

Méthodologie statistique

M 2019/01

Les méthodes de décomposition appliquées à l'analyse des inégalités

Béatrice Boutchenik, Elise Coudin, Sophie Maillard
(DMCSI)

Document de travail



Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2019/01

Les méthodes de décomposition appliquées à l'analyse des inégalités

Béatrice Boutchenik, Elise Coudin, Sophie Maillard (DMCSI)

Les auteurs remercient pour leur aide et leurs suggestions Pauline Givord, Sylvie Lagarde, Stéphane Legleye, Sophie Ponthieux, Pierre Pora, Mathilde Poulhes, Patrick Sillard et l'ensemble des participants du groupe de lecture interne à l'Insee sur les méthodes de décomposition.

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L001
88 AVENUE VERDIER - CS 70058 - 92541 MONTROUGE CEDEX - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : <http://www.insee.fr>

Ces documents de travail ne reflètent pas la position de l'Insee et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only their author's views.

Les méthodes de décomposition appliquées à l'analyse des inégalités

Béatrice Boutchenik*

Élise Coudin†

Sophie Maillard‡

Résumé

Les méthodes de décomposition sont des outils standards pour l'analyse statistique des discriminations, notamment salariales. Le modèle canonique utilisé est celui d'Oaxaca et de Blinder (Oaxaca 1973, Blinder 1973), qui propose une décomposition des écarts constatés entre deux populations (hommes et femmes par exemple) en une part expliquée par les caractéristiques observables de ces deux groupes, c'est-à-dire un effet de composition, et en une part inexplicable. Isoler cet écart inexplicable permet de mettre en avant d'éventuels phénomènes discriminatoires, sous certaines hypothèses que nous nous attachons à clarifier. Les conclusions apportées peuvent également être sensibles aux choix de spécification, dont nous cherchons à expliciter les conséquences. Plusieurs méthodes ont par ailleurs été proposées pour étendre le cadre classique d'Oaxaca et de Blinder à une analyse plus complète des écarts, en particulier pour des variables catégorielles et pour l'ensemble de la distribution de variables continues. Nous portons un intérêt particulier à cette extension aux distributions. Enfin, nous présentons différentes approches pour prendre en compte le rôle de la sélection dans les décompositions. La mise en oeuvre de ces différentes méthodes est illustrée à partir de données françaises issues de l'Enquête Emploi en continu, pour l'exemple des disparités de salaire entre hommes et femmes et entre descendants d'immigrés et personnes sans ascendance migratoire.

Mots-clés : décomposition, économétrie, analyse des inégalités, discriminations, effets hétérogènes

Classification JEL : B41, C18, J31

*Insee, Université Paris Dauphine.

†Insee, CREST.

‡Insee. Auteur correspondant : 88, avenue Verdier, CS 70058 92541 Montrouge Cedex, (+33) 1 87 69 55 75, sophie.maillard@insee.fr.

Table des matières

1 La décomposition de l'écart de moyennes entre deux groupes	5
1.1 Le modèle classique d'Oaxaca-Blinder	5
1.2 Ecart expliqué et inexpliqué de salaire : un exemple avec l'Enquête emploi	6
1.3 Références de la décomposition	11
1.4 La décomposition détaillée de l'effet de composition	13
2 La validité de l'interprétation	19
2.1 Effet causal d'appartenance à un groupe et discrimination	19
2.2 Le choix de la référence	22
2.3 La validité de la décomposition détaillée	24
2.3.1 Une hypothèse plus forte pour l'identification de la décomposition détaillée	24
2.3.2 Le problème de la modalité omise dans la décomposition détaillée de l'écart inexpliqué	25
3 Variable d'intérêt dichotomique et écart entre proportions	31
3.1 La décomposition d'Oaxaca-Blinder pour une variable dichotomique	31
3.2 Décomposition de Fairlie	31
4 La décomposition des inégalités au-delà de la moyenne	40
4.1 La méthode de repondération	44
4.2 Les décompositions par régression des quantiles non-conditionnels	48
4.3 Les méthodes d'estimation de la distribution conditionnelle - Régressions sur fonction de répartition	52
5 Le traitement de la sélection dans la décomposition des inégalités	57
5.1 Sélection exogène, sélection endogène	58
5.1.1 Décrire une sélection exogène	59
5.1.2 Traiter la sélection sur inobservables avec une fonction de contrôle	62
5.2 Traiter la sélection avec des données de panel	67
5.3 Traiter la sélection sans la modéliser	67

Liste des applications numériques

1	La décomposition agrégée	6
2	Détailler l'écart expliqué avec le package <code>Oaxaca</code>	14
3	Gérer le problème de modalité omise avec la fonction <code>oaxaca</code>	28
4	Décomposition de Fairlie	34
5	Décomposition par repondération	45
6	Décomposition par les RIF	50
7	Décomposition par les régressions sur fonction de répartition	54
8	Correction de la sélection sur observables par repondération	60
9	Correction de la sélection à l'aide d'une fonction de contrôle	64

Introduction

On appelle méthodes de décomposition l'ensemble des techniques visant à séparer une différence - par exemple de revenus - entre deux groupes en une part liée à des caractéristiques observées individuelles différentes, la *part expliquée* ou *effet de composition*, et une part résiduelle à caractéristiques observables égales, la *part inexpliquée*. Les exemples canoniques de décomposition étudient les écarts de [salaires moyens entre hommes et femmes](#) (Oaxaca, 1973) et entre individus “blancs” et “noirs” aux États-Unis (Blinder, 1973) : une partie de ces écarts pourrait par exemple être attribuée à des différences dans les niveaux d'éducation, ou d'expérience sur le marché du travail. Les méthodes de décomposition se sont imposées comme des outils essentiels dans l'étude des inégalités, du fait notamment de leur facilité de mise en œuvre et de la richesse des interprétations qu'elles permettent. Celles-ci permettent de mesurer l'amplitude des effets de composition dans l'inégalité, d'analyser de façon détaillée les contributions de plusieurs facteurs à l'écart total de revenus, et de porter ainsi un diagnostic fin sur les mécanismes de formation des inégalités.

De nombreuses applications sortant du cadre de l'étude des discriminations se prêtent d'ailleurs à la mise en œuvre de ces méthodes, selon le type de groupes que l'on souhaite comparer. On pourra par exemple décomposer l'écart entre le revenu moyen dans un département donné et le revenu moyen au niveau national, afin de comprendre les situations inégales de différents territoires (Bertran, 2017 pour un exemple sur les revenus d'activité des non-salariés). Dans le cadre d'une comparaison internationale, on sera amené à analyser l'écart pour une même mesure entre différents pays, contrastés deux à deux. Les “groupes” étudiés peuvent également être deux périodes distinctes, en lesquelles on considère une même grandeur, cherchant à comprendre les déterminants de son évolution : Audenaert et al. (2014) décomposent par exemple la croissance du salaire moyen en France dans les années 2000, isolant ce qui relève des évolutions de composition de la population salariée. Les méthodes de décomposition sont souvent liées à l'analyse des discriminations. Pour autant, le cadre dans lequel ce type d'interprétation est possible nécessite des hypothèses beaucoup plus fortes que celles requises lorsqu'on souhaite seulement isoler un effet de composition.

Les méthodes de décomposition ont connu ces dernières années un fort regain d'intérêt, notamment dans un contexte de hausse rapide des inégalités de salaire aux États-Unis (Firpo et al., 2007). Premièrement, une importante réflexion a eu lieu sur les conditions de l'identification d'une discrimination, notamment par l'analogie avec la notion d'effet de traitement, empruntée aux travaux d'évaluation de politiques publiques. Mesurer une discrimination entre deux groupes en contrôlant de leurs caractéristiques peut s'apparenter à estimer l'effet d'un traitement en comparant un groupe traité et un groupe de contrôle. La littérature sur les

décompositions s'est ainsi demandé sous quelles conditions il était possible d'interpréter la part inexpliquée de la décomposition comme l'effet pur de l'appartenance au groupe considéré, et *in fine* comme une mesure de discrimination. Deuxièmement, la réflexion sur les méthodes de décomposition a permis d'élargir la palette des outils hors du cadre initialement proposé par Blinder et Oaxaca, qui s'appuyait sur la régression linéaire et permettait la décomposition de l'écart entre moyennes par groupes pour une variable continue. Des méthodes ont ainsi été développées permettant de s'intéresser à une variable dichotomique, d'une part ; et d'analyser l'ensemble de la distribution des variables continues, d'autre part.

[Fortin, Lemieux, and Firpo \(2011\)](#) ont détaillé dans un article de référence ces nouvelles méthodes, et plus généralement le cadre théorique entourant les méthodes de décomposition¹. Nous nous appuyons ici sur ce document pour en proposer une traduction pratique, tout en insistant sur les questions qu'il est nécessaire de se poser afin d'interpréter correctement les résultats issus des méthodes de décomposition. Le cadre théorique correspondant, ainsi que certains approfondissements, sont renvoyés en encadré. Des codes R sont proposés pour les méthodes les plus facilement implémentables.

Lorsque la variable d'intérêt est continue (salaires, revenus, patrimoine, heures travaillées, notes à un examen...), on s'intéresse souvent à sa moyenne par groupe. La décomposition de l'écart entre moyennes, qui correspond au cadre classique d'Oaxaca-Blinder, est abordée dans la section 1, la section 2 interrogeant la validité de l'interprétation d'une telle décomposition. La variable d'intérêt peut également être une variable dichotomique : le fait d'être au chômage, d'être actif ou encore d'avoir un emploi stable. On cherchera par exemple à comprendre l'écart entre les taux de chômage mesurés dans un groupe et dans l'autre. Ces cas sont traités dans la section 3. Lorsque la variable est continue, on considère parfois d'autres statistiques que sa moyenne, et notamment les écarts existant en différents points de la distribution : en lien avec l'existence d'un éventuel plafond de verre, d'où provient l'écart entre le salaire au-delà duquel se situent les 10 % des hommes les mieux rémunérés, et la statistique équivalente chez les femmes ? Enfin, dans le cas de comparaisons inter-temporelles ou internationales, on peut étendre les méthodes de décomposition à des indicateurs d'inégalité. On sera par exemple amené à se demander : comment la modification des inégalités salariales entre deux périodes peut-elle être expliquée par l'évolution de la composition de la population ? Le lecteur pourra se référer à la section 4 pour les différentes méthodes de décomposition des écarts entre distributions. Enfin, la section 5 traite de la question des problèmes de sélection et de la façon de les aborder dans le cadre d'une décomposition.

1. Pour que le lecteur puisse s'y reporter plus facilement, nous conservons des notations proches de celles utilisées par [Fortin et al. \(2011\)](#).

1 La décomposition de l'écart de moyennes entre deux groupes

Dans cette partie, nous présentons la méthode la plus classique de décomposition des inégalités à la moyenne, dans le cas où la variable d'intérêt est continue : la décomposition dite d'*Oaxaca-Blinder*.

1.1 Le modèle classique d'Oaxaca-Blinder

On considère ici une variable continue Y , dont on observe un ensemble de K déterminants individuels X_1, X_2, \dots, X_K . On souhaite étudier l'écart entre les moyennes de Y pour deux groupes A et B , en lien avec le fait que ces deux groupes présentent des caractéristiques observables différentes. Par exemple, la variable Y pourrait correspondre au salaire (en log), les variables X au niveau d'éducation, à l'expérience sur le marché du travail, etc., et les groupes A et B aux hommes et aux femmes. On modélise séparément, dans le groupe A et le groupe B , une relation linéaire entre la variable Y et ses déterminants :

$$Y_i = \beta_{A0} + \sum_{k=1}^K X_{ik}\beta_{Ak} + v_{iA}, \quad \forall i \in A$$

$$Y_i = \beta_{B0} + \sum_{k=1}^K X_{ik}\beta_{Bk} + v_{iB}, \quad \forall i \in B$$

Une fois les paramètres de chacun des deux modèles estimés, on peut alors écrire, en notant \bar{Y}_B et \bar{Y}_A le salaire moyen dans chaque groupe :

$$\bar{Y}_A = \hat{\beta}_{A0} + \sum_{k=1}^K \bar{X}_{Ak}\hat{\beta}_{Ak}$$

$$\bar{Y}_B = \hat{\beta}_{B0} + \sum_{k=1}^K \bar{X}_{Bk}\hat{\beta}_{Bk}$$

Le salaire moyen peut différer d'un groupe à l'autre pour deux raisons : d'une part, parce que les *caractéristiques* moyennes ne sont pas les mêmes dans le groupe A et le groupe B ; d'autre part, parce que les *valorisations* de ces caractéristiques (les $(\hat{\beta}_{g,k})_{k=1 \dots K}, g = A, B$), ainsi que les constantes des deux modèles, sont différentes. Une façon de décomposer l'écart entre \bar{Y}_B et \bar{Y}_A s'écrit :

$$\bar{Y}_B - \bar{Y}_A = \hat{\beta}_{B0} + \sum_{k=1}^K \bar{X}_{Bk}\hat{\beta}_{Bk} - \hat{\beta}_{A0} - \sum_{k=1}^K \bar{X}_{Ak}\hat{\beta}_{Ak}$$

$$= \underbrace{\sum_{k=1}^K (\overline{X_{Bk}} - \overline{X_{Ak}}) \hat{\beta}_{Bk}}_{\hat{\Delta}_X \text{ (expliqué)}} + \underbrace{\left(\hat{\beta}_{B0} - \hat{\beta}_{A0} \right) + \sum_{k=1}^K \overline{X_{Ak}} (\hat{\beta}_{Bk} - \hat{\beta}_{Ak})}_{\hat{\Delta}_S \text{ (inexpliqué)}}, \quad (1)$$

où l'on a introduit $\hat{\beta}_{B0} + \sum_{k=1}^K \overline{X_{Ak}} \hat{\beta}_{Bk}$ afin de mener la décomposition. $\hat{\Delta}_X$ renvoie à la partie de l'écart de salaire liée à l'écart de caractéristiques observables entre les deux groupes, écart que l'on valorise ici selon les paramètres $\hat{\beta}_{Bk}$ estimés pour le groupe B : on appellera cette grandeur l'écart expliqué (ou *effet de composition*). $\hat{\Delta}_S$ correspond à la part liée à l'écart de valorisation des caractéristiques (et à l'écart de constante), valorisations qui ici sont appliquées aux caractéristiques du groupe A . On désignera ce terme comme “écart inexpliqué”, puisque les écarts de caractéristiques observables ne permettent pas d'en rendre compte.

Encadré 1 : L'ambiguïté du terme d'*effet de structure*

La littérature des méthodes de décomposition appelle aussi l'écart inexpliqué $\hat{\Delta}_S$ le “wage structure effect”. Cette appellation vient de l'hypothèse qu'il existe une fonction structurelle des salaires qui diffère entre les groupes comparés. Autrement dit, la structure à laquelle il est fait référence dans ce terme est une structure de *valorisation des caractéristiques*, et non une structure de *caractéristiques*- comme l'entend le plus souvent le français courant dans le terme d’“effet de structure”. Cela pouvant être source de confusion, on évitera ici de parler d'effet structurel.

La notion de structure est aussi centrale dans d'autres méthodes de décomposition, comme les approches dites structurelles géographiques. Celles-ci permettent d'analyser les différences d'évolution entre territoires, entre ce qui tient des *structures* sectorielles spécifiques et des effets résiduels (compétitivité locale, capacités d'innovation, etc.) qui correspondent à la différence entre la croissance de l'ensemble des territoires et de la zone d'intérêt à *structure* productive donnée. Pour plus d'éléments sur ces méthodes, on pourra se reporter à [Kubrak \(2018\)](#).

1.2 Ecart expliqué et inexpliqué de salaire : un exemple avec l'Enquête emploi

Application 1 : La décomposition agrégée

On illustre la décomposition d'Oaxaca-Blinder par l'étude des différences de salaires entre hommes et femmes, à partir de l'enquête Emploi en continu entre 2013 et 2016. Si l'on se réfère à la décomposition de l'équation (1), et que l'on souhaite décomposer $\overline{Y_B} - \overline{Y_A}$, B correspondra aux hommes et A aux femmes. La variable d'intérêt est le logarithme du salaire

mensuel net². Celui-ci vaut en moyenne 7.572 chez les hommes, et 7.273 chez les femmes, soit un écart de 0.299.

On introduit les variables explicatives suivantes :

- l'expérience potentielle qui mesure pour un individu le nombre d'années écoulées depuis la fin de sa formation initiale (*exp_mtra*) et son carré (*exp_mtra2resc*),
- le niveau d'études (*ddipl*) en 6 postes : diplôme supérieur à baccalauréat + 2 ans (modalité 1), baccalauréat + 2 ans (3), baccalauréat ou brevet professionnel ou autre diplôme de ce niveau (4), CAP, BEP ou autre diplôme de ce niveau (5), brevet des collèges (6) et certificat d'études primaires ou aucun diplôme (7 - la modalité de référence retenue).
- le secteur d'activité en 10 modalités (*secteur*) : agriculture, sylviculture et pêche (AZ), industrie manufacturière, industries extractives et autres (BE), construction (FZ), Commerce de gros et de détail, transports, hébergement et restauration (GI - la modalité de référence retenue), information et communication (JZ), activités financières et d'assurance (KZ), activités immobilières (LZ), activités spécialisées, scientifiques et techniques et activités de services administratifs et de soutien (MN), administration publique, enseignement, santé humaine et action sociale (OQ) et autres activités de services (RU).
- l'ancienneté dans l'entreprise en 4 modalités : 10 ans ou plus (modalité *ancentr44*), de 5 à moins de 10 ans (*ancentr43*), de 1 à moins de 5 ans (*ancentr42*) ou moins d'un an (*ancentr41* – modalité de référence retenue).
- une indicatrice pour le fait d'être à temps partiel (*tpartiel*).

