UNIVERSITÉ DE RENNES 1

Discrimination salariale : Décomposition d'Oaxaca

Master 1 Mathématiques Appliquées Statistique

Projet d'économétrie

Tsang Guy

Champain Mathilde

2018-2019

Enseignante : Thelen Véronique

Table des matières

I Introduction 1

1 Présentation du sujet 1

2 Décomposition d'Oaxaca : principe et fonctionnement 3

II Préparation de la base de données 6

3 Sélection des variables pertinentes 6

4 Restrictions sur les individus retenus 7

5 Création des variables indicatrices (dummies) 8

5.1 `SEXE` : le genre de l'individu . . . 8
5.2 `CSER` : la catégorie socioprofessionnelle . . . 8
5.3 `NIVP` : le niveau d'étude . . . 9
5.4 `MATRI` : statut matrimonial légal . . . 10
5.5 `TPPRED` : temps complet ou temps partiel . . . 10
5.6 `NUITC` : travail de nuit . . . 11
5.7 `PUB3FP` : caractère public ou privé de l'entreprise . . . 11
5.8 `NAFG10N` : secteur économique de l'entreprise . . . 12
5.9 `PAYNEU27` : pays de naissance et Union Européenne . . . 12
5.10 `LNAIS` : pays de naissance et France . . . 13
5.11 `NFR` : nationalité française ou étrangère . . . 13
5.12 `TUU` : commune urbaine ou rurale . . . 13
5.13 `REG` : région de résidence . . . 14
5.14 `NBAGENF` : le nombre d'enfants à charge . . . 15
5.15 `CONTRA` : le type de contrat du poste . . . 16

6 Statistiques descriptives 17

6.1 Statistiques descriptives univariées . . . 17
6.2 Statistiques descriptives bivariées . . . 22
6.3 Mesures de corrélations et de similarités . . . 29
6.3.1 Corrélations entre variables continues . . . 29
6.3.2 Similarités entre variables dichotomiques . . . 30

III Résultats empiriques **32**

7 Modèles estimés par MCO **32**
7.1 Estimations pour les hommes 32
7.2 Estimations pour les femmes 35
7.3 Détection d'hétéroscédasticité 38

8 Estimations MCO avec correction de White **39**
8.1 Estimations pour les hommes 39
8.2 Estimations pour les femmes 45

9 Décomposition d'Oaxaca **49**
9.1 Tableau récapitulatif 49
9.2 Enseignements 52

Conclusion **53**

Annexes **I**
Annexe 1 : Calcul de similarité entre deux vecteurs binaires II
Annexe 2 : Liste des variables pertinentes de la base INSEE III
Annexe 3 : Liste des variables utilisées dans les régressions IX
Annexe 4 : Étude de référence : Oaxaca R., 1973 XIII

Bibliographie **XXXII**

* * *

Liens utiles :

données : `https://www.insee.fr/fr/statistiques/2415221`

code `R` : `https://github.com/Psqrt/projet_econometrie_m1s7`

article de référence : `https://www.jstor.org/stable/2525981`

Liste des tableaux

1 Détails des modalités de la variable CSER 8

2 Détails des modalités de la variable NIVP 9

3 Détails des modalités de la variable MATRI 10

4 Détails des modalités de la variable NUITC 11

5 Détails des modalités de la variable PUB3FP 11

6 Détails des modalités de la variable NAFG10N 12

7 Détails des modalités de la variable REG 14

8 Détails des modalités de la variable NBAGENF 15

9 Détails des modalités de la variable CONTRA 16

10 Matrice des similarités 30

11 Résultats des MCO sur l'échantillon des hommes 33

11 Résultats des MCO sur l'échantillon des hommes (suite) 34

12 Résultats des MCO sur l'échantillon des femmes 36

12 Résultats des MCO sur l'échantillon des femmes (suite) 37

13 Résultats des MCO corrigés de White sur l'échantillon des hommes 43

13 Résultats des MCO corrigés de White sur l'échantillon des hommes (suite) 44

14 Résultats des MCO corrigés de White sur l'échantillon des femmes 47

14 Résultats des MCO corrigés de White sur l'échantillon des femmes (suite) 48

15 Tableau récapitulatif de la décomposition d'Oaxaca (hommes et femmes) 50

15 Tableau récapitulatif de la décomposition d'Oaxaca (hommes et femmes) (suite) . . 51

16 Tableau de contingence et S_{ij} II

17 Liste des indicateurs de similarité II

18 Détail des variables pertinentes de la base INSEE III

19 Détail des variables utilisées dans les régressions IX

Liste des graphiques

1 Histogramme des âges des individus . . . 17
2 Histogramme des salaires mensuels . . . 17
3 Ancienneté en années . . . 18
4 Temps de travail hebdomadaire . . . 18
5 Salaire horaire . . . 18
6 Répartition des genres . . . 19
7 Répartition des domaines . . . 19
8 Répartition province/Ile-de-France . . . 19
9 Répartition Milieu urbain/rural . . . 19
10 Pays de naissance . . . 20
11 Nationalité . . . 20
12 Région de naissance . . . 20
13 Diplômes . . . 20
14 Nombre d'enfants . . . 20
15 Situation maritale . . . 20
16 Types de contrats . . . 21
17 Catégorie Socio-Professionnelle . . . 21
18 Salaire selon l'âge . . . 22
19 Salaire selon l'ancienneté . . . 22
20 Salaire horaire en logarithme selon le temps de travail . . . 22
21 Situation maritale selon le sexe . . . 23
22 Etudes selon le sexe . . . 23
23 Contrat selon le sexe . . . 23
24 CSP selon le sexe . . . 23
25 Activité de l'entreprise selon le sexe . . . 24

26 Secteur selon le sexe . . . 24
27 Âge selon le sexe . . . 25
28 Salaire selon le sexe . . . 25
29 Nombre d'heures travaillées selon le sexe . . . 26
30 Ancienneté selon le sexe . . . 26
31 Salaire selon le type de contrat . . . 27
32 Salaire selon les études effectuées . . . 27
33 Salaire selon le lieu de résidence . . . 27
34 Salaire selon le pays de naissance . . . 27
35 Salaire selon le domaine . . . 27
36 Salaire selon le lieu . . . 27
37 Salaire selon la nationalité (gauche) . . . 27
38 Salaire selon la situation maritale (droite) . . . 27
39 Salaire selon l'activité . . . 28
40 Matrice des corrélations . . . 29

Première partie

Introduction

1 Présentation du sujet

En première année de Master, dans le cadre du cours d'économétrie, un projet est proposé afin de mettre en pratique les acquis de ce cours. Le choix de sujet pour ce projet s'est tourné vers l'étude des inégalités salariales entre les hommes et les femmes.

Pour le même poste et les mêmes caractéristiques individuelles qu'un homme,
une femme perçoit-elle le même salaire ?

Si toutes les statistiques nationales démontrent une discrimination salariale envers les femmes, les chiffres avancés ne représentent pas nécessairement l'écart de salaire entre deux groupes homogènes en termes de caractéristiques individuelles. Les chiffres obtenus sans cette considération peuvent être surestimés si par exemple, les hommes ont plus tendance à occuper des postes mieux payés que les femmes. Il semble donc cohérent de chercher à quantifier la discrimination à postes égaux et à caractéristiques personnelles égales. Ronald Oaxaca [Oax73] est l'un des premiers à s'être intéressé à cette problématique et propose une méthode de décomposition en 1973. Celle-ci est très similaire à celle utilisée par Alan Blinder [Bli73] la même année pour la même problématique. Ainsi, cette méthode est très souvent appelée « Décomposition d'Oaxaca-Blinder » et peut être utilisée pour comparer deux groupes pour quelconque problématique [1].

Pour traiter ce sujet, le champ d'étude est restreint sur la population française, en 2012. Des modèles économétriques vont être construits puis estimés par la méthode des Moindres Carrés Ordinaires (MCO). Finalement, la méthode de décomposition d'Oaxaca (ou Oaxaca-Blinder)

1. Par exemple, Owen O'Donnell et *al.* (2008) l'ont utilisée pour analyser les inégalités de santé entre les pauvres et les riches

va être appliquée sur les résultats obtenus afin de conclure sur la part jouée par la discrimination sexiste sur les différences de salaire entre les hommes et les femmes.

Cette étude sera basée sur des données accessibles librement sur le site de l'Institut National de la Statistique et des Études Économiques (INSEE). Ces données concernent l'année 2012 et ont été collectées dans le cadre de l'enquête « Emploi en continu ». Les données brutes contiennent 422 133 observations et 555 variables. Ces dernières portent des informations notamment sur diverses caractéristiques concernant l'individu, sa situation professionnelle, ses études et formations, ou encore son logement.

Dans un premier temps, une présentation et une analyse descriptive des données seront effectuées. Dans un second temps, l'étude économétrique sera menée et aboutira à des estimations, suivie de la décomposition d'Oaxaca.

En conclusion, des enseignements en seront tirés et les limites de cette méthode seront exposées.

2 Décomposition d'Oaxaca : principe et fonctionnement

On parle de discrimination contre les femmes lorsque le salaire relatif des hommes est supérieur à celui des femmes dans le cas où les deux groupes sont payés selon les mêmes critères (caractéristiques du poste, caractéristiques individuelles, etc.).

De façon formalisée, un **coefficient de discrimination** D peut être créé pour mesurer empiriquement une quelconque discrimination salariale :

$$D \equiv \frac{W_m/W_f - (W_m/W_f)^0}{(W_m/W_f)^0} \tag{1}$$

avec (W_m/W_f) le rapport salarial observé entre les hommes et les femmes et $(W_m/W_f)^0$ le rapport salarial théorique en l'absence de discrimination.

Par une transformation logarithmique, on peut aboutir à une expression plus parlante :

$$\begin{aligned} D = \frac{W_m/W_f - (W_m/W_f)^0}{(W_m/W_f)^0} &\iff D = \frac{W_m/W_f}{(W_m/W_f)^0} - 1 \\ &\iff D + 1 = \frac{W_m/W_f}{(W_m/W_f)^0} \\ &\iff \ln(D+1) = \ln(W_m/W_f) - \ln\left((W_m/W_f)^0\right) \end{aligned} \tag{2}$$

Selon la théorie de l'économie du travail, les travailleurs sont payés à hauteur de leur productivité marginale, d'où :

$$\left(\frac{W_m}{W_f}\right)^0 = \frac{\mathrm{PM}_m}{\mathrm{PM}_f}$$

avec PM_m et PM_f respectivement la productivité marginale des hommes et des femmes.

L'estimation du coefficient de discrimination D passe entièrement par l'estimation de $(W_m/W_f)^0$ puisque c'est notre seule inconnue, W_m/W_f étant empirique. Dans l'hypothèse d'absence de discrimination, la structure salariale actuelle des femmes s'applique également aux hommes et la structure salariale actuelle des hommes s'applique également aux femmes.

La structure salariale d'un groupe peut être estimée à partir des Moindres Carrés Ordinaires sur un modèle log-linéaire de la forme :

$$\ln(W_i) = X_i'\beta + u_i \ , \qquad i = 1, \ldots, n \tag{3}$$

avec W_i le salaire horaire du travailleur i du groupe, X_i' le vecteur des caractéristiques

individuelles du travailleur, β le vecteur des coefficients à estimer et u_i le bruit du modèle. On aura alors deux structures salariales à estimer, une pour les hommes et l'autre pour les femmes.

De la forme en (2), on en tire que la différence de salaire peut se décomposer en une partie liée à la discrimination, et une autre liée aux différences des caractéristiques individuelles.

On pose :

$$G \equiv \frac{\overline{W}_m - \overline{W}_f}{\overline{W}_f} \tag{4}$$

qui équivaut, après une transformation logarithmique à :

$$\ln(G+1) = \ln(\overline{W}_m) - \ln(\overline{W}_f) \tag{5}$$

avec $\overline{W}_m$ et $\overline{W}_f$ respectivement les salaires horaires moyens des hommes et des femmes. Selon les propriétés des Moindres Carrés Ordinaires, l'estimation par le centre de gravité des variables explicatives tombe sur la valeur moyenne de la variable à expliquer, c'est-à-dire :

$$\ln(\overline{W}_m) = \overline{X}'_m \widehat{\beta}_m \tag{6}$$

$$\ln(\overline{W}_f) = \overline{X}'_f \widehat{\beta}_f \tag{7}$$

avec $\overline{X}_m$ et $\overline{X}_f$ les vecteurs des moyennes des variables explicatives pour respectivement les hommes et les femmes. $\widehat{\beta}_m$ et $\widehat{\beta}_f$ sont les coefficients estimés par la méthode des Moindres Carrés Ordinaires sur le modèle (3).

En remplaçant (6) et (7) dans (5), on obtient :

$$\ln(G+1) = \overline{X}'_m \widehat{\beta}_m - \overline{X}'_f \widehat{\beta}_f \tag{8}$$

On pose :

$$\Delta\overline{X}' \equiv \overline{X}'_m - \overline{X}'_f \tag{9}$$

$$\Delta\widehat{\beta} \equiv \widehat{\beta}_f - \widehat{\beta}_m \tag{10}$$

De (10), on a $\widehat{\beta}_m = \widehat{\beta}_f - \Delta\widehat{\beta}$. De ce résultat, on remplace dans (8) :

$$\ln(G+1) = \overline{X}'_m\widehat{\beta}_m - \overline{X}'_f\widehat{\beta}_f \Longleftrightarrow \ln(G+1) = \overline{X}'_m\left(\widehat{\beta}_f - \Delta\widehat{\beta}\right) - \overline{X}'_f\widehat{\beta}_f$$

$$\Longleftrightarrow \ln(G+1) = \left(\overline{X}'_m - \overline{X}'_f\right)\widehat{\beta}_f - \overline{Z}'_m\Delta\widehat{\beta}$$

$$\Longleftrightarrow \ln(G+1) = \Delta\overline{X}'\widehat{\beta}_f - \overline{X}'_m\Delta\widehat{\beta} \quad (11)$$

$$\Longleftrightarrow \ln(\overline{W}_m) - \ln(\overline{W}_f) = \underbrace{\Delta\overline{X}'\widehat{\beta}_f}_{(a)} - \underbrace{\overline{X}'_m\Delta\widehat{\beta}}_{(b)} \quad (12)$$

La partie (a) représente la différence de salaire liée aux différences dans les caractéristiques individuelles des individus (partie non discriminatoire) [2]. La partie (b) représente la différence de salaire liée à la discrimination [3] De cette décomposition et selon l'hypothèse d'égalité structurelle des salaires entre les deux groupes faite précédemment, on a :

$$\widehat{\ln\left(\frac{W_m}{W_f}\right)}^0 = \Delta\overline{X}'\widehat{\beta}_f \quad (13)$$

Notre mesure de la discrimination est alors :

$$\widehat{\ln(D+1)} = -\overline{X}'_m\Delta\widehat{\beta} \quad (14)$$

ce qui équivaut, d'après (12) à :

$$\widehat{\ln(D+1)} = \ln(G+1) - \Delta\overline{X}'\widehat{\beta}_f = \ln(\overline{W}_m) - \ln(\overline{W}_f) - \Delta\overline{X}'\widehat{\beta}_f \quad (15)$$

De façon parfaitement symétrique, plutôt que d'utiliser (10) pour remplacer $\widehat{\beta}_m$ dans (8), on utilise (10) pour remplacer $\widehat{\beta}_f$ dans (8). On obtient alors, à travers le même raisonnement établi avant :

$$\widehat{\ln(D+1)} = \ln(G+1) - \Delta\overline{X}'\widehat{\beta}_m = \ln(\overline{W}_m) - \ln(\overline{W}_f) - \Delta\overline{X}'\widehat{\beta}_m \quad (16)$$

Les équations (15) et (16) détermineront la part de discrimination entre les hommes et les femmes. Ceci passe par l'estimation des modèles pour les hommes et pour les femmes dans un premier temps. Après avoir déterminé les deux valeurs de $\widehat{\ln(D+1)}$, une simple moyenne arithmétique de ces derniers aboutira à la part de différence salariale expliquée par la discrimination.

2. On a supposé que la structure salariale est la même pour les femmes et pour les hommes, donc le choix du vecteur des $\widehat{\beta}$ est indifférent : $\widehat{\beta}_f = \widehat{\beta}_m$ en cas de non discrimination.

3. La discrimination s'exprime à travers des valeurs différentes au sein des couples $\left(\widehat{\beta}_m^{[k]}, \widehat{\beta}_f^{[k]}\right)$ avec $\widehat{\beta}^{[k]}$ l'estimation des Moindres Carrés Ordinaires du coefficient associé à la k$^{\text{ème}}$ variable explicative.

Deuxième partie

Préparation de la base de données

3 Sélection des variables pertinentes

Le premier travail à réaliser pour débuter ce projet est de sélectionner les variables pertinentes à l'étude. Si la liste de variables retenues est très proche de celle de l'étude de référence, des variables supplémentaires ont pu être considérées grâce à l'enquête « Emploi en continu » de l'INSEE.

Les variables conservées sont les suivantes :

— `SALRED`, le salaire mensuel net redressé des non-réponses (y compris les primes mensualisées et redressées des non-réponses) ;
— `SEXE`, le sexe du sondé ;
— `CSER`, la catégorie socio-professionnelle pour les actifs (niveau agrégé, PCS 2003) ;
— `NIVP`, niveau d'enseignement selon une version approuvée par le comité interministériel de la formation continue ;
— `MATRI`, le statut matrimonial légal ;
— `TPPRED`, le temps de travail redressé dans l'emploi principal ;
— `NBHEUR`, le nombre d'heures correspondant au salaire déclaré ;
— `NUITC`, travail de nuit (entre minuit et cinq heures du matin) ;
— `NAFG10N`, l'activité de l'établisssement actuel ;
— `PUB3FP`, caractère public ou privé de l'employeur au sens de l'OEP ;
— `PAYNEU27`, pays de naissance (l'union européenne des 27) ;
— `LNAIS`, lieu de naissance ;
— `NFR`, code de nationalité ;
— `TUU`, commune urbaine ou rurale ;
— `REG`, région de résidence ;
— `NBAGENF`, nombre et âge des enfants du logement au 31 décembre de l'année d'enquête dans

le logement ;

- `AGE`, âge détaillé au dernier jour de la semaine de référence ;
- `ANCENTR`, ancienneté dans l'entreprise ou dans la fonction publique en mois ;
- `STATUT`, statut détaillé mis en cohérence avec la profession ;
- `CONTRA`, type de contrat de travail.

Cela représente donc 20 variables conservées. Des variables indicatrices seront créées dans un second temps afin de pouvoir appliquer la méthode économétrique choisie.

4 Restrictions sur les individus retenus

Afin d'obtenir une base exploitable et cohérente, des restrictions sont appliquées sur les individus. Plusieurs filtres ont été mis en place.

Tout d'abord, les individus n'ayant pas de salaire renseigné ont été supprimés de la base (variable `SALRED`). Les individus n'ayant pas de volume horaire travaillé renseigné ont également été retirés de la base (variable `NBHEUR`).

Le choix de ne conserver que les individus dont l'âge est compris entre 18 et 71 ans inclus s'est imposé (variable `AGE`).

Aussi, les individus catégorisés en tant que chômeurs n'ayant jamais travaillé ou dont la catégorie socio-professionnelle n'est pas renseignée ont été exclus de la base (variable `CSER`, modalités `"0"` et `"8"`).

Après études préliminaires, ont été catégorisés comme individus aberrants et donc supprimés de la base les individus ayant un volume horaire de travail supérieur à 260 heures par mois (variable `NBHEUR`) ; ainsi que les individus dont le salaire mensuel est inférieur à 600 EUR ou supérieur à 20 000 EUR (variable `SALRED`).

Ces restrictions ont permis l'obtention d'une base propre et conforme à l'étude. L'échantillon final compte 30 399 observations caractérisés par 43 variables.

5 Création des variables indicatrices (dummies)

5.1 `SEXE` : le genre de l'individu

La variable `SEXE` prend initialement les valeurs `1` et `2`, pour désigner respectivement les hommes et les femmes. Elle est transformée en variable indicatrice prenant la valeur `1` si l'individu correspondant est une femme, `0` sinon. Ce choix s'explique par l'angle pris par l'étude : il s'agit de quantifier la discrimination salariale **envers les femmes**.

Dans la base d'origine, il y a 200 043 hommes et 222 090 femmes, soit des parts respectivement égales à 47.39% et 52.61%.

5.2 `CSER` : la catégorie socioprofessionnelle

La variable `CSER` indique la CSP [4] des individus sondés. Elle prend 8 modalités différentes :

Valeur	Correspondance	Effectif
0	Non renseigné	77
1	Agriculteurs exploitants	4 523
2	Artisans, commerçants et chefs d'entreprise	13 807
3	Cadres et professions intellectuelles supérieures	37 220
4	Professions intermédiaires	52 500
5	Employés	64 537
6	Ouvriers	50 699
8	Chômeurs n'ayant jamais travaillé	3 688

Table 1 – Détails des modalités de la variable `CSER`

Sans intention directe à vouloir hiérarchiser les différentes CSP dans cette étude, la modalité de référence est celle des **Employés**, pour son effectif (64 537 employés sondés) supérieur aux autres. Les individus prenant pour valeur `0` ou `8` pour la variable `CSER` ne seront pas considérés dans l'analyse (voir section 4).

4. catégorie socio-professionnelle (selon la nomenclature PCS-2003).

Cinq variables indicatrices sont construites à partir des modalités restantes : `CSP_AGRI`, `CSP_ARTI`, `CSP_CADRE`, `CSP_INTERM`, `CSP_OUVRI`. Si un individu prend la valeur `1` dans l'une de ces variables, cela indique son appartenance à la CSP associée.

Étant donné que le groupe de référence est celui des **Employés**, l'interprétation des coefficients estimés passe par la comparaison entre la CSP **Employés** et une autre CSP. Par exemple, le coefficient associé à la variable indicatrice `CSP_CADRE` correspond à la différence (de salaire horaire en logarithme ici) lorsqu'un travailleur est **Cadre** plutôt qu'**Employé**, *ceteris paribus*.

5.3 `NIVP` : le niveau d'étude

La variable `NIVP` indique le niveau d'enseignement des individus sous forme de « cap atteint » au maximum. Elle prend 10 modalités différentes :

Valeur	Correspondance	Effectif
10	Diplôme bac+5 et plus	25 503
20	Diplôme niveau licence, maîtrise	28 170
30	Diplôme niveau bac+2	41 305
40	Bac et enseignement supérieur sans diplôme bac+2	29 159
41	Niveau bac sans études supérieures	65 011
50	Niveau terminale CAP-BEP, lycée	121 353
60	Troisième, année non terminale CAP-BEP	33 771
71	Collège	10 371
72	Enseignement primaire	60 818
73	Pas d'études	6 672

Table 2 – Détails des modalités de la variable `NIVP`

Pour simplifier cette nomenclature, plusieurs modalités seront regroupées en un seul groupe de façon à créer 4 variables indicatrices : `BAC5` qui regroupe les **diplômes bac+5 et plus** (modalité `10` de `NIVP`), `BAC3` qui regroupe les **diplômes bac+3 bac+4** (modalité `20`), `BAC2` qui regroupe les **diplômes bac+2** (modalité `30`) et `BAC` qui regroupe les **bacheliers**, avec ou sans études sans diplôme supérieur (modalités `40` et `41`). Les **non bacheliers** constitueront le groupe de référence de façon à alléger l'interprétation économique. Ce groupe est formé à partir des modalités `50`,

60, 71, 72 et 73. Il est également le groupe avec l'effectif le plus élevé parmi les nouvelles classes constituées.
Les coefficients estimés, associés à ces variables dichotomiques correspondront à l'effet engendré par la possession d'un diplôme universitaire correspondant à la variable interprétée, *ceteris paribus*.

5.4 MATRI : statut matrimonial légal

Le statut matrimonial légal des individus est indiqué par la variable MATRI. Elle prend 4 modalités différentes :

Valeur	Correspondance	Effectif
1	Célibataire	151 690
2	Marié(e) ou remarié(e)	206 154
3	Veuf(ve)	31 715
4	Divorcé(e)	32 571

Table 3 – Détails des modalités de la variable MATRI

Chaque situation matrimoniale étant assez éloignée l'une des autres, notamment en terme de conséquences, 3 variables indicatrices sont créées pour distinguer ces différents effets. La situation de **célibataire** sera la référence pour l'interprétation par la suite. Les variables créées sont MARRIED, WIDOW et DIVORCED. Prendre la valeur 1 sur la variable MARRIED indique que l'individu est marié, de même pour les autres variables.

5.5 TPPRED : temps complet ou temps partiel

La variable d'origine TPPRED prend les valeurs 1 et 2 pour indiquer, respectivement que l'individu travaille à temps complet ou à temps partiel. Il y a initialement 167 374 travailleurs à temps complet contre 37 757, à temps partiel. Cette variable est transformée de façon à prendre les valeurs 0 et 1, avec comme modalité de référence, le **temps complet**. Ainsi, l'interprétation du coefficient estimé est l'effet sur le salaire de travailler à temps partiel, plutôt qu'à temps complet, *ceteris paribus*.

5.6 NUITC : travail de nuit

La variable NUITC est une variable indiquant si l'individu travaille de nuit, et dans le cas échéant, occasionnellement ou habituellement :

Valeur	Correspondance	Effectif
1	Travaille de nuit habituellement	14 607
2	Travaille de nuit occasionnellement	18 469
3	Ne travaille jamais de nuit	217 075

Table 4 – Détails des modalités de la variable NUITC

Par souci de simplicité, cette variable est rendue dichotomique en regroupant les modalités 1 et 2 en une seule pour constituer le groupe des individus qui travaillent de nuit (occasionnellement ou habituellement). Le groupe de référence est celui de ceux qui ne travaillent pas de nuit, pour pouvoir étudier l'effet du travail de nuit sur le salaire. En clair, un individu travaillant de nuit prendra la valeur 1 sur la variable NUITC.

5.7 PUB3FP : caractère public ou privé de l'entreprise

La variable PUB3FP indique si l'entreprise dans laquelle travaille l'individu est publique ou privée. Elle prend 4 modalités :

Valeur	Correspondance	Effectif
1	État	16 867
2	Collectivités locales	15 192
3	Hôpitaux publics	8 727
4	Secteur privé	140 349

Table 5 – Détails des modalités de la variable PUB3FP

Une seule variable dichotomique est construite en distinguant simplement si l'entreprise appartient au secteur public (modalités 1, 2 et 3 regroupées) ou au secteur privé (modalité 4). Celle-ci prendra la valeur 1 lorsque l'entreprise est dans le secteur public.

Ainsi, le coefficient estimé correspondra à l'effet sur le salaire lorsqu'un individu travaille dans une entreprise publique plutôt que privée, *ceteris paribus*.

5.8 NAFG10N : secteur économique de l'entreprise

L'activité principale de l'établissement actuel des individus se lit à travers la variable NAFG10N. Celle-ci établit une nomenclature en 10 postes :

Valeur	Correspondance	Effectif
AZ	Agriculture, sylviculture et pêche	6 504
BE	Industrie manufacturière, industries extractives et autres	29 099
FZ	Construction	13 984
GI	Commerce de gros et de détail, transports, hébergement et restauration	42 944
JZ	Information et communication	5 559
KZ	Activités financières et d'assurance	6 499
LZ	Activités immobilières	2 327
MN	Activités spécialisées, scientifiques et techniques et activités de services administratifs et de soutien	21 489
OQ	Administration publique, enseignement, santé humaine et action sociale	62 627
RU	Autres activités de services	13 275

Table 6 – Détails des modalités de la variable NAFG10N

Sans motivation économique à prendre un groupe en particulier comme référence, c'est la modalité **Administration publique, enseignement, santé humaine et action sociale** ("OQ") qui sera retenue pour son fort effectif. 9 variables indicatrices découlent alors des modalités restantes : SECT_ENT_AZ, SECT_ENT_BE, SECT_ENT_FZ, SECT_ENT_GI, SECT_ENT_JZ, SECT_ENT_KZ, SECT_ENT_LZ, SECT_ENT_MN et SECT_ENT_RU.

