IAE NANTES
ÉCONOMIE & MANAGEMENT

UNIVERSITÉ DE NANTES

Introduction au logiciel R

Inégalités des salaires selon l'origine

Ariinui TERIITEHAU

Master 1 EKAP

Sommaire :

I. Introduction

Au début des années 2000, grâce aux données statistiques (INSEE) on a pu constater une importante disparités salariales selon l'origine migratoire de l'individu sur le marché du travail français[1]. Malgré des réformes politique sur la parité des personnes, on assiste toujours à une forte inégalité salariale. Ces écarts sont dûs à des phénomènes discriminatoires et donnent une indication sur cette ampleur par de nombreux facteurs : l'origine, le sexe ou le lieu de résidence.

Ce phénomène peut s'expliquer par une ségrégation des employeurs qui évince certains type d'individu. Par exemple, les individus n'ayant pas la nationalité françaises et qui ne sont pas originaires d'un pays de l'Union Européenne sont exclus d'environ 5,4 millions d'emplois soit un travailleur sur cinq et essentiellement dans le secteur publique. De plus, les personnes provenant d'un ménage d'immigré ont des salaires inférieurs aux non-immigrés car la majorité occupe des emplois précaires et moins qualifiés qui sont plus accessible à cette catégorie de personne.

Nombreuses théories peuvent illustrer cette disparités salariales, par exemple, avec les travaux de Becker (1957) qui sont fondées sur des préférences discriminatoires lequel il explicite l'origine de ces comportements et mesure cette ampleur. Avec Arrow (1971)[2] qui décrit un modèle simple selon lequel un employeur peut acheter de la main d'oeuvre de personne de couleur à un prix fixe. Il désigne un salaire par rapport un point sur une courbe d'indifférence entre les salaires et la proportion de personnes blancs dans l'entreprise.

Il est important de montrer de quelle manière l'origine des personnes peuvent impacter sur l'écart des salaires. Ainsi, on peut se poser la question suivante : Quelles sont les différentes facteurs de cette disparité salariale ?

Dans l'étude, on verra les différents déterminants de l'inégalités salariales des individus selon leurs origines par une analyse **des données provenant de l'INSEE en 2015**. On procédera dans un premier temps à une démarche descriptives des données. Puis, on présente une démarche empirique du modèle utilisé, où on va expliciter les différentes méthodes utilisées et pour finir une conclusion générale de notre problématique sur les différentes causes des ces inégalités selon notre étude et les limites du modèle.

II. Présentation des variables

On va présenté un certain nombre de variables grâce au « Dictionnaire des variables du fichier de données individuelles de l'enquête emploi » de l'INSEE dans le tableau suivant commenter brièvement.

Tableau 1

	Variable	Description	
Variable expliquée	Salaire mensuel SALMEE	Variable quantitative continue importante pour illustrer notre modèle qui représente le salaire mensuel de l'individu.	
Variables explicatives	Âge (age)	Variable quantitative discrète qui représente l'âge de l'individu. Dans le monde professionnel, on remarque qu'une personne âgée gagne plus que les jeunes actifs dû à l'ancienneté dans l'entreprise.	
	Ancienneté (ancentr)	Variable quantitative discrète. Elle indique le nombre d'année passé à travailler dans son entreprise. En effet, si l'individu a une ancienneté importante dans l'entreprise, alors il est fort probable que son salaire est aussi élevé.	
	Nombre de personne dans le logement (nbind)	Variable quantitative discrète, On remarque que les ménage issu d'immigration sont susceptible de vivre à plusieurs dans un ménage. Alors c'est un indicateur important pour voir le salaire est corrélé avec le nombre de personne dans le logement.	
	Sexe (sexe)	Variable qualitative binaire qui désigne le sexe de l'individu. D'après certaines études, on a prouvé une disparité de salaire entre les hommes et les femmes sur le marché du travail	- *Homme* - *Femme*
	Nationalité étrangère (immi)	Variable qualitative binaire. Elle nous indique si la personne est né à l'étranger. Cette variable peut être efficace pour illustrer notre modèle selon l'origine de l'individu et l'impact sur le salaire.	- *Immigrant* - *Non immigrant*
	Descendance étrangère (origine)	Une autre variable binaire que l'on va créer. C'est une variable très complexe c'est-à-dire on cherche à voir si l'individu interrogé a un de ces parents qui n'est pas né sur le sol français ou n'est pas de nationalité française.	- *Origine* - *Pas d'origine*
	Publique ou privé (pp)	Une variable binaire. Elle nous indique si l'individu travaille dans un secteur privé ou publique. Ainsi, on avait vu dans l'introduction que l'accès en emplois publique était difficile pour la population issue d'immigration. De plus, d'après des études, on constate aussi une disparité de salaire entre le publique et le privé.	- *Publique* - *Privé*
	Zone urbaine sensible (zus)	Variable binaire, elle nous montre si l'individu provient d'une zone urbaine sensible. Une variable important pour montrer ces discrimination salariale. Être d'une zone urbaine sensible est « une possible barrières dans l'accès à certains emplois et notamment aux emplois de cadre ».	- *ZUS* - *Pas ZUS*

Tableau 1-2

	Variable	Description	
Variables explicative	Catégorie socio-professionnel (cser)	Variable facteur montrant les différentes « catégories socio-professionnelles pour les actifs ». Le salaire n'est pas le même au gré de la catégorie socio-professionnelle. Ainsi, un individu étant cadre a un salaire supérieure qu'un simple ouvrier.	- *Non renseigné* - *Agriculteur* - *Artisans, commerçant…* - *Cadres …* - *Professions intermédiaires* - *Employés* - *Ouvriers*
	Nature du contrat de travail (ccontr)	Variable facteur. « Type de contrat de travail à l'entrée dans l'entreprise … ». Cette variable est importante dans notre modèle, une personne qui possède un CDI a plus de probabilité d'avoir un salaire plus élevé qu'une personne ayant un contrat d'intérim.	- *Pas de contrat de travail ou pas renseigner* - *Cont. D'apprentissage* - *Cont d'Interim* - *Cont saisonnier* - *CDD* - *CDI*
	Niveau du diplôme (nivp)	Variable facteur qui représente le niveau d'enseignement atteint par l'individu sur 10 postes. Selon le diplôme atteint, il existe évidemment une disparité importante de salaire. Par exemple, la différence de revenu entre un « BAC +2 » et un diplôme supérieur est en moyenne de 953 euros. (3)	- *NIV 1* - *NIV 2* - *NIV 3* - *NIV 4* - *NIV 5*
	Tranche de taille regroupée de l'unité urbaine2010 du logement de résidence (TUU2010R)	Variable facteur. Elle indique l'emplacement du logement de l'individu par unité urbaine selon différente caractéristiques selon le nombre d'habitants.	- *Communauté rurale* - *Unité urbaine de moins de 20 000 hab* - *Unité urbaine entre 20 000 hab et 200K hab* - *Unité urbaine de plus 200K hab* - *Agglo. Parisienne*
	Nationalité (nation)	Variable facteur. Montre si l'individu est de nationalité : française, maghrébine, provenance du continent africain, de l'UE ou Europe, autres.	- *Français* - *Européen* - *Maghreb* - *Africaine* - *Reste du monde*

Exploration de la base de données.

Dans cette partie, on fait le traitement de la base de donnée provenant de « L'enquête emploi » éditer en 2015. Contenant de nombreuse données (110 049 observations pour 689 variables), on va filtrer avec différents critères en utilisant le logiciel R. Tout d'abord, on sélectionne des individus qui nous intéresse avec différentes caractéristiques : individu ayant un salaire et étant actif.

Puis, on doit créer nos trois variables binaire et factoriel (*Origine, pp* et *nation).* Pour la variable *Origine*, on avait vu que c'est une variable qui possède certaine condition : l'individu a un des parents n'est pas née en France **ou** à une nationalité étrangère. Par ailleurs, il y a une limite dans la connaissance de la base car on ne sait pas si l'individu a des origines étrangères, si il est français ou si ses grands-parents sont des immigrés.

Le *PP* on utilise la variable *PUB3FP* où on combine les secteur publique (Etat, collectivité locale, Hôpitaux publiques) et le secteur privé pour avoir une variable binaire.

Pour *nation*, on utilise la variable de la base *nat14* où on lie avec 5 modalités différentes : France, pays du Magreb, de l'Europe, pays d'Afrique et les autres pays mais aussi que l'un des parents sont nées aussi dans des pays différents de la France

Niveau de Diplôme, on catégorise à cinq niveau tel que le *niv 1* représente les personnes ayant un niveau collège, *le niv 2* ayant un niveau bac, *niv 3* les personnes qui ont un niveau bac+2, *niv 4* c'est un niveau bac +3 et *niv 5* pour un niveau bac +5 et plus.

Ensuite, on a intérêt de recoder toutes les variables qu'on attribue pour les facteur, binaire, entier ou continu (voir tableau 1). On utilise la fonction *summary* qui nous résume la moyenne, médiane, l'écart-type, les inter-quartiles et le max-min des échantillonnages. On remarque que le minimum de la base est 1. Ainsi, on décide de sélectionner un salaire supérieur au smic net en 2015 (1 135,99€) vu dans l'INSEE .

III. Analyse statistique univariée

Avant de commencer l'étude de nos données, on procède au nettoyage de la base lequel on supprime les valeurs nulle '*NA*' pour chaque observation puis on nomme cette nouvelle base *BD2.*

De plus, on doit vérifier si il existe des valeurs atypiques pouvant potentiellement fausser notre modélisation. Il convient donc de se débarrasser de ces valeurs. Pour détecter ces valeurs, on utilise des boîtes à moustaches avec la fonction *boxplot* (Annexe1). Ces graphiques nous résument de manière simple la répartition des observations.

On peut employer aussi différent test d'identification de valeur aberrante: *Test de Rosner ou ESD* (Extreme Studentized Deviate), qui sont des fonction dans le programme R. Toutes les variables (*salmee, ancentr, nbind*) à l'exception pour *age* présentaient des valeurs atypiques. On se doit alors de les retirer de la base qu'on appellera *BD3* la nouvelle base et on se retrouve avec une base de 9725 observation et 14 variables.

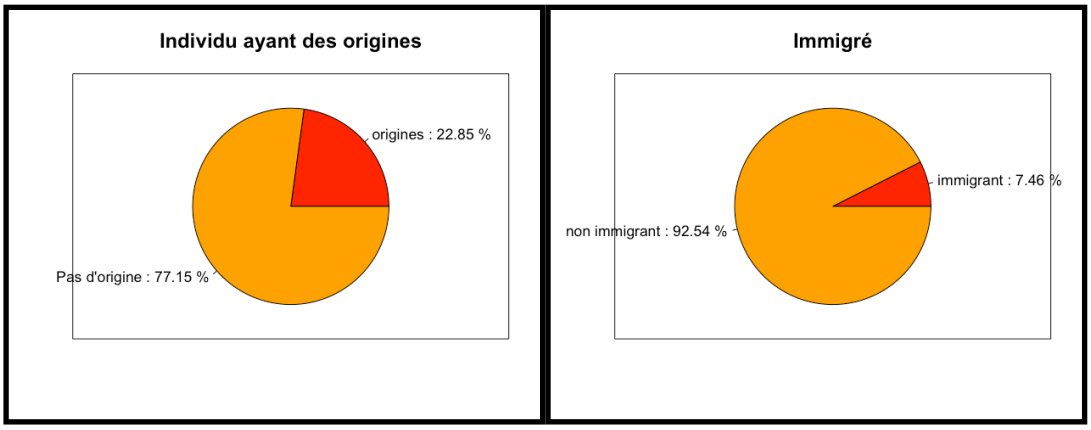
Etude statistique

On fait une étude statistique élémentaire des variables (Annexe2) pour qu'il nous donne la répartition de chaque variable quantitative et qualitative.

A. Etude des variables quantitatives

Pour commencer, on s'intéresse aux nombres d'individus qui ont des origines ou venant de l'immigration. D'après le graphique 2, la population ayant des origines représente 22.85% de l'échantillon avec 2 223 personne contre 77.15% pour 7 507 personne sans origines et il y a 7.46% venant de l'immigration soit 726 personne.

Graphique 2



D'après les résultats dans Annexe2, on remarque que les écart-types des variable quantitative (*salmee, age, nbind*) sont faibles par rapport à la moyenne alors on peut considérer que ces variables sont homogènes. Tandis que pour la variable *ancentr*, on a une variation assez élevé entre la moyenne et l'écart-type mais elle reste aussi homogène.

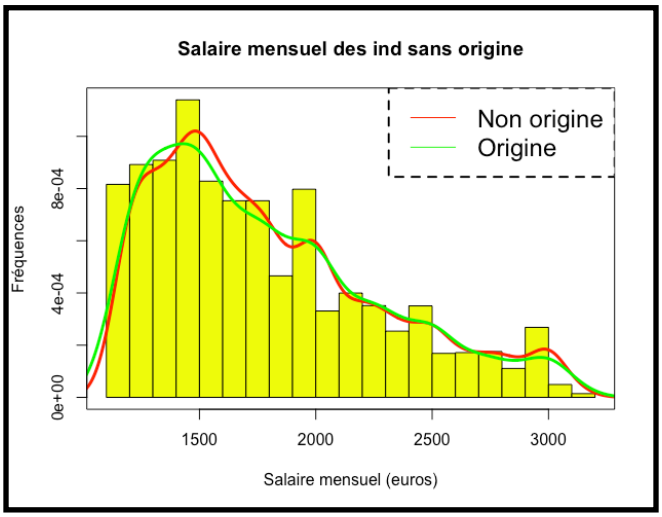
Ensuite, on fait l'analyse de corrélation entre les variables quantitatives. On calcul la matrice de corrélation avec la fonction *cor* (Annexe4). On a une relation dépendance positive supérieur à 50% pour *ancentr* et *age.* Cela peut s'expliquer que dans notre échantillonnage, on a une population assez âgée avec une moyenne de 43 ans et une ancienneté dans l'entreprise moyenne de 15 ans donc on s'attend à cette corrélation importante de 64%, puisque plus l'ancienneté dans une entreprise est importante, plus l'individu est âgé. Autrement, les autres variables se corrèle faiblement entre eux.

Puis, on étudie la répartition de la population pour chaque variable quantitative grâce à des histogramme (annexe5) et leurs courbes de densités. Effectivement, la plus part de la population gagne entre 1250 et 2000 et on a essentiellement des individu âges entre 30 et 55 ans. Dans cette échantillonnage, on trouve une ancienneté dans l'entreprise plutôt récente et qui diminue au fil des années. Enfin, le nombre de personne dans le logement de l'individu varie essentiellement entre 1 et 4.

Ce qui est important pour notre sujet, c'est de s'intéresser comment est réparti le salaire par exemple selon l'origine des personnes afin de comparer cette discrimination alors on va créer deux nouvelles bases pour comparer les populations différentes selon la variable *origine* dont *BD_NORI* qui sont les individus n'ayant pas d'origine et la base *BD_ORIGINE,* les individus ayant des origines.

Ainsi, on va comparer les moyenne des salaires entre les deux bases précédentes (annexe6). Pour une population sans origine, on a une moyenne de 1802 euros alors que pour une population avec des origines, elle n'est que de 1789, soit une différence de 13 euros. On trouve aussi des écart-types qui sont pratiquement la même.
Selon les deux courbes de densités (pour chaque base) du **graphique 3** représentant la répartition des salaires totales, nous montrent que les deux courbes se confondent quasiment entre elles, mis à part quand le salaire se trouve entre 1500 et 1900.

