Université de Montréal

Rapport de recherche

Étude sur les déterminants de la poursuite d'études supérieures avec décomposition d'Oaxaca-Blinder

Rédigé par :
Blais, Frédéric

Dirigé par :
Yves Richelle

Département de sciences économiques
Faculté des arts et des sciences

08/07/2013

Étude sur les déterminants de la poursuite d'études supérieures avec décomposition d'Oaxaca-Blinder

Travail présenté à M. Yves Richelle
Atelier de maîtrise : Projet de recherche
par Frédéric Blais
BLAF23099008

Université de Montréal
Le 8 juillet 2013

Résumé :

Dans le cadre de ce projet, nous tentons de déterminer les facteurs principaux entrant en ligne de compte dans la décision de s'enrôler dans des études universitaires chez les canadiens de moins de 24 ans. Un modèle Logit binomial est utilisé et une décomposition de Blinder-Oaxaca est implementé sur ce dernier. L'importance des parents et du sexe de l'individu d'intérêt sur la probabilité d'enrôlement est appuyée par nos résultats.

Table des matières

1 Introduction

Depuis le virage de la macroéconomie vers la prise de racines chez la microéconomie, un vaste courant d'érudits a voulut modéliser l'effet du comportement des agents et des firmes rationnelles sur des variables macroéconomiques. Entre autre, la croissance économique fait partie des sujets de recherche majeurs en macroéconomie. Derrière cette dernière peuvent se cacher divers facteurs économiques et d'autres leviers prenant pour source d'autres sciences comme la politique. Néanmoins, un point central de cette étude repose sur l'exploration de l'accumulation de capital physique et humain. La décision d'accumulation de capital humain, lorsqu'étudier dans un cadre macroéconomique, suppose une optimisation de part et d'autre des décideurs en jeu dans le modèle et une fonction de transition pour chaque type de capital entre les périodes. Cependant, il est clair que, en prenant un individu donné, ses motivations dans la poursuite d'études outrepassent en complexité toute équation simplificatrice. Cette sophistication dans le processus décisionnel le rend captivant et elle justifie le fait de s'y attarder.

De la même façon que des racines de la macroéconomie sont prises en microéconomie, nous nous demandons dans cette étude quels sont les facteurs qui poussent un individu moyen, membre de l'agrégat, à poser la décision d'investir dans sa propre personne. Ce regard microéconomique sur l'accumulation de capital humain désire non seulement jeter de la lumière sur les forces en présence, mais aussi les décortiquer dans leurs détails. Il est raisonnable d'imaginer que l'effet d'une variable donnée sur le processus décisionnel implicite de notre voisin peut être différent de l'effet de cette même variable sur le nôtre. Dans ce contexte précis, nous cherchons à estimer ces différences dans la probabilité qu'à un individu de faire le choix explicite d'investir dans ses propres capacités. Le point focal de cette recherche est les déterminants sous-jacents à cette décision et comment leur effet change selon le groupe d'individu dont nous faisons partie.

L'apport de cet article sera à la fois de scruter les effets de variables, comme les fonds disponibles au moment de la décision ou la cadre dans lequel un individu a évolué pendant les temps précédents sa décision, mais aussi d'observer quelle part dans la différence des probalités moyennes prédites pour les observations faisant parties de divers groupes provient réellement de différence des estimateurs des régression et quelle part serait inexpliquée et issu des différences dans les régresseurs chez les individus entre ces groupes. Selon le groupe traité, cette source externe pourrait être vue comme une sorte de discrimination. Notons, avant de poursuivre, que nous utilisons «discrimination» au sens large du terme tout au long de ce texte. Cette recherche fait suite à une étude effectuée en 2011 par Benoît, Blais, Desjardins et Pierre dans le cadre du cours ECN3950 «Atelier en économie appliquée» à l'Université de Montréal. Nous abordons dans l'étude en cours la question avec plus de détail, des données plus récentes et divers méthodes pour décomposer les effets en présence. À l'aide de régressions par strates et de décompositions de Blinder-Oaxaca [1], tributaire des textes phares d'Oaxaca (1973) et de Blinder (1973), appliquées dans le cadre d'un modèle Logit binomial sur diverses variables, comme le sexe du répondant, nous serons en mesure d'évaluer la direction et l'amplitude de l'effet sur la probabilité de fréquenter l'université conditionnel au fait de faire partie d'un groupe ou d'un autre.

Au fil du texte, nous amènerons de la manière la plus auto-contenu possible les théories et observations nécessaires à l'approche de ce sujet. Dans la section 2, nous effectuerons une revue de littérature en traitant des textes sur lesquels nos méthodes pratiques se basent et des textes sur lesquels notre raisonnement théorique prend ses fondations. Par la suite, nous expliciterons dans la section 3 le cadre théorique que nous utiliserons plus tard et nous décrirons quelles variables nous avons eu à notre disposition et sous quelles formes dans la section 4. La section 5 se verra comme un résumé concis des méthodes économétriques utilisées; le traitement se voudra rapide, mais aussi détaillé que possible. Finalement, nous exposerons les résultats que nous aurons obtenus dans la section 6. Cette dernière sera divisée en une partie abordant les régressions standards avec diverses modifications et une seconde, tributaire de la première quant à l'intuition et aux explications, dans lesquels les décompositions de B-O seront exposées à proprement parlé.

1. Nous noterons cette décomposition par «décomposition de B-O» pour le reste de notre traitement du sujet

2 Revue de littérature

Plusieurs articles dans la littérature abordent les méthodes et arguments utilisés dans nos régressions et dans les décompositions qui s'en suivent. Dans les sous-sections à venir, nous exposerons en détails les principaux articles qui nous auront guidés à travers ce projet. Dans les sous-sections 2.1 et 2.2 nous observerons des articles traitant de la méthode de décomposition de Blinder-Oaxaca non-linéaire et de son application à travers le logiciel STATA. Notons que les articles traitant de cette décomposition sont tous tributaires des articles phares de Ronald L. Oaxaca (1973) et d'Alan S. Blinder (1973). À travers les sous-sections 2.3 et 2.4, nous examinerons de plus près cette méthode de décomposition dans des modèles de régression logistique et avec des applications similaires en essence à celles que nous mettrons à terme dans ce projet. Enfin, dans les sous-sections 2.5 et 2.6, nous traiterons d'articles ayant le raisonnement général derrière l'intuition économique et la théorie économique qui caractériseront notre analyse de la demande d'éducation supérieure au fil de notre étude.

2.1 Décomposition de B-O pour des modèles non-linéaires

Ces deux articles exposent par leurs manières respectives une façon d'attaquer les décompositions de Blinder-Oaxaca dans des contextes généraux. Bien que le premier fut paru après le second, ce dernier le généralise. Nous commençons donc par la revu d'un cadre général d'attaquer les décompositions non-linéaires pour ensuite se concentrer sur les cas particuliers des modèles logit ou probit binomiaux.

2.1.1 Thomas K. Bauer et Mathias Sinning (2007)

Dans cet article, les chercheurs T.K. Bauer et M. Sinning exposent une méthode générale de traiter la décomposition de Blinder-Oaxaca dans le cadre de modèles non-linéaires comme, par exemple, dans un modèle logistique. En plus d'une section dédiée aux modèles ayant une variable dépendante discrète, cet article aborde aussi l'application la décomposition dans des modèles à variable dépendante restreinte, tel des modèles censurés ou tronqués.

Supposons que nous avons des groupes A et B, indexés par $g \epsilon \{A, B\}$, où chaque individu membre du groupe g est indexé par $i = 1, 2, ...N_g$ et où $N_A + N_B = N$ avec le modèle linéaire $Y_{ig} = \mathbf{X}_{ig}\beta_g + \varepsilon_{ig}$ qui respecte les hypothèses de base. Postulons que$\mathbf{X}_{ig}$ et β_g sont des vecteurs de dimension (1 x K) et (K x 1) où $K \epsilon \mathbb{N}^*$. La décomposition Blinder-Oaxaca (B-O) bipartite (en opposition à la décomposition tripartite) de ce modèle linéaire des moindres carrées ordinaires (MCO) est alors :

$$\overline{Y}_A - \overline{Y}_B = \Delta_{MCO} = \left[\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_B\right] \hat{\beta}_A + \overline{\mathbf{X}}_B \left[\hat{\beta}_A - \hat{\beta}_B\right]$$

où $\overline{Y}_g$ et $\overline{\mathbf{X}}_g$ sont les moyennes échantillonales de la variable dépendante et du vecteur des régresseurs, pour le groupe $g \epsilon \{A, B\}$, et où $\hat{\beta}_g$ est de le vecteur des estimateurs MCO. Cette décomposition est dû au fait que dans le cadre linéaire $E[Y_{ig}|\mathbf{X}_{ig}] = \mathbf{X}_{ig}\beta_g$ et donc $E[Y_{ig}] = \overline{\mathbf{X}}_g\hat{\beta}_g$. Par conséquent, en dérivant les différences des moyennes des groupes, nous arrivons aux résultats ci-dessus.

En opposition, il est clair que si $E[Y_{ig}|\mathbf{X}_{ig}] \neq \mathbf{X}_{ig}\beta_g$, comme dans le cas des modèles non-linéaires, alors la décomposition explicitée ci-dessus n'est plus valable. De façon générale, nous pouvons tout de même passer par les différences d'espérances conditionnelles entre les groupes pour obtenir la décomposition escomptée. Dans ce cas-ci, si le groupe A agit comme groupe de référence, nous avons :

$$\Delta_{A,NL} = [E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB})] + [E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})]$$

et si le groupe B sert de groupe de référence, nous avons :

$$\Delta_{B,NL} = [E_{\beta_B}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})] + [E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_B}(Y_{iA}|\mathbf{X}_{iA})]$$

où $E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA})$ est l'espérance conditionnelle de Y_{iA} et $E_{\beta_B}(Y_{iA}|\mathbf{X}_{iA})$ est l'espérance conditionnelle deY_{iA} évaluée à au vecteur de paramètre de l'autre groupe β_B, avec une interprétation symétrique pour les autres termes.

Dans un contexte général, nous allons alors vouloir estimer les valeurs $E_{\beta_g}(Y_{ig}|\mathbf{X}_{ig})$ et $E_{\beta_h}(Y_{ig}|\mathbf{X}_{ig})$, où $g, k \epsilon \{A, B\}$ avec $g \neq k$, puisque ce sont des quantités propre à la population que nous n'en observons qu'un échantillon. Des analogues empiriques que nous noterons $S(\hat{\beta}_g, \mathbf{X}_{ig})$ et $S(\hat{\beta}_k, \mathbf{X}_{ig})$ serviront d'estimateurs pour les espérances. Ces derniers prendront la forme de moyenne échantillonale

Après avoir explicité le cadre général, Bauer et Sinning élaborent alors les formulations pour divers modèles non-linéaires. Nous traiterons seulement de la méthode pour des modèles Logit et Probit puisque c'est cette dernière que nous utiliserons plus tard. Pour bien comprendre le raisonnement en oeuvre dans ces traitements, nous allons tracer le cadre d'un modèles ordonnés et, dans la revue de l'article suivant, nous resserons notre attention sur le contexte d'un modèle variable qualitative binaire.

Logit et Probit ordonnées

Premièrement, observons la décomposition de B-O dans le cas d'un modèle qualititif logistique ou probit ordonné avec $J \epsilon \mathbb{N}^*$résultats possibles. Posons le modèle de variable latente suivant avec l'index "O" sur le vecteur des paremètres pour le différencier des vecteurs précédents :

$$Y_{ig}^* = \mathbf{X}_{ig}\beta_{g,O} + \varepsilon_{ig,O}$$

où Y_{ig}^* n'est pas observable. Nous observons plutôt Y_{ig} qui prend des valeurs allant de 1 à J selon le système d'égalités suivant :

$$Y_{ig} = \begin{cases} 0 & si\, Y_{ig}^* \leq 0 \\ 1 & si\, 0 \leq Y_{ig}^* \leq \theta_1 \\ 2 & si\, \theta_1 \leq Y_{ig}^* \leq \theta_2 \\ (...) & \\ J & si\, Y_{ig}^* \leq \theta_{J-1} \end{cases}$$

où $\theta_1, \theta_2, ..., \theta_{J-1}$et le vecteur $\beta_{g,O}$ de dimension (1xK) doivent être estimés. En développant l'espérance conditionnelle de Y_{ig} évaluée au vecteur de paramètre, nous écrivons :

$$\begin{aligned} E_{\beta_{g,O}}(Y_{ig}|\mathbf{X}_{ig}) &= \sum_{j=1}^{J-1} j\{F(\theta_j - \mathbf{X}_{ig}\beta_{g,O}) - F(\theta_{j-1} - \mathbf{X}_{ig}\beta_{g,O})\} \\ &+ J\{1 - F(\theta_{J-1} - \mathbf{X}_{ig}\beta_{g,O})\} \end{aligned}$$

où $\theta_0 = 0$. Ces quantités ne sont pas observables; elles doivent alors être estimées de façon consistente. Les problèmes de consistence ne sont pas abordés dans cet article, alors nous supposons que les hypothèses nécessaires pour évoquer une loi faible des grands nombres. Par conséquent, l'analogue empirique de l'espérance conditionnelle de Y_{ig} sera :

$$\begin{aligned} S\left(\hat{\beta}_{g,O}, \mathbf{X}_{ig}\right) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J-1} j\left\{F(\hat{\theta}_j - \mathbf{X}_{ig}\hat{\beta}_{g,O}) - F(\hat{\theta}_{j-1} - \mathbf{X}_{ig}\hat{\beta}_{g,O})\right\} \\ &+ J\left\{1 - F(\hat{\theta}_{J-1} - \mathbf{X}_{ig}\hat{\beta}_{g,O})\right\} \end{aligned}$$

Cela nous donne alors une expression que l'on pourra utiliser pour obtenir un estimateur de la différence des moyennes selon les groupes. Notons qu'en remplaçant $\hat{\beta}_g$ par l'estimateur $\hat{\beta}_h$ dans $S\left(\hat{\beta}_{g,O}, \mathbf{X}_{ig}\right)$, où $h, g \epsilon \{A, B\}$ et $h \neq g$, nous obtenons les termes croisés. En utilisant ces estimateurs pour les différentes combinaisons de groupe, toujours avec les 2 groupes que nous avons supposé au départ, nous obtenons :

$$\hat{\Delta}_O = \left[S\left(\hat{\beta}_{A,O}, \mathbf{X}_{iA}\right) - S\left(\hat{\beta}_{A,O}, \mathbf{X}_{iB}\right)\right] + \left[S\left(\hat{\beta}_{A,O}, \mathbf{X}_{iB}\right) - S\left(\hat{\beta}_{B,O}, \mathbf{X}_{iB}\right)\right]$$

Cette égalité découle directement de la définition de Δ_O où $\hat{\Delta}_O$ est sont analogue empirique composé lui-même des analogues empiriques des espérances.

Remarquons finalement que les modèles logit et probit binomiaux sont un cas particulier des modèles logit et probit ordonnés. Ainsi, à partir de cette formulation de l'analogue empirique, nous pourrons obtenir l'estimateur correspondant pour des sous-cas.

En concluant, il est important de mentionné un problème relevé par les auteurs. La méthode de décomposition (B-O) ne donnera généralement pas les mêmes estimés selon notre choix du groupe de référence que le modèle de base soit linéaire ou non. Nous verrons dans des textes ultérieurs des solutions potentitelles à ce problème.

