalpha		2.002***		-0.674***
		(36.76)		(-10.57)
inflate				
HouseInc -womanWage			0.0195*	0.0195*
			(2.30)	(2.30)
years of schooling			-0.218**	-0.218**
			(-3.02)	(-3.02)
Actual experience			-0.196**	-0.196**
			(-2.75)	(-2.75)
ExperienceSq			0.00305**	0.00305**
			(2.98)	(2.98)
woman's age in yrs			0.0896***	0.0896***
			(6.12)	(6.12)
kids < 6y			1.458***	1.458***
			(7.14)	(7.14)
kids 6 - 18y			-0.0691	-0.0691
			(-0.92)	(-0.92)
hours worked by husband, 1975			0.000225	0.000225
			(1.52)	(1.52)
Interaction educ exper			-0.000609	-0.000609
			(-0.12)	(-0.12)
Constant			-1.010	-1.010
			(-0.87)	(-0.87)
Observations	753	753	753	753
Pseudo R^2	0.264	0.006		
AIC	630063.3	8662.0	197582.6	7721.4
BIC	630109.5	8712.9	197675.1	7818.5
p	0	9.26e-09	0	0.00134

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Commentaires de la table de résultats.

Le meilleur modèle.

Tous les modèles sont globalement significatifs (dernière ligne du tableau p), il est donc question de trouver quel est le modèle le plus adapté pour modéliser le nombre d'heures travaillées par les femmes en 1975.

Le modèle de Poisson indique un R^2 plutôt bon pour un modèle non-linéaire (supérieur à 20%). Néanmoins, nous soulever plusieurs limites de ce modèle :

— Le modèle est construit sur une hypothèse d'*équidispersion* (c'est-à-dire que si on considère que $y \backsim Poisson(\lambda)$ suit une loi de Poisson alors $E(y) = V(y) = \lambda$) ;
— Le modèle de Poisson met en valeur beaucoup de 0 (*zero excess burden*).

Pour résoudre ces deux problèmes, nous utilisons le modèle négative binomial qui permet de lever l'hypothèse d'équidispersion, la variance peut être alors différente de la moyenne. Nous pouvons réduire la surepresentation de zéro à l'aide d'un modèle *zero inflated*. Il apparait que le paramètre *lnalpha* est significatif dans les deux modèles considérant une regression binomiale négative. Cela indique que l'hypothèse nulle du modèle de poisson est rejetée. Nous pouvons donc écarter les modèles de Poisson.

En comparant les critères d'information des quatre modèles, le modèle 4 (ZINB) est le meilleur modèle puisqu'il minimise les critères d'information AIC et BIC. Ce modèle (ZINP) malgré un pseudo R2 faible supposé (nous nous appuyons sur le R^2 de la modélisation négative binomial), permet de traiter à la fois l'*équidispersion* posé par le modèle de Poisson et la problématique lié à la *surpondérance de zéro* dans la prédiction.

L'interprétation des coefficients.

Pour la regression de Poisson, le coefficient associé au revenu de transfert de la femme (revenu du conjoint ainsi que les revenus du patrimoine), peut être estimé de la manière suivante : une augmentation d'une unité du revenu de transfert conduit à 0,0061% d'heures travaillées en moins. Une année d'éducation supplémentaire, conduite à un accroissement du nombre d'heures travaillées de 0,0725%.

Si nous interprétons le modèle *Zero Inflated Negativie Binomial*, seules deux variables sont significatives au seuil de 5% : l'âge de la femme ainsi que le nombre d'enfants ayant moins de 6 ans. L'interprétation est particulière : pour les individus qui travaillent (des heures travaillées supérieures à 0), être agé d'une année supplémentaire conduit à une diminution de 0,015% d'heures travaillées toute chose égale par ailleurs. Les femmes qui travaillent, qui ont un enfant agé de moins de 6 ans, ont 0,289% d'heures travaillées en moins. Le reste des coefficients ne sont pas statistiquement significativement différent de 0, et ne donc pas interprétable.

2 Education et emploi des hommes aux États-Unis.

Présentation de la base *keane*.

Cette seconde base présente un échantillon d'hommes sur leur historiques éducatif et d'emploi. Cette base est composée de 12 723 observations et de 7 variables : l'éducation en année *educ*, la couleur de peau *black*, le salaire (exprimé en log) avec la variable *lwage*, le niveau d'expérience sur le marché du travail *exper* et le statut professionnel *status*.

