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

15 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 `education`, 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'`experience` 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'expérience à zéro puisque c'est l'expérience 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.

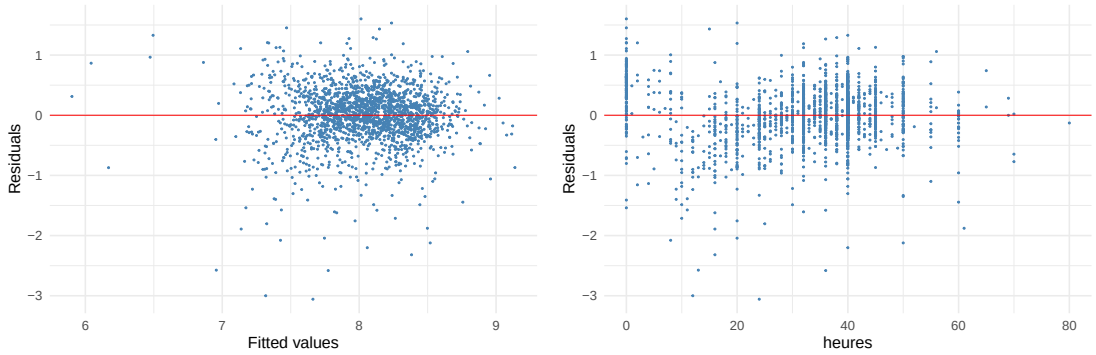

FIGURE 1 – Répartition des résidus

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.

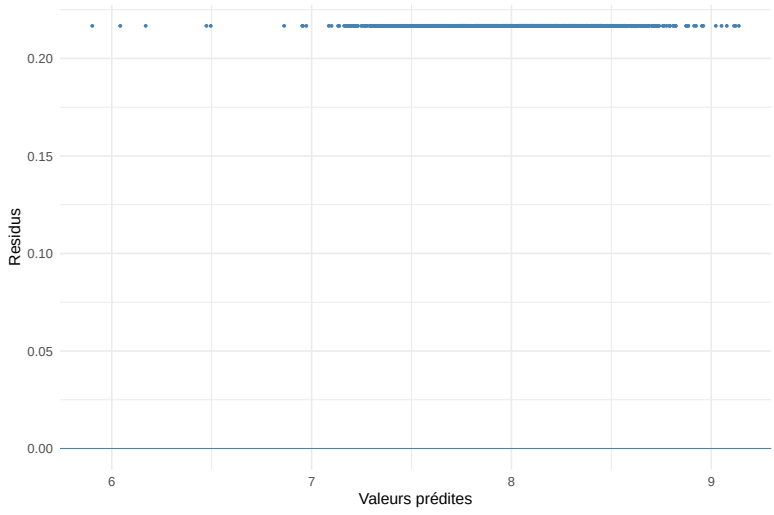

FIGURE 2 – Répartition des résidus en fonction des valeurs prédites

1.3 Détection de l'endogénéité et pistes de correction

2 Résultats principaux

2.1 Analyse et discussion des paramètres

TABLE 1 – Tableau des résidus

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.8617	0.0959	61.13	0.0000
data$age	0.0078	0.0010	7.77	0.0000
data$genre	-0.3036	0.0222	-13.70	0.0000
data$heures	0.0140	0.0008	17.01	0.0000
data$experience	0.0030	0.0011	2.65	0.0081
data$nbenfants	-0.0072	0.0095	-0.76	0.4483
data$education	0.0922	0.0046	19.93	0.0000

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 0,14% le log du revenu (voir figure 3).