Graphique 3



B. Etude des variables qualitatives

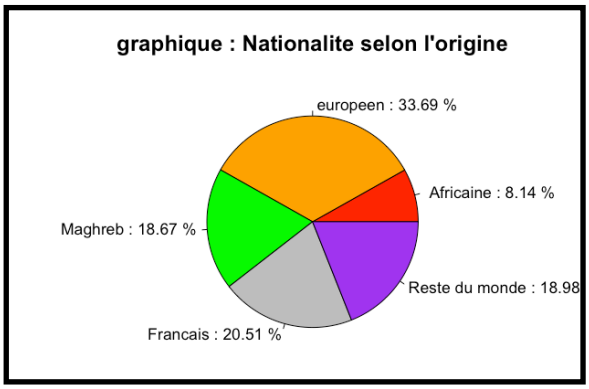
Dans cette partie, on va étudier notre échantillonnage en se référant aux variables qualitatives du modèle avec des croisement de variable. On se réfère comme précédemment avec deux bases (*BD_NORI et BD_ORIGINE*).

On fait deux croisement selon la base et les autres variables qualitatives. Pour représenter ces graphiques, on utilise des « *Pie Chart* » et on crée une fonction « *makepie* » pour faciliter la création de graphique.

- On se renseigne sur la *nation* des personnes selon leurs origines avec le **graphique 4.** Majorité des nationalités étrangères dans l'échantillon sont européenne avec 33.69%,

suivi du Maghreb 18.67 %, puis « le reste du monde » 18.98% et enfin ayant des origines africaine avec 8.14%.

Graphique 4



- On regarde le *nivp* pour les deux populations (annexe9). Ici, on constate pour *BD_ORIGINE* on a 47.95% des individus qui ont un niveau BAC (niv 2) et 9.22% des personnes ont un niveau collège. Tandis que *BD_NORI* , il y a 53.5% pour un niv 2 et 3.12%. Il existe donc une inégalité au niveau de l'éducation pour les personnes immigrées.

- La variable *cser* (annexe10), on trouve pratiquement le même pourcentage pour les deux bases. Mis à part des professions intermédiaires étant plus élevés pour la *BD_NORI* et les ouvriers pour *BD_ORIGINE.*

- Pour *ccontr* (*annexe11*), les individus avec un CDI représentent une part importante dans les deux cas. C'est une volonté d'avoir une sécurité de l'emploi. Mais, les contrat d'Intérim et d'apprentissage se distinguent, *BD_ORIGINE* a 8.55% contre 6.22%, cela s'explique par des emplois peu qualifié et plus accessible pour les migrants.

- *tuu2010r1* (*annexe12*), *BD_ORIGINE* a une population plus importante dans l'agglomération parisienne (28.97%) pour la même raison c'est-à-dire par la facilité d'avoir accès à un emploi dans la capital pour les migrants. Alors que pour *BD_NORI*, on a une population plus concentré dans la communauté rurale (25.96%) et dans l'unité urbaine de moins de 20K habitants (20.46%).

On peut s'intéresser aussi comment la femme est discriminé dans l'accès à l'emploi selon *ccontr*, *tpp* :

- Pour les *ccontr*, les deux bases sont assez similaires, mais d'après les graphiques (*annexe13*) un individu Homme a plus de chance d'avoir un emploi stable q'une femme qui montré par rapport à la proportion des contrats de CDI.

- *tpp*, on constate une même répartition pour les deux bases, mais une fois de plus on voit une discrimination sur le type d'emploi. En effet, les hommes ont plus de contrat

de travail dans un établissement privé que les femmes, tandis que les femmes sont plus représenter dans des entreprises publiques. (*annexe15*)

Après une exploration des données, on fait une reconfiguration des variables qualitatives à plusieurs facteurs afin de les nommer pour avoir une représentation plus explicite. Ensuite, on souhaite les modéliser avec plusieurs régression linéaire que nous allons voir dans la partie suivante.

IV. Régression linéaire multiple

Dans un premier temps, on procède à une régressions linéaire multiple en général puis une régression pour chacun des groupes selon la variable *origine* afin de montrer l'écart des salaires selon l'origine, Cette méthode visent à analyser de façon plus claire les variations de salaires en fonction des multiples facteurs. On a choisis comme références pour les variables qualitatives suivantes : nivdip NIV 2, nation Français, ZUS NON, immi Non immigrant, cser Employés, tuu2010r Unité urbaine de 200k habitants ou plus.

Modele 1

D'après les résultats dans *l'annexe16* La majorité de nos variables sont significatives par rapport aux test de Student dans cette régression mis à part pour les variables : *ccontr, nation* et *pp.* On remarque un coefficient positif pour la variable origines. Par conséquent, on décide de procéder à une nouvelle régression en retirant la variable origine afin d'avoir un modèle plus significatif pour illustrer la disparité salariale. On portera plus d'intérêt à la variable immi dans la suite de nos modèles.

Modele 2

Ce modèle MCO se base donc sur cette équation avec le logarithme pour avoir une forme fonctionnel linéaire plus appropriée :

$$ln(salmee) = \alpha + \beta_1\, age + \beta_2\, ancentr + \beta_3\, nbind + \beta_4\, sexe + \beta_5\, nivp + \beta_5\, zus + \beta_6\, immi + \beta_7\, cser + \beta_8\, tuu2010r + \beta_9\, pp + \epsilon$$

Après avoir procédé différents test classique pour montrer la significativité du modèle (*voir annexe 16)*.On s'intéresse au résultats des coefficients des variables. *Age, ancentr, nbind, nivp (Niv 5,4 et 3), nation (européen), cser, tuu2010r (Agglo, Unit urbaine 20k-200K)* ont des effets positives sur le salaire. Tandis que le reste des variables ont des effets

négatifs. On remarque un écart de salaire marquant et négative selon le sexe de la personne ou une personne provenant de l'immigration. Ce sont des résultats attendus dans notre modèle pour montrer la discrimination raciale mais aussi celui du genre.

De plus, les variables agissant positivement et fortement sur le salaire sont le cser Cadres et Prof… et nivp NIV 5, alors faire de grandes études assure évidemment un meilleur emploi avec un salaire plus élevée.

Table 2

	coefficient	EcartType	Signi
(Intercept)	7.152249	0.011954	Significative
age	0.003350	0.000254	Significative
ancentr	0.000422	0.000021	Significative
nbind	0.007725	0.001597	Significative
sexe1Femme	-0.098503	0.004490	Significative
nivp1NIV 5	0.135809	0.009169	Significative
nivp1NIV 4	0.103855	0.007441	Significative
nivp1NIV 3	0.072492	0.005619	Significative
nivp1NIV 1	-0.079845	0.010433	Significative
nation1 europeen	0.033094	0.008062	Significative
nation1 Maghreb	-0.006400	0.011171	Non significative
nation1 Africaine	-0.019427	0.016643	Non significative
nation1Reste du monde	0.000182	0.010390	Non significative
zus1ZUS OUI	-0.038382	0.010398	Significative
immi1immigrant	-0.028770	0.009950	Significative
cser1Cadres et professions intellectuelles supérieures	0.305972	0.008032	Significative
cser1Ouvriers	0.001556	0.006144	Non significative
cser1Professions intermédiaires	0.141497	0.005581	Significative
cser1Non renseigne	0.132642	0.042345	Significative
tuu2010r1Agglomeration parisienne	0.055455	0.006974	Significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	0.001850	0.006111	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	-0.006692	0.006439	Non significative
tuu2010r1Communaute rurale	-0.007244	0.006214	Non significative

Les coefficients de la **table.3** indique l'écart de salaire des individus dans notre modèle. Par exemple, une personne étant un immigrant vivant dans une zone sensible a un revenu inférieur de 6.45% par rapport à un français. $e^{-0.028-0.038} - 1 = -0.0645.$

Un écart de salaire en faveur des personnes provenant d'Europe, ils ont un revenu supérieur de 3.32%. En effet, les personnes venant de l'Europe sont essentiellement de l'Union Européenne qui immigrent en France pour de meilleur opportunité professionnelle avec un savoir-faire déjà acquis.

De plus, on constate des coefficients négatifs pour des individus ayant des origines africaines et maghrébines. Les populations provenant de l'Afrique ont plus de chances d'avoir des revenus plus faible que la population de référence.

Malgré les tests qu'on a effectué (*voir annexe 16)* , il faudrait faire attention à l'interprétation des coefficients du modèle. En effet, on avait vu que les hypothèses de la MCO ne sont pas vérifiés mais que l'analyse de la covariance pouvait valider nos variables avec une marges d'erreurs. Il est plus approprié d'utiliser d'autres méthodes de régression.

Modèle 3

Pour nos modèles suivant , on procède à deux régression linéaire multiple selon le groupe immi des population afin de faire de comparer l'inégalité des salaire dont on retire les variables immi et nation (*voir annexe 18)*.

Puis on a un alpha différent pour nos deux modèle dont l'un est égale à 7.3186 pour les migrant et l'autre à 7.1129 pour les non migrants. Donc il sera difficile de juger par rapport à ces modèles l'écart salariale. On décide de procéder à une nouvelle régression avec une nouvelle base pour des individus ayant des origines arabes pour démontrer l'écart de salaire en France. On trouve un nouvel alpha qui est égale 7.156 pour la population arabe, qui réduit l'écart avec celle de la population française.

Cependant, les résultats obtenus ne correspondent pas aux attentes pour les coefficients des variables. Par exemple, pour la variable ZUS, le coefficient du modèle de la base française est inférieure à celle du modèle arabe, ce qui suppose que les individus d'origine arabe pourraient gagner plus que les personnes vivant hors des zones urbaines sensibles. De même, pour la variable ccontr, le coefficient est plus important pour le modèle arabe que l'autre modèle. Donc, il n'est pas possible d'interpréter ces résultats. Le problème peut être causer à la combinaison de l'utilisation des variables, de même avec la variables origine dans le modèle 1.

Ainsi, après quelque recherche dans la littérature, on a trouvé une modélisation plus performante pour mettre en évidence la différence de salaire selon la condition de provenance ou pas d' une immigration ou si avoir des origines peut affecter le salaire avec le modèle classique d**'Oaxaca-Blinder**[3]**.**

Modèle Oaxaca-Blinder

Cette méthode a été utilisé pour décomposé les écarts de salaires moyens entre les hommes et femmes (Oaxaca, 1973). Pour notre modèle, on va décomposer selon la variable origine qu'on avait mis de coté.

Avant de lancer notre nouveau modèle, il faut appliquer une modification des variables qualitative à plusieurs facteurs en les mettant à une seule variable binaire et avoir une différence de alpha à l'avantage de la population non immigrante. Ainsi, on a dû retirer nombreuses variables pour avoir un modèle cohérent dans notre étude. (annexe 19)

Avec la fonction Oaxaca, on a pu procéder à une analyse plus performante que celle qui ont été faite précédemment. D'après les résultat du modèle Oaxaca (annexe 19), la différence de salaires est expliqué par le niveau du diplôme et l'appartenance de la catégorie socio-professionnel. Effectivement, une personne avec origine ne possède pas

de diplôme a plus de chance d'avoir de revenu faible et d'être de la classe ouvrière lequel l'individu n'a pas besoin d'avoir un très haut niveau de qualification pour pratiquer son emploi.

Par ailleurs, il faut faire attention à l'interprétation des résultats des modèles. Il faudrait procéder à d'autres tests pour justifier la disparité salarial.

V. Conclusion

Cette étude de cas était prévu pour pratiquer le logiciel R mais aussi de s'intéresser à la disparité de salaires selon l'appartenance d'origine d'une personne. Lors de l'étude, on a procédé à une exploration des données avec des commandes de R afin d'avoir une meilleur analyse sur le phénomène lors des régressions.

Avec une échec de combinaison de plusieurs variable dans la première modélisation pour un coefficient incohérent selon origine, on a su s'adapter pour avoir un modèle plus logique.

Ainsi, d'après les résultats des modèles économétriques, on a fort constater que la majorité des individus dans l'échantillon provenant de l'immigration ou ayant des origines gagne effectivement moins qu'une personne française selon différent facteur. La tendance pour une personne typé dans l'échantillon, tient au fait que la personne provient d'une classe ouvrière et très peu qualifié académiquement lequel le modèle de Oaxaca expliquer ce phénomène. Ainsi pour inverser cette tendance, comme pour les personnes migrantes ont plus d'intérêt de s'installer en île de France comme le montre le graphique dans l'annexe 20 pour avoir un accès plus facile à un emploi.

Cependant, on avait constaté de faible écart de salaire moyens dans le modèle Oaxaca. Cela pourrait s'expliquer lors de l'exploration des données dont on avait sélectionner un échantillon avec un salaire minimum qui est égale au smic de 2015. Mais aussi sur le manque de choix des variables, c'est-à-dire si l'individu était à temps partiel, si il occupé plusieurs emploi pour subvenir à ces besoins. Pour améliorer nos modèles, il faudrait réajuster ces paramètres pour avoir un modèle encore plus performants pour expliquer ce phénomène d'inégalité salariales.

Bibliographie

1. Elika Athari, Jérôme Lê (Insee), Yaël Brinbaum (Cnam-Lise-CEET), « *Le rôle des origines dans la persistance des inégalités d'emploi et de salaire* »,Insee.
https://www.insee.fr/fr/statistiques/4175267?sommaire=4182950

2. Yves de Curraize et Réjane Hugounenq OFCE | « Revue de l'OFCE »
2004/3 n° 90 | pages 193 à 224
ISSN 1265-9576 ISBN 2-7246-2995-7
https://www.cairn.info/revue-de-l-ofce-2004-3-page-193.htm

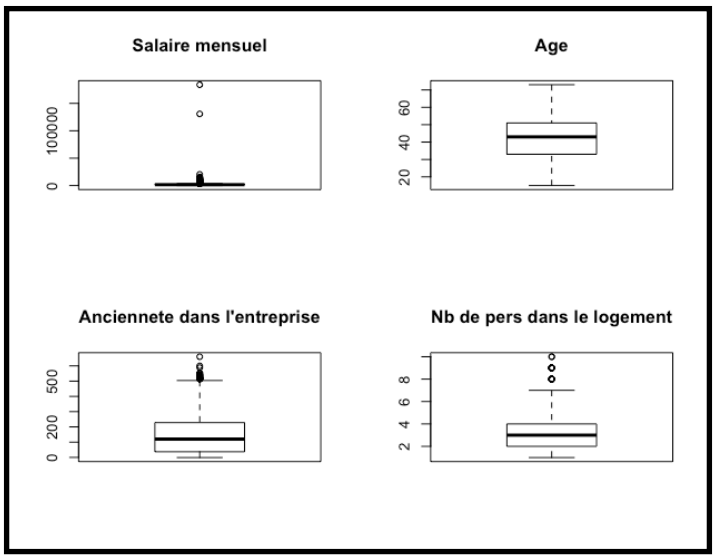
3. 13es Journées de méthodologie statistique de l'Insee (JMS) / 12-14 juin 2018 / PARIS
http://jms-insee.fr/2018/S02_1_ACTE_MAILLARD_JMS2018.pdf

4. Dominique Meurs Sophie Ponthieux, Economie et Statistique , 2000, 337-338, pp. 135-158. https://www.persee.fr/doc/estat_0336-1454_2000_num_337_1_7501

5. Hlavac, Marek (2018). oaxaca: Blinder-Oaxaca Decomposition in R.
R package version 0.1.4. https://CRAN.R-project.org/package=oaxaca