Ce travail consiste à interpréter les déterminants du statuts professionnels entre être à un étudiant (*school*), être un travailleur (*work*) et être ni l'un, ni l'autre (*home*). Nous chercherons à expliquer le choix du statut par l'expérience (+ l'expérience au carré) et la couleur de peau.

2.1 L'hypothèse : *Independance of Irrelevant Alternatives* (IIA).

Le modèle Logit conditionnel est construit sur l'hypothèse d'indépendance des alternatives non pertinente (Independance of Irrelevant Alternatives). L'IIA que la probabilité conditionnelle ne dépend pas des autres alternatives. En d'autres termes, le rapport de probabilités entre deux modalités est indépendant des autres modalités de la variable d'intérêts. Le corolaire de cette propriété est : ajouter une modalité ou bien en supprimer une ne modifie par les rapports de probabilités.

Dans notre cas, nous travaillons sur le statut professionnel, variable qui prendre trois modalités : *school*, *home* et *work*. Pour estimmer la probabilité de chaque modalité, nous utilisons la formule suivante :

$$p_i = P(y_i = j, Xi = x_i) = \frac{e^{x_{ij} \cdot \beta_j}}{\sum_{j=1}^{K} e^{x_{ij} \cdot \beta_j}} \tag{4}$$

Nous allons rajouter une nouvelle catégorie : *unemployment* et nous estimons à nouveau le modèle avec cette nouvelle catégorie. Nous notons les probabilités associées aux modalités respectivement p_{work}, p_{home}, p_{school} et $p_{unemployment}$. L'hypothèse IIA est acceptée si l'ensemble des rapports des différents probabilités ne se modifie pas entre la période 1 (3 modalités) et la période 2 (4 modalités). Économétriquement, nous posons les hypothèses suivantes :

$$\begin{aligned} H_0 : & \frac{p_{home}}{p_{school\,1}} = \frac{p_{home}}{p_{school\,2}} \; et \; \frac{p_{home}}{p_{work\,1}} = \frac{p_{home}}{p_{work\,2}} \; et \; \frac{p_{work}}{p_{school\,1}} = \frac{p_{work}}{p_{school\,2}} \\ H_1 : & \; la\; condition\; n\; est\; pas\; v\acute{e}rifi\acute{e}e \end{aligned}$$

Pour tester cette hypothèse, nous procédons aux tests *Hausman et Mac Fadden* et celui de *Small-Hsiao* (qui est plus robuste que le premier). La statistique du test est comparée à la valeur critique de la table du χ^2 à K degrés de liberté (avec K le nombre de composante du vecteur β_c dans le test de Hausman). Si la statistique du test est supérieure à la valeur critique alors, l'hypothèse nulle $H_0 : IIA$ est rejetée.

2.2 Estimation d'un Logit multinomial.

Avant de commencer à interpréter à la table de résultats, nous testons la validité du modèle et plus particulièrement l'hypothèse iia. Nous disposons de plusieurs de tests dont deux de Hausman et le test de Small-Hsiao. Le test de Hausman rejette l'hypothèse nulle $H_0 : IIA$ au seuil de 5%. Le test de Small Hsiao, plus robuste, ne rejette pas l'hypothèse nulle. Le modèle est alors valide en s'appuyant sur ce test. Nous obtenons la table de résultat suivante :

```
Iteration 0:   log likelihood = -12620.959
Iteration 1:   log likelihood = -10202.023
Iteration 2:   log likelihood = -10029.003
Iteration 3:   log likelihood = -10021.733
Iteration 4:   log likelihood = -10021.689
Iteration 5:   log likelihood = -10021.689

Multinomial logistic regression                 Number of obs     =     12,665
                                                LR chi2(6)        =    5198.54
                                                Prob > chi2       =     0.0000
Log likelihood = -10021.689                     Pseudo R2         =     0.2059
```

status	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
school	(base outcome)					
home						
exper	1.139577	.0578639	19.69	0.000	1.026166	1.252989
expersq	-.120484	.010237	-11.77	0.000	-.1405481	-.1004199
black	1.253806	.058931	21.28	0.000	1.138304	1.369309
_cons	-.644524	.0401709	-16.04	0.000	-.7232576	-.5657905
work						
exper	2.086142	.0546137	38.20	0.000	1.979101	2.193183
expersq	-.2014872	.009233	-21.82	0.000	-.2195835	-.1833909
black	.5170278	.0602732	8.58	0.000	.3988945	.635161
_cons	-.6110416	.0393023	-15.55	0.000	-.6880726	-.5340105

Avec l'estimation de ce modèle, nous avons perdu des observations : la base comprenait 12 723 observations et en compte désormais 12 665 (il manque au moins une valeur dans les variables pour cette observations). Néanmoins, la taille de l'échantillon reste important.