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

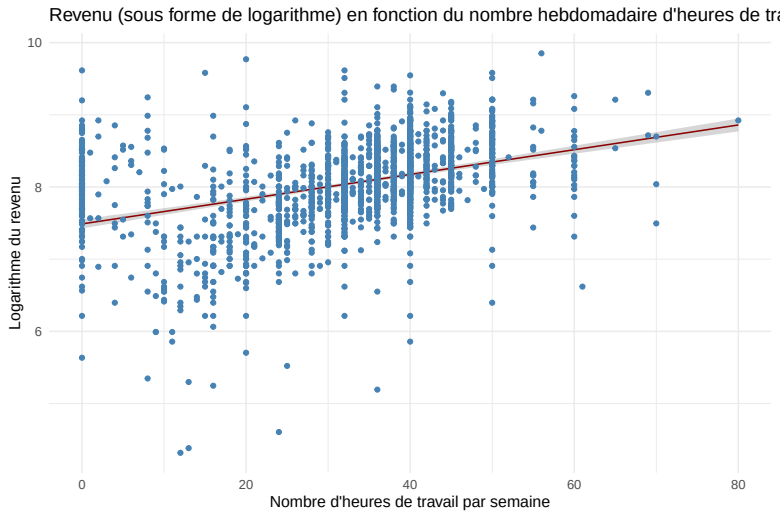

FIGURE 3 – Log revenu en fonction du nombre d'heures de travail

L'expérience 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'expérience, 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.

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. Premièrement la variable "éducation" au travers du nombre d'années d'éducation n'est pas exogène. Comme Becker l'avait déjà pointé, tous les individus n'effectuent pas les mêmes choix d'éducation car tous n'ont pas les mêmes ressources pour assumer leurs choix. Une année supplémentaire d'études représente un investissement financier dépendant des ressources à disposition. Par conséquent, au travers du nombre d'années d'étude, c'est probablement également le revenu des parents qui peut se lire. En parallèle, soulignons l'importance et l'influence du parcours scolaire des parents sur celui de leurs enfants, notamment lorsque les parents ont effectué des études supérieures. Des parents diplômés pousseront certainement leurs enfants à faire plus d'études. Une deuxième variable inconnue jouant sur le nombre d'années d'éducation et l'influence sur le salaire indirectement serait donc les ressources disponibles pour l'individu à différents moments de ses études mais aussi le parcours scolaire de ses parents. Deuxièmement, l'expérience est évidemment également endogène. Le coefficient de corrélation entre l'âge et l'expérience dans l'entreprise est de 0,52, le plus élevé de tous les coefficients entre variables (soulignons que l'expérience est mesurée comme expérience avec son employeur actuel, une période de chômage remettant cette variable à zéro). Un individu jeune, de 25 ans par exemple, ne pourra évidemment pas accumuler une dizaine d'années d'expérience et au travers de l'expérience, l'âge influe durablement. 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'expérience pourrait être une option intéressante à tester. Une deuxième source d'endogénéité pesant sur l'expérience 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 sorti des bancs de l'école ? Un individu avec beaucoup d'expérience 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'expérience (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).

Et ici on peut écrire ... et insérer des blocs de code qui s'éxécutent, avec le code et le résultat qui s'affichent

```
a <-  2+2
a

## [1] 4
```

ou juste le résultat :

```
## [1] 6
```

ou totalement invisibles :

Et ensuite on peut citer les résultats : à première vue $4 < 6$ mais je crois que c'est 8 qui est le plus grand.

```
## [1] 5615
## [1] 0
```

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

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 :

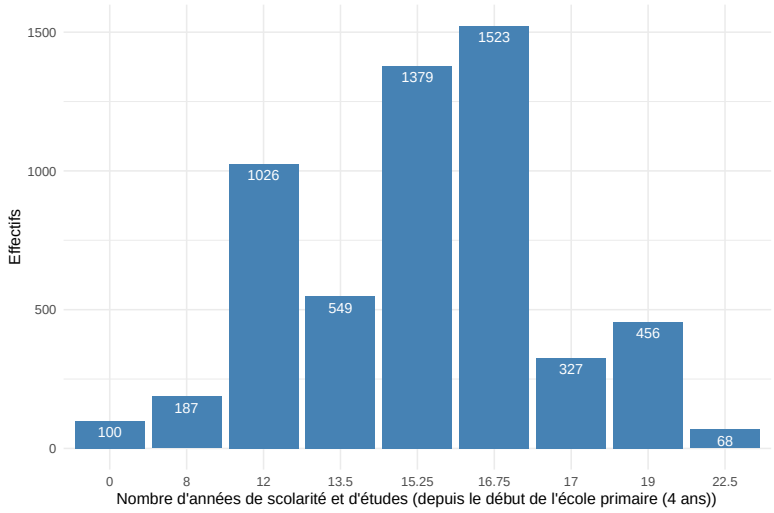

FIGURE 4 – Niveau d'éducation (avec diplôme) des individus de l'échantillon

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 `experience`

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.