Dans la même logique que les variables créées à partir des CSP, l'interprétation des coefficients associés aux variables liées au secteur économique de l'entreprise se fera en comparant la modalité étudiée à la modalité de référence.

5.9 PAYNEU27 : pays de naissance et Union Européenne

Cette variable indique si oui ou non l'individu est né dans un pays extérieur à l'Union Européenne (des 27). La modalité (transformée) 1 (respectivement 0) indique que le pays de naissance est non

membre de l'Union Européenne (respectivement membre de l'Union Européenne). Il y a 384 659 personnes nées au sein de l'Union Européenne contre 37 474 personnes dans l'échantillon de départ. L'estimation permettra de conclure sur l'effet de ne pas être né au sein de l'Union Européenne sur le salaire perçu, *ceteris paribus*.

5.10 `LNAIS` : pays de naissance et France

Dans la même logique qu'avec la variable `PAYSNEU27`, la variable `LNAIS` indique si oui ou non l'individu est né dans un pays qui n'est pas la France. Elle prend la modalité `1` lorsque le sondé est né en dehors du territoire français. Il y a 369 506 personnes nées en France contre 52 627 dans l'échantillon.
En ayant le groupe des « nés en France » comme référence, l'étude permet de voir l'effet d'être né ailleurs sur le salaire, *ceteris paribus*.

5.11 `NFR` : nationalité française ou étrangère

Dans la continuité des deux précédentes variables, la variable `NFR` indique si l'individu est de nationalité française ou non. Plus exactement, elle prend la modalité `1` lorsqu'il est de nationalité étrangère et `0` sinon. Il y a 397 534 français (de naissance ou par naturalisation) contre 24 598 étrangers dans l'échantillon.
L'impact d'être de nationalité étrangère sur le salaire en France se lira à partir du coefficient estimé.

5.12 `TUU` : commune urbaine ou rurale

La variable `TUU` indique si l'individu habite dans une commune urbaine ou rurale. Elle prend la modalité `1` si celle-ci est urbaine, `0` sinon. On dénombre `105 339` habitants ruraux contre `316 764` habitants urbains dans les données.
Cette distinction permettra de lire l'effet d'être un urbain sur le salaire perçu.

5.13 `REG` : région de résidence

La variable `REG` permet de localiser la région de résidence de chaque individu. Elle prend 22 modalités différentes :

Valeur	Correspondance	Effectif
11	Ile-de-France	72 259
21	Champagne-Ardenne	13 076
22	Picardie	12 780
23	Haute-Normandie	12 290
24	Centre	16 598
25	Basse-Normandie	11 109
26	Bourgogne	12 036
31	Nord-Pas de Calais	28 098
41	Lorraine	15 264
42	Alsace	13 955
43	Franche-Comté	9 660
52	Pays de la Loire	24 677
53	Bretagne	20 974
54	Poitou-Charentes	12 028
72	Aquitaine	21 920
73	Midi-Pyrénées	17 429
74	Limousin	9 875
82	Rhône-Alpes	40 379
83	Auvergne	9 233
91	Languedoc-Roussillon	17 134
93	Provence-Alpes-Côte-d'Azur	29 649
94	Corse	1 710

Table 7 – Détails des modalités de la variable `REG`

On note que les départements et territoires d'Outre-Mer ne sont pas visés par cette enquête INSEE. Par souci de simplicité, une seule variable dichotomique `IDF` est construite pour distinguer la région **Ile-de-France** des autres. Elle prendra la valeur `1` lorsque l'individu prend la modalité `11` sur la

variable `REG`, `0` sinon. Il y a alors 72 259 individus résidant en Ile-de-France contre 349 874 dans l'échantillon de l'enquête.
L'estimation du coefficient associé à la variable correspondra alors à l'impact d'être un résident d'Ile-de-France sur le salaire, par rapport aux provinciaux, *ceteris paribus*.

5.14 `NBAGENF` : le nombre d'enfants à charge

La variable `NBAGENF` indique le nombre d'enfants à charge (en plus d'une indication sur l'âge du plus jeune). Elle prend 10 modalités :

Valeur	Correspondance	Effectif
0	Pas d'enfant de moins de 18 ans	279 943
1	Un enfant de 6 à 17 ans	47 113
2	Un enfant de 3 à 5 ans	6 746
3	Un enfant de moins de 3 ans	10 078
4	Deux enfants, dont le plus jeune a de 6 à 17 ans	34 601
5	Deux enfants, dont le plus jeune a de 3 à 5 ans	10 367
6	Deux enfants, dont le plus jeune a moins de 3 ans	9 145
7	Trois enfants ou plus, dont le plus jeune a de 6 à 17 ans	11 036
8	Trois enfants ou plus, dont le plus jeune a de 3 à 5 ans	6 805
9	Trois enfants ou plus, dont le plus jeune a moins de 3 ans	6 299

Table 8 – Détails des modalités de la variable `NBAGENF`

Trois variables indicatrices sont construites : `NBENF1` (modalités `1`, `2` et `3`), `NBENF2` (modalités `4`, `5` et `6`) et `NBENF3PLUS` (modalités `7`, `8` et `9`), désignant respectivement, 1 enfant, 2 enfants et 3 enfants ou plus. Le groupe de référence est donc celui des individus sans enfants (modalité `0`) pour faciliter l'interprétation économiques des coefficients à estimer.
On dénombre respectivement, 63 937, 54 113 et 24 140 individus avec un, deux et trois ou plus enfants à charge.

5.15 `CONTRA` : le type de contrat du poste

Enfin, la variable `CONTRA` distingue les différents contrats possibles des individus. Elle prend 5 modalités :

Valeur	Correspondance	Effectif
1	Contrat à durée indéterminée (y compris contrat Nouvelles Embauches)	120 970
2	Contrat à durée déterminée autre que saisonnier	15 779
3	Contrat saisonnier	1 018
4	Contrat d'intérim ou de travail temporaire	3 961
5	Contrat d'apprentissage ou contrat en alternance	3 309

Table 9 – Détails des modalités de la variable `CONTRA`

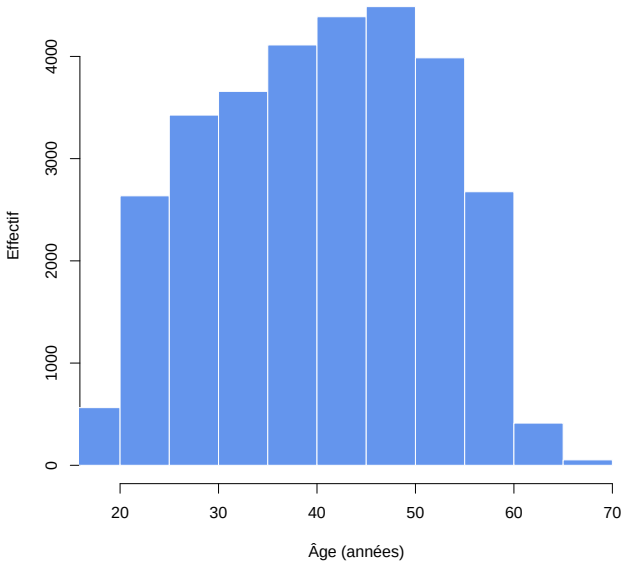
Deux variable indicatrices sont construites pour simplifier cette nomenclature. On prend la modalité `CDI` comme référence pour son effectif surpassant les autres modalités (minoritaires). On a : `CDD` qui correspond à la modalité `2` uniquement et `AutreCDD` qui regroupe les modalités restantes : `3`, `4` et `5`. Ainsi, un individu en CDD autre que saisonnier prendra la valeur 1 sur la variable `CDD` construite par exemple. Il y a 120 970 travailleurs en CDI, 15 779 en CDD autre que saisonnier et 8 288 en contrat saisonnier, d'intérim, d'alternance ou d'apprentissage.

La valeur estimée du coefficient associé à la variable `CDD` quantifie l'effet sur le salaire d'avoir une CDD (non saisonnier) plutôt qu'un CDI dans notre modèle, *ceteris paribus*. La même interprétation se transpose sur la variable `AutreCDD`.

6 Statistiques descriptives

Des statistiques descriptives sont faites afin d'explorer les données avant l'application des méthodes économétriques. Dans un premier temps, il s'agira de statistiques descriptives univariées, puis dans un second temps des statistiques descriptives bivariées seront présentées.

6.1 Statistiques descriptives univariées



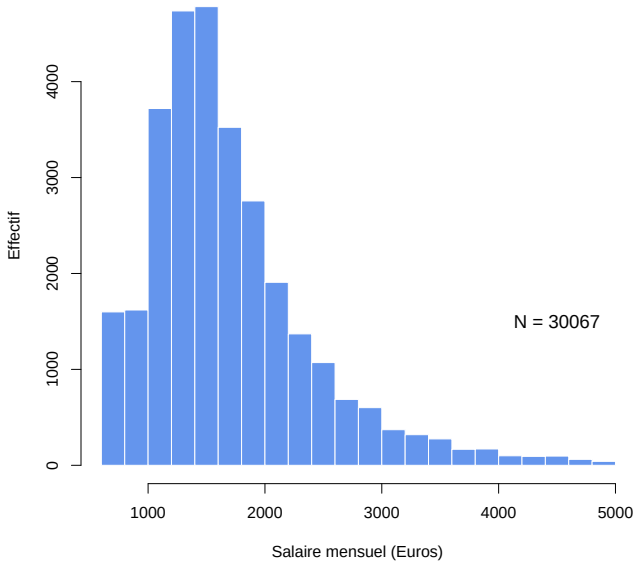


Figure 1 – Histogramme des âges des individus Figure 2 – Histogramme des salaires mensuels

L'âge des individus sondés se distribue de façon normale, sans qu'une classe particulière surpasse les autres en terme d'effectif. La tendance centrale semble se positionner entre 40 et 50 ans. La distribution est légèrement décalée à droite, donnant un faible étalement vers la gauche. Les jeunes travailleurs représentent une part plus importante dans l'échantillon.

La répartition des salaires quant à elle est plutôt inégale, la majorité des individus gagne moins de 1 600 Euros par mois (la médiane est de 1 550 Euros). Au-dessus de la classe modale, le nombre d'observations par tranche de salaire diminue rapidement. De plus, pour ce graphique le choix de restreindre aux individus gagnant moins de 5 000 Euros par mois a été fait, afin d'en garantir la lisibilité. Le graphique représente donc 30 067 individus au lieu des 30 399 que compte la base.

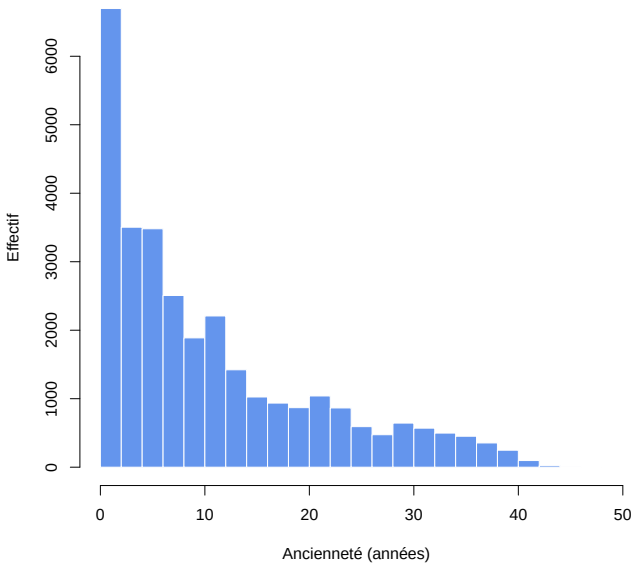


Figure 3 – Ancienneté en années

L'ancienneté n'est pas très élevée en général parmi les sondés, La classe modale est celle entre 0 et 2 ans, avec plus de 6 000 individus. La médiane est de 8 ans, c'est-à-dire qu'au moins 50% des travailleurs ont moins de 8 ans d'ancienneté dans leur entreprise. Environ 1 200 personnes ont une ancienneté supérieure à 35 ans.

La quasi-totalité des sondés travaille entre 35 et 40 heures par semaine, ce qui est cohérent avec la durée légal du travail pour un temps complet en France. Les travailleurs en temps partiel représentent une minorité dans l'échantillon.

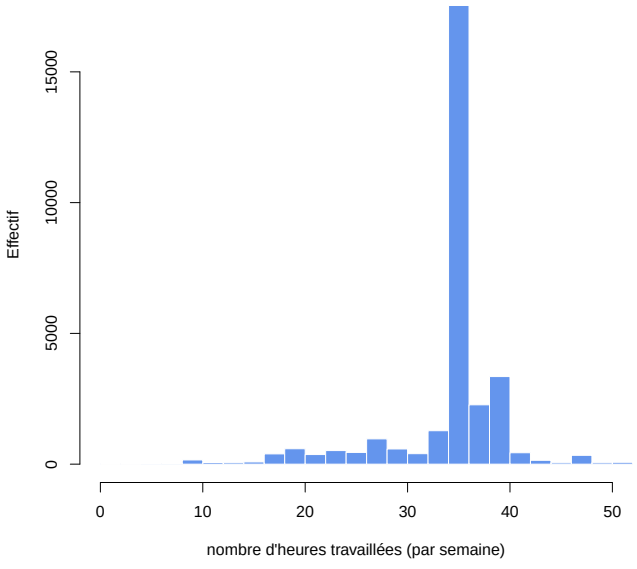


Figure 4 – Temps de travail hebdomadaire

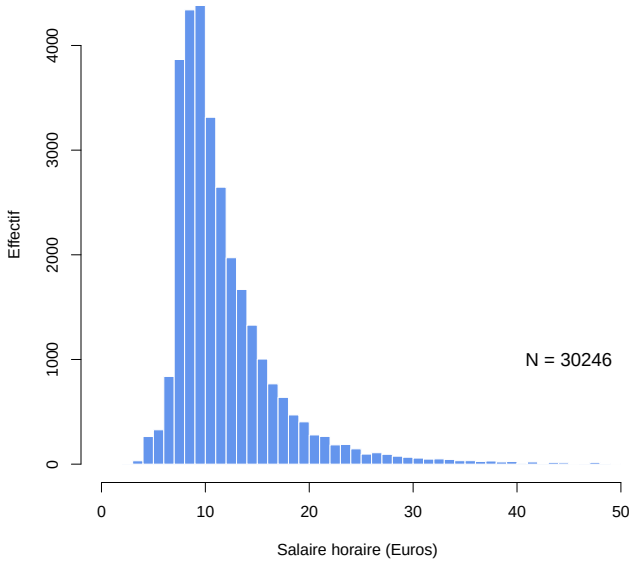


Figure 5 – Salaire horaire

Le salaire horaire a tendance à être compris entre 8 et 14 euros de l'heure. Comme pour le salaire mensuel, un choix a été fait de restreindre les individus afin de garantir la pertinence et la lisibilité du graphique. Ici, ont été conservés les individus gagnant moins de 50 euros par heure travaillée, soient 30 246 individus.

Selon les camemberts suivants, la base est composée à 46% de femmes et à 54% d'hommes. Une majorité des travailleurs habite en Province, 16% des individus vivent en Ile-de-France. Par ailleurs, seuls 6% des travailleurs de l'échantillon travaillent dans le secteur public. Enfin, 73% des individus vivent dans un milieu urbain.

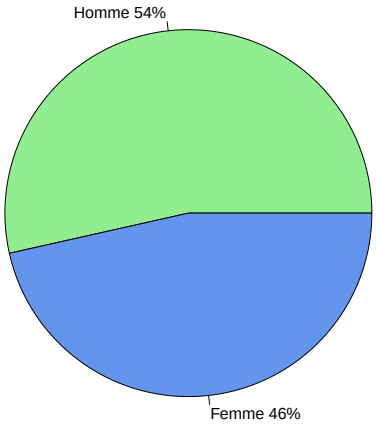


Figure 6 – Répartition des genres

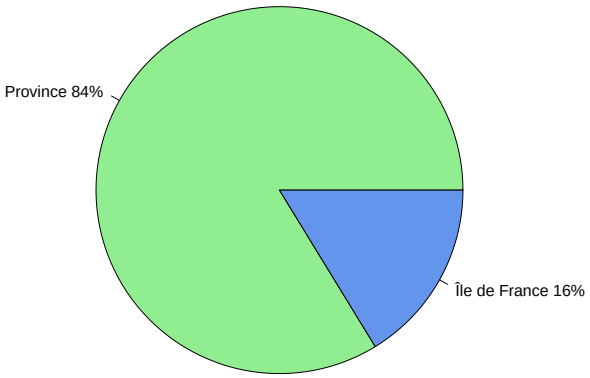


Figure 8 – Répartition province/Ile-de-France

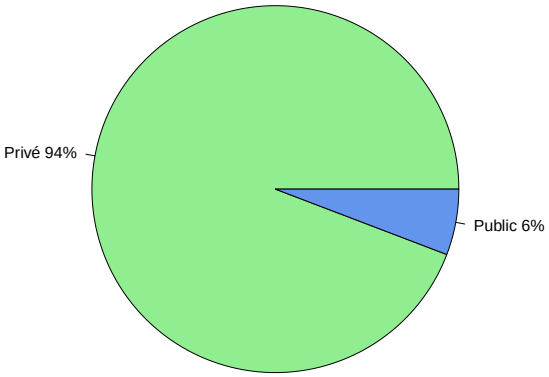


Figure 7 – Répartition des domaines

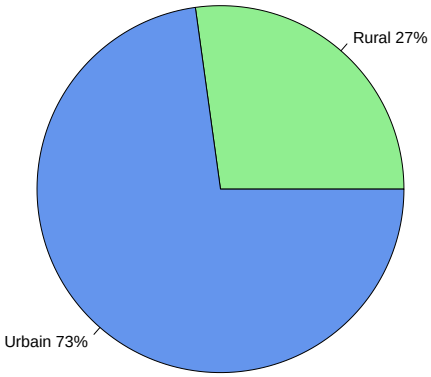


Figure 9 – Répartition Milieu urbain/rural

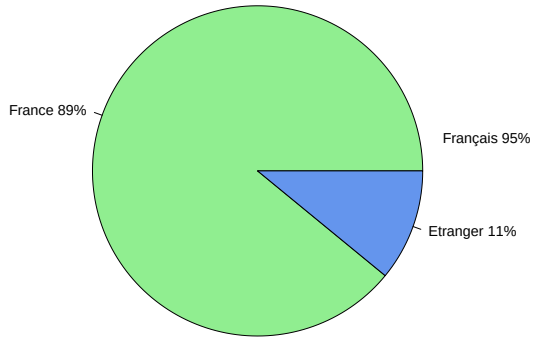


Figure 10 – Pays de naissance

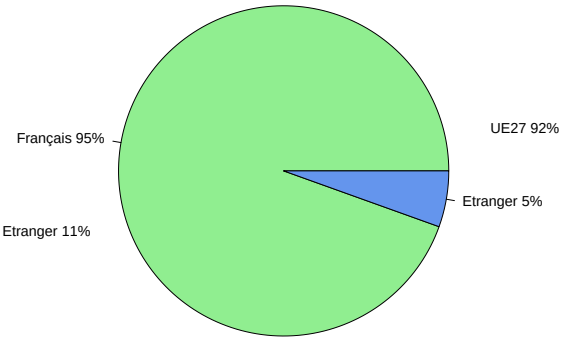


Figure 11 – Nationalité

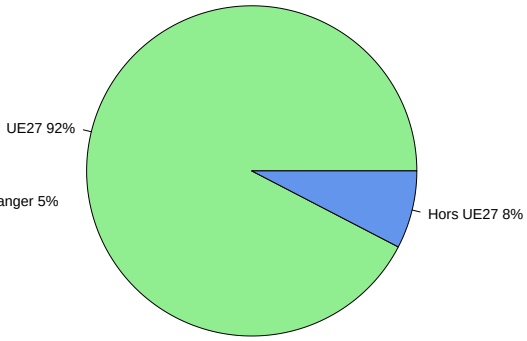


Figure 12 – Région de naissance

Si 89% des individus sont nés en France, 95% de l'échantillon entier est de nationalité française. Par ailleurs, 92% des travailleurs sont nés dans l'un des pays membres de l'Union Européenne.

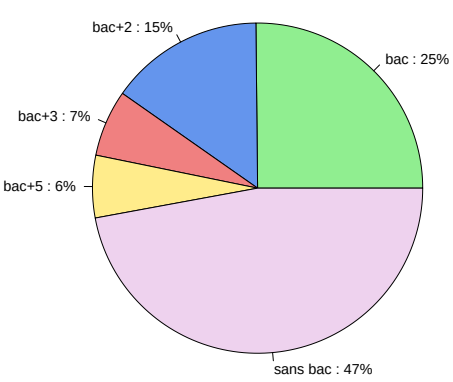


Figure 13 – Diplômes

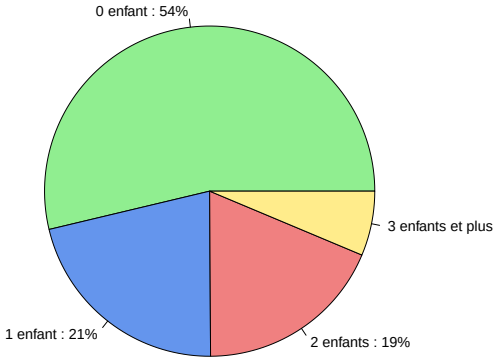


Figure 14 – Nombre d'enfants

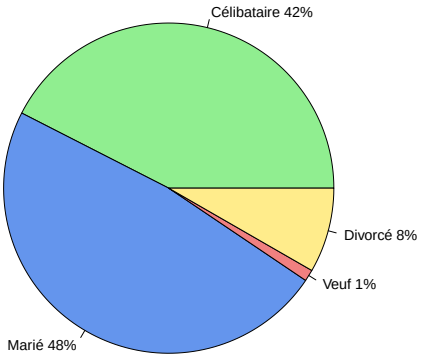


Figure 15 – Situation maritale

Presque la moitié de l'échantillon ne possède pas le baccalauréat et un quart s'est arrêté avant d'obtenir un diplôme de niveau BAC+2. La majorité de la population n'a pas d'enfant (54%), contre 40% qui en ont entre 1 et 2 enfants et 6% qui ont au moins 3 enfants. Environ 2 travailleurs sur 5 sont célibataires. Les individus divorcés restent très minoritaires dans la population française (8%).

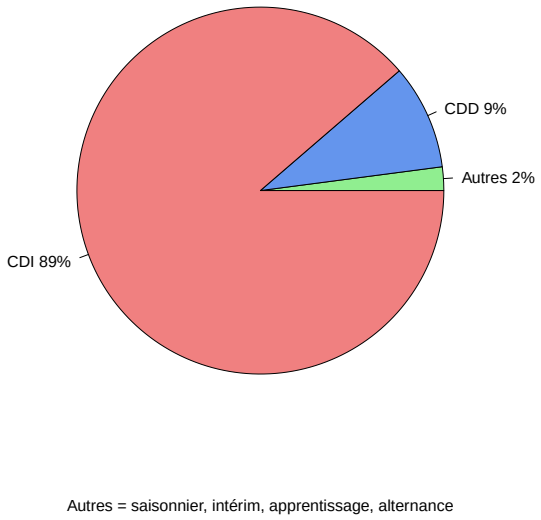


Figure 16 – Types de contrats

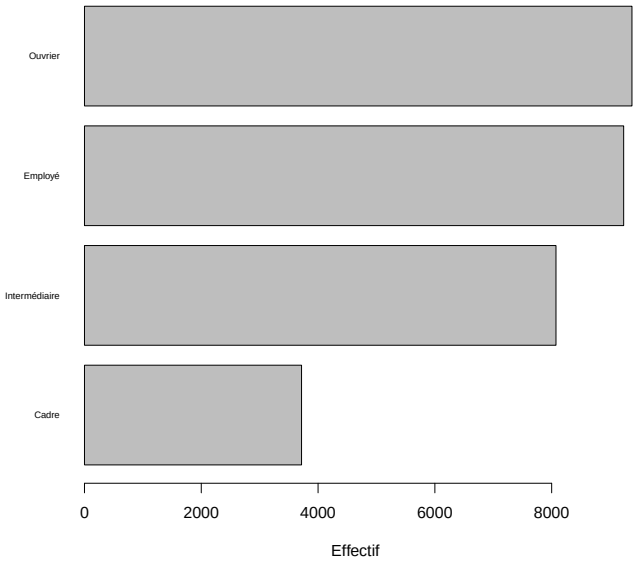


Figure 17 – Catégorie Socio-Professionnelle

9 individus retenus sur 10 sont en CDI et les CDD représentent 9% des contrats signés. La majorité des individus est ouvrière ou employée. Environ 8 000 travailleurs exercent une profession intermédiaire, et ils sont moins de 4000 à être cadres.

* * *

Ces différents graphiques descriptifs posent le panorama des caractéristiques de la population française active en 2012. Une analyse bivariée serait plus pertinente quant à l'identification des différences entre les hommes et les femmes, mais également à l'identification de la relation entre le salaire et les différentes variables qualitatives.