Ce modèle multinomial est estimé à l'aide de la méthode du maximum de vraisemblance. Il est globalement significatif (la pvalue du modèle est inférieur au seuil de 5%). De plus, ce modèle dispose d'un pouvoir explicatif satisfaisant (le Pseudo R2 est supérieur à 20%).

La modalité de référence de ce modèle est être à l'école (outcome 3). L'ensemble des coeffients pour les trois variables sont significativement différent de 0. Nous ne pouvons pas interpréter davantage, les coefficients sont construits sur deux élements non séparément identifiés. Nous pourrons interpréter les résultats grâce aux rapports des risques relatifs.

2.3 Les rapports des risques relatifs du Logit multinomial.

L'estimation du Logit multinomial en mettant en évidence les Rapports des Risques Relatives (RRR), nous donne la table de résultat suivante :

Nous allons directement interpréter les résultats (les autres remarques sur la qualité du modèle ont été présenté dans la question précédente. Tous les coefficients sont significatifs et donc interprétables.

Lorsque le niveau d'expérience augmente d'une année, alors il y a 3,12 fois plus (ou 212% de plus) de chance d'être à la maison plutôt que d'être en cours toute chose égale par ailleurs. Il y a plus de 8 fois plus de chance de se trouver au travail plutôt qu'à l'école, lorsque le niveau d'expérience augmente d'une année.

Le RRR associé à la variable expérience au carré est inférieur à 1 et indique donc une relation négative, mettant en évidence alors une relation non constante et négative de la variable expérience (elle admet un maximum). En d'autres termes, il y existe un niveau d'expérience à partir duquel, une année d'expérience supplémentaire accroit la chance d'appartenir à la catégorie maison ou travail par rapport à l'école mais mois que les années d'année d'expérience précédente. Nous pouvons remarquer que le RRR associé à la catégorie maison est supérieur à celui du travail (toujours par rapport à la catégorie de référence).

Le RRR de la variable black indique qu'il y a 3,5 fois plus (250% de plus) de chance d'être à la