Annexe

Annexe1

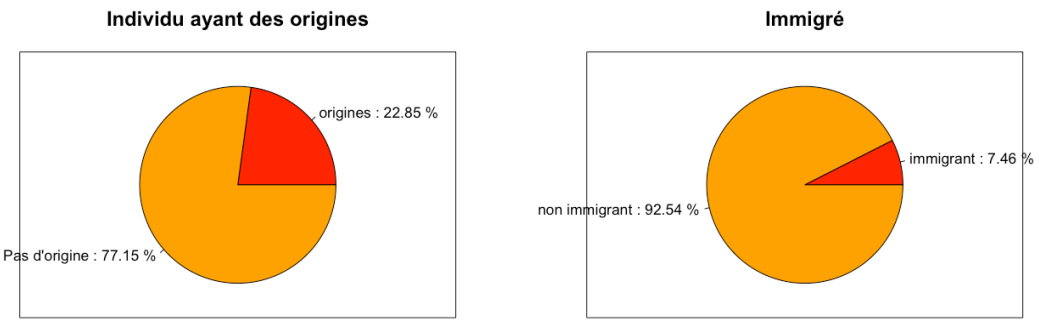


Annexe 2

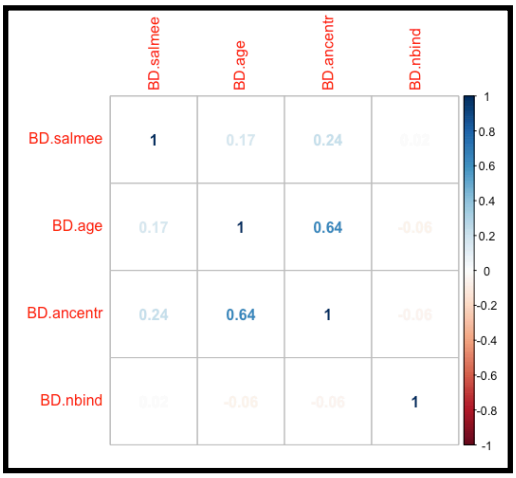
BD.salmee	BD.age	BD.ancentr	BD.nbind	BD.sexe	BD.immi	BD.ORIGINE	BD.pp
Min. :1137	Min. :17.0	Min. : 0.0	Min. :1.000	0:4967	0:8999	0:7503	0:2743
1st Qu.:1400	1st Qu.:33.0	1st Qu.: 48.0	1st Qu.:2.000	1:4758	1: 726	1:2222	1:6982
Median :1700	Median :42.0	Median :132.0	Median :3.000				
Mean :1799	Mean :41.8	Mean :155.4	Mean :2.878				
3rd Qu.:2100	3rd Qu.:51.0	3rd Qu.:240.0	3rd Qu.:4.000				
Max. :3150	Max. :71.0	Max. :480.0	Max. :7.000				

```
> describeBy(BD3)
             vars    n    mean     sd median trimmed    mad  min  max range  skew kurtosis   se
BD.salmee       1 9725 1799.34 486.99   1700 1748.13 444.78 1137 3150  2013  0.81    -0.16 4.94
BD.age          2 9725   41.80  10.66     42   41.92  13.34   17   71    54 -0.07    -0.97 0.11
BD.ancentr      3 9725  155.37 125.94    132  141.65 130.47    0  480   480  0.77    -0.37 1.28
BD.nbind        4 9725    2.88   1.28      3    2.82   1.48    1    7     6  0.27    -0.58 0.01
```

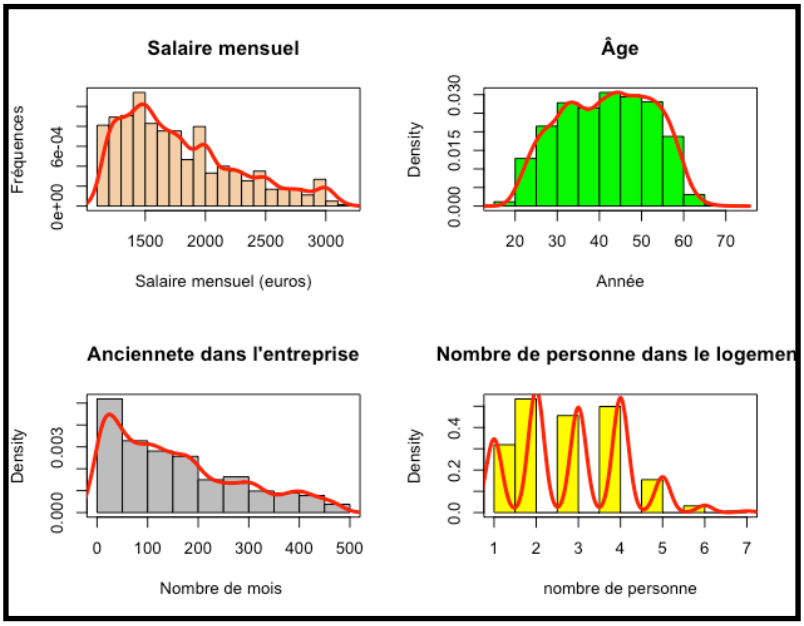
Annexe3



Annexe4



Annexe5

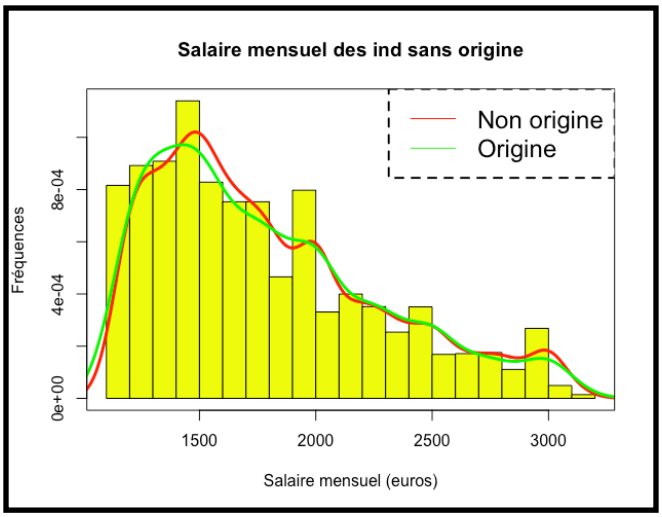


Annexe6

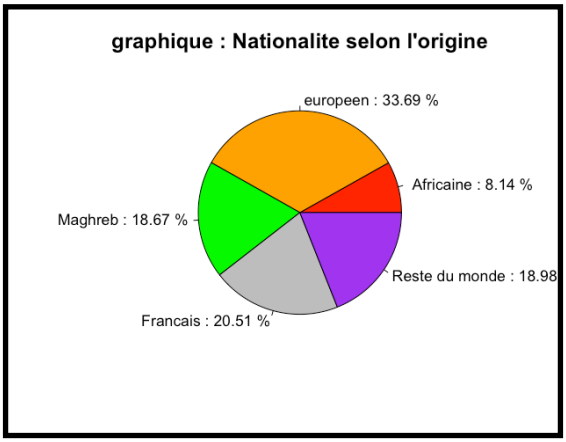
```
> summary(BD_NORI$BD.salmee)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1136    1400    1700    1802    2100    3150
> summary(BD_ORIGINE$BD.salmee)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1136    1400    1700    1789    2100    3150
```

```
> describeBy(BD_NORI$BD.salmee)
   vars    n    mean     sd median trimmed    mad  min  max range skew kurtosis   se
X1    1 7507 1801.82 487.22   1700 1750.34 444.78 1136 3150  2014 0.81    -0.15 5.62
Warning message:
In describeBy(BD_NORI$BD.salmee) : no grouping variable requested
> describeBy(BD_ORIGINE$BD.salmee)
   vars    n    mean     sd median trimmed    mad  min  max range skew kurtosis    se
X1    1 2223 1789.49 486.69   1700 1739.26 444.78 1136 3150  2014  0.8    -0.17 10.32
Warning message:
In describeBy(BD_ORIGINE$BD.salmee) : no grouping variable requested
```

