L'accès à un haut revenu aux États-Unis

Léa Aumagy[1], **Dhalil Bello**[2], **Hédi Bouchneb**[3]

[1,3]Etudiants en Master 1 Economie à Aix-Marseille School of Economics
[2]Etudiant en Master 1 Economie et en Magistère 2 Ingénieur Economiste à Aix-Marseille School of Economics

22 janvier 2023

Résumé

Le **revenu**, la rémunération de la force de travail d'un individu, est un facteur de richesse. Cependant, il engendre de nombreuses **inégalités** dues à un niveau différent selon l'emploi occupé, ce qui crée une hiérarchisation de la société. Les **États-Unis** est un pays où l'on retrouve une grande dispersion des revenus due à leur système social, que l'on mesure avec l'indice de Gini. Malgré le fait que ce soit un pays très développé, il y a la présence de très nombreuses inégalités. Ces différentes inégalités entre les individus provoquent de nombreuses **discriminations** que ce soit ethnique ou encore de genre. Néanmoins, il existe de nombreux facteurs de **richesse**, qui permettent une élévation sociale des individu tel que la famille ou encore l'école, mais les individus peuvent avoir plus de facilité (difficultés) grâce (à cause) de leur **origine sociale**. Nous allons étudier ces différents effets par le biais d'une étude économétrique à l'aide d'une base de données.

Mots clés : *Revenu, Inégalités, États-Unis, Discrimination, Richesse, Origine sociale*

Table des matières

1 Introduction

Chaque ménage possède des avoirs lui permettant de disposer de ressources futures, c'est ce qu'on appelle le patrimoine brut. Ce patrimoine brut comprend les actifs financiers, les actifs professionnels et les biens immobiliers possédés. Dans le patrimoine, on y trouve donc le revenu, qui est un facteur de richesse. Les individus qui perçoivent un certain montant de salaire se distinguent dans différentes classes sociales. D'un ménage à un autre, le revenu perçu diffère selon l'emploi occupé, mais aussi par différentes caractéristiques de l'individu. Des personnes provenant de hautes classes sociales ou ayant certaines caractéristiques auront plus de facilités à accéder à de hauts postes dans la hiérarchie d'une entreprise. De même, il est possible que pour un même poste d'une personne à l'autre le revenu peut différer, à cause du profil de l'individu, ce qui engendre des inégalités et de la discrimination.

Le revenu engendre de nombreuses inégalités que l'on peut mesurer grâce à l'indice de Gini, qui permet de rendre compte de la répartition du revenu, du patrimoine ou du salaire, au sein d'une population. Les Etats-Unis est l'un des pays les plus développés, mais contient au haut indice de Gini, qui, selon la Banque Mondiale, est de 0,415 contre 0,324 pour la France. Cependant, afin d'observer ce phénomène, il est intéressant d'analyser le revenu médian. Le revenu médian est le revenu qui permet de diviser la population en deux groupes de même taille, ceux qui perçoivent un salaire inférieur au revenu médian et ceux qui perçoivent un salaire supérieur au revenu médian. C'est-à-dire que le revenu médian est la valeur du niveau de vie qui partage la population en deux parties égales : 50% des individus ont un niveau de vie en dessous et 50% au-dessus. Le seuil de pauvreté est donc égal à 50% de la médiane de la distribution des niveaux de vie. Le coefficient de Gini peut s'observer grâce à la courbe de Lorenz, où l'on peut voir le fait que les individus ne perçoivent pas les mêmes salaires, donc les inégalités de salaires. Si l'indice de Gini est de 0, alors la courbe de Lorenz est une courbe d'égalité parfaite de la répartition des salaires, et donc tous les individus perçoivent le même salaire, mais ce n'est pas le cas.

On pourrait donc se demander quelles sont les différentes caractéristiques d'un individu percevant un revenu supérieur au revenu médian, notamment aux États-Unis.

Le revenu est un facteur d'inégalité, qui engendre une hiérarchisation de la société. Cependant, chaque individu est doté d'un capital culturel transmis par la famille et par l'éducation. L'éducation est connue comme une voie d'accès à des postes plus ou moins rémunérateurs selon le niveau d'éducation. On peut donc analyser si le niveau d'éducation est le seul facteur du revenu, ou il existe d'autres facteurs qui influencent le salaire, tel que la discrimination a l'entrée du marché du travail qui engendre plus de difficultés à certaines personnes d'accéder à des postes de hauts niveaux.

Pour traiter notre sujet, nous allons d'abord étudier les faits dans la littérature. Nous pouvons dans un premier temps analyser de quelle façon le revenu est un facteur de stratification sociale, et puis comment les individus peuvent exercer une mobilité sociale pour permettre une élévation de leur niveau social en accédant à des postes de hautes rémunérations. Enfin, nous allons analyser la présence de la discrimination au sein de certains postes de haut niveau, qui serait un frein à certains types de profil d'individu, et donc de percevoir un salaire élevé. Une étude empirique réalisée aux États-Unis va être étudiée pour analyser les différents phénomènes, et anticiper les effets dans une analyse économétrique à l'aide d'une base de données. Dans notre étude économétrique, nous allons étudier différents profils d'individus qui perçoivent plus ou moins de 50.000 dollars par an, qui sera un revenu annuel supérieur au revenu médian de la population. Nous allons effectuer des comparaisons entre les individus, évaluer selon les modalités de chaque variable quelles sont celles qui ont une forte relation avec le fait de percevoir un revenu élevé. Puis nous effectuerons une étude économétrique avec les variables significatives. Nous pourrons donc connaître quels sont les caractéristiques d'un individus face à l'accès à un haut revenu aux États-Unis.

2 Présentation du modèle économique

2.1 Revue théorique

2.1.1 Le revenu, un facteur d'inégalité

Dans une société, chaque individu qui est actif sur le marché du travail et en âge de travailler perçoit un revenu en échange de sa force de travail, comme la rémunération d'une activité ou d'un travail. Cependant, chaque individu ne perçoit pas le même montant de revenu, que ce soit mensuel, annuel ou autre, cela provoque des inégalités entre eux. Ces différences de revenu et de patrimoine sont des facteurs de stratification sociale qui engendre une hiérarchisation des individus au sein de la société. La stratification sociale désigne la division des sociétés humaines en catégories sociales, dans lesquelles il existe la présence d'une certaine homogénéité résultant de l'ensemble des différences sociales associées aux différences de richesses, de pouvoir, de prestige ou de connaissance.

Les inégalités de revenus représentent donc les différences de revenu entre des individus, des ménages, des groupes sociaux ou des espaces géographiques. Ces inégalités monétaires portent sur le flux de revenu que ce soit mensuel ou encore annuel, et non sur le stock de patrimoines. Par ailleurs, le revenu doit être différencié du revenu disponible, c'est-à-dire du revenu restant après paiement des impôts et des prestations sociales. Les écarts de revenus sont principalement liés à la situation professionnelle des particuliers ou aux biens qu'ils possèdent. L'influence d'autres facteurs tels que le niveau d'éducation, le sexe, l'origine sociale ou géographique peut être mise en évidence. Outre l'analyse des écarts de revenus dans le temps (inégalités « transversales »), les inégalités peuvent également être étudiées dans d'autres dimensions, par exemple sous l'angle de la mobilité sociale, par exemple d'une génération à l'autre (on parle alors d'"inégalités longitudinales").

Dans notre étude, nous allons donc distinguer ces inégalités de revenu selon différents comportements selon l'origine sociale et la discrimination.

2.1.2 Inégalités de revenu, une hiérarchisation de la société

Au sein de la société, chaque individu perçoit un salaire différent selon différents critères. Cependant, le revenu perçu permet de distinguer les individus et engendre une hiérarchisation de la société par le biais d'une stratification sociale. Cependant, toutes ces différences sont fortement corrélées.

Différents sociologues ont étudié le phénomène de classes sociales, tel que Karl Marx[5], qui pour lui les sociétés capitalistes sont divisées en classes sociales, en fonction d'un critère économique : la position des individus dans le processus de production. Si l'individu possède du

capital, il fait partie de la bourgeoisie, mais s'il ne possède que sa force de travail, il fait partie du prolétariat. Pour lui, comme la position dans le processus de production d'un individu détermine sa classe sociale, alors le revenu est le facteur principal de la classe sociale. Cependant, selon Max Weber, la classe sociale dépend du niveau de revenu selon qu'il vient du travail ou du patrimoine.

La société est donc stratifiée dans différentes classes sociales, chaque famille appartient à une classe sociale distincte. Cependant, il se peut que la classe sociale évolue au sein d'une famille, ou que l'individu change de classe sociale au cours de sa vie. On appelle ce phénomène la mobilité sociale. Cependant, il existe différents déterminants qui poussent les individus à garder leur statut ou à évoluer.

Les ménages ont des comportements différents selon leur origine sociale, qui s'explique par la volonté de l'élévation sociale ou du maintien de celle-ci grâce au revenu perçu. Les individus peuvent donc avoir une classe sociale qui diffère de leur origine, ou bien, ils peuvent évoluer au cours de leur vie. On appelle ce phénomène la mobilité sociale. Cette mobilité sociale est causée par différents déterminants comme la famille ou encore l'école, ces déterminants sont donc des facteurs du revenu.

L'école permet aux individus d'accéder à des qualifications favorisant leur ascension sociale, ou d'obtenir un diplôme qui facilite l'accès à un emploi stable, et donc à un revenu plus ou moins élevé. C'est est un lieu d'une certaine mixité sociale qui favorise les socialisations anticipatrices, les individus peuvent adopter les aspirations d'un groupe de référence différent de son groupe d'appartenance, voir même y nouer des relations qui permettent sa mobilité sociale. Les individus d'une classe inférieure qui tissent des liens avec des camarades issus de la classe supérieure, pourront avoir une ascension sociale grâce à l'entraide et l'encouragement. Néanmois, au sein de l'école, il peut y avoir une forme de discrimination, certains élèves préfèrent rester avec d'autres élèves qui leur ressemblent, dû à leur origine sociale, qui se distingue par les vêtements, la culture, etc. Pierre Bourdieu[2], par exemple, a observé qu'un enfant issu de la grande bourgeoisie a plus de chances de faire des études longues et prestigieuses qu'un enfant issu des classes populaires qui se retrouvera plus fréquemment dans des filières modestes, comme des bacs professionnels. C'est un phénomène d'auto-sélection. Pour lui, l'école et les enseignants sont responsables, ils auraient tendance à valoriser, par leurs appréciations et notations, les élèves les mieux dotés en capital culturel, soit les élèves provenant de classe supérieure. L'école permet d'accéder à de hauts postes dans la hiérarchie, cependant, l'origine sociale influence fortement le niveau d'éducation qu'auront les individus. À la fin des études, il y aura une proportion plus élevée d'étudiants issus de hautes classes, que d'étudiants issus de familles dites du prolétariat selon Marx.