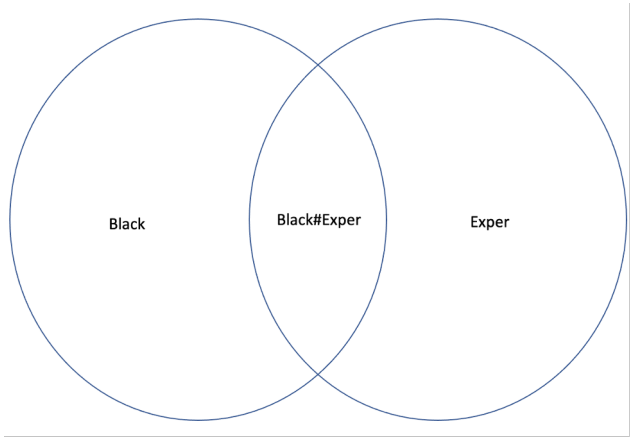
```
. mlogit status $x3list, baseoutcome(1) rrr

Iteration 0:   log likelihood = -12620.959
Iteration 1:   log likelihood = -10202.023
Iteration 2:   log likelihood = -10029.003
Iteration 3:   log likelihood = -10021.733
Iteration 4:   log likelihood = -10021.689
Iteration 5:   log likelihood = -10021.689

Multinomial logistic regression                 Number of obs     =     12,665
                                                LR chi2(6)        =    5198.54
                                                Prob > chi2       =     0.0000
Log likelihood = -10021.689                     Pseudo R2         =     0.2059

------------------------------------------------------------------------------
      status |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
school       |  (base outcome)
-------------+----------------------------------------------------------------
home         |
       exper |   3.125447   .1808506    19.69   0.000     2.790348    3.50079
     expersq |   .8864913    .009075   -11.77   0.000     .8688819    .9044576
       black |   3.503654   .2064739    21.28   0.000     3.121469    3.932633
       _cons |   .5249123   .0210862   -16.04   0.000     .4851692    .5679111
-------------+----------------------------------------------------------------
work         |
       exper |   8.053782   .4398468    38.20   0.000     7.236234    8.963696
     expersq |    .817514   .0075481   -21.82   0.000     .8028531    .8324427
       black |   1.677036   .1010803     8.58   0.000     1.490176    1.887326
       _cons |   .5427852   .0213327   -15.55   0.000     .5025437    .5862491
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

FIGURE 5 – Effet joint de la couleur de peau et de l'expérience.



maison plutôt que d'être à l'école quand nous sommes de couleur noir par rapport être blanc. Il y a 1,67 fois plus de chance d'être au travail plutôt qu'à l'école quand nous sommes noirs plutôt que blanc.

2.4 Expérience et être noir : l'effet sur le statut.

Ayant mis l'expérience dans le modèle et la couleur de peau, l'ajout de cette variable va permettre de capter un effet spécifique. Graphiquement, nous avons 5

La variable expérience dans le modèle va capter l'effet de l'expérience (seul) et la variable black va capter l'effet relatif à la couleur de peau (seul). La zone entre les deux cercles donne l'effet joint des deux variables.

Il existe de fortes inégalités aux États-Unis basé sur la couleur de peau à la fois sur le revenu, le patrimoine et le statut social. Nous allons donc pouvoir mesurer cette discrimination grâce à la variable d'interaction. Nous re-estimons le modèle en rajoutant cette variable. Nous obtenons la table qui suit :

Avant d'interpréter les coefficients, nous réalisons à nouveau le test IIA. Ici aussi, le test de Small-

```
Iteration 0:   log likelihood = -12620.959
Iteration 1:   log likelihood = -10262.617
Iteration 2:   log likelihood = -10035.879
Iteration 3:   log likelihood = -10021.797
Iteration 4:   log likelihood =  -10021.63
Iteration 5:   log likelihood =  -10021.63

Multinomial logistic regression                 Number of obs     =     12,665
                                                LR chi2(8)        =    5198.66
                                                Prob > chi2       =     0.0000
Log likelihood =  -10021.63                     Pseudo R2         =     0.2060
```

status	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
school	(base outcome)					
home						
exper	3.147197	.1932371	18.67	0.000	2.790361	3.549666
expersq	.8863355	.0090715	-11.79	0.000	.8687328	.9042948
black	3.534249	.2267093	19.68	0.000	3.116705	4.007732
black#c.exper						
1	.9791541	.067109	-0.31	0.759	.8560747	1.119929
_cons	.5228852	.0218778	-15.50	0.000	.4817166	.5675721
work						
exper	8.077449	.4631285	36.44	0.000	7.218878	9.038132
expersq	.8175036	.0075488	-21.82	0.000	.8028413	.8324337
black	1.677935	.1137064	7.64	0.000	1.46924	1.916273
black#c.exper						
1	.9862156	.0655368	-0.21	0.835	.8657792	1.123406
_cons	.5429189	.0221435	-14.98	0.000	.5012078	.5881012

Note: _cons estimates baseline relative risk for each outcome.

Hsiao indique l'hypothèse nulle d'IIA n'est pas rejetée. Le modèle est globalement significatif et le pseudo R2 est satisfaisant (très légèrement supérieur à celui du précédent modèle.

La variable d'interaction n'apparait pas comme significativement différentes de 0 dans le cadre de ce modèle. Ce résultat apparait comme surprenant. Les RRR se sont vu modifié à la marge. Par ce modèle, nous ne pouvons pas assurer la mesure d'une discrimination ou du moins une particularité du l'expérience chez les personnes noires.

2.5 Estimation par un modèle séquentiel.

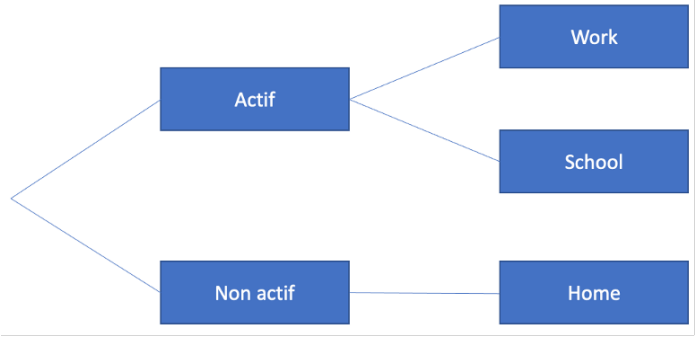
Pourquoi utiliser un modèle séquentiel ?

Le modèle séquentiel permet de traiter une situation dans laquelle les modalités sont classés en parties. Ce modèle permet de découper les choix de décision en séquences, à chaque séquence, l'individu optimise son choix. L'avantage de ces modèles est de pouvoir relâcher l'hypothèse contrainte d'IIA.

La construction de l'arbre.

Dans notre cas, nous pouvons imaginer deux arbres séquentiels. Le premier traite les individus en fonction qu'ils soient actifs ou non (soit à l'école ou au travail ou bien à la maison). Le second arbre traite des individus qu'ils travaillent ou non (work contre school et home). Nous allons modéliser le premier arbre 6 :

FIGURE 6 – Arbre pour la construction du modèle Logit séquentiel.



Estimation et interprétation du modèle logit séquentiel.

Nous obtenons la table de résultats suivante :

