Les déterminants du salaire au Pays-Bas

Projet d'économétrie — Département de Sciences Humaines et Sociales

École normale supérieure Paris-Saclay

Louis Bourges, Jean-Baptiste Lagrange-Dupuis et Luc Letonturier

16 mai 2023

Introduction

Depuis Becker et sa théorie du capital humain en 1964, les travaux économiques visant à expliquer les différences de revenu entre les individus se sont multipliées. Becker a théorisé l'existence d'un calcul coût-avantage microéconomique, qui conduit les individus à arbitrer entre le coût d'une année supplémentaire d'études et le gain espéré à long terme [Becker 1964]. Mincer, une décennie plus tard, a enrichi cette approche en incluant l'expérience accumulée au cours des années de travail dans le capital humain [Mincer 1974].

Dans notre étude, nous tenterons de mesurer les effets de ces variables mais aussi d'autres paramètres, à l'instar du genre, de la présence d'enfants, mais aussi des heures travaillées ou de l'âge. Nous nous baserons sur deux enquêtes du LISS [1] menées aux Pays-Bas respectivement en mai 2022 et en septembre 2022. Il s'agira, après une régression classique permettant de comprendre l'influence des différentes variables, de tester la présence d'hétéroscédasticité dans le modèle et, le cas échéant, de la corriger ; de mener un test de Chow pour tenter d'identifier d'éventuels effets de "paliers" quant au lien entre salaire et éducation ainsi que de discuter de la présence d'endogénéité dans le modèle et des moyens à notre disposition pour la corriger. Nous replacerons notre travail dans le contexte de la littérature existante et discuterons aussi de ses limites.

1 Présentation du modèle et de ses limites

1.1 Présentation des variables utilisées

Nous avons sélectionné plusieurs variables au sein de l'enquête *Work and Schooling* et de la base *Background variables*. La variable `éducation`, issue d'un recoupement de plusieurs variables, correspond au nombre d'années de scolarité et d'études achevée (c'est-à-dire ayant conduit à l'obtention d'un diplôme), ses valeurs sont comprises entre 0 (personne n'étant jamais allée à l'école) à 22.5 (personne titulaire d'un doctorat, sachant que la scolarité débute à l'âge de 4 ans aux Pays-Bas). La variable `genre` sépare la population en deux groupes : hommes et femmes, les autres identités de genre ayant été écartées car très peu nombreuses et ayant été jugées difficilement interprétables et `age` indique l'âge des enquêtés. La variable `revenu` prend en compte le revenu brut mensuel autodéclaré, que nous avons préféré au revenu net, plus dépendant des politiques fiscales et de redistribution. La variable `heures` correspond au nombre d'heures de travail effectuées en moyenne chaque semaine tandis qu'`ancienneté` mesure l'ancienneté des salariés

1. *Longitudinal Internet studies for the Social Sciences*, les questionnaires sont administrées par Centerdata

dans leur entreprise (en années) ; il est à noter qu'un licenciement ou une démission remet ce compteur d'ancienneté à zéro puisque c'est l'ancienneté dans l'entreprise actuelle. Enfin, `nbenfants` indique le nombre d'enfants présents dans le foyer.

1.2 Détection et correction de l'hétéroscédasticité

Afin de vérifier la présence ou non d'hétéroscédasticité au sein de notre modèle, nous avons réalisé les deux variantes du test de Breusch-Pagan (avec un test de Fisher et avec un test du rapport de vraisemblance) que nous avons également vérifié manuellement pour nous assurer d'obtenir les mêmes valeurs (p-value). Tous concordent et corroborent la présence d'hétéroscédasticité, qui est par ailleurs observable graphiquement : la répartition des résidus en fonction des données prédites n'est pas homogène et l'on observe une forte variabilité de ces résidus en fonction de certaines variables du modèle, notamment la variable heures (FIGURE 1) où la dispersion semble être plus forte entre 30 et 40 heures qu'avant ou après. Avec le test de Breusch-Pagan avec le rapport de vraisemblance, dont nous avons vérifié les résultats via la commande `bptest()`, nous obtenons une statistique de 75.6 et une p.value de 2.9×10^{-14}, qui nous permet de rejeter l'hypothèse nulle au seuil significatif de 0,01.