6.2 Statistiques descriptives bivariées

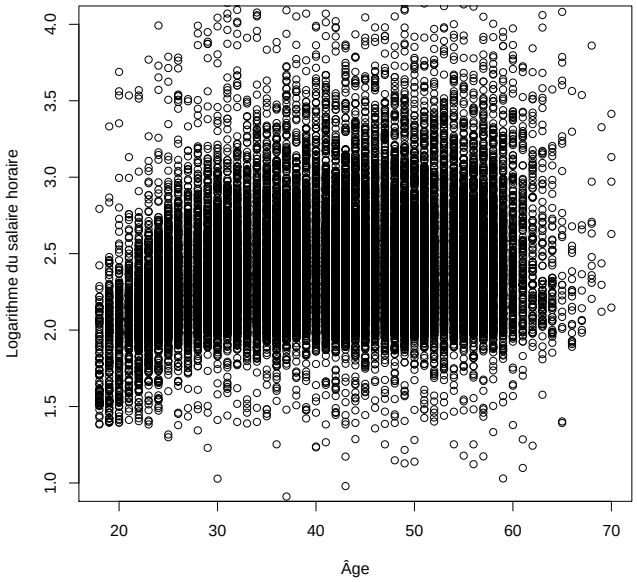


Figure 18 – Salaire selon l'âge

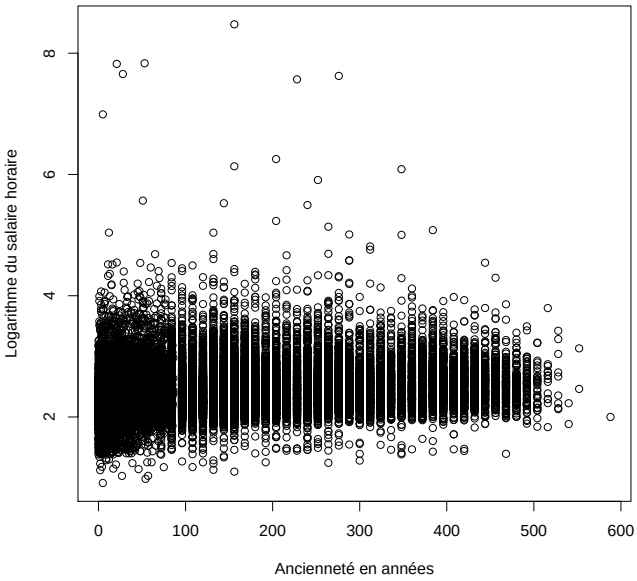


Figure 19 – Salaire selon l'ancienneté

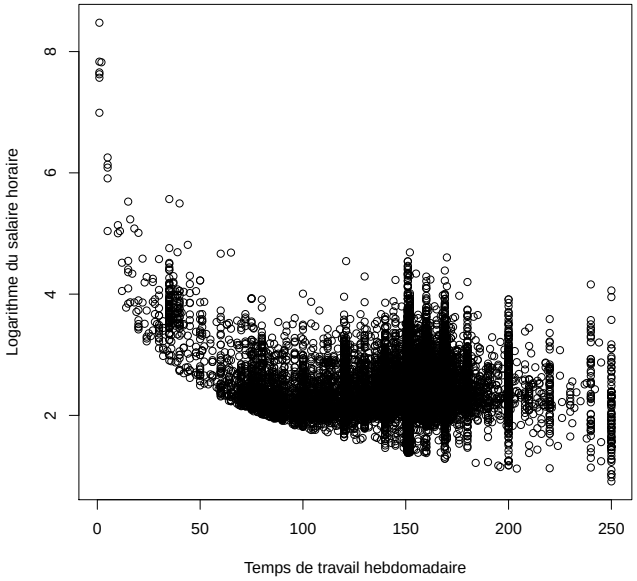


Figure 20 – Salaire horaire en logarithme selon le temps de travail

Il semble y avoir une légère relation positive entre le salaire horaire en logarithme et l'âge. Cette relation semble être courbée, induisant une relation quadratique entre ces deux variables.

Il ne semble pas y avoir de relation entre le salaire horaire en logarithme et l'ancienneté de l'individu au sein de l'entreprise qui l'emploie. La variable sur l'ancienneté sera quand même conservée dans l'analyse économétrique en raison des apports de la littérature économique sur la relation avec le salaire mensuel cette fois-ci.

Une relation négative apparaît entre le salaire horaire en logarithme et le temps de travail hebdomadaire. La forme de la relation préconise une transformation de la variable `NBHEUR` par la suite. Les restrictions faites sur ces deux variables ont influencé sur ce nuage.

À présent, les femmes et les hommes vont être comparés par le biais de 6 variables. L'intérêt est de voir les différences de structures entre ces deux groupes.

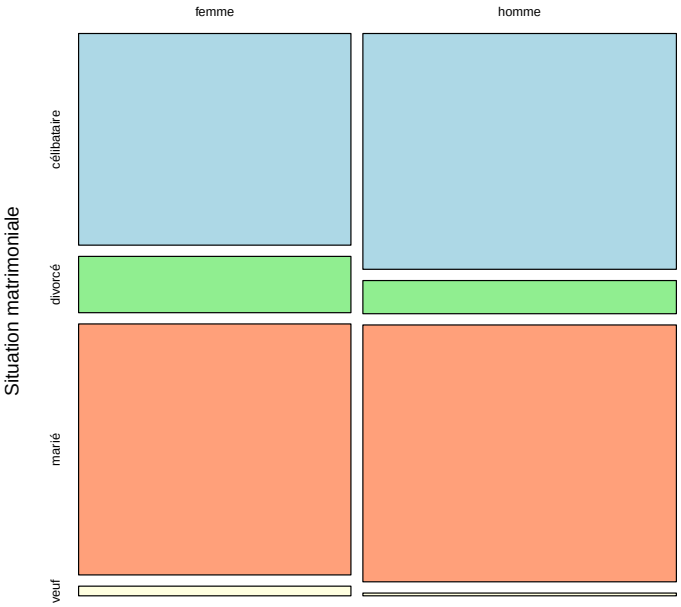


Figure 21 – Situation maritale selon le sexe

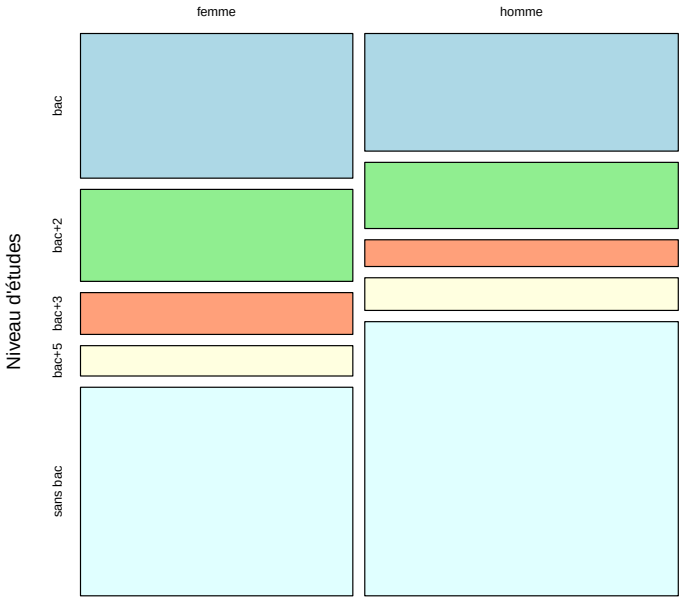


Figure 22 – Etudes selon le sexe

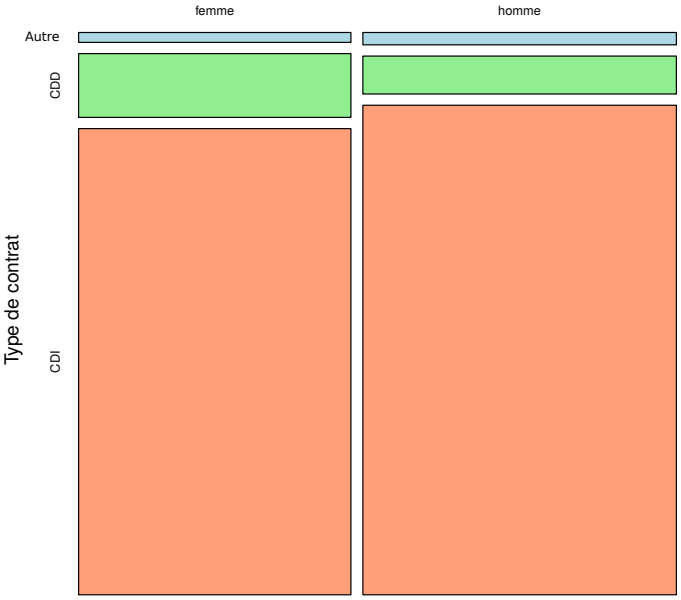


Figure 23 – Contrat selon le sexe

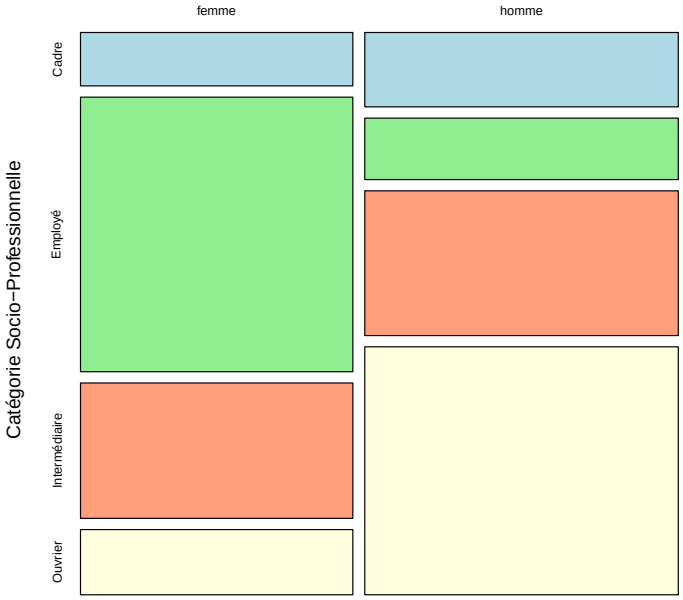


Figure 24 – CSP selon le sexe

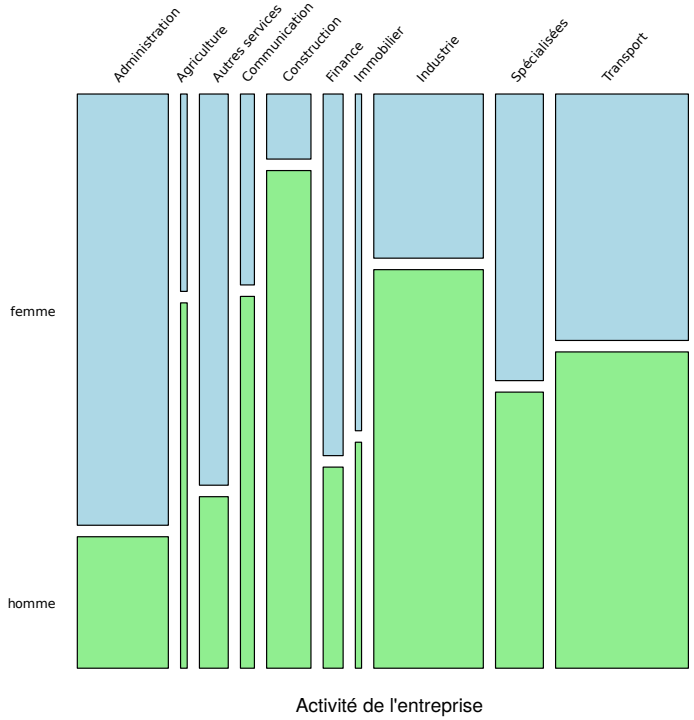


Figure 25 – Activité de l'entreprise selon le sexe

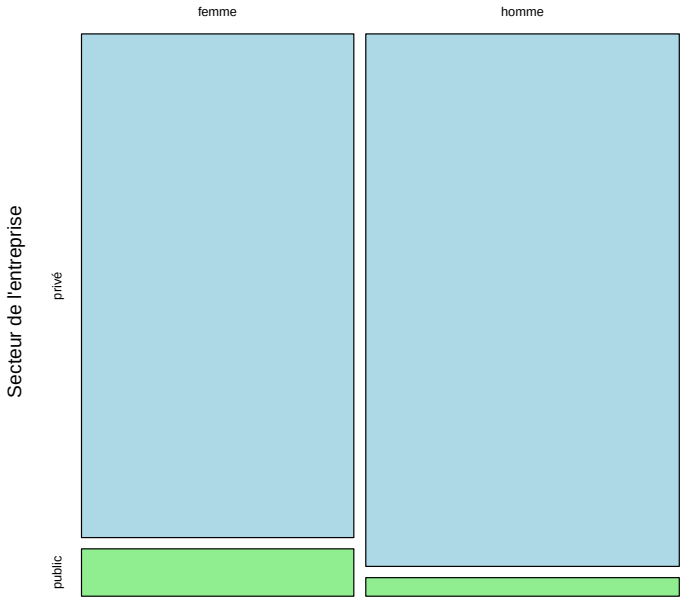


Figure 26 – Secteur selon le sexe

Sur le graphique 21, on peut voir que les situations maritales des individus sont quasiment identiques des deux côtés, le sexe de l'individu ne l'influence pas de manière significative.

On peut noter dans le graphique 22, que les hommes ont tendance à avoir fait moins d'études que les femmes. La part des hommes non bacheliers dans l'échantillon masculin est supérieure à celle des femmes et la part des femmes diplômées d'un bac+5 dans l'échantillon féminine est supérieure à celle des hommes. Par ailleurs, on note une légère différence dans la répartition des types de contrat selon le sexe (graphique 23). Pour un effectif égal, les hommes sont plus nombreux à être en CDI que les femmes.

La différenciation de structuration selon le sexe se remarque sur la répartition des CSP au sein des deux groupes. Près d'une femme sur deux est une employée, tandis que chez les hommes, les employés masculins sont minoritaires. À l'opposé, la part d'ouvriers chez les hommes est 3 fois plus grande que celle chez les femmes. Les deux autres catégories restent relativement similaires. Les agriculteurs et artisans ne sont pas représentés dans l'échantillon, ce qui explique leur absence sur le graphique 24.

Les femmes ont plus tendance à travailler dans l'administration, la finance ou l'immobilier que les hommes. Quant aux hommes, ils sont majoritaires dans les secteurs de l'agriculture, la construction, l'industrie ou le transport.

Les femmes ont plus tendance à se tourner vers le secteur public que les hommes. La part de femmes travaillant dans ce secteur reste cependant très faible dans la population féminine.

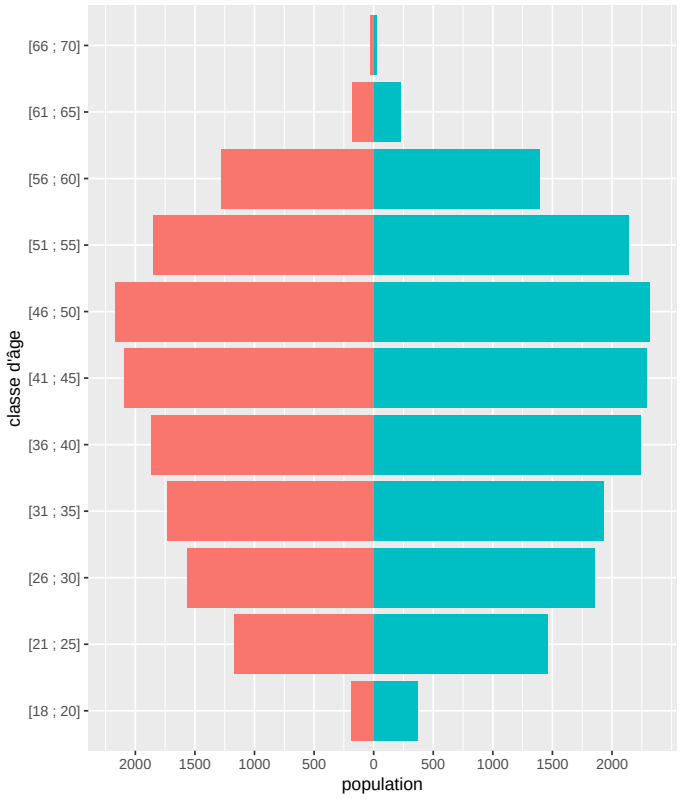


Figure 27 – Âge selon le sexe

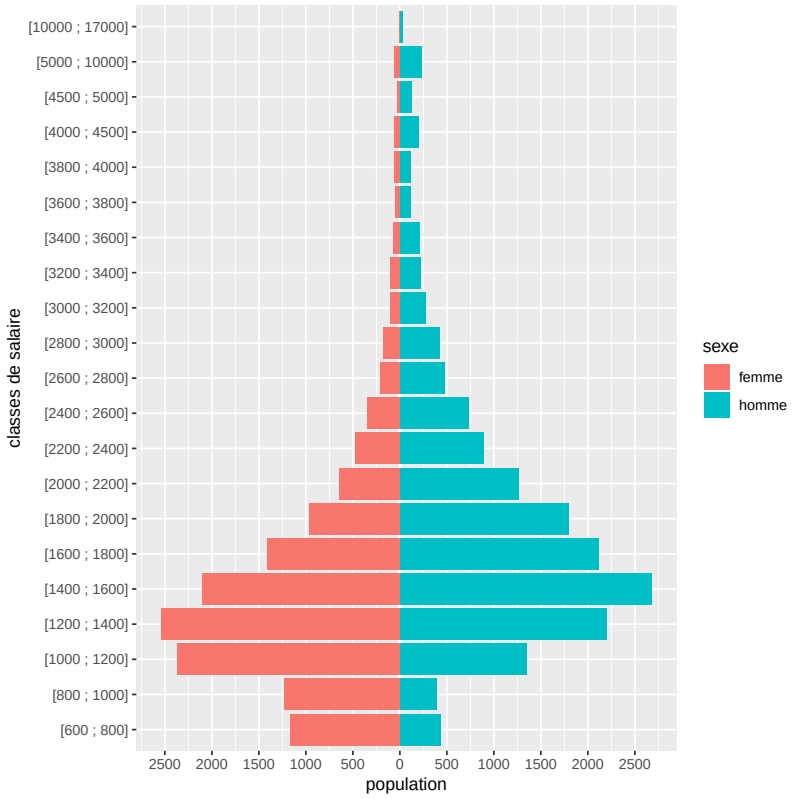


Figure 28 – Salaire selon le sexe

Selon le graphique 27, le sexe et l'âge des sondés n'ont pas de relation. Les hommes actifs ne sont pas plus âgés que les femmes actifs, et réciproquement. Les distributions, d'un côté ou d'un autre restent très similaires.

Quant à l'étude graphique de la distribution des salaires selon le sexe, on note que les femmes ont tendance à être plus nombreuses sur les bas salaires et que les hommes ont tendance à être davantage présents en haut de la grille des salaires. Les femmes ont tendance à être moins payées que les hommes. L'inversion du rapport de force s'effectue à partir de 1 400 euros de salaire mensuel.

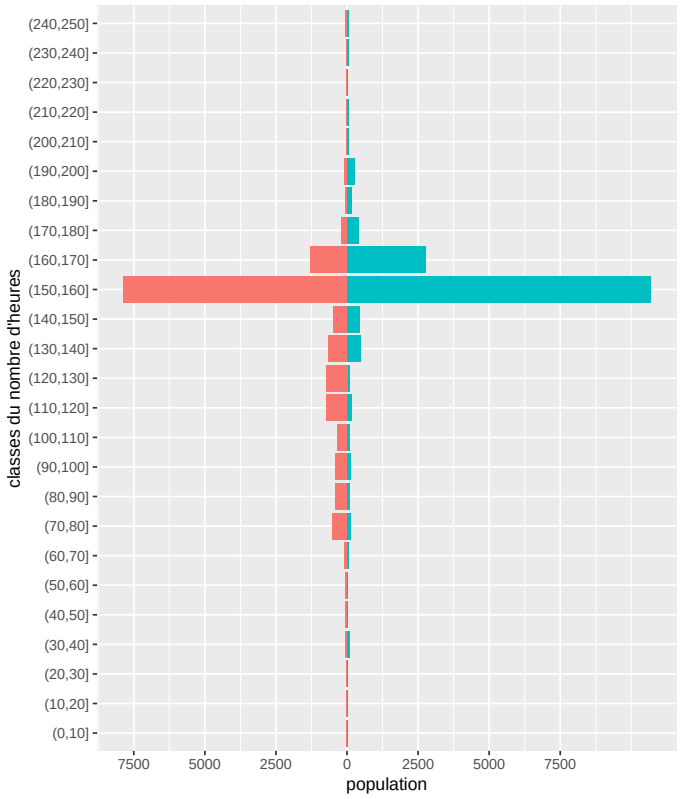


Figure 29 – Nombre d'heures travaillées selon le sexe

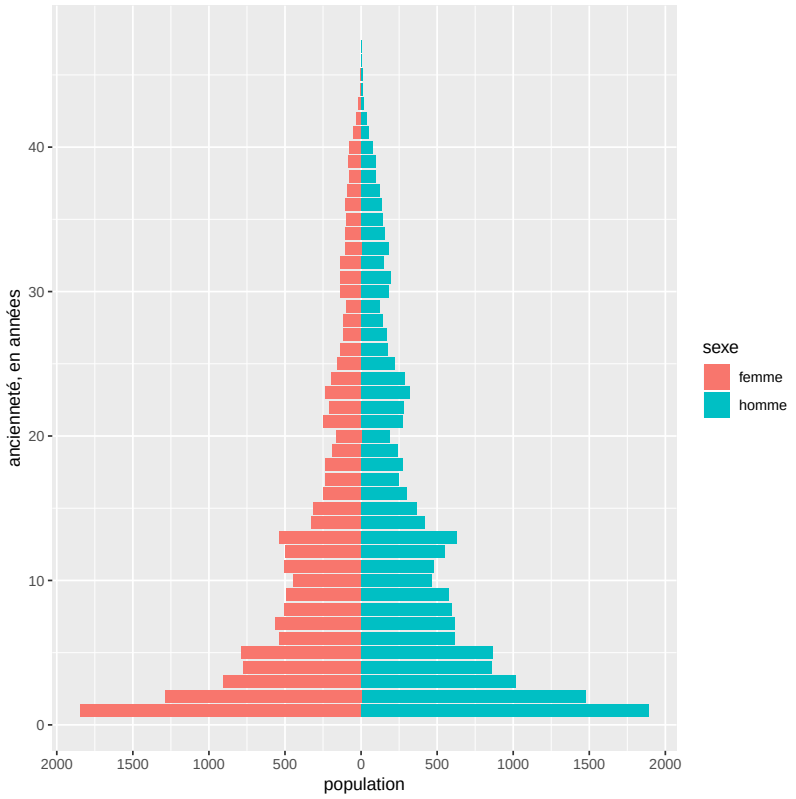


Figure 30 – Ancienneté selon le sexe

Le même constat peut être porté sur la distribution des heures travaillées par mois selon le sexe. Sachant que la durée légale de travail est de 35 heures par semaine, soit 150 heures par mois, on note que les femmes sont plus nombreuses à occuper des emplois à temps partiel (nombre d'heures travaillées inférieure à 150 par mois). Les hommes se situent essentiellement du côté haut de la pyramide. Enfin, on ne note pas d'enseignement intéressant tiré de l'étude du graphique 30, qui distribue l'ancienneté en années des hommes et des femmes.

Par la suite, le salaire mensuel est étudié selon les autres variables qualitatives.

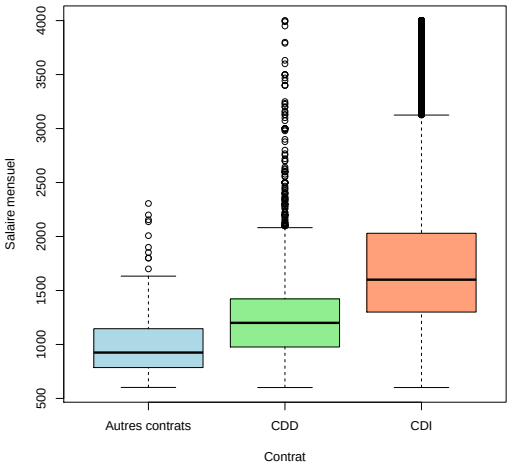


Figure 31 – Salaire selon le type de contrat

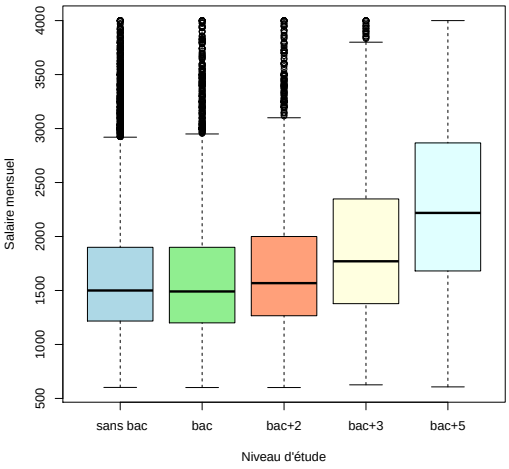


Figure 32 – Salaire selon les études effectuées

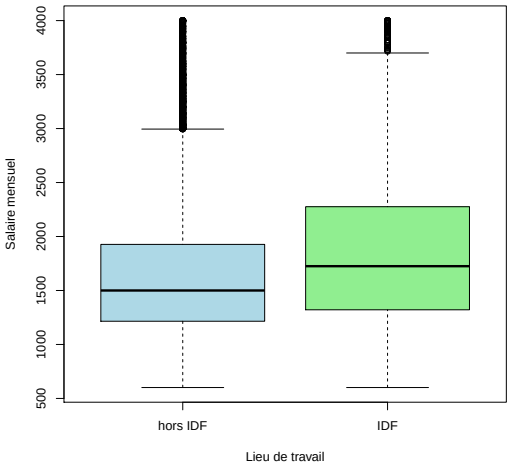


Figure 33 – Salaire selon le lieu de résidence

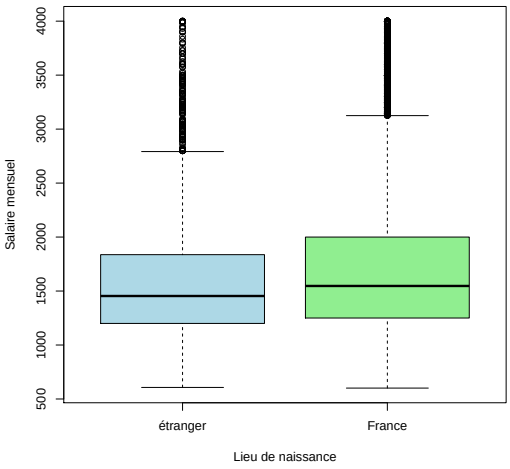


Figure 34 – Salaire selon le pays de naissance

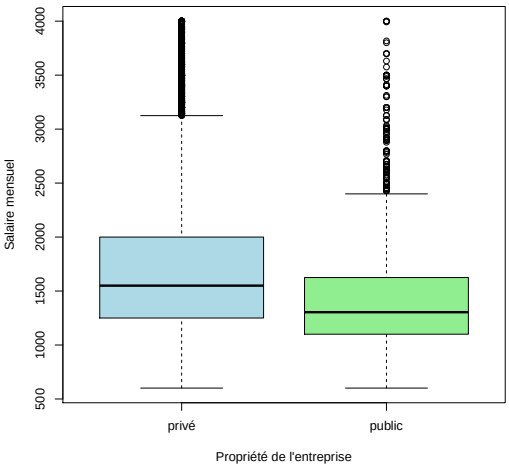


Figure 35 – Salaire selon le domaine

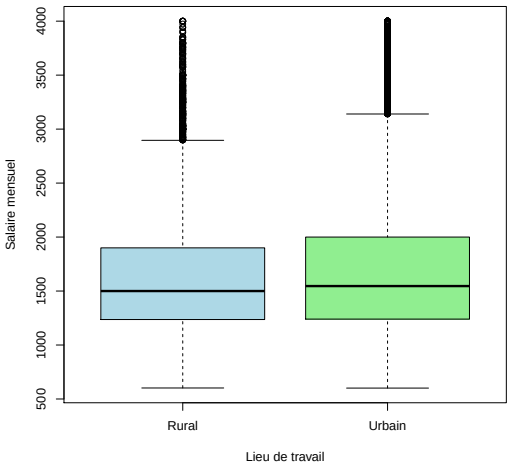
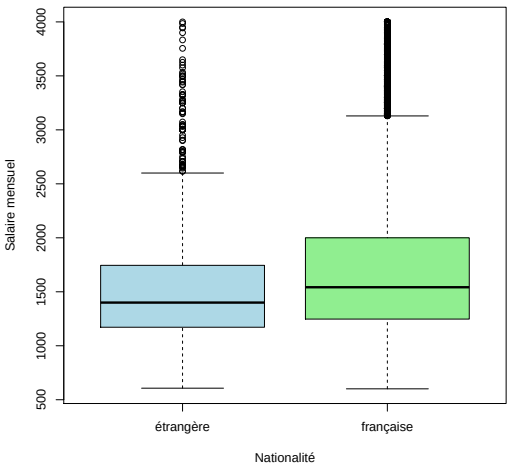


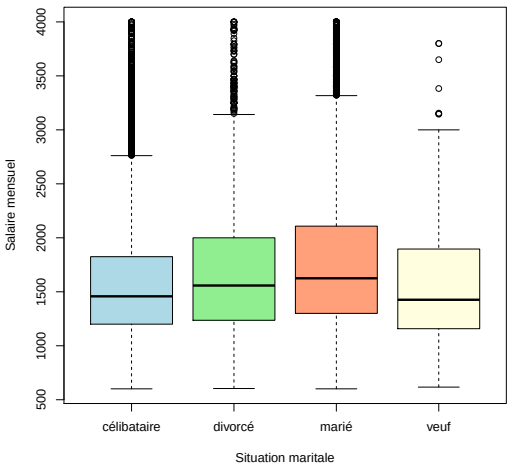
Figure 36 – Salaire selon le lieu



⇐

Figure 37 – Salaire selon la nationalité (gauche)

Figure 38 – Salaire selon la situation maritale (droite) ⇒



Selon le graphique 31, le type de contrat semble avoir une influence sur le salaire, un individu en CDI aura tendance à être mieux payé qu'un individu en CDD par exemple.
Par ailleurs, le graphique 32 montre qu'à partir d'un BAC+3, l'effet du diplôme se ressent davantage sur les salaires. Les salaires évoluent positivement avec le degré d'études.
Les résidents en Ile-de-France perçoivent en tendance, un salaire plus élevé que les autres, pour d'une part compenser le coût de la vie dans cette région et parce que les postes pourvus sont souvent plus sensibles qu'ailleurs (graphique 33).
La médiane des salaires est quasiment la même pour un travailleur né en France que pour un travailleur né à l'étranger, toutefois le troisième quantile est plus élevé pour le travailleur français (graphique 34). Cette remarque s'applique également pour la nationalité du travailleur (graphique 37).