```
Iteration 0:   log pseudolikelihood = -10019.359
Iteration 1:   log pseudolikelihood = -10019.359

                                                Number of obs     =     12,665
                                                Wald chi2(4)      =    1102.60
Log pseudolikelihood = -10019.359               Prob > chi2       =    0.0000
```

status	Odds Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
_1_3v2						
exper	1.483548	.0586411	9.98	0.000	1.372953	1.603052
expersq	.9819846	.007272	-2.45	0.014	.9678347	.9963413
black	.3648661	.0192971	-19.06	0.000	.3289387	.4047176
black#c.exper						
1	1.090121	.0320051	2.94	0.003	1.029163	1.15469
_cons	2.753359	.1033938	26.97	0.000	2.557989	2.963651
_3v1						
exper	8.268568	.4990802	35.00	0.000	7.346033	9.306958
expersq	.8137375	.006619	-25.34	0.000	.8008674	.8268144
black	1.591929	.1094268	6.76	0.000	1.391276	1.82152
black#c.exper						
1	1.036354	.0731082	0.51	0.613	.902529	1.190022
_cons	.5401861	.0221735	-15.00	0.000	.498429	.5854415

Ce modèle est estimé par la méthode du maximum de vraisemblance. Comme pour le logit multinomial, le modèle ne considère pas l'ensemble de la population (perte d'observations du fait de valeurs manquantes). Le modèle est globalement significatif (la probabilité critique est inférieure à 5%). Le modèle est estimé à l'aide des ODD ratio pour faciliter l'interprétation.

Nous obtenons deux tables : la première présente la possibilité de prendre le choix école ou bien travail au lieu de choisir être à la maison. Dans cette première régression, tous les Odd Ratio sont significatifs au seuil de 5% (y compris la variable d'interaction). Choisir d'être actif par rapport à ne pas être actif augmente de 48% lorsque nous avons une année d'expérience supplémentaire. La relation de l'expérience est là aussi inférieure à 1 indique une relation non linéaire qui admet un maximum. Être noir par rapport à être non noir augmente le risque de ne pas être actif de 63,52% toutes choses égale par ailleurs. Enfin, le terme d'interaction est supérieur à 1. Nous pouvons alors l'interpréter, lorsque nous somme de couleur noir, si l'expérience augmente d'une année, la chance de choisir d'être

actif par rapport à ne pas être actif augmente de 9% toutes choses égale par ailleurs.

Dans la seconde table, nous comparons le fait de choisir le travail par rapport à l'école. Une année d'éducation supplémentaire augmente la chance d'être au travail de 726% (odd-ratio = 8,2). La relation de l'expérience au carré est là aussi négative (inférieur à 1). Il existe donc bien un minimum. Être noir augmente la chance de travail de 59% par rapport à choisir l'école. Le terme d'interaction n'est pas significativement différent de 0 dans cette table. Alors, il n'est pas possible d'interpréter le coefficient.