Afin d'effectuer la décomposition d'Oaxaca-Blinder correspondante, on estime les coefficients de l'équation de salaire dans chacun des groupes.

On s'intéresse dans un premier temps au partage *global* entre effet de composition $\hat{\Delta}_X$ et écart inexpliqué $\hat{\Delta}_S$: c'est la décomposition agrégée. Pour ce faire, on a en fait seulement besoin d'estimer le modèle chez les hommes³ (*sex=0*), mais la comparaison étant d'intérêt on procède aussi à l'estimation chez les femmes (*sex=1*).

```
#Estimation dans le groupe A (femmes) et enregistrement des coefficients
modele.A <- lm(logsal ~ exp_mtra + exp_mtra2resc +
  as.factor(ddipl) + tpartiel + secteurOQ +
  secteurBE + secteurRU + secteurFZ + secteurMN +
  secteurAZ + secteurKZ + secteurJZ + secteurLZ +
  ancentr44 + ancentr43 + ancentr42,
```

2. Il serait sans doute plus pertinent de considérer le logarithme du salaire horaire net. Cependant, à des fins pédagogiques et pour faciliter l'interprétation des résultats, on utilise le salaire mensuel.

3. voir la fin de la section 1.2 pour plus de détails.

```

data = data[data$sex==1,])
coeffs.A <- modele.A$coefficients

#Estimation dans le groupe B (hommes)
modele.B <- lm(logsal ~ exp_mtra + exp_mtra2resc +
                  as.factor(ddipl) + tpartiel + secteur0Q +
                  secteurBE + secteurRU + secteurFZ + secteurMN +
                  secteurAZ + secteurKZ + secteurJZ + secteurLZ +
                  ancentr44 + ancentr43 + ancentr42,
                  data = data[data$sex==0,])
coeffs.B <- modele.B$coefficients

#Comparaison des estimateurs obtenus dans les 2 groupes
round(cbind(coeffs.A,coeffs.B),3)

##                                     coeffs.A coeffs.B
## (Intercept)          6.721    6.827
## exp_mtra           0.010    0.025
## exp_mtra2resc     -0.016   -0.035
## as.factor(ddipl)1  0.684    0.718
## as.factor(ddipl)3  0.504    0.434
## as.factor(ddipl)4  0.314    0.287
## as.factor(ddipl)5  0.194    0.134
## as.factor(ddipl)6  0.175    0.169
## tpartiel          -0.510   -0.672
## secteur0Q          -0.044   -0.101
## secteurBE          0.091    0.060
## secteurRU          -0.245   -0.154
## secteurFZ          0.030    0.036
## secteurMN          0.008    0.001
## secteurAZ          -0.151   -0.119
## secteurKZ          0.136    0.183
## secteurJZ          0.197    0.124
## secteurLZ          0.011    -0.060
## ancentr44          0.367    0.202
## ancentr43          0.209    0.119
## ancentr42          0.113    0.083

```

On calcule ensuite les moyennes pour chaque variable, pour chacun des deux groupes. Dans le

cas des variables catégorielles introduites dans le modèle sous forme de facteur (`as.factor`), ici le diplôme, on a besoin des proportions pour chacune des modalités (hors référence). On réécrit également les variables catégorielles comme autant d'indicatrices qu'il y a de modalités, afin de pouvoir associer directement les proportions dans chacune des modalités aux coefficients correspondants estimés plus haut. Pour faire cette transformation automatiquement, on peut utiliser la fonction `model.matrix`.

```
#Les variables catégorielles sont transformées en indicatrices
X.A <- model.matrix(~ exp_mtra + exp_mtra2resc
+ as.factor(ddipl) + tpartiel + secteur0Q
+ secteurBE + secteurRU + secteurFZ + secteurMN
+ secteurAZ + secteurKZ + secteurJZ + secteurLZ
+ ancentr44 + ancentr43 + ancentr42,
data = data[data$sex==1,])

#On applique la fonction moyenne pour chaque variable
#(donc pour chaque colonne - d'où le paramètre margin=2)
X.moy.A<-apply(X.A, 2, mean)

#idem dans le groupe B
X.B <- model.matrix(~ exp_mtra + exp_mtra2resc
+ as.factor(ddipl) + tpartiel + secteur0Q
+ secteurBE + secteurRU + secteurFZ + secteurMN
+ secteurAZ + secteurKZ + secteurJZ + secteurLZ
+ ancentr44 + ancentr43 + ancentr42,
data = data[data$sex==0,])

X.moy.B<-apply(X.B, 2, mean)

#On compare les caractéristiques moyennes dans les groupes A et B
round(cbind(X.moy.A,X.moy.B),3)

## X.moy.A X.moy.B
## (Intercept) 1.000 1.000
## exp_mtra 22.498 22.230
## exp_mtra2resc 6.413 6.273
## as.factor(ddipl)1 0.218 0.192
## as.factor(ddipl)3 0.186 0.141
## as.factor(ddipl)4 0.201 0.185
```

```

## as.factor(ddipl)5 0.232 0.304
## as.factor(ddipl)6 0.055 0.049
## tpartiel 0.306 0.056
## secteurOQ 0.482 0.211
## secteurBE 0.086 0.227
## secteurRU 0.065 0.030
## secteurFZ 0.013 0.100
## secteurMN 0.086 0.089
## secteurAZ 0.007 0.016
## secteurKZ 0.041 0.028
## secteurJZ 0.017 0.037
## secteurLZ 0.014 0.011
## ancentr44 0.505 0.510
## ancentr43 0.172 0.171
## ancentr42 0.220 0.219

```

Pour retrouver l'effet de composition défini dans l'équation (1), il reste seulement à appliquer les coefficients estimés chez les hommes aux différences entre caractéristiques moyennes chez les hommes et chez les femmes et à sommer pour toutes les variables. Cela donne :

```

#Ecarts de caractéristiques moyennes valorisés comme dans le groupe B
sum((X.moy.B - X.moy.A)*coeffs.B)

## [1] 0.177

```

à rapporter à un écart total de log salaire de 0.299 entre hommes et femmes. L'effet de composition représente ainsi 59.1 % de l'écart total de salaire observé entre les sexes. Autrement dit, 59.1 % de l'écart de salaire observé entre hommes et femmes à partir de l'enquête Emploi peut être attribué à des caractéristiques moyennes différentes entre les sexes. On peut vérifier que, mécaniquement, l'écart inexpliqué correspond bien à 0.299 - 0.177 soit 0.122 :

```

#Différence de coefficients appliquée aux caractéristiques moyennes des A
sum(X.moy.A*(coeffs.B-coeffs.A))

## [1] 0.122

```

On peut voir que dans l'exemple précédent, où l'on s'intéresse uniquement à la décomposition agrégée, il suffit en fait d'estimer le jeu de coefficients $(\hat{\beta}_{B,k})_{k=1 \dots K}$ des hommes pour obtenir la décomposition souhaitée. En effet, on peut réécrire :

$$\overline{Y_B} - \overline{Y_A} = \underbrace{\overline{Y_B} - \sum_{k=0}^K \overline{X_{Ak}} \hat{\beta}_{Bk}}_{\hat{\Delta}_X} + \underbrace{\sum_{k=0}^K \overline{X_{Ak}} \hat{\beta}_{Bk} - \overline{Y_A}}_{\hat{\Delta}_S} \quad (2)$$

On ne s'appuie ici que sur les $\hat{\beta}_{Bk}$, et non sur les $\hat{\beta}_{Ak}$: cette formulation de la décomposition “agrégée” est utile lorsque le groupe A compte peu d’observations, ce qui conduirait à des résultats identiques en moyenne mais moins précis si l’on devait s’appuyer également sur les coefficients estimés dans ce groupe.

1.3 Références de la décomposition

On a implicitement introduit un salaire de référence valant $\hat{\beta}_{B0} + \sum_{k=1}^K \overline{X_{Ak}} \hat{\beta}_{Bk}$. Il correspond au salaire prédit pour les caractéristiques observables moyennes du groupe A valorisées selon l’équation de salaire estimée dans le groupe B. On note ce salaire $Y^{C,A}$, et on le désignera également comme salaire “contrefactuel”⁴. La question posée par ce salaire contrefactuel peut en effet se formuler ainsi : que gagneraient les individus du groupe A si leurs caractéristiques étaient valorisées de la même manière que pour les B ? L’écart entre ce terme $Y^{C,A}$ et le salaire moyen du groupe B, $\hat{\beta}_{B0} + \sum_{k=1}^K \overline{X_{Bk}} \hat{\beta}_{Bk}$, résulte uniquement de différences de caractéristiques : on retrouve l’effet de composition. L’écart entre $Y^{C,A}$ et le salaire moyen du groupe A correspond à l’écart inexpliqué.

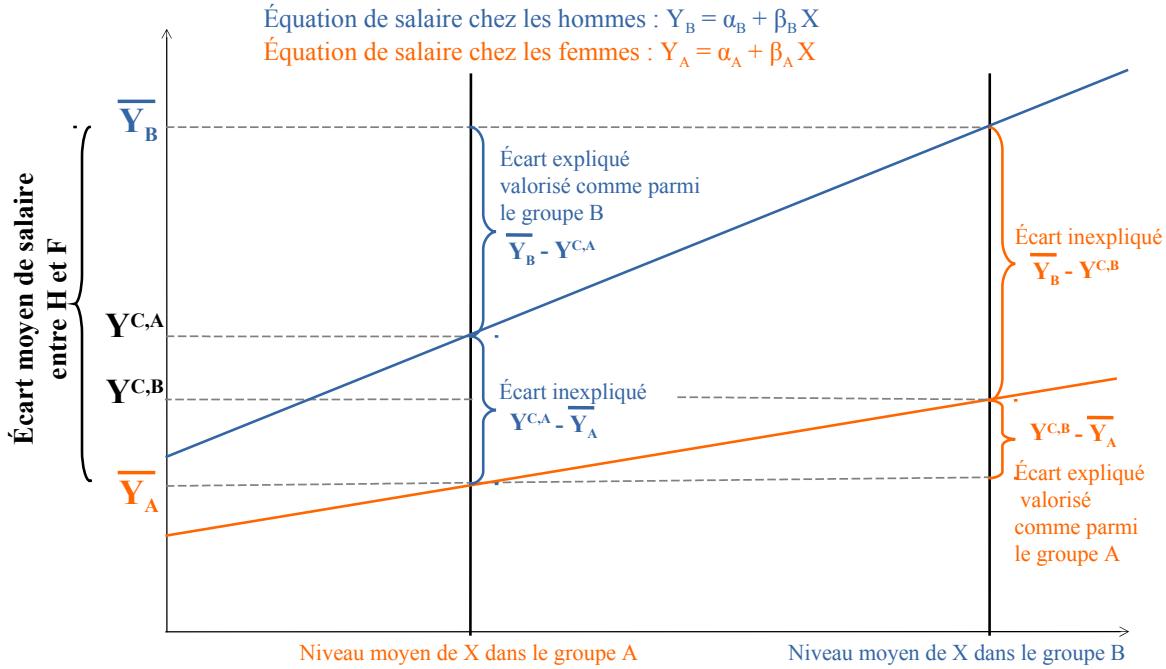
Ceci est illustré en figure 1 qui présente un cas simple où l’on dispose d’une seule variable observable X . En bleu (respectivement orange) est représentée l’équation de salaire chez les hommes (resp. les femmes). $\overline{Y_B}$ correspond au salaire moyen chez les hommes, qui ont un niveau moyen de caractéristiques plus élevé que les femmes, dont le salaire moyen est noté $\overline{Y_A}$. En bleu, la décomposition va s’articuler autour du point $Y^{C,A}$, qui valorise les caractéristiques moyennes du groupe A selon l’équation de salaire du groupe B, ce qui est explicité par l’expression de la décomposition en équation (1) :

$$\overline{Y_B} - \overline{Y_A} = \underbrace{\sum_{k=1}^K (\overline{X_{Bk}} - \overline{X_{Ak}}) \hat{\beta}_{Bk}}_{\hat{\Delta}_X \text{ (expliqué)}} + \underbrace{\left(\hat{\beta}_{B0} - \hat{\beta}_{A0} \right) + \sum_{k=1}^K \overline{X_{Ak}} (\hat{\beta}_{Bk} - \hat{\beta}_{Ak})}_{\hat{\Delta}_S \text{ (inexpliqué)}}. \quad (1)$$

Un contrefactuel alternatif à $Y^{C,A}$ correspondrait au salaire qu’aurait le groupe B si ses caractéristiques étaient valorisées comme celles du groupe A, c’est-à-dire $\hat{\beta}_{A0} + \sum_{k=1}^K \overline{X_{Bk}} \hat{\beta}_{Ak}$. On note ce contrefactuel $Y^{C,B}$ et on dessine en bleu les accolades illustrant la décomposition suivant ce contrefactuel sur la figure 1. La décomposition correspondante est la suivante :

4. On emploie ici le terme contrefactuel à la façon de Fortin et al. (2011) pour désigner le salaire de référence de la décomposition- celui qu’auraient par exemple les femmes si, à caractéristiques observables inchangées, celles-ci étaient valorisées comme parmi les hommes. Ce salaire de référence ne s’interprète pas de façon causale.

Figure 1 – Décomposition d'écart moyen de salaire entre les groupes B et A



$$\bar{Y}_B - \bar{Y}_A = \underbrace{\sum_{k=1}^K (\bar{X}_{Bk} - \bar{X}_{Ak}) \hat{\beta}_{Ak}}_{\hat{\Delta}_X \text{ (expliqué)}} + \underbrace{\left(\hat{\beta}_{B0} - \hat{\beta}_{A0} \right) + \sum_{k=1}^K \bar{X}_{Bk} (\hat{\beta}_{Bk} - \hat{\beta}_{Ak})}_{\hat{\Delta}_S \text{ (inexpliqué)}} \quad (3)$$

Ici, l'écart de caractéristiques entre les deux groupes est donc valorisé selon les coefficients $\hat{\beta}_A$, qui constituent les valorisations de référence, et non selon les $\hat{\beta}_B$ comme c'était le cas dans la décomposition présentée en équation (1).

Rien n'empêche de considérer n'importe quel autre vecteur de coefficients β_Ω comme la référence de la décomposition. L'écart inexplicable comprend alors un terme supplémentaire, la décomposition s'écrivant :

$$\bar{Y}_B - \bar{Y}_A = \left(\hat{\beta}_{B0} - \hat{\beta}_{A0} \right) + \underbrace{\sum_{k=1}^K \bar{X}_{Bk} (\hat{\beta}_{Bk} - \hat{\beta}_{\Omega k}) + \sum_{k=1}^K \bar{X}_{Ak} (\hat{\beta}_{\Omega k} - \hat{\beta}_{Ak})}_{\hat{\Delta}_S}$$

$$+ \underbrace{\sum_{k=1}^K (\bar{X}_{Bk} - \bar{X}_{Ak}) \hat{\beta}_{\Omega k}}_{\hat{\Delta}_X} \quad (4)$$

On pourra par exemple choisir comme coefficients β_Ω ceux estimés sur l'ensemble de la population. L'écart inexpliqué obtenu dans une telle approche, qui impose que les valorisations des caractéristiques soient communes dans les deux groupes, correspond à l'estimateur $\hat{\beta}_0$ obtenu dans l'équation :

$$Y_i = \beta'_0 + \sum_{k=1}^K X_{ik} \beta'_k, \quad \forall i \in A \cup B. \quad (5)$$

Alternativement, on peut intégrer à cette équation de salaire en population générale une indicatrice d'appartenance à l'un des deux groupes :

$$Y_i = \beta''_0 + \sum_{k=1}^K X_{ik} \beta''_k + \mathbf{1}_{i \in B} \gamma_B, \quad \forall i \in A \cup B. \quad (6)$$

On a là encore une valorisation de référence commune aux deux groupes, sauf pour la constante qu'on autorise à différer (via le coefficient $\hat{\gamma}_B$). Dans ce cas, mener une décomposition d'Oaxaca-Blinder à partir des $\hat{\beta}''$ et calculer l'écart inexpliqué fera retomber précisément sur $\hat{\gamma}_B$ estimé dans l'équation (6).

La question du salaire de référence permet ainsi de faire le lien entre méthodes de décomposition et une autre méthode courante d'analyse des écarts de salaire, qui est l'interprétation directe de l'équation (6) : $\hat{\gamma}_B$ y est simplement lu comme une estimation de l'écart inexpliqué entre les deux groupes, en contrôlant des différences de caractéristiques observables. Autrement dit, la méthode de l'indicatrice dans une équation de salaire en population générale est un cas particulier de la méthode d'Oaxaca-Blinder, où les valorisations de référence sont celles de l'équation (6). On détaille en section 2.2 les questions à se poser pour choisir la référence de la décomposition.