La famille est un autre déterminant de la mobilité sociale, qui permet à un individu issu d'une classe inférieure d'accéder à un haut statut dans la hiérarchie sociale. La famille influence de plu-

sieurs façons la mobilité des individus. Eric Maurin[6], montre que plus le nombre d'enfants par chambre est élevé, plus la réussite scolaire est contrariée. Les familles peuvent avoir des comportements stratégiques : le choix du lieu d'habitation, de l'établissement fréquenté, des cours particuliers, etc. La famille est centrale, car elle transmet le capital culturel et économique, et participe aux stratégies scolaires de l'enfant. Mais les familles peuvent freiner la mobilité sociale. Monique Pinçon-Charlot et Michel Pinçon[8] analysent les stratégies établies par les familles de la grande bourgeoisie pour conserver leur "entre-soi", notamment la sélection des lieux et personnes que fréquentent leurs enfants, et la mobilisation de leur capital social. Par exemple, les familles aisées auront tendance à inscrire leur enfant dans des écoles privées. Raymond Boudon[4] montre qu'à résultats scolaires équivalents, les familles populaires attendent moins de leurs enfants que celles les mieux dotées. Cependant, la pression familiale est très importante dans les familles de classe supérieure, par exemple, les familles riches attendent plus de leurs enfants donc ils seront amenés à entrer dans de grandes écoles et à accéder à de hauts postes. Il y a également la possibilité qu'un individu issu d'une classe inférieure renonce à une promotion sociale afin de ne pas perdre l'affection des proches qu'il laisserait derrière lui, dans le cas où il serait amené à faire des études loin du domicile familial. La famille permet d'aider l'individu à accéder à de hauts postes dans la hiérarchie, mais peut avoir un effet différent selon l'origine sociale.

L'origine sociale est facteur du revenu, mais différents déterminants influencent également l'ascension sociale. Néanmoins, l'origine sociale peut créer des discriminations entre les individus et certains peuvent faire face à plus de difficultés que d'autres, et peuvent renoncer à leur élévation ainsi qu'à une rémunération importante. En outre, d'autres discriminations ont lieu directement sur le marché du travail qui sont un frein à certains individus, tel que l'origine ethnique et le genre.

2.1.3 Le plafond de verre, discrimination sur le marché du travail

Il y a la présence de difficultés d'accès à certains emplois, selon le profil des individus, ce qui engendrerait souvent les mêmes types de profils qui perçoivent une rémunération élevé. Ce phénomène est appelé le plafond de verre. Le "plafond de verre" désigne les inégalités entre les hommes et les femmes dans les organisations et dans les professions, notamment sur l'accès aux fonctions supérieures et postes de pouvoir. Le "plafond de verre" concerne également les minorités ethniques, qui représente une barrière inobservable mais tellement importante qui empêche les femmes et les individus d'une origine ethnique autre que américaine d'accéder à la hiérarchie managériale, malgré la qualification de ces personnes.

Discrimination ethnique aux Etats-Unis

Les discriminations ethniques sont encore présentes dans le monde. Des politiques d'équité en matière d'emploi ont été mises en place, notamment aux États-Unis, pour lutter contre ce phénomène. Mais face à des candidats de même qualification, il est difficile de dire si les em-

ployeurs ont tendance à favoriser les ”blancs” par rapport aux afro-américains. Il est possible que les employeurs ont des préjugés liés à l'ethnie de la personne sur son niveau de productivité, qui engendre la discrimination. Il se peut, également, que la discrimination n'ait pas lieu sur le marché du travail, car elle aurait disparu avec le temps grâce à des politiques comme la mise en place de la discrimination positive ou encore par le fait que les entreprises cherchent simplement à maximiser leur profit.

Cependant, malgré les avancées en matière de discrimination, les inégalités sont toujours présentes sur le marché du travail aux Etats-Unis. Deux chercheurs américains de l'université de Chicago, Marianne Bertrand et Sendhil Mullainathan, ont fait une étude sur le niveau de discrimination ethnique[1]. Ils ont envoyé plus de 5.000 CV à 1.300 offres d'emploi dans différents domaines. Les CV présentent des profils différents, notamment sur l'ethnie qui se caractérise par le choix du prénom et du nom, mais aussi sur la qualité. Pour chaque offre d'emploi, ils ont envoyé quatre CV de bonnes qualités et d'autres de mauvaise qualité. À la suite des retours fait par les recruteurs, les deux chercheurs ont classé chaque CV selon sa capacité à engendrer un appel téléphonique ou un email.

Les résultats de leur étude ont montré que l'origine ethnique reste toujours un facteur de discrimination sur le marché du travail. Les individus d'origines afro-américains, ont révélé une influence négative sur leurs chances d'être embauchés. Mais les CV dont le prénom et le nom étaient à consonance anglo-saxonne, ont eu des retours de propositions d'entretien deux fois plus élevés. Ils ont également analysé où se situent les entreprises ayant répondu, et ils se sont aperçus que les entreprises se situant dans les quartiers qui regroupent un fort taux de personnes d'ethnie afro-américaine sont beaucoup moins discriminantes. Les chercheurs ont conclu de leur étude qu'une personne ayant un prénom et nom anglo-saxons équivalait à huit ans d'expérience en plus qu'une personne afro-américaine.

Les chercheurs ont également analysé les retours des CV ayant un haut niveau de qualification. Ils ont constaté qu'une personne afro-américaine, quel que soit son niveau de qualification, ne reçoit pas un nombre supplémentaire significatif de retour comparé aux CV à consonance anglo-saxons, qui ont une augmentation de 30% de retour pour un haut niveau de qualification. De plus, les personnes habitant dans des quartiers à la population plus riche, donc d'une certaine classe sociale, ont plus majoritairement un plus haut taux de retour, que les autres personnes pour un même niveau de qualification.

L'origine ethnique peut être un frein d'accès à certains emploi hautement rémunérateur, sous la forme de discrimination. Il existe d'autres discriminations sur le marché, tel que l'inégalité de genre.

Discrimination de genre

La place des femmes sur le marché du travail aux États-Unis révèle un problème fondamental. Ces dernières années, les États-Unis mènent un combat pour l'égalité des femmes avec les hommes sur le marché du travail. Cela est rendu grâce aux mouvements féministes et de l'avancement de la demande de main d'œuvre féminine pour permettre d'augmenter la place de la femme dans certains postes. Cependant, la discrimination de genre ainsi que les inégalités de revenus ont fortement diminué mais reste toujours présente dans certains emplois.

Selon une sociologue américaine, Ruth Milkman[7], les systèmes de services sociaux américains sont un obstacle aux femmes à l'entrée du marché du travail, notamment en termes de congés parentaux. Certaines entreprises ont plus de difficultés à recruter des femmes que des hommes, car les femmes auraient plus tendance à s'absenter pour répondre aux besoins de leurs enfants et donc être moins productives qu'un homme. Les femmes qui travaillent subissent donc une sanction salariale à chaque fois qu'elles ont un enfant, ce qui ne fait qu'augmenter l'écart de revenus entre les deux sexes. Les femmes ont plus tendance à occuper un emploi d'un niveau inférieur auquel elle serait capable d'accéder pour consacrer plus de temps à leur vie de famille, et donc percevoir un revenu moindre.

Cette brève revue de la littérature, nous a permis d'identifier un ensemble de facteurs impactant le revenu. En effet, on constate qu'il existe des déterminants fondamentaux du niveau de richesse, soit la classe de travail, l'éducation, la famille, l'ethnie, ainsi que le genre. Ainsi, la littérature, nous a permis de trouver un cadre théorique et de préciser les variables que nous allons analyser par la revue empirique.

2.2 Revue empirique : Étude du revenu médian aux États-Unis

Pour analyser les différents niveaux de revenu, nous pouvons prendre en compte le revenu médian, qui est le montant de revenu qui divise une population en deux groupes égaux, la partie inférieure et supérieure au revenu médian. Cette méthode permet d'observer si l'individu perçoit un salaire supérieur ou inférieur par rapport aux autres.

Pour anticiper les effets attendus de notre modèle économétrique, on peut analyser une étude réalisée par Census Bureau[3], qui est une administration publique américaine qui dépend du département du Commerce des États-Unis. L'étude a été réalisée à partir du recensement du revenu des ménages en 1994. Nous pouvons analyser le tableau de "comparaison des mesures sommaires du revenu selon certaines caractéristiques aux États-Unis en 1994".

TABLE 1 – Income Summary Measures By Selected Characteristics in USA in 1994, Census Bureau

Characteristics	1994 Median Income ($) USA
HOUSEHOLDS	
All households	**32.264**
Region :	
Northeast	34.926
Midwest	32.505
South	30.021
West	34.452
Race and Hispanic origin of householder :	
White	34.028
White, not Hispanic	35.126
Black	21.027
Other races	32.283
Asian and Pacific Islander	40.482
Hispanic origin	23.421
Age of Householder :	
15 to 24 years	19.34
25 to 34 years	33.151
35 to 44 years	41.667
45 to 54 years	47.261
55 to 64 years	35.232
65 years and over	18.095
Type of Household :	
Family households	39.39
Married-couple familie	45.041
Male householder, no wife present	30.472
Female householder, no husband present	19.872
Nonfamily households	18.947
Male householder	24.593
Female householder	14.948
EARNINGS OF YEAR-ROUND, FULL-TIME WORKERS	
Male	30.854
Female	22.205
PER CAPITA INCOME	
All races	16.555
White	17.611
Black	10.65
Asian and Pacific Islander	16.902
Hispanic origin	9.435

Le revenu médian aux États-Unis, en 1994, s'élève à 32 264 dollars. Pour qu'une variable soit un facteur pour percevoir un revenu élevé, il faut que le revenu médian de cette variable soit plus élevé que le revenu médian global des États-Unis.