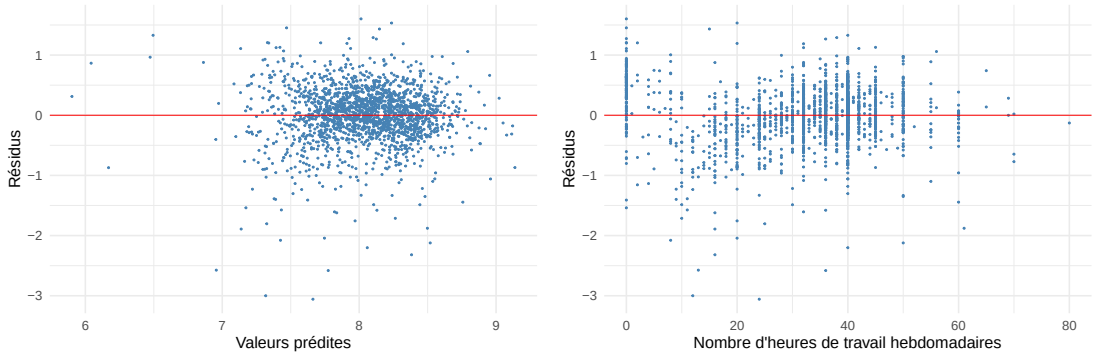

FIGURE 1 – Répartition des résidus avant toute correction

Nous avons corrigé l'hétéroscédasticité en utilisant la méthode de white (variante HC1), via la commande `coeftest`. Après correction, nous obtenons le graphique de répartition des résidus suivant, confirmant notre correction avec une variance des résidus d'environ 0,22 pour l'ensemble des observations. Nous remarquons dans la régression linéaire classique que les valeurs des paramètres des régresseurs n'ont pas changé, ces dernières n'étant pas influencées par l'hétéroscédasticité. 2.

1.3 Détection de l'endogénéité et pistes de correction

2 Résultats principaux

2.1 Analyse et discussion des paramètres

Nous avons tout d'abord préféré effectuer une analyse en log-niveau, le log limitant les risques d'hétéroscédasticité en « écrasant » les observations sans les fausser. À première vue, les paramètres de nos 6 régresseurs de base sur la variable expliqués paraissent plutôt cohérents à ce qui est observé dans la littérature, comme le montre le tableau 1, l'âge a une influence positive mais assez négligeable (paramètre de 0,008), le genre a une forte influence négative sur le salaire pour les femmes. Les heures de travail ont évidemment une influence positive forte : 1 heure travaillée en plus par semaine augmente d'environ 1,4% le log du revenu (voir figure 3).

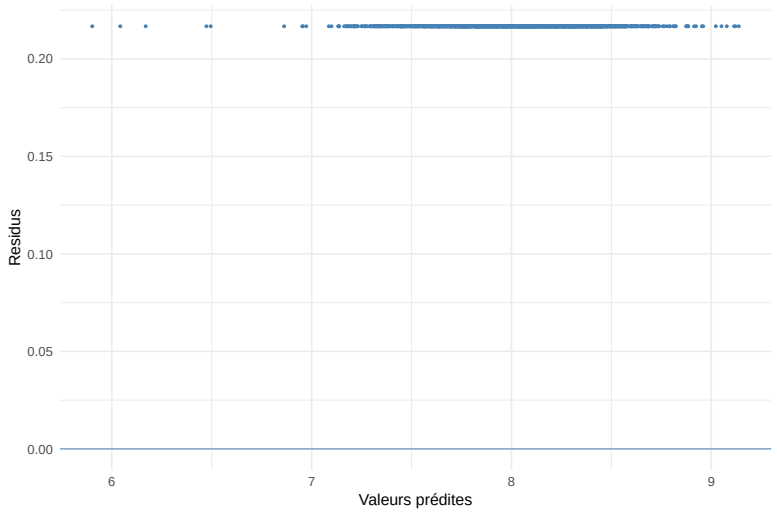

FIGURE 2 – Répartition des résidus en fonction des valeurs prédites, après correction de l'hétéroscédasticité