1.4 La décomposition détaillée de l'effet de composition

Afin d'avoir une vision plus fine des mécanismes jouant sur l'effet de composition, il est possible de détailler celui-ci variable par variable. Ainsi, on peut considérer un à un au sein de $\hat{\Delta}_X$,

chacun des termes liés à une variable explicative X_k en particulier :

$$\hat{\Delta}_X = \sum_{k=1}^K \hat{\Delta}_{X_k},$$

où pour chaque covariable X_k , $\hat{\Delta}_{X_k}$ désigne sa contribution à l'écart expliqué :

$$\hat{\Delta}_{X_k} = (\overline{X_{Bk}} - \overline{X_{Ak}}) \hat{\beta}_{Bk}, \quad k \in 1 \dots K.$$

Comme dans le cas simple de la décomposition agrégée, on n'a besoin d'estimer que les valorisations des caractéristiques du groupe B pour calculer chacun des termes de l'effet de composition.

Application 2 : Détailler l'écart expliqué avec le package `Oaxaca`

On utilise à présent le package `Oaxaca` qui permet d'automatiser les calculs des écarts expliqué et inexpliqué, de comparer différentes références et de détailler l'analyse variable par variable. On pourra se reporter à [Hlavac \(2014\)](#) pour plus de détails. L'exemple d'application est le même que précédemment.

```
library("oaxaca")
```

On utilise la fonction `oaxaca` pour renseigner le modèle linéaire sur lequel est fondé la décomposition (on a au préalable passé les variables catégorielles en variables dichotomiques, ce qu'impose la fonction) et la variable permettant de distinguer les deux groupes à comparer. Par défaut, les erreurs sont calculées par bootstrap, à partir de 100 réplications. On peut modifier le nombre de réplications du bootstrap en spécifiant le paramètre `R`.

```
#On utilise les mêmes variables qu'en section 1.2.
results <- oaxaca(formula = logsal ~ exp_mtra + exp_mtra2resc
+ ddip16 + ddip15 + ddip14 + ddip13 + ddip11
+ tpartiel + secteur0Q + secteurBE + secteurRU
+ secteurFZ + secteurMN + secteurAZ + secteurKZ
+ secteurJZ + secteurLZ + ancentr44 + ancentr43
+ ancentr42 | sex , data = data, R=10)
```

Une fois les paramètres de la décomposition estimés, on peut afficher différentes sorties, comme la composante `n` qui renvoie le nombre d'observations dans les deux groupes ou `y` qui donne les salaires moyens dans chaque groupe et la différence entre les deux. Pour avoir un aperçu de toutes les sorties on peut utiliser la commande `str(results$twofold)`. Plus intéressant, on peut afficher les résultats de la décomposition agrégée :

```
# On ne montre ici que les colonnes 1, 2 et 4 de la table de sortie.
```

```
# Les colonnes 3 et 5 contiennent les écarts-types
```

```
round(results$twofold$overall[, c(1, 2, 4)], 3)
```

	group.weight	coef(explained)	coef(unexplained)
[1,]	0.000	0.133	0.166
[2,]	1.000	0.177	0.122
[3,]	0.500	0.155	0.144
[4,]	0.489	0.154	0.144
[5,]	-1.000	0.176	0.123
[6,]	-2.000	0.144	0.155

La colonne `group.weight` indique à partir de quelle référence est calculée la décomposition. Plus précisément elle donne le poids relatif accordé au groupe tel que la variable `sex` soit égale à zéro (ici, les hommes) par rapport au groupe pour lequel elle vaut 1 (ici, les femmes) :

- pour la ligne 0, les coefficients de référence sont estimés chez les femmes (les hommes ont un poids relatif de 0), ce qui correspond à la décomposition de l'équation (3). L'effet de composition obtenu avec cette décomposition correspond donc à l'écart de salaire entre hommes et femmes lié à leur différence de caractéristiques lorsqu'on les valorise comme chez les femmes.
- pour la ligne 1, on a les résultats de la décomposition avec les coefficients de référence estimés chez les hommes (voir équation (1)). On peut remarquer qu'on retrouve bien le même résultat que dans l'application 1, qui utilisait les valorisations des hommes comme référence.
- 0.5 : moyenne (non pondérée) des coefficients estimés séparément dans chacun des groupes. Implicitement, cela revient à se rapporter à une valorisation moyenne des caractéristiques (Reimers, 1983).
- 0.489 : moyenne pondérée des coefficients estimés dans chaque groupe (Cotton, 1988). Le poids relatif des hommes est légèrement inférieur à 0,5 ce qui traduit le fait qu'on observe un peu plus de femmes que d'hommes dans les données.
- -1 : coefficients de référence estimés sur l'ensemble de la population, sans indicatrice de groupe (Neumark, 1988) (formule (1.3), avec les coefficients de l'équation (5)). Cette décomposition revient à considérer comme valorisation de référence une valorisation strictement identique des caractéristiques entre les deux groupes.
- -2 : coefficients de référence estimés sur l'ensemble de la population mais avec indicatrice de groupe (Jann, 2008) (formule (1.3), avec les coefficients de l'équation (6)). La référence considérée autorise donc seulement la constante du modèle à différer entre les deux groupes comparés. Cette méthode correspond à celle présentée à la fin de la

section 1.3.

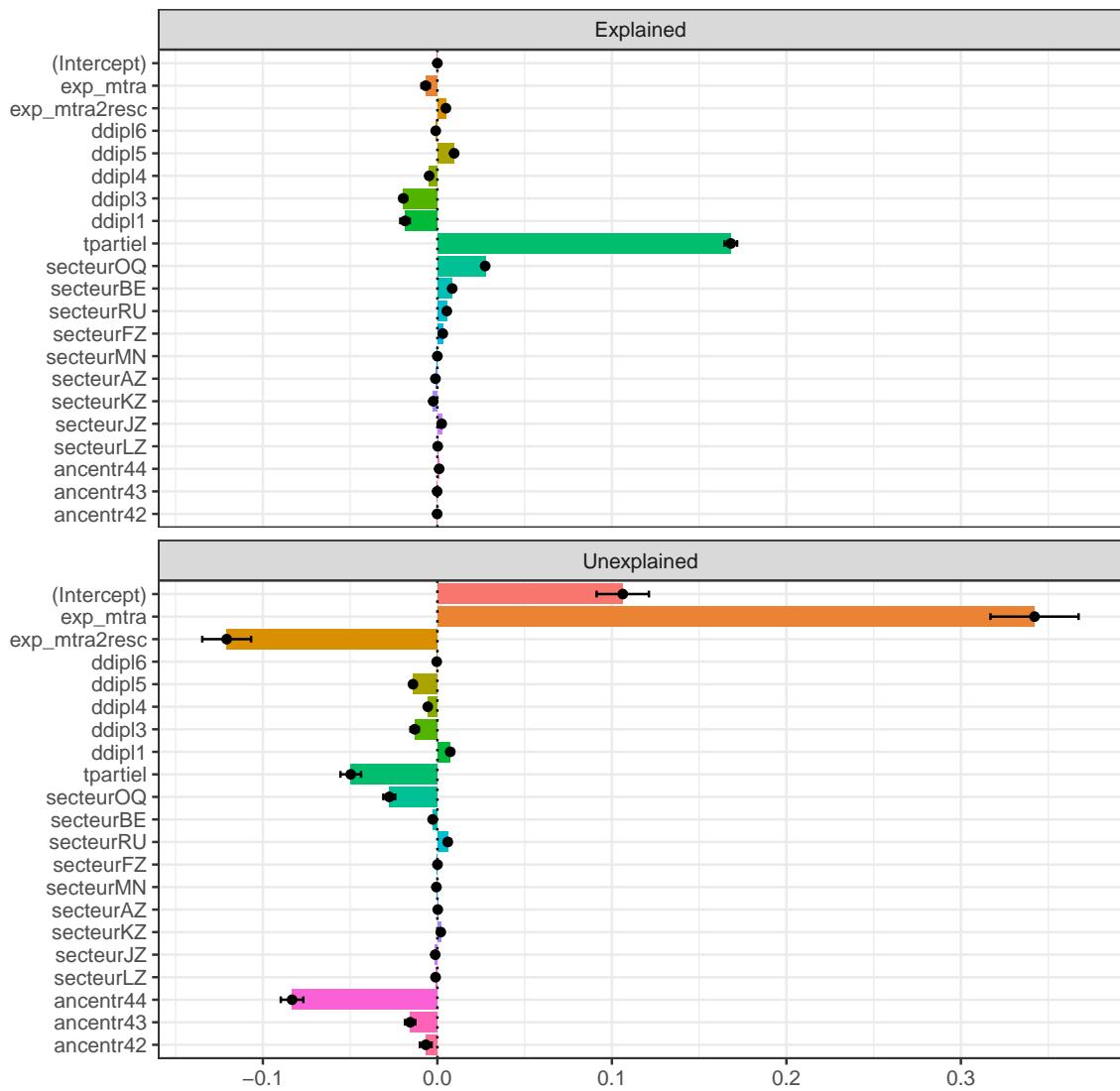
Ainsi, pour le modèle `group.weight= 1`, on retrouve comme précédemment que l'écart expliqué vaut 0.177 et correspond à la différence entre le salaire moyen des hommes et le salaire que toucheraient les femmes si leurs caractéristiques étaient valorisées comme celles des hommes. L'écart inexpliqué, de 0.122 point, correspond à la différence entre le salaire que toucherait les femmes si leurs caractéristiques étaient valorisées comme celles des hommes, et le salaire moyen effectivement observé chez les femmes. Cette répartition entre expliqué et inexpliqué varie avec la référence retenue. Par exemple, on trouve, en considérant comme *contre-factuel* de la décomposition le salaire que toucheraient les hommes si leurs caractéristiques étaient valorisées comme celles des femmes (`group.weight= 0`), un effet de composition de 0.133 et un écart inexpliqué de 0.166.

Pour chaque ensemble de coefficients de référence on peut ensuite afficher les résultats détaillés, c'est-à-dire la contribution de chaque variable à l'écart expliqué et inexpliqué, soit en graphique, soit en tableau.

	coef(explained)	se(explained)	coef(unexplained)	se(unexplained)
(Intercept)	0.000	0.000	0.106	0.008
exp_mtra	-0.007	0.001	0.342	0.013
exp_mtra2resc	0.005	0.001	-0.121	0.007
ddipl6	-0.001	0.000	0.000	0.000
ddipl5	0.010	0.000	-0.014	0.001
ddipl4	-0.005	0.001	-0.006	0.001
ddipl3	-0.020	0.001	-0.013	0.001
ddipl1	-0.019	0.001	0.007	0.001
tpartiel	0.168	0.002	-0.050	0.003
secteurOQ	0.027	0.001	-0.028	0.002
secteurBE	0.009	0.000	-0.003	0.000
secteurRU	0.005	0.000	0.006	0.001
secteurFZ	0.003	0.000	0.000	0.000
secteurMN	0.000	0.000	-0.001	0.000
secteurAZ	-0.001	0.000	0.000	0.000
secteurKZ	-0.002	0.000	0.002	0.000
secteurJZ	0.002	0.000	-0.001	0.000
secteurLZ	0.000	0.000	-0.001	0.000

ancentr44	0.001	0.000	-0.083	0.003
ancentr43	0.000	0.000	-0.016	0.001
ancentr42	0.000	0.000	-0.007	0.002

```
#La figure de la décomposition détaillée avec group.weight= 1
plot(results, decomposition = "twofold", group.weight = 1)
```



La partie supérieure du graphique présente la contribution de chaque variable à l'écart expliqué. La variable qui contribue le plus positivement à l'écart expliqué est l'indicatrice de temps partiel, avec une contribution de 0.168 soit 95.2 % de l'écart expliqué total. Autrement dit, presque l'intégralité de l'effet de composition tient au fait que les femmes sont plus sou-

vent en emploi à temps partiel. On notera que l'inclusion de certaines variables peut réduire l'écart expliqué : c'est par exemple le cas pour certains niveaux de diplôme. En effet, quand les femmes sont dotées de caractéristiques plus favorables en termes de salaire que les hommes, contrôler de ces caractéristiques réduit la part des écarts qui peut être imputée aux X . Ainsi, on ne peut pas exclure d'obtenir un effet de composition de signe contraire à celui de l'écart initial que l'on cherche à décomposer (et par conséquent, un écart inexpliqué plus élevé que l'écart initial). Ce type de cas, s'il invite à s'interroger sur l'interprétation de la décomposition, ne remet pas en cause sa validité.

La partie inférieure du graphique ventile l'écart inexpliqué par variable : de même qu'il est possible de détailler les contributions de chaque variable à l'effet de composition, on peut aussi obtenir le détail de l'écart inexpliqué. Cependant, des hypothèses supplémentaires et des précautions particulières sont nécessaires pour analyser et interpréter ces résultats. On renvoie le lecteur à la section 2.3 pour plus d'éléments sur la décomposition détaillée de l'écart inexpliqué.

2 La validité de l'interprétation

2.1 Effet causal d'appartenance à un groupe et discrimination

Les méthodes de décomposition sont fréquemment utilisées dans le but de mesurer une discrimination entre deux groupes, soit une différence de traitement qui n'est dûe qu'au fait d'appartenir à un groupe plutôt qu'à l'autre. Dans ce cas, l'objectif est d'isoler un effet causal d'appartenance au groupe. Sous quelles conditions un écart inexplicable peut-il être interprété comme un effet causal de l'appartenance à un groupe plutôt qu'à l'autre - et donc comme une discrimination ?

Pour assimiler l'écart inexplicable à un effet causal, il faut être en mesure d'affirmer qu'aucune différence de caractéristiques inobservées ne subsiste entre les deux groupes, une fois qu'on a contrôlé des caractéristiques observables (encadré 2 pour une définition formelle). C'est une hypothèse forte. Prenons l'exemple des écarts de salaire entre hommes et femmes, lorsque l'on dispose comme variables de contrôle de l'âge, du diplôme et du fait d'être cadre. Une partie de l'écart de salaire entre hommes et femmes est liée aux différences d'âge, de diplôme et de statut entre les hommes et les femmes présents sur le marché du travail. On ne pourra interpréter le reste de l'écart comme de la discrimination que si, pour chaque niveau d'âge, de diplôme et de statut, les hommes et les femmes ont bien un niveau de compétences identique.

Plusieurs raisons peuvent conduire à ce que cette hypothèse d'indépendance conditionnelle ne soit pas vérifiée :

- De façon générale, les deux groupes peuvent différer selon une variable que l'on n'observe pas (variable omise). L'expérience effective sur le marché du travail pourrait par exemple être plus élevée, à âge donné, chez les hommes que chez les femmes. Dans ce cas, l'écart inexplicable sur-estime le niveau de discrimination car il est en réalité gonflé par une composante qui devrait appartenir à l'écart expliqué.
- L'existence d'une sélection différenciée conduit fréquemment à un tel problème de variable omise, alors même que les deux groupes peuvent être initialement comparables. Si les femmes accèdent plus difficilement à l'emploi que les hommes, les femmes sélectionnées sur le marché du travail pourraient avoir une motivation plus forte que les hommes d'âge et diplôme identiques, motivation qui ne serait pas rétribuée, ou dont la rétribution serait à tort attribuée à d'autres caractéristiques. Dans un tel cas, l'écart attribuable à de la discrimination sera sous-estimé.
- Les variables explicatives incluses dans le modèle peuvent elles-mêmes faire l'objet d'une sélection : si les femmes sont par exemple plus rigoureusement sélectionnées pour accéder au statut de cadre, et qu'on contrôle par le fait d'être cadre, on pourra conclure à l'absence de discrimination alors même que les femmes ont une motivation

plus grande à niveau d'observables donné.

Encadré 2 : Décompositions, modèle de Rubin, discrimination

Un individu i est doté des caractéristiques X_i . Soit un “traitement” binaire $T : T_i = 0$ si $i \in \mathcal{P}_0$, $T_i = 1$ si $i \in \mathcal{P}_1$. Les *outcomes* (par exemple les salaires) potentiels s'écrivent :

- $Y_i(0)$ pour l'individu i si $T_i = 0$,
- $Y_i(1)$ si $T_i = 1$.

La variable de traitement T correspond, dans le cadre des décompositions, à la variable d'appartenance à un groupe plutôt qu'à l'autre (variable de sexe par exemple). Pour un individu donné, on n'observe pas les deux salaires potentiels, mais seulement la réalisation effective de la variable d'intérêt selon que l'individu appartient à l'un ou l'autre groupe, soit :

$$Y_i = (1 - T_i)Y_i(0) + T_iY_i(1).$$

Si le modèle est linéaire de la forme $\mathbb{E}(Y | X) = X\beta$ et que l'hypothèse $Y_i(0), Y_i(1) \perp T_i | X_i, \forall i$ (indépendance conditionnelle, voir section 2.1) est vérifiée, alors $\bar{X}_1\hat{\beta}_0$, soit les caractéristiques moyennes dans le groupe traité valorisées comme celles des individus non traités, est un estimateur convergent de $\mathbb{E}(Y(0) | T = 1)$. Alors, la décomposition de Oaxaca-Blinder :

$$\bar{Y}_1 - \bar{Y}_0 = (\bar{X}_1\hat{\beta}_1 - \bar{X}_1\hat{\beta}_0) + (\bar{X}_1 - \bar{X}_0)\hat{\beta}_0$$

peut être vue comme la contrepartie empirique de :

$$\mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) = \mathbb{E}(Y(1) | T = 1) - \mathbb{E}(Y(0) | T = 1) + \mathbb{E}(Y(0) | T = 1) - \mathbb{E}(Y(0) | T = 0).$$

La mesure de discrimination (écart inexpliqué dans Oaxaca-Blinder) correspond ainsi à l'*average treatment effect on the treated* (ATT), soit l'effet du traitement une fois que l'on a contrôlé des différences de caractéristiques entre groupe traité et groupe de contrôle. Cette mesure de discrimination quantifie un effet causal sous l'hypothèse d'indépendance conditionnelle.