Parmi les différents types de ménage qui constituent une famille, ce sont les ménages mariés qui perçoivent un revenu médian supérieur aux autres. Cependant, les hommes au foyer sans femme ont un salaire médian supérieur au salaire médian américain. Les femmes au foyer sans mari reçoivent un salaire médian inférieur aux hommes au foyer et inférieur au revenu médian américain. Nous pouvons déjà constater que les hommes ont un niveau de richesse supérieur aux femmes. Cet effet peut être causé par une inégalité de salaire sur les mêmes types de postes ou

bien par le fait que les hommes occupent plus des emplois à responsabilités plus élevés que les femmes. Pour les types de ménages sans famille que ce soit des femmes ou des hommes, le revenu médian est inférieur au revenu médian américain, cela peut s'expliquer par le fait qu'un individu vivant seul n'aura pas les mêmes dépenses qu'une mère ou qu'un père de famille et donc il y a moins d'attente sur le salaire. En outre, parmi les individus, à temps plein, de travail, ce sont les hommes qui ont un revenu médian supérieur à celui des femmes, ce qui confirme l'idée d'inégalité d'occupation des postes à haut niveau.

Concernant l'âge de l'agent économique, on peut observer que le type d'individu qui perçoit un plus haut revenu est la tranche d'âge de 45 à 54 ans. Ce phénomène s'explique par le fait qu'à cette catégorie d'âge, l'individu a pu gravir les différents échelons hiérarchiques et donc il y a plus d'individus qui accèdent à des hauts postes. Cependant, après cet âge, le revenu médian est inférieur, cela est dû au fait qu'après un potentiel licenciement, l'approche de l'âge en retraite, ou bien la maladie, l'individu va privilégier des emplois moins rémunérateurs qui demandent moins de temps à consacrer au travail.

De plus, le revenu médian selon l'ethnie est également variable. On peut s'apercevoir que les personnes blanches ont un revenu médian supérieur aux autres types, soit 17 611 dollars contre 10 650 dollars contre les personnes de couleurs noires.

Désormais, nous allons étudier une base de données regroupant ces différentes variables, afin d'étudier le revenu perçu par les différents agents économiques.

3 Présentation des données et méthodologie

3.1 La source et description des données

Pour effectuer notre étude, nous avons utilisé une base de données nommée “Adult Data Set” issus du site web UCI Machine Learning, ces données ont été recensées grâce à l’institut Census Bureau en 1994. Il s’agit d’une base de données multivariées, contenant 15 attributs, dont 6 attributs quantitatifs et 9 attributs catégoriels. De plus, ce jeu de données contient 32561 observations, et 4262 valeurs manquantes représenté par des “ ?”.

Dans notre tableau de données on retrouve les variables quantitatives suivantes :

— **age**
— **fnlwgt** : qui représente le nombre de fois que le type de profil ayant chacun d’eux les mêmes caractéristique apparaît (le poids)
— **education_num**
— **capital_gain**
— **capital_loss**
— **hours_per_week** : le nombre d’heures travaillées par semaine.

On y retrouve également les variables qualitatives suivante :

— **workclass** : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
— **education** : Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
— **marital_status** : Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
— **occupation** : Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
— **relationship** : Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
— **race** : White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
— **sex** : Female, Male.
— **native_country** : United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad, Tobago, Peru, Hong, Holand-Netherlands.
— **income** : $> 50K$, $\leq 50K$

Pour mener notre étude, nous avons effectué des modifications sur la base de données.

Pour commencer, nous avons ajouté les noms de chaque colonne correspondante à chaque variable de notre jeu de données. De plus, les données manquantes sont représentées sous R par ”NA”, nous avons donc remplacé tous les $\ll ? \gg$ de notre base de données par des NA.

Nous avons également modifié certaines de nos variables. Nous avons commencé par recodé notre variable expliquée “income” en une variable binaire ”income_r”, en remplaçant tous les “$> 50K$” par 1 et tous les “$\leq 50K$” par 0. Nous avons également recodé notre variable “marital-status” en une variable “marital_status_r”, en remplaçant tous les “Married-civ-spouse”, “Married-spouse-absent”, “Married-AF-spouse” en “Married” afin d'obtenir seulement l'effet global des personnes mariées sur la probabilité de percevoir un revenu élevé. En outre, la variable ”native_coutry” regroupe les pays d'origine de chaque individu, cependant, il y a une trop grandes proportions d'individus d'origine américaine comparé autres que l'on a regroupé en une modalité ”other”. Pour finir, nous avons recodé la variable “hours_per_week” en une variable binaire “hours_per_week_r”, en remplaçant les heures de travail par semaine inférieur ou égal à 40 par “NormalWorkLoad” et toutes les heures supérieur à 40 par “HugeWorkLoad”.

Nous avons donc 3 variables supplémentaires :
— **income_r** : 1, 0
— **marital_status_r** : Married, Divorced, Never-married, Separated, Widowed
— **hours_per_week_r** : NormalWorkLoad, HugeWorkLoad

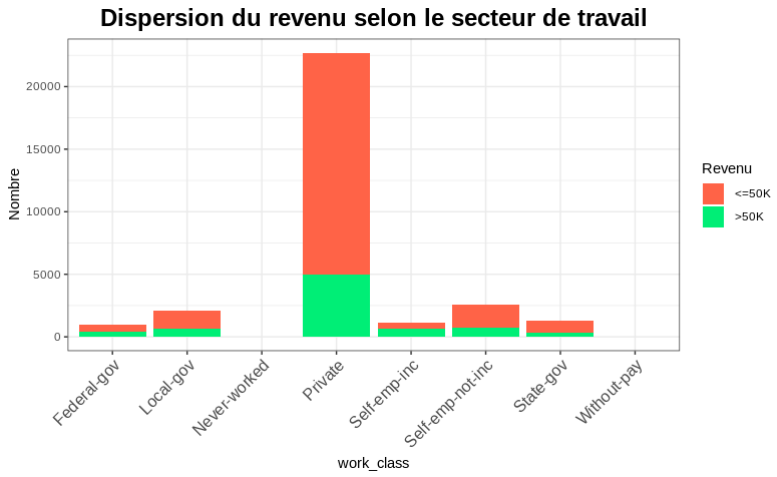
Désormais, nous pouvons analyser les différentes variables.

3.2 Statistiques descriptives et relations

3.2.1 Statistiques descriptive

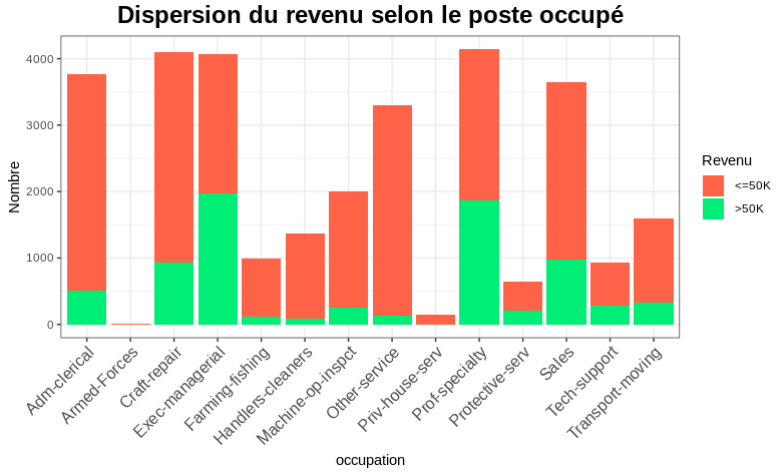
Il est important de notifier, dans un premier temps, que l'analyse descriptive concerne uniquement nos données. Cela peut, dans certains cas, ne pas refléter la réalité. Afin d'analyser et d'avoir une vue d'ensemble de nos données, nous utiliserons des diagrammes en barres à l'aide de la géométrie geom_bar de la fonction ggplot sur R. De plus, nous considérerons par la suite que les personnes ayant un revenu supérieur à 50.000 dollars par an sont des personnes ayant un “haut revenu”.

Dispersion du revenu selon le secteur de travail



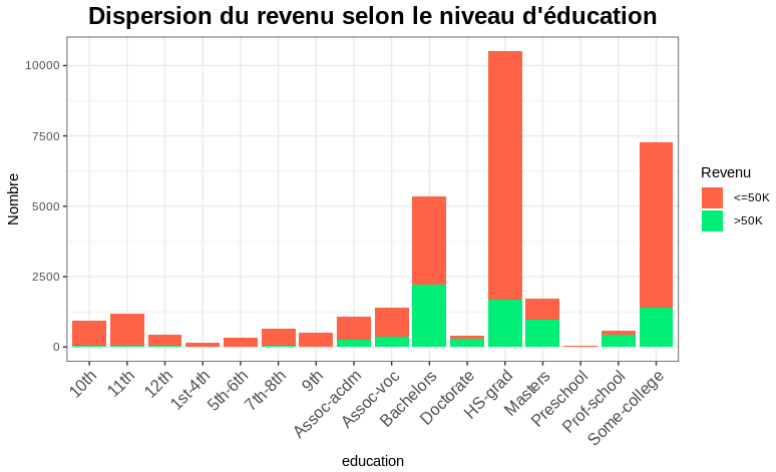
On observe que c'est le secteur privé, qui emploie le plus de personnes, et qui compte le plus grand nombre de personnes ayant un revenu supérieur à 50.000 dollars par an.