D'après le graphique 35, la rémunération est en tendance, plus importante dans le secteur privé que dans le secteur public. 75% des salaires mensuels du secteur privé sont au-dessus d'environ 1 300 euros contre 50% dans le secteur public.
Le fait de vivre en milieu urbain ou rural n'a pas de grande différence en terme de rémunération pour un individu (graphique 36). De même pour la situation maritale (graphique 38).

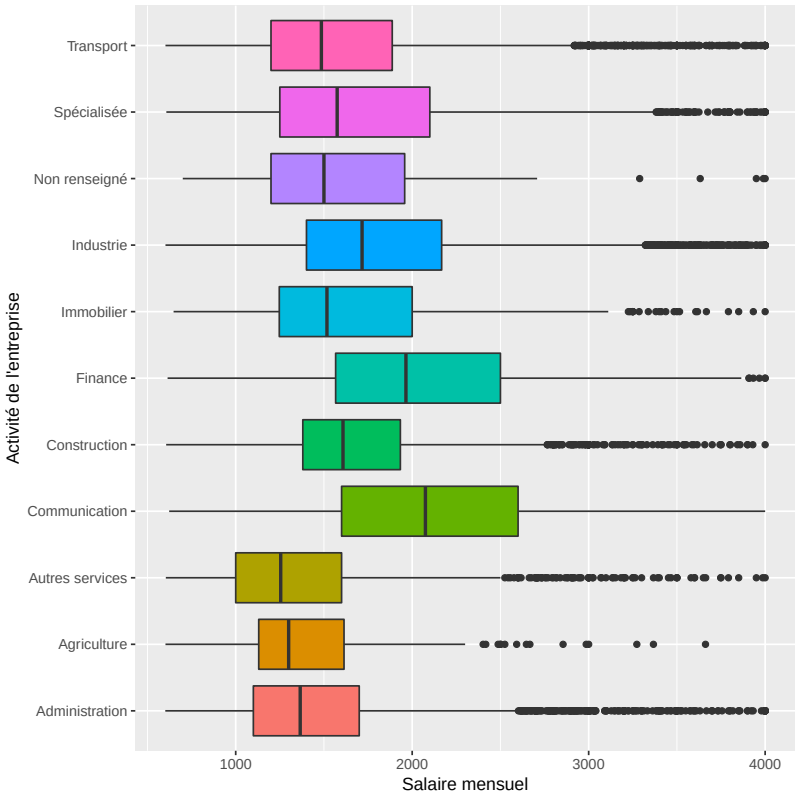


Figure 39 – Salaire selon l'activité

Enfin, d'après le graphique à droite, le secteur d'activité de l'entreprise joue un rôle sur la rémunération. En effet, Certains secteurs comme le secteur de la communication ou de la finance offrent globalement un salaire plus élevé que les secteurs tels que l'agriculture ou de l'administration par exemple.

6.3 Mesures de corrélations et de similarités

6.3.1 Corrélations entre variables continues

L'étude de la corrélation entre variables continues telles que `SALRED` ou `ANCENTR` peut se faire à travers le coefficient de corrélation de Pearson. On obtient alors les résultats suivants :

Figure 40 – Matrice des corrélations

Un coefficient positif (respectivement négatif) met en évidence une relation dans le même sens (respectivement dans le sens contraire) des deux variables. Plus ce coefficient est élevé (respectivement faible) en valeur absolue, plus l'intensité de la relation est forte (respectivement faible). Ainsi, on note qu'il existe peu de corrélation entre les variables continues. La corrélation de `0.6` entre `ANCENTR` et `AGE` est naturelle. Si on note que `SALRED` est corrélée positivement avec toutes les autres variables continues, il faut remarquer que `SALHR` n'est pas corrélée avec `AGE` et `ANCENTR`. De plus, `SALHR` est corrélée négativement avec `NBHEUR`.

6.3.2 Similarités entre variables dichotomiques

En ce qui concerne la « corrélation » entre variables binaires, il est davantage intéressant de s'intéresser à leur similarité. Comme pour le coefficient de corrélation, il existe plusieurs indicateurs. Ici, on prendra celui de Sokal-Michener [5].

On obtient alors cette matrice des similarités suivante (sur deux pages) :

Table 10 – Matrice des similarités

	SEXE	BAC	BAC2	BAC3	BAC5	NFR	LNAIS	PAYNEU27	TUU	IDF	MARRIED	DIVORCED	WIDOW	NBENF1	NBENF2	NBENF3PLUS
SEXE	1	0.37	0.38	0.37	0.36	0.35	0.35	0.35	0.32	0.36	0.33	0.38	0.37	0.37	0.35	0.35
BAC	0.37	1	0.43	0.52	0.52	0.56	0.53	0.55	0.24	0.5	0.32	0.54	0.59	0.48	0.49	0.56
BAC2	0.38	0.43	1	0.64	0.65	0.67	0.61	0.65	0.21	0.58	0.34	0.65	0.72	0.54	0.58	0.68
BAC3	0.37	0.52	0.64	1	0.78	**0.8**	0.73	0.77	0.19	0.67	0.34	0.75	**0.86**	0.6	0.64	0.79
BAC5	0.36	0.52	0.65	0.78	1	**0.81**	0.73	0.78	0.19	0.69	0.35	0.76	**0.87**	0.6	0.64	0.79
NFR	0.35	0.56	0.67	**0.8**	**0.81**	1	**0.89**	**0.87**	0.19	0.71	0.36	0.77	**0.88**	0.61	0.64	0.81
LNAIS	0.35	0.53	0.61	0.73	0.73	**0.89**	1	**0.94**	0.22	0.69	0.38	0.71	0.79	0.57	0.59	0.74
PAYNEU27	0.35	0.55	0.65	0.77	0.78	**0.87**	**0.94**	1	0.2	0.7	0.37	0.75	**0.84**	0.59	0.62	0.79
TUU	0.32	0.24	0.21	0.19	0.19	0.19	0.22	0.2	1	0.26	0.3	0.19	0.16	0.23	0.2	0.17
IDF	0.36	0.5	0.58	0.67	0.69	0.71	0.69	0.7	0.26	1	0.34	0.64	0.71	0.54	0.55	0.66
MARRIED	0.33	0.32	0.34	0.34	0.35	0.36	0.38	0.37	0.3	0.34	1	0.28	0.34	0.35	0.4	0.38
DIVORCED	0.38	0.54	0.65	0.75	0.76	0.77	0.71	0.75	0.19	0.64	0.28	1	**0.83**	0.58	0.6	0.76
WIDOW	0.37	0.59	0.72	**0.86**	**0.87**	**0.88**	0.79	**0.84**	0.16	0.71	0.34	**0.83**	1	0.64	0.67	0.86
NBENF1	0.37	0.48	0.54	0.6	0.6	0.61	0.57	0.59	0.23	0.54	0.35	0.58	0.64	1	0.43	0.57
NBENF2	0.35	0.49	0.58	0.64	0.64	0.64	0.59	0.62	0.2	0.55	0.4	0.6	0.67	0.43	1	0.6
NBENF3PLUS	0.35	0.56	0.68	0.79	0.79	**0.81**	0.74	0.79	0.17	0.66	0.38	0.76	**0.86**	0.57	0.6	1
PUBLIC	0.39	0.56	0.68	**0.8**	**0.81**	**0.81**	0.73	0.78	0.18	0.66	0.34	0.77	**0.87**	0.6	0.63	0.8
TPPRED	0.46	0.52	0.6	0.68	0.68	0.7	0.64	0.67	0.2	0.58	0.36	0.67	0.74	0.54	0.58	0.69
NUITC	0.3	0.51	0.59	0.67	0.68	0.69	0.63	0.67	0.2	0.58	0.34	0.66	0.73	0.54	0.57	0.69
CDD	0.38	0.55	0.64	0.75	0.76	0.76	0.7	0.74	0.19	0.63	0.32	0.72	**0.81**	0.58	0.6	0.75
AutreCDD	0.36	0.6	0.71	**0.85**	**0.85**	**0.86**	0.77	**0.83**	0.16	0.7	0.34	**0.81**	**0.94**	0.63	0.66	0.85
CSP_CADRE	0.34	0.5	0.64	0.74	**0.83**	0.71	0.65	0.69	0.21	0.65	0.36	0.69	0.77	0.56	0.6	0.71
CSP_INTERM	0.34	0.46	0.58	0.57	0.53	0.53	0.5	0.52	0.25	0.5	0.34	0.53	0.57	0.46	0.49	0.54
CSP_OUVRI	0.21	0.39	0.39	0.46	0.46	0.53	0.5	0.51	0.23	0.42	0.33	0.49	0.52	0.43	0.44	0.51
SECT_ENT_AZ	0.36	0.59	0.72	**0.86**	**0.86**	**0.88**	0.79	**0.84**	0.16	0.7	0.35	**0.83**	**0.95**	0.64	0.67	0.86
SECT_ENT_BE	0.29	0.45	0.53	0.58	0.59	0.59	0.55	0.57	0.21	0.49	0.36	0.58	0.63	0.49	0.51	0.6
SECT_ENT_FZ	0.3	0.54	0.63	0.74	0.75	0.78	0.71	0.74	0.18	0.63	0.35	0.72	**0.82**	0.58	0.61	0.76
SECT_ENT_GI	0.34	0.47	0.48	0.53	0.53	0.54	0.52	0.54	0.25	0.5	0.33	0.53	0.57	0.47	0.49	0.55
SECT_ENT_JZ	0.36	0.58	0.72	**0.85**	**0.86**	**0.85**	0.77	**0.82**	0.17	0.71	0.35	**0.81**	**0.93**	0.63	0.66	0.84
SECT_ENT_KZ	0.38	0.57	0.71	**0.83**	**0.83**	**0.83**	0.75	**0.8**	0.17	0.7	0.35	0.79	**0.91**	0.62	0.65	0.82
SECT_ENT_LZ	0.37	0.59	0.73	**0.86**	**0.87**	**0.88**	0.79	**0.84**	0.16	0.71	0.35	**0.83**	**0.95**	0.64	0.67	0.86
SECT_ENT_MN	0.37	0.54	0.65	0.75	0.77	0.76	0.7	0.74	0.2	0.65	0.34	0.72	**0.81**	0.58	0.61	0.75
SECT_ENT_RU	0.39	0.57	0.68	**0.8**	**0.8**	**0.82**	0.74	0.78	0.18	0.67	0.35	0.77	**0.87**	0.6	0.63	0.79

5. voir annexe 1 pour la formulation de cet indicateur

Les variables similaires entre elles sont mises en relief avec un fond orange (pour un seuil minimal arbitraire de `0.8`) et celles très similaires sont sur un fond rouge (pour un seuil minimal arbitraire de `0.9`). Cet indicateur présente la mauvaise propriété[6] de s'approcher facilement de `1` lorsque les deux vecteurs étudiés présentent très peu de valeur `1` (ici, c'est-à-dire que les deux variables présentent peu d'individus vérifiant la modalité). Ceci explique la forte similarité de `WIDOW` avec la majorité des autres variables.

PUBLIC	TPPRED	NUITC	CDD	AutreCDD	CSP_CADRE	CSP_INTERM	CSP_OUVRI	SECT_ENT_AZ	SECT_ENT_BE	SECT_ENT_FZ	SECT_ENT_GI	SECT_ENT_JZ	SECT_ENT_KZ	SECT_ENT_LZ	SECT_ENT_MN	SECT_ENT_RU	
0.39	0.46	0.3	0.38	0.36	0.34	0.34	0.21	0.36	0.29	0.3	0.34	0.36	0.38	0.37	0.37	0.39	SEXE
0.56	0.52	0.51	0.55	0.6	0.5	0.46	0.39	0.59	0.45	0.54	0.47	0.58	0.57	0.59	0.54	0.57	BAC
0.68	0.6	0.59	0.64	0.71	0.64	0.58	0.39	0.72	0.53	0.63	0.48	0.72	0.71	0.73	0.65	0.68	BAC2
0.8	0.68	0.67	0.75	**0.85**	0.74	0.57	0.46	**0.86**	0.58	0.74	0.53	**0.85**	**0.83**	**0.86**	0.75	**0.8**	BAC3
0.81	0.68	0.68	0.76	**0.85**	**0.83**	0.53	0.46	**0.86**	0.59	0.75	0.53	**0.86**	**0.83**	**0.87**	0.77	**0.8**	BAC5
0.81	0.7	0.69	0.76	**0.86**	0.71	0.53	0.53	**0.88**	0.59	0.78	0.54	**0.85**	**0.83**	**0.88**	0.76	**0.82**	NFR
0.73	0.64	0.63	0.7	0.77	0.65	0.5	0.5	0.79	0.55	0.71	0.52	0.77	0.75	0.79	0.7	0.74	LNAIS
0.78	0.67	0.67	0.74	**0.83**	0.69	0.52	0.51	**0.84**	0.57	0.74	0.54	**0.82**	**0.8**	**0.84**	0.74	0.78	PAYNEU27
0.18	0.2	0.2	0.19	0.16	0.21	0.25	0.23	0.16	0.21	0.18	0.25	0.17	0.17	0.16	0.2	0.18	TUU
0.66	0.58	0.58	0.63	0.7	0.65	0.5	0.42	0.7	0.49	0.63	0.5	0.71	0.7	0.71	0.65	0.67	IDF
0.34	0.36	0.34	0.32	0.34	0.36	0.34	0.33	0.35	0.36	0.35	0.33	0.35	0.35	0.35	0.34	0.35	MARRIED
0.77	0.67	0.66	0.72	**0.81**	0.69	0.53	0.49	**0.83**	0.58	0.72	0.53	**0.81**	0.79	**0.83**	0.72	0.77	DIVORCED
0.87	0.74	0.73	**0.81**	**0.94**	0.77	0.57	0.52	**0.95**	0.63	**0.82**	0.57	**0.93**	**0.91**	**0.95**	**0.81**	**0.87**	WIDOW
0.6	0.54	0.54	0.58	0.63	0.56	0.46	0.43	0.64	0.49	0.58	0.47	0.63	0.62	0.64	0.58	0.6	NBENF1
0.63	0.58	0.57	0.6	0.66	0.6	0.49	0.44	0.67	0.51	0.61	0.49	0.66	0.65	0.67	0.61	0.63	NBENF2
0.8	0.69	0.69	0.75	**0.85**	0.71	0.54	0.51	**0.86**	0.6	0.76	0.55	**0.84**	**0.82**	**0.86**	0.75	0.79	NBENF3PLUS
1	0.72	0.69	**0.85**	**0.86**	0.72	0.55	0.47	**0.87**	0.57	0.75	0.51	**0.84**	**0.82**	**0.87**	0.74	**0.8**	PUBLIC
0.72	1	0.58	0.69	0.73	0.61	0.48	0.42	0.74	0.49	0.64	0.5	0.72	0.71	0.74	0.65	0.72	TPPRED
0.69	0.58	1	0.65	0.72	0.61	0.5	0.51	0.73	0.58	0.64	0.51	0.72	0.69	0.73	0.64	0.68	NUITC
0.85	0.69	0.65	1	**0.8**	0.67	0.52	0.47	**0.81**	0.54	0.71	0.5	0.79	0.77	**0.81**	0.7	0.76	CDD
0.86	0.73	0.72	**0.8**	1	0.75	0.56	0.53	**0.94**	0.63	**0.81**	0.57	**0.91**	**0.89**	**0.94**	**0.8**	**0.86**	AutreCDD
0.72	0.61	0.61	0.67	0.75	1	0.44	0.4	0.76	0.55	0.67	0.49	0.78	0.76	0.77	0.7	0.71	CSP_CADRE
0.55	0.48	0.5	0.52	0.56	0.44	1	0.27	0.57	0.47	0.51	0.42	0.57	0.56	0.57	0.53	0.54	CSP_INTERM
0.47	0.42	0.51	0.47	0.53	0.4	0.27	1	0.54	0.54	0.57	0.42	0.5	0.48	0.52	0.47	0.48	CSP_OUVRI
0.87	0.74	0.73	**0.81**	**0.94**	0.76	0.57	0.54	1	0.62	**0.82**	0.57	**0.92**	**0.9**	**0.95**	**0.8**	**0.87**	SECT_ENT_AZ
0.57	0.49	0.58	0.54	0.63	0.55	0.47	0.54	0.62	1	0.53	0.35	0.61	0.59	0.63	0.52	0.57	SECT_ENT_BE
0.75	0.64	0.64	0.71	**0.81**	0.67	0.51	0.57	**0.82**	0.53	1	0.48	0.79	0.77	**0.82**	0.69	0.75	SECT_ENT_FZ
0.51	0.5	0.51	0.5	0.57	0.49	0.42	0.42	0.57	0.35	0.48	1	0.55	0.53	0.57	0.47	0.51	SECT_ENT_GI
0.84	0.72	0.72	0.79	**0.91**	0.78	0.57	0.5	**0.92**	0.61	0.79	0.55	1	**0.87**	**0.92**	0.78	**0.84**	SECT_ENT_JZ
0.82	0.71	0.69	0.77	**0.89**	0.76	0.56	0.48	**0.9**	0.59	0.77	0.53	**0.87**	1	**0.9**	0.76	**0.82**	SECT_ENT_KZ
0.87	0.74	0.73	**0.81**	**0.94**	0.77	0.57	0.52	**0.95**	0.63	**0.82**	0.57	**0.92**	**0.9**	1	**0.81**	**0.87**	SECT_ENT_LZ
0.74	0.65	0.64	0.7	**0.8**	0.7	0.53	0.47	**0.8**	0.52	0.69	0.47	0.78	0.76	**0.81**	1	0.74	SECT_ENT_MN
0.8	0.72	0.68	0.76	**0.86**	0.71	0.54	0.48	**0.87**	0.57	0.75	0.51	**0.84**	**0.82**	**0.87**	0.74	1	SECT_ENT_RU

6. liée directement à sa formule de construction (voir annexe 1).

Troisième partie

Résultats empiriques

7 Modèles estimés par MCO

L'estimation des effets des variables retenues sur le salaire (logarithme du salaire horaire) se fait à partir de modèles log-linéaires et à partir de la méthode des Moindres Carrés Ordinaires. L'estimation se fait en deux temps, un pour les hommes et l'autre pour les femmes.

7.1 Estimations pour les hommes

Le modèle à estimer chez les hommes est le suivant :

$$\begin{aligned}\text{SALHR\_LN}_i = {} & \beta_0 + \beta_1 \text{BAC}_i + \beta 2 \text{BAC2}_i + \beta_3 \text{BAC3}_i + \beta_4 \text{BAC5}_i + \beta_5 \text{AGE}_i + \beta_6 \text{AGE}_i^2 \\ & + \beta_7 \text{NFR}_i + \beta_8 \text{LNAIS}_i + \beta_9 \text{PAYNEU27}_i + \beta_{10} \text{TUU}_i + \beta_{11} \text{IDF}_i + \beta_{12} \text{MARRIED}_i \\ & + \beta_{13} \text{DIVORCED}_i + \beta_{14} \text{WIDOW}_i + \beta_{15} \text{PUBLIC}_i + \beta_{16} \text{NBHEUR}_i + \beta_{17} \text{NBHEUR}_i^2 \\ & + \beta_{18} \text{TPPRED}_i + \beta_{19} \text{NUITC}_i + \beta_{20} \text{ANCENTR}_i + \beta_{21} \text{ANCENTR}_i^2 + \beta_{22} \text{CDD}_i \\ & + \beta_{23} \text{AutreCDD}_i + \beta_{24} \text{CSP\_CADRE}_i + \beta_{25} \text{CSP\_INTERM}_i + \beta_{26} \text{CSP\_OUVRI}_i \\ & + \beta_{27} \text{SECT\_ENT\_AZ}_i + \beta_{28} \text{SECT\_ENT\_BE}_i + \beta_{29} \text{SECT\_ENT\_FZ}_i + \beta_{30} \text{SECT\_ENT\_GI}_i \\ & + \beta_{31} \text{SECT\_ENT\_JZ}_i + \beta_{32} \text{SECT\_ENT\_KZ}_i + \beta_{33} \text{SECT\_ENT\_LZ}_i + \beta_{34} \text{SECT\_ENT\_MN}_i \\ & + \beta_{35} \text{SECT\_ENT\_RU}_i + \varepsilon_i \qquad , i = 1, \ldots, n_H, \ \text{SEXE}_i = 0\end{aligned}$$

Ceci constitue un modèle log-linéaire. Il est composé 35 variables explicatives et de 36 coefficients à estimer. On supposera que les bruits ε_i sont distribués de façon gaussienne afin de pouvoir appliquer des tests statistiques ensuite.

Par une estimation par les moindres carrés ordinaires, on obtient les résultats présentés dans la table 11. Les coefficients ne seront pas dans un premier temps interprétés, il est fort probable, au vue des graphiques fournis dans la section 6, qu'un phénomène d'hétéroscédasticité soit présent dans les modèles. L'échantillon des hommes est constitué de 16 268 observations.

Table 11 – Résultats des MCO sur l'échantillon des hommes

variable	label	Estimation	Erreur-type	t-stat	p-value	signif [7]
(Intercept)	Constante du modèle	3.8198	0.0452	84.5441	< 0.0001	***
BAC	Possession d'un baccalauréat sans plus	0.0646	0.0057	11.3586	< 0.0001	***
BAC2	Possession d'un BAC+2 sans plus	0.1164	0.0075	15.5654	< 0.0001	***
BAC3	Possession d'un BAC+3 sans plus	0.1196	0.0108	11.1141	< 0.0001	***
BAC5	Possession d'un BAC+5 sans plus	0.2226	0.0111	19.9941	< 0.0001	***
AGE	Age de l'individu	0.0163	0.0016	10.2886	< 0.0001	***
AGE_SQUARE	Age au carré de l'individu	-1e-04	< 0.0001	-7.747	< 0.0001	***
NFR	... est de nationalité étrangère	-0.042	0.0124	-3.3973	7e-04	***
LNAIS	... est né(e) à l'étranger	0.0657	0.0146	4.5115	< 0.0001	***
PAYNEU27	... est né(e) en dehors de l'UE27	-0.1274	0.014	-9.1121	< 0.0001	***
TUU	... vit en commune urbaine	0.0077	0.0049	1.5786	0.1144	
IDF	... vit en Ile-de-France	0.0756	0.0061	12.3171	< 0.0001	***
MARRIED	... est marié(e)	0.0529	0.0052	10.162	< 0.0001	***
DIVORCED	... est divorcé(e)	0.032	0.0095	3.3463	8e-04	***
WIDOW	... est veuf(ve)	0.0434	0.0295	1.4703	0.1415	
NBENF1	... a exactement 1 enfant	–	–	–	–	
NBENF2	... a exactement 2 enfants	–	–	–	–	
NBENF3PLUS	... a 3 enfants ou plus	–	–	–	–	
PUBLIC	entreprise publique	0.0269	0.0147	1.8323	0.0669	.
NBHEUR	nombre d'heures de travail par mois	-0.0242	4e-04	-59.2323	< 0.0001	***
NBHEUR_SQUARE	NBHEUR au carré	1e-04	< 0.0001	46.2156	< 0.0001	***
TPPRED	... travaille en temps partiel	-0.4133	0.0129	-32.045	< 0.0001	***

7. Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 : correspond au seuil de significativité des différentes variables.

Table 11 – Résultats des MCO sur l'échantillon des hommes (suite)

ANCENTR	ancienneté en mois dans l'entreprise	8e-04	1e-04	11.984	< 0.0001	***
ANCENTR_SQUARE	ANCENTR au carré	< 0.0001	< 0.0001	-6.4607	< 0.0001	***
CDD	contrat de type CDD	-0.0994	0.0096	-10.3253	< 0.0001	***
AutreCDD	contrat de type autre que CDD hors CDI	-0.3502	0.0153	-22.9437	< 0.0001	***
CSP_AGRI	Agriculteur	–	–	–	–	
CSP_ARTI	Artisan	–	–	–	–	
CSP_CADRE	Cadre	0.5004	0.0098	51.1559	< 0.0001	***
CSP_INTERM	Profession intermédiaire	0.1806	0.0078	23.2683	< 0.0001	***
CSP_OUVRI	Ouvrier	0.0051	0.0076	0.6727	0.5011	
SECT_ENT_AZ	Agriculture, syviculture et pêche	0.0519	0.0192	2.7052	0.0068	**
SECT_ENT_BE	Industrie manufacturière, industries extractives et autres	0.1515	0.0101	14.971	< 0.0001	***
SECT_ENT_FZ	Construction	0.1488	0.011	13.4814	< 0.0001	***
SECT_ENT_GI	Commerce de gros et de détail, transports, hébergement et restauration	0.0914	0.01	9.1551	< 0.0001	***
SECT_ENT_JZ	Information et communication	0.0815	0.0148	5.5044	< 0.0001	***
SECT_ENT_KZ	Activités financières et d'assurance	0.2228	0.0159	14.0229	< 0.0001	***
SECT_ENT_LZ	Activités immobilières	0.0369	0.0236	1.5629	0.1181	
SECT_ENT_MN	Activités spécialisées, scientifiques, techniques et services administratifs	0.0756	0.0115	6.5881	< 0.0001	***
SECT_ENT_RU	Autres activités de services	0.029	0.0145	1.9929	0.0463	*
		# Obs	R^2	$\overline{R^2}$	F-test (global)	
		16 268	0.5721	0.5712	p-value $<$ 2.2e-16	

7.2 Estimations pour les femmes

Le modèle à estimer chez les femmes est le suivant :

$$\begin{aligned}\text{SALHR\_LN}_i = {} & \beta_0 + \beta_1 \text{BAC}_i + \beta 2 \text{BAC2}_i + \beta_3 \text{BAC3}_i + \beta_4 \text{BAC5}_i + \beta_5 \text{AGE}_i + \beta_6 \text{AGE}_i^2 \\ & + \beta_7 \text{NFR}_i + \beta_8 \text{LNAIS}_i + \beta_9 \text{PAYNEU27}_i + \beta_{10} \text{TUU}_i + \beta_{11} \text{IDF}_i + \beta_{12} \text{MARRIED}_i \\ & + \beta_{13} \text{DIVORCED}_i + \beta_{14} \text{WIDOW}_i + \beta_{15} \text{NBENF1} + \beta_{16} \text{NBENF2} + \beta_{17} \text{NBENF3PLUS} \\ & + \beta_{18} \text{PUBLIC}_i + \beta_{19} \text{NBHEUR}_i + \beta_{20} \text{NBHEUR}_i^2 + \beta_{21} \text{TPPRED}_i + \beta_{22} \text{NUITC}_i \\ & + \beta_{23} \text{ANCENTR}_i + \beta_{24} \text{ANCENTR}_i^2 + \beta_{25} \text{CDD}_i + \beta_{26} \text{AutreCDD}_i + \beta_{27} \text{CSP\_CADRE}_i \\ & + \beta_{28} \text{CSP\_INTERM}_i + \beta_{29} \text{CSP\_OUVRI}_i + \beta_{30} \text{SECT\_ENT\_AZ}_i + \beta_{31} \text{SECT\_ENT\_BE}_i \\ & + \beta_{32} \text{SECT\_ENT\_FZ}_i + \beta_{33} \text{SECT\_ENT\_GI}_i + \beta_{34} \text{SECT\_ENT\_JZ}_i + \beta_{35} \text{SECT\_ENT\_KZ}_i \\ & + \beta_{36} \text{SECT\_ENT\_LZ}_i + \beta_{37} \text{SECT\_ENT\_MN}_i + \beta_{38} \text{SECT\_ENT\_RU}_i + \varepsilon_i \\ & , i = 1, \ldots, n_F, \ \text{SEXE}_i = 1\end{aligned}$$

Ceci constitue un modèle log-linéaire. Il est composé 38 variables explicatives et de 39 coefficients à estimer. On supposera que les bruits ε_i sont distribués de façon gaussienne afin de pouvoir appliquer des tests statistiques ensuite.