TABLE 1 – Paramètres estimés du modèle (hétéroscédasticité corrigée)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.8617	0.1255	46.6890	0.0000
age	0.0078	0.0012	6.7330	0.0000
genre	-0.3036	0.0223	-13.6003	0.0000
heures	0.0140	0.0011	13.0212	0.0000
anciennete	0.0030	0.0011	2.7550	0.0059
nbenfants	-0.0072	0.0108	-0.6637	0.5070
education	0.0922	0.0062	14.9240	0.0000

```
## `geom_smooth()` using formula = 'y ~ x'
```

L'ancienneté joue enfin positivement mais faible, le nombre d'enfants négativement mais faiblement, et l'éducation enfin est la variable qui joue le plus fortement avec le nombre d'heures travaillées, avec un paramètre de 0,09. Fait notable, le nombre d'enfants semble assez peu significatif, cette variable étant la seule dans notre régression initiale avec une p-value supérieure à 0,05 (pour atteindre 0,45). Les données du LISS semblent donc confirmer à première vue ce que la littérature a déjà pointé du doigt : une différence structurelle de salaire (brut ici) entre hommes et femmes, une influence positive de l'éducation, de l'ancienneté, de l'âge... Ayant souhaité nous concentrer plus particulièrement sur l'effet de l'éducation, nous avons cependant souhaité vérifier l'existence d'effets de seuils de l'éducation sur la salaire au moment de l'obtention de diplôme. Nous allons désormais évoquer les résultats d'un test de Chow effectué sur l'éducation.

2.2 Test de chow sur la variable `éducation`

L'intérêt de cette partie est d'identifier un éventuel niveau d'études qui fait rupture dans le modèle en menant un test de Chow sur la variable `éducation`. Nous avons décidé de tester l'existence d'un changement structurel dans notre modèle au-delà de 16 ans d'études. Il s'agit de savoir si notre modèle est stable au sein de deux sous-populations, qui se distinguent par la durée de leur scolarité et de leurs études. Notons qu'ici peu importe que l'égalité soit stricte ou pas, étant donné qu'`éducation` n'est pas continue. Formellement, en écrivant notre modèle comme suit :

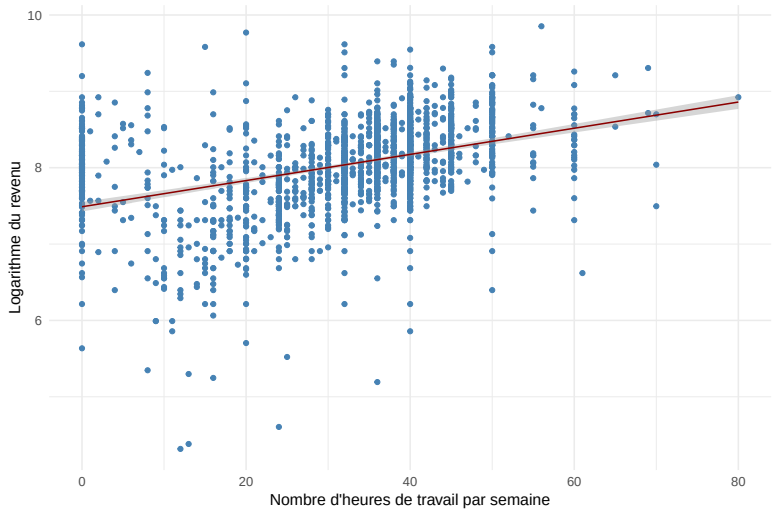

FIGURE 3 – Log revenu en fonction du nombre d'heures de travail

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6$$

où β_i est le coefficient associé au régresseur X_i, nous testons l'hypothèse nulle H_0 contre l'hypothèse alternative H_1 :

$$\begin{cases} H_0 & : \forall i = 1, \ldots, 6, \quad \beta_{i_{\text{éducation}<16}} = \beta_{i_{\text{éducation}>16}} = \beta_i \\ H_1 & : \exists i = 1, \ldots, 6 \ / \quad \beta_i \neq \beta_{i_{\text{éducation}<16}} \text{ ou } \beta_i \neq \beta_{i_{\text{éducation}>16}} \end{cases}$$