Ces limites de la validité de l'hypothèse d'indépendance conditionnelle doivent être prises en compte dans le choix des variables explicatives (pour d'autres précautions quant au choix des explicatives, voir l'encadré 3). Il y a ainsi un équilibre à trouver en pratique entre l'introduction de contrôles ayant un pouvoir explicatif important et/ou qui sont intéressants pour l'analyse, et la prudence quant aux facteurs qui pourraient fragiliser la condition d'identification. Il faut donc être attentif à ne pas “trop” contrôler et à questionner le choix des variables explicatives incluses dans le modèle : est-ce que pour l'ensemble des X introduits la

comparaison des deux groupes a bien un sens ? En général, les variables résultant d'un choix de l'individu violent l'hypothèse d'indépendance conditionnelle. Un procédé utile lorsqu'on a recours à de telles variables est d'introduire les explicatives au fur et à mesure : on commence par les *pre-market factors* -les caractéristiques des individus déterminées avant leur entrée sur le marché du travail-, puis on ajoute les variables résultant (en partie) de choix individuels comme la catégorie sociale. On peut ainsi présenter les deux décompositions et préciser que dans la deuxième il est difficile d'assimiler l'écart inexpliqué à une discrimination.

Par ailleurs les cas suivants, peu ou pas pertinents dans le cas hommes/femmes, peuvent être rencontrés et rendre invalide l'hypothèse d'indépendance conditionnelle :

- Le fait que l'appartenance au groupe soit le résultat d'une décision de l'individu (auto-sélection), par exemple si l'on cherche à étudier les écarts entre public et privé ou encore entre groupes définis selon leur lieu de résidence. Ainsi, les salariés qui choisissent de travailler dans le secteur privé y ont un intérêt plus grand (une espérance de salaires plus élevée par exemple), ce qui se traduit par des inobservables différentes. De même, les choix résidentiels peuvent être liés aux caractéristiques inobservées des individus - par exemple si les salariés résidant près de certaines zones d'emploi sont plus motivés à niveau de caractéristiques observables donné.
- L'inclusion de variables ne mesurant pas le même phénomène selon le groupe considéré : par exemple lorsque l'on compare immigrés et non-immigrés, ou deux pays dans le cadre d'une comparaison internationale, la variable de diplôme ne reflète pas nécessairement le même niveau de compétences selon le pays dans lequel l'individu a fait ses études.

L'hypothèse d'indépendance conditionnelle autorise que l'effet d'une variable sur le salaire soit mesurée avec biais sur chaque sous-groupe – par exemple l'effet du diplôme sur le salaire capte également l'effet d'une motivation croissante – tant que la structure de corrélation entre diplôme et motivation est la même dans les deux groupes (à niveau de diplôme donné, hommes et femmes ont la même motivation)⁵. Attention, cela n'est plus vrai dès lors que l'on cherche à isoler la contribution de chaque variable dans la décomposition détaillée, par exemple connaître la part effectivement liée aux écarts d'éducation dans les écarts de salaire (sans capter par la même occasion la part liée aux écarts de motivation). Cette question sera à nouveau abordée dans la section 2.3.

Encadré 3 : Questions de support commun

5. Ainsi on autorise au total des différences de caractéristiques inobservées (de motivation par exemple) entre les deux groupes, tant que ces différences sont uniquement liées aux différences de caractéristiques observables (les plus diplômés sont plus motivés, or l'un des groupes est plus diplômé).

Les groupes A et B peuvent différer largement en termes de caractéristiques observables. Que se passe-t-il si l'un des groupes prend des valeurs très atypiques pour l'une des variables considérées ?

- dans le cas d'une variable catégorielle :
 - tant qu'on ne souhaite pas procéder à la décomposition détaillée de l'écart inexpiqué, on utilise uniquement les valorisations de référence (par exemple celles du groupe B). Il faut dans ce cas que chacune des modalités dans lesquelles on observe des individus du groupe A soit également représentée parmi le groupe B. Pour les écarts de salaire hommes-femmes, il pourrait par exemple être problématique de ne pas observer d'hommes exerçant le métier de maïeuticien (sage-femme) si l'on souhaite introduire cette modalité dans le modèle. Le risque de rencontrer un tel problème de support commun sera d'autant plus fort que l'on cherche à examiner des modalités à un niveau très fin, et que la taille de l'échantillon est réduite.
 - si l'on cherche à effectuer la décomposition détaillée de l'écart inexpiqué, on a besoin des deux jeux de valorisations : il faut dans ce cas que chacune des modalités soit connue par chacun des deux groupes.
- dans le cas d'une variable continue, la régression linéaire conduit à “extrapoler” pour des valeurs qui seraient en dehors du support commun aux deux groupes.

Certaines conditions fondamentales sont ainsi requises pour pouvoir intégrer des variables explicatives dans l'analyse :

- variables définies dans les deux groupes. Par exemple si l'on veut comparer des immigrés à des personnes nées en France, il est problématique d'introduire l'année d'arrivée en France.
- existence d'une variabilité dans chacun des groupes. Là encore si l'on considère immigrés et natifs français, on ne peut pas utiliser le pays de naissance dans la décomposition.

2.2 Le choix de la référence

Le choix des valorisations de référence est crucial, notamment pour bien interpréter les résultats de la décomposition. Dans le cas de l'analyse des inégalités entre une majorité et une minorité, un contrefactuel assez naturel consiste à retenir les valorisations du groupe majoritaire. Cela revient implicitement à considérer qu'en l'absence de discrimination salariale entre les deux groupes, tous les salariés seraient rémunérés à la façon dont l'est le groupe en majorité. Les résultats obtenus permettent de répondre à la question de l'existence et de l'ampleur d'une discrimination négative. Une extension sur ce sujet utilisant les méthodes

de décomposition pour mesurer des inégalités intra- et inter- entreprises est renvoyée en l’encadré 4. A l’inverse, en retenant les valorisations estimées dans la minorité, on interroge plutôt l’existence de discrimination positive. D’autres options introduites plus haut consistent à raisonner en référence à une moyenne pondérée de $\hat{\beta}_A$ et $\hat{\beta}_B$, ou bien à des coefficients estimés sur l’ensemble de la population avec inclusion d’une indicatrice d’appartenance à l’un des groupes. En procédant ainsi, on tient donc compte de possibles effets d’équilibre. Cela peut par exemple être pertinent pour étudier des inégalités de sexe : dans un marché du travail où il n’existerait plus de discrimination selon le sexe, les salariés ne seraient peut-être pas payés au même niveau que les hommes le sont actuellement. Dans ce cas, il peut être utile de considérer un autre point de comparaison que les valorisations des caractéristiques observées chez les hommes, par exemple les valorisations moyennes sur l’ensemble de la population.

Encadré 4 : Une extension des méthodes de décomposition aux écarts d’effets entreprises

Les méthodes de décomposition peuvent être utilisées dans des modèles à effets fixes. C'est particulièrement utile en économie du travail où l'on s'attend à ce que les caractéristiques inobservables, à la fois des salariés et des entreprises, jouent un rôle essentiel dans les inégalités ([Abowd et al., 1999; Lentz and Mortensen, 2010](#)).

[Card et al. \(2016\)](#) s’inspirent des décompositions à la Oaxaca-Blinder pour séparer ce qui, dans l’influence des entreprises sur les inégalités salariales hommes-femmes, provient de la ségrégation des hommes et des femmes dans certaines entreprises et ce qui provient du fait qu’une même entreprise ne rémunère pas de la même manière ses salariés hommes et ses salariées femmes, même si leurs caractéristiques individuelles (compétences) sont identiques. Pour cela, ils proposent un modèle à doubles effets fixes dans l'esprit d'[Abowd et al. \(1999\)](#), pour lequel deux effets fixes sont associés à chaque entreprise, le premier représentant la “prime” que cette entreprise verse à ses salariés hommes et le deuxième celle qu’elle verse à ses salariés femmes. Ces “primes” définissent comment le partage de la richesse se fait au sein de chaque entreprise indépendamment des compétences individuelles des salariés.

Soit $w_{it}^{G(i)}$ le (log) salaire d'un individu à la date t , de sexe $G(i) = g \in \{F, M\}$ et travaillant à la date t dans l'entreprise $J(i, t)$:

$$w_{it}^{G(i)} = \alpha_i + X'_{it}\beta^{G(i)} + \psi_{J(i,t)}^{G(i)} + r_{it}, \quad (7)$$

avec r_{it} composé d'un terme d'erreur individuel et des éléments variant dans le temps du

surplus de l'entreprise.

Une telle écriture décompose le salaire en fonction d'un effet fixe individuel α_i , d'un effet entreprise pour les hommes et pour les femmes $\psi_{J(i,t)}^{G(i)}$, et de covariables aux rendements spécifiques pour les hommes et pour les femmes. Avec $\psi_{J(i,t)}^g$ l'effet fixe spécifique pour l'entreprise $J(i,t)$ pour le sexe g , on peut réécrire l'écart moyen entre effets entreprises moyens des hommes et des femmes de la manière suivante :

$$\begin{aligned} \mathbb{E} [\psi_{J(i,t)}^M | g = M] - \mathbb{E} [\psi_{J(i,t)}^F | g = F] &= \underbrace{\mathbb{E} [\psi_{J(i,t)}^M - \psi_{J(i,t)}^F | g = M]}_{\text{Effet } bargaining} \\ &\quad + \underbrace{\mathbb{E} [\psi_{J(i,t)}^F | g = M] - \mathbb{E} [\psi_{J(i,t)}^F | g = F]}_{\text{Effet } sorting} \end{aligned}$$

Le premier terme de cette décomposition correspond à un écart intra-entreprise : c'est la différence d'effet fixe entreprise moyen chez les hommes et chez les femmes, *si les femmes travaillaient dans les mêmes entreprises que les hommes*, soit la différence de captation du surplus de l'entreprise entre collègues masculins et féminins. Le second élément de la décomposition correspond à un écart inter-entreprise : c'est la différence entre l'effet entreprise moyen pour les femmes *si elles travaillaient dans les mêmes entreprises que les hommes*, et leur véritable effet entreprise moyen, étant donné leur répartition dans les entreprises. Cela correspond donc à la pénalité salariale liée au fait que les femmes travaillent pour des employeurs qui ne rémunèrent pas leurs salariés comme les employeurs des hommes.

La méthode de décomposition est ici utilisée comme un outil de mesure des composantes d'inégalités intra-entreprises et inter-entreprises entre hommes et femmes. Elle repose cependant sur des hypothèses supplémentaires, notamment de mobilité exogène des salariés entre employeurs.

2.3 La validité de la décomposition détaillée

2.3.1 Une hypothèse plus forte pour l'identification de la décomposition détaillée

L'hypothèse d'indépendance conditionnelle est nécessaire afin d'interpréter l'écart inexpliqué comme un effet causal de l'appartenance au groupe, et donc pour l'interprétation de la partition entre observé et inobservé (décomposition agrégée).

La décomposition détaillée demande toutefois de formuler des hypothèses supplémentaires, afin d'identifier le rôle des $(X_k)_{k=1\dots K}$ à la fois dans $\hat{\Delta}_S$ et dans $\hat{\Delta}_X$. Si l'on veut pouvoir at-

tribuer une part donnée de l'écart à une covariable X_k , il est nécessaire d'estimer l'effet causal de cette covariable en particulier sur la variable d'intérêt Y. On revient alors à l'hypothèse nécessaire à l'estimation sans biais des coefficients dans les équations de salaire initiales : l'hypothèse d'espérance conditionnelle nulle. Il ne suffit donc plus de supposer que les corrélations entre caractéristiques observables et inobservées sont les mêmes dans les deux groupes (cf. hypothèse d'indépendance conditionnelle), il est nécessaire de postuler qu'elles sont nulles.

2.3.2 Le problème de la modalité omise dans la décomposition détaillée de l'écart inexpliqué

On a évoqué précédemment la possibilité, comme pour l'effet de composition, de détailler terme à terme les contributions de chaque variable à l'écart inexpliqué $\hat{\Delta}_S$:

$$\hat{\Delta}_S = \sum_{k=0}^K \hat{\Delta}_{S_k},$$

où pour chaque variable explicative X_k dont la constante, $\hat{\Delta}_{S_k}$ correspond à sa contribution à l'écart inexpliqué, autrement dit :

$$\hat{\Delta}_{S_k}^\nu = \overline{X_{Ak}} (\hat{\beta}_{Bk} - \hat{\beta}_{Ak}), \quad k \in 0 \dots K.$$

La décomposition détaillée de l'écart inexpliquée peut être difficile à interpréter, notamment lorsque certaines caractéristiques X introduites dans la décomposition sont catégorielles. En effet, les composantes de la part inexpliquée des inégalités seront dépendantes de la catégorie de référence omise dans la régression : pour une variable X_k , les parts attribuées à β_0 et à β_k varient. Cette difficulté peut aussi apparaître pour une variable continue dont le zéro n'aurait pas d'interprétation naturelle. Il n'existe pas de solution générale au problème : un équilibre entre interprétabilité et comparabilité doit être trouvé.

Considérons un exemple où l'on décrit les différences de salaires moyens hommes-femmes en fonction d'une simple indicatrice de diplôme : les individus sont diplômés (D) ou non-diplômés (ND), et l'on cherche à effectuer la décomposition des écarts de salaires en fonction de cette seule variable explicative. On examinera la décomposition selon que l'on considère comme modalité de référence, dans l'équation de salaire, le fait d'être non-diplômé, ou bien le fait d'être diplômé.

Dans un premier temps, on considère comme modalité de référence le fait d'être non-diplômé. Ecrivons la décomposition (1) découlant de l'équation de salaire correspondante $Y_i = \beta_0 + \beta_1 D_i$:

$$\overline{Y^B} - \overline{Y^A} = \underbrace{\hat{\beta}_1^B(\overline{D^B} - \overline{D^A})}_{\text{Effet de composition}} + \underbrace{(\hat{\beta}_0^B - \hat{\beta}_0^A) + (\hat{\beta}_1^B - \hat{\beta}_1^A)\overline{D^A}}_{\text{Ecart inexpliqué}}.$$

en notant $\overline{D^B}$ la proportion de diplômés parmi les hommes, etc. Or on a dans ce cas $\hat{\beta}_0^g = \overline{Y}_{ND}^g$ et $\hat{\beta}_1^g = \overline{Y}_D^g - \overline{Y}_{ND}^g$, avec $g = A, B$: l'estimateur de β_0 dans le groupe g est simplement le salaire moyen parmi les non-diplômés du groupe g , et l'estimateur de β_1 est l'écart de salaire moyen entre diplômés et non-diplômés dans le groupe g , qui mesure le rendement du diplôme dans ce groupe. Dans ce cas, la décomposition détaillée de l'écart inexpliqué fait donc intervenir :

- Une part associée à la constante : $\hat{\beta}_0^B - \hat{\beta}_0^A = \overline{Y}_{ND}^B - \overline{Y}_{ND}^A$,
 - Une part associée à la variable de diplôme :
- $$(\hat{\beta}_1^B - \hat{\beta}_1^A)\overline{D^A} = [(\overline{Y}_D^B - \overline{Y}_{ND}^B) - (\overline{Y}_D^A - \overline{Y}_{ND}^A)]\overline{D^A}.$$

Le tableau 1 ci-dessous compare ces grandeurs avec celles obtenues dans le cas où l'on considère les diplômés comme modalité de référence, auquel cas l'équation de salaire s'écrit $Y_i = \alpha_0 + \alpha_1 ND_i$ et la décomposition correspondante devient :

$$\overline{Y^B} - \overline{Y^A} = \underbrace{\hat{\alpha}_1^B(\overline{ND^B} - \overline{ND^A})}_{\text{Effet de composition}} + \underbrace{(\hat{\alpha}_0^B - \hat{\alpha}_0^A) + (\hat{\alpha}_1^B - \hat{\alpha}_1^A)\overline{ND^A}}_{\text{Ecart inexpliqué}}.$$

TABLE 1 – Décomposition en fonction de la modalité de référence choisie

	Référence = Non-diplômés	Référence = Diplômés
Equation de salaire	$Y_i = \beta_0 + \beta_1 D_i$	$Y_i = \alpha_0 + \alpha_1 ND_i$
Estimateurs associés à la constante	$\hat{\beta}_0^g = \overline{Y}_{ND}^g$	$\hat{\alpha}_0^g = \overline{Y}_D^g$
Estimateurs associés à l'explicative	$\hat{\beta}_1^g = \overline{Y}_D^g - \overline{Y}_{ND}^g$	$\hat{\alpha}_1^g = \overline{Y}_{ND}^g - \overline{Y}_D^g$
Effet de composition	$(\overline{Y}_D^B - \overline{Y}_{ND}^B)(\overline{D^B} - \overline{D^A})$	$(\overline{Y}_{ND}^B - \overline{Y}_D^B)(\overline{ND^B} - \overline{ND^A})$
Ecart inexpliqué associé à la constante	$\overline{Y}_{ND}^B - \overline{Y}_{ND}^A$	$\overline{Y}_D^B - \overline{Y}_D^A$
Ecart inexpliqué associé à la variable explicative	$[(\overline{Y}_D^B - \overline{Y}_{ND}^B) - (\overline{Y}_D^A - \overline{Y}_{ND}^A)]x\overline{D^A}$	$[(\overline{Y}_{ND}^B - \overline{Y}_D^B) - (\overline{Y}_{ND}^A - \overline{Y}_D^A)]x\overline{ND^A}$

La répartition générale entre effet de composition et écart inexpliqué est, conformément à ce qui est attendu, identique, quelle que soit la modalité de référence considérée (on a en effet $\overline{D^g} = 1 - \overline{ND^g}$). Seule change la décomposition détaillée de l'écart inexpliqué. Notamment, la part inexpliquée associée à la variable explicative prend nécessairement des signes opposés selon la modalité de référence considérée : $[(\overline{Y}_D^B - \overline{Y}_{ND}^B) - (\overline{Y}_D^A - \overline{Y}_{ND}^A)] = -[(\overline{Y}_{ND}^B - \overline{Y}_D^B) - (\overline{Y}_{ND}^A - \overline{Y}_D^A)]$, or les proportions $\overline{D^A}$ et $\overline{ND^A}$ sont toutes deux positives.

