Analyse des données du Panel Européen des Ménages

Le salaire selon le niveau d'étude

1. **Analyse des distributions des différentes variables**

Les log-salaires et l'expérience professionnelle (mesurée en mois) sont des données numériques. Tout d'abord, il y a une raison pour laquelle on utilise le log-salaire plutôt que le salaire. En effet, le logarithme est une fonction mathématique qui nous sert d'outil pour réduire l'effet des très grands salaires. Ces derniers pourraient tronquer notre analyse, il est donc préférable d'utiliser le log-salaire. Pour pouvoir comprendre visuellement et facilement la répartition de ces deux variables, il est utile d'utiliser des histogrammes afin d'observer en un coup d'œil combien de fois chaque valeur apparaît. Pour les log-salaires, les salaires semblent suivre une courbe en cloche qu'on appelle "distribution normale", c'est-à-dire qu'il y a beaucoup de valeur proche de la moyenne et peu, voire pas de valeur extrême. Cependant, on peut voir sur l'histogramme qu'il existe bien des valeurs extrêmes, du moins plus que si le log-salaire suivait une distribution normale. Aussi, on peut effectuer des tests statistiques qui nous permettent de vérifier si la distribution des log-salaires suit une distribution normale, et ces tests nous montrent qu'elle ne suit effectivement pas ce type de distribution. Pour l'expérience professionnelle, on voit directement que sa distribution ne ressemble pas une cloche, donc elle ne suit pas une distribution normale. Il y a beaucoup de monde avec peu d'expérience et il existe quelques personnes avec une longue expérience. Une autre variable numérique est la variable "mois". Sa répartition est parfaitement uniforme, chaque mois a exactement le même nombre d'occurrence. L'utilisation d'un diagramme est idéale pour voir cette distribution.

Enfin, les deux dernières variables regroupent les personnes en catégories. D'abord, le sexe contient deux catégories : homme ou femme. Le diagramme en camembert est une bonne manière d'observer la répartition entre homme et femme dans l'échantillon, et on observe une majorité de femme. Le niveau d'études contient plusieurs catégories (primaire, secondaire, professionnel court, professionnel long, cycle 2 et cycle 3). Ici, on utilise également un diagramme en barre qui illustre de grandes différences entre le nombre de personnes pour chaque niveau d'étude. Par exemple, il existe beaucoup de personnes qui ont suivi un cursus professionnel court tandis que peu ont atteint un doctorat.

2. **Comparaison des log-salaires moyens selon le niveau d'études**

Ici on continue de raison en log-salaire. Le log-salaire moyen dans l'échantillon est égal à 3,89 approximativement. Lorsqu'on regarde le log-salaire moyen pour chaque niveau d'étude, nous pouvons séparer ces moyennes en deux catégories. Premièrement, les niveaux d'études dont le log-salaire moyen est plus élevé que la moyenne de l'échantillon. Cela concerne les personnes ayant été au bout du deuxième cycle avec un log-salaire moyen d'environ 4,07 et ceux ayant fini des études de troisième cycle (log-salaire moyen de 4,29 environ). Deuxièmement, les niveaux d'études dont le log-salaire moyen est inférieur à la moyenne de l'échantillon. D'abord la catégorie professionnelle long et secondaire qui ont un log-salaire moyen plutôt proche de la moyenne de l'échantillon

(approximativement 3,86 et 3,84 respectivement). Ensuite les catégories un peu plus éloignées avec les études professionnel court et son log-salaire moyen d'environ 3,73, puis la catégorie primaire (environ 3,6). Ce qu'on peut comprendre de ceci, c'est que le niveau d'étude a un impact sur la rémunération. Les personnes ayant suivi des longues études semblent être mieux rémunérées que celles qui ont suivi des études courtes.

$$\frac{2292*4{,}070447 + 924*3{,}603021 + 2664*3{,}726121 + 732*3{,}857681 + 1308*3{,}835276 + 936*4{,}287887}{2292+924+2664+732+1308+936} = 3{,}888761$$

Si l'on reprend le log-salaire mentionnée dans le tableau des paramètres de la loi normale, alors la moyenne pondérée par les effectifs pertinents du log-salaire moyen par niveau d'étude et le log-salaire moyen dans l'échantillon sont égales. Si on prend la moyenne non pondérée qu'on peut observer dans le dernier modèle de régression, le log-salaire moyen serait égale à 3,89674 (environ 3,90) ce qui est légèrement plus élevé que la moyenne pondérée.