2.1.2 Robert W. Fairlie (2003)

Cet article de Fairlie élabore sur la décomposition lorsqu'elle est appliquée sur des modèles Logit ou Probit binomiaux. En posant $J = 1$ dans $E_{\beta_{g,O}}(Y_{ig}|\mathbf{X}_{ig})$ et $S(\hat{\beta}_{g,O}, \mathbf{X}_{ig})$, vue dans le cas (a) de la revue de l'article précédent, nous remarquons que nous obtenons :

$$\begin{aligned} E_{\beta_{g,O}}\left(Y_{ig}|\mathbf{X}_{ig}\right) &= \{F(\theta - \mathbf{X}_{ig}\beta_{g,O}) - F(-\mathbf{X}_{ig}\beta_{g,O})\} \\ &+ \{1 - F(\theta - \mathbf{X}_{ig}\beta_{g,O})\} \\ &= 1 - F(-\mathbf{X}_{ig}\beta_{g,O}) \\ &= 1 - \{1 - F(\mathbf{X}_{ig}\beta_{g,O})\} \\ &= F(\mathbf{X}_{ig}\beta_{g,O}) \end{aligned}$$

où l'on obtient la troisième égalité ci-dessus si la distribution est symétrique. Pour l'équivalent empirique de ce terme, nous dérivons de la même façon :

$$\begin{aligned} S\left(\hat{\beta}_{g,O}, \mathbf{X}_{ig}\right) &= \frac{1}{N_g}\sum_{i=1}^{N}\left\{F(\hat{\theta} - \mathbf{X}_{ig}\hat{\beta}_{g,O}) - F(-\mathbf{X}_{ig}\hat{\beta}_{g,O})\right\} + \left\{1 - F(\hat{\theta} - \mathbf{X}_{ig}\hat{\beta}_{g,O})\right\} \\ &= \frac{1}{N_g}\sum_{i=1}^{N}\left[1 - F(-\mathbf{X}_{ig}\hat{\beta}_{g,O})\right] \\ &= \frac{1}{N_g}\sum_{i=1}^{N}\left[1 - \left\{1 - F(\mathbf{X}_{ig}\hat{\beta}_{g,O})\right\}\right] \\ &= \frac{1}{N_g}\sum_{i=1}^{N}\left\{F(\mathbf{X}_{ig}\hat{\beta}_{g,O})\right\} \end{aligned}$$

où, encore une fois, l'on obtient la troisième égalité ci-dessus si la distribution est symétrique.

Après cette simplification, nous nous retrouvons dans le cas exacte décrit par Fairlie. En d'autres mots, avec les groupes $g \epsilon \{A, B\}$, si $\hat{Y}_{ig} = F(\mathbf{X}_{ig}\hat{\beta}_{g,O})$ nous obtenons :

$$\begin{aligned}\overline{Y}_A - \overline{Y}_B &= \left[\frac{1}{N_A}\sum_{i=1}^{N_A} F(\mathbf{X}_{iA}\hat{\beta}_A) - \frac{1}{N_B}\sum_{i=1}^{N_B} F(\mathbf{X}_{iB}\hat{\beta}_A)\right] \\ &+ \left[\frac{1}{N_B}\sum_{i=1}^{N_B} F(\mathbf{X}_{iB}\hat{\beta}_A) - \frac{1}{N_B}\sum_{i=1}^{N_B} F(\mathbf{X}_{iB}\hat{\beta}_B)\right] \\ &= \hat{\Delta}_{binom.}\end{aligned}$$

Notons qu'il ne s'agit pas des mêmes notations que celles utilisées dans le papier de Fairlie, nous les avons adaptés pour assurer la transition entre les textes. Encore une fois, nous pouvons trouvons un résultat symétrique si nous cherchons $\overline{Y}_B - \overline{Y}_A$. Le terme dans les premières braquettes du côté droit de l'égalité peut être interprété comme la différence estimée dans les moyennes dû aux disparités entre les groupes dans leurs régresseurs tandis que le terme dans les deuxièmes braquette peut être interprété comme la différence estimée inexpliquée qui serait dû à des caractéristiques propre aux groupes. C'est cette deuxième partie que l'on associe typiquement à la «discrimination» entre deux groupe selon le contexte.

2.2 Traitement la décomposition par STATA

Ces deux articles, parus de pair dans la même édition de ''The STATA Journal'', traitent, dans le cas du premier, du procédé par lequel on opère une décomposition de Blinder-Oaxaca dans le cas d'un modèle linéaire sur le logiciel STATA et, dans le cas du second, de la façon par laquelle on peut opérer cette même décomposition dans le cadre d'un modèle non-linéaire encore une fois sur STATA.

2.2.1 Ben Jann (2008)

L'article de Jann, bien que non applicable directement dans la situation que l'on cherchera d'étudier, fournit une quantité intéressante d'intuition quant à la façon dont la décomposition prend lieu. En effet, en plus d'expliciter la démarche dans une cadre simplifié, ce papier nous met aussi en garde face aux divers problèmes qui peuvent se présenter lors de la décomposition et de la mise en place du cadre intuitif des interprétations.

Par définition, la différence de moyenne, que l'on indexe par "lin" pour rapeller que l'on parle du cas linéaire, est $\Delta_{lin} = E(Y_A) - E(Y_B)$. Par conséquent, si $Y_g = X_g\beta_g + \varepsilon_g$, où $g \epsilon \{A, B\}$ et où l'absence d'indice i sur Y_g et X_g indique que Y_g est un vecteur $(N_g \mathrm{x}\, 1)$ et que X_g est une matrice $(N_g \mathrm{x}\, K)$, nous pouvons développer Δ_{lin} . En supposant que $E(\varepsilon_g) = 0$, nous obtenons :

$$\Delta_{lin} = E(X_A\beta_A + \varepsilon_A) - E(X_B\beta_B + \varepsilon_B) = E(X_A)'\beta_A - E(X_B)'\beta_B$$

Après de simples manipulations, nous pouvons réécrire Δ_{lin}comme :

$$\begin{aligned}\Delta_{lin} &= [E(X_A) - E(X_B)]'\,\beta_B + E(X_B)'\,[\beta_A - \beta_B] \\ &+ [E(X_A) - E(X_B)]'\,[\beta_A - \beta_B]\end{aligned}$$

Cette décomposition est analogue à celle montrée plus haut dans le cas non-linéaire dans l'équation (1) où l'on a simplifier plusieurs termes en utilisant la linéarité du modèle. Notons qu'elle prend le groupe B comme groupe de référence, mais nous pouvons toujours effectuer une décomposition d'une manière symétrique en utilisant le groupe A comme groupe de référence. Dans un article subséquent, il sera question des problèmes que l'on encourt lorsque l'on utilise cette méthode plutôt que la méthode propre aux cas non-linéaires ; de là se trouve l'importane d'explicité cette décomposition.

Un seconde forme est courrament utilisée selon Jann. Cette dernière fais intervenir un paramètre que nous noterons β^*. Ce coefficient est un vecteur non-discriminatoire que l'on désigne pour exprimer l'apport des différences dans les paramètres prédits pour les groupes A et B. Avec ce vecteur, nous réécrivons la décomposition comme :

$$\begin{aligned}\Delta^*_{lin} &= \left\{[E(X_A) - E(X_B)]' \beta^*\right\} \\ &+ \left\{E(X_B)' [\beta_A - \beta^*] + E(X_B)' [\beta_A - \beta^*]\right\}\end{aligned}$$

Cette deuxième méthode permet de diviser notre terme en deux parties. À l'intérieur de la première braquette du côté droit de l'équation (4), nous avons la partie de la différence que l'on attribue à la disparité dans les valeurs des régresseurs entre les groupes. Le terme à l'intérieur du deuxième ensemble de braquet du côté droit de (4) est, quant à lui, représenté comme la partie non-expliquée de la différence. On associe cette valeur à la discrimination qui survient entre les groupes observés. En effet, cette représentation est fréquemment utilisée dans la littérature qui traite de la discrimination (par exemple entre les sexes ou entre les ethnies).

Actuellement, bien qu'il existe des façons prévalentes de définir β^*, il n'y a pas de consencus quant à la méthode pour l'obtenir. Dans les cas où il serait raisonnable de croire que la discrimination est dirigée principalement pour (ou contre) un groupe, on suggère de poser $\beta^* = \beta_g$ où g serait le groupe discriminé. Cordelia W. Reimers en 1983 proposent de poser $\beta^* = 0.5\beta_A + 0.5\beta_B$ tandis que Jeremiah Cotton en 1988 suggère de le définir comme $\beta^* = \frac{N_A}{N}\beta_A + \frac{N_B}{N}\beta_B$. Finalement, une autre avenue serait de définir β^* comme la valeur de l'estimateur obtenue en effectuant la régression avec les deux groupes que l'on désire observer.

2.2.2 Mathias Sinning, Markus Hahn et Thomas K. Bauer (2008)

L'article de Sinning, Hahn et Bauer agit, quant à lui, comme une extension à la fois à l'article de Jann, ci-dessus, et à l'article de Bauer et Sinning, exposé dans la section précédente. Ce prolongement nous explique la façon générale d'aborder sur STATA la décomposition convoitée dans un cadre non-linéaire. Cet article sera d'une grande utilitée dans notre étude lorsqu'il viendra le temps d'appliquer le modèle théorique pour obtenir les résultats. Par ailleurs, l'exposition de ce dernier mets en évidence le fait que l'approche non-linéaire réflete l'approche linéaire.

Les auteurs expriment un analogue non-linéaire plus direct à l'équation (3) en manipulant quelque peu l'équation (1). En considérant encore la décomposition avec le groupe A comme groupe de référence, ils obtiennent alors :

$$\begin{aligned}\overline{Y}_A - \overline{Y}_B &= \left[E_{\beta_B}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\right] \\ &+ \left[E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\right] \\ &+ \left[E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_B}(Y_{iA}|\mathbf{X}_{iA})\right] + \left[E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\right]\end{aligned}$$

Cette représentation est d'un grand interêt, car elle permet de voir les liens entre le cas linéaire et non-linéaire. Ils interprétent alors les termes de cette égalité selon la provenance de ces effets. Le premier terme du côté droit de cette égalité est interprétable comme «l'effet de dotation». Les différences provenants des différentes valeurs des régresseurs sont incorporées dans ce terme. Le deuxième terme comprend l'effet dans la différence des moyennes venant de l'écart des paramètres des régressions séparés selon le groupe. La dernière ligne de l'égalité ci-dessus quant-à-elle représente la partie de la différence venant de l'interaction entre l'effet de la dotation et l'effet des coefficients.

La décomposition effectuée dans l'équation (4) de ce texte a elle aussi un analogue pour les modèles non-linéaires. Nous l'obtenons de la même façon que celle par laquelle nous avons effectué les développement précedents l'équation (2). Notons le Δ^*_{NL}

$$\begin{aligned}\Delta^*_{NL} &= \overline{Y}_A - \overline{Y}_B \\ &= \{E_{\beta^*}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta^*}(Y_{iB}|\mathbf{X}_{iB})\} \\ &+ \{E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta^*}(Y_{iA}|\mathbf{X}_{iA})\} \\ &+ \{E_{\beta^*}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\}\end{aligned}$$

Les élements de l'équation (5) ont des interprétations similaires à ceux de l'équation (4). Notons que $E_{\beta^*}(Y_{ig}|\mathbf{X}_{ig})$ représente l'espérance conditionelle de Y_{ig} évaluée à β^*. Tel que nous l'avons vu dans la revue de l'article précédent, la façon de dériverβ^* varie selon les opinions des chercheurs, mais il s'agit toujours d'une moyenne pondérée des vecteurs des paramètres de la régression des deux groupes. C'est avec ce point de vue que les auteurs définissent $\beta^* = \Omega\beta_A + (I - \Omega)\beta_B$ où Ω est une matrice de poids, I est une matrice identité et ces deux matrices sont de dimensions (K x K). Notons que par définition d'une matrice de poids, $\Omega_{ij}\,\epsilon[0,1]\ \forall i,\ j\,\epsilon\{1,2,3,...,K\}$. Cette notation permet d'exprimer les divers cas que nous avons énoncé dans le résumé précédent en plus de combinaisons plus complexes en utilisant les valeurs appropriées dans la matrice de poids.

En addition, il est important de remarquer qu'avec cette décomposition, en plus des quatres analogues empiriques dont nous avons fait état du besoin plus haut, nous aurons besoins d'estimateurs consistents également pour $E_{\beta^*}(Y_{iB}|\mathbf{X}_{iB})$ et $E_{\beta^*}(Y_{iA}|\mathbf{X}_{iA})$. Pour le reste de notre exposé, nous les noterons $S(\beta^*, \mathbf{X}_{ig})$ pour $g\,\epsilon\{A, B\}$.

2.3 Application pratique de la décomposition de B-O

2.3.1 Vani Borooah et Sriya Iyer (2005)

Ce document de Borooah et Iyer propose une analyse par la méthode de Blinder-Oaxaca appliquée dans le cadre d'un modèle logistique de l'enrôlement de garçons dans des écoles en Indes selon certaines caractéristiques socio-démographiques comme la religion et la provenance géographique. En plus de constituer une application instructive de cette décomposition dans un cadre non-linéaire, cet exposé propose également une solution au problème du choix du coefficient de $\beta^\star$. Cet article est intéressant, car il expose de façon concrète la modèlisation d'une décomposition dans le cadre Logit.

Dans le modèle de régression de cet article, nous notons J le nombre de régresseurs, $k\,\epsilon\{1, 2..., K\}$ un des groupes (tous étant mutuellement exclusifs et collectivement exhaustifs) et N_k le nombre d'individus dans le groupe k. Définissons ENR_i^k comme une variable binaire prennant la valeur 1 si l'individu i membre du groupe k est inscrit à l'école et 0 dans le cas contraire ; ceci est la variable dépendante du modèle général de Borooah et Iyer. Si nous posons les hypothèses propres au modèle Logit et supposons implicitement un modèle latent, nous pouvons alors écrire la probabilité que cet individu arbitraire soit inscrit à l'école comme :

$$P(ENR_i^k = 1) = \frac{exp(\mathbf{X}_i^k\beta^k)}{1 + exp(\mathbf{X}_i^k\beta^k)} = F(\mathbf{X}_i^k\beta^k)$$

où $\mathbf{X}_i^k = \left(X_{i1}^k,\ X_{i2}^k,\ ...,\ X_{iJ}^k\right)$ est le vecteur des régresseurs de l'individu i faisant partie du groupe k et où β^k est le vecteur (J x 1) des paramètres de la régression sous-jacente.

Étant donné cette représentation, il est alors direct de définir la probabilité prédite moyenne pour un individu du groupe k d'être inscrit à l'école comme :

$$\overline{ENR}_k \equiv \overline{P}(\mathbf{X}_i^k, \hat{\beta}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k} F(\mathbf{X}_i^k\hat{\beta}_k)$$

où le côté droit de l'identité fait référence à la moyenne échantillonale de la probabilité prédite et où $\hat{\beta}_k$ est le vecteur (J x 1) des paramètres estimés de la régression Logit sur un échantillon restreint aux observations membrent du groupe k. Avec l'exposé que nous avons donné plus haut, nous pouvons

exprimer la décomposition d'Oaxaca-Blinder dans ce contexte Logit de la sorte pour deux groupes mutuellement exclusifs et collectivement exhaustifs que nous notons K et H :

$$\begin{aligned}\overline{ENR}_H - \overline{ENR}_K &= \overline{P}(\mathbf{X}_i^H, \hat{\beta}_H) - \overline{P}(\mathbf{X}_i^K, \hat{\beta}_K) \\ &= \left\{\overline{P}(\mathbf{X}_i^K, \hat{\beta}_H) - \overline{P}(\mathbf{X}_i^K, \hat{\beta}_K)\right\} \\ &+ \left\{\overline{P}(\mathbf{X}_i^H, \hat{\beta}_H) - \overline{P}(\mathbf{X}_i^K, \hat{\beta}_H)\right\}\end{aligned}$$

et de façon symétrique si l'on désire l'autre groupe comme référence. Il s'agit là d'un exemple concret des développements explicités plus haut. Tout comme l'interprétation faite dans le cas général, le premier terme à la droite de la deuxième égalité représente la part de la différence dans les probabilités moyennes prédites attribuables aux différences dans les coefficients prédits et le deuxième terme représente la part qui serait issue de la différence dans les régresseurs entre les groupes.