Supposons que le rendement associé au fait d'être diplômé est plus fort chez les hommes que chez les femmes ($\overline{Y}_D^B - \overline{Y}_{ND}^B > \overline{Y}_D^A - \overline{Y}_{ND}^A$), comme décrit dans la première colonne du tableau

2. En prenant comme modalité de référence les non-diplômés, la part inexpliquée associée à la variable de diplôme sera positive. En effet, le fait de passer de non-diplômé à diplômé vient dans ce cas creuser l'écart salarial entre hommes et femmes (et y contribue donc positivement). Au contraire, si l'on prend les diplômés comme modalité de référence, le passage de la modalité "diplômés" à la modalité "non-diplômés" viendra ici diminuer les écarts salariaux hommes-femmes, d'où une contribution négative. Cet exemple simple permet de souligner la forte dépendance de la décomposition détaillée de l'écart inexpliqué à la modalité de référence choisie. Dans un tel cas, le signe de la contribution à l'inexpliqué de la variable X en dépend directement.

TABLE 2 – Signe de la contribution de la variable X à l'inexpliqué en fonction des rendements relatifs et de la modalité de référence choisie

	Référence = Non-diplômés	Référence = Diplômés
Rendement du diplôme plus élevé chez les hommes $\bar{Y}_D^B - \bar{Y}_{ND}^B > \bar{Y}_D^A - \bar{Y}_{ND}^A$	Contribution de X à l'inexpliqué > 0	Contribution de X à l'inexpliqué < 0
Rendement du diplôme plus élevé chez les femmes $\bar{Y}_D^B - \bar{Y}_{ND}^B < \bar{Y}_D^A - \bar{Y}_{ND}^A$	Contribution de X à l'inexpliqué < 0	Contribution de X à l'inexpliqué > 0

De façon plus générale, quelles que soient les variables explicatives incluses dans l'analyse, les résultats de la décomposition détaillée de l'inexpliqué sont modifiés par un changement de modalité de référence, avec un transfert entre la contribution à l'inexpliqué de la constante et celle de la variable X .

- La part inexpliquée associée à la constante représente l'écart salarial qui existerait si tous les individus se trouvaient dans cette situation de référence.
- Puis, pour une variable explicative X donnée, la part inexpliquée qui lui est attribuée traduit la façon dont le fait de s'éloigner de la modalité de référence pour aller vers les caractéristiques effectives de la population vient creuser, ou au contraire diminuer, l'écart de référence.

Ainsi pour [Jones and Kelley \(1984\)](#), l'interprétation détaillée de la décomposition n'a de sens que pour des variables catégorielles ayant une modalité de référence naturelle⁶. Cette question d'identification a aussi été discutée par [Oaxaca and Ransom \(1999\)](#); [Gardeazabal and Ugidos \(2004\)](#) ou encore [Yun \(2005, 2008\)](#), qui proposent de procéder à une normalisation des coefficients pour éliminer de la constante l'effet de la modalité omise, par exemple en

6. Dans le cas où l'on suit dans le temps l'évolution d'une décomposition détaillée, cette modalité de référence "naturelle" doit avoir le même sens au cours de la période. Ce point pose problème pour de nombreuses variables, par exemple celles de diplôme, qui peuvent être valorisées de façon assez différente suivant les années.

constraining à zéro la somme des coefficients de la variable catégorielle. Yun (2005) propose lui de considérer la moyenne des contributions obtenues pour chaque modalité de référence possible associée à chaque variable catégorielle du modèle.

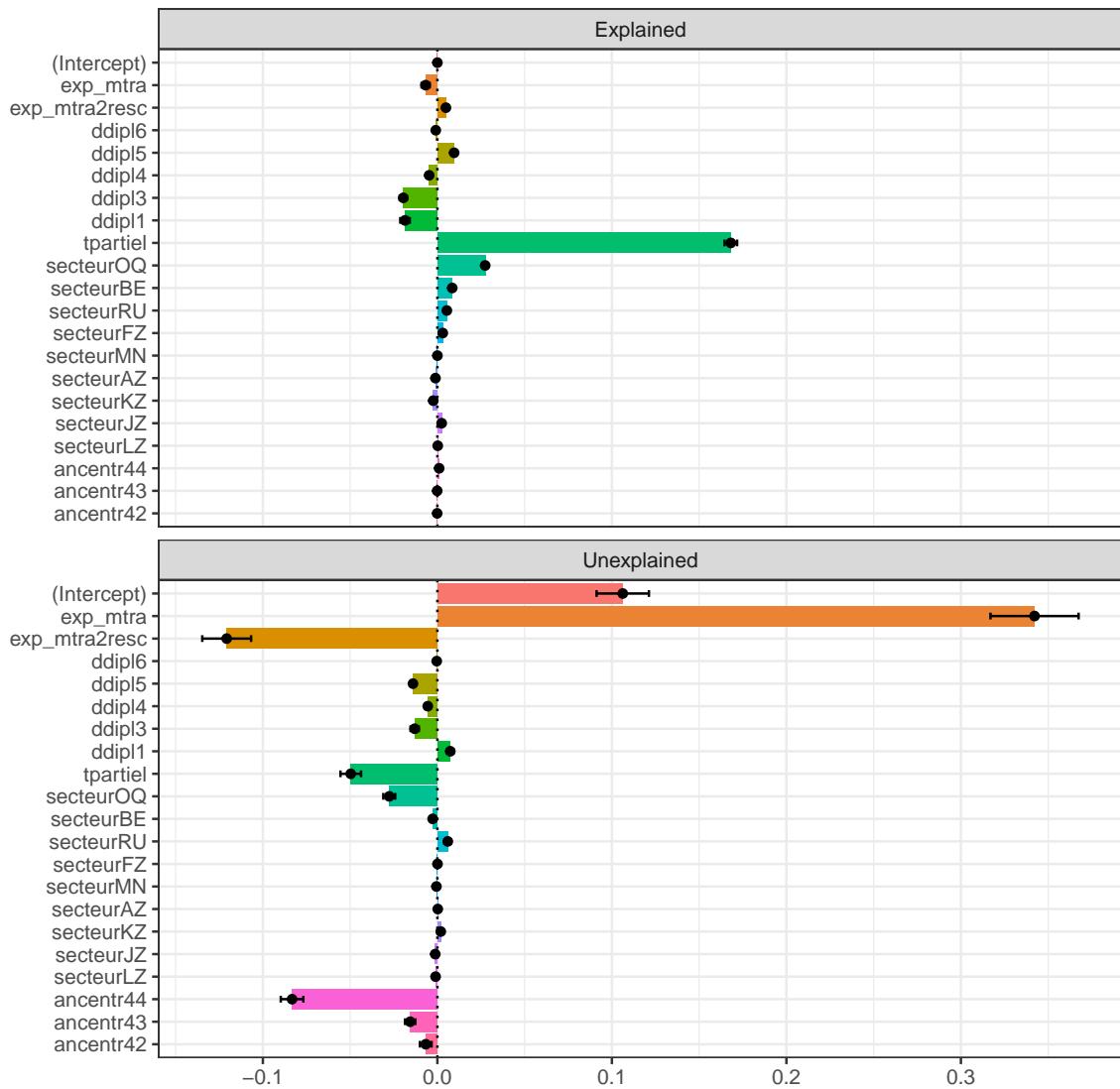
Application 3 : Gérer le problème de modalité omise avec la fonction `oaxaca`

La solution proposée par Yun peut être utilisée simplement dans le package `Oaxaca` en renseignant les modalités (sauf une) d'une variable catégorielle de la façon suivante, dans l'appel de la fonction `oaxaca` :

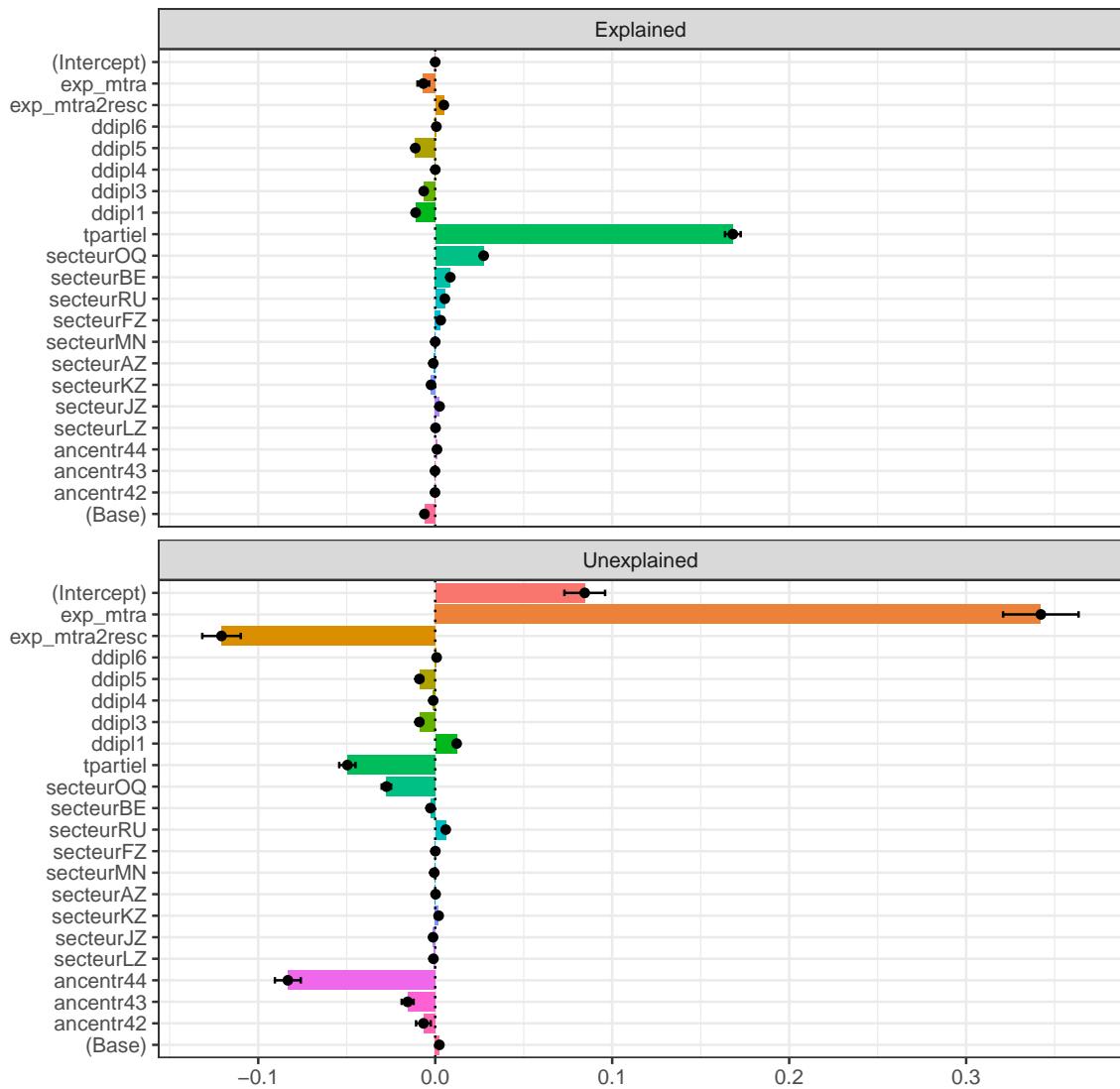
```
#Après avoir renseigné le groupe, on renseigne les modalités de la variable
#à laquelle on veut appliquer la méthode de Yun
results_mod_omise <- oaxaca(formula = logsal ~ exp_mtra + exp_mtra2resc
+ ddip16 + ddip15 + ddip14 + ddip13 + ddip11
+ tpartiel + secteurOQ + secteurBE + secteurRU
+ secteurFZ + secteurMN + secteurAZ + secteurKZ
+ secteurJZ + secteurLZ + ancentr44 + ancentr43
+ ancentr42 | sex | ddip16 + ddip15 + ddip14
+ ddip13 + ddip11, data = data, R=10)
```

Cela modifie quelque peu la décomposition de l'écart expliqué :

```
#La figure de la décomposition détaillée avec group.weight= 1
plot(results, decomposition = "twofold", group.weight = 1)
```



```
plot(results_mod_omise, decomposition = "twofold", group.weight = 1)
```



Malgré cette correction, la difficulté de l'interprétation n'est pas complètement levée, ce qui rend l'utilisation de la décomposition détaillée de la part inexpliquée délicate.

3 Variable d'intérêt dichotomique et écart entre proportions

On a jusqu'ici étudié l'écart entre deux groupes selon une variable d'intérêt continue, par exemple le salaire. De nombreuses variables sont toutefois dichotomiques : si l'on considère par exemple le fait d'être au chômage, on sera amené à décomposer l'écart de taux de chômage entre les deux sous-populations considérées.

3.1 La décomposition d'Oaxaca-Blinder pour une variable dichotomique

La décomposition présentée jusqu'ici pour le cas d'une variable Y continue peut être directement transposée à une variable d'intérêt dichotomique, dès lors qu'il est raisonnable de modéliser celle-ci par une régression linéaire. Cela présente l'avantage de la simplicité. Par ailleurs, la décomposition détaillée variable par variable est immédiate (voir section 2.3) ce qui n'est plus le cas dès lors qu'on s'éloigne du modèle linéaire.

Lorsque l'approximation linéaire s'avère problématique (événements rares) ou lorsque les différences de caractéristiques observables entre les deux groupes sont marquées (contrefactuel dénué de sens), on peut se tourner vers une modélisation non linéaire de type logit ou probit.

3.2 Décomposition de Fairlie

Fairlie (2005) propose de décomposer un écart moyen d'une variable dichotomique, noté $\overline{Y_B} - \overline{Y_A}$ ⁷. La contrepartie empirique de cet écart est $\frac{1}{N_B} \sum_{i \in B} Y_i - \frac{1}{N_A} \sum_{i \in A} Y_i$ (avec N_A et N_B les effectifs dans les deux groupes). Pour la décomposer, on utilise un contrefactuel fondé sur l'estimation d'un modèle probit ou logit. Les trois étapes de la méthode de Fairlie sont les suivantes, en prenant le groupe B comme référence et comme variable d'intérêt le fait d'être au chômage (qui vaut 0 ou 1) :

- (1) On modélise la probabilité d'être au chômage au sein du groupe B par un modèle logit ou probit : $P_B(Y = 1|X) = F(X\beta_B)$, où $F(\cdot)$ est la fonction de répartition de la loi normale (modèle probit) ou de la loi logistique (modèle logit).
- (2) On calcule alors, pour chaque individu i du groupe A , sa probabilité prédictive d'être au chômage en appliquant le modèle précédent aux caractéristiques observables de i . Par exemple dans le cas d'un écart entre hommes et femmes, on calcule pour chaque femme la probabilité qu'elle soit au chômage si ses caractéristiques effectivement observées l'exposaient au chômage comme parmi les hommes. Pour une femme avec des caractéristiques X_i , on obtient donc $\hat{P}_B(Y_i = 1|X_i) = F(\hat{\beta}_{B0} + \sum_{k=1}^K X_{ik}\hat{\beta}_{Bk})$.
- (3) On calcule la moyenne de ces probabilités prédictives pour l'ensemble des individus du groupe A : $\frac{1}{N_A} \sum_{i \in A} \hat{P}_B(Y_i = 1|X_i)$, soit un estimateur de $P_B(Y = 1 | X_A)$ - la probabilité

7. Ou de façon équivalente, $P_B(Y = 1 | X_B)$ et $P_A(Y = 1 | X_A)$.

d'être au chômage pour les femmes, si leurs caractéristiques étaient valorisées comme chez les hommes (pour plus de précision sur la notation, voir section 4).