Dispersion du revenu selon le poste occupé



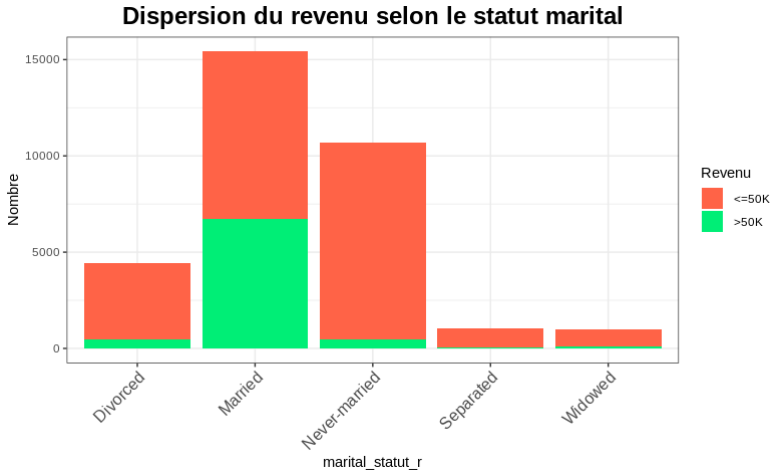
Ce graphique permet d'observer que les postes comptant le plus de personnes percevant un haut revenu sont les managers et les professeurs.

Dispersion du revenu selon le niveau d'éducation



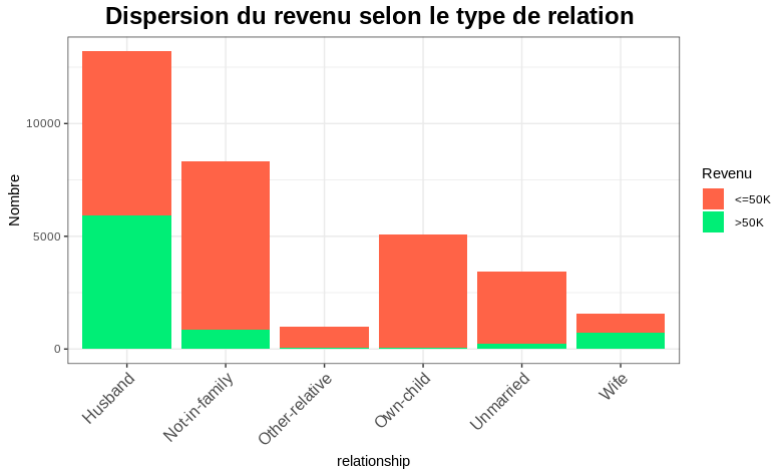
On observe que la majorité des individus de notre base de données ont un niveau d'éducation "High school Graduation". En majorité, les personnes qui perçoivent un revenu supérieur à 50.000 dollars par an, ont un niveau d'éducation type Bachelor, High school graduation, some-college et Masters. On peut donc s'apercevoir qu'un niveau d'éducation élevé permet d'augmenter la probabilité d'avoir un revenu élevé.

Dispersion du revenu selon le statut marital



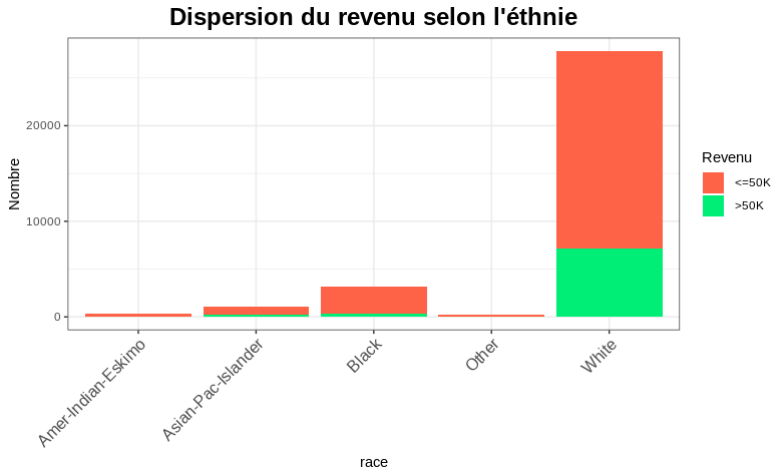
Le graphique nous indique que le mariage semble être une clé pour obtenir un haut revenu. En effet, les mariés ont la plus grande proportion de personnes percevant un revenu de plus de 50.000 dollars par an.

Dispersion du revenu selon le type de relation



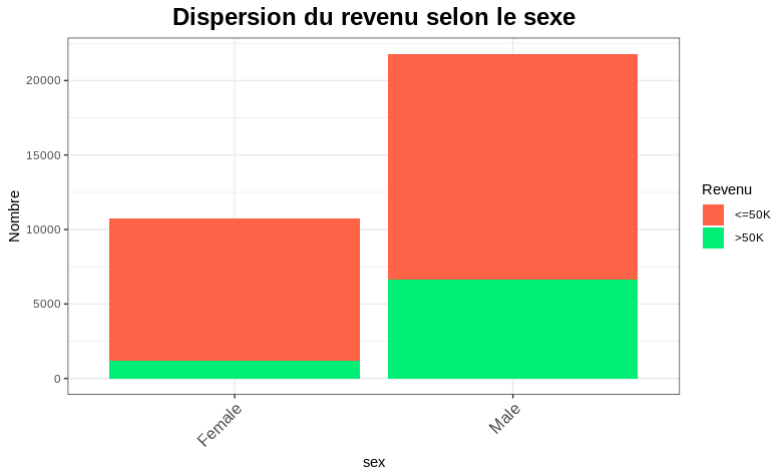
Ce graphique nous indique, comme précédent, que les personnes mariées sont les plus nombreux a percevoir un revenu supérieur à 50.000 dollars par an. On peut donc s'apercevoir que les variables "marital_statut_r" et "relationship" sont très semblable.

Dispersion du revenu selon l'ethnie



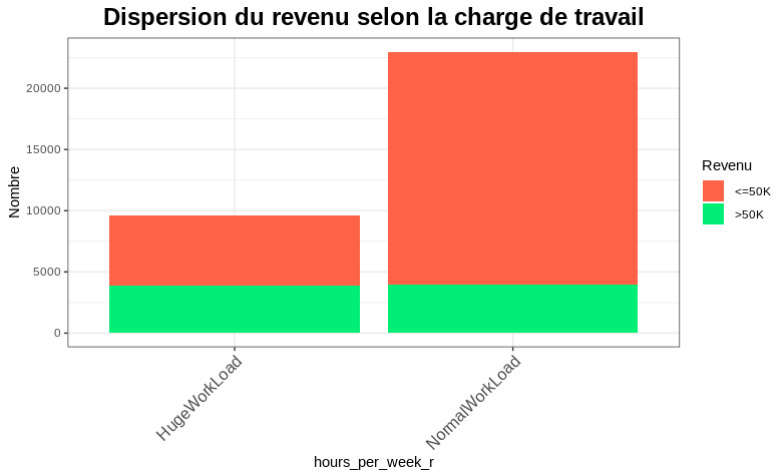
Nous pouvons voir que la modalité "white" a le plus grande nombre d'individus possédant un revenu supérieur à 50.000 dollars par an. Cependant, nous remarquons immédiatement que nos données sont en grande partie accès sur les blancs, avec une faible représentation des autres ethnies.

Dispersion du revenu selon le sexe



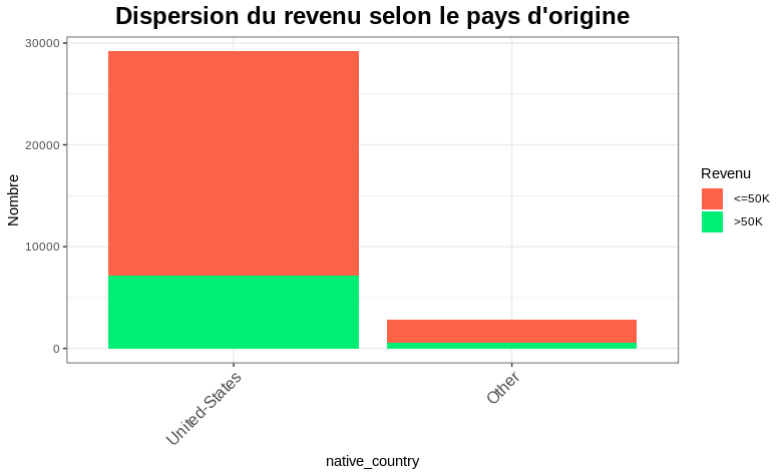
On observe que les personnes qui perçoivent un revenu supérieur à 50.000 dollars par an, sont principalement des hommes. La proportion des hommes qui perçoivent un revenu supérieur à 50.000 dollars est plus élevée que celle des femmes.

Dispersion du revenu selon la charge de travail



La proportion de revenus supérieure à 50.000 dollars par an est plus importante chez les individus qui ont une charge de travail élevée que chez ceux qui ont une charge de travail ”normale”. La charge de travail semble donc être un facteur qui influence la probabilité d'obtenir un haut revenu.

Dispersion du revenu selon l'origine sociale



Nous pouvons observer que cette variable n'est pas très pertinente à analyser. On peut voir que la majorité de nos données sont des individus d'origine américaine, et qu'il y a une très faible représentation des autres origines.

3.2.2 Relation entre variables explicatives et variable d'intérêt, le revenu

Variable d'intérêt et variables catégorielles

Dans le but d'analyser la relation entre notre variable d'intérêt et nos variables explicatives qualitatives, nous avons utilisé un test de χ^2 pour mesurer l'intensité de la relation entre les variables et un graphique en mosaïque qui représente graphiquement le tableau de contingence entre les deux variables. Le graphique en mosaïque nous donne une information similaire aux diagrammes précédents mais en raisonnant cette fois-ci en termes de proportions.