Cet exemple concret est intéteressant, car ce dernier nous permet d'observer le raisonnement sous-jacent à la décomposition. L'article de Borooah présente aussi des méthodes pour effectuer des décompositions dans des contextes où plus nous avons plus de deux groupes d'intéret. Cette dernière ne sera pas explicitée ici, mais nous notons néanmoins qu'elle pourrait être pertinente dans des développements subséquents au modèle que nous exposons plus bas.

2.4 Littérature sur l'étude de la demande d'éducation supérieure

Au-delà de la méthode de traitement, il est aussi primordial de pouvoir justifier par la littérature existante la structure que va prendre notre analyse de la demande d'éducation supérieure, ainsi que des variables et divers points qu'il sera important de considérer. Dans cette sous-section, nous commencons par étudier l'article de Finnie et Mueller, car notre analyse en sera grandement tributaire dû à la similarité des données et de la région traitée. Pour conclure la revue de littérature, nous effectuons une synthèse sommaires des modèles et intuitions développées à cet effet.

2.4.1 Ross Finnie et Richard E. Mueller (2008)

Cette recherche, utilisant des données de la ''Youth in Transition Survey'' étudie les déterminants de l'enrôlement dans l'éducation supérieure chez les jeunes canadiens. Elle suggère un cadre de travail général duquel nous nous inspirerons pour étudier les facteurs sociodémographiques décisifs dans l'enrôlement d'un individu dans des programmes d'éducation supérieure. Par ailleurs, bien qu'elle n'utilise pas de méthode de décomposition à proprement parler, c'est-à-dire autre que l'usage de variables nominales, elle fournit un apport théorique qui pourra guider nos efforts et, par la suite, permettre la comparaison de nos résultats avec les siens pour, ainsi, en découler des interprétations.

Le cadre de base de leur recherche est un modèle Logit multinomial où la variable dépendante prend trois valeurs selon la décision de l'individu d'intégrer une université à la suite de ses études dans une High School, d'intégrer un centre de formation professionnelle ou encore de ne pas prendre part à des études supplémentaires. Nous désirons s'inspirer de ce traitement dans notre séparation de la variable dépendante à l'exception du fait que l'on serait plutôt intéressé à observer la réalisation «s'intègre à un programme d'étude universitaire» ou «ne s'intègre pas à un programme d'étude universitaire».

Parmi les variables indépendantes utilisées dans ce texte, on compte la scolarité parentale, le revenu parental, la localisation de leur High School (à savoir dans quelle province elle se trouve et si elle est située en milieu urbain), la langue de l'individu, des facteurs démographiques de l'individu (à savoir s'il est citoyen canadien, s'il fait partie d'une minorité visible ou la langue qu'il utilise à la maison). L'interprétation de l'effet de ces variables s'avère être le point central de leur recherche.

Outre les variables décrites ci-dessus, deux autres types de variables sont ajoutées dans le modèle de Finnie et Mueller. Le premier type est relié à l'environnement dans lequel le répondant a évolué. Parmi ces variables, nous comptons notamment des mesures du comportement des parents et des mesures de la qualité de l'école pré-universitaire de l'individu. Le deuxième type est constitué de mesures des

qualités idiosyncratiques de l'individu tel ses résultats dans des tests de lecture et de sciences ou des pointages de tests tentant de refléter le comportement (par exemple l'estime de soi) du répondant.

À l'aide d'une base de données exceptionnelle, ils parviennent à tirer un bon nombre de résultats. Entres autres, notons qu'ils trouvent que le fait que les deux parents du répondant vivent ensemble, que le fait que le répondant vive en milieu urbain et que le répondant fasse partie d'un minorité ethnique ont tous des effets positifs sur la probabilité prédite d'atteinte d'éducation supérieure. De plus, on observe que cette même probabilité prédite est significativement croissante avec le revenu parental et la quantité d'éducation des parents du répondant.

Finalement, un point important à noter, dû à l'absence de ces variables dans la base de données à notre disponibilité, est l'importance des caractéristiques intellectuelles et familiales des répondants. Finnie et Mueller observent, tel que l'on aurait pu supposer, que la probabilité prédite de poursuivre ses études augmente de façon statistiquement et économiquement significative lorsque les mesures du bon comportement des parents sont hautes et lorsque les mesures de capacités cognitives de l'individu sont élevées. Un tel constat jumelé à l'absence de telles données à notre disposition nous indique qu'il sera impératif de s'assurer que la corrélation théorique entre ces variables et celle de notre modèle soient nulles. Dans le cas contraire, les coefficients que nous reporterons pourraient être biaisés. Cette éventualité devra alors potentiellement être traitée, par exemple, à travers des variables instrumentales.

2.4.2 Sommaire de textes sélectionnés

Avant de terminer cette revue de littérature, notons brièvement certains faits intéressants dans l'étude de la demande d'éducation supérieure et certains articles clés qui ont su guider notre réflexion.

En premier lieu, Jiménez et Salas-Velasco (2000) soulèvent, de leur côté, non seulement l'importance des capacités académiques du répondant, mais aussi sa perception de ses propres capacités. Un tel constat est cohérent avec les mesures d'estime de soi utilisée par Finnie et Mueller. Par ailleurs, les autres valeurs d'intérêt qu'ils posent dans leur modèle sont similaires à celles utilisées par ces mêmes auteurs. Cela vient appuyer l'importance de ces dernières dans un tel modèle.

Deuxièmement, Drolet (2005) remarque dans une étude échelonnée de 1993 à 2001 au Canada qu'une mesure potentiellement meilleure dans l'évaluation de l'impact du revenu familial sur la probabilité d'enrôlement universitaire des enfants serait le revenu permanent plutôt que le revenu annuel du ménage. Un tel constat est intuitif, car le revenu permanent d'un ménage pourrait être supposé comme plus évocateur du niveau de vie et des perspectives économiques d'un famille qu'un revenu annuel. Ses résultats réitèrent, encore une fois, l'importance des caractéristiques familiales des répondants dans ce type d'étude. En particulier, l'éducation parental, les moyens financiers et le type de famille des répondants s'avèrent à la fois statistiquement et économiquement significatifs.

À la lumière des observations faites par les textes présentés ci-dessus, nous pouvons accorder une certaine crédibilité au cadre dans lequel nous allons effectuer cette recherche. Plusieurs facteurs observés comme important dans ces régressions ne seront toutefois pas à notre disposition. Nous devons, par conséquent, nous assurer tout au long de cette recherche que cela n'entraîne pas d'effets indésirables.

3 Modèle théorique

La théorie évoquée dans la littérature portant sur l'investissement en capital humain, et plus spécialement celle arguée dans les textes que nous avons mentionnés dans la section précédente, nous aidera à définir quelles variables explicatives devront être inclusent ou non dans notre modèle. Ces considérations passent par le modèle à partir duquel nous allons effectuer les régressions et elles se prolongent jusqu'aux décompositions que nous allons faire. Derrière ces dernières, nous chercherons à voir si les effets de certaines variables d'intérêt, comme le revenu, changent si l'agent moyen fait parti ou non d'un groupe, par exemple les agents dont la famille présente des antécédents universitaires. Notons que ces questionnements se reposent sur des intuitions économiques et des observations empiriques. Il

s'agit donc de ces dernières que nous exposons dans les paragraphes suivants.

3.1 Accumulation de capital humain (LUCAS & KRUSSEL & Gibbons)

Plusieurs modèles macroéconomiques et microéconomiques proposent des cadres simplificateurs qui décrive le processus décisionnel d'un décideur faisant face à un choix d'investir ou non dans du capital humain. Des modèles en économie du travail et en théorie des jeux proposent aussi des structures dans lesquels on peut voir les coûts encourus par un individu qui désir étudier selon son abilité innée. Ces derniers s'avèrent intéressantes pour encadrer notre réflexion à ce sujet.

Entre autres, dans la théorie de la croissance endogène, en macroéconomie, le modèle proposé par Robert Lucas Jr. (1988) évoque une économie à un nombre de période infinie dans laquelle un consommateur représentatif prend une décision d'accumulation de capital humain à chaque période. Dans le cadre du modèle, ce choix peut être représenté comme une décision dans la formation. Le modèle proposé dans cet article comprend une infinité de période, une quantité continue de capital humain pouvant être choisit et un taux de dépréciation à ce capital (pouvant être vu comme l'imperfection de la mémoire ou de la capacité de rétention d'information). En guise de mise en contexte, résumons rapidement ce modèle. Examinons le cadre d'un planificateur social, puisque la dynamique de marché n'est pas notre principal soucit ; nous cherchons à représenter l'arbitrage sous-jacent à la décision d'étudier, et donc d'accumuler du capital humain, ou non à la période t. Dans une version simplifiée en temps discret tributaire de Krussel (2004) pour faciliter l'exposition, ce dernier comprend :

1. Une fonction de production $y_t = F(H_t, K_t)$ où H_t et K_t représentent la quantité de capital humain et de capital physique respectivement dont la décision provient de la période précédente

2. Des équations de transition pour le capital humain et physique de la sorte :

$$\begin{aligned} K_{t+1} &= (1-\delta_K)\,K_t + I_t^K \\ H_{t+1} &= (1-\delta_H)\,H_t + I_t^H \end{aligned}$$

où I_t^K et I_t^H représente l'investissement en capital physique et l'investissement en capital à la période t respetivement et δ_g, $g \epsilon \{K, H\}$, représente la dépréciation de ce capital. Par exemple, la dépreciation en capital humain pourrait être vu comme l'imperfection de la capacité de rétention d'information d'un individu.

3. Une fonction d'utilité totale comme une somme escomptée de fonctions d'utilités instantanés additivement séparables dans leurs arguments comme par exemple :

$$U\left(\{c_t\}_{t=1}^{\infty}\right) = \sum_{t=1}^{\infty} \beta^t u(c_t)$$

où c_t est la consommation de notre consommateur représentatif à la période t dont il choissira la séquence pour $t = 0, 1, 2, ...$ afin de maximiser son utilité totale et où $\beta \epsilon (0, 1)$ est le facteur d'escompte.

4. Des contraintes dont une contrainte de ressources pour chaque période du type :

$$c_t + I_t^K + I_t^H = F(H_t, K_t) \qquad \forall\, t = 1, 2, 3, ...$$

Étant donné ces contraintes et ces fonctions de transition, le planificateur choisit la séquence de capital humain, de capital physique et de consommation qui maximisera l'utilité totale du consommateur représentatif. Cette représentation simpliste explicite bien les coûts et bénifices en jeu dans la décision d'investir ou non dans du capital humain à chaque période. À la période t, le choix d'investir plus en capital humain implique une moins grande quantité de consommation courante ou d'investissement en capital physique, mais une plus grande production par la suite dû à la quantité accrue de capital

humain dans la fonction de production. Nous ne résoudrons pas le problème de maximisation explicitement, mais l'intuition économique nous suggère que ces choix dépenderont de la fonction de production et des rendement marginaux de chaque capital. Ce modèle précédent explicite le raisonnement quant au coût d'opportunité présent lors de la décision d'accumulation de capital humain et c'est de là que vient son principal intérêt dans le cadre qui nous concerne.

La situation que nous désirons représenter comprend plusieurs différences avec le modèle décrit ci-dessus. En effet, la quantité de capital humain observable est discrète (un nombre fini et dénombrable de diplôme sont obtenable pour tout individus) et le bénéfice suivant la décision d'accumuler du capital humain est décallée (le temps d'obtenir le diplôme ou de compléter ses études) tandis que les coûts en termes de consommation sacrifiée, d'expérience de travail décalée ou d'accumulation de capital physique diminuée entrent en vigueur dès le début des études et ne sont pas nécessairement constants. Aussi à prendre en compte, tel que plusieurs modèles de la théorie des jeux le suggèrent [2], est l'hypothèse que le coût en terme d'effort pour accumuler du capital humain peut varier selon le niveau d'abilité inné de l'individu. La modélisation de cette dernière parenthèse nécessiterait d'incorporer un coût d'effort dans la fonction de transition et potentiellement une valorisation du loisir dans la fonction d'utilité instantannée. Comme il s'agirait de considérations idiosyncratiques, il serait nécessaire alors de développer ce modèle de façon à ce qu'il prenne en compte des agents hétérogènes.

Malgré ces différences, les dynamiques en place restent les mêmes et la structure entourant la décision est identique. Le choix de l'individu se résume à décider de la quantité d'éducation maximisant son utilité étant donné les facteurs lui étant spécifiques. Par ailleurs, dès qu'un individu entame des études universitaire, ce dernier dévoile sa décision d'optimisation à ce stade-ci de sa vie. Cette décision, ou plutôt les facteurs ayant poussé cette dernière à prendre forme, est le point focale de notre étude. Avec cela en tête, nous pouvons définir un cadre similaire à celui exposé ci-dessus, mais prenant en compte notre désir de déterminer les facteurs idiosyncratiques motivants ces décisions. Notre intérêt est l'observation d'un choix binomial dans le temps dans un contexte où, une fois que l'individu à atteint l'université, ce dernier dévoile son désir de poursuivre ou non ses études. Nous explicitons alors le cadre, tributaire du modèle de croissance endogène de Lucas Jr. , qui encradre notre raisonnement :

Supposons que les agents sont rationnels, c'est-à-dire que leur préférences sont complètes et transitives, et que leurs goûts sont représentable par une fonction d'utilité de von Neumann-Morgenstern strictement croissante dans la quantité de monnaie disponible et dans la qualité de leurs perpectives financières. La théorie de l'accumulation de capital humain propose comme piste d'explication l'idée que l'individu encourt un coût en utilité au temps présent (où ce dernier peut être vu comme un coût d'opportunité autant au point de vue de consommation renoncée que du point de vue d'une moins grande quantité de loisir aux périodes suivant la prise de la décision) en échange d'un revenu espéré plus grand dans le futur. Dans ce cadre, la décision d'étudier ou non est schématisée de la même façon que le serait une décision d'investissement.

Nous postulons alors qu'un individu donné i, dont les préférences seraient représentées par une fonction $U_i : \mathbb{R} \to \mathbb{R}$, choisirait $H_i^{optimal}$ tel que :

$$H_i^{optimal} = \underset{H_i \epsilon \{0,1\}}{argmax} \, \mathbb{E}\left\{U_i(H_i)\right\}$$

où $U_i(H_i) = U(H_i|x_1, x_2, ..., x_k)$ et où $H = 0$ représente la décision de ne pas étudier à l'université tandis que $H = 1$ représente l'évènement complément. En d'autres termes, nous supposons que chaque individu présente des préférences de choix d'éducation représentées par une même fonction $U : \mathbb{R} \to \mathbb{R}$ à une série de paramètres idiosyncratiques près. Le moment de décision serait modélisé comme $t = 0$. Sans aller plus en détails, nous supposons que cette fonction représente l'utilité totale espérée associée à la décision prise. L'espérance dans l'identité ci-dessus représente le fait que les coûts et bénéfices, au moment de la décision, sont incertains ; nous supposons que l'individu a des espérances rationnelles à cet égard. Les considérations subséquentes reliées aux facteurs motivant la quantité de diplômes universitaires choisit ou d'années d'université sont laissées à une étude subséquente. Nous arguons que seule la décision d'entamer des études universitaires est abordée dans le cadre que nous posons et que d'observer le sujet de la sorte est temporellement rationnel ; nous ne tentons pas de déterminer qu'est-ce qui pousse un individu à faire un doctorat plutôt qu'un baccalauréat. Finalement, il est

2. Par exemple, tel qu'exposé dans Gibbons (1992)

raisonnable selon nous de supposer que les facteurs affectant la décision de commencer un baccalauréat sont similaires, jusqu'à un certain point, à ceux qui influencent le choix de continuer ou non vers une maîtrise ou un doctorat par la suite.