Comme pour la décomposition d'Oaxaca-Blinder, on obtient ainsi un contrefactuel répondant à la question suivante : quel serait le taux de chômage des individus du groupe A si leurs caractéristiques étaient valorisées de la même manière que pour le groupe B ? Ces étapes permettent d'obtenir la décomposition agrégée, qui s'écrit :

$$\overline{Y_B} - \overline{Y_A} = \underbrace{\frac{1}{N_B} \sum_{i \in B} Y_i - \frac{1}{N_A} \sum_{i \in A} \hat{P}_B(Y_i = 1 | X_i)}_{\text{Effet de composition (lié aux } X\text{)}} + \underbrace{\frac{1}{N_A} \sum_{i \in A} \hat{P}_B(Y_i = 1 | X_i) - \frac{1}{N_A} \sum_{i \in A} Y_i}_{\text{Ecart inexpliqué (à } X\text{ donnés)}}$$

L'effet de composition correspond à une variation seulement des caractéristiques, à modèle identique ; inversement, l'écart inexpliqué est calculé à caractéristiques données, comme une différence de valorisation des X entre les deux groupes. La version détaillée de la décomposition de Fairlie est plus délicate à établir que dans le cadre linéaire d'Oaxaca-Blinder - sa mise en œuvre est détaillée dans l'encadré 5.

Encadré 5 : La décomposition détaillée au-delà du cas linéaire

Si la décomposition agrégée du modèle de Fairlie est facile à obtenir, la version détaillée est nettement plus difficile à calculer. En effet, dans la construction du contrefactuel pour la décomposition agrégée $F(\hat{\beta}_{B0} + \sum_{k=1}^K X_{ik}\hat{\beta}_{Bk})$ implicitement on "remplace" les X des hommes par les X des femmes. Pour détailler l'effet de composition il faut donc remplacer successivement chaque X_k des hommes par les X_k des femmes. Prenons l'exemple d'un cas à trois variables explicatives. Pour calculer la contribution de X_3 à l'effet de composition, il faut prendre la différence entre :

$$F(\hat{\beta}_{0B} + X_{1B}\hat{\beta}_{1B} + X_{2B}\hat{\beta}_{2B} + X_{3B}\hat{\beta}_{3B})$$

$$\text{et } F(\hat{\beta}_{0B} + X_{1B}\hat{\beta}_{1B} + X_{2B}\hat{\beta}_{2B} + \textcolor{red}{X_{3A}}\hat{\beta}_{3B}).$$

Pour un homme à X_1 et X_2 donnés, par quelle valeur remplace-t-on son X_3 ? On voudrait une sorte d'appariement entre les hommes et les femmes pour tenir compte de la structure de corrélation entre les variables X_1 , X_2 et X_3 . Pour ce faire, la solution proposée par Fairlie (2005) consiste en quatre étapes, par exemple si la variable d'intérêt est la probabilité d'être au chômage pour les actifs :

- (1) On tire un échantillon dans la population des hommes, de même taille que celle

des femmes.

- (2) Au sein de chaque échantillon, on classe les individus selon leur propension à être au chômage (probabilité prédictive d'être au chômage selon un modèle $P(Y = 1 | X)$ commun aux deux groupes).
- (3) On apparie l'homme ayant la plus forte propension à être au chômage à la femme ayant la plus forte propension à être au chômage, etc.
- (4) Pour un homme donné, on remplace la valeur de X_k considérée par celle prise par l'individu femme apparié. La moyenne des probabilités prédictives ainsi calculées est comparée à la probabilité initiale chez les hommes, pour obtenir la contribution de X_k à l'effet de composition.

On reproduit ces étapes un grand nombre de fois, en tirant à chaque fois un nouvel échantillon.

Cette procédure est très intensive en calcul, et elle ne résout pas le problème de l'impossibilité d'avoir une décomposition détaillée additive et non sensible à l'ordre.

Deux propriétés sont en effet particulièrement souhaitables pour la décomposition détaillée : l'additivité et l'invariance à l'ordre. On entend par additivité le fait que les contributions à l'expliqué de chaque variable se somment bien en la part expliquée totale, autrement dit : $\Delta_X = \sum_{k=1}^K \Delta_{X_k}$. Cette propriété est satisfaite dans le cadre linéaire simple mais elle n'est pas forcément garantie hors de celui-ci, par exemple dans les approches présentées dans la section 4.

Elle sera généralement satisfaite dans une procédure séquentielle consistant à remplacer la distribution de X_1 puis de X_2 etc., jusqu'à ce que la distribution des X ait été entièrement remplacée. Mais comme l'impact du changement d'une variable donnée dépend généralement de la distribution des autres variables, on peut alors avoir une décomposition détaillée qui dépend de l'ordre dans lequel on la réalise. L'invariance à l'ordre n'est donc pas respectée.

Pour une approche plus simple de la décomposition détaillée, on pourra préférer l'approximation de [Yun \(2004\)](#) : celle-ci consiste à repartir de l'effet de composition agrégé estimé selon Fairlie et à le désagréger selon un système de poids attribuant à chaque variable le poids $\frac{(\bar{X}_{kB} - \bar{X}_{kA})\hat{\beta}_{kB}}{\sum_k (\bar{X}_{kB} - \bar{X}_{kA})\hat{\beta}_{kB}}$ avec les $\hat{\beta}_{kB}$ estimés par logit ou probit dans la première étape de la décomposition de Fairlie. Cette méthode peut cependant poser problème lorsque les

prédictions se prêtent mal à l'approximation linéaire, typiquement quand elles sont hors de l'intervalle entre 0 et 1 et/ou lorsqu'il existe de fortes différences dans les X entre les deux groupes.

Application 4 : Décomposition de Fairlie

On propose d'appliquer la méthode de Fairlie aux écarts de probabilité d'accès à des postes d'encadrement (variable *encadr* valant 1 ou 0) entre hommes et femmes, toujours à partir des données de l'enquête Emploi en continu. On cherche à expliquer ces écarts par des différences de caractéristiques. Comme en section 1.2 on utilise comme variables explicatives le niveau d'études *ddipl* et une variable d'ancienneté (*anctr4*). On introduit également dans le modèle l'expérience potentielle de l'individu sur le marché du travail (notée *exp*) et son carré (*exp2*), ainsi que la quotité de travail en 6 modalités, *tppred* (1 pour moins d'un mi-temps, 2 pour un mi-temps, 3 pour 50 à 80 %, 4 pour 80 %, 5 pour plus de 80 % et 6 pour un temps plein).

```
# Quelle proportion d'hommes (groupe B) occupe des fonctions d'encadrement ?
prop.B<-mean(base_empl$encadr [base_empl$sex=="1"])
prop.B

## [1] 0.251

# Et parmi les femmes ?
prop.A<-mean(base_empl$encadr [base_empl$sex=="2"])
prop.A

## [1] 0.136

# Ecart hommes-femmes
prop.B-prop.A

## [1] 0.115
```

On estime pour chaque sexe un modèle logit d'accès à des fonctions d'encadrement :

```
#On estime un modèle logit dans le groupe B (hommes)
logitB<-glm(encadr ~ as.factor(anctr4) + as.factor(tppred)
              + exp+ exp2 + as.factor(ddipl), family=binomial (link='logit'),
              data=base_empl[base_empl$sex=="1",])
```

Mettons que l'on souhaite connaître la probabilité contrefactuelle d'encadrement parmi les hommes, s'ils avaient les caractéristiques des femmes (ou dit autrement, la probabilité contrefactuelle d'encadrement des femmes si leurs caractéristiques étaient valorisées comme celles

des hommes). Il suffit pour cela de calculer les prédictions individuelles selon le modèle des hommes, puis de calculer la moyenne de ces probabilités prédites parmi les femmes.

```
#On prédit la proportion à partir des estimateurs des hommes (B)
#appliqués aux caractéristiques des femmes (A)
base_empl$pB<-predict(logitB, base_empl,type='response')
prop.cA<-mean(base_empl$pB[base_empl$sexe=="2"],na.rm=TRUE)
prop.cA

## [1] 0.23

#Ecart expliqué
expl<- prop.B - prop.cA
expl

## [1] 0.0209

#Ecart inexpliqué
inexpl<-prop.cA - prop.A
inexpl

## [1] 0.0941
```

Ce contrefactuel nous permet de mesurer un écart expliqué de 0.021 et un écart inexpliqué de 0.094 : l'écart expliqué se calcule comme l'écart entre la probabilité contrefactuelle et la proportion d'encadrement mesurée parmi les hommes, car celui-ci provient bien uniquement de différences de caractéristiques. Au contraire pour l'écart inexpliqué, on raisonne à caractéristiques données (celles des femmes).

On peut effectuer la même décomposition en repartant, non pas d'un logit, mais d'un modèle de probabilité linéaire (ie. simples MCO). A noter : ce code permet d'obtenir rapidement la décomposition d'Oaxaca-Blinder agrégée, pour une variable continue également.

```
#On estime un modèle MCO de la proba d'encadrement chez les hommes (B)
lpmB<-lm(encadr ~ as.factor(ancentr4) + as.factor(tppred)
           + exp+ exp2 + as.factor(ddipl), data=base_empl[base_empl$sexe=="1",])

#On prédit un taux d'encadrement à partir des estimateurs
base_empl$pB<-predict(lpmB, base_empl)
```

```

#Le contrefactuel est la prediction moyenne chez les femmes (A)
mean(base_empl$pB[base_empl$sexe=="2"] ,na.rm=TRUE)

[1] 0.238

#d'où l'écart expliqué (proba des hommes - proba contrefactuelle)
mean(base_empl$encadr[base_empl$sexe=="1"] ,na.rm=TRUE) -
mean(base_empl$pB[base_empl$sexe=="2"] ,na.rm=TRUE)

[1] 0.0125

#et inexplicable (proba contrefactuelle - proba des femmes)
mean(base_empl$pB[base_empl$sexe=="2"] ,na.rm=TRUE) -
mean(base_empl$encadr[base_empl$sexe=="2"] ,na.rm=TRUE)

[1] 0.102

```

On trouve des ordres de grandeur comparables pour l'écart expliqué (0.021 avec la méthode de Fairlie vs 0.012 avec le modèle de probabilité linéaire) et pour l'écart inexplicable (0.094 vs 0.102). Toutefois, dans un nombre de cas non-négligeable, la probabilité prédite individuelle d'être au chômage est en dehors de [0,1] avec ce modèle linéaire.

```

#Décompte des prédictions hors de [0;1]
table(base_empl$pB<0 | base_empl$pB>1)

##
##  FALSE    TRUE
## 190949    7910

```

Pour la décomposition détaillée, on va procéder à l'approximation de Yun (voir encadré 5). Pour cela, on a besoin de récupérer le vecteur des $(\bar{X}_B^k - \bar{X}_A^k)\hat{\beta}_B^k$, $k = 1\dots K$, qu'on nomme ci-dessous `delta.X.beta`.

```

#On récupère le vecteur des coefficients chez les hommes,
#ainsi que le vecteur des X moyens dans les deux groupes
coeffs.B<-logitB$coefficients

X.B <- model.matrix(~ as.factor(anctr4) + as.factor(tppred) + exp + exp2 +
                     as.factor(ddipl), data=base_empl[base_empl$sexe=="1",])
X.moy.B<-apply(X.B,2,mean)

```

```

X.A <- model.matrix(~ as.factor(ancentr4) + as.factor(tppred) + exp + exp2 +
                     as.factor(ddipl), data=base_empl[base_empl$sex=="2",])
X.moy.A<-apply(X.A,2,mean)

#On calcule alors delta.X.beta
delta.X.beta<-(X.moy.B- X.moy.A)*coeffs.B

# Part liée à l'ancienneté :
# 4 modalités qui correspondent aux éléments 2 à 5 de delta.X.beta
part.ancentr<-expl*sum(delta.X.beta[grep("ancentr", names(delta.X.beta))])/
  sum(delta.X.beta)
part.ancentr

## [1] -0.00335

#Idem pour temps partiel (6), expérience potentielle (7) et diplôme (8 à 12)
part.quotite<-expl*sum(delta.X.beta[grep("tppred", names(delta.X.beta))])/
  sum(delta.X.beta)
part.quotite

## [1] 0.0358

part.exp<-expl*sum(delta.X.beta[grep("exp", names(delta.X.beta))])/
  sum(delta.X.beta)
part.exp

## [1] -0.00101

part.ddipl<-expl*sum(delta.X.beta[grep("ddipl", names(delta.X.beta))])/
  sum(delta.X.beta)
part.ddipl

## [1] -0.0106

#On retrouve bien en sommant ces 4 composantes l'écart expliqué total
part.ancentr+part.quotite+part.exp+part.ddipl

## [1] 0.0209

expl

## [1] 0.0209

```

On peut comparer ces résultats à ce qu'on obtient en utilisant un modèle d'Oaxaca-Blinder sur la probabilité linéaire :

```

library(oaxaca)
#les hommes sont à 0, les femmes à 1
base_empl$sexenum <- as.numeric(base_empl$sexe)-1
results <- oaxaca(formula = encadr ~ as.factor(ancentr4) + as.factor(tppred)
+ exp + exp2 + as.factor(ddipl) | sexenum , data = base_empl, R=10)

results$twofold$overall[2,2]

## coef(explained)
##          0.0125

#La part expliquée liée à l'ancienneté (on somme les contributions
#des différentes modalités de la variable ancienneté)
sum(results$twofold$variables[[2]][ grep("ancentr4",
names(results$twofold$variables[[2]][,1])) ,2])

## [1] -0.00199

#La part expliquée liée à la quotité
sum(results$twofold$variables[[2]][ grep("tppred",
names(results$twofold$variables[[2]][,1])) ,2])

## [1] 0.0303

#La part expliquée liée à l'expérience
sum(results$twofold$variables[[2]][ grep("exp",
names(results$twofold$variables[[2]][,1])) ,2])

## [1] -0.00116

#La part expliquée liée au diplome
sum(results$twofold$variables[[2]][ grep("ddipl",
names(results$twofold$variables[[2]][,1])) ,2])

## [1] -0.0147

```

On retrouve un écart expliqué de 0.012 avec le modèle de probabilité linéaire, dont une contribution de l'ancienneté de -16 % (-16 % dans le modèle de Fairlie), de la quotité de travail

de 244 % (172 % dans le modèle de Fairlie), de l'expérience de -9 % (-22 % dans le modèle de Fairlie) et du diplôme de -118 % (-12 % dans le modèle de Fairlie). Autrement dit, la décomposition détaillée obtenue par l'approximation de Yun et celle obtenue avec la méthode d'Oaxaca-Blinder pour un modèle de probabilité linéaire donnent des résultats similaires dans cet exemple. La quotité de travail explique une majeure partie de l'écart de taux d'accès à l'encadrement entre hommes et femmes, tandis que le niveau d'études tend au contraire à diminuer cette différence.

4 La décomposition des inégalités au-delà de la moyenne

Lorsque la variable d'intérêt Y est continue, on la résume souvent par sa moyenne : on cherche alors à expliquer l'écart entre moyennes calculées pour chacun des deux groupes. On peut toutefois souhaiter aller “au-delà de la moyenne” et s'intéresser à des inégalités en certains endroits de la distribution de Y , ou plus généralement à d'autres statistiques que la moyenne : en termes de salaires par exemple, il peut exister un phénomène de type *plafond de verre* lorsqu'un des deux groupes ne parvient pas aux salaires les plus élevés. Dans ce cas, il sera plus pertinent de s'intéresser au sommet de la distribution des salaires, plutôt qu'au salaire moyen. De même, lorsqu'on effectue une comparaison intertemporelle ou internationale, c'est souvent à une statistique caractérisant les inégalités que l'on s'intéresse (par exemple écart interdécile, le coefficient de Gini, etc.), pour chaque période ou pour chaque pays, plutôt qu'à la seule moyenne.

Dans ce cas plus général, on va s'intéresser à l'écart entre la distribution de Y observée dans le groupe B , et celle observée dans le groupe A . Là encore, on introduit un objet contrefactuel, ici une distribution contrefactuelle, répondant par exemple à la question suivante : quelle serait la distribution des salaires chez les femmes (groupe A), si leurs caractéristiques étaient valorisées comme celles des hommes (groupe B) ? De façon analogue à la décomposition d'Oaxaca-Blinder ou de Fairlie, l'écart entre cette distribution contrefactuelle et la distribution des salaires dans le groupe B correspondra à l'effet de composition, tandis que l'écart avec la distribution des A sera inexpliqué.

La construction de cette distribution contrefactuelle est moins directe que dans les cas vus précédemment. Pour y parvenir, on va utiliser la notion de *distribution conditionnelle*, la fonction qui associe à un ensemble de caractéristiques X , la distribution des Y pour ces caractéristiques. Si par exemple on considère une unique variable explicative binaire X (par exemple, le fait d'être cadre ou non), la distribution des salaires conditionnelle au statut d'encadrement est notée $F_{Y|X}$. Cette distribution conditionnelle dans le groupe A s'écrit $F_{Y_A|X}$, et associe à $X = 1$ la distribution des salaires parmi les cadres du groupe A, et à $X = 0$ la distribution des salaires parmi les non-cadres du groupe A.