Une moyenne pondérée légèrement plus faible peut se comprendre car il y a beaucoup de personnes qui il y a plus de personnes qui ont un niveau d'étude dont le log salaire moyen est plus faible que la moyenne d'échantillon. En effet, le nombre de personnes ayant un diplôme de deuxième ou troisième cycle s'élève à 3480, alors que la somme des personnes des autres catégories s'élève à 5628. Donc si on pondère la moyenne, qu'on accorde de l'importance aux effectifs pour chaque niveau d'étude, celle-ci sera légèrement plus faible que la moyenne non pondérée.

3. Problèmes de multicolinéarité et solutions pour l'estimation des salaires

Dans le premier modèle on essaye d'estimer le log-salaire pour tous les niveaux d'études. Seulement nous faisons face à un problème appelé multicolinéarité car les différents niveaux d'étude sont fortement liés entre elles, ce qui rend l'estimation du modèle impossible. Pour résoudre ce problème, une solution consiste a construire un modèle d'estimation similaire mais en retirant l'une des catégories d'études. Ici, le programme a décidé de retirer la catégorie troisième cycle, ce qui fait que ce modèle est identique au modèle estimé pour tous les niveaux d'études sauf pour le troisième cycle. Ainsi, ce niveau d'étude devient la référence et le modèle estime les différences de salaire entre les autres niveaux d'étude et cette référence.

Le modèle estimé pour tous les niveaux d'études sauf primaire est similaire à ces deux modèles, à la différence que la référence est ici le niveau d'étude primaire. Donc ce modèle estime les différences de salaire entre les niveaux d'étude et ceux s'étant arrêtés après le primaire.

Une autre solution pour adresser le problème de multicolinéarité consiste à retirer la constante du modèle. Précédemment quand on prenait une catégorie comme référence, la moyenne du log-salaire de cette catégorie se retrouvait dans la constante ce qui nous permettait d'estimer le salaire des autres catégories par rapport à cette constante. Dans ce modèle-ci, on retire cette constante de sorte de pouvoir estimer le salaire de chaque niveau d'étude sans prendre de référence et sans avoir de problème de multicolinéarité. Enfin, une dernière méthode pour corriger le problème de multicolinéarité consiste à prendre la moyenne des log-salaires comme référence. Un des modèles utilise la moyenne pondérée calculée précédemment, et un autre utilise la moyenne non pondérée. Donc on estime les salaires des différents niveaux d'études par rapport à la moyenne, pondérée ou non selon le modèle. Une différence avec les premiers modèles cependant, c'est qu'ici on impose une contrainte. La somme des coefficients, la valeur qui nous permet d'estimer la différence entre chaque niveau d'étude et la référence, doit être égale à 0

4. **Interprétation et prédictions des modèles de régression**

Dans presque tous les modèles le R^2 est égale à 0.1939 et le R^2 ajustée à 0.1934, à l'exception du modèle estimé sans constante où le R^2 et le R^2 ajustée sont les deux égales à 0.9885. Cela veut dire que pour presque tous les modèles, les niveaux d'études expliquent 19,39% de la variation des salaires dans l'échantillon. L'interprétation du R^2 pour un modèle sans constante est plus compliquée, donc on ne peut l'interpréter de la même manière et dire que ce modèle est une meilleure estimation que les autres.

Ensuite, dans tous les modèles, l'impact des différents niveaux d'études sont tous importants pour l'estimation du log-salaire.

Dans tous les modèles le log-salaire moyen estimé est de 3,87 environ, ce qui est légèrement plus bas que le log-salaire observé qui est de 3,89 approximativement. De plus, on observe une corrélation modérée entre les log-salaires observés et prédits par les modèles, ce qui indiquerait que les prédictions du modèle sont assez proches des valeurs réelles malgré les quelques différences entre prédiction et observation. Enfin, les estimations respectent la tendance des salaires reçus par rapport au niveau d'étude. Les personnes ayant suivi de longues études peuvent s'attendre à des salaires plus élevés que ceux ayant suivi des études courtes.