Par une estimation par les moindres carrés ordinaires, on obtient les résultats présentés dans la table 12. Les coefficients ne seront pas dans un premier temps interprétés, il est fort probable, au vue des graphiques fournis dans la section 6, qu'un phénomène d'hétéroscédasticité soit présent dans les modèles.

L'échantillon des femmes est constitué de 14 131 observations.

Table 12 – Résultats des MCO sur l'échantillon des femmes

variable	label	Estimation	Erreur-type	t-stat	p-value	signif [8]
(Intercept)	Constante du modèle	3.05	0.0414	73.7247	< 0.0001	***
BAC	Possession d'un baccalauréat sans plus	0.0743	0.0056	13.2301	< 0.0001	***
BAC2	Possession d'un BAC+2 sans plus	0.155	0.007	22.264	< 0.0001	***
BAC3	Possession d'un BAC+3 sans plus	0.1625	0.0091	17.7882	< 0.0001	***
BAC5	Possession d'un BAC+5 sans plus	0.2219	0.0113	19.5784	< 0.0001	***
AGE	Age de l'individu	0.0138	0.0017	7.915	< 0.0001	***
AGE_SQUARE	Age au carré de l'individu	-1e-04	< 0.0001	-6.099	< 0.0001	***
NFR	... est de nationalité étrangère	0.0067	0.0137	0.4909	0.6235	
LNAIS	... est né(e) à l'étranger	-0.0124	0.0145	-0.8551	0.3925	
PAYNEU27	... est né(e) en dehors de l'UE27	-0.0438	0.0147	-2.9792	0.0029	**
TUU	... vit en commune urbaine	0.0094	0.005	1.8905	0.0587	.
IDF	... vit en Ile-de-France	0.0906	0.006	15.0109	< 0.0001	***
MARRIED	... est marié(e)	0.0073	0.0054	1.3666	0.1718	
DIVORCED	... est divorcé(e)	0.0189	0.008	2.3462	0.019	*
WIDOW	... est veuf(ve)	-7e-04	0.0166	-0.0445	0.9645	
NBENF1	... a exactement 1 enfant	0.0102	0.0057	1.7814	0.0749	.
NBENF2	... a exactement 2 enfants	0.0258	0.0067	3.8572	1e-04	***
NBENF3PLUS	... a 3 enfants ou plus	-0.001	0.0106	-0.0986	0.9215	
PUBLIC	entreprise publique	0.0159	0.0092	1.7331	0.0831	.
NBHEUR	nombre d'heures de travail par mois	-0.0139	4e-04	-38.4587	< 0.0001	***
NBHEUR_SQUARE	NBHEUR au carré	< 0.0001	< 0.0001	23.6311	< 0.0001	***
TPPRED	... travaille en temps partiel	-0.2174	0.0069	-31.6613	< 0.0001	***
NUITC	... travaille de nuit	0.0829	0.0081	10.1833	< 0.0001	***

8. Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 : correspond au seuil de significativité des différentes variables.

Table 12 – Résultats des MCO sur l'échantillon des femmes (suite)

ANCENTR	ancienneté en mois dans l'entreprise	9e-04	1e-04	13.882	< 0.0001	***
ANCENTR_SQUARE	ANCENTR au carré	< 0.0001	< 0.0001	-5.431	< 0.0001	***
CDD	contrat de type CDD	-0.0814	0.008	-10.1183	< 0.0001	***
AutreCDD	contrat de type autre que CDD hors CDI	-0.2518	0.0167	-15.1166	< 0.0001	***
CSP_AGRI	Agriculteur	–	–	–	–	
CSP_ARTI	Artisan	–	–	–	–	
CSP_CADRE	Cadre	0.4521	0.0088	51.1681	< 0.0001	***
CSP_INTERM	Profession intermédiaire	0.1608	0.0056	28.6525	< 0.0001	***
CSP_OUVRI	Ouvrier	-0.0492	0.0076	-6.4466	< 0.0001	***
SECT_ENT_AZ	Agriculture, syviculture et pêche	0.0119	0.0222	0.5362	0.5918	
SECT_ENT_BE	Industrie manufacturière, industries extractives et autres	0.1001	0.008	12.5894	< 0.0001	***
SECT_ENT_FZ	Construction	0.0794	0.015	5.3019	< 0.0001	***
SECT_ENT_GI	Commerce de gros et de détail, transports, hébergement et restauration	0.0214	0.0063	3.4075	7e-04	***
SECT_ENT_JZ	Information et communication	0.0702	0.0157	4.4632	< 0.0001	***
SECT_ENT_KZ	Activités financières et d'assurance	0.1177	0.0103	11.3701	< 0.0001	***
SECT_ENT_LZ	Activités immobilières	0.0253	0.0174	1.4574	0.145	
SECT_ENT_MN	Activités spécialisées, scientifiques, techniques et services administratifs	0.0529	0.0081	6.5305	< 0.0001	***
SECT_ENT_RU	Autres activités de services	-0.0321	0.0085	-3.7696	2e-04	***
		# Obs	R^2	$\overline{R^2}$	F-test (global)	
		14 131	0.5324	0.5312	p-value $<$ 2.2e-16	

7.3 Détection d'hétéroscédasticité

Les estimateurs MCO sont non biaisés et à variance minimale parmi les estimateurs linéaires selon Gauss-Markov (théorème) à condition que l'hypothèse d'**homoscédasticité** soit respectée. Celle-ci passe par la vérification de la propriété suivante :

$$\mathbb{E}(\varepsilon\varepsilon') = \sigma^2\mathbb{I}_n$$

avec ε les bruits du modèle, σ^2 le terme de variance et $\mathbb{I}_n$ la matrice identité de taille n. Si on est en présence d'**hétéroscédasticité**, alors on vérifie plutôt :

$$\mathbb{E}(\varepsilon\varepsilon') = \sigma^2\Omega, \qquad \Omega \neq \mathbb{I}_n$$

dans ce cas, la matrice Ω de taille $(n \times n)$ est différente de la matrice identité, donc la variance des bruits varie en fonction des variables explicatives du modèle.
Sans l'hypothèse d'homoscédasticité, les estimateurs MCO restent non biaisés mais ne sont plus efficients (variance non minimale). Par conséquent, les tests d'inférence de student sont erronés.
Le principal test statistique de détection lorsqu'on ne connaît, ni la source, ni la forme que prend l'hétéroscédasticité, est celui de Breusch-Pagan. Il confronte l'hypothèse nulle **H0** : $\mathbb{E}(\varepsilon\varepsilon') = \sigma^2\mathbb{I}_n$ à l'hypothèse alternative **H1** : $\mathbb{E}(\varepsilon\varepsilon') = \sigma^2\Omega, \quad \Omega \neq \mathbb{I}_n$. Si le test rejette l'hypothèse nulle pour un seuil de confiance de 5%, alors on peut affirmer qu'il y a un phénomène d'hétéroscédasticité dans le modèle.

```
R > bptest(MCO_hommes)

---

studentized Breusch-Pagan test

data:  MCO_hommes

BP = 1270.7, df = 35, p-value < 2.2e-16
```

```
R > bptest(MCO_femmes)

---

studentized Breusch-Pagan test

data:  MCO_femmes

BP = 1004.6, df = 38, p-value < 2.2e-16
```

L'application du test sur le logiciel `R` conduit à rejeter l'hypothèse nulle pour un seuil de confiance de 5% dans les deux modèles. Ceci note la présence d'hétéroscédasticité dans les données.
Plutôt qu'essayer de remédier au problème en passant par les **Moindres Carrés Quasi-**

Généralisées pour estimer une matrice des poids et appliquer la **transformation d'Aitken**[9], il a été préférable d'appliquer la **Correction de White** sur la matrice de variance-covariance des estimateurs MCO de façon qu'elle soit à nouveau conforme aux tests statistiques. Cette méthode ne change ni les estimations, ni le pouvoir explicatif du modèle : seule la matrice de variance-covariance des estimateurs change. Certaines variables auparavant significatives au modèle ne le seront plus, d'autres pourraient le devenir.

8 Estimations MCO avec correction de White

8.1 Estimations pour les hommes

Le modèle à estimer reste exactement le même. Les résultats sont présentés dans la table 13.
Sur les 35 variables explicatives considérées dans le modèle, 5 ne sont pas significatives au seuil de 5%. Ces variables sont `TUU` (commune urbaine ou rurale de résidence), `WIDOW` (statut de veuf), `PUBLIC` (entreprise publique), `SECT_ENT_LZ` (activités immobilières) et `SECT_ENT_RU` (autres activités de service).
On notera que les variables préalablement créées `CSP_AGRI` et `CSP_ARTI` ne sont pas considérées dans le modèle. Ceci s'explique par l'absence d'observations appartenant à ces CSP dans l'échantillon des hommes. De même, les variables sur le nombre d'enfants `NBENF1`, `NBENF2` et `NBENF3PLUS` ne sont pas considérées. Ce choix est justifié par l'impact négligeable d'un nouveau-né dans la famille du côté de l'homme sur sa carrière professionnelle, contrairement aux femmes (perte d'une année d'expérience ou de formation et/ou arbitrages à faire).
Ces variables non significatives sont compréhensibles. À l'instar de vivre ou non en région Parisienne, l'impact de vivre ou non en commune urbaine sur le salaire n'est pas direct. Bien souvent, les habitants ruraux se déplacent pour travailler en ville. Par ailleurs, l'impact d'être veuf sur le salaire perçu est délicat à analyser, par intuition, il n'y a pas de raison de connecter cette situation au salaire.

9. consistant à multiplier l'ensemble du modèle (variable à expliquée comme variables explicatives et bruits) par une matrice estimée $\widehat{\Omega}^{-0.5}$.

L'écart des salaires entre les fonctionnaires et le secteur privé, bien qu'il a longtemps existé, est de moins en moins visible. Les salaires du secteur privé ont réussi à faire concurrence au secteur public. Ceci pourrait justifier la non-significativité de la variable `PUBLIC`.
Enfin, les variables `SECT_ENT_LZ` et `SECT_ENT_RU` ne sont pas significatives. Ceci peut provenir du fait que le groupe de référence est celui des Administrations publiques, enseignement, santé humaine et action sociale. Ces trois modalités concernent des activités de services et il n'est pas étonnant qu'en passant d'un secteur de services de ces genres à un autre, les salaires rencontrés ne soient pas beaucoup différents.

* * *

Le modèle estimé se base sur 16 268 observations et présente un coefficient de détermination ajusté de 0.5712. Ainsi, 57.12% du salaire horaire en logarithme des hommes est expliqué par le modèle. La qualité d'ajustement est plutôt satisfaisante pour un agrégat économique à expliquer telle que le salaire horaire.
En ce qui concerne le signe pris par les estimations des coefficients significatifs, il est intéressant d'y passer un temps rapidement en comparant avec l'intuition économique *a priori*.
Les variables sur le diplôme ont toutes un coefficient positif. Le groupe de référence étant les non-bacheliers, l'intuition économique *a priori* est conforme. De plus, l'effet marginal est de plus en plus élevé selon le niveau d'études, passant de `0.0646` pour les `BAC` à `0.2226` pour les `BAC+5`. La théorie du capital humain développée par Gary Becker [CCZ14] est vérifiée ici. L'éducation constitue un investissement rationnel des individus en vue d'accroître le salaire espéré par ces derniers une fois sur le marché du travail.
La littérature économique admet fréquemment que l'âge a un effet parabolique par rapport au salaire. Ce dernier est augmente au départ d'une carrière et diminue au moment du déclin d'une carrière professionnelle habituelle. Cette situation est observable sur les résultats empiriques, bien que le coefficient associé à `AGE_SQUARE` est très faible (tout en restant significativement différent de 0).
Du côté des variables liées à la nationalité, le pays de naissance ou la région de résidence des indi-

vidus, elles présentent des signes coïncidant avec l'intuition économique que chacun pourrait avoir, à l'exception de la variable `LNAIS`. Elle est associée à un coefficient positif, induisant l'idée qu'être né à l'étranger a un effet positif sur le salaire en France, chez les hommes. Une intuition *a posteriori* serait que la France est aux premiers rangs sur certains domaines de pointe (aéronautique, aéronavale, spatial, etc.), qui nécessitent une main d'œuvre très qualifiée, avec des salaires à hauteur des compétences exigées. Bien souvent, les étrangers répondent à l'appel pour occuper ces postes. En revanche, d'autres secteurs peu rémunérateurs tels que le secteur du bâtiment ont un fort taux d'emploi d'étrangers (bien qu'il existe un marché du travail parallèle et non déclaré dans ces secteurs).

Les variables liées au statut matrimonial ont des coefficients positifs. Ceci est intuitif, mais la question du sens du lien de causalité est à étudier. L'intuition voudrait qu'un homme se marie parce qu'il a une situation financière correcte et non le sens contraire.

Le salaire horaire décroit selon le nombre d'heures travaillées (`NBHEUR`) selon les résultats. Les personnes qui travaillent davantage d'heures le font souvent par nécessité. En occupant des postes laborieux ou en accumulant les emplois, ces travailleurs gagnent moins en moyenne à l'heure.

Les variables liées à la caractérisation du travail (`TPPRED`, `NUITC`, `CDD` et `AutreCDD`) présentent des coefficients à signes concordant avec l'intuition économique. Brièvement, l'un gagne moins en ayant un travail à temps partiel, gagne plus pour travailler de nuit et gagne moins avec un CDD plutôt qu'un CDI. L'ancienneté joue, sans surprise, positivement sur le salaire perçu.

Par ailleurs, la « *hiérarchie des CSP* » se vérifie ici. Le salaire moyen d'un ouvrier n'est pas significativement différent de celui d'un employé mais les professions intermédiaires ont tendance à être mieux rémunérées. Ceci est davantage vrai pour les Cadres. Cette augmentation graduée des coefficients se conforte à la réalité économique.

Enfin, selon le secteur d'activité de l'entreprise, l'effet sur le salaire se quantifie différemment. On note que tous les coefficients sont positifs, bien que deux variables sont non significatives. De plus, c'est le secteur des activités financières et d'assurance qui a l'impact le plus élevé sur les salaires.

Pour conclure sur l'analyse de ces résultats empiriques, il est important de rappeler que le modèle prédit est de la forme log-linéaire. Ceci induit une interprétation des coefficients en tant que semi-élasticités. Pour une valeur de $\widehat{\beta}_h$ donnée, une augmentation de 1 unité de la variable X_h entraîne

une variation de Y (variable expliquée) de $100 * \widehat{\beta}_h\%$, *ceteris paribus*.

Cette interprétation est valable pour les variables « continues » telles que `AGE`, `NBHEUR` ou `ANCENTR`.

Les coefficients associés aux variables élevées au carré s'interprètent d'une façon particulière. Le coefficient estimé correspond à la variation marginale en pourcentage comme précédemment, *ceteris paribus*, mais pour un niveau donné de la variable.

Enfin, en ce qui concerne l'interprétation des coefficients associés à des variables indicatrices, on a une variation de $100 * \widehat{\beta}_h\%$ de la variable expliquée lorsque l'individu prend la valeur `1` sur la variable (c'est-à-dire lorsqu'il est concerné par la modalité représentée par cette variable), *ceteris paribus*.

À partir de là, tous les coefficients dans la table 11 peuvent être interprétés, mais ce n'est pas l'objet de l'étude.

L'`erreur-type` correspond à l'écart-type de l'estimation. La `statistique` t est la statistique de test de Student et est égale à :

$$t = \frac{\widehat{\beta}}{\widehat{\sigma}_\beta} = \frac{\text{Estimation}}{\text{Erreur-type}} \qquad \text{sous H0}$$

Le test de Student permet de déterminer la significativité individuelle des coefficients. Il prend pour hypothèse nulle **H0** : $\beta = 0$ contre **H1** : $\beta \neq 0$. Selon si la statistique de test tombe ou non dans la région critique, on rejettera ou non l'hypothèse nulle. Une p-value inférieure au seuil de confiance de 5% permet de rejeter l'hypothèse nulle (pour une confiance de 5%) et donc d'affirmer que le coefficient est significativement différent de 0 (hypothèse alternative). Le codage en étoiles proposé par le logiciel `R` facilite l'identification de variables non significatives au modèle.

Enfin, le test de significativité globale de Fisher permet de savoir si au moins une variable est globalement significative au modèle. L'hypothèse nulle **H0** correspond à la situation où tous les coefficients, à l'exception de la constante du modèle sont simultanément non significativement différents à 0 (comprendre qu'aucune variable considérée n'apporte de l'information au modèle). L'hypothèse alternative **H1** correspond à la situation où au moins l'un des coefficients (toujours à l'exception de la constante) est significativement différent de 0. La `p-value` de ce test est donnée en dernier dans la table 11. Celle-ci est inférieure à 5% donc le modèle est globalement significatif.

Table 13 – Résultats des MCO corrigés de White sur l'échantillon des hommes

variable	label	Estimation	Erreur-type	t-stat	p-value	signif [10]
(Intercept)	Constante du modèle	3.8198	0.1133	33.7265	< 0.0001	***
BAC	Possession d'un baccalauréat sans plus	0.0646	0.0054	11.9168	< 0.0001	***
BAC2	Possession d'un BAC+2 sans plus	0.1164	0.0077	15.1692	< 0.0001	***
BAC3	Possession d'un BAC+3 sans plus	0.1196	0.0123	9.7229	< 0.0001	***
BAC5	Possession d'un BAC+5 sans plus	0.2226	0.0143	15.5652	< 0.0001	***
AGE	Age de l'individu	0.0163	0.0018	9.2267	< 0.0001	***
AGE_SQUARE	Age au carré de l'individu	-1e-04	< 0.0001	-6.6178	< 0.0001	***
NFR	... est de nationalité étrangère	-0.042	0.0135	-3.1153	0.0018	**
LNAIS	... est né(e) à l'étranger	0.0657	0.0162	4.0554	1e-04	***
PAYNEU27	... est né(e) en dehors de l'UE27	-0.1274	0.0154	-8.2713	< 0.0001	***
TUU	... vit en commune urbaine	0.0077	0.0047	1.6314	0.1028	
IDF	... vit en Ile-de-France	0.0756	0.0066	11.4485	< 0.0001	***
MARRIED	... est marié(e)	0.0529	0.0052	10.1985	< 0.0001	***
DIVORCED	... est divorcé(e)	0.032	0.0094	3.4098	7e-04	***
WIDOW	... est veuf(ve)	0.0434	0.0324	1.34	0.1803	
NBENF1	... a exactement 1 enfant	–	–	–	–	
NBENF2	... a exactement 2 enfants	–	–	–	–	
NBENF3PLUS	... a 3 enfants ou plus	–	–	–	–	
PUBLIC	entreprise publique	0.0269	0.017	1.5851	0.113	
NBHEUR	nombre d'heures de travail par mois	-0.0242	0.0013	-18.2158	< 0.0001	***
NBHEUR_SQUARE	NBHEUR au carré	1e-04	< 0.0001	15.5881	< 0.0001	***
TPPRED	... travaille en temps partiel	-0.4133	0.0224	-18.4766	< 0.0001	***
NUITC	... travaille de nuit	0.0886	0.0054	16.4914	< 0.0001	***

10. Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 : correspond au seuil de significativité des différentes variables.

Table 13 – Résultats des MCO corrigés de White sur l'échantillon des hommes (suite)

ANCENTR	ancienneté en mois dans l'entreprise	8e-04	1e-04	11.687	< 0.0001	***
ANCENTR_SQUARE	ANCENTR au carré	< 0.0001	< 0.0001	-6.2287	< 0.0001	***
CDD	contrat de type CDD	-0.0994	0.0105	-9.4628	< 0.0001	***
AutreCDD	contrat de type autre que CDD hors CDI	-0.3502	0.0152	-23.0341	< 0.0001	***
CSP_AGRI	Agriculteur	–	–	–	–	
CSP_ARTI	Artisan	–	–	–	–	
CSP_CADRE	Cadre	0.5004	0.0109	45.8134	< 0.0001	***
CSP_INTERM	Profession intermédiaire	0.1806	0.0074	24.4956	< 0.0001	***
CSP_OUVRI	Ouvrier	0.0051	0.0071	0.7181	0.4727	
SECT_ENT_AZ	Agriculture, syviculture et pêche	0.0519	0.0203	2.5581	0.0105	*
SECT_ENT_BE	Industrie manufacturière, industries extractives et autres	0.1515	0.0117	12.9995	< 0.0001	***
SECT_ENT_FZ	Construction	0.1488	0.0122	12.1797	< 0.0001	***
SECT_ENT_GI	Commerce de gros et de détail, transports, hébergement et restauration	0.0914	0.0116	7.8727	< 0.0001	***
SECT_ENT_JZ	Information et communication	0.0815	0.0169	4.8321	< 0.0001	***
SECT_ENT_KZ	Activités financières et d'assurance	0.2228	0.0201	11.1089	< 0.0001	***
SECT_ENT_LZ	Activités immobilières	0.0369	0.0251	1.4712	0.1413	
SECT_ENT_MN	Activités spécialisées, scientifiques, techniques et services administratifs	0.0756	0.0131	5.7535	< 0.0001	***
SECT_ENT_RU	Autres activités de services	0.029	0.018	1.609	0.1076	
		# Obs	R^2	$\overline{R^2}$	F-test (global)	
		16 268	0.5721	0.5712	p-value $< 2.2e\text{-}16$	

8.2 Estimations pour les femmes

Le modèle à estimer reste exactement le même. Les résultats sont présentés dans la table 14.
Celui-ci est augmenté de trois nouvelles variables : `NBENF1`, `NBENF2` et `NBENF3PLUS` qui désignent respectivement la prise en charge d'exactement 1 enfant, 2 enfants ou 3 enfants et plus. Le nombre d'enfants a un impact direct sur le salaire des femmes. Ces dernières doivent généralement s'occuper d'eux, en laissant potentiellement de côté certains aspects de leur carrière professionnelle (réduction du temps de travail, congé maternité, augmentation de l'incertitude pour l'employeur, etc.).
On notera à nouveau que les variables `CSP_AGRI` et `CSP_ARTI` ne sont pas incluses dans le modèle, pour les mêmes raisons énoncées dans le modèle des hommes (section 8.1).
Ainsi, ce sont au total 38 variables explicatives considérées dans l'équation à estimer chez les femmes. 9 coefficients ne sont pas significativement différents de 0 pour un seuil de confiance de 5%. Ces variables sont : `NFR`, `LNAIS`, `TUU`, `MARRIED`, `WIDOW`, `NBENF1`, `NBENF3PLUS`, `PUBLIC`, `SECT_ENT_AZ` et `SECT_ENT_LZ`. Cette liste se rapproche beaucoup de celle des hommes. L'ajout des variables liées au nombre d'enfant est mitigé pour l'information apportée, bien que deux sur les trois variables soient significatives pour un seuil de 10%.
Le coefficient de détermination ajusté de l'ajustement est de 0.5312. Par conséquent, 53.12% du salaire horaire en logarithme des femmes est expliqué par les variables retenues dans ce modèle. La qualité d'ajustement est légèrement inférieure à celle des hommes mais reste au-dessus de 50%.
L'échantillon analysé est construit à partir de 14 131 femmes.
Le test de significativité globale de Fisher aboutit à un rejet de l'hypothèse nulle. Au moins un coefficient du modèle, à l'exception de la constante, est simultanément [11] différent de 0 et le modèle est globalement significatif.
L'intuition économique derrière chaque signe des coefficients est donnée dans la section précédente (section 8.1). On passera rapidement sur les résultats en notant uniquement des points importants.
Dans un premier temps, on retrouve l'augmentation séquencée de l'effet du diplôme sur les salaires : on passe de `0.0743` pour un `BAC` à `0.2219` pour un `BAC+5` en passant par `0.155` et `0.1625` pour respectivement les `BAC+2` et `BAC+3`. Ensuite, en ce qui concerne les variables sur l'âge et les aspects

11. c'est-à-dire qu'en présence des autres variables

spatiaux de l'individu, on retrouve les mêmes enseignements (à l'exception des variables (`NFR` et `LNAIS`) sur les origines qui sont devenues non significatives).

D'autre part, la logique des CSP est à nouveau vérifiée ici. On note que le coefficient lié à la variable `CSP_OUVRI` est passé au négatif, les femmes ouvrières ont tendance à gagner moins que les femmes employées, *ceteris paribus*. Les conclusions (pour les hommes) sur le secteur d'activité s'appliquent ici.

Enfin, si on s'intéresse uniquement aux variables significatives aux modèles (hommes et femmes), on note que les coefficients associés sont tous du même signe, d'une régression à une autre. De plus, les valeurs des estimations sont relativement proches.

* * *

Si on constate que les modèles estimés ont des coefficients assez proches, on pourrait penser que la discrimination envers les femmes est faible. Cependant, ces méthodes de régression ne prennent pas en compte l'effet de différence structurelle au sein des groupes. Par exemple, le salaire moyen des femmes pourrait être tiré vers le bas parce que les femmes ouvrières sont en nombre important dans l'échantillon. À l'inverse, le salaire moyen des hommes pourrait être tiré vers le haut parce que les hommes cadres travaillant dans le secteur financier sont en grand nombre dans l'échantillon. Tout l'intérêt de l'étude est de savoir si une femme ayant les mêmes caractéristiques individuelles (caractéristiques personnelles, familiales, professionnelles, poste, entreprise, etc.) qu'un homme gagnera en moyenne moins que ce dernier.