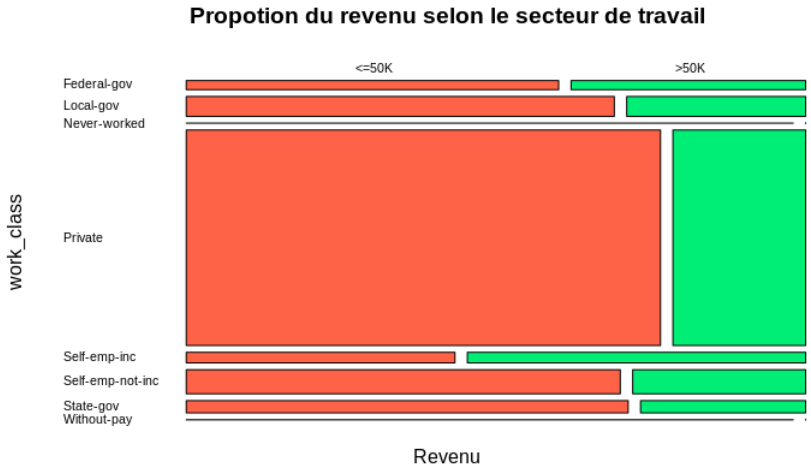
Revenu et secteur travail

Les résultats du test de Khi-deux réalisé entre le revenu et le secteur de travail se présente comme suit :

```
        Pearson's Chi-squared test

data:  df$work_class and df$income_r
X-squared = 827.72, df = 7, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et le secteur de travail indiquent une forte relation entre ces deux variables. En effet, la p-value associée à la statistique de test $(2.2e\text{-}16)$ est significative au seuil de 1% et justifie cette relation.



L'observation du graphique en mosaïque fait ressortir que le secteur privé est le plus représenté avec une proportion très élevée dans la population. Cependant, au sein des différents secteurs, celui avec une plus forte proportion d'individus ayant un revenu supérieur à 50.000 dollars par an est le secteur des entrepreneurs suivi du secteur gouvernemental.

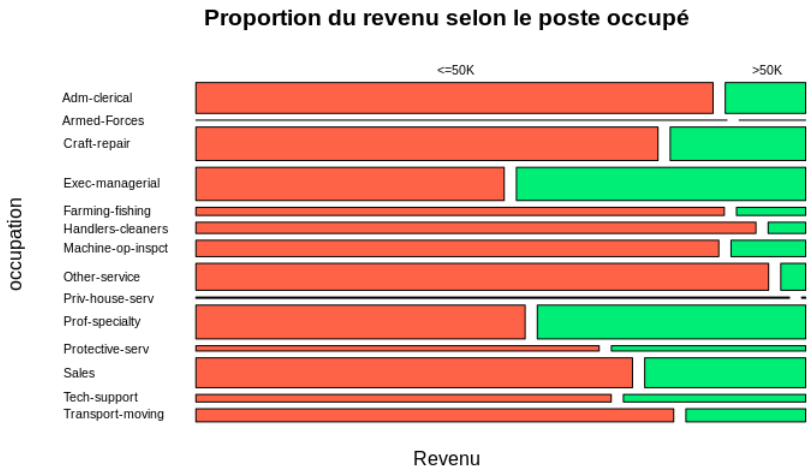
Revenu et poste occupé

Les résultats du test de χ^2 réalisé entre le revenu et le poste occupé se présente comme suit :

```
        Pearson's Chi-squared test

data:  df$occupation and df$income_r
X-squared = 3744.9, df = 13, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et le poste occupé indiquent une forte relation entre ces deux variables. En effet la p-value associée à la statistique de test ($2.2e$-16) est significative au seuil de 1% et justifie cette relation.



Ce graphique nous indique la proportion de personnes percevant un haut revenu selon le poste occupé. Nous remarquons qu'environ la moitié des individus occupant un poste de manageur et un poste de professeur ont un revenu supérieur à 50.000 dollars par an.

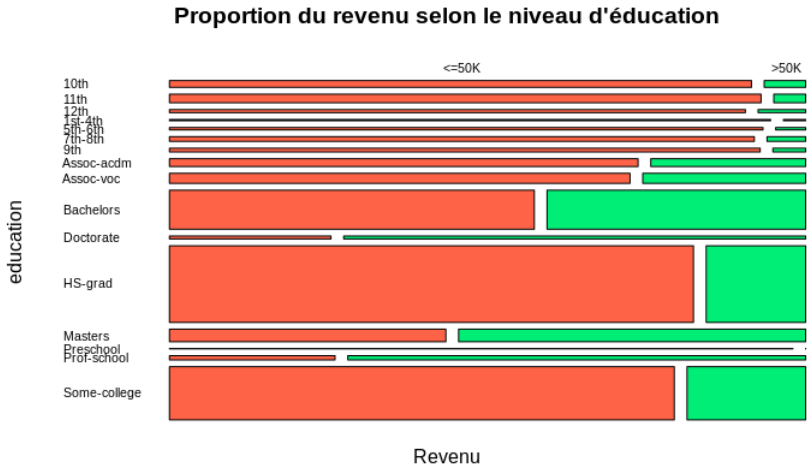
Revenu et niveau d'éducation

Les résultats du test de χ^2 réalisé entre le revenu et le niveau d'éducation se présente comme suit :

```
        Pearson's Chi-squared test

data:  df$education and df$income_r
X-squared = 4429.7, df = 15, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et le niveau d'éducation montre que le revenu est fortement lié au niveau d'éducation. En effet la p-value associée à la statistique de test ($2.2e$-16) est significative au seuil de 1% et justifie cette relation.



L'observation du graphique fait ressortir que le niveau High School grade (Baccalauréat) est le plus représenté dans la population. En revanche, la proportion d'individus ayant un revenu supérieur à 50.000 dollars par an est plus élevé au niveau des individus ayant les grades de doctorat, professeur d'université et master. Ce résultat nous indique une relation positive entre le niveau d'éducation et le fait d'avoir un revenu supérieur à 50.000 dollars par an.

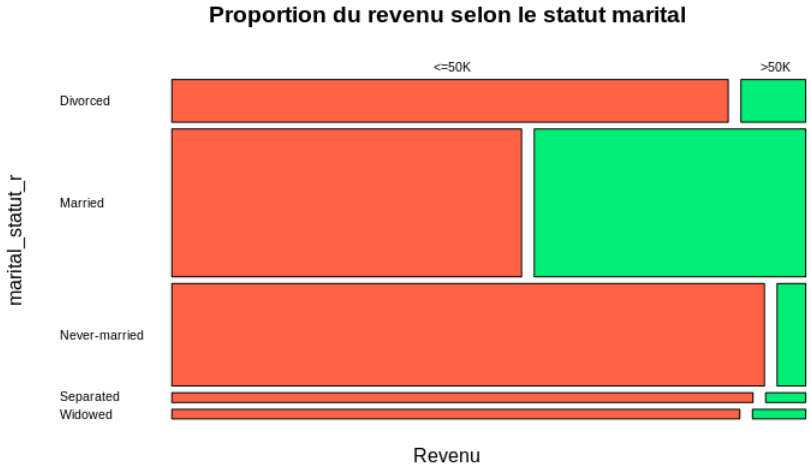
Revenu et statut marital

Les résultats du test de χ^2 réalisé entre le revenu et le statut marital se présente comme suit :

```
        Pearson's Chi-squared test

data:  df$marital_statut_r and df$income_r
X-squared = 6220.6, df = 4, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et le statut marital montre que le revenu est fortement lié au statut marital. En effet la p-value associée à la statistique de test ($2.2e$-16) est significative au seuil de 1% et justifie cette relation.



L'observation du graphique fait ressortir que le statut de marié est le plus représenté dans la population suivie par le statut de jamais marié. On note également que la proportion d'individus ayant un revenu supérieur à 50.000 dollars par an est beaucoup plus élevée au niveau des individus mariés qu'au niveau de tous les autres statuts. Le fait d'être marié est donc positivement lié au fait d'avoir un revenu supérieur à 50.000 dollars par an.

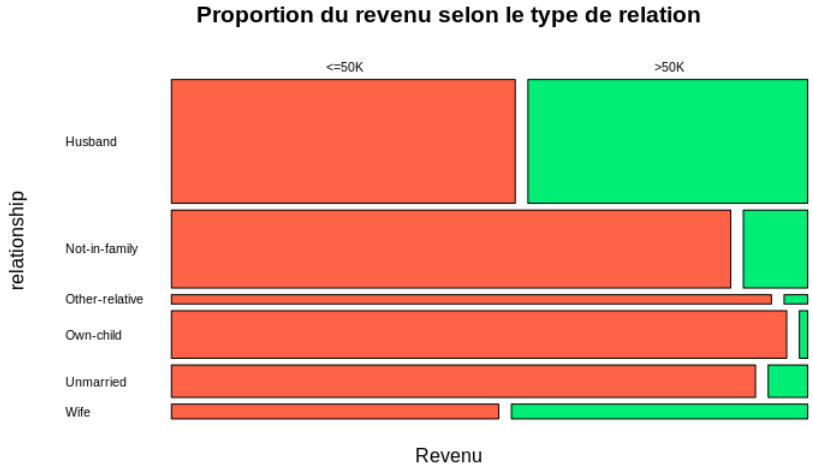
Revenu et type de relation

Les résultats du test de χ^2 réalisé entre le revenu et le type de relation se présente comme suit :

```
        Pearson's Chi-squared test

data:  df$relationship and df$income_r
X-squared = 6699.1, df = 5, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et le type de relation montre que le revenu est fortement lié au type de relation. En effet la p-value associée à la statistique de test ($2.2e$-16) est significative au seuil de 1% et justifie cette relation.



L'observation du graphique en mosaïque, nous indique, tout comme le graphique précédent, que la proportion de personnes percevant un haut revenu est supérieur chez les personnes mariés (”Husband and Wife”), l'hypothèse d'une relation positive entre le fait d'être marié et la probabilité de percevoir un haut revenu semble exister.

Revenu et ethnie

Les résultats du test de χ^2 réalisé entre le revenu et l'ethnie se présente comme suit :

```
        Pearson's Chi-squared test

data:  df$race and df$income_r
X-squared = 330.92, df = 4, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et la race indiquent une forte relation entre le revenu et la race. En effet la p-value associée à la statistique de test $(2.2e\text{-}16)$ est significative au seuil de 1% et justifie cette relation.

L'observation du graphique en mosaïque fait ressortir que l'ethnie la plus représentée au sein de la population est celle des blancs suivis de celle des noirs. Cependant on observe que la proportion d'individus ayant un revenu supérieur à 50.000 dollars par an est plus élevée au niveau des individus de d'ethnie Asian-Pac-Islander (API) que chez ceux d'ethnie blanche. Cependant, cette ethnie est faiblement représentée, il est donc difficile de dire si cet effet reflète la réalité.