D'après les estimations, les personnes ayant un diplôme primaire peuvent s'attendre à un salaire horaire de 36,7 francs ($e^{3,60302} \approx 36,7$). Les personnes ayant obtenu un doctorat peuvent espérer un salaire 98% plus élevé, et 60% de plus pour celles ayant un diplôme de cycle 2 ($e^{0,68487} - 1 \approx 98\%$; $e^{0,46743} - 1 \approx 60\%$). Les personnes ayant suivi un cursus professionnel long peuvent prétendre à un salaire 29 % supérieur, tandis que celles ayant suivi un cursus professionnel court peuvent espérer 13 % de plus ($e^{0,25466} - 1 \approx 29\%$; $e^{0,46743} - 1 \approx 13,10\%$). Enfin, ceux ayant arrêté après le secondaire peuvent s'attendre à un salaire 26 % plus élevé que ceux de la catégorie primaire ($e^{0,23226} - 1 \approx 26\%$).

A titre personnel, je m'attendais à une rémunération un peu plus élevée pour ceux qui ont obtenu un diplôme de cycle 2 par rapport à la catégorie primaire. Plus précisément, je pensais que ce serait plus proche du salaire de ceux ayant eu un doctorat. Je ne pensais qu'en France on ne jugeait pas les personnes ayant un doctorat à leur juste valeur, et notamment qu'un bac+8 n'aidait pas à l'emploi dans l'industrie (je reprécise en France).

Aussi, je m'attendais à ceux qui suivent un cursus pro court aurait un salaire plus élevé que ceux ayant arrêté après le bac, car les formations du générale ou technologique du secondaire ne forment pas directement à un métier mais plutôt à des études supérieures (BTS, licence etc…) tandis que les formations courtes permettent de se spécialiser et travailler directement. Donc un salaire plus élevé pour cette dernière catégorie. Cependant je suis probablement biaisé par ma vision moderne des choses, et la situation en 1995 était sûrement très différente.

5. **Comparaison des modèles de régression et choix du modèle optimal pour l'analyse des salaires**

Les modèles sont tous les mêmes et prédisent la même chose sauf un. En effet, chaque modèle utilise une référence et estime la différence du salaire par rapport à cette référence, à l'exception du modèle sans constante. Pour les modèles qui utilisent une référence, l'interprétation est facile. On y

observe la différence de salaire par rapport à la référence. Cependant, les modèles où on utilise une moyenne comme référence compliquent l'interprétation car on utilise une contrainte.

Pour le modèle sans constante, l'interprétation peut être considérée comme plus simple étant donné qu'on n'a pas besoin de calculer les différences de salaire par rapport à la référence, mais c'est également un défaut car on perd cette interprétation en pourcentage de différence par rapport à une référence qui est économiquement plus intéressante car l'intuition est plus facile à comprendre et expliquer. On arrive mieux à s'imaginer les différences. De plus, on ne peut comparer le R^2 d'un modèle sans constante car il peut être surestimé.

Une différence entre chaque modèle est la valeur du test t de chaque variable qui n'est absolument pas la même entre tous les modèles. En valeur absolue, le modèle sans constante a les valeurs du test t les plus élevées. Il faut noter que la valeur de F de ce modèle est également beaucoup plus élevée que les autres modèles. Néanmoins, je dirai que le modèle qui fait le plus sens économiquement parlant est le modèle estimé pour tous les niveaux d'études sauf primaire. En effet, il facilite l'explication des différences de salaires entre les niveaux d'études en s'exprimant en pourcentages, cela fait aussi plus de sens économiquement parlant. De plus, prendre le niveau "primaire" comme référence est pertinent, car il représente le niveau d'études le plus bas, offrant une base simple pour comparer les autres catégories. J'ajouterai que cela permet aussi d'obtenir des pourcentages positifs pour les autres niveaux d'études, ce qui rend ces différences plus intuitives et faciles à s'imaginer.