Avec $\beta_{i_{\text{éducation}<16}}$ les coefficients du régresseur X_i pour la sous-population la moins diplômée. Les effectifs de nos deux sous-populations sont de tailles suffisantes (1076 personnes d'un côté (les plus diplômés) et 831 personnes de l'autre). Le test (réalisé manuellement, puis confirmé avec la commande `sctest()`) nous donne une statitique de 12.2 et une p-value de 2.6×10^{-15}, nous pouvons donc conclure à l'existence d'un changement structurel dans la relation entre nos régresseurs et le log revenu entre l'échantillon des moins diplômés et celui des plus diplômés. Pour analyser plus finement ce changement, examinons les tableaux des paramètres estimés pour nos deux sous-groupes :

TABLE 2 – Paramètres estimés du modèle pour les sous-groupes des moins diplômés (à gauche) et des plus diplômés (à droite)

	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.5637	0.1422	46.1515	0.0000	6.2337	0.2089	29.8446	0.0000
age	0.0047	0.0015	3.1557	0.0017	0.0106	0.0013	7.9129	0.0000
genre	-0.3778	0.0336	-11.2534	0.0000	-0.2523	0.0288	-8.7473	0.0000
heures	0.0137	0.0012	11.3171	0.0000	0.0138	0.0011	12.5308	0.0000
experience	0.0055	0.0015	3.6817	0.0002	0.0003	0.0016	0.2059	0.8369
nbenfants	-0.0334	0.0140	-2.3833	0.0174	0.0169	0.0126	1.3477	0.1781
education	0.0512	0.0082	6.2682	0.0000	0.0660	0.0109	6.0709	0.0000

En quelques mots, on remarque que l'effet du genre sur le est bien plus faible chez les plus diplômés, ce qui peut s'expliquer par des règles plus strictes pour encadrer les salaires (fonction publique) et pas une plus grande part de femmes au sein des métiers occupés par les plus diplômés, chaque année d'éducation est également légèrement plus valorisée pour les plus diplômés, de même que l'âge. On note également une rupture au niveau des effets du nombre d'enfants : si la variable joue négativement chez les moins diplômés (chaque enfant en plus fait baisser le salaire de 3,3%), à l'inverse elle joue positivement chez les plus diplômés (chaque enfant augmente le salaire de 1,7%).

3 Analyse et mise en perspective des résultats

3.1 Endogénéité et utilisation de variables instrumentales

Une discussion sur la possible présence d'endogénéité dans notre modèle s'impose, dans la mesure où certaines de nos variables, à l'instar d'`éducation`, ne sont pas exogènes. Si l'on suit l'analyse de Becker, les choix d'éducation — la poursuite des études, qui représente un investissement financier, ou leur arrêt, qui permet d'entrer sur le marché du travail — dépendent des ressources des individus. Par conséquent, le nombre d'années d'études dépend au moins pour partie du revenu des parents. Soulignons également l'importance et l'influence du parcours scolaire des parents sur celui de leurs enfants : des parents diplômés pourront inciter leurs enfants à faire des études plus longues, mais surtout les aider (aide aux devoirs notamment). Malheureusement, nous ne disposons d'aucune information ni sur les ressources des parents ni sur leur niveau d'études, qui auraient pu faire des variables instrumentales appropriées.

```
## `geom_smooth()` using formula = 'y ~ x'
```