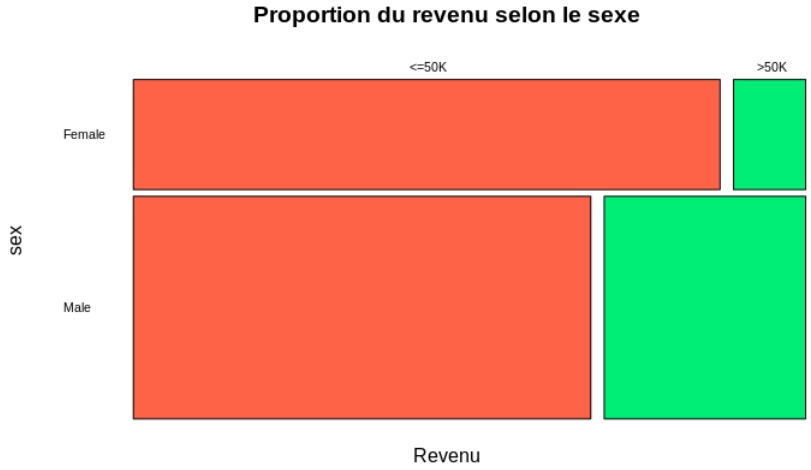
Revenu et sexe

Les résultats du test de χ^2 réalisé entre le revenu et le sexe se présente comme suit :

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df$sex and df$income_r
X-squared = 1517.8, df = 1, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et le sexe indique une forte relation entre le revenu et le sexe. En effet la p-value associée à la statistique de test $(2.2e\text{-}16)$ est significative au seuil de 1% et justifie cette relation.



L'observation du graphique en mosaïque fait ressortir que la proportion des hommes dans la population est beaucoup plus élevée que celle des femmes. On note également que la proportion des hommes ayant un revenu supérieur à 50.000 dollars par an au sein des hommes est beaucoup plus élevée que celle des femmes ayant un revenu supérieur à 50.000 dollars pas an au sein des femmes. Ce résultat justifie la relation positive entre le fait d'avoir un revenu supérieur à 50.000 dollars par an et le fait d'être un homme par rapport à une femme.

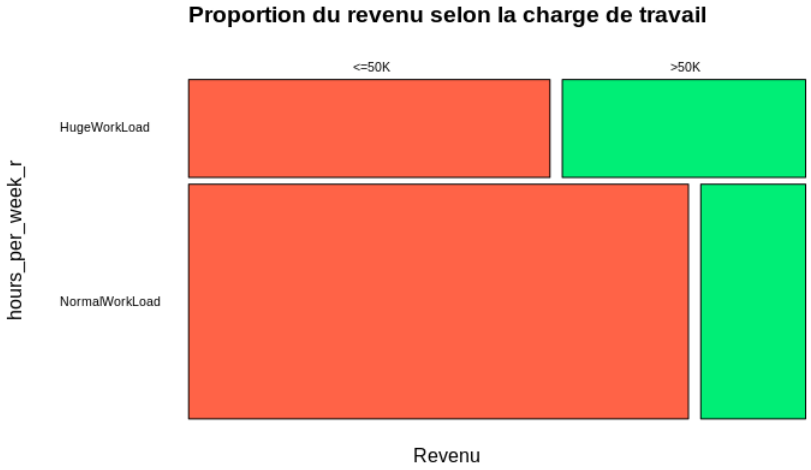
Revenu et charge de travail

Les résultats du test de χ^2 réalisé entre le revenu et la charge de travail se présente comme suit :

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df$hours_per_week_r and df$income_r
X-squared = 1939.2, df = 1, p-value < 2.2e-16
```

Les résultats du test de χ^2 entre le revenu et la charge de travail indiquent une forte relation entre ces deux variables. En effet la p-value associée à la statistique de test ($2.2e$-16) est significative au seuil de 1% et justifie cette relation.



L'observation du graphique en mosaïque fait ressortir que la proportion des individus avec une charge de travail normale dans la population est beaucoup plus élevée que celle des individus avec une charge de travail rude. En revanche, la proportion des individus ayant un revenu supérieur à 50.000 dollars par an est beaucoup plus élevée du côté de ceux ayant une rude charge de travail par rapport à ceux qui ont une charge de travail normale. Ce résultat justifie la relation positive entre le fait de travailler dur et le fait d'avoir un revenu supérieur à 50.000 dollars par an.

Revenu et pays d'origine

Les résultats du test de χ^2 réalisé entre le revenu et le pays d'origine se présente comme suit :

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df$native_country and df$income_r
X-squared = 48.845, df = 1, p-value = 2.771e-12
```

Les résultats du test de χ^2 entre le revenu et la charge de travail indiquent une forte relation entre ces deux variables. En effet la p-value associée à la statistique de test ($2.2e$-16) est significative au seuil de 1% et justifie cette relation.

Le graphique en mosaïque nous montre, comme précédemment notifié, que le pays d'origine « United-states » est majoritairement représenté. En effet, 95% des individus de notre base de données ont comme pays d'origine les États-Unis. Il est donc difficile de déduire l'effet du pays d'origine sur la probabilité de gagner plus de 50.000 dollars par an avec une telle distribution.

Résumé des tests de χ^2

Pour finir, nous avons créé un tableau récapitulatif de nos tests de χ^2 avec les valeurs exactes des p-value, afin d'avoir une vue d'ensemble et plus clair de l'intensité des relations entre les variables explicatives et la variable expliquée, le revenu.

TABLE 2 – Synthèse des tests de Khi-deux

	X-squared	df	P-value
Income/hours_per_week_r	$1,939.209$	1	$0.000000e$+00
Income/race	330.920	4	$2.305961e$-70
Income/marital_statut_r	$6,220.580$	4	$0.000000e$+00
Income/education	$4,429.653$	15	$0.000000e$+00
Income/work_class	827.718	7	$1.933848e$-174
Income/sex	$1,517.813$	1	$0.000000e$+00
Income/native_country	48.845	1	$2.770515e$-12
Income/occupation	$3,744.899$	13	$0.000000e$+00
Income/relationship	$6,699.077$	5	$0.000000e$+00

Nous pouvons donc observer que toutes les p-value associée à la statistique de test ($2.2e$-16) sont significative au seuil de 1%.

Variable d'intérêt et variables quantitatives discrètes

Dans le but d'analyser la relation entre notre variable d'intérêt et nos variables explicatives quantitatives discrètes, nous avons procédé à une analyse de la variance à un facteur (ANOVA), pour comparer les différentes moyennes d'âge relativement aux différentes modalités de notre variable d'intérêt. Nous avons également utilisé une boîte à moustache (Box-plot) pour observer graphiquement la distribution des âges au sein des deux groupes de revenus.

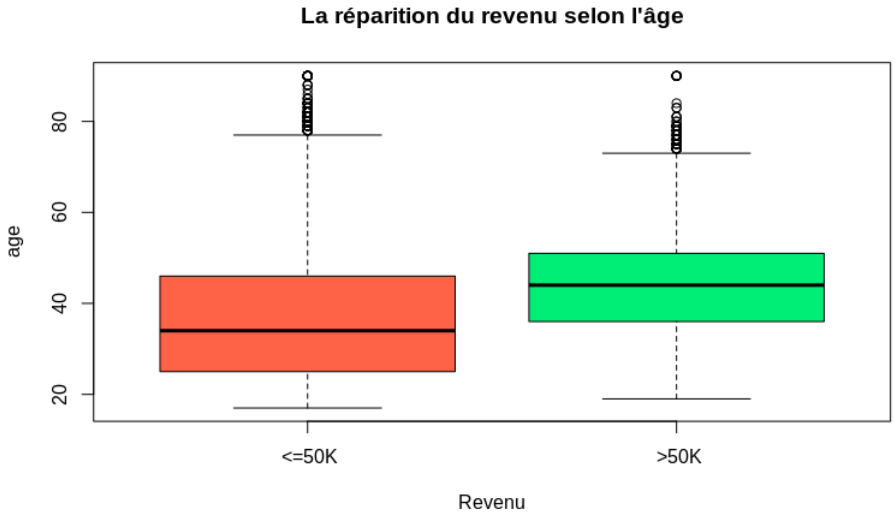
Revenu et âge

Les résultats de l'ANOVA réalisé entre l'âge et le revenu se présente comme suit :

TABLE 3 – Test ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income_r	1	331826	331826	1887	$< 2.2e\text{-}16^{***}$
Residuals	32559	5726333	176		

La p-value associée à la statistique F étant significative au seuil de 1%, on rejette l'hypothèse nulle d'égalité des moyennes d'âge relative aux groupes de revenus. Ces moyennes sont donc significativement différentes ressortant l'existence d'un lien entre l'âge et le revenu.



L'observation de la boîte à moustache permet de faire ressortir la différence de distribution des âges au sein des deux groupes de revenus. En effet l'âge médian des individus à revenu inférieur à 50.000 dollars pas an est d'environ la trentaine tandis que l'âge médian des individus à revenu supérieur à 50.000 dollars par an dépasse la quarantaine. Cela montre donc que le revenu augmente relativement avec l'âge.

Après avoir assimilé les différentes variables de notre base de données, nous pouvons maintenant passer à notre modèle économétrique afin de voir si nos estimations sont en accord avec les anticipations faites précédemment.

3.3 Variables retenues dans l'étude

Dans un premier temps, nous cherchons à définir quels facteurs permettent d'augmenter les chances de percevoir un revenu supérieur à 50.000 dollars par an. Nous sommes donc en présence d'un modèle à choix discret puisque notre variable expliqué est une variable binaire.

Pour répondre à notre problématique, nous avons exclu certaines variables afin de ne retenir que les variables souhaitées.

Pour commencer, nous avons décidé d'exclure de notre modèle des variables offrant de l'information redondante. En effet, la variable « relationship » donne une information similaire à la variable « marital status », nous avons donc choisi de garder seulement cette dernière. Nous avons ensuite exclu de notre modèle la variable « occupation » puisque nous nous intéresserons déjà à l'effet de la variable « work_class » sur la probabilité de percevoir un haut revenu. Concernant la variable « native_country », nous avons décidé de ne pas la retenir pour notre modèle, suite aux analyses descriptives faites précédemment.

Au final notre modèle spécifié s'écrit :

$$income\_r = \beta_0 + \beta_1 age + \beta_2 education + \beta_3 work\_class + \beta_4 hours\_per\_week\_r +$$

$$\beta_5 marital\_status\_r + \beta_6 sex + \beta_7 race$$

Le modèle a été défini, nous pouvons maintenant passer à notre modèle économétrique.