La méthode de décomposition d'Oaxaca permet d'évincer ces effets structurels liés aux différences individuelles entre les deux groupes. Le résidu après écartement de ces effets correspondra à une estimation de la part de différence de salaire liée à la discrimination directe [12].

L'ensemble des résultats de décomposition est détaillé dans la section suivante (section 9). Ces résultats seront ensuite étudiés pour en tirer des enseignements et conclure l'étude.

12. le qualificatif « directe » est important, ceci constitue une limite à la méthode employée. Ce point sera détaillé en conclusion de l'étude.

Table 14 – Résultats des MCO corrigés de White sur l'échantillon des femmes

variable	label	Estimation	Erreur-type	t-stat	p-value	signif [13]
(Intercept)	Constante du modèle	3.05	0.0915	33.3371	< 0.0001	***
BAC	Possession d'un baccalauréat sans plus	0.0743	0.0054	13.6698	< 0.0001	***
BAC2	Possession d'un BAC+2 sans plus	0.155	0.007	22.1751	< 0.0001	***
BAC3	Possession d'un BAC+3 sans plus	0.1625	0.0095	17.0999	< 0.0001	***
BAC5	Possession d'un BAC+5 sans plus	0.2219	0.0131	16.8749	< 0.0001	***
AGE	Age de l'individu	0.0138	0.0018	7.4687	< 0.0001	***
AGE_SQUARE	Age au carré de l'individu	-1e-04	< 0.0001	-5.621	< 0.0001	***
NFR	... est de nationalité étrangère	0.0067	0.0141	0.4759	0.6342	
LNAIS	... est né(e) à l'étranger	-0.0124	0.0149	-0.8301	0.4065	
PAYNEU27	... est né(e) en dehors de l'UE27	-0.0438	0.0157	-2.7899	0.0053	**
TUU	... vit en commune urbaine	0.0094	0.0049	1.9258	0.0542	.
IDF	... vit en Ile-de-France	0.0906	0.0066	13.772	< 0.0001	***
MARRIED	... est marié(e)	0.0073	0.0052	1.4239	0.1545	
DIVORCED	... est divorcé(e)	0.0189	0.0081	2.32	0.0204	*
WIDOW	... est veuf(ve)	-7e-04	0.0169	-0.0435	0.9653	
NBENF1	... a exactement 1 enfant	0.0102	0.0057	1.8087	0.0705	.
NBENF2	... a exactement 2 enfants	0.0258	0.0067	3.8586	1e-04	***
NBENF3PLUS	... a 3 enfants ou plus	-0.001	0.0112	-0.0926	0.9262	
PUBLIC	entreprise publique	0.0159	0.0094	1.6948	0.0901	.
NBHEUR	nombre d'heures de travail par mois	-0.0139	0.0011	-12.6807	< 0.0001	***
NBHEUR_SQUARE	NBHEUR au carré	< 0.0001	< 0.0001	8.5026	< 0.0001	***
TPPRED	... travaille en temps partiel	-0.2174	0.0114	-19.1193	< 0.0001	***
NUITC	... travaille de nuit	0.0829	0.0082	10.1087	< 0.0001	***

13. Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 : correspond au seuil de significativité des différentes variables.

Table 14 – Résultats des MCO corrigés de White sur l'échantillon des femmes (suite)

ANCENTR	ancienneté en mois dans l'entreprise	9e-04	1e-04	13.2252	< 0.0001	***
ANCENTR_SQUARE	ANCENTR au carré	< 0.0001	< 0.0001	-5.1809	< 0.0001	***
CDD	contrat de type CDD	-0.0814	0.0081	-10.0914	< 0.0001	***
AutreCDD	contrat de type autre que CDD hors CDI	-0.2518	0.0178	-14.1687	< 0.0001	***
CSP_AGRI	Agriculteur	–	–	–	–	
CSP_ARTI	Artisan	–	–	–	–	
CSP_CADRE	Cadre	0.4521	0.0105	42.9836	< 0.0001	***
CSP_INTERM	Profession intermédiaire	0.1608	0.0056	28.8787	< 0.0001	***
CSP_OUVRI	Ouvrier	-0.0492	0.0073	-6.7678	< 0.0001	***
SECT_ENT_AZ	Agriculture, syviculture et pêche	0.0119	0.0213	0.5589	0.5762	
SECT_ENT_BE	Industrie manufacturière, industries extractives et autres	0.1001	0.0084	11.8835	< 0.0001	***
SECT_ENT_FZ	Construction	0.0794	0.0153	5.2042	< 0.0001	***
SECT_ENT_GI	Commerce de gros et de détail, transports, hébergement et restauration	0.0214	0.0062	3.4692	5e-04	***
SECT_ENT_JZ	Information et communication	0.0702	0.0156	4.4919	< 0.0001	***
SECT_ENT_KZ	Activités financières et d'assurance	0.1177	0.0104	11.3225	< 0.0001	***
SECT_ENT_LZ	Activités immobilières	0.0253	0.0188	1.3452	0.1786	
SECT_ENT_MN	Activités spécialisées, scientifiques, techniques et services administratifs	0.0529	0.0093	5.7162	< 0.0001	***
SECT_ENT_RU	Autres activités de services	-0.0321	0.0084	-3.8028	1e-04	***
		# Obs	R^2	$\overline{R^2}$	F-test (global)	
		14 131	0.5324	0.5312	p-value $<$ 2.2e-16	

9 Décomposition d'Oaxaca

La méthode de décomposition a nécessité l'estimation des coefficients du modèle, regroupés dans les tables 13 et 14. Pour rappel, les équations à estimer sont :

$$\widehat{\ln(D+1)} = \ln(G+1) - \Delta\overline{X}'\widehat{\beta}_f = \ln(\overline{W}_m) - \ln(\overline{W}_f) - \Delta\overline{X}'\widehat{\beta}_f \tag{17}$$

et

$$\widehat{\ln(D+1)} = \ln(G+1) - \Delta\overline{X}'\widehat{\beta}_m = \ln(\overline{W}_m) - \ln(\overline{W}_f) - \Delta\overline{X}'\widehat{\beta}_m \tag{18}$$

Pour rappel également, D est le **coefficient de discrimination** posé en section 2. De plus, $\Delta\overline{X}' \equiv \overline{X}'_m - \overline{X}'_f$ et $\ln(\overline{W}_m)$ est la moyenne des logarithmes des salaires horaires chez les hommes. $\overline{X}'_m$ et $\overline{X}'_f$ sont respectivement les vecteurs moyens des variables explicatives pour les hommes et les femmes. Enfin, $\widehat{\beta}_m$ et $\widehat{\beta}_f$ sont respectivement les vecteurs des coefficients estimés du modèle pour les hommes et pour les femmes.
La table 15 de la sous-section 9.1 regroupe tous ces éléments et décompose la partie (en absolu et en pourcentage) de la différence salariale expliquée, variable par variable.

9.1 Tableau récapitulatif

Dans la table 15, les valeurs des coefficients et des moyennes sont arrondies à 10^{-4}. Les pourcentages sont arrondis à 10^{-2}. Les colonnes (1) et (3) correspondent aux estimations MCO obtenues dans les sections 8.1 et 8.2. Les colonnes (2) et (4) sont constituées des moyennes respectivement des hommes et des femmes des variables explicatives considérées dans les modèles. Les colonnes (5) et (6) constituent l'étape intermédiaire de l'estimation de la discrimination. Elles seront ôtées à l'écart brut du salaire horaire en logarithme. Ce dernier est indiqué en première ligne du tableau. Il est égal à l'écart entre le salaire horaire moyen en logarithme moyen des hommes moins le salaire horaire moyen en logarithme moyen des femmes.
La dernière ligne du tableau donne les estimations des $\widehat{\ln(D+1)}$ que l'on cherche à déterminer depuis le départ. Les colonnes (7) et (8) indiquent le pourcentage restant après éviction des effets des différences de caractéristiques individuelles entre les groupes.

Table 15 – Tableau récapitulatif de la décomposition d'Oaxaca (hommes et femmes)

variable	(1) $\widehat{\beta}_h$	(2) $\overline{x}_h$	(3) $\widehat{\beta}_f$	(4) $\overline{x}_f$	(5) $(\overline{x}_h - \overline{x}_f)\widehat{\beta}_f$	(6) $(\overline{x}_h - \overline{x}_f)\widehat{\beta}_h$	(7) $\%_f$ [a]	(8) $\%_m$ [b]
Écart SALHR_LN	–	**0.1322**	–	**0.1322**	–	–	**100**	**100**
BAC	0.0646	0.2273	0.0743	0.2796	-0.0039	-0.0034	+2.94	+2.56
BAC2	0.1164	0.1279	0.155	0.1778	-0.0077	-0.0058	+5.84	+4.39
BAC3	0.1196	0.0517	0.1625	0.0808	-0.0047	-0.0035	+3.58	+2.63
BAC5	0.2226	0.0631	0.2219	0.0587	0.001	0.001	-0.75	-0.75
AGE	0.0163	40.6675	0.0138	41.2042	-0.0074	-0.0088	+5.6	+6.62
AGE_SQUARE	-1e-04	1779.9245	-1e-04	1818.418	0.005	0.0057	-3.79	-4.3
NFR	-0.042	0.0642	0.0067	0.0441	1e-04	-8e-04	-0.1	+0.64
LNAIS	0.0657	0.1164	-0.0124	0.1013	-2e-04	0.001	+0.14	-0.75
PAYNEU27	-0.1274	0.0829	-0.0438	0.0684	-6e-04	-0.0018	+0.48	+1.39
TUU	0.0077	0.7234	0.0094	0.7336	-1e-04	-1e-04	+0.07	+0.06
IDF	0.0756	0.1585	0.0906	0.1673	-8e-04	-7e-04	+0.6	+0.5
MARRIED	0.0529	0.4859	0.0073	0.4748	1e-04	6e-04	-0.06	-0.44
DIVORCED	0.032	0.0628	0.0189	0.1065	-8e-04	-0.0014	+0.62	+1.06
WIDOW	0.0434	0.0052	-7e-04	0.0183	< 0.0001	-6e-04	-0.01	+0.43
NBENF1	–	0.199	0.0102	0.2303	-3e-04	–	+0.24	–
NBENF2	–	0.1871	0.0258	0.1846	1e-04	–	-0.05	–
NBENF3PLUS	–	0.0746	-0.001	0.0502	< 0.0001	–	+0.02	–
PUBLIC	0.0269	0.0337	0.0159	0.086	-8e-04	-0.0014	+0.63	+1.06
NBHEUR	-0.0242	153.331	-0.0139	141.3335	-0.1672	-0.2905	+126.42	+219.71
NBHEUR_SQUARE	1e-04	23963.4615	< 0.0001	20775.58	0.1044	0.2105	-78.97	-159.21
TPPRED	-0.4133	0.0405	-0.2174	0.2579	0.0473	0.0899	-35.75	-67.96

[a] pourcentage de la colonne (5) sur la différence de salaire (indiquée en première ligne)

[b] pourcentage de la colonne (6) sur la différence de salaire (indiquée en première ligne)

Table 15 – Tableau récapitulatif de la décomposition d'Oaxaca (hommes et femmes) (suite)

variable	(1) $\widehat{\beta}_h$	(2) $\overline{x}_h$	(3) $\widehat{\beta}_f$	(4) $\overline{x}_f$	(5) $(\overline{x}_h - \overline{x}_f)\widehat{\beta}_f$	(6) $(\overline{x}_h - \overline{x}_f)\widehat{\beta}_h$	(7) $\%_f$[a]	(8) $\%_m$[b]
NUITC	0.0886	0.2044	0.0829	0.0744	0.0108	0.0115	-8.15	-8.71
ANCENTR	8e-04	135.416	9e-04	127.1401	0.0073	0.0063	-5.54	-4.75
ANCENTR_SQUARE	< 0.0001	34192.831	< 0.0001	31109.6628	-0.0024	-0.0029	+1.85	+2.19
CDD	-0.0994	0.0703	-0.0814	0.1181	0.0039	0.0048	-2.95	-3.6
AutreCDD	-0.3502	0.0229	-0.2518	0.0183	-0.0011	-0.0016	+0.86	+1.2
CSP_AGRI	–	–	–	–	–	–	–	–
CSP_ARTI	–	–	–	–	–	–	–	–
CSP_CADRE	0.5004	0.1407	0.4521	0.1011	0.0179	0.0198	-13.56	-15.01
CSP_INTERM	0.1806	0.2739	0.1608	0.2561	0.0029	0.0032	-2.16	-2.42
CSP_OUVRI	0.0051	0.4691	-0.0492	0.1233	-0.017	0.0018	+12.86	-1.34
SECT_ENT_AZ	0.0519	0.016	0.0119	0.01	1e-04	3e-04	-0.05	-0.24
SECT_ENT_BE	0.1515	0.2884	0.1001	0.1367	0.0152	0.023	-11.48	-17.37
SECT_ENT_FZ	0.1488	0.1457	0.0794	0.0219	0.0098	0.0184	-7.43	-13.93
SECT_ENT_GI	0.0914	0.2775	0.0214	0.2489	6e-04	0.0026	-0.46	-1.98
SECT_ENT_JZ	0.0815	0.0342	0.0702	0.0202	0.001	0.0011	-0.74	-0.86
SECT_ENT_KZ	0.2228	0.0265	0.1177	0.0548	-0.0033	-0.0063	+2.52	+4.78
SECT_ENT_LZ	0.0369	0.0093	0.0253	0.0159	-2e-04	-2e-04	+0.13	+0.19
SECT_ENT_MN	0.0756	0.0872	0.0529	0.1042	-9e-04	-0.0013	+0.68	+0.97
SECT_ENT_RU	0.029	0.0326	-0.0321	0.0857	0.0017	-0.0015	-1.29	+1.16
$\ln(\widehat{\mathbf{D}+\mathbf{1}})$	–	–	–	–	**0.1189**	**0.06**	**89.88**	**45.37**

[a] pourcentage de la colonne (5) sur la différence de salaire (indiquée en première ligne)

[b] pourcentage de la colonne (6) sur la différence de salaire (indiquée en première ligne)

9.2 Enseignements

Les résultats empiriques aboutissent à 89.88% pour les femmes et 45.37% pour les hommes. La moyenne arithmétique de ces deux pourcentages permet d'obtenir la part de la différence de salaire non expliquée par les différences dans les caractéristiques individuelles entre les deux groupes. Ainsi, on obtient, un pourcentage final de 67.62%.

On peut alors dire que 67.62% de la différence de salaire entre les hommes et les femmes est due à la discrimination dans la structure salariale en France. Dans notre base, les hommes gagnent en moyenne 1 947,228 EUR par mois contre 1 543,11 EUR chez les femmes. La différence en absolu est de 404,12 EUR en faveur des hommes. Les femmes gagnent en moyenne, 20,75% moins que les hommes en France en 2012.

À cette différence, on peut attribuer une part de 67.62% liée à la discrimination. Par conséquent, 273,26 EUR de différence s'expliquent par la discrimination envers les femmes, et le pourcentage associé est de 14.03%.

En termes de salaire horaire, les hommes gagnent en moyenne 13,86 EUR par heure, contre 11,41 EUR pour les femmes. L'écart est de 2,44 EUR en faveur des hommes. Les femmes gagnent en moyenne, 17.64% moins que les hommes par heure travaillée.

La part liée à la discrimination est égale à 67.62% donc au final, les femmes sont discriminées à 11.93% de leur salaire horaire. Ceci équivaut à une différence de 1,65 EUR par heure.

Conclusion

L'étude réalisée avait pour objectif de quantifier la discrimination salariale envers les femmes en France sur l'année 2012. Le moyen mis en œuvre dans ce but a été la méthode de décomposition d'OAXACA. Celle-ci permet de prendre en compte les différences de caractéristiques individuelles entre les hommes et les femmes. Elle permet de quantifier l'écart de salaire entre un homme et une femme avec exactement les mêmes caractéristiques (personnelles comme professionnelles).

À partir de la base de données fournie par l'INSEE, un écart brut de 20.75% sur le salaire mensuel en ressort en faveur des hommes. Par conséquent, un homme gagnait en moyenne 404,12 euros de plus qu'une femme en 2012. Dans cet écart de 20.75%, la décomposition d'OAXACA attribue une part de 67.62% liée à la discrimination « directe ». Ainsi, un écart inexpliqué de 14.03% peut être attribué à la discrimination.

En matière de salaires horaires, ils sont en moyenne supérieurs de 2,44 euros pour les hommes, ce qui équivaut à une différence de 17.64%. En appliquant le pourcentage de 67.62%, on trouve que 1,65 euro est attribué à la discrimination.

En ce qui concerne les études portant sur le même sujet en France, Bernard THIRY [Thi85] (*Annales de l'INSEE n.58*, 1985) a construit deux modèles aboutissant à deux écarts assez éloignés. D'un côté, pour un écart brut de 30%, la part inexpliquée est de 76%, réduisant l'écart à 23% et de l'autre côté, toujours pour le même écart brut, il arrive à un écart de 15%, soit une part inexpliquée de 50%. Par la suite, Michel GLAUDE [Gla87] (*Données sociales*, 1987), à partir de l'enquête Emploi 1985, trouve un écart brut de salaire mensuel moyen de 25%. Il aboutit à un écart inexpliqué par son modèle de 14.5% en faveur des hommes, soit 58% de l'écart total. Plus récemment, Alain BAYET [Bay96] (*Données sociales*, 1996), à partir de l'enquête sur la structure des salaires en 1992 trouve un écart inexpliqué (du salaire mensuel moyen entre les hommes et les femmes) par son modèle de 14% en faveur des hommes. Les études tendent à se positionner autour d'un écart net de 14-15%.

Toutefois, la méthode comporte des limites. En effet, un certain nombre de paramètres ne peuvent être quantifiés et pris en compte dans l'étude, comme la discrimination subie par la femme tout au long de son parcours scolaire et professionnel. Par exemple, il est difficile de quantifier la discrimination à l'embauche, tout comme il est impossible de mesurer les écarts de trajectoire entre une femme et un homme dans les sphères éducatrice et domicile. Les caractéristiques non observables et la discrimination rétroactive ne sont donc pas considérées dans cette étude.

Si, ici, on a essentiellement considéré des variables pouvant jouer sur la productivité des travailleurs (théorie du capital humain), les nouvelles théories de l'économie du travail (A. PERROT) [Per98], dans les années 1990, montrent que des écarts de salaires peuvent provenir de caractéristiques non productives telles que la gestion interne de la main d'œuvre ou le taux de syndicalisation.

Annexes

Les annexes sont :

1. Explications brèves du calcul de similarité entre deux vecteurs binaires.
2. Liste des variables de la base INSEE retenues et leur label
3. Liste des variables construites/transformées et utilisées et leur label
4. Article original d'Oaxaca (1973)

ANNEXE 1 : CALCUL DE SIMILARITÉ ENTRE DEUX VECTEURS BINAIRES

Soient deux vecteurs binaires X et Y de dimensions N :

$$X = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N, \qquad Y = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^N$$

où $x_i, y_i \in \{0, 1\},\ \forall i \in \{1, 2, \ldots, N\}$.

On pose I le vecteur unitaire de taille N tel qu'il est entièrement composé de 1 :

$$I = (1, 1, \ldots, 1) \in \mathbb{R}^N.$$

Les compléments de X et de Y sont donnés par $\overline{X} = I - X$ et $\overline{Y} = I - Y$.

Similarité (*source* : [W+08, ZS03])

Soit $S_{ij_{i,j\in\{0,1\}}}$ le nombre d'apparition d'associations de i dans X avec j dans Y à la même position. On démontre alors que $S_{11}(X, Y) = X \cdot Y$ et $S_{00}(X, Y) = \overline{X} \cdot \overline{Y}$.

De façon claire, un tableau de contingence permet de mettre en évidence les S_{ij} :

		Vecteur Y	
		Valeur 1	Valeur 0
Vecteur X	Valeur 1	S_{11}	S_{10}
	Valeur 0	S_{01}	S_{11}

TABLE 16 – Tableau de contingence et S_{ij}

À partir de S_{00}, S_{01}, S_{10} et S_{11}, on peut calculer des indicateurs de similarité $S(X, Y)$ entre X et Y :

Indicateur de ...	$S(X, Y)$	Indicateur de ...	$S(X, Y)$
JACCARD-NEEDHAM	$\frac{S_{11}}{S_{11} + S_{10} + S_{01}}$	SOKAL-MICHENER	$\frac{S_{11} + S_{00}}{N}$
DICE	$\frac{S_{11}}{2S_{11} + S_{10} + S_{01}}$	ROGERS-TANMOTO	$\frac{S_{11} + S_{00}}{S_{11} + S_{00} + 2S_{10} + 2S_{01}}$
YULE	$\frac{S_{11}S_{00} - S_{10}S_{01}}{S_{11}S_{00} + S_{10}S_{01}}$	KULZINSKY	$\frac{S_{11}}{S_{10} + S_{01}}$
RUSSELL-RAO	$\frac{S_{11}}{N}$		

TABLE 17 – Liste des indicateurs de similarité

ANNEXE 2 : LISTE DES VARIABLES PERTINENTES DE LA BASE INSEE [14]

TABLE 18 – Détail des variables pertinentes de la base INSEE

Variable	Libellé	Modalités
SALRED	Salaire mensuel net redressé des non réponses (y compris les primes mensualisées et redressées des non- réponses)	montant en euros
SEXE	Sexe	— 1 - Masculin — 2 - Féminin
CSER	CSP pour les actifs (niveau agrégé, PCS 2003)	— - Sans objet (inactif) — 0 - Non renseigné — 1 - Agriculteurs exploitants — 2 - Artisans, commerçants et chefs d'entreprise — 3 - Cadres et professions intellectuelles supérieures — 4 - Professions intermédiaires — 5 - Employés — 6 - Ouvriers — 8 - Chômeurs n'ayant jamais travaillé
DDIPL	Diplôme le plus élevé obtenu agrégé	— - Non renseigné — 1 - Diplôme supérieur à baccalauréat + 2 ans — 3 - Baccalauréat + 2 ans — 4 - Baccalauréat ou brevet professionnel ou autre diplôme de ce niveau — 5 - CAP, BEP ou autre diplôme de ce niveau — 6 - Brevet des collèges — 7 - Aucun diplôme ou certificat d'études primaires
TPPRED	Temps de travail redressé dans l'emploi principal	— - Sans objet (ACTOP='2') ou non renseigné — 1 - Temps complet — 2 - Temps partiel

14. La liste complète des variables d'origine est disponible en téléchargement ici (235 pages) : `https://www.insee.fr/fr/statistiques/fichier/2415221/contenu_eec12_indiv12.pdf`

CSPIP	CSP de l'emploi occupé un an auparavant (niveau intermédiaire, PCS 2003)	— - Sans objet (sans emploi un an auparavant ou réinterrogation) — 00 - Non renseigné — 10 - Agriculteurs exploitants — 21 - Artisans — 22 - Commerçants et assimilés — 23 - Chefs d'entreprises de 10 salariés ou plus — 31 - Professions libérales — 32 - Cadres de la fonction publique, professions intellectuelles et artistiques — 36 - Cadres d'entreprises — 41 - Professions intermédiaires de l'enseignement, de la santé, de la fonction publique et assimilés — 46 - Professions intermédiaires administratives et commerciales des entreprises — 47 - Techniciens — 48 - Contremaîtres, agents de maîtrise — 51 - Employés de la fonction publique — 54 - Employés administratifs d'entreprise — 55 - Employés de commerce — 56 - Personnels des services directs aux particuliers — 61 - Ouvriers qualifiés — 66 - Ouvriers non qualifiés — 69 - Ouvriers agricoles
MATRI	Statut matrimonial légal	— - Sans objet (moins de 15 ans) — 1 - Célibataire — 2 - Marié(e) ou remarié(e) — 3 - Veuf(ve) — 4 - Divorcé(e)
TPP	Temps de travail dans l'emploi principal	— - Sans objet (ACTOP='2') ou non renseigné — 1 - A temps complet — 2 - A temps partiel — 3 - Sans objet (pour les personnes non salariées qui estiment que cette question ne s'applique pas à elles)

EMPNBH	Nombre d'heures au cours de la semaine référence effectuées dans l'emploi principal	— - Sans objet (ACTOP='2') ou non renseigné — 0.0 à 99.59 - Nombre d'heures effectuées la semaine de référence
HHC	Nombre moyen d'heures par semaine dans l'emploi principal (emploi régulier)	— - Sans objet (ACTOP='2') ou non renseigné — 0.30 à 99.59 - Nombre moyen d'heures par semaine
NBHEUR	Nombre d'heures correspondant au salaire déclaré	— - Sans objet (en interrogation intermédiaire) ou non déclaré — 1 à 250 - Nombre d'heures correspondant au salaire
NUITC	Travail de nuit (entre minuit et cinq heures du matin)	— - Sans objet (ACTOP='2') — 1 - Habituellement — 2 - Occasionnellement — 3 - Jamais
NAFG10N	Activité de l'établissement actuel (NAF rév2 en 10 postes)	— - Sans objet — 00 - Non renseigné — AZ - Agriculture, sylviculture et pêche — BE - Industrie manufacturière, industries extractives et autres — FZ - Construction — GI - Commerce de gros et de détail, transports, hébergement et restauration — JZ - Information et communication — KZ - Activités financières et d'assurance — LZ - Activités immobilières — MN - Activités spécialisées, scientifiques et techniques et activités de services administratifs et de soutien — OQ - Administration publique, enseignement, santé humaine et action sociale — RU - Autres activités de services
PUB3FP	Caractère public ou privé de l'employeur au sens de l'OEP	— - Sans objet (non salarié STAT2 ¡¿ 2) — 1 - Etat — 2 - Collectivités locales — 3 - Hôpitaux publics — 4 - Secteur privé

EFEN	Effectif salarié de l'entreprise	— - Non renseigné — 000000 à 999999 - Effectif salarié de l'entreprise
PAYNEU27	Pays de naissance (l'Union européenne des 27)	— 0 - Non — 1 - Oui
LNAIS	Lieu de naissance	— 1 - France — 2 - A l'étranger
NFR	Code de nationalité	— - Non renseigné — 1 - Français de naissance, y compris par réintégration — 2 - Français par naturalisation, mariage, déclaration ou option à sa majorité — 3 - Etranger
TAU10	Indicateur de tranche de taille d'aire urbaine	— 00 - Commune hors aire urbaine — 01 - Aire urbaine de moins de 15 000 habitants — 02 - Aire urbaine de 15 000 à 19 999 habitants — 03 - Aire urbaine de 20 000 à 24 999 habitants — 04 - Aire urbaine de 25 000 à 34 999 habitants — 05 - Aire urbaine de 35 000 à 49 999 habitants — 06 - Aire urbaine de 50 000 à 99 999 habitants — 07 - Aire urbaine de 100 000 à 199 999 habitants — 08 - Aire urbaine de 200 000 à 499 999 habitants — 09 - Aire urbaine de 500 000 à 9 999 999 habitants — 10 - Aire urbaine de Paris
TUU	Commune urbaine ou rurale	— 0 - Commune — 1 - Commune appartenant à une unité urbaine