Avec une telle représentation, nous postulons que le décideur choisira d'entreprendre des études supérieures s'il estime que de poser cette décision lui fournira une plus grande utilité espérée escomptée que celle qu'il aurait en choisissant de ne pas étudier.

3.2 Le modèle théorique

Finnie et Mueller (2008) proposent un cadre évoquant plusieurs variables explicatives dans la décision d'études supérieures. Outre cet article, nous avons aussi observé des modèles proposés par Jiménez et Salas-Velasco (2000) et Drolet (2005). À travers la littérature, nous dénotons un certain nombre de variables récurrentes utilisées dans ces modèles. En utilisant ces dernières, nous avons fait un triage nous permettant de formuler un modèle théorique à la fois cohérent avec l'intuition et avec les articles duquel il est tributaire.

La structure de base du modèle latent que nous avons défini, en se basant sur la littérature[3], prend la forme suivante :

$$y \quad = \quad \alpha + \mathbf{F^T}\beta + \mathbf{D^T}\delta + \mathbf{G^T}\gamma + \varepsilon$$

où y est la décision binaire en éducation, où $\mathbf{F}$ est un vecteur contenant des variables comprenant les considérations familiales, telles le niveau de richesse de la famille, le niveau de valorisation de l'éducation ou le type de famille d'un individu arbitraire, de dimension $(K_1 \times 1)$, où $\mathbf{D}$ est un vecteur contenant des variables comprenant les considérations démographiques, telles la langue parlée par le répondant, son ethnie, son sexe ou son âge, de dimension $(K_2 \times 1)$ et où $\mathbf{G}$ est un vecteur contenant des variables comprenant les considérations géographiques, telles la province de résidence, le pays de provenance ou la proximité des grands centres, de dimension $(K_3 \times 1)$. Le terme α représente la constante du modèle tandis que β, δ et γ, respectivement de dimension $(K_1 \times 1)$, $(K_2 \times 1)$ et $(K_3 \times 1)$, représentent les vecteurs de paramètres associés aux groupes des variables familiales, des variables démographiques et des variables géographiques.

En résumé, le modèle sous-jacent à notre étude suppose alors un lien de cause à effet entre la réalisation, ou l'absence de réalisation, «prendre part à des études universitaires» et le revenu du ménage dans lequel l'individu à grandit, le niveau d'étude des parents de ce dernier, la proximité des centres urbains où les universités sont majoritairement centrées, le sexe de l'individu et d'autres variables moins prépondérantes. Les membres de ce groupe font partie des variables explicatives d'intérêt de notre étude.

3.3 Les prédictions de la théorie

La théorie économique et les articles que nous avons exposés plus haut nous suggèrent avant même de procéder à l'estimation de nos modèles certaines directions que pourraient prendre des coefficients et des discriminations.

Premièrement, il est raisonnable de supposer qu'un individu dont les parents sont familiers avec le milieu universitaire aurait à sa disposition plus de ressources pour effectuer un choix éclairé toutes chose étant égales par ailleurs. De plus, on pourrait conjecturer que, en moyenne, si des individus ont poursuivi des études universitaires ces derniers dévoilent une certaine valorisation l'éducation qu'ils sont suceptible de transmettre à leur progéniture. Pour ces raisons, on serait à même de s'attendre, au minimum, un effet non-négatif de l'éducation parentale sur la probabilité d'un individu moyen de fréquenter l'université. Dans la même veine de raisonnement, si un individu vit avec ses parents pendant la période de prise de décision, il n'est pas impossible que ce dernier fasse face à des figures

3. Principalement Finnie et Mueller (2008) et Drolet (2005)

d'autorités plus présentent et à un cadre plus stricte. Il serait alors légitime de croire que la présence parentale aurait un effet non-négatif dans la probabilité de réaliser des études universitaires.

Deuxièmement, le revenu disponible dans un ménage est potentiellement facilitateur dans l'accessibilité d'options de consommation et d'investissement. Comme il existe un coût financier aux études, pour une relation parent-enfant donnée, il est légitime d'imaginer que les fonds disponibles à cet individu auraient un effet non-négatif dans sa décision d'éducation, car ces derniers pourraient venir atténuer le coût d'opportunité de cette décision. Cela nous mène à prévoir une probabilité de réalisation croissante avec le revenu parental. Dans un système de prêts et bourses où les prêts étudiants sont fonction du revenu parental, il sera cependant nécessaire d'être vigilant. Imaginons un individu arbitraire dont les parents sont bien nantis, mais dont ces derniers ne contribuent pas financièrement aux études de leur enfant. Cet enfant n'aura potentiellement pas accès à des prêts qui auraient été nécessaires à la poursuite de ses études. L'effet peut alors être constitué de deux partis contradictoires. Outre ces considération, une grande fortune familiale pourrait aussi avoir un effet négatif dans la perception de la nécessité du travail chez la progéniture de ce ménage fictif. Une autre considération vient de la corrélation évidente entre le revenu du ménage et du nombre de parents vivant dans ce dernier. Étant donné ces points, il sera intéressant de faire des variables d'interaction pour capturer les divers effets en jeu à ce niveau.

Troisièmement, on observe typiquement une présence accrue des dames dans les université comparativement aux hommes. Par conséquent, nous attendons *ceteri paribus* de voir une probabilité prédite plus grande pour une femme moyenne de prendre part à des études universitaires comparativement à un homme. Dans un autre ordre d'idée, si un individu vit près des sources d'éducation universitaire, nous nous attendrons que, toutes choses étant égales par ailleurs, sa probabilité de réaliser des études universitaires sera plus grande que quelqu'un vivant en milieu rural, car le coût d'opportunité lié au déplacement sera moindre. Par exemple, un individu vivant à Fermont encourrait plus de coûts à assister à des cours universitaires qu'un montréalais,*ceteri paribus*.

Finalement, nous examinons si certaines formes de discrimination, par exemple en faveur des gens mieux nantis ou en faveur des femmes, prennent place selon notre modèle. Nous nous attendons que l'appartenance à certains groupes entraîne des différences entre la probabilité moyenne prédite de ces individus et celle des gens ne faisant pas partie de ces groupes donnés. Entre autre, nous supposons que le fait d'être une femme et le fait d'avoir des parents connaissant le milieu universitaire entraînera une probabilité d'études universitaires prédite plus grande que les gens ne faisant pas partie de ces groupes. Il est aussi raisonnable d'imaginer la présence de certaines particularités non-observées propre aux membres de ces groupes respectifs. Par conséquent, nous suspectons que les femmes et les gens ayant des parents universitaires verront une «discrimination» positive en leur endroit ou, entre d'autre termes, qu'il existe une différence prositive dans la probabilité prédite qui n'est pas attribuable à la différence entre les régresseurs des membres et des non-membres de ces groupes

Dans la section suivante, nous résumerons de quelle façon ces diverses variables d'intérêt sont disponible pour le cadre de notre étude. Par la suite, nous procéderons à l'explicitation des méthodes économétriques que nous utiliserons. Nous finirons par discuter des résultats que nous avons obtenus.

4 Statistiques descriptives

L'étude ci-présente utilise les données de l'enquête sociale générale (ESG) du Canada réalisée en 2011. Cette enquête, réalisée par Statistique Canada, comporte plus de 950 variables et plus de 22 000 observations. Dans la section des appendices de ce travail, une tableau contenant la description des diverses variables utilisées dans cette recherche ainsi que leurs aronymes peut être trouvée.

4.1 Population observée

Canadiens selon une série de groupes d'âge

Cette étude fait face à une limitation dans son pouvoir explicatif dû au type de données disponibles. En effet, puisque les données sont statiques et que les questions posées ne s'intéressent qu'à l'état actuel des répondant (revenu de l'année en cours, lieu de résidence actuel et ainsi de suite), les facteurs déterminants leur décision d'étudier ne sont pas nécessairement observés selon leur âge. En effet, utiliser le salaire annuel, ou le lieu de résidence, d'un individu ayant complété ses études depuis longtemps, par exemple, pose problème intuitif majeur. En effet, selon l'individu qu'on observe, il se peut que les données recueillient aient affecté son choix ou qu'elles soient au contraire des conséquences de son choix.

Au niveau intuitif, nous aurions pu postuler que, pour tous les gens âgés n'ayant pas réalisés d'études supérieures, la période de décision reste en vigueur indéfiniment et tenter de traiter cette décision de la sorte. Cependant, il s'agit d'un raisonnement problématique qui implique une grande difficulté de traitement aussi bien du point de vue théorique que du point de vue pratique. Nous avons par conséquent choisit un traitement statique et binaire de la question. Pour que la démarche soit conséquente avec ce raisonnement, nous avons restreint nos observations dans un premier temps aux gens âgées de 17 à 21 ans pour le Canada anglais (puisque les gens y vivant font face à la décision d'éducation à partir de 17 ans) et de 18 à 22 ans pour les résidents du Québec (car le choix s'y fait un ans plus tard). Ensuite, nous avons observer les canadiens de 18 et 24 ans afin de capter le plus de gens possible en arguant que pour bien des gens les études peuvent être relativement longues et que les variables d'intérêts restent semblable pour une certain période de temps. Pour cette raison, cette tranche d'âge est celle priviligiée dans notre étude. À l'instar de Finnie et Mueller (2008), un tel traitement nous a permis de suggérer des implications intéressantes. Avec ces choix de populations, nous concentrons nos résultats sur les gens qui sont dans leur période typique de décision de réalisation d'études supérieures.

Ces définitions des populations cibles nous ont permis d'obtenir divers résultats selon nos intérêt sans que ces derniers ne soient faussée par des variables n'ayant pas la même interprétation pour tout les répondants. D'autres considérations sur la sélection des variables dans notre modèle sont évoquées ci-dessous.

4.2 La variable dépendante

Le plus haut niveau d'étude atteint par le répondant

À travers ce projet, nous désirons identifier les facteurs déterminants dans la demande d'éducation supérieure. À proprement parler, nous n'observons pas la demande en elle-même. Ce qui est observable est plutôt la réalisation ou non d'un évènement donné à un moment précis. Par souci de rigueur, il est alors plus juste de dire que nous cherchons les déterminants de l'entreprise d'un diplôme universitaire.

Pour cette étude, nous définissons alors une variable dépendante binomiale. Cette dernière prendra la valeur de 1 si le répondant à débuter un programme universitaire et 0 dans le cas non-échéant.

4.3 Variables indépendantes

4.3.1 Variables familiales

Notre base de données nous donne accès à diverses variables familiales. Ces dernières sont considérées comme le point focal de cette recherche. En effet, comme nous tentons de trouver les facteurs déterminants de la réalisation d'études universitaire et que dans la littérature une grande attention est portée sur les caractéristiques de la famille dans laquelle à évoluée le répondant, ces variables sont suceptibles de révéler une grande quantité d'information.

Famille et Éducation parentale

En particulier, l'éducation des parents est une variable que nous observons avec un grand intérêt. Comme nous l'avons vu dans une section précédente, la littérature suggère que le fait que des parents

aient été ou non à l'université pourrait affecter le désir de l'individu d'y accéder et les moyens à sa porter pour le faire. Nous soupçonnons que ces deux éléments affectent le coût d'opportunité de l'individu observé. De plus, ces variables sont parmi celles qui sont le plus facilement utilisable. Nous avons le degré maximal atteint par chacun des deux parents et, dû à la différence d'âge entre un individu et ses parents, nous pouvons estimer qu'en général ces variables n'ont pas changé depuis le début de la majorité de l'individu (soit le moment à partir duquel il peut normalement prendre sa décision d'étude supérieur).

Nous avons créé de nouvelles variables révelant l'appartenance ou non à des sous-groupe à partir de ces variables de départ. Par exemple, nous avons créé des variables binomiales pour l'éducation de chaque parent et, ensuite, une variable prenant la valeur 1 si le plus haut niveau scolaire atteint entre les deux parents est un baccalauréat ou plus haut et 0 autrement. Une tentative avec le minimum plutôt que maximum a être tentée, mais cette avenue fût jugée moins défendable intuitivement.

Une autre variable à notre disposition est le type de famille dans lequel le répondant à grandit. Cette dernière a été transformée en binomial représentant 1 pour une famille à deux parents et 0 pour une famille monoparentale ou recomposée.

Type de ménage et revenu du ménage

Le revenu familial du répondant au moment de sa décision est présumément un facteur majeur dans la décision de poursuivre ou non ses études. Cependant, ces valeurs ne sont pas disponibles. Nous avons le revenu de l'individu au moment où il répond au questionnaire. Bien entendu, ces valeurs ne représentent pas réellement l'information que l'on désirerait, mais par contre en ciblant la population observée tel que nous l'avons décrit plus haut, nous pouvons nous en approcher.

Le type de ménage dans lequel l'individu vit en ce moment sera utilisé dans cette étude en plus du revenu familial au moment de l'enquête. Nous observons alors avec quel parents l'individu vit au moment du sondage et de la braquette de revenu leur ménage se situe. Dans un contexte où la population n'aurait pas été restreinte, nous aurions observé des individus ayant completer leur études depuis longtemps. Cela aurait amené des problèmes dans le sens où, à partir d'un certain âge, ils ont leur propre revenu, qui serait potentiellement une conséquence de leurs études, et ils sont les principaux pourvoyeurs du ménage. Notons que même dans un tel cas, ou même dans la situation extrême où les revenus au cours de la période de décision observée seraient disponible, les avoirs en capital immobilier et autres de la famille des répondants ne sont pas nécessairement pris en compte, car ils ne constituent pas des revenus de travail. Comme une grande part du coût d'opportunité pourrait être attribuable à ces derniers, nous n'aurions pas nécessairement l'heure juste à ce niveau même dans le cas où les données seraient présentent. C'est dans ce point que se trouve l'intérêt principal de notre sélection de la population observée. Cette dernière nous permet de s'approcher de la réalité et, ainsi, de tirer des interprétations pertinentes.

Notons avant de poursuivre que, pour la durée de temps requise avant l'obtention du diplôme recherché et le départ du la maison, le revenu du ménage ne sera pas conséquence des études. De plus, nous pouvons supposer que le revenu familial, bien que variable, reste suffisament stable en moyenne pour que, dans la tranche d'âge que nous observons, le revenu familial actuel consiste en un suffisament bonne approximation du revenu familial au moment de la décision de l'individu.

4.3.2 Variables géographiques

Dans les modèles cherchant à déterminer les facteurs qui influencent l'enrôlement dans des études supérieurs, des variables représentant le milieu de vie des répondants sont fréquemment utilisés. En effet, divers facteurs géographiques peuvent avoir une influence sur les décisions des individus observés. À travers notre base de données, nous avons accès à la province de résidence, à la région de résidence et au milieu de résidence du répondant. Nous utiliserons les informations reliées à la province et au milieu, mais nous omettrons la région. En effet, comme il y a un grand nombre de région avec un nombre relativement faible d'habitant, cette variable s'avère moins intéressante que les autres et quelque peu redondant lorsque l'on utilise les deux autres.