4 Le modèle économétrique

Nous pouvons désormais analyser les effets des variables explicatives retenus sur le revenu grâce à une régression Logistique. Nous analyserons ensuite la qualité et la pertinence de notre modèle grâce au test du rapport de vraisemblance et à la matrice de confusion. Enfin, nous pourrons observer les effets marginaux de chaque variable retenus sur le revenu, ainsi l'impact de la présence d'une telle variable par le ratio ODDS, afin d'interpréter nos résultats.

4.1 Le modèle Logit

Afin de pouvoir connaître l'impact des variables explicatives retenues dans notre modèle, sur la probabilité de percevoir un revenu supérieur à 50.000 dollars par an, nous utiliserons le modèle logit. En effet, cette méthode est idéale pour les modèles à choix binaire, elle permet de prédire la probabilité qu'un événement arrive, ou non, à partir de l'optimisation des coefficients de régression.

La probabilité de succès du modèle logistique est défini par

$$P(y_i = 1|x_i) = \Lambda(x_i'\beta) = \frac{\exp x_i'\beta}{1 + \exp x_i'\beta}$$

où la probabilité de succès est égale à la fonction de répartition de la loi logistique. Les paramètres sont estimés par la méthode du Maximum de Vraisemblance.

La régression logistique de notre modèle, réalisé sur R grâce à la fonction glm se présente comme suit :

TABLE 4 – Estimation du modèle Logit

term	estimate	std.error	statistic	p.value
(Intercept)	$-4,6463$	0.279	-16.650	$< 2e\text{-}16^{***}$
age	$0,0291$	0.002	19.224	$< 2e\text{-}16^{***}$
education 11th	$0,0842$	0.200	0.420	0.674
education 12th	$0,5174$	0.248	2.085	0.037^{*}
education 1st-4th	$-0,9873$	0.448	-2.205	0.027^{*}
education 5th-6th	$-0,6926$	0.314	-2.204	0.028^{*}
education 7th-8th	$-0,5379$	0.225	-2.395	0.017^{*}
education 9th	$-0,4143$	0.253	-1.637	0.102
education Assoc-acdm	$1,7094$	0.166	10.292	$< 2e\text{-}16^{***}$
education Assoc-voc	$1,6381$	0.160	10.258	$< 2e\text{-}16^{***}$
education Bachelors	$2,4642$	0.147	16.727	$< 2e\text{-}16^{***}$
education Doctorate	$3,6047$	0.199	18.136	$< 2e\text{-}16^{***}$
education HS-grad	$0,9483$	0.146	6.512	$< 2e\text{-}16^{***}$
education Masters	$2,9599$	0.156	18.993	$< 2e\text{-}16^{***}$
education Preschool	$-11,7190$	114.183	-0.103	0.918
education Prof-school	$3,6332$	0.185	19.644	$< 2e\text{-}16^{***}$
education Some-college	$1,4118$	0.147	9.601	$< 2e\text{-}16^{***}$
work_class Local-gov	$-0,6851$	0.102	-6.726	$1.74e\text{-}11^{***}$
work_class Never-worked	$-11,5322$	295.307	-0.039	0.969
work_class Private	$-0,6152$	0.086	-7.183	$6.81e\text{-}13^{***}$
work_class Self-emp-inc	$-0,1672$	0.112	-1.491	0.136
work_class Self-emp-not-inc	$-1,1543$	0.099	-11.626	$< 2e\text{-}16^{***}$
work_class State-gov	$-0,8388$	0.116	-7.239	$4.52e\text{-}13^{***}$
work_class Without-pay	$-14,3950$	196.466	-0.073	0.942
hours_per_week_rNormalWorkLoad	$-0,7121$	0.035	-20.447	$< 2e\text{-}16^{***}$
marital_statut_rMarried	$1,9629$	0.061	32.433	$< 2e\text{-}16^{***}$
marital_statut_rNever-married	$-0,5488$	0.075	-7.309	$2.69e\text{-}13^{***}$
marital_statut_rSeparated	$-0,1435$	0.146	-0.983	0.326
marital_statut_rWidowed	$-0,1068$	0.138	-0.773	0.440
sex Male	$0,2452$	0.046	5.341	$9.26e\text{-}08^{***}$
race Asian-Pac-Islander	$0,3109$	0.226	1.374	0.170
race Black	$0,3958$	0.217	1.828	0.068
race Other	$-0,3346$	0.328	-1.020	0.308
race White	$0,5757$	0.207	2.775	0.006^{**}
Null deviance :	34487 on 30724 degrees of freedom			
Residual deviance :	23156 on 30691 degrees of freedom			
(1836 observations deleted due to missingness)				
AIC :	23221			
Number of Fisher Scoring iteractions :	14			

Nous remarquons que 1836 observations ont été supprimées à cause des valeurs manquantes. Étant donné que notre base de données contient 32561 observations, ce nombre d'observations supprimées nous semble acceptable.

De plus, il est important de rappeler que notre modèle dispose de nombreuses variables qualitatives, de nombreuses dummy ont donc été créé pour chaque modalités de ces variables qualitatives excepté les modalités dites de référence que nous allons préciser ci dessous :

— education : modalité de référence → 10th
— work_class : modalité de référence → Federal-gov
— hours_per_week_r : modalité de référence → HugeWorkLoad
— marital_status : modalité de référence → Divorced
— sex : modalité de référence → Female
— race : modalité de référence → Amer-Indian-Eskimo

Ces modalités de référence nous serviront par la suite, afin d'interpréter nos résultats.
Nous remarquons aussi que plusieurs de ces dummy ne sont pas significatives, nous ne prendrons pas en compte ces modalités et interpréterons uniquement celles ayant une p-value supérieur à 0,05.

4.2 Modèle d'évaluation

4.2.1 Test du rapport de vraisemblance

Le test du rapport de vraisemblance permet de comparer la qualité de l'ajustement de deux modèles de régression imbriqués en fonction du rapport de leurs vraisemblances, en particulier l'un obtenu par maximisation sur l'ensemble de l'espace des paramètres et l'autre obtenu après avoir imposé une certaine contrainte.

Nous considérerons dans notre cas, le modèle imbriqué de notre modèle globale portant les contraintes :

$$hours\_per\_week = age = race = 0$$

Ce modèle s'écrit donc :

$$income\_r = \beta_0 + \beta_2 education + \beta_3 work\_class + \beta_5 marital\_status\_r + \beta_6 sex$$

Pour voir si ce modèle diffère significativement de notre modèle global, nous pouvons utiliser un test du rapport de vraisemblance avec les hypothèses nulles et alternatives suivantes :
— H_0 : Le modèle global et le modèle imbriqué s'ajustent aussi bien aux données. Par conséquent, on doit utiliser le modèle imbriqué.
— H_1 : Le modèle global surpasse de manière significative le modèle imbriqué en termes d'ajustement des données. Par conséquent, on doit utiliser le modèle global.

Le critère de décision est si la valeur p du test est inférieure au seuil de significativité de 0,05, nous pouvons rejeter l'hypothèse nulle et conclure que le modèle global fournit un ajustement significativement meilleur.

Les résultats du LR test effectué à partir des deux modèles se présente comme suit :

TABLE 5 – Test du rapport de vraisemblance

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	34	-11,578.110			
2	28	-11,979.430	-6	802.638	$< 2.2e\text{-}16^{***}$

La p-value de la statistique de test étant inférieur à 0,05, nous rejetterons l'hypothèse nulle. Notre modèle global fournit donc un ajustement significativement meilleur de nos données.

4.2.2 Matrice de confusion

La matrice de confusion est un résumé des résultats de prédiction pour un problème particulier de classification. Elle permet d'évaluer les performances d'un modèle de classification en comparant les données réelles pour une variable cible à celles prédites par un modèle. Les prédictions justes et fausses sont révélées et réparties par classe, ce qui permet de les comparer avec des valeurs définies.

La matrice de confusion associée à notre modèle se présente comme suit :

TABLE 6 – Matrice de confusion

	0	1
FALSE	21272	3699
TRUE	1803	3951

De cette table de confusion, il ressort que notre modèle prédit efficacement 21272 individus qui ont effectivement un revenu inférieur à 50.000 dollars par an et 3951 individus qui ont effectivement un revenu supérieur à 50.000 dollars par an. Par contre il prédit à tort 3699 individus ayant en réalité un revenu supérieur à 50.000 dollars par an et 1803 individus ayant en réalité un revenu inférieur à 50.000 dollars par an.

Notre modèle présente donc 5502 $(= 3699 + 1803)$ prédictions incorrectes sur un total de 30725, soit un taux de mauvais classement de $17,9\%$. Le taux de bon classement de $82,1\%$ de notre modèle témoigne donc de son fort pouvoir prédictif.

4.3 Statistiques inférentielles et interprétations

Après avoir évalué notre modèle et défini le pouvoir explicatif de ce dernier, nous pouvons maintenant nous intéresser à nos résultats.

Contrairement aux modèles de régression linéaire, les coefficients estimés de notre régression logistique ne peuvent pas être interprété comme tel. Cependant le signe de ces coefficients nous indique, avant toute chose, le sens de la relation.

Prenons par exemple la variable ”age” dont son coefficient estimé est positif (0.0291). Cela nous indique qu'un individu plus âgé ayant les mêmes caractéristiques qu'un individu plus jeune, aura en moyenne une probabilité plus élevée de percevoir un revenu supérieur à 50.000 dollars par an. Ou encore un homme, ayant les mêmes caractéristiques qu'une femme, aura en moyenne une probabilité plus élevé qu'une femme d'obtenir un haut revenu.

Afin de quantifier ces relations positives ou négatives entre nos variables explicatives et variable notre expliquée, nous avons décidé d'utiliser les ODDS ratio. L'ODDS ratio également appelé rapport des chances, permet d'obtenir : pour les variables quantitatives, l'impact sur la côte de la variable expliquée d'une augmentation de 1 unité de la variable, mais encore pour les variable catégorielle l'impact sur la côte de la variable expliquée du fait d'appartenir à la catégorie indiquée par la variable muette par rapport au fait d'appartenir à la catégorie de référence.

L'odds ratio est toujours positif, la valeur 1 sert de référence et nous indique l'absence de changement. En utilisant un modèle logit pour notre étude, l'odds ratio est obtenu en calculant l'exponentiel de nos coefficients estimés. Nous pouvons donc calculer l'odds ratio pour chacune de nos variables.