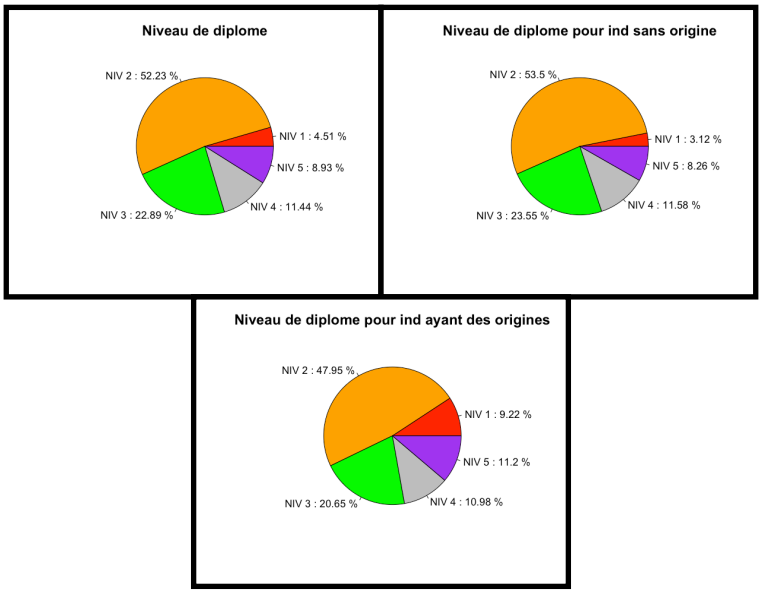
Annexe7



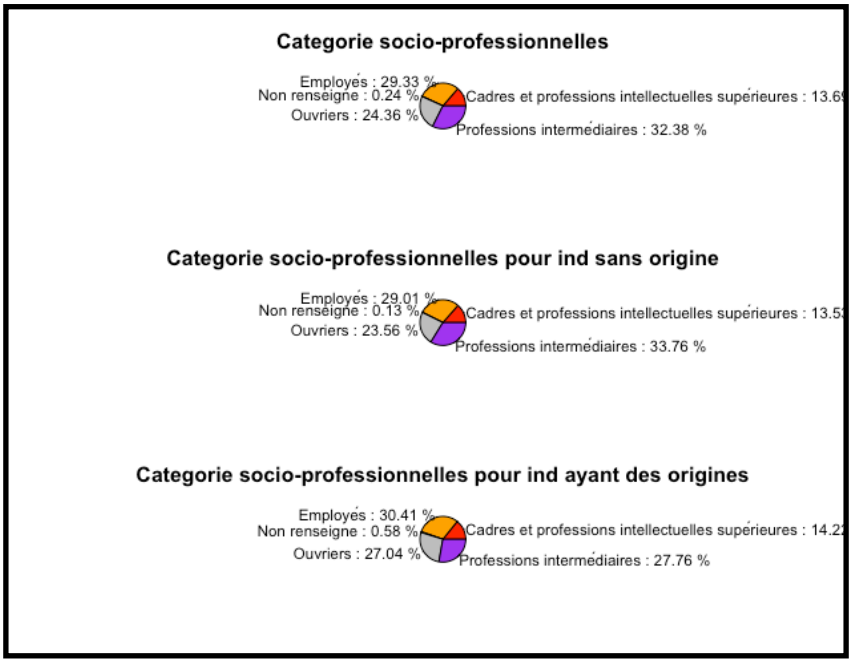
Annexe8



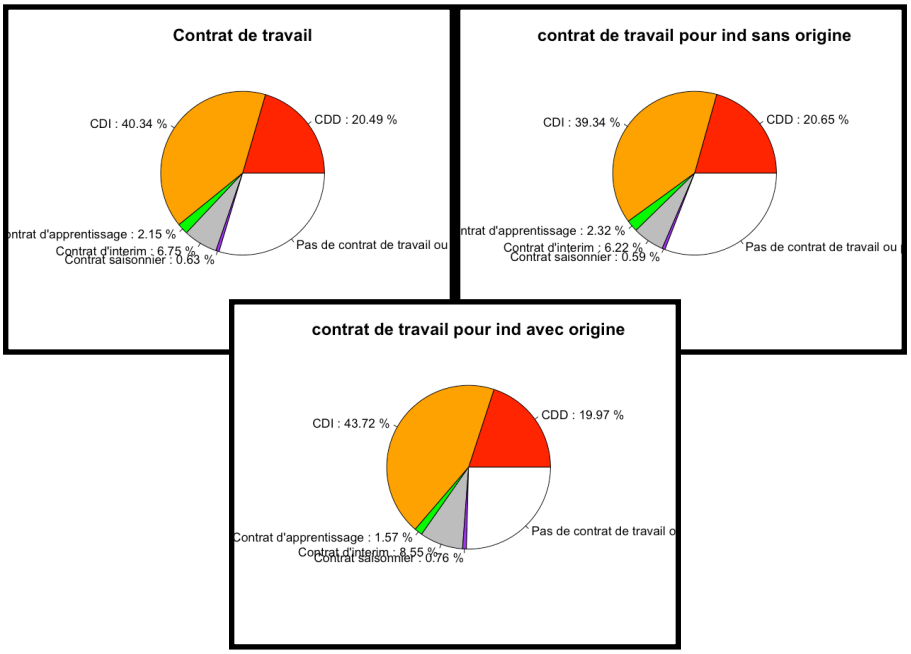
Annexe9



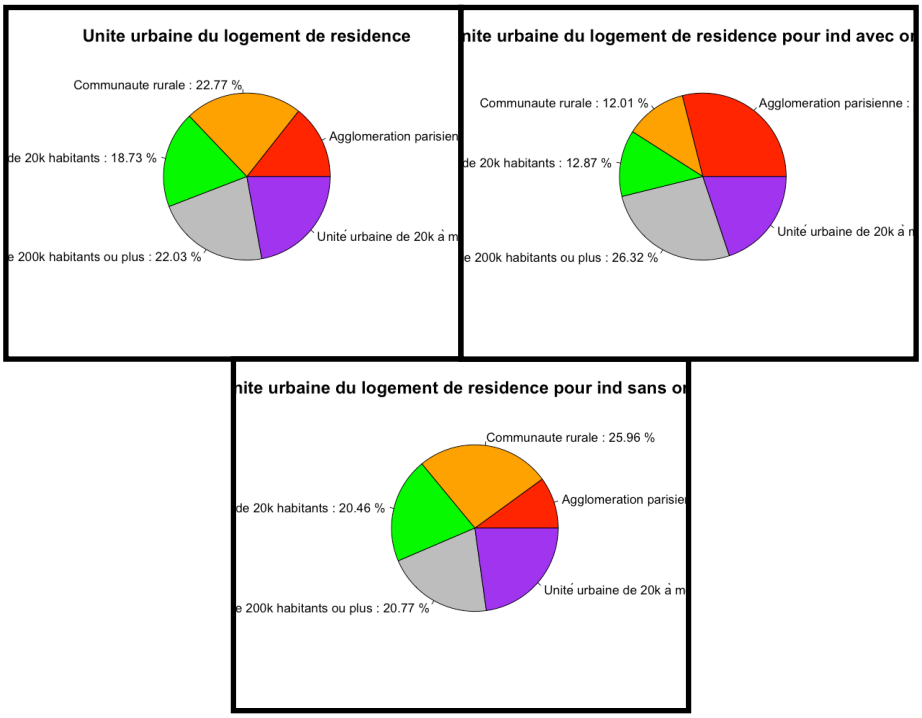
Annexe10



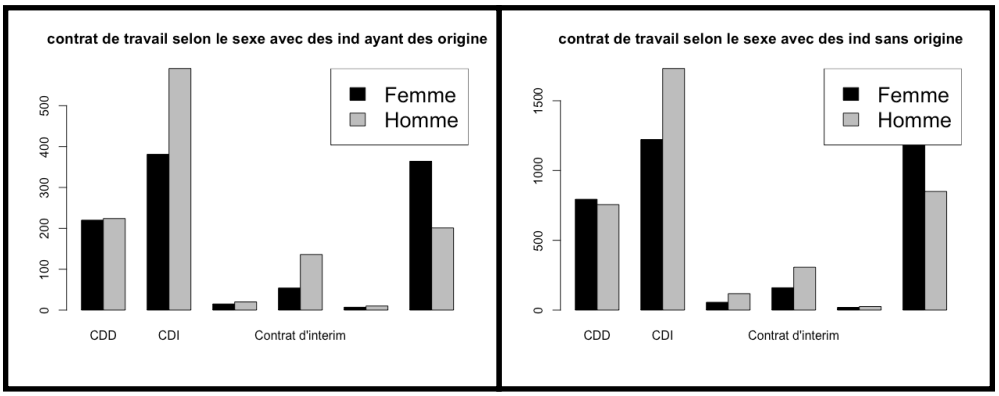
Annexe11



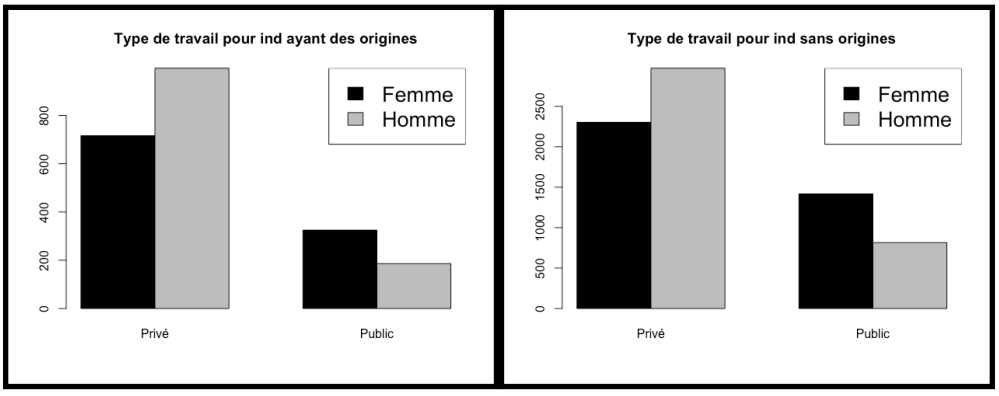
Annexe12



Annexe13



Annexe15



Annexe16

	coefficient	EcartType	Signi
(Intercept)	7.149997	0.011996	Significative
age	0.003338	0.000254	Significative
ancentr	0.000419	0.000021	Significative
nbind	0.007597	0.001597	Significative
sexe1Femme	-0.098691	0.004500	Significative
nivp1NIV 5	0.134298	0.009213	Significative
nivp1NIV 4	0.102144	0.007545	Significative
nivp1NIV 3	0.072048	0.005621	Significative
nivp1NIV 1	-0.080062	0.010432	Significative
nation1 europeen	0.014168	0.011524	Non significative
nation1 Maghreb	-0.025551	0.013898	Non significative
nation1 Africaine	-0.038455	0.018531	Significative
nation1Reste du monde	-0.018328	0.013109	Non significative
zus1ZUS OUI	-0.038670	0.010396	Significative
immi1immigrant	-0.029305	0.009957	Significative
origine1origines	0.021354	0.009196	Significative
cser1Cadres et professions intellectuelles supérieures	0.306645	0.008061	Significative
cser1Ouvriers	0.003162	0.006223	Non significative
cser1Professions intermédiaires	0.142046	0.005587	Significative
cser1Non renseigne	0.132123	0.042379	Significative
tuu2010r1Agglomeration parisienne	0.055646	0.006977	Significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	0.002432	0.006118	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	-0.005954	0.006447	Non significative
tuu2010r1Communaute rurale	-0.006170	0.006227	Non significative
pp1Public	0.005652	0.004794	Non significative

Annexe17

```
Call:
lm(formula = log(salmee) ~ age + ancentr + nbind + sexe1 + nivp1 +
    nation1 + zus1 + immi1 + cser1 + tuu2010r1 + pp1, data = BD3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76676 -0.13766 -0.01335  0.12737  0.86168

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     7.151e+00  1.199e-02 596.586  < 2e-16 ***
age                                             3.355e-03  2.543e-04  13.192  < 2e-16 ***
ancentr                                         4.175e-04  2.147e-05  19.445  < 2e-16 ***
nbind                                           7.665e-03  1.597e-03   4.799 1.62e-06 ***
sexe1Femme                                     -9.887e-02  4.501e-03 -21.968  < 2e-16 ***
nivp1NIV 5                                      1.348e-01  9.212e-03  14.629  < 2e-16 ***
nivp1NIV 4                                      1.024e-01  7.546e-03  13.568  < 2e-16 ***
nivp1NIV 3                                      7.235e-02  5.621e-03  12.872  < 2e-16 ***
nivp1NIV 1                                     -8.003e-02  1.043e-02  -7.670 1.89e-14 ***
nation1 europeen                                3.329e-02  8.064e-03   4.128 3.69e-05 ***
nation1 Maghreb                                -6.342e-03  1.117e-02  -0.568 0.570204
nation1 Africaine                              -1.952e-02  1.664e-02  -1.173 0.240941
nation1Reste du monde                           2.397e-04  1.039e-02   0.023 0.981591
zus1ZUS OUI                                    -3.839e-02  1.040e-02  -3.692 0.000224 ***
immi1immigrant                                 -2.846e-02  9.953e-03  -2.859 0.004252 **
cser1Cadres et professions intellectuelles supérieures  3.068e-01  8.062e-03  38.053  < 2e-16 ***
cser1Ouvriers                                   2.702e-03  6.221e-03   0.434 0.664090
cser1Professions intermédiaires                 1.418e-01  5.587e-03  25.379  < 2e-16 ***
cser1Non renseigne                              1.346e-01  4.238e-02   3.175 0.001501 **
tuu2010r1Agglomeration parisienne               5.573e-02  6.978e-03   7.987 1.54e-15 ***
tuu2010r1Unité urbaine de 20k à moins de 200k habitants  1.733e-03  6.112e-03   0.284 0.776720
tuu2010r1Unité urbaine de - de 20k habitants   -6.735e-03  6.439e-03  -1.046 0.295604
tuu2010r1Communaute rurale                     -7.138e-03  6.214e-03  -1.149 0.250761
pp1Public                                       5.605e-03  4.795e-03   1.169 0.242523
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1978 on 9701 degrees of freedom
Multiple R-squared:  0.4139,    Adjusted R-squared:  0.4125
F-statistic: 297.9 on 23 and 9701 DF,  p-value: < 2.2e-16
```

Après la régression du modèle, on procède aux tests sur la forme linéaire, l'homoscédasticité et on remarque que aucun des deux tests ne sont valides. En effet, on a des X_i qualitatives (multi factorielles) et quantitatives, il est préférable de procéder à une analyse de la covariance des variables avec la fonction Anova sous R. Ainsi, d'après la **table 2**, on a toutes nos variables qui ont un effet sur notre variable Y salmee.

Puis, les résidus du modèle suit une loi normale et on n'a aucun problème de multicolinéarité entre les variables X.

Table 2

```
Analysis of Variance Table

Response: log(salmee)
            Df Sum Sq Mean Sq  F value    Pr(>F)
age          1  19.74  19.743 504.7694 < 2.2e-16 ***
ancentr      1  20.25  20.250 517.7289 < 2.2e-16 ***
nbind        1   0.65   0.646  16.5039 4.893e-05 ***
sexe1        1  11.92  11.922 304.8089 < 2.2e-16 ***
nivp1        4 137.70  34.425 880.1367 < 2.2e-16 ***
nation1      4   1.13   0.281   7.1928 8.864e-06 ***
zus1         1   0.80   0.797  20.3691 6.461e-06 ***
immi1        1   0.75   0.753  19.2575 1.154e-05 ***
cser1        4  71.28  17.820 455.5988 < 2.2e-16 ***
tuu2010r1    4   3.70   0.925  23.6502 < 2.2e-16 ***
pp1          1   0.05   0.053   1.3660    0.2425
Residuals 9701 379.44   0.039
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le table 3 résume la significativité, le coefficient et l'écart-type des variables. On a un R^2 =41%, alors le modèle explique 41% les variations de salaires. On pourrait l'améliorer en ajoutant d'autres variables ou d'observations.

Table 3

	coefficient	EcartType	Signi
(Intercept)	7.152249	0.011954	Significative
age	0.003350	0.000254	Significative
ancentr	0.000422	0.000021	Significative
nbind	0.007725	0.001597	Significative
sexe1Femme	-0.098503	0.004490	Significative
nivp1NIV 5	0.135809	0.009169	Significative
nivp1NIV 4	0.103855	0.007441	Significative
nivp1NIV 3	0.072492	0.005619	Significative
nivp1NIV 1	-0.079845	0.010433	Significative
nation1 europeen	0.033094	0.008062	Significative
nation1 Maghreb	-0.006400	0.011171	Non significative
nation1 Africaine	-0.019427	0.016643	Non significative
nation1Reste du monde	0.000182	0.010390	Non significative
zus1ZUS OUI	-0.038382	0.010398	Significative
immi1immigrant	-0.028770	0.009950	Significative
cser1Cadres et professions intellectuelles supérieures	0.305972	0.008032	Significative
cser1Ouvriers	0.001556	0.006144	Non significative
cser1Professions intermédiaires	0.141497	0.005581	Significative
cser1Non renseigne	0.132642	0.042345	Significative
tuu2010r1Agglomeration parisienne	0.055455	0.006974	Significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	0.001850	0.006111	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	-0.006692	0.006439	Non significative
tuu2010r1Communaute rurale	-0.007244	0.006214	Non significative

Annexe18

Base immigration

```
Call:
lm(formula = log(salmee) ~ age + ancentr + nbind + sexe1 + nivp1 +
    zus1 + pp1 + ccontr1 + cser1 + tuu2010r1, data = BD_IMMI)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66227 -0.15266 -0.02474  0.14186  0.76532

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     7.319e+00  5.175e-02 141.420  < 2e-16 ***
age                                            -2.379e-04  9.633e-04  -0.247  0.80498
ancentr                                         4.111e-04  9.135e-05   4.500 7.96e-06 ***
nbind                                           7.083e-04  5.624e-03   0.126  0.89981
sexe1Femme                                     -7.111e-02  1.902e-02  -3.739  0.00020 ***
nivp1NIV 5                                      6.598e-02  3.396e-02   1.943  0.05244 .
nivp1NIV 4                                      3.467e-02  3.273e-02   1.059  0.28980
nivp1NIV 3                                      5.047e-02  2.592e-02   1.947  0.05196 .
nivp1NIV 1                                     -6.711e-02  2.244e-02  -2.991  0.00288 **
zus1ZUS OUI                                    -5.198e-02  2.218e-02  -2.343  0.01939 *
pp1Public                                      -4.927e-02  3.539e-02  -1.392  0.16425
ccontr1CDI                                      2.659e-02  2.144e-02   1.241  0.21517
ccontr1Contrat d'apprentissage                 -9.667e-02  9.930e-02  -0.974  0.33064
ccontr1Contrat d'interim                        2.370e-02  3.252e-02   0.729  0.46638
ccontr1Contrat saisonnier                      -3.109e-02  8.393e-02  -0.370  0.71114
ccontr1Pas de contrat de travail ou pas renseigne  3.502e-02  3.578e-02  0.979  0.32800
cser1Cadres et professions intellectuelles supérieures  3.572e-01  3.620e-02  9.868  < 2e-16 ***
cser1Ouvriers                                  -9.498e-04  2.312e-02  -0.041  0.96724
cser1Professions intermédiaires                 1.222e-01  2.566e-02   4.763 2.31e-06 ***
cser1Non renseigne                              1.965e-01  1.560e-01   1.260  0.20827
tuu2010r1Agglomeration parisienne               3.747e-05  2.217e-02   0.002  0.99865
tuu2010r1Unité urbaine de 20k à moins de 200k habitants  9.752e-03  2.594e-02  0.376  0.70709
tuu2010r1Unité urbaine de - de 20k habitants    2.625e-02  3.163e-02   0.830  0.40697
tuu2010r1Communaute rurale                     -1.629e-03  3.613e-02  -0.045  0.96405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2143 on 702 degrees of freedom
Multiple R-squared:  0.3431,    Adjusted R-squared:  0.3215
F-statistic: 15.94 on 23 and 702 DF,  p-value: < 2.2e-16
```

Base non immigration

```
Call:
lm(formula = log(salmee) ~ age + ancentr + nbind + sexe1 + nivp1 +
    zus1 + pp1 + ccontr1 + cser1 + tuu2010r1, data = BD_NOIMMI)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70532 -0.13619 -0.01176  0.12514  0.89642