REG	Région de résidence	— 11 - Ile-de-France — 21 - Champagne-Ardenne — 22 - Picardie — 23 - Haute-Normandie — 24 - Centre — 25 - Basse-Normandie — 26 - Bourgogne — 31 - Nord-Pas de Calais — 41 - Lorraine — 42 - Alsace — 43 - Franche-Comté — 52 - Pays de la Loire — 53 - Bretagne — 54 - Poitou-Charentes — 72 - Aquitaine — 73 - Midi-Pyrénées — 74 - Limousin — 82 - Rhône-Alpes — 83 - Auvergne — 91 - Languedoc-Roussillon — 93 - Provence-Alpes-Côte-d'Azur — 94 - Corse
NBAGENF	Nombre et âge des enfants du logement au 31 décembre de l'année d'enquête dans le logement.	— 0 - Pas d'enfant de moins de 18 ans — 1 - Un enfant de 6 à 17 ans — 2 - Un enfant de 3 à 5 ans — 3 - Un enfant de moins de 3 ans — 4 - Deux enfants, dont le plus jeune a de 6 à 17 ans — 5 - Deux enfants, dont le plus jeune a de 3 à 5 ans — 6 - Deux enfants, dont le plus jeune a moins de 3 ans — 7 - Trois enfants ou plus, dont le plus jeune a de 6 à 17 ans — 8 - Trois enfants ou plus, dont le plus jeune a de 3 à 5 ans — 9 - Trois enfants ou plus, dont le plus jeune a moins de 3 ans
AGE	Age détaillé au dernier jour de la semaine de référence	— 14 à 98 - age détaillé — 99 - 99 ans et plus

ANCENTR	Ancienneté dans l'entreprise ou dans la fonction publique en mois	— - Sans objet (ACF='2','3') ou non renseigné — 0 à 60 - Nombre de mois — Plus de 60 - Nombre de mois entre l'année d'entrée et l'année de collecte
STATUT	Statut détaillé mis en cohérence avec la profession	— - Sans objet (ACT='2','3') — 11 - Indépendants — 12 - Employeurs — 13 - Aides familiaux — 21 - Intérimaires — 22 - Apprentis — 33 - CDD (hors Etat, coll.loc.), hors contrats aides — 34 - Stagiaires et contrats aides (hors Etat, coll.loc.) — 35 - Autres contrats (hors Etat, coll.loc.) — 43 - CDD (Etat, coll.loc.), hors contrats aides — 44 - Stagiaires et contrats aides (Etat, coll.loc.) — 45 - Autres contrats (Etat, coll.loc.)
CONTRA	Type de contrat de travail	— - Sans objet (sans emploi ou fonctionnaire) — 1 - Contrat à durée indéterminée (y compris contrat Nouvelles Embauches) — 2 - Contrat à durée déterminée autre que saisonnier — 3 - Contrat saisonnier — 4 - Contrat d'intérim ou de travail temporaire — 5 - Contrat d'apprentissage ou contrat en alternance
NIVP	Niveau d'enseignement selon une version approuvée par le comité interministériel de la formation continue	— - Non renseigné — 10 - Diplôme bac+5 et plus — 20 - Diplôme niveau licence, maîtrise — 30 - Diplôme niveau bac+2 — 40 - Bac et enseignement supérieur sans diplôme bac+2 — 41 - Niveau bac sans études supérieures — 50 - Niveau terminale CAP-BEP, lycée — 60 - Troisième, année non terminale CAP-BEP — 71 - Collège — 72 - Enseignement primaire — 73 - Pas d'études

ANNEXE 3 : LISTE DES VARIABLES UTILISÉES DANS LES RÉGRESSIONS

TABLE 19 – Détail des variables utilisées dans les régressions

Variable	Label	Modalités
BAC	Possession d'un baccalauréat sans plus	— 0 - Ne possède pas le baccalauréat comme diplôme le plus élevé — 1 - Possède le baccalauréat comme diplôme le plus élevé
BAC2	Possession d'un BAC+2 sans plus	— 0 - Ne possède pas de diplôme BAC+2 comme diplôme le plus élevé — 1 - Possède un diplôme BAC+2 comme diplôme le plus élevé
BAC3	Possession d'un BAC+3 sans plus	— 0 - Ne possède pas de diplôme BAC+3 comme diplôme le plus élevé — 1 - Possède un diplôme BAC+3 comme diplôme le plus élevé
BAC5	Possession d'un BAC+5 sans plus	— 0 - Ne possède pas un diplôme BAC+5 ou plus comme diplôme le plus élevé — 1 - Possède un diplôme BAC+5 ou plus
AGE	Age de l'individu	
AGE_SQUARE	Age au carré de l'individu	
NFR	... est de nationalité étrangère	— 0 - Est de nationalité française — 1 - Est de nationalité étrangère
LNAIS	... est né(e) à l'étranger	— 0 - Est né en France — 1 - Est né à l'étranger
PAYNEU27	... est né(e) en dehors d'un pays de l'UE27	— 0 - Est né au sein de l'Union Européenne des 27 — 1 - Est né en dehors de l'Union Européenne des 27

TUU	... vit en commune urbaine	— 0 - Vit en commune rurale — 1 - Vit en commune urbaine
IDF	... vit en Ile-de-France	— 0 - Vit en Province — 1 - Vit en Ile-de-France
MARRIED	... est marié(e)	— 0 - N'est pas marié — 1 - Est marié
DIVORCED	... est divorcé(e)	— 0 - N'est pas divorcé — 1 - Est divorcé
WIDOW	... est veuf(ve)	— 0 - N'est pas veuf — 1 - Est veuf
NBENF1	... a exactement 1 enfant	— 0 - N'a pas exactement 1 enfant — 1 - A exactement 1 enfant
NBENF2	... a exactement 2 enfants	— 0 - N'a pas exactement 2 enfants — 1 - A exactement 2 enfants
NBENF3PLUS	... a 3 enfants ou plus	— 0 - N'a pas au moins 3 enfants — 1 - A au moins 3 enfants
PUBLIC	entreprise publique	— 0 - L'entreprise est privée — 1 - L'entreprise est publique
NBHEUR	nombre d'heures de travail par mois	
NBHEUR_SQUARE	NBHEUR au carré	
TPPRED	... travaille en temps partiel	— 0 - Travaille en temps complet — 1 - Travaille en temps partiel

NUITC	... travaille de nuit	— 0 - Ne travaille pas de nuit — 1 - Travaille de nuit
ANCENTR	ancienneté en mois dans l'entreprise	
ANCENTR_SQUARE	ANCENTR au carré	
CDD	contrat de type CDD	— 0 - N'a pas de CDD — 1 - A un CDD
AutreCDD	contrat de type autre que CDD hors CDI	— 0 - N'a pas de CDD saisonnier, contrat saisonnier, d'intérim, d'apprentissage ou d'alternance — 1 - A un CDD saisonnier, contrat saisonnier, d'intérim, d'apprentissage ou d'alternance
CSP_AGRI	Agriculteur	— 0 - N'est pas agriculteur ou exploitant — 1 - Est agriculteur ou exploitant
CSP_ARTI	Artisan	— 0 - N'est pas artisan, commerçant ou chef d'entreprise — 1 - Est artisan, commerçant ou chef d'entreprise
CSP_CADRE	Cadre	— 0 - N'est pas cadre ou de profession intellectuelle supérieure — 1 - Est cadre ou de profession intellectuelle supérieure
CSP_INTERM	Profession intermédiaire	— 0 - N'est pas de profession intermédiaire — 1 - Est de profession intermédiaire
CSP_OUVRI	Ouvrier	— 0 - N'est pas ouvrier — 1 - Est ouvrier
SECT_ENT_AZ	Agriculture, syviculture et pêche	— 0 - L'entreprise n'est pas dans le secteur de l'agriculture, syviculture ou de pêche — 1 - L'entreprise est dans le secteur de l'agriculture, syviculture ou de pêche

SECT_ENT_BE	Industrie manufacturière, industries extractives et autres	— 0 - L'entreprise n'est pas dans le secteur de l'industrie manufacturière, extractive ou autre — 1 - L'entreprise est dans le secteur de l'industrie manufacturière, extractive ou autre
SECT_ENT_FZ	Construction	— 0 - L'entreprise n'est pas dans le secteur de la construction — 1 - L'entreprise est dans le secteur de la construction
SECT_ENT_GI	Commerce de gros et de détail, transports, hébergement et restauration	— 0 - L'entreprise n'est pas dans le secteur du commerce de gros et de détail, transports, hébergement ou restauration — 1 - L'entreprise est dans le secteur du commerce de gros et de détail, transports, hébergement ou restauration
SECT_ENT_JZ	Information et communication	— 0 - L'entreprise n'est pas dans le secteur de l'information et communication — 1 - L'entreprise est dans le secteur de l'information et communication
SECT_ENT_KZ	Activités financières et d'assurance	— 0 - L'entreprise n'est pas dans le secteur des activités financières et d'assurance — 1 - L'entreprise est dans le secteur des activités financières et d'assurance
SECT_ENT_LZ	Activités immobilières	— 0 - L'entreprise n'est pas dans le secteur des activités immobilières — 1 - L'entreprise est dans le secteur des activités immobilières
SECT_ENT_MN	Activités spécialisées, scientifiques, techniques et services administratifs	— 0 - L'entreprise n'est pas dans le secteur des activités spécialisées, scientifiques, techniques ou des services administratifs — 1 - L'entreprise est dans le secteur des activités spécialisées, scientifiques, techniques ou des services administratifs
SECT_ENT_RU	Autres activités de services	— 0 - L'entreprise n'est pas dans le secteur des autres services — 1 - L'entreprise est dans le secteur des autres services

ANNEXE 4 : ÉTUDE DE RÉFÉRENCE : OAXACA R., 1973

INTERNATIONAL ECONOMIC REVIEW
Vol. 14, No. 3, October, 1973

MALE-FEMALE WAGE DIFFERENTIALS IN URBAN LABOR MARKETS*

BY RONALD OAXACA[1]

CULTURE, TRADITION, AND OVERT DISCRIMINATION tend to make restrictive the terms by which women may participate in the labor force. These influences combine to generate an unfavorable occupational distribution of female workers vis-à-vis male workers and to create pay differences between males and females within the same occupation. The result is a chronic earnings gap between male and female full-time, year-round workers. Unfortunately, explanations at this level of generality are mainly descriptive. It is the purpose of this paper to estimate the average extent of discrimination against female workers in the United States and to provide a quantitative assessment of the sources of male-female wage differentials.

1. INTRODUCTION

In his study of sex differentials from the 1950 U. S. Census, Sanborn [**13**] found the female/male annual income ratio to be .58 which implies a male-female wage differential (as a proportion of the female wage) of .72. Sanborn's stated objective was to consider discrimination only in the context of equal pay for equal work and not to deal with discrimination stemming from occupational barriers. Using both male and female adjustment weights, Sanborn adjusted the income ratio for occupational distribution, annual hours of work, education, urbanness, race, turnover, absenteeism, and work experience. In his attempt to approximate equal work, Sanborn controlled for 262 detailed occupations. These adjustments brought the income ratio up to .87—.88. The residual difference was therefore .13 and thus about 18% of the original differential. As an estimate of the upper limits of the effects of discrimination it is rather low, but this is not surprising in view of the aspect of discrimination under study.

Using data from the 1960 U. S. Census, Fuchs [**6**] calculated the hourly earnings of females relative to males to be .60 which implies a male-female wage differential of .66. Fuchs does not believe that inherent differences in physical strength account for the sizeable pay difference simply because jobs requiring heavy labor are in a minority in the present day occupational structure. He contended that role differentiation stemming from social attitudes and discrimination affects the determinants of earnings. The earnings ratio was raised to .66

* Manuscript received May 1, 1972; revised November 13, 1972.

[1] The author wishes to thank Orley Ashenfelter, Daniel Hamermesh, Albert Rees, Sherwin Rosen, and a referee for comments on an earlier version of this paper. Naturally, the responsibility of any remaining errors lies with the author. Financial assistance for this study was provided by the U.S. Department of Labor, Manpower Administration, under the provisions of Title 1 of the Manpower Development and Training Act, Public Law 87-415, as amended.

after adjustments for color, schooling, age, city size, marital status, class of worker, and length of trip to work. This adjusted ratio implies a residual differential of .51, which is 77% of the original differential. From the results of his regressions of hourly earnings across occupations, Fuchs concluded that nearly all of the wage differential could be explained away if one chooses sufficiently narrow occupational categories. As Fuchs observed, this merely casts the problem interms of why occupational distributions are so different between males and females.

In Cohen's study of sex differentials [**5**], an annual difference of $5,000 for full-time employed males and females was calculated from a 1969 working conditions survey. Cohen adjusted the difference in pay by excluding those under the age of 22 and over 64, the self-employed, persons without a steady job, and professionals. He also adjusted for differences in annual hours worked, fringe benefits, absenteeism, seniority, education, and unionization. The income difference was reduced to $2,550 or 51% of the original difference. Cohen attributed this large residual to the concentration of women in lower paying jobs.

The Malkiels' study [**9**] of professional workers revealed male-female wage differentials that ranged from .48 to .51 for the years 1966, 1969–1971. Adjustment for differences in schooling and experience yielded residual differentials varying between 37% to 49% of the original wage differentials. For one of the years the unexplained residual was reduced to only 3% of the original wage difference by adjusting for differences in schooling, experience, job level, critical area of study, and publications. This clearly indicates that unequal pay for equal work is not a significant source of discrimination. The problem of discrimination centers around the assignment of job levels. The Malkiels found that 53% of the male-female difference in job levels could not be explained by differences in personal characteristics.

2. A MEASURE OF DISCRIMINATION

Discrimination against females can be said to exist whenever the relative wage of males exceeds the relative wage that would have prevailed if males and females were paid according to the same criteria. We can formalize this notion by proposing the concept of a discrimination coefficient (D) as a measure of discrimination:

$$D = \frac{W_m/W_f - (W_m/W_f)^0}{(W_m/W_f)^0}; \tag{1}$$

where

(W_m/W_f) = the observed male-female wage ratio;

and

$(W_m/W_f)^0$ = the male-female wage ratio in the absence of discrimination.

An equivalent expression in natural logarithms is

$$(2) \qquad \ln (D+1) = \ln (W_m/W_f) - \ln (W_m/W_f)^0 .$$

Assuming that employers in a nondiscriminating labor market adhere to the principle of cost minimization, we have

$$\left(\frac{W_m}{W_f}\right)^0 = \frac{MP_m}{MP_f} ;$$

where MP_m and MP_f are the marginal products of males and females, respectively.

In [2] Becker defined the market discrimination coefficient as the percentage wage differential between two types of perfectly substitutable labor. For those cases in which the two factors were not necessarily perfect substitutes, Becker defined the discrimination coefficient as the simple difference between the observed wage ratio and the wage ratio in the absence of discrimination. The discrimination coefficient defined by (1) is simply Becker's generalized measure divided by the wage ratio in the absence of discrimination. The generalized measures admit perfect substitutes as a special case and thus afford more flexibility for empirical work.

3. ESTIMATION PROCEDURES

Since $(W_m/W_f)^0$ is unknown, the estimation of D is equivalent to estimating $(W_m/W_f)^0$. On the basis of either of two assumptions, we can estimate the male-female wage ratio that would exist in the absence of discrimination: If there were no discrimination, 1) the wage structure currently faced by females would also apply to males; or 2) the wage structure currently faced by males would also apply to females. Assumption one (two) says that females (males) would on average receive in the absence of discrimination the same wages as they presently receive, but that discrimination takes the form of males (females) receiving more (less) than a nondiscriminating labor market would award them.

Ordinary least squares estimation of a wage equation for any given group of workers provides an estimate of the wage structure applicable to that group. The wage equation to be estimated separately for each race-sex group has the semi-log functional form

$$(3) \qquad \ln (W_i) = Z_i'\beta + u_i , \qquad i = 1, \ldots, n$$

where

W_i = the hourly wage rate of the i-th worker,
Z_i' = a vector of individual characteristics,
β = a vector of coefficients,
u_i = a disturbance term.

When the male-female wage differential is expressed in natural logarithms, the formulation of the discrimination coefficient in (2) and our alternative assumptions about which wage structure would prevail in the absence of discrimination together imply that the wage differential can be decomposed into the

effects of discrimination and the effects of differences in individual characteristics.

Let

$$G = \frac{\bar{W}_m - \bar{W}_f}{\bar{W}_f},$$

then

$$\ln (G + 1) = \ln (\bar{W}_m) - \ln (\bar{W}_f) \tag{4}$$

where $\bar{W}_m$ and $\bar{W}_f$ are the average hourly wages for males and females, respectively.[2] From the properties of ordinary least squares estimation, we have

$$\ln (\bar{W}_m) = \bar{Z}_m'\hat{\beta}_m \tag{5}$$

$$\ln (\bar{W}_f) = \bar{Z}_f'\hat{\beta}_f \tag{6}$$

where

$\bar{Z}_m'$ and $\bar{Z}_f'$ = the vectors of mean values of the regressors for males and females, respectively.

$\hat{\beta}_m$ and $\hat{\beta}_f$ = the corresponding vectors of estimated coefficients.

Upon substitution of (5) and (6) into (4), we obtain

$$\ln (G + 1) = \bar{Z}_m'\hat{\beta}_m - \bar{Z}_f'\hat{\beta}_f . \tag{7}$$

If we let

$$\Delta\bar{Z}' = \bar{Z}_m' - \bar{Z}_f' \tag{8}$$

$$\Delta\hat{\beta} = \hat{\beta}_f - \hat{\beta}_m \tag{9}$$

and substitute $\hat{\beta}_m = \hat{\beta}_f - \Delta\hat{\beta}$ in (7), then the male-female wage differential can be written as

$$\ln (G + 1) = \Delta\bar{Z}'\hat{\beta}_f - \bar{Z}_m'\Delta\hat{\beta} . \tag{10}$$

On the basis of equation (2) and the assumption that the current female wage structure would apply to both males and females in a nondiscriminating labor market, it can be shown that

$$\ln \left(\frac{\widehat{W}_m}{W_f}\right)^0 = \Delta\bar{Z}'\hat{\beta}_f \tag{11}$$

$$\widehat{\ln (D + 1)} = -\bar{Z}_m'\Delta\hat{\beta} . \tag{12}$$

Thus expressions (11) and (12) represent the decomposition of the wage differential into the estimated effects of differences in individual characteristics and the estimated effects of discrimination, respectively.

An alternative decomposition of the wage differential is obtained by substi-

[2] These wage figures are computed as geometric means, i.e.,

$$\bar{W} = \exp \left\{ \left[\sum_{i=1}^{n} \ln (W_i) \right] \Big/ n \right\}$$

tuting $\hat{\beta}_f = \Delta\hat{\beta} + \hat{\beta}_m$ in (7):

$$\ln (G + 1) = \Delta\bar{Z}'\hat{\beta}_m - \bar{Z}_f'\Delta\hat{\beta}\,. \tag{13}$$

On the basis of (2) and the assumption that the current male wage structure would apply to both males and females in the absence of discrimination, it can be shown that

$$\ln \left(\frac{\widehat{W_m}}{W_f}\right)^0 = \Delta\bar{Z}'\hat{\beta}_m \tag{14}$$

$$\ln \widehat{(D + 1)} = -\bar{Z}_f'\Delta\hat{\beta}\,. \tag{15}$$

Our method of estimating the effects of discrimination involves the familiar index number problem. Therefore, the separate estimates obtained from using both the male and female regression weights establish a range of possible values.

4. SPECIFICATION OF THE CONTROL VARIABLES

In accordance with the post-schooling investment model of human capital formation as developed in [**8**] and [**10**], a quadratic experience variable is included in the wage equations. The corresponding coefficients measure the combined effects of the average rate of return to on-the-job training, the initial proportion of time allocated to OJT, and the length of the investment horizon.

Since data on the actual number of years of work experience for a large sample of workers are generally unavailable, we define a proxy for actual work experience:

$$X_i = A_i - E_i - 6 \tag{16}$$

where

X_i = potential experience,
A_i = the age of the i-th individual,
E_i = the number of years of schooling completed by the i-th individual.

When work experience is acquired without interruption after the completion of formal schooling, potential and actual experience coincide. Potential experience is a reasonable proxy for actual experience in the case of males since males on average exhibit a strong attachment to the labor force. However, potential experience overstates the actual years of work experience of females to the extent that many female workers have left the labor force for some period in the past due to their household and childbearing activities. The difficulty this presents for estimation is of course the errors in variables problem. It is not clear what the net effect would be on our estimates of discrimination. It seems reasonable to suppose that the potential experience-wage profile would be flatter than the actual experience-wage profile. If the estimator of the coefficient on the linear experience term were biased downward for females, then $-\Delta\hat{\beta}$ would be biased upward in this instance. Consequently, there would be a bias toward finding

discrimination.[3]

As a rough attempt to handle the problem of lost experience, we have controlled for the number of children (C) born to the female. The linear children variable reflects the cost of lost experience due to child care, including the costs from the depreciation of skills during the periods of absence from the labor force. Accordingly, we expect the estimated coefficient ($\hat{\beta}_c$) to have a negative sign. There is some difficulty associated with the introduction of this variable since it is obviously correlated with the proxy experience variable; however, in the absence of data on actual work experience, the use of the children variable seems justified.[4]

The remaining control variables are briefly described:

Education: years of schooling completed (linear and quadratic terms);

Class of Worker: dummy variables for union membership (privately employed wage and salary worker), government employed, and self-employed with non-union private wage and salary workers as the reference group;

Industry: dummy variables for U.S. Census two digit industries with retail trade as the reference group;

Occupation: dummy variables for U.S. Census two digit occupations with sales workers as the reference group;

Health Problems: dummy variable = 1 if the individual reports health problems that affect the kind or amount of work he or she can perform, and '0' otherwise;

Part-Time: dummy variable = 1 if the individual works less than thirty-five hours a week, and '0' otherwise;

Migration: a) dummy variable = 1 if the individual has maintained a residence more than fifty miles from his or her current address since the age of seventeen, and '0' otherwise, b) *YRSM*: number of years since the individual last migrated (linear and quadratic terms);

Marital Status: dummy variables for spouse present, spouse absent, widowed, and divorced (or separated) with never married individuals as the reference group;

Size of Urban Area: dummy variables for residence in Standard Metropolitan Statistical Areas less than 250,000 (SMSA<250), greater than or equal to 250,000 but less than 500,000 (SMSA 250–500), greater than or equal to 500,000 but less than 750,000 (SMSA 500–750), and greater than or equal to 750,000 (SMSA 750+) with urban, non-SMSA's as the reference group; and

Region: dummy variables for U.S. Census regions North East, North Central, and West with South as the reference group.[5]

One difficulty with the present formulation of the wage equation is that it controls for what many would consider to be major sources of discrimination

[3] These considerations also apply to wage models that specify an age variable in lieu of experience.

[4] The consequences of omitting this variable are discussed in footnote 6.

[5] The rationale for including most of these variables in a wage regression is fairly well known; nevertheless a more detailed discussion is given in [**11**].

against women. By controlling for broadly defined occupation, we eliminate some of the effects of occupational barriers as sources of discrimination. As a result, we are likely to underestimate the effects of discrimination. Therefore, we estimate another set of equations that do not control for occupation, industry, and class of worker. We shall refer to this set of regressions as the personal characteristics wage regressions, and to the original set as the full-scale wage regressions.

It is clear that the magnitude of the estimated effects of discrimination crucially depends upon the choice of control variables for the wage regressions. A researcher's choice of control variables implicitly reveals his or her attitude toward what constitutes discrimination in the labor market. If it were possible to control for virtually all sources of variation in wages, one could pretty well eliminate labor market discrimination as a significant factor in determining wage differentials by sex (or race). This is because $\Delta\hat{\beta}$, and therefore ln $(D+1)$, would be very small. The result is that whatever the wage differential observed, it is completely justified on the grounds of alleged productivity differences. The other extreme is to control for virtually nothing and thereby minimize the role of productivity differences, $\Delta\bar{Z}$. This is tantamount to declaring at the outset that the two labor inputs are near perfect substitutes and therefore attributing virtually all of the observed wage differential to labor market discrimination, i.e., relatively high values of $\Delta\hat{\beta}$.

In reference to the type of regression approach we have adopted, it is impractical to control for detailed occupation. When we control for broadly defined two digit categories, we are not assuming that the conditions of equal work are met. Although the full-scale wage regressions eliminate male-female differences in broad occupational attachment as possible sources of discrimination, they can still reflect job and pay discrimination within each two digit category. The personal characteristics wage equations were specified with the intention of examining the issue of equal pay for roughly similar personal characteristics and not just for equal work.

5. EMPIRICAL RESULTS

The data for the study are from the 1967 Survey of Economic Opportunity. The particular subsample used for this study consists of the intersection of the following sets: those individuals who show an hourly wage for the week preceding the survey; adults sixteen years or older; those who live in urban areas; and those who report their race as either White or Negro.