De façon générale, en considérant un ensemble de caractéristiques X plus vaste, on peut écrire la distribution des salaires effectivement observée dans le groupe A (la distribution *non-conditionnelle* F_{Y_A} , qu'on pourra également noter $F_{Y_A|X_A}$) comme la résultante de la distribution conditionnelle $F_{Y_A|X}$, correspondant à la structure de valorisation des caractéristiques des A, appliquée à la répartition des caractéristiques X dans le groupe A (c'est-à-dire, intégrée

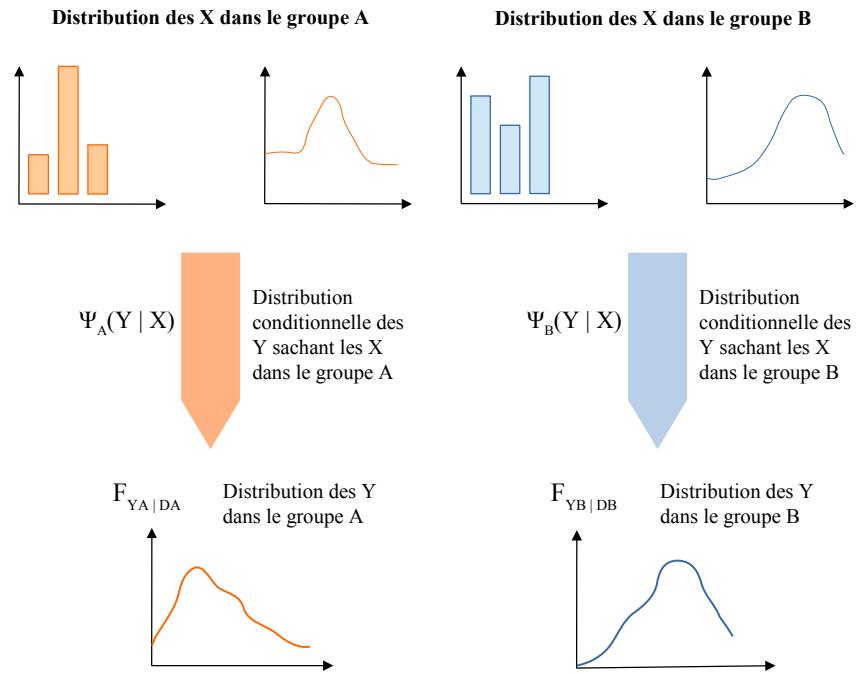
sur ces caractéristiques). On a ainsi⁸ :

$$F_{Y_A} (= F_{Y_A|X_A}) = \int F_{Y_A|X}(y|x)dF_{X_A}(x)$$

avec F_{X_A} la distribution des caractéristiques observables dans le groupe A .

La figure 2 met en avant le passage, pour chacun des groupes A et B , entre distributions des caractéristiques observables X (que celles-ci soient discrètes ou continues), distributions conditionnelles de Y qui valorisent ces caractéristiques, et distributions non-conditionnelles. L'écart entre les distributions observées F_{Y_A} et F_{Y_B} peut ainsi trouver deux sources : un écart entre distribution des caractéristiques observables F_{X_A} et F_{X_B} , ou un écart entre distributions conditionnelles $F_{Y_A|X}$ et $F_{Y_B|X}$, c'est-à-dire entre valorisations des X en termes de distributions de salaires.

Figure 2 – Distribution jointe de X et Y dans chaque groupe



La notion de distribution conditionnelle va aider à expliciter le contrefactuel qui permettra de dissocier effet de composition et écart inexpliqué. Par exemple, si on applique la distribution

8. Les notations proposées ici sont légèrement simplifiées par rapport à celles de Fortin et al. (2011). On ne reprend notamment pas l'indicatrice d'appartenance au groupe, $D_g, g = A, B$ et on indice directement les distributions en désignant le groupe concerné.

conditionnelle du groupe B aux caractéristiques des A, on obtient la distribution de salaires contrefactuelle suivante :

$$F_{Y_B|X_A} = \int F_{Y_B|X}(y|x) dF_{X_A}(x).$$

Cela correspond à la distribution qui s'appliquerait si les caractéristiques présentes dans le groupe A étaient valorisées comme chez les B. Supposons que l'on s'intéresse à une statistique ν de la distribution en particulier, par exemple le dernier décile : on souhaiterait décomposer l'écart entre le dernier décile de salaire dans le groupe B , et le dernier décile de salaire dans le groupe A . On peut décomposer l'écart de ν entre groupes B et A de la façon suivante :

$$\nu(F_{Y_B}) - \nu(F_{Y_A}) = \underbrace{[\nu(F_{Y_B|X_B}) - \nu(F_{Y_B|X_A})]}_{\text{Effet de composition}} + \underbrace{[\nu(F_{Y_B|X_A}) - \nu(F_{Y_A|X_A})]}_{\text{Écart inexpliqué}} \quad (8)$$

Le premier terme correspond à l'effet de composition : on voit en effet apparaître un écart lié aux caractéristiques observables (X_A vs. X_B), valorisées dans les deux cas par la même distribution conditionnelle $F_{Y_B|X}$. Pour le deuxième terme au contraire, on raisonne à caractéristiques données (X_A) : il s'agit de l'écart inexpliqué.

Les méthodes de décomposition des écarts entre distributions reposent ainsi sur la construction de la distribution contrefactuelle $F_{Y_B|X_A}$ ⁹. On peut distinguer deux façons de parvenir à celle-ci :

- On part de la distribution des salaires dans le groupe B ($F_{Y_B|X_B}$), mais on modifie la distribution des caractéristiques observables des individus de ce groupe de façon à ce qu'elle soit la même que dans le groupe A (on “remplace” ainsi F_{X_B} par F_{X_A}). Cela correspond notamment aux méthodes par repondération (DiNardo, Fortin, and Lemieux, 1996)¹⁰. Ce procédé est représenté dans la partie gauche de la figure 3, et présenté dans la section suivante.
- On estime directement la distribution conditionnelle du groupe B ($F_{Y_B|X}$), et on l'applique ensuite aux caractéristiques X du groupe A . Cela correspond aux méthodes d'estimation de la distribution conditionnelle (Chernozhukov et al., 2013; Machado and Mata, 2005). Ce procédé est représenté dans la partie droite de la figure 3 et présenté en section 4.3.

La méthode des régressions de quantiles non-conditionnels, présentée en section 4.2, diffère légèrement de ces approches en ce qu'elle ne s'appuie pas sur la construction d'une distribution

9. Comme dans le cas de la décomposition d'Oaxaca-Blinder, d'autres distributions contrefactuelles peuvent bien sûr être envisagées, en premier lieu $\nu(F_{Y_A|X_B})$.

10. D'autres méthodes pourraient être envisagées afin de rendre les deux sous-populations comparables en termes de caractéristiques observables, notamment les méthodes d'appariement (*matching*) sur score de propension ou sur caractéristiques exactes. Pour une application sur données françaises, voir par exemple Duvivier et al. (2016).

contrefactuelle.

4.1 La méthode de repondération

Afin de construire la distribution contrefactuelle $F_{Y_B|X_A}$ correspondant à la distribution des Y du groupe B , si celui-ci présentait les mêmes caractéristiques observables que celles du groupe A , DiNardo et al. (1996) proposent d'ajuster les poids des observations du groupe B afin de rendre leurs caractéristiques observables similaires à celles des individus du groupe A . Par exemple, si l'on souhaite décomposer l'écart entre les distributions de salaire des hommes et des femmes en contrôlant du statut de cadre, et que l'on suppose que les hommes accèdent plus souvent au statut de cadre que les femmes, on va repondérer à la baisse les observations des hommes exerçant des fonctions d'encadrement ; et à la hausse celles des hommes qui n'ont pas de fonction d'encadrement. À partir de la distribution des salaires pour les observations des hommes ainsi repondérées (qui correspond ici à la distribution contrefactuelle $F_{Y_H|X_F}$), on peut calculer très facilement n'importe quelle statistique ν et parvenir à la décomposition 8. L'étape de calcul des poids de repondération peut elle-même s'effectuer très aisément.

En effet, le facteur de repondération $\Psi(X)$ qui, appliqué à chaque observation du groupe B , permet de rendre la distribution des caractéristiques du groupe B similaire à celle du groupe A s'écrit (en notant $g = A, B$ la variable d'appartenance au groupe) :

$$\Psi_{DFL}(X) = \frac{P(X|g = A)}{P(X|g = B)} = \frac{P(g = A|X)}{P(g = B|X)} \cdot \frac{P(g = B)}{P(g = A)} = \frac{P(g = A|X)}{1 - P(g = A|X)} \cdot \frac{1 - P(g = A)}{P(g = A)}.$$

$P(g = A)$ correspond simplement à la proportion d'individus du groupe A dans la population. Afin d'obtenir une estimation de $P(g = A|X)$, on modélise la probabilité d'appartenir au groupe A , sur l'ensemble de l'échantillon, en fonction des caractéristiques observables X . L'estimation peut être faite par logit ou probit¹¹. Ce modèle fournit directement pour chaque individu de caractéristiques X la probabilité prédictive d'appartenir au groupe A , c'est-à-dire $\hat{P}(g = A|X)$. On calcule alors le facteur de repondération $\hat{\Psi}_{DFL}(X)$ de façon très simple :

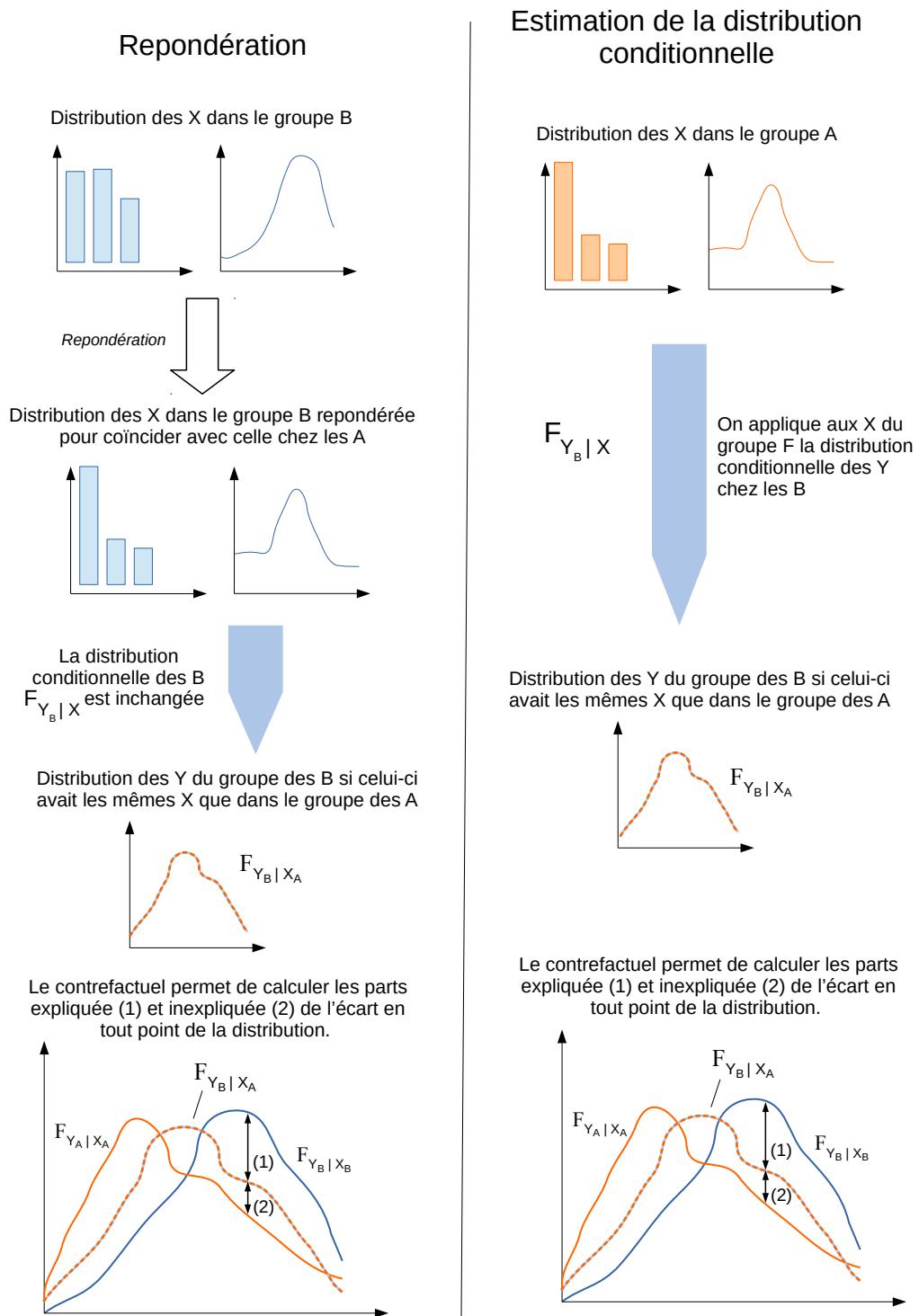
$$\hat{\Psi}_{DFL}(X) = \frac{\hat{P}(g = A|X)}{1 - \hat{P}(g = A|X)} \cdot \frac{1 - \hat{P}(g = A)}{\hat{P}(g = A)} \text{ 12.}$$

Bien que très simple à mettre en œuvre, cette méthode doit être utilisée avec précaution en cas de problème de support commun, car le facteur de repondération peut alors avoir un comportement erratique. Notamment, si $P(g = B|X) \rightarrow 0$ et $P(g = A|X) \rightarrow 1$, ce qui sera le

11. Hirano et al. (2003) proposent alternativement l'emploi d'un modèle non-paramétrique, permettant de tenir compte de façon plus flexible de la structure de corrélation entre les variables.

12. Le terme $\frac{1 - \hat{P}(g = A)}{\hat{P}(g = A)}$ est constant d'un individu à l'autre, l'inclure ou non dans la repondération ne modifie donc pas la distribution contrefactuelle obtenue.

Figure 3 – Comparaison des méthodes de décomposition au-delà de la moyenne : repondération ou estimation de la distribution conditionnelle



cas si une caractéristique est très rare au sein du groupe B relativement au groupe A , $\Psi(X)$ peut devenir très grand pour les individus B détenant cette caractéristique. Ces observations repondérées risquent alors de porter à elles seules toute la distribution contrefactuelle. Il est ainsi nécessaire de s'assurer, lorsque l'estimation du facteur de repondération pour chacun des B est effectuée, que celui-ci ne prend pas de valeur anormalement élevée ou faible. En pratique, on peut regarder la façon dont les poids après repondération sont distribués.

La méthode de repondération initialement proposée par DiNardo et al. (1996) permet d'isoler la participation à l'effet de composition d'une variable binaire¹³. Toutefois, la décomposition détaillée obtenue est non-additive, si l'on remplace pour chaque X_k la distribution au sein du groupe B par celle du groupe A , tout en conservant pour les autres explicatives la distribution des B . Si l'on procède plutôt de façon séquentielle en remplaçant successivement la distribution de X_1 , puis de X_2 , et ainsi de suite jusqu'à ce que la distribution de l'ensemble des X soit celle du groupe A , la décomposition détaillée obtenue est additive mais dépendante de l'ordre dans lequel on procède.

Application 5 : Décomposition par repondération

Considérons l'écart entre la distribution des salaires des descendants d'immigrés maghrébins (groupe A , pour lequel $magh=1$), et celle des non-descendants (groupe B , $magh=0$) tel que mesuré dans les données de l'Enquête Emploi en continu. On souhaite repondérer les non-descendants pour qu'ils ressemblent, en termes d'expérience potentielle et de diplôme, aux descendants d'immigrés maghrébins. On va ainsi être amené à augmenter le poids des non-descendants dont les caractéristiques sont courantes parmi les descendants (par exemple, les individus jeunes) relativement à celui des non-descendants dont les X sont rares parmi les descendants d'immigrés maghrébins.

La première étape consiste alors à estimer la probabilité conditionnelle à X d'appartenir au groupe des descendants d'immigrés maghrébins, $P(g = A|X)$, en fonction de l'expérience potentielle, de son carré et du niveau d'études des individus (les notations sont les mêmes que dans l'application 4).

```
#Appartenance au groupe A conditionnellement aux X
logit<-glm(magh ~ exp + exp2 + as.factor(ddipl),
            family=binomial (link='logit'),
            data=base)
summary(logit)$coefficients
```

13. Des travaux ultérieurs, notamment Altonji et al. (2012), ont proposé des extensions à des variables catégorielles ou continues.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.46485	0.057493	-60.265	0.00e+00
exp	0.02627	0.005849	4.492	7.05e-06
exp2	-0.00188	0.000148	-12.665	9.26e-37
as.factor(ddipl1)3	0.02399	0.056621	0.424	6.72e-01
as.factor(ddipl1)4	0.32650	0.050984	6.404	1.51e-10
as.factor(ddipl1)5	0.32497	0.051548	6.304	2.90e-10
as.factor(ddipl1)6	0.65246	0.081772	7.979	1.47e-15
as.factor(ddipl1)7	1.06533	0.061558	17.306	4.24e-67

On voit par exemple qu'à expérience donnée, les descendants d'immigrés maghrébins sont $e^{1.07} = 2.9$ fois plus susceptibles d'être sans diplôme (modalité 7) plutôt que diplômés d'un Bac+3 ou plus (modalité de référence), relativement aux non-descendants. On sera donc amené à répondre à la hausse les observations des non-descendants sans diplôme.