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     7.113e+00  1.290e-02 551.413  < 2e-16 ***
age                                             3.483e-03  2.680e-04  12.995  < 2e-16 ***
ancentr                                         4.244e-04  2.211e-05  19.198  < 2e-16 ***
nbind                                           8.325e-03  1.658e-03   5.022 5.22e-07 ***
sexe1Femme                                     -9.900e-02  4.624e-03 -21.409  < 2e-16 ***
nivp1NIV 5                                      1.395e-01  9.554e-03  14.605  < 2e-16 ***
nivp1NIV 4                                      1.086e-01  7.710e-03  14.088  < 2e-16 ***
nivp1NIV 3                                      7.517e-02  5.727e-03  13.126  < 2e-16 ***
nivp1NIV 1                                     -9.097e-02  1.224e-02  -7.434 1.15e-13 ***
zus1ZUS OUI                                    -3.343e-02  1.190e-02  -2.809  0.00498 **
pp1Public                                       1.077e-02  9.040e-03   1.192  0.23341
ccontr1CDI                                      4.995e-02  5.683e-03   8.789  < 2e-16 ***
ccontr1Contrat d'apprentissage                 -6.751e-03  1.464e-02  -0.461  0.64477
ccontr1Contrat d'interim                        4.546e-02  9.396e-03   4.839 1.33e-06 ***
ccontr1Contrat saisonnier                      -4.364e-02  2.700e-02  -1.617  0.10598
ccontr1Pas de contrat de travail ou pas renseigne  2.870e-02  9.494e-03  3.023  0.00251 **
cser1Cadres et professions intellectuelles supérieures  2.954e-01  8.252e-03  35.798  < 2e-16 ***
cser1Ouvriers                                   4.178e-04  6.550e-03   0.064  0.94914
cser1Professions intermédiaires                 1.385e-01  5.715e-03  24.243  < 2e-16 ***
cser1Non renseigne                              1.388e-01  4.464e-02   3.108  0.00189 **
tuu2010r1Agglomeration parisienne               6.223e-02  7.401e-03   8.408  < 2e-16 ***
tuu2010r1Unité urbaine de 20k à moins de 200k habitants  1.676e-03  6.246e-03  0.268  0.78840
tuu2010r1Unité urbaine de - de 20k habitants   -7.858e-03  6.522e-03  -1.205  0.22832
tuu2010r1Communaute rurale                     -6.503e-03  6.261e-03  -1.039  0.29903
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1951 on 8975 degrees of freedom
Multiple R-squared:  0.4267,    Adjusted R-squared:  0.4252
F-statistic: 290.4 on 23 and 8975 DF,  p-value: < 2.2e-16
```

	coefficient	EcartType	Signi
(Intercept)	7.318562	0.051750	Significative
age	-0.000238	0.000963	Non significative
ancentr	0.000411	0.000091	Significative
nbind	0.000708	0.005624	Non significative
sexe1Femme	-0.071112	0.019017	Significative
nivp1NIV 5	0.065983	0.033963	Non significative
nivp1NIV 4	0.034668	0.032725	Non significative
nivp1NIV 3	0.050468	0.025924	Non significative
nivp1NIV 1	-0.067108	0.022435	Significative
zus1ZUS OUI	-0.051981	0.022182	Significative
pp1Public	-0.049275	0.035389	Non significative
ccontr1CDI	0.026592	0.021435	Non significative
ccontr1Contrat d'apprentissage	-0.096673	0.099305	Non significative
ccontr1Contrat d'interim	0.023699	0.032519	Non significative
ccontr1Contrat saisonnier	-0.031094	0.083932	Non significative
ccontr1Pas de contrat de travail ou pas renseigne	0.035020	0.035777	Non significative
cser1Cadres et professions intellectuelles supérieures	0.357226	0.036201	Significative
cser1Ouvriers	-0.000950	0.023118	Non significative
cser1Professions intermédiaires	0.122242	0.025663	Significative
cser1Non renseigne	0.196463	0.155984	Non significative
tuu2010r1Agglomeration parisienne	0.000037	0.022171	Non significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	0.009752	0.025943	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	0.026249	0.031635	Non significative
tuu2010r1Communaute rurale	-0.001629	0.036126	Non significative

	coefficient	EcartType	Signi
(Intercept)	7.112871	0.012899	Significative
age	0.003483	0.000268	Significative
ancentr	0.000424	0.000022	Significative
nbind	0.008325	0.001658	Significative
sexe1Femme	-0.098999	0.004624	Significative
nivp1NIV 5	0.139538	0.009554	Significative
nivp1NIV 4	0.108613	0.007710	Significative
nivp1NIV 3	0.075171	0.005727	Significative
nivp1NIV 1	-0.090975	0.012238	Significative
zus1ZUS OUI	-0.033426	0.011898	Significative
pp1Public	0.010773	0.009040	Non significative
ccontr1CDI	0.049948	0.005683	Significative
ccontr1Contrat d'apprentissage	-0.006751	0.014642	Non significative
ccontr1Contrat d'interim	0.045463	0.009396	Significative
ccontr1Contrat saisonnier	-0.043644	0.026996	Non significative
ccontr1Pas de contrat de travail ou pas renseigne	0.028700	0.009494	Significative
cser1Cadres et professions intellectuelles supérieures	0.295406	0.008252	Significative
cser1Ouvriers	0.000418	0.006550	Non significative
cser1Professions intermédiaires	0.138545	0.005715	Significative
cser1Non renseigne	0.138773	0.044645	Significative
tuu2010r1Agglomeration parisienne	0.062230	0.007401	Significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	0.001676	0.006246	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	-0.007858	0.006522	Non significative
tuu2010r1Communaute rurale	-0.006503	0.006261	Non significative

Base arabe

	coefficient	EcartType	Signi
(Intercept)	7.156471	0.056938	Significative
age	0.001093	0.001173	Non significative
ancentr	0.000530	0.000127	Significative
nbind	0.013864	0.007026	Significative
sexe1Femme	-0.070716	0.021997	Significative
nivp1NIV 5	0.135135	0.037857	Significative
nivp1NIV 4	0.135888	0.038407	Significative
nivp1NIV 3	0.076383	0.025823	Significative
nivp1NIV 1	-0.066016	0.035989	Non significative
zus1ZUS OUI	-0.023831	0.024098	Non significative
pp1Public	-0.025321	0.044244	Non significative
ccontr1CDI	0.060766	0.025510	Significative
ccontr1Contrat d'apprentissage	0.050569	0.076493	Non significative
ccontr1Contrat d'interim	0.052297	0.035552	Non significative
ccontr1Contrat saisonnier	0.122179	0.101177	Non significative
ccontr1Pas de contrat de travail ou pas renseigne	-0.002463	0.046015	Non significative
cser1Cadres et professions intellectuelles supérieures	0.333529	0.040575	Significative
cser1Ouvriers	0.020037	0.028763	Non significative
cser1Professions intermédiaires	0.099015	0.027113	Significative
tuu2010r1Agglomeration parisienne	0.026609	0.023090	Non significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	-0.017916	0.027895	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	0.012280	0.041003	Non significative
tuu2010r1Communaute rurale	0.060106	0.047864	Non significative

	coefficient	EcartType	Signi
(Intercept)	7.112871	0.012899	Significative
age	0.003483	0.000268	Significative
ancentr	0.000424	0.000022	Significative
nbind	0.008325	0.001658	Significative
sexe1Femme	-0.098999	0.004624	Significative
nivp1NIV 5	0.139538	0.009554	Significative
nivp1NIV 4	0.108613	0.007710	Significative
nivp1NIV 3	0.075171	0.005727	Significative
nivp1NIV 1	-0.090975	0.012238	Significative
zus1ZUS OUI	-0.033426	0.011898	Significative
pp1Public	0.010773	0.009040	Non significative
ccontr1CDI	0.049948	0.005683	Significative
ccontr1Contrat d'apprentissage	-0.006751	0.014642	Non significative
ccontr1Contrat d'interim	0.045463	0.009396	Significative
ccontr1Contrat saisonnier	-0.043644	0.026996	Non significative
ccontr1Pas de contrat de travail ou pas renseigne	0.028700	0.009494	Significative
cser1Cadres et professions intellectuelles supérieures	0.295406	0.008252	Significative
cser1Ouvriers	0.000418	0.006550	Non significative
cser1Professions intermédiaires	0.138545	0.005715	Significative
cser1Non renseigne	0.138773	0.044645	Significative
tuu2010r1Agglomeration parisienne	0.062230	0.007401	Significative
uu2010r1Unité urbaine de 20k à moins de 200k habitants	0.001676	0.006246	Non significative
tuu2010r1Unité urbaine de - de 20k habitants	-0.007858	0.006522	Non significative
tuu2010r1Communaute rurale	-0.006503	0.006261	Non significative

Remarque : la modélisation avec la base Arabe, on trouve le même problème du coefficient de alpha (intercept) lequel il est supérieur à celui de la base non immigrante. Il n'est pas possible de comparer avec ces résultats. Il faudrait retirer des variables et garder des facteurs plus cohérent pour modéliser l'écart de salaire.

Annexe19

Modèle Oaxaca

D'après une analyse des variables qualitative et des résultat précédents par rapport aux coefficients, on décide de choisir les variables binaire selon la variable *origine*. Ainsi, en regardant les graphiques dans la partie Analyse statistique univariée, on décide de garder les variables suivantes :

- le niveau de diplôme : *NIV1,NIV2,NIV3,NIV4*.
- type de contrat de travail : *CDI, CDD, contrat d'intérim(CI), contrat saisonnier(CA).*
- catégorie socio-professionnel : *employé (EMPL), ouvrier (OUV),*
- la taille l'unité urbaine : *agglo. parisienne, unité urbaine de moins 20 000 hab., entre 20 000 et 20 000 hab., et plus de 20 000 hab.*

Avant de lancer notre nouveau modèle, il faut appliquer une modification des variables qualitative à plusieurs facteurs en les mettant à une seule variable binaire et avoir une différence de alpha à l'avantage de la population non origine. Ainsi, on a dû retirer nombreuses variables pour avoir un modèle cohérent dans notre étude.

Ensuite, on peut modéliser distinctement deux modèle de MCO qu'on nomme le groupe A et B:

$$Y_A = \beta_{A0} + \sum X_{i,k}\beta_{A,k} + \epsilon_i \quad , \quad \forall i \in A, \text{ population avec origine.}$$

$$Y_B = \beta_{B0} + \sum X_{i,k}\beta_{B,k} + \epsilon_i \quad , \quad \forall i \in B, \text{ population sans origine.}$$

Les coefficients dans la table 4 des deux modèles paraissent cohérents par rapport au modèle 3, on a un alpha avantagé pour le groupe B sans origine qui est de 7.75 et A avec origine qui est égale à 7.72.

Après l'estimation des deux modèles, on calcul la moyenne des salaires des deux groupes qui sont différentes entre elles tel que :

$$\overline{Y_A} = \widehat{\beta_{AO}} + \sum_{k=1}^{k} \overline{X}_{Ak} + \widehat{\beta_{Ak}} \ , \qquad \overline{Y_B} = \widehat{\beta_{BO}} + \sum_{k=1}^{k} \overline{X}_{Bk} + \widehat{\beta_{Bk}}$$

On fait la différences des deux moyennes $\overline{Y_B} - \overline{Y_A}$ afin de bien mettre en évidence la disparité des salaires. Et on trouve une partie de l'équation qui est expliqué (effet de composition), l'autre part qui inexpliqué (écart de constant).

$$\overline{Y_B} - \overline{Y_A} = 0.00738 \text{ avec } \overline{Y}_A = 7.455262 \text{ et } \overline{Y}_B = 7.462645$$

Table 5

group.weight	coef(explained)	se(explained)	coef(unexplained)	se(unexplained)
0	0.015	0.005	-0.008	0.006
1	-0.001	0.005	0.009	0.006
0.5	0.007	0.004	0	0.005
0.772	0.002	0.004	0.005	0.006
-1	0.005	0.004	0.002	0.005
-2	0.005	0.004	0.002	0.005

Table.5 indique les résultat de la décomposition agrégée. La deuxième ligne montre l'écart expliqué qui vaut -0.001 (vu pour l'effet de composition) et correspond à la différence entre le salaire que toucherait une personne typé avec les caractéristiques moyennes d'une personne non typé. Toutefois, l'écart expliqué est très faible.

Table 4

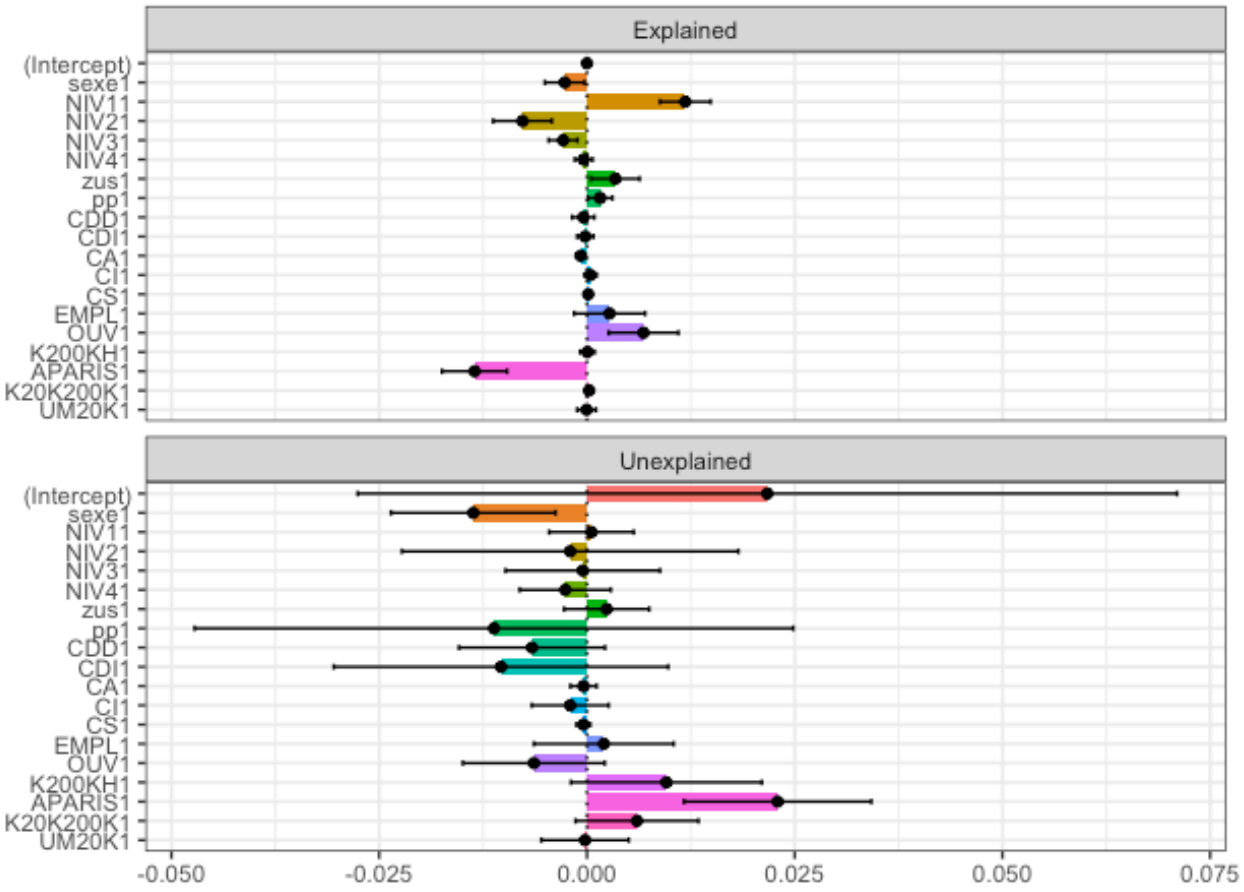
	coeffs.A	coeffs.B
(Intercept)	7.728	7.75
sexe1	-0.07	-0.099
NIV11	-0.2	-0.194
NIV21	-0.135	-0.139
NIV31	-0.098	-0.1
NIV41	-0.041	-0.065
zus1	-0.063	-0.041
pp1	-0.009	-0.024
CDD1	-0.03	-0.063
CDI1	0.028	0.005
CA1	-0.074	-0.101
CI1	0.005	-0.019
CS1	-0.032	-0.088
EMPL1	-0.199	-0.192
OUV1	-0.174	-0.198
K200KH1	-0.037	-0.001
APARIS1	-0.008	0.071
20K200K1	-0.021	0.009
UM20K1	0.001	-0.001

Table 6

	coef(explained)	se(explained)	coef(unexplained)	se(unexplained)
(Intercept)	0	0	0.02170939	0.02515323
sexe1	-0.00267329	0.00122133	-0.01367821	0.00505437
NIV11	0.01183291	0.0015586	0.00053998	0.00259828
NIV21	-0.0077555	0.00180449	-0.00202335	0.01033221
NIV31	-0.00287393	0.00088857	-0.00050333	0.00474623
NIV41	-0.00039176	0.00054794	-0.00261299	0.00280207

Le table.6 montre la contribution de chaque variable à l'écart expliqué et inexpliqué

Graphique A



Ensuite, on illustre dans le graphique A le modèle Oaxaca avec la fonction sous R pour une meilleur compréhension.
Les variables qui contribue fortement à l'écart expliqué sont les variables : NIV1 et OUV1, avec une contribution respective de 0.01 et 0.006. En conséquence, presque l'intégralité de la différence de salaires moyens entre la provenance une personne avec des origine ne sont pas diplômé et provient d'une catégorie professionnel d'ouvrière.

Les variables qui contribuent moyennement sont les variables empl et zus. Ainsi, le fait d'être dans une zone urbaine sensible et d'être un employé ne favorise pas le salaire d'un individu ayant des origines.

Il est possible que l'inclusion de certaines variables peut réduire l'écart inexpliqué : par exemple le cas pour certains niveaux de diplôme ou le fait d'habiter en agglomération parisienne. En effet, quand une personne avec des origines vit en région parisienne et a un niveau de diplôme élevé, alors en termes de rémunération sera plus important pour lui.

- Décomposition par repondération

De plus, pour montrer cette discrimination salarial, on peut procéder à une **décomposition par repondération** en étudiant la distribution des salaires d'individu originaire du Maghreb et non typés. Dans un premier temps, on estime la probabilité conditionnelle d'appartenir au groupe d'origine avec la régression logit dans R en gardant les mêmes variables que dans le modèle Oaxaca-Blinder.

Table 7

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.79230243	0.3046840	-15.7287631	9.606944e-56
sexe1	-0.16022001	0.1178384	-1.3596583	1.739381e-01
NIV11	0.02764290	0.2570104	0.1075555	9.143483e-01
NIV21	-0.49257541	0.1917232	-2.5692002	1.019335e-02
NIV31	-0.21958061	0.1878884	-1.1686759	2.425343e-01
NIV41	-0.46149344	0.2290501	-2.0148140	4.392414e-02
zus1	1.53266936	0.1406905	10.8939099	1.232334e-27
pp1	0.09480042	0.2415871	0.3924068	6.947577e-01
CDD1	0.26767050	0.2501838	1.0698955	2.846663e-01
CDI1	0.11620196	0.2421019	0.4799712	6.312479e-01
CA1	0.06627199	0.4576511	0.1448090	8.848617e-01
CI1	0.66038266	0.2854660	2.3133499	2.070341e-02
CS1	0.94561626	0.6020606	1.5706329	1.162679e-01
EMPL1	0.15078959	0.1446715	1.0422897	2.972774e-01
OUV1	0.46394635	0.1605896	2.8890178	3.864472e-03
K200KH1	1.78306095	0.2532590	7.0404631	1.916020e-12
APARIS1	2.61164842	0.2516855	10.3766364	3.167385e-25
K20K200K1	1.28656861	0.2616786	4.9165987	8.806081e-07
UM20K1	0.66929386	0.2975319	2.2494862	2.448158e-02

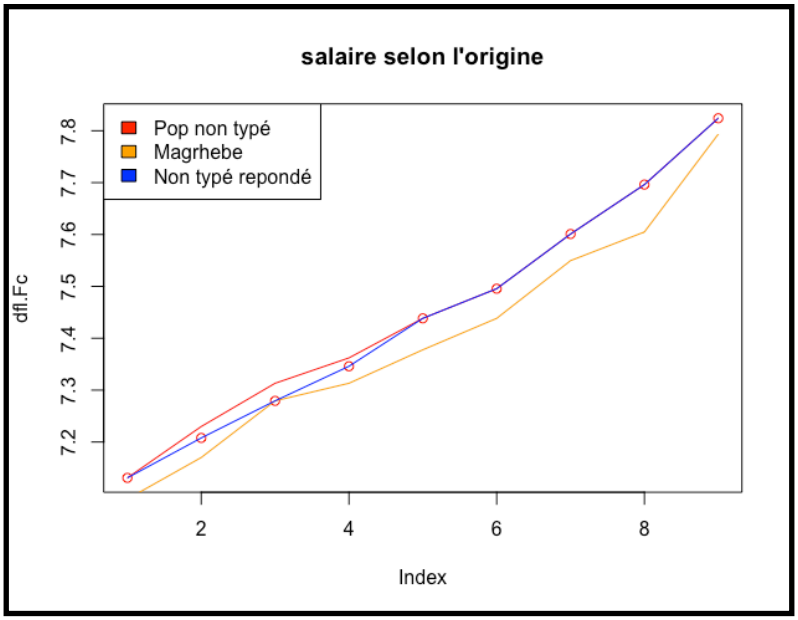
La table 7 résume la régression logit et les coefficients indique la probabilité conditionnelle cité dans le paragraphe précédent. Ainsi, il est 13.4 fois plus probables qu'une personne maghrébine vit au agglomération parisienne et 4.48 fois pour être dans une zone urbaine sensible par rapport à une population témoin ($e^{2.6} = 13.4$).

On effectue le calcul de la pondération pour avoir la distribution de salaire sous certaines conditions. La fonction sous R calcule les quantiles en incluant les pondérations. On l'illustre dans le graphique B , il représente les déciles de log-salaire pour la distribution contrefactuelle (repondération de la distribution des salaires selon la population témoin).

Pour la première décile et à partir de la cinquième décile, on remarque que la distribution contrefactuelle est éloigné de celle de la courbe des personnes originaires du Maghreb et se confond avec la courbe de la population témoin. On constate que l'écart de l'effet de composition est nul alors les disparités de lieu d'habitat entre les deux groupes n'expliquent pas entièrement les différences de salaires.

Par ailleurs, entre le décile 2 et 4 avec un salaire d'environ de 1210 et 1495 euros, l'écart de salaire est mieux expliqué par l'effet de composition. Cette intervalle correspond à l'analyse vu dans le graphe 3 p.8.

Graphique B



Annexe20

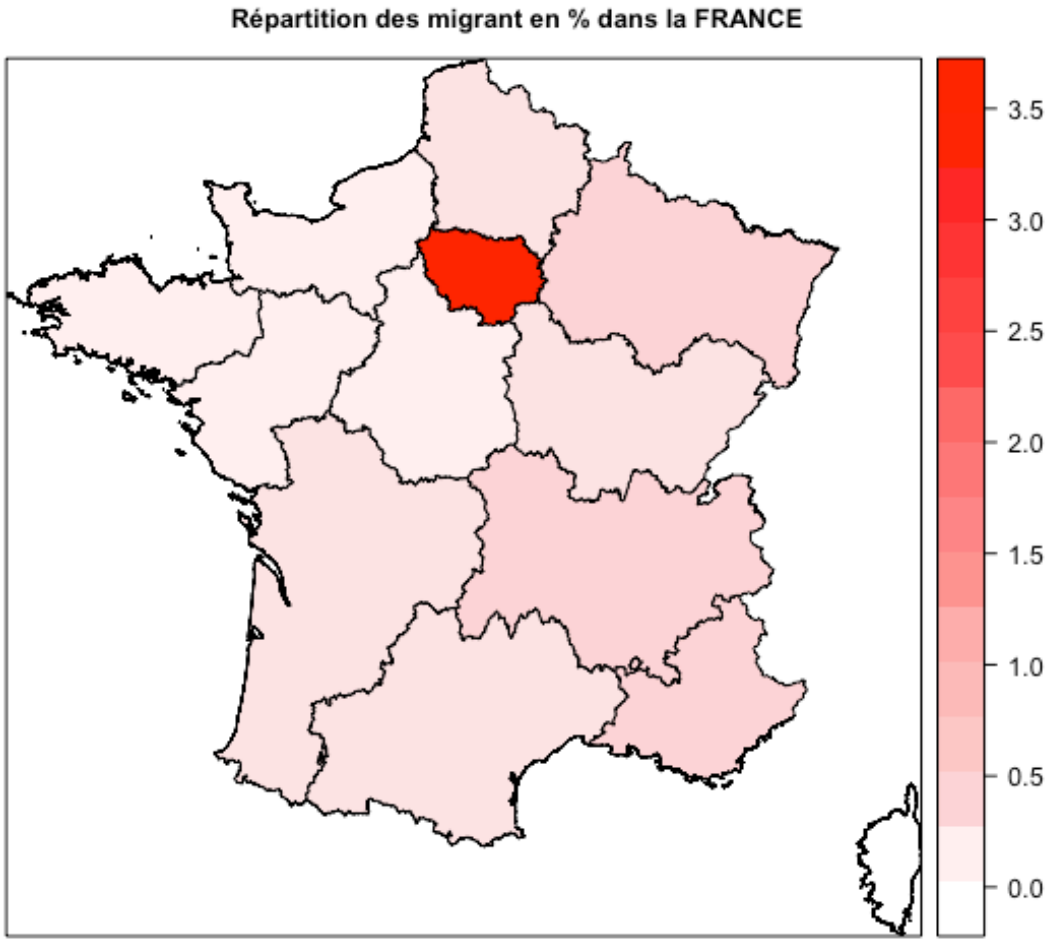

Répartition des migrant en % dans la FRANCE
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Table des matières

Script R