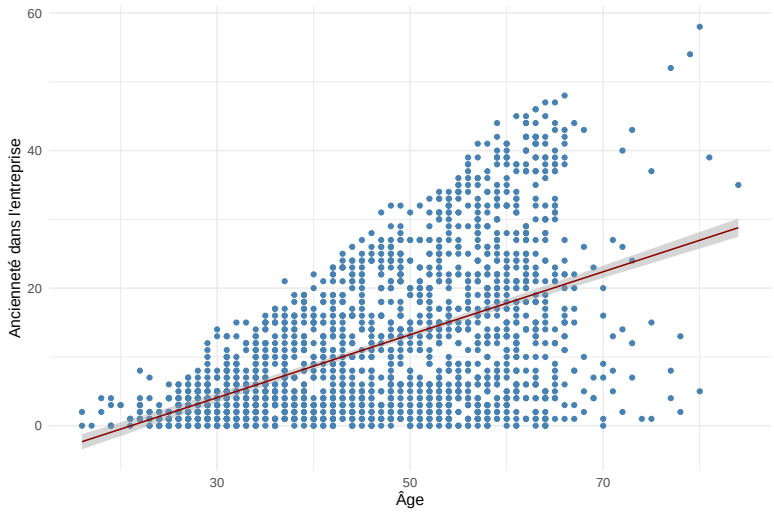

FIGURE 4 – Ancienneté de l'entreprise en fonction de l'âge

Autre variable endogène : l'`ancienneté`. Le coefficient de corrélation entre l'âge et l'ancienneté dans l'entreprise est de 0,52, le plus élevé de tous (c.f. le graphique 4). Soulignons qu'un changement d'employeur remet cette valeur à 0. Un individu jeune, de 25 ans par exemple, ne pourra évidemment pas accumuler une dizaine d'années d'ancienneté. L'âge est quant à lui une variable exogène par excellence ; à partir de là, prendre l'âge comme variable instrumentale pour corriger l'endogénéité associée à l'ancienneté pourrait être une option intéressante à tester. Une deuxième source d'endogénéité pesant sur l'ancienneté serait intéressante à étudier : celle de la formation continue. Ici, la variable éducation se contente d'étudier la durée des études de l'individu, mais qu'en est-il pour ceux qui effectuent, au sein de leur entreprise, une formation complémentaire pour monter en grade ? Ou qui, du fait d'une reconversion professionnelle, effectuent une formation une fois sortis des bancs de l'école ? Un individu avec beaucoup d'ancienneté a d'autant plus de chances, au cours de ses années passées avec son même employeur, d'avoir effectué une formation complémentaire qui n'est pas prise en compte dans l'éducation mais joue pourtant sur le salaire au travers de l'ancienneté (nous faisons ici la supposition qu'un individu avec le même employeur depuis longtemps tend statistiquement à plus réaliser de formation en continu, le lien entre stabilité de l'emploi et propension à faire des formations étant confirmé empiriquement).

3.2 Limites et inscription dans la littérature existante

Plusieurs limites peuvent évidemment être pointées du doigt quant au modèle utilisé, au choix des variables et à la méthodologie. Nous avons fait le choix d'utiliser des données en coupe instantanée, c'est-à-dire pour une seule période. Mais comme le mettent en avant S. Lollivier et P. Pollet, les coupes instantanées sont risquées puisqu'elles peuvent ne pas prendre en compte des phénomènes conjoncturels affectant les données [Lollivier et Pollet 2003]. Comment interpréter par exemple une crise économique ou une vague de licenciements massifs ? Ou une nouvelle loi sur le temps de travail ? Pour nos données datant de 2022, ce pourrait être la guerre en Ukraine qui influerait sur certains secteurs et emplois comme choc non pris en compte dans des coupes instantanées. Concernant l'évolution de l'influence de certaines variables explicatives, l'étude réalisée par D. Witkowska sur les déterminants des salaires en Pologne de 2001 à 2009 est exemplaire. Elle y montre montre qu'entre 2001 et 2009, l'écart entre la part de la rémunération imputable aux études chez les personnes ayant été à l'université d'une part et d'autre part chez celles et ceux s'étant arrêté avant *(general secondary* ou *post secondary)* tend à se réduire : le paramètre pour les individus ayant été jusqu'à l'université était 4 fois plus élevé en 2001 que ceux s'étant arrêté au lycée ; en 2009 le rapport n'était plus que de 2 [Witkowska 2014]. Précisons que ceci ne signifie pas que les individus ayant été à l'université gagnent 2 fois plus, d'une part le salaire est en log et d'autre part cela représente seulement les différences de salaire imputables à l'éducation). Nos données en coupe instantanée ne permettent pas de mesurer de telles évolutions.