Urbain/Rural

Il est raisonnable de penser que le fait de vivre en milieu urbain, où l'accès aux principales universités est simple d'un point de vue de proximité, affecte non pas nécessairement le désir de poursuivre des études supérieures, mais plutôt le coût d'opportunité de compléter ces dites études. En opposition, on suspecte qu'un individu vivant en milieu rural encourrait des plus grands coûts de renonciations s'il désire poursuivre ses études.

Dans cette ligne de pensée, il serait idéal d'avoir accès à des données représentant la catégorie de l'endroit de résidence du répondant lors du moment de la décision. Cependant, certains problèmes se posent à ce niveau. Le choix d'entamer des études supérieures ne survient pas à un seul moment dans la vie de chaque individus et ces moments diffèrent pour chacun. Selon le cheminement que les individus prennent et selon leurs valeurs, un individu peut avoir l'occasion de décider de débuter ses études universitaires quelques années de suites avec un coût d'opportunité variant à chaque année ou tout simplement être en retard par rapport aux individus du même âge (notons que ce problème semble plus présent chez les jeunes québécois étant donné la flexibilité qu'accorde le CEGEP).

De plus, les données disponibles nous indiquent si un répondant vit en milieu rural ou urbain au moment même du sondage. Par conséquent, il faut être prudent dans la sélection de la population que l'on observe. Un individu de, disons, 45 ans, vivant en milieu rural n'a pas nécessairement pris sa décision d'étudier ou non en vivant dans ce milieu. Il est important de prendre en compte la migration des gens et du moment du sondage. Nous désirons tout de même utiliser cette variable, mais cela doit être fait avec parcimonie.

Étant donné ces considérations, il est important de remarquer les limitations au niveau des interprétations de ce modèle. Encore une fois, le choix de la population observée s'avère crucial.

Province de résidence et de naissance

La base de données nous offre aussi des observations quant à la province de résidence des répondants. Le lieu de résidence incorpore plusieurs facteurs pouvant influencer de désir de poursuivre ses études. Certains facteurs économiques, sociaux et systémiques peuvent potentiellement être la source d'une part de l'effet prédit capter par ces variables.

Premièrement, observons les facteurs sociaux. Nous pouvons postuler qu'il serait raisonnable de croire que le tissu social diffère selon les provinces. S'il y a une différence de valorisation dans les études supérieures qui est propre à certaines provinces et que cette dernière est significative dans la décision des individus, il est raisonnable d'imaginer qu'elle serait responsable d'une part des coefficients.

En second lieu, observons des facteurs économiques comme le dynamisme du marché de l'emploi et des perspectives dans cette province. Entre différentes provinces, il est raisonnable de considérer des disparités entre les perspectives d'emplois postérieurs à la complétion d'un diplôme. Cette hétérogénéité dans les avenues suivant les études impliquerait alors des différences dans les coûts d'opportunités de poursuivre ou non des études supérieures. Notons que même si un travailleur formé au pays verra ses compétences reconnus partout à travers le Canada, un déplacement vers une province plus dynamique impliquera tout de même en général un coût. Même si ce dernier varie selon les préférences des individus, nous estimons qu'il est raisonnable de croire qu'il sera tout de même présent en moyenne. Il est alors rationnel de croire que le coût d'opportunité de poursuivre ses études est plus grand pour quelqu'un vivant (ou projetant de vivre) dans une province plus moribonde que celui vivante dans une province dynamique. Comme notre analyse est centrée sur le niveau microéconomique, nous ne contrôlerons pas la régression pour les performances relatives. Conséquemment, ce facteur fera partie de l'effet de provenir d'une province donnée.

Finalement, des facteurs systémiques peuvent potentiellement être incorporés dans les coefficients de ces régresseurs. Les systèmes scolaires pré-universitaires canadiens sont gérés par les gouvernements provinciaux. Il est alors légitime de soupçonner une certaine variation dans la préparation des étudiants. Par ailleurs, des différences dans les systèmes entre les provinces anglophones et dans le système québécois pourraient avoir un effet sur le désir de poursuivre des études. Il serait intéressant de comparer les performances du système québécois relativement au reste du Canada. À ce point-ci, il

est crucial de remarquer qu'en incluant des variables pour les provinces tout en incluant une variable pour la langue des répondants nous courrons le risque d'une grande relation entre la langue « français » et la province « Québec ». Ce point a été considérer dans les diverses régressions que nous avons effectuer.

Le fait que ces divers facteurs prennent part dans l'effet de ces régresseurs ne pose pas nécessairement un problème. Cependant, lorsque l'on interprète l'effet de la province sur la complétion d'un diplôme universitaire, il est crucial de reconnaitre que ce sont en grande partie ces effets que l'on observe. Cela est cohérent, car ce sont des facteurs propres aux provinces.

Encore une fois, il aurait été idéal d'observer la province de résidence lors de la décision du répondant. Les problèmes sont néanmoins analogues à ceux que nous avons cités dans le cas de la variable du milieu. Un autre problème qui se présente potentiellement est le faible nombre de répondant dans les provinces de l'extrême Est du Canada. Comme nous cherchons à observer la décision de compléter ou non un diplôme universitaire et que, typiquement, un faible pourcentage de la population en mène un à terme, nous avons considéré l'éventualité d'un problème à ce niveau pour les provinces moins populeuses. Une piste de solution à cet égard que nous avons choisi est d'amalgamer les provinces à l'Est du Québec comme une seule entité : « Les provinces maritimes ». En procédant de la sorte, nous postulons implicitement que les facteurs caractéristiques propre aux provinces sont similaires au sein de ce groupe.

Ces facteurs ne sont que postulés pour l'instant. Lors de la phase des résultats, nous pourrons supporter ou miner la crédibilité des affirmations précédentes. Bien qu'intéressant, ces variables occupent tout de même l'arrière scène de notre étude et ne seront qui succinctement évoquées dans la section des résultats.

4.3.3 Variables démographiques

Au sein de notre base de donnés, nous avons aussi accès à plusieurs variables démographiques. Parmi ces dernières, nous avons particulièrement utilisé l'ethnicité du répondant (à savoir si ce dernier fait partie d'une minorité visible ou non), sa langue parlée et son sexe. Dans la revue de littérature, nous avons pu remarquer que ce type de variable est fréquemment jugée comme importante dans les diverses recherches faites sur l'enrôlement dans l'éducation supérieure. Il est raisonnable d'imaginer qu'il existe des valeurs prédominantes dans certaines cultures qui le sont moins dans d'autre. Additionnellement, certaines cultures peuvent faire face à certains facteurs non-observable qui affecterait leur probabilité d'entamer des études universitaire. Dans cette ligne de pensée, il est intéressant d'observer l'impact de l'inclusion à ce groupe sur la probabilité prédite de réalisation d'étude supérieures.

De même, la langue et le sexe d'un individu sont d'autres facteurs propices à ce type de discrimination au sens large. Ce sont donc des groupes entre lesquels il est intéressant d'effectuer la décomposition de B-O à la recherche d'indices appuyant ou minant l'hypothèse de différence de probabilité moyenne non-expliquée.

5 Modèle économétique

5.1 Régression Logit binomiale (WOOLDRIDGE)

5.1.1 Cadre théorique

Tel qu'exposé plus haut, notre variable d'interêt dans cette étude est la décision observée de poursuite ou non des études universitaire. En d'autres termes, nous cherchons à trouver les variables qui affectent principalement la probabilité d'un individu donné de s'engager dans de telles études ou non. Par conséquent, un modèle économétrique particulièrement propice à l'estimation serait un modèle Logit binomiale. Avant de poursuivre, nous traçons formellement les lignes directrices derrière un tel

modèle [4].

Soit le modèle suivant où nous notons la réalisation d'avoir été ou non à l'université chez l'individus i par Y_i où $i = 1, 2, ...n$ dans un échantillon de n individus. Avant de continuer, définissons ces objets. Nous avons dans un premier lieu la matrices :$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]^T$, dans laquelle on retrouve les vecteurs-colonnes $\mathbf{x_j} = [x_{,j,1},\ x_{j,2}, ...,\ x_{j,K}]^T$, contenant la valeur d'un régresseur donné pour chaque individu en guise de colonne et les observations de chaque régresseurs pour un individu donné en guise de ligne. Le vecteur des paramètres du modèle de régression sera exprimée par $\beta = [\beta_1, \beta_2, ..., \beta_K]^T$, le vecteur de la variable dépendante non-observable sera $Y = [Y_1,\ Y_2, ...\ Y_n]^T$ et le vecteur des observations de la variable dépendante dans la modèle latent sera notée $Y^* = [Y_1^*,\ Y_2^*, ...\ Y_n^*]^T$.

En nous basant sur l'intuition économique et le concensus prévalent dans la littérature à ce sujet, nous avons choisit comme régresseurs l'éducation parentale, le revenu du ménage, le statut du ménage du répondant ainsi que d'autres contrôles. L'éducation parentale sera traiter de façon à utiliser le plus haut diplôme haut diplôme obtenue entre les 2 parents du répondant. Le revenu du ménage sera décomposer en 3 palliers, soit de 0$ à 50 000$, de 50 000$ à 65 000$ et de 65 000$ allant jusqu'au plus grand. Nous avons construit des variables binaires pour les provinces de résidence et de naissance, pour le milieu urbain ou rural et pour des variables démographiques telles l'appartenance ou non à une minorité visible ou la langue utilité dans le ménage du répondant. Ce sont de ces variables et d'autres moins centrales, observées pour chaques membres de notre échantillon, que la matrice $\mathbf{X}$ sera constituée.

Dans le cadre d'un modèle Logit binomial, nous définissons la probabilité que notre varaible dépendante prenne la valeur 1 par $P(Y_i = 1|\mathbf{x_i}) = \Lambda(\mathbf{x_i^T}\beta)$ où la fonction $\Lambda :\ \mathbb{R} \to [0,\ 1]$, connue sous le nom de fonction Logit, est définie par :

$$\Lambda(x)\ := \frac{e^x}{1 + e^x}$$

Pour cela, nous définissons donc le modèle latent $Y^* = \mathbf{X}\beta + \varepsilon$, où ε est le vecteur $(N \,\mathrm{x}\, 1)$ où ε suit une distribution Logit de moyenne nulle et de variance unitaire, avec $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{E}[Y^*|\mathbf{X}] = \mathbf{X}\beta$. De façon équivalente, pour un individu donné nous écrirons $Y_i^* = \mathbf{x_i^T}\beta + \varepsilon_i$. Notre vrai variable dépendante est alors décrite de la sorte par cette fonction caractéristique : $Y_i = \mathbb{I}_{\{Y_i^* > 0\}}$. Étant donné ce modèle latent décrivant la valeur de Y par la suite, nous voyons que la probabilité conditionnelle que Y_i prenne la valeur de 1 est :

$$\begin{aligned} P(Y_i = 1|\mathbf{x_i}) &= P(Y^* > 0|\mathbf{x_i}) \\ &= P(\varepsilon > -\mathbf{x_i^T}\beta|\mathbf{x_i}) \\ &= 1 - \Lambda(-\mathbf{x_i^T}\beta)\ = \Lambda(\mathbf{x_i^T}\beta) \end{aligned}$$

où la dernière égalité vient du fait que $\Lambda(-x) = 1 - \Lambda(x)$. Comme la réalisation de la variable dépendante est 0 ou 1, nous constatons que $\mathbb{E}[Y_i\ |\mathbf{x_i}] = P(Y_i = 1|\mathbf{x_i}) = \Lambda(\mathbf{x_i^T}\beta)$. L'espérance conditionnelle de Y est donc estimée par la méthode du maximum de vraisemblance en trouvant le vecteur $\hat{\beta}$ tel que :

$$\hat{\beta} = \underset{\beta\ \epsilon\mathbb{R}^K}{argmax} \prod_{i=1}^{N} \Lambda(\mathbf{X}\beta)$$

ou de façon équivalente :

$$\hat{\beta} = \underset{\beta\ \epsilon\mathbb{R}^K}{argmax} \sum_{i=1}^{N} ln\left\{\Lambda(\mathbf{X}\beta)\right\}$$

Les deux égalités sont équivalentes par la monotonicité de la fonction logarithmique. À partir de cet estimateur $\hat{\beta}$, nous pouvons alors trouver la valeur, ou plutôt la probabilité dans le cas qui nous concerne, prédite $\hat{Y}_i = \Lambda(\mathbf{x_i^T}\hat{\beta})$ pour un individu i arbitraire et un vecteur $\mathbf{x_i}$ donnés. C'est ce problème

4. La section qui suit se base sur le traitement proposé par Wooldrigde (2002) et Greene (2011) sur le sujet

explicite qui sera implicitement résolu par le logiciel que nous utiliserons lorsqu'il sera temps d'effectuer l'estimation.

Notons que l'effet partiel d'un changement de z à $z + \Delta$, avec $z, \Delta \epsilon \mathbb{R}$, de la variable x_j, où $j \epsilon \{1, 2, 3, ..., K\}$, dans la probabilité prédite sera :

$$\Lambda(\beta_1 x_1 + ... + \beta_j(z + \Delta) + ... + \beta_K x_K) - \Lambda(\beta_1 x_1 + ... + \beta_j z + ... + \beta_K x_K)$$

Bien que cela soit trivial, remarquons que les coefficients estimés ainsi ne sont pas directement interprétables comme des effets marginaux contrairement aux coefficients dans d'autres modèles comme les MCO.

Dans le cadre dans lequel nous travaillons, nos variables d'interêts sont principalement celles décrivant les caractéristique du ménage dans lequel le répondant aura évolué. Comme nous l'avons évoqué plus haut, le point focal de l'étude sera l'effet de l'éducation parentale et du revenu parental dans la poursuite des études de l'enfant. Le sexe est aussi une partie intéressante de l'analyse.

5.1.2 Implémentation sur STATA

Décrivons les grandes lignes de l'application du cadre décrit ci-dessus avec nos données sur le logiciel STATA. Avant de débuter, notons que dans les appendices de ce texte seront jointes les codes associés au diverses opérations informatiques associées à notre étude. Après la manipulation des données brutes pour créer les variables que nous avons utilisées dans notre modèle $\mathbf{x}_1$, $\mathbf{x}_2$,...,$\mathbf{x}_K$, l'opération de maximisation est effectué à l'aide de ce logiciel.

Cette étape est relativement rapide. Après avoir généré les variables dont nous avons besoin à partir de la base de donnée, il nous suffit d'utiliser la commande *logit* où nous utiliserions normallement la commande *reg* dans une régression de moindre carrés. De plus, si nous désirons que les coefficients obtenus soient directement interprétables comme des effets marginaux, nous utiliser ajouter la commande mfx à la suite de notre régression. De cette façon, d'autres transformations ne seront pas requises pour expliquer l'effet prédit de nos diverses variables. Un fait à noter est que, comme nous le verrons plus bas, la décomposition d'Oaxaca-Blinder utilisera les coefficients estimés de la régression sans que ces derniers n'aient été transformé pour représenter des effets marginaux.

L'implémentation s'avère assez directe pour obtenir l'estimation d'un modèle Logit. Nous utiliserons implicitement les méthodes explicitées dans cette sous-section lorsque nous effectuerons les diverses décompositions qui nous intéressent.

5.2 Décomposition Oaxaca-Binder non-linéaire

5.2.1 Cadre théorique

Maintenant, supposons que l'on désire appliquer la méthode de décomposition d'Oaxaca dans le modèle décrit ci-dessus. L'interêt de cette dernière est de permettre de comparer la valeur moyennes de la variable dépendante entre deux groupe; ainsi, de trouver quelle part de cette différence serait attribuable à des différences dans les régresseurs et quelle part serait inexpliquée par ces derniers et donc traitée comme une sorte de discrimination entre les groupes. Pour conclure cette section, nous décrivons la théorie derrière cette décomposition dans le cadre d'un modèle Logit binomial et de la façon de l'implémenter à l'aide du logiciel $STATA$.