```
### DOSSIER INTRODUCTION A R
### DOSSIER INTRODUCTION A R


###############
#fonction

#creation de la fonction recodage
recodage <- function(varx, code){
  library(carData)
  library(car)
  varx <- recode(varx,code)
  varx <- as.factor(varx)
}


#creer graphique Histogramme
makehisto <- function(variable,x_titre,y_titre, main_titre, couleur1){
  hist(variable,col=couleur1,xlab=x_titre,ylab=y_titre,main=main_titre,
       freq=FALSE)
  box()
  densite <- density(variable)
  lines(densite, col = "red",lwd=3)
}


#creer des pie charts
makepie<-function(a,b,c){

  categoriesi = a
  pcti=round((b/sum(b)*100),2)
  lblsi=paste(a ,":",pcti,"%")
  couleurs=c("red","orange","green","grey","purple","white")
```

```
  pie(b,labels = lblsi,main=c, col=couleurs,cex=1.1, cex.main=1.5)
  box()

}

#creer barplot
makebarplot <- function(a,b,c){
  tablecroise <- table(a,b)
  barplot(tablecroise, beside=T,col=c("black","gray"), main = c)
}



#Creer un Tableau significative
makesign <- function(coefficient,EcartType,P_signi){
  taille <- length(coefficient)
  Signi <- rep(NA, taille)
  for (i in 1:taille){
    if (P_signi[i]<0.05){
      Signi[i] <- "Significative"

    }
    else{
      Signi[i] <- "Non significative"
    }
  }
  tableau <- data.frame(coefficient,EcartType,Signi)
  return (tableau)

}



#creer un PNG pour sortie de tableau
sortie_tab <- function(x,y,z,w){
```

```
  library(gridExtra)
  png(x, height = z*nrow(y), width =
        w*ncol(y))
  grid.table(y)
  p1 <- tableGrob(head(y))
  p2 <- grid.arrange(p1)
  return(p2)
}


#Importe les donnees
library(foreign)
donnee<- read.dta("INDIV151.dta", convert.factors=FALSE)


#on garde que les individu avec un salaire et "actif occupe" selon la BIT
BD <- donnee
BD <- BD[BD$salmee!="      0",]
BD <- BD[BD$salmee!="9999999",]
BD <- BD[BD$salmee!="9999998",]
BD <- BD[BD$acteu6=="1",]
table(BD$acteu6)
dim(BD)

#convertir en numerique
BD$salmee<-as.numeric(BD$salmee)

#on selectionne les individu avec un salaire egale ou supérieur que le smic
BD = BD[BD$salmee>1136,]

rm(donnee)



#recodage
```

```
BD$salmee <-as.numeric(BD$salmee)
BD$age <-as.integer(BD$age)
BD$ancentr <-as.integer(BD$ancentr)
BD$nbind <- as.integer(BD$nbind)

#fonction recodage créer
table(BD$sexe)
BD$sexe <- recodage(BD$sexe,"'1'=0;'2'=1")
table(BD$sexe)

BD$immi <- recodage(BD$immi,"'1'=1;'2'=0")
table(BD$immi)

BD$nivp <- recodage(BD$nivp, "'73'=0;'72'=0;'71'=0;'60'=1;
                '50'=1;'41'=1;'40'=2;'30'=2;'20'=3;'10'=4" )
table(BD$nivp)

BD$ccontr <- recodage(BD$ccontr,"''=0;'0'=0;'5'=1;'4'=2;'3'=3;'2'=4;'1'=5")
table(BD$ccontr)

BD$zus <- recodage(BD$zus,"''=0" )

BD$ccontr <- as.factor(BD$ccontr)
BD$tuu2010r <- as.factor(BD$tuu2010r)



#creation de la nouvelle variable nation
#avec un des parents ayant une nationnalité etrangere
BD$nat14 <- recode(BD$nat14,"''=99")



BD$nation[BD$nat14=="10"]="0"
BD$nation[BD$nat14=="11"]="1"
BD$nation[BD$nat14=="12"]="1"
BD$nation[BD$nat14=="13"]="1"
```

```
BD$nation[BD$nat14=="14"]="1"
BD$nation[BD$nat14=="21"]="2"
BD$nation[BD$nat14=="22"]="2"
BD$nation[BD$nat14=="23"]="2"
BD$nation[BD$nat14=="24"]="3"
BD$nation[BD$nat14 == "31"]="4"
BD$nation[BD$nat14 == "32"]="4"
BD$nation[BD$nat14 == "41"]="4"
BD$nation[BD$nat14 == "51"]="4"
BD$nation[BD$nat14 == "99"]="4"

BD$nation[BD$natper == "06" | BD$natmer == "06"]="2"
BD$nation[BD$natper == "07" | BD$natmer == "07"]="3"
BD$nation[BD$natper == "03" | BD$natmer == "03"]="1"
BD$nation[BD$natper == "04" | BD$natmer == "04"]="1"
BD$nation[BD$natper == "05" | BD$natmer == "05"]="1"
BD$nation[BD$natper == "08" | BD$natmer == "08"]="4"
BD$nation[BD$natper == "09" | BD$natmer == "09"]="4"
BD$nation[BD$natper == "10" | BD$natmer == "10"]="4"
BD$nation[BD$natper == "99" | BD$natmer == "99"]="4"

BD$nation <- as.factor(BD$nation)

#creation de la nouvelle variable ORIGINE
BD$ORIGINE[BD$naimer!="1" | BD$naiper!="1" |BD$nat14!="10"]="1"
BD$ORIGINE[BD$naimer=="1" & BD$naiper=="1" &BD$nat14=="10"]="0"



#creation de la variable binaire type public ou prive
BD$pp[BD$pub3fp=="1" ]="0"
BD$pp[BD$pub3fp=="2" ]="0"
BD$pp[BD$pub3fp=="3" ]="0"
BD$pp[BD$pub3fp=="4" ]="1"
```

```
#nouvelle base avec l'exploitation des variables pour le modele
BD1 <- data.frame(BD$salmee,BD$age,BD$ancentr,BD$nbind, BD$sexe, BD$immi,
BD$ORIGINE, BD$pp, BD$zus,
                  BD$cser , BD$ccontr, BD$nivp, BD$tuu2010r,
BD$nation,row.names=row.names(BD))
vect_name <- c("salmee","age","ancentr","nbind","sexe","immi","origine",
               "pp","zus","cser","ccontr","nivp","tuu2010r","nation")
names(BD1) <- vect_name


#nettoyage de la base
BD2<-BD1
BD2<-na.omit(BD2)


#nommer les variable qualitative
BD2$immi1[BD2$immi==0] <- "non immigrant"
BD2$immi1[BD2$immi==1] <- "immigrant"
table(BD2$immi1)

BD2$sexe1[BD2$sexe==0] <- "Homme"
BD2$sexe1[BD2$sexe==1] <- "Femme"
table(BD2$sexe1)

BD2$nivp1[BD2$nivp==""] <- "NIV 1"
BD2$nivp1[BD2$nivp==0] <- "NIV 1"
BD2$nivp1[BD2$nivp==1] <- "NIV 2"
BD2$nivp1[BD2$nivp==2] <- "NIV 3"
BD2$nivp1[BD2$nivp==3] <- "NIV 4"
BD2$nivp1[BD2$nivp==4] <- "NIV 5"

table(BD2$nivp1)

BD2$zus1[BD2$zus == 0 ] <- "ZUS NON"
```

```
BD2$zus1[BD2$zus == 1 ] <- "ZUS OUI"

BD2$ccontr1[BD2$ccontr== 0]<- "Pas de contrat de travail ou pas renseigne"
BD2$ccontr1[BD2$ccontr== 1]<- "Contrat d'apprentissage"
BD2$ccontr1[BD2$ccontr== 2]<- "Contrat d'interim"
BD2$ccontr1[BD2$ccontr== 3]<- "Contrat saisonnier"
BD2$ccontr1[BD2$ccontr== 4]<- "CDD"
BD2$ccontr1[BD2$ccontr== 5]<- "CDI"

BD2$nation1[BD2$nation==0] <- "Francais"
BD2$nation1[BD2$nation==1] <- " europeen"
BD2$nation1[BD2$nation==2] <- " Maghreb"
BD2$nation1[BD2$nation==3] <- " Africaine"
BD2$nation1[BD2$nation==4] <- "Reste du monde"

BD2$origine1[BD2$origine==0] <- "Pas d'origine"
BD2$origine1[BD2$origine==1] <- "origines"

BD2$pp1[BD2$pp==0] <- "Public"
BD2$pp1[BD2$pp==1] <- "Privé"


BD2$cser1[BD2$cser ==0 ] <-"Non renseigne"
BD2$cser1[BD2$cser ==1 ] <-"Agriculteurs exploitants"
BD2$cser1[BD2$cser ==2 ] <-"Artisans, commercants et chefs d'entreprises"
BD2$cser1[BD2$cser ==3 ] <-"Cadres et professions intellectuelles supérieures"
BD2$cser1[BD2$cser ==4 ] <-"Professions intermédiaires"
BD2$cser1[BD2$cser ==5 ] <-"Employés"
BD2$cser1[BD2$cser ==6 ] <-"Ouvriers"


BD2$tuu2010r1[BD2$tuu2010r ==1 ] <-"Communaute rurale"
BD2$tuu2010r1[BD2$tuu2010r ==2 ] <-"Unité urbaine de - de 20k habitants"
BD2$tuu2010r1[BD2$tuu2010r ==3 ] <-"Unité urbaine de 20k à moins de 200k habitants"
BD2$tuu2010r1[BD2$tuu2010r ==4 ] <-"Unité urbaine de 200k habitants ou plus"
```