Les deux auteurs recommandent aussi, dans leur papier, de prendre en compte non seulement le revenu, mais également les opportunités d'emploi et le fait de ne pas se retrouver au chômage dans l'analyse : des formations différentes donnent, selon eux, des accès et opportunités différentes d'accès à l'emploi qui ne sont pas pleinement retranscrites par le biais du salaire. Pour un même salaire, des individus dans des branches différentes d'emploi peuvent avoir une certitude tout à fait différente quant à la stabilité de leur poste ou des possibilités d'embauche à la sortie, ce qui serait un complément intéressant à l'analyse de Becker et pourrait être pris en compte par une variable telle que le nombre des périodes de chômage ou encore une variable associant à un secteur du monde du travail la probabilité de se retrouver au chômage. Une autre critique pourrait être faite quant aux choix de nos variables, somme toute pertinentes mais auxquelles nous aurions pu en ajouter d'autres. Ainsi le papier de D. Witkowska prend également en compte le lieu de résidence (urbain, rural, nombre d'habitants), le type de contrat ou encore la taille de l'entreprise. Nous avons tenté d'inclure plus de variables (notamment concernant la question du lieu de vie et de la possession d'une maison/appartement ou bien sa location) mais cela nécessitait la fusion d'autres enquêtes du LISS et nous descendions rapidement à un nombre trop restreint d'observations. Somme toute notre modèle ne remet pas en question l'abondante littérature à ce sujet, il s'inscrit dans leur droite ligne quant aux résultats que nous pouvons en tirer. Ce qui ressort est plutôt la possibilité d'usage d'autres variables, d'autres méthodes et données pour obtenir des résultats plus fins et cette fois, peut être pouvoir nuancer et se distinguer de la littérature déjà existante.

Références

[1] Gary S. Becker. « Human capital : a theoretical and empirical analysis, with special reference to education ». In : *Chicago : University of Press* 19.6 (1964), p. 4.

[2] Stéfan Lollivier et Pascale Pollet. « Impact de la formation initiale sur les rémunérations au cours de la vie active ». FR. In : *Revue d'économie politique* 113.6 (2003), p. 801-827.

[3] Jacob Mincer. « Schooling, Experience, and Earnings. » In : *Human Behavior & Social Institutions* 2 (1974).

[4] Dorota Witkowska. « Determinants of wages in Poland ». In : *Metody Ilościowe w Badaniach Ekonomicznych* 15.1 (2014), p. 192-208.

Annexes

A Précisions sur les variables utilisées et leur codage

A.1 La variable `age`

Il s'agit de la variable `leeftijd`, issue du questionnaire *Background variables*, reprise telle quelle. Les valeurs vont de 16 à 84 ans.

A.2 La variable `education`

Il s'agit du regroupement de deux variables : nous avons d'abord récupéré les données de la variable `oplmet` du questionnaire *Background variables* et avons converti les différentes modalités en nombre d'années d'éducation (scolarité et études), comme suit :

`oplmet` : Highest level of education with diploma

Valeur d'origine	Valeur de remplacement
1. primary school [8 ans]	8 (de 4 à 12 ans)
2. vmbo (intermediate secondary education, US : junior high school) [4 ans]	12 (après l'école primaire)
3. havo/vwo (higher secondary education/preparatory university education, US : senior high school) [5-6 ans]	13.5 (après l'école primaire)
4. mbo (intermediate vocational education, US : junior college) [1-4 ans]	15.25 (après VMBO, HAVO ou VWO, soit en moyenne 2.5 + 12.75)
5. hbo (higher vocational education, US : college)[4 ans]	16.75 (après VMBO, HAVO ou VWO, 4 + 12.75)
6. wo (university) [3 ans]	17 (après 1ere année HBO ou après VWO, 14 + 3)
7. other	suppression des observations
8. Not (yet) completed any education	0
9. Not yet started any education	0

La variable `oplmet` ne proposant pas de modalité *master* ou *Ph.D.*, nous nous sommes ensuite appuyés sur la variable `cw22o005` du questionnaire *Work and schooling*. Cette variable, très complète, comprend 28 modalités, ce qui la rendait trop difficile à coder étant donné que nous ne maîtrisons pas les subtilités du système universitaire néerlandais, mais nous nous sommes contentés de récupérer les trois modalités qui nous intéressent, afin de compléter le codage de la variable `education`, comme suit :

`cw22o005` : What is the highest level of education that you have completed with diploma or certificate ?