Posons deux groupes mutuellement exclusifs et complètement exhaustifs; par exemple le groupe A des répondants dont au moins un parent aurait fréquenter un établissement universitaire et le groupe B des répondants dont aucun des deux parents n'ont été à l'université. D'autres séparations de groupes peuvent s'avérer intéressantes. Entre autre, nous pourrions effectuée des décompositions selon la richesse de la famille du répondant, son sexe et bien d'autres facteurs pour vérifier la présence de discrimination entre les groupes.

Soit le groupe $g \epsilon \{A, B\}$, nous allons définnir nos valeurs d'intérêts et nos analogues empirique. Supposons que nous ayons N_g observations dans le groupe g. Après avoir séparé notre échantillon selon les observations qui font parties du groupe A et celles qui font parties du groupe B, nous effectuons

alors les régressions décrites plus haut pour obtenirs les vecteurs de paramètres estimés $\hat{\beta}_g$, $g \epsilon \{A, B\}$, que nous utiliserons dans notre décomposition (Notons la différence entre les vecteurs d'estimés $\hat{\beta}, \hat{\beta}_A$ et $\hat{\beta}_B$).

Comme nous sommes dans le cadre d'un modèle Logit, l'espérance conditionnelle pour un groupe g arbitraire sera :

$$\begin{aligned} E_{\beta_g}\left(Y_{ig}|\mathbf{X}_{ig}\right) &= \{\Lambda(\theta - \mathbf{X}_{ig}\beta_g) - \Lambda(-\mathbf{X}_{ig}\beta_g)\} + \{1 - \Lambda(\theta - \mathbf{X}_{ig}\beta_g)\} \\ &= \Lambda(\mathbf{X}_{ig}\beta_g) \end{aligned}$$

où nous basons nos développement sur la méthode de Fairlie (2003). Étant donné nos données, nos allons nous retrouver à estimer, pour chaque groupes, l'analogue empirique de l'espérance ci-dessus. Cet estimateur prendra à forme suivante :

$$\begin{aligned} S\left(\hat{\beta}_g, \mathbf{X}_{ig}\right) &= \frac{1}{N_g}\sum_{i=1}^{N_g}\left\{\Lambda(\hat{\theta} - \mathbf{X}_{ig}\hat{\beta}_g) - \Lambda(-\mathbf{X}_{ig}\hat{\beta}_g)\right\} + \left\{1 - \Lambda(\hat{\theta} - \mathbf{X}_{ig}\hat{\beta}_g)\right\} \\ &= \frac{1}{N_g}\sum_{i=1}^{N_g}\left\{\Lambda(\mathbf{X}_{ig}\hat{\beta}_g)\right\} \end{aligned}$$

Nous estimerons alors la différence dans les moyennes de la variable d'intérêt entre les groupes. Ce terme sera définit par $\hat{\Delta}_{g,h} = \overline{Y}_g - \overline{Y}_h$, où g, $h \epsilon \{A, B\}$et $g \neq h$. Par conséquent, la différence estimée sera la différence entre les termes propres à chaques groupes de la formes ci-dessus. Par exemple, la différence estimée entre le groupe A et B sera donnée par :

$$\begin{aligned} \hat{\Delta}_{A,B} &= S\left(\hat{\beta}_A, \mathbf{X}_{i,A}\right) - S\left(\hat{\beta}_B, \mathbf{X}_{i,B}\right) \\ &= \frac{1}{N_A}\sum_{i=1}^{N_A}\left\{\Lambda(\mathbf{X}_{i,A}\hat{\beta}_A)\right\} - \frac{1}{N_B}\sum_{i=1}^{N_B}\left\{\Lambda(\mathbf{X}_{i,B}\hat{\beta}_B)\right\} \end{aligned}$$

que l'on décompose en additionnant et soustrayant la quantité $S\left(\hat{\beta}_A, \mathbf{X}_{i,B}\right)$ du côté droit de l'équation précédente :

$$\begin{aligned} \hat{\Delta}_{A,B} &= \left\{S\left(\hat{\beta}_A, \mathbf{X}_{i,A}\right) - S\left(\hat{\beta}_A, \mathbf{X}_{i,B}\right)\right\} + \left\{S\left(\hat{\beta}_A, \mathbf{X}_{i,B}\right) - S\left(\hat{\beta}_B, \mathbf{X}_{i,B}\right)\right\} \\ &= \left\{\frac{1}{N_A}\sum_{i=1}^{N_A}\left\{\Lambda(\mathbf{X}_{i,A}\hat{\beta}_A)\right\} - \frac{1}{N_B}\sum_{i=1}^{N_B}\left\{\Lambda(\mathbf{X}_{i,B}\hat{\beta}_A)\right\}\right\} \\ &+ \left\{\frac{1}{N_B}\sum_{i=1}^{N_B}\left\{\Lambda(\mathbf{X}_{i,B}\hat{\beta}_A)\right\} - \frac{1}{N_B}\sum_{i=1}^{N_B}\left\{\Lambda(\mathbf{X}_{i,B}\hat{\beta}_B)\right\}\right\} \end{aligned}$$

Le terme $\hat{\Delta}_{A,B}$ nous donne alors l'estimation de la décomposition bipartite dans le contexte de notre modèle. En utilisant les mêmes analogues empiriques et $\hat{\beta}_*$définit comme étant une moyenne pondérée des vecteurs $\hat{\beta}_A$ et $\hat{\beta}_B$ que nous aurons obtenus préalablement, nous pouvons aussi définir l'estimateur de la décomposition tripartite comme :

$$\begin{aligned} \hat{\Delta}^*_{A,B} &= \left\{S\left(\hat{\beta}_*, \mathbf{X}_{i,A}\right) - S\left(\hat{\beta}_*, \mathbf{X}_{i,B}\right)\right\} \\ &+ \left\{S\left(\hat{\beta}_A, \mathbf{X}_{i,A}\right) - S\left(\hat{\beta}_*, \mathbf{X}_{i,A}\right)\right\} \\ &+ \left\{S\left(\hat{\beta}_*, \mathbf{X}_{i,B}\right) - S\left(\hat{\beta}_B, \mathbf{X}_{i,B}\right)\right\} \end{aligned}$$

où nous définissons $S(\hat{\beta}_*, \mathbf{X}_{i,g}) \equiv \frac{1}{N_g} \sum_{i=1}^{N_g} \left\{ \Lambda(\mathbf{X}_{i,g}\hat{\beta}_*) \right\}$ pour $g \epsilon \{A\, B\}$. Nous remarquons qu'à l'instar de la décomposition bipartite, cette forme de décomposition est obtenue en additionnant et soustrayant des mêmes éléments, $S(\hat{\beta}_*, \mathbf{X}_{i,A})$ et $S(\hat{\beta}_*, \mathbf{X}_{i,B})$, du côté droit de l'égalité ci-dessus.

5.2.2 Implémentation STATA

Tel que nous en avons discuté précédement, cette section s'appuiera particulièrment sur les textes de Jann (2008) et de Sinning, Hahn et Bauer (2008). À l'aide des outils développés dans le second articles, l'implémentation de la décomposition de Blinder-Oaxaca dans un cadre non-linéaire s'avère très directe. En effet, le travail de programmation a déjà été accompli par les auteurs et il ne reste qu'à utiliser les outils, se résumant en une série de commandes, qu'ils nous fournissent de façon adéquate. Les principales commandes que nous utilisons seront *nldecompose*, *by*(.), *threefold* et *bootstrap*. Exposons leur utilisation par un exemple :

Supposons que nous désirons décomposer notre échantillon entre les hommes et les femmes, c'est-à-dire une séparation en groupes complètement exhaustifs et mutuellement exclusifs. Nous définissons alors arbitrairement une vecteur nommé S qui prendra la valeur 1 dans son i^eélément si l'individu est un homme et 0 dans le cas contraire pour $i \epsilon \{1, 2, 3..., n\}$. De plus, supposons que le modèle non-linéaire que nous désirons traiter est un modèle Logit. Finalement, supposons que nous avons déjà défini toutes les variables dépendante et indépendantes que nous désirons dans notre régression ; notons cette séquence de variable avec la variable dépendante en premier et les variables indépendantes par la suite par $\{variables\}$. Étant donné ces objets, nous écrivons la commande pour la décomposition de base par :

$$nldecompose,\, by(S) \,:\, logit\ \{variables\}$$

La commande *nldecompose* nous réfère à la décomposition de Blinder-Oaxaca non-linéaire et la commande *by*(.) spécifie les deux groupes entre lesquels nous désirons effectuer la décomposition. Dans notre exemple, mettre S dans la commande *by*(.) indique que nous désirons effectuer la décomposition entre les hommes et les femmes. Les termes suivant le deux-points représente à régression sur laquelle nous désirons appliquer la décomposition. Dans un contexte où nous désirons estimer l'écart-type des estimateur, cela sera effectué par bootstrap et il ne suffira que d'ajouter la commande *bootstrap* entre la commande *by*(.) et le deux-point de l'expression ci-dessus. Il est possible de personnaliser les caractéristiques de ce bootstrap grâce à une commande *bsoptions*(.) que l'on écrirait subséquemment. Si nous désirons une décomposition tripartite, il suffit d'ajouter la commande $threefold$ avant le deux-point. Tel que mentionné plus haut, aucun concensus n'est présent dans la littérature quant aux poids qui devraient être utilisés pour l'obtention de β^*. À cette fin, la commande *omega*(.) placée après la commande $threefold$ si présente nous permet de personnaliser les poids utilisés pour l'obtention de cet estimateur.

En concluant, il est important de mentionner qu'il ne faudrait pas inclure de variable binaire pour le sexe des individus dans la régression de cet exemple-ci puisque que cette dernière serait parfaitement correlée à la variable S que nous avons défini plus haut et que c'est par cette dernière que nous effectuons la décomposition.

6 Résultats

Dans cette section, nous exposons les divers résultats que nous avons obtenus dans cette étude. En premier lieu, nous abordons les régressions de base. Ensuite, nous discutons bièvement d'un analogue de la régression de base où nous avons remplacé les variables d'éducation parentale et de revenu familal par des variables d'interaction entre ces dernières. Le raisonnement derrière cette démarche est de prendre en compte le lien entre l'éducation des parents d'un répondant et le revenu de ces derniers et, ainsi, d'observer si l'effet de se trouver dans une tranche de revenu pour un niveau d'éducation parentale donné a un effet significatif sur la probabilité prédite d'entamer des études universitaire. De même, nous abordons tout aussi succinctement des régressions correspondantes à la régression de

base, mais selon des populations faisant partie de certaines tranches de revenu ou d'éducation parentales dans le but de comparer l'effet prédit d'une variable d'intérêt particulière étant donné que notre répondant fait partie d'un certain groupe. Par exemple, l'effet de l'éducation parental chez les mieux fortunés. Finalement, nous nous basons sur notre modèle de base pour, à partir de ce dernier, effectuer des décompositions de Blinder-Oaxaca pour, ensuite, observer quelle part de la différence de moyenne de la variable dépendante entre des individus faisant partie de deux groupes différents est expliquée par les différences dans les estimateurs des régressions et quelle part serait inexpliquée.

6.1 Modèle de base

6.1.1 Régression standard

Commencons par discuter des régressions de base sans effectuer de décomposition ou sans créer de termes d'intéraction. Ces régressions ont été effectuées sur diverses tranches de notre population cible avec l'idée sous-jacente que différents individus peuvent intégrer le milieu universitaire à différents moments dans leur vie selon divers facteurs idiosyncratiques ou systémiques propre à leur milieu.

Régression standard

Population:	18-24 ans CDN	18-24 ans Non-QC	17 -22 ans Non-QC & 18-22 ans QC
Variable	**Effet marginal**	**Effet marginal**	**Effet marginal**
Parent avec B.Sc. ou plus	0,3228145	0,3172591	0,3269279
Parent avec PSE non-univ.	0,1773064	0,1230869	0,2622984
Minorité	0,2299397	0,1817166	0,126203
Femme	0,1332734	0,1567223	0,0767859
Vivre sans ses parents	-0,1664675	-0,2258198	-
Ménage 65 000$ et +	-	-	-
Vivre en milieu urbain	-	0,1049216	-
& autres contrôles	-	-	-
Pseudo-R^2	0,1347	0,1506	0,0765
# observations	927	740	773
Significatif à 5%	Significatif à 10%		Non-significatif

FIGURE 1 – Régression standard selon divers groupes de la population

La **figure 1** ci-dessus expose nos résultats pour notre modèle phare dans les populations des 18 à 24 ans résidents du Canada, des 18 à 24 ans résidents du Canada hors Québec et des 17 à 22 ans résidents du Canada hors Québec avec les 18 à 22 ans résidents du Québec. Les études de Finnie et Mueller (2008) et de Drolet (2005) appuient notre choix de tranches d'âges observées. Ces groupes d'âge sont configurés dans le but d'inclure le plus d'observations possibles faisant partie de notre population d'intérêt dont l'interprétation des coefficients restera cohérente. L'individu de comparaison dans ces régressions est un homme, résident en Ontario, ne faisant pas partie d'une minorité visible, vivant avec ses deux parents dans un ménage dont les revenus sont de 65 000$ ou moins par années et dont aucun de ces derniers n'a de diplôme post-secondaire. Plusieurs régresseurs de contrôle sont exclus du tableau précédent pour des fins de succincteté. Selon le code de couleur, nous pouvoir voir quels coefficients estimés sont statistiquement différents de 0 et quels ne le sont pas. Notons que les coefficients reportés sont les effets marginaux et non les estimateurs initiaux.

Les résultats de la figure 1 abondent dans la même direction que l'intuition économique et les observations de la littérature. En effet, nous remarquons que plus les parents ont atteint un haut niveau d'éducation, plus notre modèle prédit une grande probabilité pour leur progéniture de s'adonner

à des études supérieures. À l'échelle du Canada chez les 18 à 24 ans, le modèle suggère qu'un individu moyen dont les parents aurait suivit des études post-secondaire non-universitaire[5] aurait 17% plus de chance de se rendre à l'université qu'un individu dont aucun parent n'aurait poursuivit d'études haut-delà du secondaire. Cette probilité prédite passe à 32% lorsque les parents du premier individus ont suivit des études universitaires. Tel que nous l'avons exposé plus haut, ce résultat est cohérent avec le raisonnement selon lequel le support et les valeurs propices au désir d'étudier au niveau universitaire sont une fonction croissante de l'éducation parentale. Ces résultats sont consistents en magnitude lorsque nous comparons les trois groupes observés dans la figure 1.

Dans la même ligne d'idée, nous observons que de faire parti d'une minorité visible et d'être du sexe féminin a un effet positif sur la probabilité prédite de se livrer a des études supérieures. Au niveau des Canadiens de 18 à 24 ans, notre modèle prédit que d'être une femme ou de faire partie d'une minorité visible augmente la probabilité de 13% et 22%, respectivement, comparativement à ceux qui ne le sont pas. Des résultats similiaires sont, entres autres, documentés dans Finnie et Mueller (2008) et Drolet (2005) et viennent appuyer ceux de ce modèle. Par ailleurs, il est fréquemment documenté que les jeunes femmes performent mieux en moyenne au niveau secondaire que les jeunes hommes[6] ; il est alors raisonnable que cela ait l'incidence que l'on observe dans ce modèle. De plus, étant donné le filtrage de l'immigration au Canada privilégiant les individus étant jugé comme ayant un certain potentiel et leur famille, il est raisonnable d'imaginer que ces derniers ou leur descendants, lors que comparer au canadien moyen, ait plus de chance de réaliser des études universitaire. Notons de même que la proximité des centres urbain a un effet prédit positif lorsque comparer au fait de vivre en milieu rural et le fait de vivre seul, comparativement à vivre en compagnie de ses parents, à un effet prédit négatif. Ces estimations sont consistente avec un raisonnement se basant sur le coût d'opportunité, dans un premier temps, et sur la valeur de l'encadrement parental dans un second temps.