```
BD2$tuu2010r1[BD2$tuu2010r ==5 ] <-"Agglomeration parisienne"


table(BD2$sexe)

#detection des valeurs aberrantes


library(EnvStats)
str(BD2$salmee)

par(mfrow=c(2,2))
boxplot(BD2$salmee, main = "Salaire mensuel", color= "Blue")
boxplot(BD2$age, main ="Age",color= "Blue")
boxplot(BD2$ancentr, main = "Anciennete dans l'entreprise")
boxplot(BD2$nbind, main = "Nb de pers dans le logement")

rosnerTest(BD2$salmee, k = 20, alpha = 0.05)
order(BD2$salmee)
sort(BD2$salmee)
BD3 = BD2[BD2$salmee<8000,]

#on utilise order pour trier par rapport au salaire dans Rstudio
#Puis on peut comparer par rapport au graphique de boxplot
#quelle valeur doit être inferieur au valeur aberrante

rosnerTest(BD3$salmee, k = 20, alpha = 0.05)
boxplot(BD3$salmee)
BD3 = BD3[BD3$salmee<3157,]
boxplot(BD3$age)

boxplot(BD3$ancentr)
rosnerTest(BD3$ancentr, k = 9, alpha = 0.05)
BD3 = BD3[BD3$ancentr<492,]
BD3 = BD3[BD3$ancentr>-1,]
```

```
boxplot(BD3$nbind)
rosnerTest(BD3$nbind, k = 9, alpha = 0.05)
BD3 = BD3[BD3$nbind<8,]


#save BD3
save(BD, file = "RData")
save(BD1, file = "BD1.RData")
save(BD2, file = "BD2.RData")
save(BD3, file = "BD3.RData")
dim(BD3)

rm(BD1)
rm(BD2)
rm(BD)

##########################
summary(BD3)
boxplot(BD3)

library(psych)
describeBy(BD3)

par(mfrow=c(2,2))
boxplot(BD3$salmee, main = "Salaire mensuel", color= "Blue")
boxplot(BD3$age, main ="Age",color= "Blue")
boxplot(BD3$ancentr, main = "Anciennete dans l'entreprise")
boxplot(BD3$nbind, main = "Nb de pers dans le logement")



########################
#matrice de correlation
library(corrplot)
```

```
m<- cor(BD3[,c(1:4)])
corrplot(m, method="number")


##########################
#Variable numerique continue histogramme
par(mfrow=c(2,2))
makehisto(BD3$salmee,"Salaire mensuel (euros)", "Frequence","Salaire
mensuel","#F5D0A9")
makehisto(BD3$age,"Annee", "Frequence","Âge","green")
makehisto(BD3$ancentr,"Nombre de mois", "Frequence","Anciennete dans
l'entreprise","Gray")
makehisto(BD3$nbind,"nombre de personne", "Frequence","Nombre de personne dans le
logement","Yellow")


############################
#tableau avec la moyenne et ecart-type
library(questionr)

library(psych)
describeBy(BD3)

library(gmodels)
CrossTable(BD3$nation,BD3$cser,prop.chisq=FALSE,chisq=FALSE,expected=FALSE)



#Creation des bases avec origine et aucune origine
BD_NORI = BD3[BD3$origine==0,]
BD_ORIGINE = BD3[BD3$origine!=0,]

#########################################################################
######
#graphique histogramme
```

```
#histogramme BD3 avec densité de base BD_NORI et BD_ORIGINE
makehisto(BD3$salmee,"Salaire mensuel (euros)", "Frequence","Salaire
mensuel","#F5D0A9")
densite <- density(BD_NORI$salmee)
lines(densite, col = "red",lwd=3)
densite1 <- density(BD_ORIGINE$salmee)
lines(densite1, col = "green",lwd=3)
box()
legend("topright", legend = c("Non origine", "Origine"),
       col= c("red","green"),lty=1:1, cex=1.5,
       box.lty=2, box.lwd=2
       )


hist(BD_ORIGINE$salmee,col="#96FD13",
     xlab="Salaire mensuel (euros)",ylab="Fréquences",main="Salaire mensuel des ind
ayant des origines", freq=FALSE)
densite1 <- density(BD_ORIGINE$salmee)
lines(densite1, col = "red",lwd=3)

som<- summary(BD_NORI$salmee)


library(psych)

describeBy(BD_NORI$salmee)
describeBy(BD_ORIGINE$salmee)



hist(BD_ORIGINE$salmee)
hist(BD_NORI$salmee)
```

```
##### graphique pie
#fonction graphique
library("questionr")

par(mfrow=c(2,1))

#nombre d'origine
t<- table(BD3$origine1)
t<- data.frame(t)
makepie(t$Var1,t$Freq,"Individu ayant des origines")
table(BD3$origine)

#nombre d'immigree
t11 <- table(BD3$immi1)
t11<- data.frame(t11)
makepie(t11$Var1,t11$Freq,"Immigré")
table(BD3$immi1)

#Nationalite avec base d'origine
t1 <- table(BD_ORIGINE$nation1)
t1 <- data.frame(t1)
titre1 <- "graphique : Nationalite selon l'origine"
makepie(t1$Var1,t1$Freq,titre1)

#Niveau de diplome
par(mfrow=c(3,1))
t2 <- table(BD3$nivp1)
t2 <- data.frame(t2)
titre2<- "Niveau de diplome"
makepie(t2$Var1,t2$Freq, titre2)

t2_2 <- table(BD_NORI$nivp1)
t2_2 <- data.frame(t2_2)
```

```
titre2_2 <- "Niveau de diplome pour ind sans origine"
makepie(t2_2$Var1,t2_2$Freq,titre2_2)

t2_1 <- table(BD_ORIGINE$nivp1)
t2_1 <- data.frame(t2_1)
titre2_1 <- "Niveau de diplome pour ind ayant des origines"
makepie(t2_1$Var1,t2_1$Freq,titre2_1)

#categorie socio-prof
par(mfrow=c(3,1))
t3 <- table(BD3$cser1)
t3 <- data.frame(t3)
titre3 <- "Categorie socio-professionnelles"
makepie(t3$Var1,t3$Freq,titre3)

t3_2 <- table(BD_NORI$cser1)
t3_2 <- data.frame(t3_2)
titre3_2 <- "Categorie socio-professionnelles pour ind sans origine"
makepie(t3_2$Var1,t3_2$Freq,titre3_2)

t3_1 <- table(BD_ORIGINE$cser1)
t3_1 <- data.frame(t3_1)
titre3_1 <- "Categorie socio-professionnelles pour ind ayant des origines"
makepie(t3_1$Var1,t3_1$Freq,titre3_1)

#contrat
par(mfrow=c(3,1))
t4 <- table(BD3$ccontr1)
t4 <- data.frame(t4)
titre4<-"Contrat de travail"
makepie(t4$Var1,t4$Freq,titre4)

t4_2 <- table(BD_NORI$ccontr1)
t4_2 <- data.frame(t4_2)
titre4_2<- "contrat de travail pour ind sans origine"
```

```
makepie(t4_2$Var1,t4_2$Freq,titre4_2)

t4_1 <- table(BD_ORIGINE$ccontr1)
t4_1 <- data.frame(t4_1)
titre4_1 <- "contrat de travail pour ind avec origine"
makepie(t4_1$Var1,t4_1$Freq,titre4_1)

#Unite urbaine 2010
par(mfrow=c(3,1))
t5 <- table(BD3$tuu2010r1)
t5 <- data.frame(t5)
titre5 <-"Unite urbaine du logement de residence"
makepie(t5$Var1,t5$Freq,titre5)

t5_2 <- table(BD_NORI$tuu2010r1)
t5_2 <- data.frame(t5_2)
titre5_2 <-"Unite urbaine du logement de residence pour ind sans origine"
makepie(t5_2$Var1,t5_2$Freq,titre5_2)

t5_1 <- table(BD_ORIGINE$tuu2010r1)
t5_1 <- data.frame(t5_1)
titre5_1 <-"Unite urbaine du logement de residence pour ind avec origine"
makepie(t5_1$Var1,t5_1$Freq,titre5_1)



### barplot

#graphique barplot
library(descr)
tit1<- "contrat de travail selon le sexe avec des ind ayant des origine"
makebarplot(BD_ORIGINE$sexe1,BD_ORIGINE$ccontr1,tit1)
legend("topright", legend = c("Femme", "Homme"),
       fill = c("black","gray"),
       cex = 1.7)
```

```
tit1_1<- "contrat de travail selon le sexe avec des ind sans origine"
makebarplot(BD_NORI$sexe1,BD_NORI$ccontr1,tit1_1)
legend("topright", legend = c("Femme", "Homme"),
       fill = c("black","gray"),
       cex = 1.7)

tit2 <- "Type de travail"
makebarplot(BD3$sexe1,BD3$pp1,tit2)
legend("topright", legend = c("Femme", "Homme"),
       fill = c("black","gray"),
       cex = 1.7)



#type de travail
tit2_1 <- "Type de travail pour ind ayant des origines"
makebarplot(BD_ORIGINE$sexe1,BD_ORIGINE$pp1,tit2_1)
legend("topright", legend = c("Femme", "Homme"),
       fill = c("black","gray"),
       cex = 1.7)

tit2_2 <- "Type de travail pour ind sans origines"
makebarplot(BD_NORI$sexe1,BD_NORI$pp1,tit2_2)
legend("topright", legend = c("Femme", "Homme"),
       fill = c("black","gray"),
       cex = 1.7)

rm(BD1)
rm(BD)
rm(BD2)
rm(donnee)

####
#regression lineaire multiple
# categoriser les variable X continue pour l'anciennete en entreprise
```

```
BD3$age1[BD3$age<35]<- "moins de 35 ans"
BD3$age1[BD3$age>=35 & BD3$age<45]<- "35 et 45 ans"
BD3$age1[BD3$age>=45 & BD3$age<55]<- "45 et 55 ans"
BD3$age1[BD3$age>=55]<- "plus de 55"
table(BD3$age1)

BD3$ancentr1<-round(BD3$ancentr/12,0)

BD3$ancentr2[BD3$ancentr1<5]<- "- de 5 ans"
BD3$ancentr2[BD3$ancentr1>=5 & BD3$ancentr1<15]<- "5 et 14 ans"
BD3$ancentr2[BD3$ancentr1>=15 & BD3$ancentr1<25]<- "15 et 24 ans"
BD3$ancentr2[BD3$ancentr1>=25 ]<- "+ 25 ans"
table(BD3$ancentr2)



#### RLM
library(MASS)
library(gridExtra)
str(BD3)

BD3$origine1[BD3$origine==0] <- "Pas d'origine"
BD3$origine1[BD3$origine==1] <- "origines"
table(BD3$origine1)

BD_ORIGINE <- BD3[BD3$origine1=="origines",]
BD_NORI <- BD3[BD3$origine1=="Pas d'origine",]

BD_IMMI <- BD3[BD3$immi1=="immigrant",]
BD_NOIMMI <- BD3[BD3$immi1=="non immigrant",]

table(BD3$immi1)
str(BD3$sexe1)

# changer l'ordre des variable quali multi fact
```

```
facteur <- function(x,y){
x <- factor(x ,levels = y)
return(x)
}

## permet d'avoir le sens des facteurs pour le referentiel
BD3$nivp1<-facteur(BD3$nivp1,c("NIV 2","NIV 5","NIV 4","NIV 3","NIV 1"))
BD3$origine1 <- facteur(BD3$origine1,c("Pas d'origine","origines"))
BD3$nation1 <- facteur(BD3$nation1 ,c("Francais"," europeen"," Maghreb","
Africaine","Reste du monde"))
BD3$immi1 <- facteur(BD3$immi1,c("non immigrant","immigrant"))
BD3$cser1 <- facteur(BD3$cser1, c("Employés","Cadres et professions intellectuelles
supérieures",
                          "Ouvriers","Professions intermédiaires","Non renseigne"))
BD3$sexe1 <- facteur(BD3$sexe1,c("Homme","Femme"))
BD3$tuu2010r1 <- facteur(BD3$tuu2010r1,c("Unité urbaine de 200k habitants ou plus"
                              ,"Agglomeration parisienne",
                              "Unité urbaine de 20k à moins de 200k habitants"
                              ,"Unité urbaine de - de 20k habitants",
                              "Communaute rurale"))
table(BD3$ancentr2)
BD3$age1 <- facteur(BD3$age1,c("35 et 45 ans","moins de 35 ans","45 et 55 ans","plus
de 55"))


#####################################
rego1 <- lm(log(salmee) ~  age+ ancentr+ nbind+ sexe1
          + nivp1+nation1+ zus1+ immi1+ origine1+ cser1+ tuu2010r1+pp1,data= BD3)
sommaire<- summary(rego1)

tabrego1 <- round(coef(rego1),6)
Ecart_type1 <- round(sommaire$coefficients[,2],6)
P_value1<- round(sommaire$coefficients[,4],3)
tableau_reg <- makesign(tabrego1,Ecart_type1,P_value1)
```

```
#sortie tableau
#Creer un Tableau significative
sortie_tab("modele1.png", tableau_reg, 50,310)




table(BD3$origine1)
hist(BD_NORI$salmee, freq=F)
hist(BD_ORIGINE$salmee, freq=F)


#Modele 2 en retirant origine
str(BD3)
rego2 <- lm(log(salmee) ~  age+ ancentr+ nbind+ sexe1
          + nivp1+nation1+ zus1+ immi1+ cser1+ tuu2010r1+pp1,data= BD3)
sommaire2<- summary(rego2)


tabrego2 <- round(coef(rego2),6)
Ecart_type2 <- round(sommaire2$coefficients[,2],6)
P_value2<- round(sommaire2$coefficients[,4],3)
tableau_reg2 <- makesign(tabrego2,Ecart_type2,P_value2)
sortie_tab("modele2.png", tableau_reg2, 50,310)


table(BD3$cser1)

### analyse du modele
#Existence de valeurs influençant les estimations via le graphique de la distance de Cook
plot(cooks.distance(rego2),type="h")
# pas d'observation influençant l'estimation du modele


#Forme linéaire
```

```
library(zoo)
library(lmtest)
reset(rego2)
#On accepte pas la forme linéaire car p_value < 0.05

#test de fisher , il existe au moins une des variables qui est significative



#test des residus
residu <- residuals(rego2)
hist(residu)
#Test de Kolmogorov pour les grands nombres.
ks.test(residu,"pnorm",mean(residu),sd(residu))
#On valide l'hypothese que les residu suit une loi normale

#H d'HMS des residus
#on refuse l'Homoscedasticite
bptest(rego2)

#La probabilité critique du test étant inférieure
# à 0,05, l'hypothèse d'homoscédascité des résidus
#du modèle n'est pas acceptée au seuil de risque de 5 %.