Valeur d'origine	Valeur de remplacement
25. academic education, bachelor [3 ans]	17 (soit la valeur que nous avions déjà)
26. academic education, master [1-3 ans]	19 (17 +2)
27. doctor's degree (Ph.D, including doctoral research program to obtain Ph.D) [3-4 ans]	22.5 (19+3.5)

À l'issue du codage, on obtient la répartition suivante :

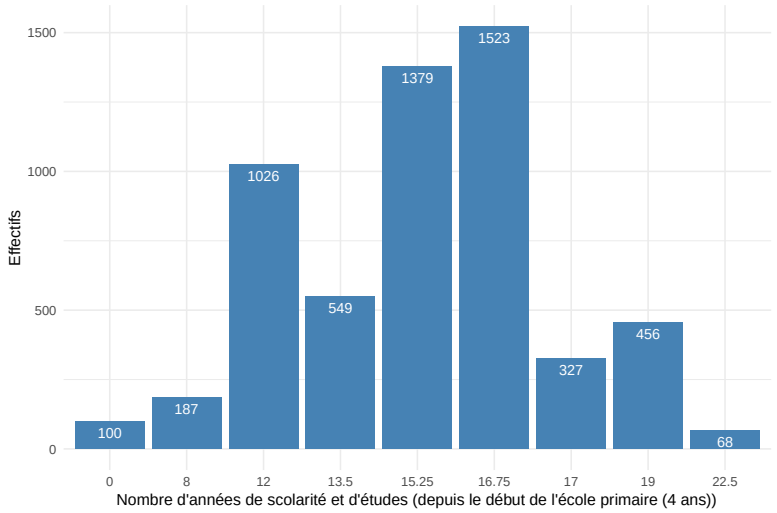

FIGURE 5 – Niveau d'éducation (avec diplôme) des individus de l'échantillon

A.3 La variable genre

Cette variable est issue de la variable `geslacht` du questionnaire *Background variables*, dont nous avons enlevé toutes les réponses autres (ni masculin, ni féminin), qui sont par ailleurs très minoritaires (4 observations, soit moins de 0,1% des observations)

A.4 La variable revenu

Issue de la variable `brutoink` du questionnaire *Background variables*, elle correspond au revenu brut mensuel individuel (*personal gross monthly income*), dans notre modèle, nous utilisons principalement le logarithme de cette variable (`log_revenu`). Nous avons décidé de supprimer les observations pour lesquelles le revenu déclaré est nul, suivant l'avertissement présent dans le *codebook* du LISS :

> Since some people prefer not to make their income information available to Centerdata, a 0 (zero) can mean two different things : (1) that there is no income at all, or (2) that a panel member does not know what the income is or does not want to make that information available to us.

A.5 La variable heures

Il s'agit de la variable `cw22o127` du questionnaire *Work and schooling*, qui correspond à la question : « How many hours per week do you work on average ? / How many hours per week did you work on average ? », que nous avons reprise telle quelle. Les valeurs s'échelonnent de 0 à 80 heures hebdomadaires, avec une médiane égale à 36 heures.

A.6 La variable ancienneté

Issue du questionnaire *Work and schooling* (variable `cw22o134`), il s'agit de la réponse à la question : « In which year did you enter into employment with your current employer ? ». Nous n'avons appliqué aucun traitement particulier aux observations, en dehors de la suppression des *NA*. Les valeurs vont de 0 à 58 ans d'ancienneté, avec une médiane à 7 ans et une moyenne à 12 années d'ancienneté.

A.7 La variable nbenfants

Issue de l'enquête *Background variables*, il s'agit du nombre d'enfants vivant au sein du foyer (« Number of living-at-home children in the household, children of the household head or his/her partner »). Nous l'avons reprise sans modifier ses valeurs.