En concluant, notons que le fait d'appartenir aux gens les plus fortunés n'a pas, à prime abord, un effet significatif sur la probabilité que nous cherchons à prédire. Il est toutefois nécessaire de rester prudent à cet égard, étant donné le lien entre les moyens financiers d'un ménage et l'éducation des chefs de familles de ce dernier. Nous abordons dans les sous-sections suivantes des régressions tentant de venir appuyer notre observation sur l'impact des moyens financiers ou, tout au moins, de venir amener une piste d'explication quant à ce constant.

6.1.2 Régression avec termes d'interactions

En remplacant les variables associées aux réalisation de revenu du ménage et d'éducation parentales, nous sommes en mesure d'isoler la provenance des effets dans la magnitude des coefficients. Par soucit de pouvoir prédictif, nous avons décidé d'appliquer une séparation dans les revenus en deux classe, soit les ménages de plus et de moins de 65 000$/ans , et une séparation dans l'éducation parentale en trois classes, soit les gens dont le niveau maximal est équivalent à un $high - school$ ou moins, ceux dont le niveau maximal est des études PSNU et ceux dont le niveau maximal consiste en des études universitaires. Cela nous donne alors 6 variables d'interactions. Les régressions, ayant les mêmes variables que celles de la section précédente aux variables d'éducation et de revenu près, sont exposées dans la figure 2 ci-dessous. Ces dernières ont comme groupe de base le même individu que nous vous avons présenté tout à l'heure à l'exception du fait que ses parents ont des études universitaires et gagnent 65 000$/ans ou plus.

5. Que nous notons dorénavant PSNU
6. Par exemple, dans l'étude de Stoet et Geary (2013)

Régressions avec termes d'interactions - Type I		
Population:	18-24 ans CDN	18-24 ans Non-QC
Variable	**Effet marginal**	**Effet marginal**
B.Sc ou + & - de 65K$	-	-
PSE non-univ. & 65K ou +	-	-
PSE non-univ. & - de 65K	-0,1597957	-0,1947028
HS ou - & 65K ou +	-0,248373	-0,2304159
HS ou - & - de 65K	-0,3409852	-0,3401878
Minorité	0,2536824	0,1948195
Femme	0,1336753	0,1552057
Vivre sans ses parents	-0,1521447	-0,2137291
Vivre en milieu urbain	-	0,119548
& autres contrôles	-	-
Pseudo-R^2	0,1179	0,135
# observations	927	740
Significatif à 5%	Significatif à 10%	Non-significatif

FIGURE 2 – Régressions avec termes d'interactions entre 2 classes de revenu et l'éducation parentale

Nous remarquons que dans ces régressions les magnitudes des effets sont encore un fois semblables entre les deux populations ciblées. De plus, nous notons que les coefficients prédits pour les variables d'interactions sont négatifs, tout comme ou devrait s'y attendre puisque le groupe de base se trouve à être les individus les mieux nantis et dont les parents sont les plus éduqués. Moins un individu est nanti en terme d'éducation parentale et de revenu familial, moins les chances que l'on prédit à ce dernier de se rendre à l'université sont grande comparativement à une personne présentant la combinaison décrite plus haut. Notons aussi que l'effet d'avoir des parents présentant des études universitaire, mais moins bien nantis, ou des parents présentant des études PSNU, mais faisant partie des mieux dotés, n'est pas significatif lorsque comparé à notre groupe de base. Un tel constat peut suggèrer une sorte de compensation dans les effets entre le revenu et l'éducation parentale. Sans trop nous avancer, nous pouvons tout de même inférer que l'effet d'avoir des parents qui se sont rendu à l'université est très positif dans la probabilité de réalisation dépendante à un tel point qu'un revenu moins grand, pour ce même niveau d'éducation, n'amène pas de différence significative. Comparer à ces derniers, les individus dont les parents présentent des PSNU partent relativement désavantager, mais le fait de faire parti des mieux dotés est suffisant pour rendre la différence non-significative. Au contraire, le fait d'avoir des parents ayant au maximum des études secondaire entraîne un effet si négatif dans la probabilité de se rendre soit même à l'université que même si ces derniers sont bien nantis nous observons quand même un effet négatif économiquement et statistiquement significatif dans la proabilité prédite. Par ailleurs, divers facteurs peuvent être en causes à cet endroit ; diverses interprétations sont alors potentiellement valides pour commenter les forces en présences. En concluant, l'intuition économique appuit dans une certaine mesure les résultats obtenus dans la figure 2.

6.1.3 Régression par strates

Avant de pousuirvre aux décompositions par groupes, nous présentons des régressions par strates d'éducation et par strates de revenu de ménage. Les régressions restent structurées de la même façon que la régression standard présentée plus haut à l'exception du fait que les variables par lesquels nous trions la population dans chaque cas de figure sont ôtées de la régression pour des raisons évidentes.

Strates de revenu de ménage

Notre première régression par strate est divisée en fonction du revenu des ménages ; les résultats associés à cette dernières sont présentés dans la figure 3 ci-dessous. Dans le groupe de base, nous retrouvons à présent les individus dont les parents ne présentent pas d'éducation post-secondaire. Nous remarquons dans un premier temps que les effets sont similaires entre les régressions faites sur les canadiens de 18 à 24 ans en général et leur analogue faite seulement sur les québécois.

Régressions par strates : Revenu

Population:	18-24 ans CDN		18-24 ans Non-QC	
Strates:	- de 65K	65K & +	- de 65K	65K & +
Variable	**Effet margin.**	**Effet margin.**	**Effet margin.**	**Effet margin.**
B.Sc ou +	0,2798763	0,2727128	0,2338882	0,2119691
PSE non-univ.	0,1996127	0,1717428	-	-
Minorité	0,3220051	0,2154791	0,2884521	0,163029
Femme	-	0,1791908	-	0,256068
Vit sans parents	-0,1402711	-0,206374	-0,209546	-0,334316
Vit milieu urbain	0,2388627	-	0,2954988	-
& autres contrôles	-	-	-	-
Pseudo-R^2	0,1638	0,1449	0,1851	0,1631
# observations	397	389	298	325
Significatif à 5%	Significatif à 10%		Non-significatif	

FIGURE 3 – Régression par strates de revenu du ménage

Parmi les principaux constats que nos estimateurs nous suggèrent, nous dénotons un effet moins grand associé au sexe féminin chez les ménages moins fortunés comparativement à celui qu'on retrouve chez les ménages mieux dotés. Dans les deux groupes, cet effet n'est pas significatif lorsque le ménage fait moins de 65 000$ et il est significativement positif chez les ménages plus fortunés. Entre d'autre termes, la différence de la probabilité prédite chez une femme et chez un homme est une fonction croissante du revenu du ménage. Bien entendu, ces résultats doivent-être prit avec réserve et il faut noter qu'on n'observe que deux palliers. Cependant, ces derniers sont conformes aux résultats typiques et peuvent être relativement bien rationalisés. Par exemple, si on suppose un interêt plus marqué dans les études chez les filles que chez les garçons, un revenu de ménage élevé pourrait fournir aux garçons comme aux filles des meilleurs moyens de se rendre à l'université et, à cause de leur interêt, les filles seraient plus porter à utiliser cet avantage quand ce dernier se présente. Notons que ces interprétations sont hypothétiques et portent sur l'individu moyen.

Avant de poursuivre, notons que la différence de coefficients associés à l'effet d'études universitaire chez les parents selon les deux palliers de revenu n'est que minime. Cela nous amène a suggérer que l'apport du revenu n'a pas d'effet significatif sur l'effet de l'éducation parental. Aussi intéressant à souligner est le fait que de vivre en milieu urbain semble plus important sur la probabilité prédite de se rendre à l'université lorsque l'on est issue d'un milieu moins fortuné que dans le cas contraire. C'est une autre estimation raisonnable dans le sens où le coût d'opportunité de faire des études supérieures lorsque l'on vit en milieu rural pourrait être partiellement éponger par les facilitations qui vont de paire avec une plus grande richesse.

Strates d'éducation

Nous avons procédé de la même façon que dans la sous-section précédente en séparant notre population selon le niveau d'éducation des parents des répondants ; les estimateurs sont exposés ci-dessous

dans la figure 4. Le groupe de base dans ces régressions incorpore à présent les individus dont le revenu familial est de moins de 65 000$/ans.

Régressions par strates : Éducation parentale				
Population:	18-24 ans CDN		18-24 ans Non-QC	
Strates:	non-Univ.	B.Sc. ou +	non-Univ.	B.Sc. ou +
Variable	**Effet margin.**	**Effet margin.**	**Effet margin.**	**Effet margin.**
65K & +	0,0946259	-	0,1310331	-
Minorité	0,3065702	0,1446058	0,2532314	0,1222417
Femme	0,1822313	-	0,2134446	-
Vit sans parents	-0,2583991	-	-0,2716079	-0,125989
Vit milieu urbain	0,1033504	-	0,1719244	-
& autres contrôles	-	-	-	-
Pseudo-R^2	0,1023	0,1183	0,1227	0,0903
# observations	562	343	444	278
Significatif à 5%	Significatif à 10%		Non-significatif	

FIGURE 4 – Régression par strates d'éducations parentales

En observant ce tableau, deux résultats nous sautent rapidement aux yeux. Premièrement, l'effet d'être fortuné, comparativement au fait de l'être moins, est significativement positif chez les gens issue d'un ménage ou aucun parent n'a fréquenté l'université, mais n'est pas significatif chez ceux dont les parents l'ont fréquenté. Cela suggère que l'effet de la richesse sur la probabilité de s'adonner à des études unversitaires est une fonction décroissante de l'éducation parentale du ménage en question. Ce constat est cohérent avec les résultats observés plus haut. En effet, on pourrait inférer qu'en présence de modèles ayant vécus l'expérience universitaire et qui la valorisent présumément les considérations financières actuelles auraient moins d'emprise sur la décision de l'individu. Deuxièmement, nous observons que l'effet d'être une femme plutôt qu'un homme sur la probabilité prédite est significativement positif dans la strate inférieure d'éducation parentale, mais que ce dernier n'est pas significatif dans la strate supérieure. Cette observation est cohérente avec la théorie et pourrait être rationnalisée de plusieurs façons différentes. À titre d'illustration, il est possible que, par un interêt inné ou aquis par la façon dont elles sont élevées, les jeunes femmes aient moins besoin de la présence d'une personne ayant vécu des études universitaires dans son entourage pour s'y adonner. Dans des études subséquentes alliant économie et psychologie, il pourrait être intéressant de développer cette question et amener d'autres pistes d'explication.

6.2 Décompositions de Blinder-Oaxaca de notre modèle

À partir du modèle décrit précédement, nous avons effectué des décompositions entre certains groupes mutuellement exhaustifs et complètement exclusifs d'interêt. La sous-section précédente était instrumentale dans notre désir de comparer les différences entre les différents groupes ; nous nous y référerons fréquemment dans l'exposé suivant. Nous commencerons par une décomposition de B-O selon l'éducation parentale, ensuite, nous procédons à la décomposition par revenu et nous concluons par la décomposition par sexe. Les décompositions ont été faites sur les sous-ensembles des gens résidents au Canada de 18 à 24 ans. En guise de comparaison leurs analogues sur des gens résidents au Canada hors Québec de 18 à 24 ans et des gens résidents au Canada hors Québec de 17 à 22 ans et résidents au Québec de 18 à 22 ans sont inclus en annexes, mais hors du texte principal dû à leur similarités.

6.2.1 Décomposition selon l'éducation parentales

Afin de voir la part de la différence dans la moyenne chez la probabilité prédite d'entamer des cours universitaires, nous avons effectué une décomposition de B-O entre les gens dont les parents ont été à l'université et ceux dont ce n'est pas le cas. Les entrées de la figure 5 ci-dessous sur la ligne nommée «Coefficient» représente la part de la différence totale attribuable aux différences dans les coefficients des régressions tandis que les entrées se trouvant sur la ligne nommé «Caractéristiques» représentent la part qui est attribuable aux différence dans les caractéristiques intrinsèques des membres du groupe en question, pour un poids donnés. [7] Les entrées de la deuxième ligne incorporent potentiellement des effets provenant de discrimination ou de variables non-observées, mais représentent principalement l'estimation de la différence provenant de l'écart dans les valeurs des régresseurs entre les groupes. Par ailleurs, nous avons seulement mit deux poids différence, en l'occurence 0 et 1, puisque qu'il y a peu d'antécédent dans l'application de la décomposition de B-O dans le contexte présent et qu'il est raisonnable d'imaginer que la discrimination que nous cherchons à observé serait concentrée vers un groupe plutôt qu'un autre. Un telle présentation nous permet d'observer d'un côté comme de l'autre.

Décomposition - Régression Standard		
Éducation parentale		
Population	18-24 ans au Canada	
Groupe de référence	Universitité chez un parent	
Poids	1	0
Coefficients	0,2913119	0,2109844
Caractéristiques	0,0087789	0,0891065
Significatif à 5%	Significatif à 10%	Non-significatif
Groupe	**Univ.**	**Non-Univ.**
# Individus	343	584

FIGURE 5 – Décomposition de B-O selon l'éducation parentale

Rappelons que pour un $poids = 1$ nous supposons implicitement $\beta^* = \beta_{groupe\ de\ référence}$ et que pour $poids = 0$ l'estimateur β^* sera l'estimateur du coefficient dans la régression sur le groupe complément au groupe de référence. En d'autres termes, si nous fonctionnons avec des groupes arbitraires $\{A\,B\}$, un $poids = 1$ accordé au groupe A et la même décomposition que celle exposée dans les sections précédentes, les estimateurs que nous obtenons sont alors ce premier terme :

$$«caratéristique» \quad \equiv \quad S(\hat{\beta}_A, X_{i,A}) - S(\hat{\beta}_A, X_{i,B})$$

c'est-à-dire l'estimation de la différence pour de mêmes coefficients avec des régresseurs différents, et ce deuxième terme :

$$«coefficient» \quad \equiv \quad S(\hat{\beta}_A, X_{i,B}) - S(\hat{\beta}_B, X_{i,B})$$

c'est-à-dire l'estimation de la différence pour de mêmes régresseurs avec des coefficients différents. Si on change le poids pour 0, le groupe de référence devient B et l'interprétation des estimateurs est analogue en interchangeant A et B dans les deux égalités ci-dessus.

Dans le cas qui nous intéresse, nous remarquons que les estimateurs sont significativement différents de zéro sauf celui de la part caractéristique lorsque le poids est de 1. Dans le cas où le poids associé au groupe de référence est $poids = 1$, 97.1% de la différence totale dans la probabilité moyenne prédite de réaliser des entre les gens dont les parents ont des études universitaire et les gens dont ce n'est pas

7. Sinning, Hahn et Bauer, 2008

le cas est attribuable au valeurs des régresseurs. La balance, soit seulement 2.9%, serait attribuale aux différences dans les caractéristiques entre les individus de ces deux groupes. La part de la différence que nous interpréterions comme une discrimination n'est pas significativement différente de zéro. Notons que poser un poids égal à 1 équivaut à poser :

$$\begin{aligned}\hat{\Delta}_{Univ.} &= \left\{S(\hat{\beta}_{Univ.}, X_{i,Univ.}) \text{ -}S(\hat{\beta}_{Univ.}, X_{i,Non-Univ.})\right\} \\ &+ \left\{S(\hat{\beta}_{Univ.}, X_{i,Non-Univ.}) - S(\hat{\beta}_{Non-Univ.}, X_{i,Non-Univ.})\right\}\end{aligned}$$

où les termes $Y_{i,g}$ et $X_{i,g}$ font référence au valeurs observées que présente un individus i membre du groupe $g \epsilon \{Univ., Non-Univ.\}$.