BD3$ancentr2

#identifier la variable explicative
ncvTest(rego2,~BD3$ancentr2) #p_value= 0.14
ncvTest(rego2,~BD3$sexe1) #P_Value= 0.054
ncvTest(rego2,~BD3$nivp1) #P_Value= 0.066
ncvTest(rego2,~BD3$zus1) #P_Value= 0.26
ncvTest(rego2,~BD3$immi1) #P_Value= 5.1e^-6 , refuse H0
ncvTest(rego2,~BD3$cser1) #P_Value= 2.5e^-5 , refuse H0
ncvTest(rego2,~BD3$tuu2010r1) #P_Value= 0.00016 , refuse H0
ncvTest(rego2,~BD3$origine1) #P_Value= 0.00051 , refuse H0
```

```
#source d'Heterosc ne fonctionne pas car variabl qualiti a plusieurs facteur
#revoir
library(car)
residualPlots(rego2)
## NBIND pas significative


str(BD3)

#correction de la matrice de white
library(sandwich)
library(lmtest)
coeftest(rego2,vcov=vcovHC(rego2,type="HC0"))
waldtest(rego2, vcov = vcovHC(rego2,type="HC0"))

vif(rego2)
#comme Y quantitative et Xqualitative a plusieurs facteur : analyse de la covariance avec
# ANCOVA
plot(rego2)
anova(rego2)
summary.aov(rego2)



#modele 3 avec base immi
rego3 <- lm(log( salmee) ~  age+ ancentr+
              nbind+ sexe1
            + nivp1+ zus1+pp1+ccontr1
            + cser1+ tuu2010r1, data=BD_IMMI)
sommaire3<-summary(rego3)
tabrego3 <- round(coef(rego3),6)
```

```
Ecart_type3 <- round(sommaire3$coefficients[,2],6)
P_value3<- round(sommaire3$coefficients[,4],3)


tableau_reg3 <- makesign(tabrego3,Ecart_type3,P_value3)
tit_modele3="modele3.png"
sortie_tab(tit_modele3, tableau_reg3, 50,310)

summary(BD_ORIGINE$salmee)
summary(BD_NORI$salmee)
summary(BD3$salmee)
str(BD3)

#modele 4 avec base sans immi
rego4 <- lm(log( salmee) ~  age+ ancentr+
              nbind+ sexe1
            + nivp1+ zus1+pp1+ccontr1
            + cser1+ tuu2010r1,data= BD_NOIMMI)
summary(rego4)
sommaire4<-summary(rego4)
tabrego4 <- round(coef(rego4),6)
dim(tabrego4)
Ecart_type4 <- round(sommaire4$coefficients[,2],6)
P_value4<- round(sommaire4$coefficients[,4],3)
signi_4 <- rep(NA, 22)


summary(exp(predict(rego3)))
summary(exp(predict(rego4)))
tableau_reg4 <- makesign(tabrego4,Ecart_type4,P_value4)
tit_modele4="modele4.png"
sortie_tab(tit_modele4, tableau_reg4, 50,310)


#illustrer les coefficient des deux groupes selon immi
```

```
test1 <- data.frame(round(coef(rego3),4),round(coef(rego4),4))
sortie_tab("uni.png",test1,70,290)



#Magrebh

table(BD3$nation1)
BD_ARABE <- BD3[BD3$nation==2,]
dim(BD_ARABE)

rego5 <- lm(log( salmee) ~  age+ ancentr+
              nbind+ sexe1
            + nivp1+ zus1+pp1+ccontr1
            + cser1+ tuu2010r1+pp1,data= BD_ARABE)
summary(rego5)

sommaire5<-summary(rego5)
tabrego5 <- round(coef(rego5),6)

Ecart_type5 <- round(sommaire5$coefficients[,2],6)
P_value5<- round(sommaire5$coefficients[,4],3)



tableau_reg5 <- makesign(tabrego5,Ecart_type5,P_value5)
tit_modele5="modele5.png"
sortie_tab(tit_modele5, tableau_reg5, 50,310)



#modele de décomposition des salaires
str(BD4)

BD4[,c(28:47)]=lapply(BD4[,c(28:47)],as.factor)
table(BD4$origine)
```

```
rego5 <- lm(log(salmee)~
          sexe+NIV1+NIV2+NIV3+NIV4+ zus+pp+CDD+
            CDI+CA+CI+CS+EMPL+OUV
          +K200KH+APARIS+K20K200K+UM20K,
          data = BD4[BD4$origine==1,])

summary(rego5)
rego6 <- lm(log(salmee)~
            sexe+NIV1+NIV2+NIV3+NIV4+ zus+pp+CDD+
            CDI+CA+CI+CS+EMPL+OUV
          +K200KH+APARIS+K20K200K+UM20K,
          data = BD4[BD4$origine==0,])
summary(rego6)



coeffs.A <- rego5$coefficients
coeffs.B <- rego6$coefficients
tablecoef <- round(cbind(coeffs.A,coeffs.B),3)
sortie_tab("tablecoef.png", tablecoef, 50,310)
mean(predict(rego5))
mean(predict(rego6))
X.A <- model.matrix(~  sexe+NIV1+NIV2+NIV3+NIV4+ zus+pp+CDD+
                      CDI+CA+CI+CS+EMPL+OUV
                    +K200KH+APARIS+K20K200K+UM20K,
                    data = BD4[BD4$origine==1,])
#on applique la fonction moyenne pour chaque variable
X.moy.A<-apply(X.A,2,mean)

X.B <- model.matrix(~  sexe+NIV1+NIV2+NIV3+NIV4+ zus+pp+CDD+
                      CDI+CA+CI+CS+EMPL+OUV
                    +K200KH+APARIS+K20K200K+UM20K,
                    data = BD4[BD4$origine==0,])
X.moy.B<-apply(X.B,2,mean)
round(cbind(X.moy.A,X.moy.B),3)
```

```
sum((X.moy.B- X.moy.A)*coeffs.B)
sum(X.moy.A*(coeffs.B-coeffs.A))
somme <- sum((X.moy.B- X.moy.A)*coeffs.B)+sum(X.moy.A*(coeffs.B-coeffs.A))

mean(predict(rego6))-mean(predict(rego5))



plot(predict(rego5)~BD_ORIGINE$salmee)
abline(lm(predict(rego5)~BD_ORIGINE$salmee),col="red")
abline(lm(predict(rego6)~BD_NORI$salmee),col="green")

#methode de OAXACA
#Hlavac, Marek (2018). oaxaca: Blinder-Oaxaca Decomposition in R.
#R package version 0.1.4. https://CRAN.R-project.org/package=oaxaca

BD4<-BD3
table(BD4$nivp1)
table(BD4$nivp)
binary <- function(facteur, nom_va, var,base_salaire){
  base_salaire$nom_va[base_salaire$var==facteur]<- 1
  base_salaire$nom_va[base_salaire$var!=facteur]<- 0
  table(base_salaire$nom_va)

}



BD4$NIV2[BD4$nivp1=="NIV 2"] <- 1
BD4$NIV2[BD4$nivp1!="NIV 2"] <- 0
table(BD4$NIV2)

BD4$NIV3[BD4$nivp==2] <- 1
BD4$NIV3[BD4$nivp!=2] <- 0
table(BD4$NIV3)
```

```
BD4$NIV4[BD4$nivp==3] <- 1
BD4$NIV4[BD4$nivp!=3] <- 0
table(BD4$NIV4)

BD4$NIV5[BD4$nivp==4] <- 1
BD4$NIV5[BD4$nivp!=4] <- 0
table(BD4$NIV5)

BD4$NIV1[BD4$nivp1 == "NIV 1"] <- 1
BD4$NIV1[BD4$nivp1 != "NIV 1"] <- 0
table(BD4$NIV1)

table(BD3$ccontr)
table(BD3$ccontr1)
BD4$CDD[BD4$ccontr==4 ] <- 1
BD4$CDD[BD4$ccontr!=4] <- 0
table(BD4$CDD)

BD4$CDI[BD4$ccontr==5] <- 1
BD4$CDI[BD4$ccontr!=5] <- 0
table(BD4$CDI)

BD4$CA[BD4$ccontr==1] <- 1
BD4$CA[BD4$ccontr!=1] <- 0
table(BD4$CA)

BD4$CI[BD4$ccontr==2] <- 1
BD4$CI[BD4$ccontr!=2] <- 0
table(BD4$CI)

BD4$CS[BD4$ccontr==3] <- 1
BD4$CS[BD4$ccontr!=3] <- 0
table(BD4$CS)

BD4$PCT[BD4$ccontr==0] <- 1
```

```
BD4$PCT[BD4$ccontr!=0] <- 0
table(BD4$PCT)

table(BD4$cser1)
table(BD4$cser)
BD4$EMPL[BD4$cser==5] <- 1
BD4$EMPL[BD4$cser!=5] <- 0
table(BD4$EMPL)

BD4$OUV[BD4$cser==6] <- 1
BD4$OUV[BD4$cser!=6] <- 0
table(BD4$OUV)


BD4$CPI[BD4$cser==3] <- 1
BD4$CPI[BD4$cser!=3] <- 0
table(BD4$CPI)

BD4$CPI[BD4$cser==3] <- 1
BD4$CPI[BD4$cser!=3] <- 0
table(BD4$CPI)

BD4$PI[BD4$cser==4] <- 1
BD4$PI[BD4$cser!=4] <- 0
table(BD4$PI)

table(BD4$tuu2010r1)
table(BD4$tuu2010r)
BD4$K200KH[BD4$tuu2010r1=="Unité urbaine de 200k habitants ou plus"] <- 1
BD4$K200KH[BD4$tuu2010r1!="Unité urbaine de 200k habitants ou plus"] <- 0
table(BD4$K200KH)

BD4$APARIS[BD4$tuu2010r1=="Agglomeration parisienne"] <- 1
BD4$APARIS[BD4$tuu2010r1!="Agglomeration parisienne"] <- 0
table(BD4$APARIS)
```

```
BD4$K20K200K[BD4$tuu2010r1=="Unité urbaine de 20k à moins de 200k habitants"] <-
1
BD4$K20K200K[BD4$tuu2010r1!="Unité urbaine de 20k à moins de 200k habitants"] <- 0
table(BD4$K20K200K)

BD4$UM20K[BD4$tuu2010r1=="Unité urbaine de - de 20k habitants"] <- 1
BD4$UM20K[BD4$tuu2010r1!="Unité urbaine de - de 20k habitants"] <- 0
table(BD4$UM20K)

BD4$COMRU[BD4$tuu2010r1=="Communaute rurale"] <- 1
BD4$COMRU[BD4$tuu2010r1!="Communaute rurale"] <- 0
table(BD4$COMRU)



library("oaxaca")
table(BD4$origine)
results <- oaxaca(formula=log(salmee)~  sexe+NIV1+NIV2+NIV3+NIV4+ zus+pp+CDD+
                    CDI+CA+CI+CS+EMPL+OUV
                  +K200KH+APARIS+K20K200K+UM20K|origine,
                    data = BD4, R=100)
summary(results)

round(results$twofold$overall[,1:5], 3)
plot(results, decomposition = "twofold", group.weight = 1)
round(results$twofold$variables[[2]][,2:5] ,3)

sortie_tab("oxaca1.png", round(results$twofold$overall[,1:5], 3), 50,310)
sortie_tab("oxaca.png", round(results$twofold$variables[[2]][,2:5] ,6), 50,310)



# selon une nationalité pour les personnnes originaires de magrehb
BD4$magh[BD4$nation==2] <- 1
BD4$magh[BD4$nation!=2] <- 0
table(BD4$nation)
```

```
logit<-glm(magh ~ sexe+NIV1+NIV2+NIV3+NIV4+ zus+pp+CDD+
          CDI+CA+CI+CS+EMPL+OUV
         +K200KH+APARIS+K20K200K+UM20K, family=binomial (link='logit'),
         data=BD4)
summary(logit)$coefficients




p<-predict(logit,type='response')
w1<-ifelse(BD4$magh==0,
         p/(1-p)*(1-mean(BD4$magh))/mean(BD4$magh), 1)



library(Hmisc)
grid<-seq(0.1,0.9,0.1)
ref<-BD4$magh==0
BD4$logsal <- log(BD4$salmee)
dfl.Fc<-wtd.quantile(BD4$logsal[ref], weights=w1[ref], probs=grid)


ref1<-BD4$magh==1
dfl.Fmagh<-wtd.quantile(BD4$logsal[ref1], weights=w1[ref1], probs=grid)


test <- rbind(dfl.Fc,dfl.Fmagh,dfl.FAFRI)

library(Hmisc)
#On calcule les d´eciles 1 `a 9 de la distribution contrefactuelle #(la distribution dans le
groupe B repond´er´ee)
dfl.Fc<-wtd.quantile(BD4$logsal[BD4$magh==0],
```

```
weights=w1[BD4$magh==0], probs=seq(0.1,0.9,0.1))
#A comparer aux d´eciles de la distribution des salaires des B
dfl.FB<-wtd.quantile(BD4$logsal[BD4$magh==0],
                     probs=seq(0.1,0.9,0.1))
#Et `a ceux de la distribution des salaires des A
dfl.FA<-wtd.quantile(BD4$logsal[BD4$magh==1],
                     probs=seq(0.1,0.9,0.1))

plot(dfl.Fc, cex = 1, pch = 1, col = "red", main ="salaire selon l'origine")
lines(dfl.FB, col = "red")
lines(dfl.FA, col = "orange")
lines(dfl.Fc, col = "blue")
legend("topleft", legend = c("Pop non typé", "Magrhebe","Non typé repondé"), fill =
c("red", "orange","blue"))



#ecart total
round(dfl.FB-dfl.FA,3)

#dont effet de composition
round(dfl.FB-dfl.Fc,3)



# ecart inexplique
round(dfl.Fc-dfl.FA,3)

#### Répartition des migrants dans la France

BD= BD[BD$reg!="01",]
BD= BD[BD$reg!="02",]
BD= BD[BD$reg!="03",]
BD= BD[BD$reg!="04",]
table(BD$reg)
table(FranceFormes$NAME_1)
```

```
BD$departement[BD$regio=="94"]="Corse"
BD$departement[BD$regio=="52"]="Pays de la Loire"
BD$departement[BD$regio=="11"]="Île-de-France"
BD$departement[BD$regio=="53"]="Bretagne"
BD$departement[BD$regio=="28"]="Normandie"
BD$departement[BD$regio=="93"]="Provence-Alpes-Côte d'Azur"
BD$departement[BD$regio=="75"]="Nouvelle-Aquitaine"
BD$departement[BD$regio=="24"]="Centre-Val de Loire"
BD$departement[BD$regio=="27"]="Bourgogne-Franche-Comté"
BD$departement[BD$regio=="76"]="Occitanie"
BD$departement[BD$regio=="32"]="Hauts-de-France"
BD$departement[BD$regio=="44"]="Grand Est"
BD$departement[BD$regio=="84"]="Auvergne-Rhône-Alpes"
table(BD$departement)

table(BD3$immi1,BD3$origine1)
rm(BD)
#### carte de la france
str(FranceFormes)
#Importation du package
library(raster)
table(BD$ORIGINE)
sum(table(BD$departement))
table <- table(BD$departement,BD$ORIGINE)
addmargins(table)
table<-data.frame(addmargins(table))
BD_TABLE <- data.frame(table$Var1,table$Freq)
BD_TABLE1<- BD_TABLE[-c(14:42),]
BD_TABLE2<- BD_TABLE[-c(1:14,28:42),-1]
BD_TABLE3 <- BD_TABLE[-c(1:28,42),-1]
BD_TABLET<- data.frame(BD_TABLE1,BD_TABLE2)
length(BD_TABLE3)
BD_TABLET$table.Freq <-100*BD_TABLET$table.Freq/9187
str(BD_TABLET)
```

```
BD_TABLET$table.Var1 <- as.character(BD_TABLET$table.Var1)

BD_TABLET



table(BD_TABLET$table.Var1)

#Découpage des régions avant 2015
FranceFormes <- getData(name="GADM", country="FRA", level=1)
plot(FranceFormes, main="Carte de la France, régions (avant 2015)")
idx <- match(FranceFormes$NAME_1, BD_TABLE1$table.Var1)
concordance <- BD[idx, "table.Freq"]
FranceFormes$immi1 <- BD_TABLE1$table.Freq/(sum(BD_TABLE1$table.Freq))*100



concordance <- BD[idx, "BD_TABLE2"]
FranceFormes$immi2 <- concordance

#établissemment de la charte des coupeurs puis tracage de la carte en utilisant
couleurs <- colorRampPalette(c('white', 'red'))
spplot(FranceFormes,"immi1",col.regions=couleurs(30),
      main=list(label="Répartition des migrants en % dans la FRANCE",cex=.8))



table(BD4$origine1,BD4$tuu2010r1)
```