The regression coefficients corresponding to the full-scale and personal characteristics wage equations are presented in Tables 1 and 2, respectively. Coefficients were not estimated in the following cases: 1) the particular characteristic served as the base group; 2) there were no observations in a particular cell: 3) the same observations were found in another cell; or 4) the regressor was left out on the basis of poor results from earlier regressions. The joint tests of significance for $\Delta\hat{\beta}$ in both the full-scale and personal characteristics wage regressions

TABLE 1

FULL-SCALE WAGE REGRESSIONS[a]

Variable	Whites			Blacks		
	Male	Female	$\Delta\beta$	Male	Female	$\Delta\beta$
Constant	.0365 (.77)	−.1024 (−1.34)	−.1389 (−1.60)	.0953 (1.71)	−.3851* (−6.35)	−.4804* (−5.85)
Experience						
Experience	.0176* (13.89)	.0138* (8.19)	−.0038 (−1.88)	.0117* (7.73)	.0067* (4.38)	−.0050* (−2.29)
Experience**2	−.000288* (−12.22)	−.000248* (−7.31)	.000040 (.98)	−.000204* (−7.59)	−.000122* (−4.33)	.000082* (2.11)
Education						
Education	.0082 (1.27)	−.0118 (−.98)	−.0200 (−1.53)	−.0308* (−4.60)	−.0175* (−1.98)	.0133 (1.19)
Education**2	.00169* (5.92)	.00194* (3.53)	.00025 (.42)	.00300* (8.23)	.00245* (5.26)	−.00055 (−.93)
Class of Worker						
Union	.1113* (9.39)	.1500* (6.70)	.0387 (1.59)	.2129* (14.15)	.0719* (3.11)	−.1410* (−5.14)
Nonunion Private Wage and Salary	—	—	—	—	—	—
Government	.0646* (3.15)	.1445* (5.89)	.0799* (2.54)	.1328* (5.44)	.1263* (5.19)	−.0065 (−.19)
Self-Employed	−.1290* (3.51)	.1137 (1.22)	.2427* (2.54)	−.0128 (−.15)	−.3437* (−2.67)	−.3309* (−2.15)
Industry						
Agriculture	.1285 (1.81)	.2847 (1.09)	.1562 (.61)	−.0067 (−.08)	−.0190 (−.21)	−.0123 (−1.18)
Mining	.3604* (6.83)	.4112* (2.02)	.0508 (.26)	.0697 (.40)	—	—
Construction	.2997* (13.72)	.2444* (3.80)	−.0553 (−.86)	.2729* (10.54)	.0395 (.22)	−.2334 (−1.30)
Manufacturing-Durable	.2398* (13.76)	.2562* (8.39)	.0164 (.48)	.2101* (9.15)	.2590* (6.46)	.0489 (1.06)
Manufacturing-Non Durable	.2086* (11.03)	.1968* (6.60)	−0.118 (−.35)	.1679* (6.85)	.2305* (6.46)	.0626 (1.45)
Transportation	.2332* (9.81)	.3154* (5.54)	.0822 (1.40)	.2182* (7.39)	.5463* (5.73)	.3281* (3.32)
Communications	.2370* (5.62)	.2290* (4.56)	−.0080 (−.12)	.1555 (1.78)	.2657* (3.71)	.1102 (.98)
Utilities	.2414* (7.32)	.2451* (2.83)	.0087 (.04)	.1433* (3.45)	.7026* (2.76)	.5593* (2.19)
Wholesale Trade	.2039* (8.45)	.1979* (4.74)	−.0060 (−.13)	.1204* (3.76)	.3065* (4.34)	.1861* (2.41)
Retail Trade	—	—	—	—	—	—
Finance	.2224* (8.25)	.1761* (5.65)	−.0463 (−1.14)	.0184 (.47)	.1593* (3.22)	.1409* (2.25)

(*Continued on next page*)

TABLE 1

(CONTINUED)

Variable	Whites			Blacks		
	Male	Female	$\Delta\beta$	Male	Female	$\Delta\beta$
Business and Repair Services	.1385* (4.44)	.1525* (3.24)	.0140 (.26)	.0766* (2.10)	.1326* (2.31)	.0560 (.83)
Personal Services	−.0618 (−1.71)	−.0183 (−.50)	.0435 (.85)	−.1055* (−3.22)	.0118 (.40)	.1173* (2.65)
Recreation	.0488 (.97)	.1527* (1.97)	.1039 (1.16)	.0020 (.04)	.1019 (1.29)	.0999 (1.05)
Professional Services	−.0629* (−2.53)	.0528* (2.01)	.1157* (3.24)	.0633* (2.13)	.1181* (4.45)	.0548 (1.38)
Public Administration	.1970* (6.58)	.2165* (4.86)	.0195 (.37)	.2374* (6.75)	.2170* (5.61)	−.0204 (−.39)
Occupation						
Professional Workers	.1563* (6.62)	.3736* (10.25)	.2173* (5.16)	.2144* (4.62)	.4631* (10.80)	.2487* (3.43)
Managers	.1822* (8.27)	.2759* (6.85)	.0937* (2.12)	.0810 (1.49)	.2792* (3.53)	.1982* (2.07)
Clerical Workers	−.0639* (−2.68)	.1665* (6.03)	.2304* (6.41)	.0208 (.54)	.1509* (4.50)	.1301* (2.55)
Sales Workers	—	—	—	—	—	—
Craftsmen	.0275 (1.28)	.0932 (1.31)	.0657 (.93)	.0733* (1.99)	.1297* (1.97)	.0564 (.75)
Operatives	−.1064* (−4.92)	.0128 (.37)	.1192* (3.00)	−.0271 (−.77)	.0236 (.62)	.0507 (.98)
Private Household Workers	−.1900 (−1.03)	−.3060* (−5.46)	−.1160 (−.58)	−.0458 (−.28)	−.1432* (−3.58)	−.0974 (−.58)
Service Workers	−.1358* (−5.19)	−.0219 (−.72)	.1139* (2.89)	−.0998 (−2.84)	−.0164 (−.53)	.0834 (1.78)
Farm Laborers	−.4570* (−5.38)	.1579 (.43)	.6149 (1.71)	−.1421 (−1.36)	—	—
Laborers	−.1540* (−5.59)	−.0166 (−.15)	.1374 (1.29)	−.0637 (−1.77)	.0317 (.37)	.0954 (1.03)
Health Problems	−.1001* (−6.08)	−.0710* (−2.70)	.0291 (.97)	−.0811* (−3.79)	−.0270 (−1.31)	.0541 (1.82)
Part-Time	−.1874* (−9.14)	−.0445* (−2.64)	.1429* (5.37)	−.1117* (−4.80)	.0034 (.21)	.1151* (4.04)
Migration						
Migration	−.0356* (−2.48)	−.1073* (−5.03)	−.0717* (−2.87)	.0052 (.44)	−.0361 (−1.94)	−.0413 (−1.88)
YRSM	.0072* (4.22)	.0087* (3.33)	.0015 (.48)	—	.0025* (2.73)	—
YRSM**2	−.000140* (−3.08)	−.000147* (−2.14)	−.000007 (−.10)	—	—	—
Marital Status						
Spouse Present	.1841* (11.88)	.0883* (4.51)	−.0958* (−3.91)	.1211* (6.43)	.0995* (5.13)	−.0216 (−.80)

(*Continued on next page*)

TABLE 1
(CONTINUED)

Variable	Whites			Blacks		
	Male	Female	$\Delta\beta$	Male	Female	$\Delta\beta$
Spouse Absent	.1124 (1.72)	.0852 (1.39)	−.0272 (−.30)	.0446 (.79)	.1050* (2.38)	.0604 (.84)
Widowed	.1030* (2.37)	.0687* (2.21)	−.0343 (−.64)	.0920* (2.13)	.0980* (3.47)	.0060 (.12)
Divorced	.0793* (2.74)	.0933* (3.38)	.0140 (.35)	.0396 (1.53)	.0607* (2.72)	.0211 (.62)
Single, Never Married	—	—	—	—	—	—
Children	—	−.0198* (−4.51)	—	—	−.0007 (−.24)	—
Size of Urban Area						
Urban, Non SMSA	—	—	—	—	—	—
SMSA<250	.0332* (1.98)	.0920* (3.86)	.0588* (2.07)	.0523 (1.54)	.1458* (4.19)	.0935 (1.93)
SMSA 250-500	.0727* (3.89)	.0956* (3.65)	.0229 (.73)	.1098* (2.83)	.1833* (4.61)	.0735 (1.32)
SMSA 500-750	.1411* (7.30)	.1524* (5.46)	.0113 (.34)	.1349* (3.55)	.1816* (4.46)	.0467 (.84)
SMSA 750+	.1745* (12.57)	.2186* (11.21)	.0441 (1.89)	.2079* (6.46)	.3643* (10.92)	.1564* (3.38)
Region						
North East	.0738* (5.63)	.0882* (4.69)	.0144 (.64)	.1366* (7.86)	.1724* (9.24)	.0358 (1.41)
North Central	.0749* (5.85)	.0646* (3.52)	−.0103 (−.47)	.1479* (9.37)	.1376* (8.00)	−.0103 (−.44)
South	—	—	—	—	—	—
West	.1200* (8.51)	.1389* (6.83)	.0189 (.79)	.2452* (12.48)	.2612* (12.07)	.0160 (.55)
F Statistic for Joint Test of Significance	128.31*	50.88*	13.28*	71.96*	97.14*	9.93*
R^2	.43	.33		.46	.56	
Standard Error of Estimate	.40	.45		.35	.36	
Number of Observations	8,123	4,962		3,897	3,502	

[a] '*t*' values in parentheses.
* Significant at the 5% level.

reveal that the wage structure for males and females are significantly different with respecct to the regressors common to both groups.

The average logarithms of the hourly wages (and the corresponding geometric mean wages) computed from our sample are as follows: 1.0806 ($2.95) for white males, .6499 ($1.92) for white females, .7721 ($2.16) for black males, and .3732 ($1.45) for black females. The values of the wage differentials in logarithmic terms, ln $(G + 1)$, are .4307 for whites and .3989 for blacks. To facilitate com-

TABLE 2

PERSONAL CHARACTERISTICS WAGE REGRESSIONS[a]

Variable	Whites			Blacks		
	Male	Female	$\Delta\beta$	Male	Female	$\Delta\beta$
Constant	−.0681* (−2.03)	.0894 (.94)	.1575 (1.87)	.1472* (2.44)	−.2325* (−3.75)	−.3797* (−4.65)
Experience						
Experience	.0222* (16.51)	.0182* (10.20)	−.0039 (−1.81)	.0195* (12.03)	.0066* (4.02)	−.0129* (−5.53)
Experience**2	−.000354* (−14.19)	−.000349* (−9.69)	.000005 (.13)	−.000340* (−11.69)	−.000133* (−4.26)	.000207* (4.86)
Education						
Education	.0342* (5.24)	−.0394* (−3.30)	−.0736* (−5.63)	−.0434* (−6.16)	−.0660* (−7.12)	−.0226 (−1.95)
Education**2	.00097* (3.51)	.00450* (8.63)	.00353* (6.24)	.00417* (11.46)	.00685* (15.35)	.00268* (4.68)
Health Problems	−.1325* (−7.57)	−.1097* (−3.89)	.0228 (.71)	−.1275* (−5.41)	−.0638* (−2.75)	.0637 (1.92)
Part-Time	−.3154* (−14.84)	−.1560* (−9.09)	.1594* (5.81)	−.1908* (−7.57)	−.1139* (−6.73)	.0769* (2.52)
Migration						
Migration	−.0316* (−2.05)	−.1262* (−5.54)	−.0946* (−3.56)	.0125 (.96)	−.0726* (−3.48)	−.0851* (−3.49)
YRSM	.0073* (3.98)	.0107* (3.81)	.0034 (1.05)	—	.0034* (3.24)	—
YRSM**2	−.000140* (−2.91)	−.000169* (−2.29)	−.000029* (−.34)	—	—	—
Marital Status						
Spouse Present	.2514* (15.44)	.1246* (5.96)	−.1268* (−4.90)	.1584* (7.66)	.0986* (4.53)	−.0598* (−1.99)
Spouse Absent	.1189 (1.71)	.0706 (1.07)	−.0483 (−.51)	.0975 (1.56)	.0964 (1.94)	−.0011 (−.01)
Widowed	.1389* (3.00)	.0804* (2.41)	−.0585 (−1.01)	.1648* (3.46)	.0754* (2.38)	−.0894 (−1.55)
Divorced	.1027* (3.34)	.1064* (3.61)	.0037 (.09)	.0511 (1.78)	.0618* (2.46)	.0107 (.28)
Single, Never Married	—	—	—	—	—	—
Children	—	−.0295* (−6.31)	—	—	−.0025 (−.80)	—
Size of Urban Area						
Urban, Non SMSA	—	—	—	—	—	—
SMSA<250	.0412* (2.30)	.1080* (4.23)	.0668* (2.20)	.0667 (1.78)	.1415* (3.60)	.0748 (1.38)
SMSA 250–500	.0845* (4.26)	.1154* (4.11)	.0309 (.92)	.1103* (2.57)	.1879* (4.20)	.0776 (1.25)
SMSA 500–750	.1739* (8.47)	.1721* (5.77)	−.0018 (−.05)	.1821* (4.34)	.1769* (3.86)	−.0052 (−.08)

(*Continued on next page*)

TABLE 2
(CONTINUED)

Variable	Whites			Blacks		
	Male	Female	$\Delta\beta$	Male	Female	$\Delta\beta$
SMSA 750+	.1972* (13.45)	.2543* (12.27)	.0571* (2.31)	.2452* (6.95)	.3888* (10.39)	.1436* (2.80)
Region						
North East	.0655* (4.73)	.1129* (5.69)	.0474* (2.01)	.1704* (8.97)	.2268* (11.07)	.0564* (2.02)
North Central	.0790* (5.91)	.0685* (3.52)	−.0105 (−.46)	.2255* (13.56)	.1996* (10.57)	−.0259 (−1.03)
South	—	—	—	—	—	—
West	.1111* (7.49)	.1174* (5.43)	.0063 (.25)	.2889* (13.44)	.3027* (12.56)	.0138 (.43)
F Statistic for Joint Test of Significance	213.29*	67.51*	45.05*	103.93*	132.01*	58.14*
R^2	.34	.22		.33	.43	
Standard Error of Estimate	.43	.49		.39	.40	
Number of Observations	8,123	4,962		3,897	3,502	

[a] t values in parentheses.
* Significant at the 5% level.

parison of our results with those of other studies, the wage differential G is also calculated. The value of G is .54 for whites and .49 for blacks.

The effects of discrimination are approximated by the residual left after subtracting the effects of differences in individual characteristics from the overall wage differential. The calculations based on the full-scale wage regressions are presented in Table 3. As a simple average of the two estimates obtained, discrimination accounts for 58.4% of the logarithmic wage differential for whites and 55.6% for blacks. The average value of the discrimination coefficient is .29 for whites and .25 for blacks. Table 4 presents the effects of discrimination calculated from the personal characteristics wage regressions. Predictably, the estimated effects of discrimination are larger than those reported in Table 3: Discrimination accounts for approximately 77.7% of the wage diffferential for whites and 93.6% for blacks. The averaged estimates of the discrimination coefficient are .40 and .45 for whites and blacks, respectively. Under both sets of regressions and for both races, sex differences in the distribution of part-time employment and marital status significantly contributed to a narrowing of the wage differential. It is evident from Tables 1 and 2 that workers with spouse present tend to earn more than others even after controlling for other factors. A smaller proportion of women workers fall into the category of spouse present, and therefore this difference reduces the wage differential due to discrimination. The difference probably reflects the competing activities of production in the home. In the case of whites, the effects of childbearing also narrow

TABLE 3

THE EFFECTS OF DISCRIMINATION ESTIMATED FROM THE FULL-SCALE WAGE REGRESSIONS

Item	Whites				Blacks			
	Female Regression Weights		Male Regression Weights		Female Regression Weights		Male Regression Weights	
	(1)[a]	(2)[b]	(3)[c]	(4)[b]	(5)[a]	(6)[b]	(7)[c]	(8)[b]
Wage differential	.4307	100.0%	.4307	100.0%	.3989	100.0%	.3989	100.0%
Adjustment for sex differences in								
Experience	−.0056	−1.3	−.0074	−1.7	−.0009	−0.2	−.0017	−0.4
Education	−.0051	−1.2	−.0037	−0.9	+.0170	+4.3	+.0140	+3.5
Class of Worker	−.0218	−5.1	−.0144	−3.3	−.0120	−3.0	−.0418	−10.5
Industry	−.0745	−17.3	−.0901	−20.9	−.0995	−24.9	−.1170	−29.3
Occupation	−.0059	−1.4	−.0338	−7.8	−.0451	−11.3	.0090	−2.3
Health Problems	+.0012	+0.3	+.0017	+0.4	−.0006	−0.2	−.0019	−0.5
Part-time	−.0065	−1.5	−.0273	−6.3	+.0006	+0.2	−.0184	−4.6
Migration	+.0030	+0.7	+.0001	0.0	+.0013	+0.3	−.0002	0.0
Marital Status	−.0078	−1.8	−.0271	−6.3	−.0070	−1.8	−.0157	−3.9
Children	−.0309	−7.2	.0000	0.0	−.0015	−0.4	.0000	0.0
Size of Urban Area	−.0015	−0.3	−.0012	−0.3	−.0030	−0.8	−.0022	−0.6
Region	+.0002	0.0	.0000	0.0	−.0045	−1.1	−.0050	−1.3
	$\ln(\widehat{D+1})=.2755$ ($\hat{D}=.32$)	63.9%	$\ln(\widehat{D+1})=.2275$ ($\hat{D}=.25$)	52.9%	$\ln(\widehat{D+1})=.2437$ ($\hat{D}=.27$)	61.1%	$\ln(\widehat{D+1})=.2000$ ($\hat{D}=.22$)	50.1%

[a] The adjustment for the j-th variable using female regression weights is $-\hat{\beta}_{fj}\Delta\bar{Z}_j$, and therefore the sum is $-\Delta\bar{Z}'\hat{\beta}_f$. This implies $\ln(\widehat{D+1}) = \ln(G+1) - \Delta\bar{Z}'\hat{\beta}_f = -\bar{Z}_m'\Delta\hat{\beta}$.

[b] Each adjustment is expressed as a percentage of the wage differential.

[c] The adjustment for the j-th variable using male regression weights is $-\hat{\beta}_{mj}\Delta\bar{Z}_j$, and therefore the sum is $-\Delta\bar{Z}'\hat{\beta}_m$. This implies $\ln(\widehat{D+1}) = \ln(G+1) - \Delta\bar{Z}'\hat{\beta}_m = -\bar{Z}_f'\Delta\hat{\beta}$.

TABLE 4

THE EFFECTS OF DISCRIMINATION ESTIMATED FROM THE PERSONAL CHARACTERISTICS WAGE REGRESSIONS

Item	Whites				Blacks			
	Female Regression Weights		Male Regression Weights		Female Regression Weights		Male Regression Weights	
	(1)[a]	(2)[b]	(3)[c]	(4)[b]	(1)[a]	(2)[b]	(3)[c]	(4)[b]
Wage differential	.4307	100.0%	.4307	100.0%	.3989	100.0%	.3989	100.0%
Adjusment for sex differences in								
Experience	−.0072	−1.7	−.0094	−2.1	−.0007	−0.2	−.0028	−0.7
Education	−.0122	−2.8	−.0008	−0.2	+.0351	+8.8	+.0190	+4.8
Health Problems	+.0018	+0.4	+.0022	+0.5	−.0015	−0.4	−.0030	−0.8
Part-Time	−.0227	−5.3	−.0459	−10.7	−.0187	−4.7	−.0314	−7.9
Migration	+.0033	+0.8	−.0002	0.0	+.0024	+0.6	−.0004	−0.1
Marital Status	−.0143	−3.3	−.0380	−8.8	−.0086	−2.2	−.0167	−4.2
Children	−.0460	−10.7	.0000	0.0	−.0052	−1.3	.0000	0.0
Size of Urban Area	−.0017	−0.4	−.0012	−0.3	−.0033	−0.8	−.0029	−0.7
Region	+.0003	+0.1	−.0001	0.0	−.0058	−1.5	−.0069	−1.7
	$\ln(\widehat{D+1})=.3320$ ($\hat{D}=.39$)	77.1%	$\ln(\widehat{D+1})=.3373$ ($\hat{D}=.40$)	78.4%	$\ln(\widehat{D+1})=.3913$ ($\hat{D}=.48$)	98.5%	$\ln(\widehat{D+1})=.3538$ ($\hat{D}=.42$)	88.7%

[a] The adjustment for the j-th variable using female regression weights is $-\hat{\beta}_{fj}\Delta\bar{Z}_j$, and therefore the sum is $-\Delta\bar{Z}'\hat{\beta}_f$. This implies $\ln(\widehat{D+1}) = \ln(G+1) - \Delta\bar{Z}'\hat{\beta}_f = -Z_m'\Delta\hat{\beta}$.

[b] Each adjustment is expressed as a percentage of the wage differential.

[c] The adjustment for the j-th variable using male regression weights is $-\hat{\beta}_{mj}\Delta\bar{Z}_j$, and therefore the sum is $-\Delta\bar{Z}'\hat{\beta}_m$. This implies $\ln(\widehat{D+1}) = \ln(G+1) - \Delta\bar{Z}'\hat{\beta}_m = -\bar{Z}_f'\Delta\hat{\beta}$.

the differential. Differences in the mean years of schooling completed widens the differential for blacks because black females complete on average almost a full year more of schooling than black males. It is clear from Table 3 that sex differences in the distributions by class of worker, industry, and occupation significantly narrow the wage differential even though industry and occupation are represented by highly aggregated categories.

Space limitations permit discussion of only a few selected aspects of the regression coefficients reported in Tables 1 and 2. The estimated experience coefficients under both sets of regressions imply that for the same rate of return to OJT, males invest more initially and for a longer period. If both sexes invest the same initially, then the pattern of differences in the coefficients imply that males earn a higher rate of return and invest for a longer period than females. The coefficients on the children variable indicate that each child lowers the white female wage by 2 or 3% but has a negligible effect on the black female wage.[6] This may suggest that black females do not stay out of the labor force as long as white females for each child born.[7] Our results are consistent with those studies of labor supply, such as [3] and [4], that find that the presence of children inhibits the labor force participation of white females significantly more than for black females. Perhaps some form of extended family arrangement in black communities provides a ready source of child care for working mothers. It may also be that lost experience and acquired skills are not very important in the kinds of jobs black females typically hold. For example, 52% of the black

[6] The personal characteristics wage equations were also estimated without the children variable. Generally, the positive coefficients were reduced in magnitude and the negative coefficients became larger in absolute value. Consequently, $-\Delta\hat{\beta}$ was biased upward, which implies a larger estimated effect of discrimination.

[7] The cost of children in terms of their effect on the hourly wage can be translated into an equivalent number of years of potential experience; 1) Set the estimated female experience profile minus the children term equal to zero; and 2) Solve the resulting quadratic for the negative root. Let R_1 be the negative root of

$$\hat{\beta}_2 X^2 + \hat{\beta}_1 X - \hat{\beta}_c C = 0 .$$

The absolute value $|R_1|$ is an estimate of the equivalent number of years of experience. The mean number of children per white female was 1.6. Using the coefficients from the personal characteristics wage regressions, we evaluated $|R_1|$ at $C = 1.6$. The value of $|R_1|$ was approximately 2.5 years. As an estimate of the number of years of experience lost to child care, 2.5 years seems low. Yet this may be a reasonable estimate when one considers that ours is a sample of employed females. A sample from the total population of adult females would include mothers not in the labor force. The years of actual work experience lost would be higher for such a sample. Since we are interested in the average years of lost experience for all employed females, the estimate of 1.6 children per female was computed using the total number of females. However the sample includes women who have never had chidren. Thus the mean number of children per mother and the estimate of their lost experience would be higher. The average ages of white males and females in our sample were 39.5 and 38.9 years, respectively. The average difference in potential experience was 0.6 years. When 0.6 is added to the estimated 2.5 years. we have a total estimated male-female experience difference of 3.1 years. In the Malkiels' study [9, (24)] the difference in actual work experience was 2.9 years in 1971 for males and females whose average ages were 40.8 and 38.9 years, respectively. Another independent source [14, (94)] shows that the difference in years of employment covered by Social Security was 3.2 years in 1960 between males and females in the 35-39 age group.

females in our sample held private household and service worker jobs whereas only 16% of the white females held these jobs.

6. CONCLUDING REMARKS

As in other studies we find the sex differential to be quite large. We are in agreement with other researchers that unequal pay for equal work does not account for very much of the male-female wage differential. Rather it is the concentration of women in lower paying jobs that produces such large differentials. Our results suggest that a substantial proportion of the male-female wage differential is attributable to the effects of discrimination.

The effects of discrimination are estimated as the residual left after adjusting the sex differential for differences in various characteristics. This methodological technique is found in other studies as well and may take the form of regression analysis or standardization analysis. There are some difficulties with this general approach which should be mentioned. Could it be that the wage structures for males and females would differ even in the absence of discrimination? For example, male-female differences in the coefficients of the experience variable suggest that the rate of return to OJT may be higher for males, and/or females invest less in OJT. One might argue that even in the absence of discrimination females may plan on a shorter working life and hence invest less than men. The result would be a difference in the parameters of the experience variables, yet these differences contribute to the effects of discrimination under our analysis. In defense of this approach it should be pointed out that occupational barriers against women deny them the opportunities to invest to the same extent as men. Also, the short work life expectancy of women may represent a rational response to anticipated discrimination in the labor market. The issue becomes one of how much of the male-female difference in the coefficients is due to discrimination.

Another difficulty with the residual approach is that it does not take into account the effects of the feedback from labor market discrimination on the male-female differences in the selected individual characteristics. The differences could reflect the adaptation of women to the biases of the labor market; yet under the residual approach all differences in the characteristics contribute to a reduction of the wage differential attributable to discrimination. The problem becomes one of how much of the observed differences in individual characteristics would exist in the absence of discrimination.

These very difficult problems have not been dealt with in this study, but they are clearly important in terms of policy prescriptions for narrowing the male-female wage differential.

University of Massachusetts, Amherst, U. S. A.

REFERENCES

[1] ASHENFELTER, ORLEY, "Racial Discrimination and Trade Unionism," *Journal of Political Economy*, LXXX (May/June, 1972), 435-464.
[2] BECKER, GARY S., *The Economics of Discrimination* (Chicago: University of Chicago Press, 1957).
[3] BOWEN, WILLIAM G., AND T. A. FINEGAN, *The Economics of Labor Force Participation* (Princeton: Princeton University Press, 1969).
[4] CAIN, GLEN G., *Married Women in the Labor Force: An Economic Analysis* (Chicago: University of Chicago Press, 1966).
[5] COHEN, MALCOM S., "Sex Differences in Compensation," *Journal of Human Resources*, VI (Fall, 1971), 434-447.
[6] FUCHS, VICTOR R., "Differences in Hourly Earnings Between Men and Women," *Monthly Labor Review*, XCIV (May, 1971), 9-15.
[7] GWARTNEY, JAMES, "Discrimination and Income Differentials, *American Economic Review*, LX (June, 1970), 396-408.
[8] JOHNSON, THOMAS, "Returns from Investment in Human Capital," *American Economic Review*, LX (September, 1970), 546-559.
[9] MALKIEL, BURTON G., AND JUDITH A. MALKIEL, "Male-Female Pay Differentials in Professional Employment," Working Paper No. 35, Industrial Relations Section, Princeton University.
[10] MINCER, JACOB, "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach," *Journal of Economic Literature*, VIII (March, 1970), 1-26.
[11] OAXACA, RONALD L., *Male-Female Wage Differentials in Urban Labor Markets*, unpublished Ph. D. dissertation, Department of Economics, Princeton University, (1971).
[12] REES, ALBERT, AND GEORGE P. SHULTZ, *Workers and Wages in an Urban Labor Market* (Chicago: Chicago University Press, 1970).
[13] SANBORN, HENRY, "Pay Differences Between Men and Women," *Industrial and Labor Relations Review*, XVII (July, 1964), 534-550.
[14] U. S. SOCIAL SECURITY ADMINISTRATION, OFFICE OF RESEARCH AND STATISTICS, *Workers Under Social Security, 1960: Annual and Work History Statistics* (Washington, D. C.: U. S. Government Printing Office, 1968).

Bibliographie

[Bay96] Alain Bayet. L'éventail des salaires et ses déterminants. *Données sociales*, 1996.

[Bli73] Alan S Blinder. Wage discrimination : reduced form and structural estimates. *Journal of Human resources*, pages 436–455, 1973.

[CCZ14] Pierre Cahuc, Stéphane Carcillo, and André Zylberberg. *Labor economics*. MIT press, 2014.

[dCH04] Yves de Curraize and Réjane Hugounenq. Inégalités de salaires entre femmes et hommes et discrimination. *Revue de l'OFCE*, (3) :193–224, 2004.

[Gla87] Michel Glaude. La structure des salaires en 1985. *Données sociales*, pages 159–171, 1987.

[Hla14] Marek Hlavac. oaxaca : Blinder-oaxaca decomposition in r. 2014.

[Mdledls01] . Ministere de l'emploi et de la solidarité. Rapport d'activité 2000. pages 94–99, 2001.

[Oax73] Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International economic review*, pages 693–709, 1973.

[Odi11] . Observatoire des inégalités. Mesurer les discriminations : méthodes et résultats. *Les dossiers de l'Observatoire n5*, 2011.

[Per98] A. Perrot. *Les nouvelles théories du marché du travail*. Collection repères. Découverte, 1998.

[PM00] Sophie Ponthieux and Dominique Meurs. Une mesure de la discrimination dans l'écart de salaire entre hommes et femmes. *Economie et Statistique*, 337(1) :135–158, 2000.

[Thi85] Bernard Thiry. La discrimination salariale entre hommes et femmes sur le marché du travail en france. In *Annales de l'INSEE*, pages 39–68. JSTOR, 1985.

[W+08] Matthijs Joost Warrens et al. *Similarity coefficients for binary data : properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. Psychometrics and Research Methodology Group, Leiden University Institute ..., 2008.

[Woo15] Jeffrey Wooldridge. *Introduction à l'économétrie : une approche moderne*. De Boeck Superieur, 2015.

[ZS03] Bin Zhang and Sargur N Srihari. Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*, volume 1, 2003.