Alternativement, nous observons qu'en posant un poids égal à 0, nous estimons que 70.3% de la différence totale dans la probabilité moyenne prédite de réaliser des études universitaire des entre les gens dont les parents ont des études universitaire et les gens dont ce n'est pas le cas est attribuable au valeurs des régresseurs. Par conséquent, 29.7% de la différence est estimée comme attribuable à des différences de caractéristiques. Cette quantité est significativement positive. Avec un poids de 0 accodé au groupe de référence dans la définition de β^*, ces estimateurs sont équivalents aux terme à l'intérieur de la première et deuxième parenthèse du côté droit de l'expression suivantes :

$$\begin{aligned}\hat{\Delta}_{Non-Univ.} &= \left\{S(\hat{\beta}_{Non-Univ.}, X_{i,Univ.}) - S(\hat{\beta}_{Non-Univ.}, X_{i,Non-Univ.})\right\} \\ &+ \left\{S(\hat{\beta}_{Univ.}, X_{i,Univ.}) - S(\hat{\beta}_{Non-Univ.}, X_{i,Univ.})\right\}\end{aligned}$$

Elle suggère une certaine discrimination positive en faveur des gens dont les parents ont été à l'université. Dans le cas présent, nous observons des valeurs positives pour les estimateurs des paramètres de coefficient et de caractéristiques que le poids accordé au groupe de référence soit 0 ou 1. Cela suggère que les caractéristiques que l'on observe et les coefficients estimés des régressions sur les goupes séparés augmentent tous deux l'écart dans la probabilité de réaliser des études universitaires entre les membres de ces groupes.

Avant de poursuivre, faisons une petite parenthèse pour expliquer pourquoi nous avons les égalités précédentes. Par exemple, pour un *poids* $= 1$ et un groupe de référence que nous nommons «A», nous avons $\beta^* = \beta_A$ où β_A est le coefficient estimé en effectuant la régression seulement sur le groupe A. Rappelons que nous avions dit dans la revue de Sinning, Hahn et Bauer (2008) que :

$$\begin{aligned}\overline{Y}_A - \overline{Y}_B &= \{E_{\beta^*}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta^*}(Y_{iB}|\mathbf{X}_{iB})\} \\ &+ \{E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta^*}(Y_{iA}|\mathbf{X}_{iA})\} \\ &+ \{E_{\beta^*}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\}\end{aligned}$$

Or, si $\beta^* = \beta_A$, cette équation peut-être réécrite comme :

$$\begin{aligned}\overline{Y}_A - \overline{Y}_B &= \{E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB})\} \\ &+ \{E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA})\} \\ &+ \{E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\} \\ &= \{E_{\beta_A}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB})\} + \{E_{\beta_A}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB}|\mathbf{X}_{iB})\}\end{aligned}$$

En définnissant A et B comme les groupes qui nous intéresse, nous obtenons automatiquement, après avoir pris leur analogue empirique, les résultats ci-dessus.

6.2.2 Décomposition selon le revenu

Nous avons effectué une décomposition selon la richesse des ménages chez les canadiens de 18 à 24 ans. Les deux groupes sont les gens issus de ménages engrageant 65 000$ ou plus par ans, que

nous notons les «>65K$» dans la figure 6 ci-dessous, et les gens issus de ménages faisant moins que cette somme que nous notons les «<65K$». Les estimateurs que nous obtenons de la sorte ne sont pas significativement différents de zéro, en se basant sur l'écart-type estimé par *bootstrap*, sauf celui de l'effet des caractéristiques lorsque le poids accordé au groupe de référence, soit les >65K% est de 0.

Décomposition - Régression Standard		
65K$ et + ou - de 65K$		
Population	18-24 ans au Canada	
Groupe de référence	65K et +	
Poids	1	0
Coefficients	0,0389988	0,030725
Caractéristiques	0,0270853	0,0353591
Significatif à 5%	Significatif à 10%	Non-significatif
Groupe	**>65K$**	**<65K$**
# Individus	389	538

FIGURE 6 – Décomposition de B-O selon le revenu du ménage

Nous remarquons que la part estimée de la différence totale lorsque $poids = 1$ est d'environ 41% attribuable aux caractéristiques et 59% accordable aux différences de coefficients. La différence totale est de 0.066 et est significativement différence de zéro pour un seuil de 5%. Cependant, comme les effets estimés sont relativement près numériquement, aucun de ces derniers n'est significatif.

Alternativement, lorsque le $poids = 0$ environ 53.5% de la différence totale dans la variable dépendante entre les individus de chaque groupe est attribuable aux différence de caractéristique tandis que 46.5% est accordé à la différence de coefficients. Seul le coefficient associé à ces premières est significativement différent de zéro.

Comme nous observons des estimateurs positif, le modèle suggère qu'ils affectent tous deux positivement l'écart (que le poids accordé au groupe de référence soit 0 ou 1), mais que seul le paramètre issu des différences dans les caractéristiques que l'on observe entre les groupes a un effet significativement différent de zéro. En d'autre terme, l'écart présent est alimenté à la fois par des différences de dotations et des différences de coefficients prédit, mais ce dernier n'est significatif statistiquement que lorsque le groupe de référence est les individus venant d'un ménage faisant moins de 65 000$/ans dans les caractéristiques leur étant propre.

Notons avant de poursuivre que d'augmenter le niveau de revenu auquel nous séparons les individus n'affecte pas en outre mesure l'ampleur des estimateurs. En effectuant la séparation à 85 000$/ans plutôt qu'à 65 000$/ans, l'effet des caractéristiques lorsque $poids = 0$ passe à 0.066 et est significativement différent de zéro à un seuil de 0.1%. Outre cela et une plus grande différence totale, le constat reste similaire.

6.2.3 Décomposition selon le sexe du répondant

Pour compléter l'exposition de nos résultats, nous explicitons une décomposition précédente à celles exposées ci-dessus dans la figure 7 ci-dessous. Le nombre d'individus dans chaque groupe est bien réparti. De plus, chaque estimateur est significativement différent de zéro.

Décomposition - Régression Standard		
Hommes - Femme		
Population	18-24 ans au Canada	
Groupe de référence	Femmes	
Poids	1	0
Coefficients	0,1031511	0,0869456
Caractéristiques	-0,0661961	-0,0499906
Significatif à 5%	Significatif à 10%	Non-significatif
Groupe	**Femmes**	**Hommes**
# Individus	455	472

FIGURE 7 – Décomposition de B-O selon le sexe du répondant

La différence totale dans ce modèle est de 0.129 et est significativement différente de zéro. Pour donner un autre de grandeur, lorsque le $poids = 1$, -179.13% de cette quantité est attribué à la différence de caractéristiques et 279.13% est attribué à la différence de coefficient. Alternativement, lorsque le $poids = 0$, -135.27% viendrait des différences de caractéristiques et 235.27% viendrait des différences de coefficients.

Nous pouvons voir que, selon le poids utilisé, la différence totale est décomposée de la sorte :

$$\begin{aligned}\hat{\Delta}_{Femme} &= \left\{S(\hat{\beta}_{Femme}, X_{i,Femme}) - S(\hat{\beta}_{Femme}, X_{i,Homme})\right\} \\ &+ \left\{S(\hat{\beta}_{Femme}, X_{i,Homme}) - S(\hat{\beta}_{Homme}, X_{i,Homme})\right\}\end{aligned}$$

ou

$$\begin{aligned}\hat{\Delta}_{Homme} &= \left\{S(\hat{\beta}_{Homme}, X_{i,Femme}) - S(\hat{\beta}_{Homme}, X_{i,Homme})\right\} \\ &+ \left\{S(\hat{\beta}_{Femme}, X_{i,Femme}) - S(\hat{\beta}_{Homme}, X_{i,Femme})\right\}\end{aligned}$$

où la première égalité représente le cas avec $Femme$ comme groupe de référence avec un $poids = 1$ et où la deuxième représente le cas avec un $poids = 0$ accordé aux femmes comme groupe de référence. Le deuxième cas est donc équivalent à avoir les $Hommes$ comme groupe de référence. Avec un poids de 0 ou de 1, nous observons que les estimations du paramètre issu de la différence dans les coefficients est positif et que celui issu de particularités spécifiques aux membres de chaque groupe est négatif.

Notre décomposition suggère alors que les écarts des coefficients a un effet positif sur la différence entre les deux groupes. Au contraire, cela suggère aussi que les différences dans les régresseurs tend à diminuer la différence dans la variable dépendante entre les groupes. La partie expliquée[8] contribue positivement à l'écart et la partie inexpliquée, quant à elle, a un effet négatif sur l'écart. Cette deuxième partie est associée à la notion de discrimation pour un groupe. Comme cette valeur a pour effet de diminuer la distance dans les valeurs de la variable dépendante entre les groupes et que les femmes ont une plus grande probabilité prédite de se rendre à l'université que les hommes, cette observation pourrait être interprétée comme un signal de la présence d'une discrimination contre les femmes. Cependant, la valeur de cette estimation lorsque le $poids = 0$ est aussi négatif. Cela viendrait potentiellement miner cette tentative d'explication. Par ailleurs, l'effet que nous observons peut aussi renfermer une part provenant de variables que l'on n'observe pas[9] alors il est nécessaire de faire des interprétations avec une grande réserve.

8. Tel qu'appelée dans Jann (2008). Il l'appel aussi «L'effet quantité».
9. Jann (2008)

7 Conclusion

Dans cette étude, nous avons avancé des estimations quant aux facteurs ayant un impact sur le processus décisionnel des individus entourant leur choix de la quantité de capital humain à accumuler. Les questions que nous avons cherchées à répondre sont : «Quels sont les facteurs déterminants dans le processus de décision binomial d'accumulation de capital humain "Étudier à l'université ou non" pour des gens d'un groupe d'âge sélectionné ?» et «Comment le fait de faire partie d'un groupe donné, par exemple homme ou femme, affecte les effets des autres variables ; quelle part de ces différences est justifiée par les données et quelle part ne l'est pas ?». Bien que nos estimations devront être contemplées avec un regard critique comme tout résultat statistique, l'intuition économique semble vouloir leur accorder une certaine légitimité.

À travers nos estimations, nous remarquons l'importance cruciale que prend l'éducation des parents des individus observés. Que ce constat soit dû à l'implentation d'un sentiment de familiarité au domaine universitaire chez le répondant, d'un système de valeur transmise par les parents qui est particulier, en moyenne, à ceux ayant fréquenté l'université ou à un accès accru à l'information et aux ressources, il n'en reste pas moins que cette variable est d'une importance primordiale dans le pouvoir explicatif de notre modèle. De plus, l'effet du sexe du répondant, surtout présent lorsque les parents de ce dernier n'ont pas été à l'université, semble constamment significativement positif. Finalement, le revenu semble avoir un effet positif, mais ce dernier jouerait un rôle plus effacé que l'éducation parentale. Nous avons, entre autre, observer que son n'effet n'était pas significatif chez les gens des groupes observés dont les parents ont fréquenté l'université et que les estimations associées à la décomposition selon le revenu ne l'étaient pas non plus.

Outre les sujets évoqués au fil du texte, il serait aussi intéressant d'observer les mêmes forces en présence, mais en prenant compte certains facteurs intrinsèques associés à la provenance des individus qui ont dû être omis dans notre étude à cause de la difficulté d'accès aux donnés. Par exemple, la qualité des écoles primaires et secondaires qu'ils ont fréquentées, des mesures de l'abilité du répondant ou le coût des études universitaires seraient des variables d'intérêt dans une telle étude. Suivre un groupe d'individus pendant un grand intervalle de temps avec de telles données permettrait d'établir des hypothèses sur le liens de cause à effet avec plus d'assurance et de précision. Par ailleurs, dans une étude subséquente, nous pourrions aussi analyser des résultats suivant une démarche du même type que celle traitée dans cet article, mais plus en profondeur à l'aide de divers test statistiques. De plus, l'étude de ce sujet, agrémentée de tests portant sur les estimations des coefficients de la décomposition de B-O d'un modèle non-linéaire, sera intéressante à aborder au fils des développements dans la littérature à ce sujet.

Bibliograhie

Bauer, Thomas K. et Mathias Sinning (2007) ''An Extension of the Blinder-Oaxaca Decomposition to nonlinear models", AStA Advances in Statistical Analysis, Springer, vol.92(2), pages 197-206, Mai

Blinder, Allan S. (1973) ''Wage Discrimination : Reduced Form and Structural Estimates," Journal of Human Resources", University of Wisconsin Press, vol. 8(4), pages 436-455.

Borooah, Vani et Sriya Iyer (2005) ''The Decomposition of Inter-Group Differences in a Logit Model : Extending the Oaxaca-Blinder Approach with an Application to School Enrolment in India" MPRA Paper 19418, University Library of Munich, Germany

Drolet, Marie (2005) ''Participation aux études postsecondaires au Canada : le rôle du revenu et du niveau de scolarité des parents a-t-il évolué au cours des années 1990 ¿' Analyse des entreprises et du marché du travail, Statistique Canada, Frévrier

Fairlie, Robert W. (2003) ''An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models" Working Papers 873, Economic Growth Center, Yale University

Finnie, Ross et Richard E. Mueller (2008) ''The Effects of Family Income, Parental Education and Other Background Factors on Access to Post-Secondary Education in Canada", MESA, Soumit au Canadian Journal of Higher Education

Gibbons, Robert (1992) '' Game Theory for Applied Economists '' Princeton University Press, Reprint edition

Greene, William (2011) '' Econometric Analysis'' Prentice Hall; 7^{e} édition

Jann, Ben (2008) ''A Stata implementation of the Blinder-Oaxaca decomposition'', Stata Journal, StataCorp LP, vol. 8(4), pages 453-479, Décembre

Jimenez, Juan de Dios et Manuel Salas-Velasco (2000) '' Modeling Educational Choices. A Binomial Logit Model Applied to the Demand for Higher Education'', Higher Education, Springer, Vol. 40, No. 3 (Oct., 2000), pp. 293-311

Krussel, Per (2004) ''Lecture notes for Macroeconomics I'' Non-publié (Utilisé à titre de support de cours)

Lucas Jr., Robert E. (1988) ''On the mechanics of economic development'' Journal of Monetary Economics, Elsevier, vol. 22(1), pages 3-42, Juillet''

Oaxaca, Ronald (1973) ''Male-Female Wage Differentials in Urban Labor Markets", International Economic Review, Department of Economics, University of Pennsylvania and Osaka University Institute of Social and Economic Research Association, vol. 14(3), pages 693-709, Octobre

Sinning, Mathias et Markus Hahn et Thomas K. Bauer (2008). ''The Blinder–Oaxaca decomposition for nonlinear regression models", Stata Journal, StataCorp LP, vol. 8(4), pages 480-492, Décembre

Stoet, Gijsbert et David C. Geary (2013) ''Sex Differences in Mathematics and Reading Achievement Are Inversely Related : Within- and Across-Nation Assessment of 10 Years of PISA Data''Eshel Ben-Jacob, Tel Aviv University, Israel, Mars

Wooldridge, Jeffrey M. (2010) '' Econometric Analysis of Cross Section and Panel Data '' The MIT Press; 2^{e} édition