Nous pouvons interpréter l'ODDS ration à l'aide de la colonne ”estimate”. Rappelons que nous interpréterons uniquement les variables significatives ayant une P-value inférieur à 0,05. Il est maintenant simple d'avoir une interprétation quantitative de nos résultats.

TABLE 7 – ODDS Ratio

term	estimate	std.error	statistic	P-value
(Intercept)	0.010	0.279	-16.650	$< 2e\text{-}16^{***}$
age	1.030	0.002	19.224	$< 2e\text{-}16^{***}$
education 11th	1.088	0.200	0.420	0.674
education 12th	1.678	0.248	2.085	0.037^{*}
education 1st-4th	0.373	0.448	-2.205	0.027^{*}
education 5th-6th	0.500	0.314	-2.204	0.028^{*}
education 7th-8th	0.584	0.225	-2.395	0.017^{*}
education 9th	0.661	0.253	-1.637	0.102
education Assoc-acdm	5.526	0.166	10.292	$< 2e\text{-}16^{***}$
education Assoc-voc	5.145	0.160	10.258	$< 2e\text{-}16^{***}$
education Bachelors	11.754	0.147	16.727	$< 2e\text{-}16^{***}$
education Doctorate	36.772	0.199	18.136	$< 2e\text{-}16^{***}$
education HS-grad	2.581	0.146	6.512	$< 2e\text{-}16^{***}$
education Masters	19.296	0.156	18.993	$< 2e\text{-}16^{***}$
education Preschool	0.00001	114.183	-0.103	0.918
education Prof-school	37.834	0.185	19.644	$< 2e\text{-}16^{***}$
education Some-college	4.103	0.147	9.601	$< 2e\text{-}16^{***}$
work_class Local-gov	0.504	0.102	-6.726	$1.74e\text{-}11^{***}$
work_class Never-worked	0.00001	295.307	-0.039	0.969
work_class Private	0.541	0.086	-7.183	$6.81e\text{-}13^{***}$
work_class Self-emp-inc	0.846	0.112	-1.491	0.136
work_class Self-emp-not-inc	0.315	0.099	-11.626	$< 2e\text{-}16^{***}$
work_class State-gov	0.432	0.116	-7.239	$4.52e\text{-}13^{***}$
work_class Without-pay	0.00000	196.466	-0.073	0.942
hours_per_week_rNormalWorkLoad	0.491	0.035	-20.447	$< 2e\text{-}16^{***}$
marital_statut_rMarried	7.120	0.061	32.433	$< 2e\text{-}16^{***}$
marital_statut_rNever-married	0.578	0.075	-7.309	$2.69e\text{-}13^{***}$
marital_statut_rSeparated	0.866	0.146	-0.983	0.326
marital_statut_rWidowed	0.899	0.138	-0.773	0.440
sex Male	1.278	0.046	5.341	$9.26e\text{-}08^{***}$
race Asian-Pac-Islander	1.365	0.226	1.374	0.170
race Black	1.486	0.217	1.828	0.068
race Other	0.716	0.328	-1.020	0.308
race White	1.778	0.207	2.775	0.006^{**}

Commençons par notre seul variable quantitative, “age”, dont son odds ratio est d’environ 1,03. Cela signifie qu’un an supplémentaire permet d’augmenter en moyenne de 3% la probabilité d’obtenir un revenu supérieur à 50.000 dollars par an, toute chose égale par ailleurs.

Passons maintenant à l’interprétation des résultats de nos variables qualitatives. L’odds ratio de ces variables qualitative nous permet d’obtenir, l’impact sur la probabilité de percevoir un revenu supérieur à 50.000 dollars par an, d’appartenir à une certaine cette catégorie par rapport à la catégorie de référence.

Concernant la variable ”sex”, nos résultats nous permettent de déduire, qu’en moyenne un

homme a 27% de chance en plus de percevoir un revenu supérieur à 50.000 dollars par an qu'une femme, toute choses égales par ailleurs.

Pour la variable "hours_per_week", nos résultats nous indiquent que les individus travaillant 40 heures ou moins par semaine, voient leurs chances de percevoir un revenu supérieur à 50.000 dollars par an divisé par 2 par rapport aux individus travaillant plus de 40 heures par semaine, toute chose égale par ailleurs.

La variable "education" a comme référence la modalité "10th" qui correspond à la première année de lycée. Il est intéressant de notifier que la probabilité d'obtenir un haut revenu, baisse lorsque les années d'étude sont inférieures à la référence et augmente lorsque les années d'études sont supérieures à cette dernière, toute chose égale par ailleurs. Par exemple, un individu possédant un master et ayant les mêmes caractéristiques qu'un individu s'étant arrêté à la première année de lycée, a environ 19 fois plus de chances de percevoir un revenu supérieur à 50.000 dollars par an. Inversement, d'après nos résultats, un individu ayant arrêté l'école en primaire ("1st"-"4th") et ayant les mêmes caractéristiques qu'un individu s'étant arrêté à la première année de lycée, voit ses chances de percevoir un revenu supérieur à 50.000 dollars par an diminuer d'environ 63%.

Concernant la variable "work_class", nous remarquons que le fait de travailler dans le secteur privé, dans le niveau le plus bas de l'administration publique ("local-gov"), dans le gouvernement d'état ("state-gov"), ainsi que dans l'entreprenariat ("self-emp-not-inc"), fait diminuer les chances de percevoir un revenu supérieur à 50.000 dollars par par rapport aux individus travaillant dans le secteur de référence ("federal-gov"), toute chose égale par ailleurs.

Le statut marital a aussi un impact sur la probabilité d'obtenir un haut revenu, en effet nos résultats nous indiquent qu'un individu marié a environ 7 fois plus de chance de percevoir un revenu supérieur à 50.000 dollars par an qu'un individu divorcé, toutes choses égales par ailleurs.

Pour finir, concernant l'ethnie, nous remarquons, sans grande surprise au vu des discriminations présente dans le pays, qu'un individu blanc a plus de chance de percevoir un haut revenu que l'ethnie de référence ("Amer-Indian-Eskimo"), toutes choses égales par ailleurs.

A la suite de tous ces résultats, il paraît intéressant de savoir quel individu de notre base de données à la plus haute probabilité de percevoir un revenu supérieur à 50.000 dollars par an. Pour cela, nous avons calculé cette probabilité pour chaque individu de notre base de données. Il s'est avéré que la plus haute probabilité de percevoir un revenu supérieur à 50.000 dollars par an de notre échantillon est de 0,9775. Cette probabilité est détenu par un homme blanc de 90 ans, marié, ayant comme niveau d'éducation "prof-school" et travaillant plus de 40 heures par semaine dans le secteur privé. Ce résultat semble cohérent avec nos précédents résultats et analyses.

5 Conclusion

Le revenu est un indicateur de richesse. Cependant, chaque individu perçoit un revenu qui diffère selon différents facteurs. Cependant, il existe des facteurs individuels qui permettent aux individus d'accéder plus ou moins facilement à un certain niveau de richesse.

À l'aide de notre étude économétrique, on a pu connaître quelles sont les différentes caractéristiques d'une personne ayant une rémunération de plus de 50.000 dollars par an. Nous avons donc constater qu'un homme, blanc, marié, qui a fait de longues études. Cependant, dans la base de données, l'individu qui perçoit la plus grande probabilité d'avoir un revenu au-dessus de 50.000 dollars par an, est un homme âgé de 90 ans, marié, ayant comme niveau d'éducation "prof-school" et travaillant plus de 40 heures par semaine dans le secteur privé.

Cependant, nous n'avons pas pu avoir accès à l'ensemble des données issues de l'étude de Census bureau datant de 1994, ce qui aurait été plus pertinent pour compléter nos interprétations. Il aurait été intéressant d'étudier notamment la localisation de l'individu aux États-Unis. Le revenu n'est pas le même, selon le lieu géographique, car ce n'est pas le même mode de vie, d'un État à un autre et d'une ville à une autre.

On peut également se poser le problème de la transitivité : est-ce que ces facteurs engendrent un certain niveau de revenu ou est-ce le niveau de revenu qui engendre les différents phénomènes. Par exemple, est-ce la situation maritale qui engendre un revenu élevé ou bien, c'est le fait d'avoir un revenu élevé qui engendre le fait que les individus se marient. On pourrait penser que l'origine sociale de l'individu a un effet sur cette question. Un individu issu d'une haute classe sociale, aura plus tendance à se marier dû aux mœurs dans celle-ci. On ne connaît pas les antécédents des individus, on ne sait pas si les personnes qui perçoivent plus de 50.000 dollars par an viennent tous d'une famille aisée ou non. Il aurait été pertinent de montrer, selon l'origine sociale, si l'individu perpétue sa hiérarchisation sociale, donc si la classe sociale est réellement un facteur du revenu, on peut seulement le déduire sans en être sûr grâce aux données.

6 Bibliographie

Références

[1] Marianne BERTRAND et Sendhil MULLAINATHAN. ≪ Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination ≫. In : *American economic review* 94.4 (2004), p. 991-1013.

[2] P BOURDIEU et JC PASSERON. ≪ Reproduction in Education, society and culture (Londres, Beverly Hills) ≫. In : (1977).

[3] Carmen DENAVAS et al. *Income, Poverty, and Valuation of Noncash Benefits: 1994*. 1996.

[4] Jean-Claude FORQUIN. ≪ Boudon (Raymond). — L'inégalité des chances. La mobilité sociale dans les sociétés industrielles ≫. In : *Revue française de pédagogie* 32.1 (1975), p. 74-78.

[5] Karl MARX et Friedrich ENGELS. *Manifeste du parti communiste*. Le livre de poche, 2012.

[6] Eric MAURIN. *L'égalité des possibles: la nouvelle société française*. Seuil, 2002.

[7] Ruth MILKMAN et Hélène TRONC. ≪ 13. Genre et marché du travail aux États-Unis ≫. In : *Travail et genre dans le monde*. La Découverte, 2013, p. 130-139.

[8] Michel PINÇON et Monique PINÇON-CHARLOT. *Dans les beaux quartiers*. FeniXX, 1989.