Le facteur de répondération $\widehat{\Psi}(X) = \frac{\widehat{P}(g=A|X)}{1-\widehat{P}(g=A|X)} \cdot \frac{1-\widehat{P}(g=A)}{\widehat{P}(g=A)}$ est calculé de la façon suivante :

```
#Prediction à partir de l'estimation précédente
p<-predict(logit,type='response')

#On corrige les poids des B à partir de cette prédiction
w1<-ifelse(base$magh==0,
            p/(1-p)*(1-mean(base$magh))/mean(base$magh), 1)
```

On peut s'assurer que cette opération a bien rendu comparable les deux populations selon les dimensions observables considérées. Par exemple, la proportion d'individus sans diplôme était initialement de 0.095 parmi les non-descendants contre 0.133 parmi les descendants. Elle est de 0.134 parmi les non-descendants répondrés.

Une fois les pondérations obtenues, on pourra directement calculer la statistique d'intérêt sur la distribution contrefactuelle, c'est-à-dire ici sur la distribution de salaire des non-descendants répondrés pour ressembler aux descendants en termes de caractéristiques observables. La fonction `wtd.quantile` du package `Hmisc` permet notamment de calculer des quantiles en incluant des pondérations. On écrira par exemple, pour obtenir les déciles du log-salaire dans la population des non-descendants répondrés :

```

library(Hmisc)

#On calcule les déciles 1 à 9 de la distribution contrefactuelle
#(la distribution dans le groupe B répondue)
dfl.Fc<-wtd.quantile(base$logsal[base$magh==0] ,
  weights=w1[base$magh==0] , probs=seq(0.1,0.9,0.1))

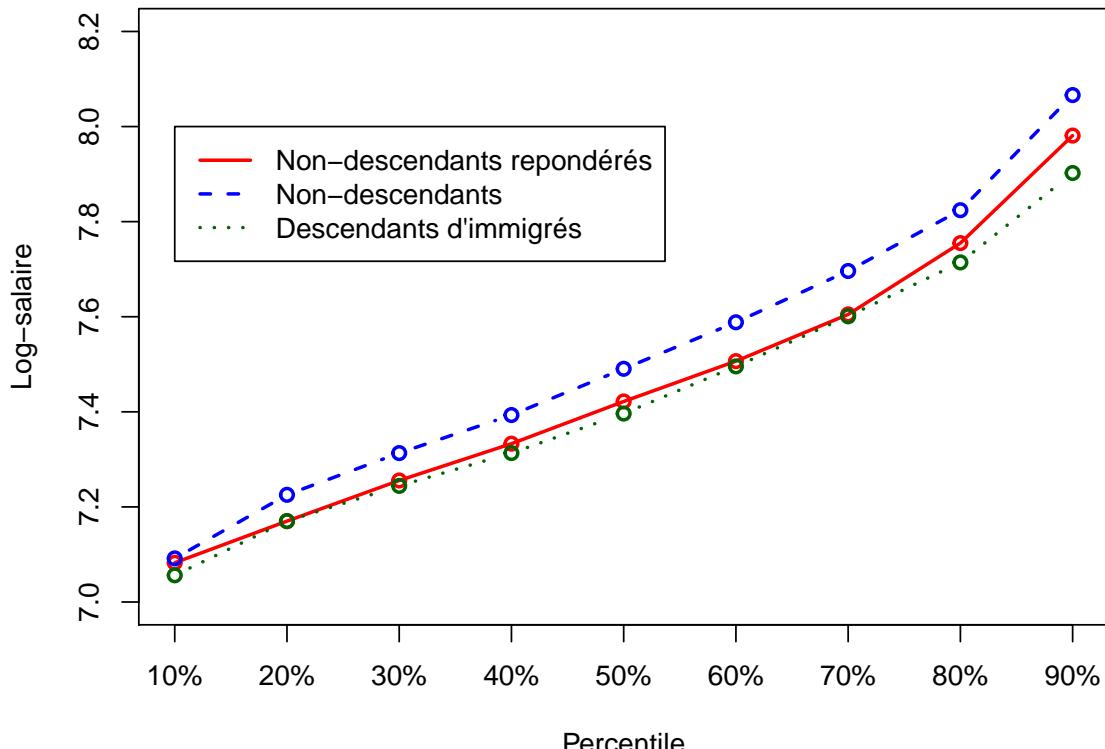
#A comparer aux déciles de la distribution des salaires des B
dfl.FB<-wtd.quantile(base$logsal[base$magh==0] ,
  probs=seq(0.1,0.9,0.1))

#Et à ceux de la distribution des salaires des A
dfl.FA<-wtd.quantile(base$logsal[base$magh==1] ,
  probs=seq(0.1,0.9,0.1))

```

Les déciles de log-salaire ainsi obtenus pour la distribution contrefactuelle, ainsi que les distributions initiales, sont présentées à la figure 4.

FIGURE 4 – Distributions de log-salaire selon le groupe considéré



Jusqu'au septième décile de revenus, la distribution contrefactuelle est très proche de celle des descendants d'immigrés maghrébins. Les différences d'expérience et de diplôme entre les deux groupes expliquent presque entièrement les différences de salaires mesurées entre ces quantiles. Toutefois, pour les déciles de salaire les plus élevés, l'écart inexplicable entre les deux groupes se creuse.

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

 10%   20%   30%   40%   50%   60%   70%   80%   90%
0.036 0.055 0.069 0.080 0.094 0.093 0.095 0.110 0.164

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

 10%   20%   30%   40%   50%   60%   70%   80%   90%
0.009 0.055 0.058 0.060 0.069 0.082 0.091 0.069 0.085

#Dont écart inexplicable
round(dfl.Fc-dfl.FA,3)

 10%   20%   30%   40%   50%   60%   70%   80%   90%
0.026 0.000 0.011 0.020 0.025 0.011 0.004 0.041 0.079
```

4.2 Les décompositions par régression des quantiles non-conditionnels

Pour détailler le rôle de chacune des variables dans la décomposition de façon à la fois additive et indépendante à l'ordre, [Firpo et al. \(2007\)](#) proposent une solution qui se rapproche de l'esprit de la décomposition d'Oaxaca-Blinder, mais adaptée au cas où l'on considère d'autres statistiques que la moyenne, notamment les quantiles de la distribution. Notons que dans le cas de la moyenne, le modèle de régression linéaire permettra d'écrire la moyenne empirique de Y comme $\bar{Y} = \bar{X}\hat{\beta}$, ce qui autorise ensuite à procéder à la décomposition d'Oaxaca-Blinder. Or si l'on considère le quantile d'ordre τ de la distribution de Y (on le note $Q^\tau(Y)$), il existe une modélisation qui permettra *in fine* d'exprimer le quantile empirique $\hat{Q}^\tau(Y)$ comme une fonction linéaire des X moyens, c'est-à-dire comme $\hat{Q}^\tau(Y) = \bar{X}\hat{\gamma}^\tau$: c'est la méthode des régressions de quantiles non-conditionnels, proposée par [Firpo et al. \(2009\)](#).

Les valorisations γ_A^τ et γ_B^τ qu'on estime au sein de chaque sous-population pour un quantile d'ordre τ donné sont l'équivalent des valorisations β_A , β_B dans le cas de la décomposition de la moyenne. On effectuera une décomposition pour chaque $\tau \in [0, 1]$ auquel on s'intéresse (par

exemple $\tau = 0.9$ si l'on souhaite se pencher sur le haut de la distribution) – en règle générale, on s'intéressera à des points tout au long de la distribution). γ^τ correspond à la valorisation des caractéristiques X pour un quantile d'ordre τ donné de la distribution des salaires, mais c'est bien aux X moyens *dans toute la (sous-)population* qu'on les applique – même si, par exemple, on estime l'effet d'être sans diplôme au quantile d'ordre $\tau = 0.9$ de la distribution de salaire¹⁴.

Pour un τ donné, une fois les valorisations γ_A^τ et γ_B^τ estimées, on a d'une part $\hat{Q}_A^\tau(Y) = \bar{X}_A \hat{\gamma}_A^\tau$, d'autre part $\hat{Q}_B^\tau(Y) = \bar{X}_B \hat{\gamma}_B^\tau$. Encore une fois, on voit que l'écart entre les quantiles d'ordre τ dans les groupes A et B peut provenir soit d'une différence de caractéristiques X entre les deux sous-populations, soit d'une différence dans la valorisation de ces caractéristiques moyennes en un point donné de la distribution. La décomposition de l'écart entre $\hat{Q}_B^\tau(Y)$ et $\hat{Q}_A^\tau(Y)$ s'écrit alors comme :

$$\hat{Q}_B^\tau(Y) - \hat{Q}_A^\tau(Y) = \underbrace{\sum_{k=1}^K (\bar{X}_{Bk} - \bar{X}_{Ak}) \hat{\gamma}_{Bk}^\tau + \hat{\gamma}_{B0}^\tau - \hat{\gamma}_{A0}^\tau}_{\hat{\Delta}_X^\tau} + \underbrace{\sum_{k=1}^K \bar{X}_{Ak} (\hat{\gamma}_{Bk}^\tau - \hat{\gamma}_{Ak}^\tau)}_{\hat{\Delta}_S^\tau}$$

Notons qu'on introduit ainsi le contrefactuel $\bar{X}_A \hat{\gamma}_B^\tau$, qui correspond à la façon dont les caractéristiques moyennes des individus du groupe A seraient valorisées par les “rendements” que connaissent les B au quantile (non-conditionnel) d'ordre τ . La décomposition détaillée obtenue est bien, tout comme la décomposition d'Oaxaca-Blinder, additive, et indépendante à l'ordre.

Pour estimer les γ , on a recours aux régressions de quantiles non-conditionnels, ou régressions sur RIF (pour *Recentered Influence Function*, ou fonction d'influence recentrée). La fonction d'influence, outil classique en statistique robuste, appréhende la façon dont une observation particulière Y_i influence une statistique donnée. Autrement dit, elle renvoie au “rôle” d'une observation particulière Y_i sur la valeur d'une statistique. Dans le cas où la statistique considérée est le quantile d'ordre τ de la distribution de Y (qu'on note Q^τ), la fonction d'influence recentrée associe à Y_i la grandeur suivante :

$$RIF(Y_i; Q^\tau) = Q^\tau + \frac{\tau - \mathbf{1}\{Y_i \leq Q^\tau\}}{f_Y(Q^\tau)}.$$

Pour un quantile Q^τ donné, cette fonction ne prendra que deux valeurs selon que Y_i se situe en-dessous ou au-dessus de Q^τ . Si l'on considère par exemple une distribution de salaire dont la médiane ($\tau = 0.5$) vaut 1700, la fonction d'influence recentrée vaut pour chaque Y_i :

14. La démarche de [Firpo et al. \(2009\)](#) consiste en effet à trouver un paramètre γ^τ tel qu'on puisse écrire $Q^\tau(Y) = X\gamma^\tau$, ce qui permet de revenir à une forme linéaire de décomposition. C'est donc bien un effet moyen sur l'ensemble de la (sous-)population qui intervient ici.

$$1700 + \frac{0.5 - \mathbf{1}\{Y_i \leq 1700\}}{f_Y(1700)}.$$

Une régression de quantile non-conditionnel (au quantile d'ordre τ) correspond ensuite simplement à une régression par MCO de la grandeur $RIF(Y_i; Q^\tau)$ sur X . Ce faisant, on modélise comme dans une régression linéaire classique $\mathbb{E}([RIF(Y, Q^\tau)|X] = X.\gamma^\tau + \epsilon$. Or la RIF permet d'écrire $\mathbb{E}[RIF(Y, Q^\tau)] = Q^\tau$, et de là $Q^\tau = \mathbb{E}[RIF(Y, Q^\tau)] = \mathbb{E}_X[\mathbb{E}([RIF(Y, Q^\tau)|X]] = \mathbb{E}[X].\gamma^\tau$. La contrepartie empirique de cette expression, $\hat{Q}^\tau = \bar{X}\hat{\gamma}^\tau$, permet la décomposition présentée plus haut. L'obtention des valorisations γ^τ se fait donc à travers deux étapes simples : transformation de chaque Y_i en $RIF(Y_i; Q^\tau)$; puis régression linéaire de $RIF(Y_i; Q^\tau)$ sur les X .

Cette méthode est très simple à mettre en œuvre pour les quantiles : les γ peuvent être directement estimés à l'aide du package `uqr` (pour *unconditional quantile regression*), voir l'application 6. L'emploi de la RIF peut également être élargi à d'autres statistiques distributionnelles que les quantiles¹⁵, notamment des statistiques s'appuyant sur les quantiles (par exemple rapport interdécile ou taux de pauvreté relative). Un des inconvénients de la méthode des RIF est que celle-ci nécessite d'utiliser une approximation locale qui peut être de mauvaise qualité en présence de points de masse.

Application 6 : Décomposition par les RIF

On étudie à nouveau les différences de salaires entre descendants d'immigrés maghrébins et personnes sans ascendance migratoire. Le package `uqr` permet d'implémenter des régressions de quantiles non-conditionnels sous R, à travers la fonction `urq`. On spécifie le(s) quantile(s) au(x)quel(s) on souhaite appliquer ces régressions grâce à l'option `tau=`. On effectue cette décomposition séparément pour la population de référence (`ref<-base$magh==0`) d'une part, et pour les descendants d'immigrés (`!ref`) d'autre part.

```
library(uqr)
rif.B<-urq(formula=logsal ~ exp + exp2 + as.factor(ddipl),
            data=as.data.frame(base[,ref]),
            tau=seq(0.1,0.9,0.1))

#On obtient par exemple pour les 3 premiers déciles :
rif.B$coefficients[,1:3]

            tau= 0.1   tau= 0.2   tau= 0.3
(Intercept) 6.954515 7.115559 7.243411
exp          0.022417 0.022688 0.021713
```

15. Pour la moyenne, on retombe sur la régression standard de Y sur X .

```

exp2           -0.000346 -0.000327 -0.000296
as.factor(ddipl)3 -0.064896 -0.089849 -0.106747
as.factor(ddipl)4 -0.158389 -0.213384 -0.246136
as.factor(ddipl)5 -0.198570 -0.270140 -0.318173
as.factor(ddipl)6 -0.237468 -0.302222 -0.341726
as.factor(ddipl)7 -0.343564 -0.412768 -0.453157

rif.A<-urq(formula=logsal ~ exp + exp2 + as.factor(ddipl),
            data=as.data.frame(base[!ref,]),
            tau=seq(0.1,0.9,0.1))

```

Une fois les valorisations γ obtenues pour chaque groupe, on peut procéder à une décomposition "classique" de type Oaxaca-Blinder, à partir des vecteurs de moyennes calculés pour chaque variable, dans la population de référence et parmi les descendants d'immigrés.

```

#Calcul des X moyens dans chaque groupe
X<-model.matrix(logsal ~ exp + exp2 + as.factor(ddipl), base)
moy.B<-apply(X[,ref,],2,mean)
moy.A<-apply(X[!,ref,],2,mean)

#Calcul des écarts expliqués par chaque variable
expl.detail<-apply(rif.B$coefficients, 2, "*",
                    moy.B-moy.A)
expl.detail[,1:3]

tau= 0.1 tau= 0.2 tau= 0.3
(Intercept)      0.00000 0.00000 0.00000
exp             0.11113 0.11247 0.10764
exp2            -0.07797 -0.07379 -0.06682
as.factor(ddipl)3 -0.00135 -0.00188 -0.00223
as.factor(ddipl)4  0.00484 0.00653 0.00753
as.factor(ddipl)5 -0.00840 -0.01143 -0.01347
as.factor(ddipl)6  0.00109 0.00138 0.00156
as.factor(ddipl)7  0.01313 0.01577 0.01732

```

On obtient alors l'ensemble des contributions détaillées à l'effet de composition, pour chaque variable, en chaque décile. On calcule de là les parts expliquées et inexpliquées totales en chaque décile, qu'on pourra notamment comparer aux résultats de la décomposition agrégée obtenus par repondération.

```

rif.expl<-apply(expl.detail,2,sum)

#Ecart expliqués totaux (effet de composition)
round(rif.expl,3)

tau= 0.1 tau= 0.2 tau= 0.3 tau= 0.4 tau= 0.5 tau= 0.6 tau= 0.7 tau= 0.8
    0.042      0.049      0.052      0.063      0.070      0.073      0.079      0.080
tau= 0.9
    0.092

#De même pour les écarts inexpliqués
inexpl.detail<-apply(rif.B$coefficients-rif.A$coefficients,
                      2, "*",moy.A)
rif.inexpl<-apply(inexpl.detail,2,sum)

#Ecart inexpliqués totaux
round(rif.inexpl,3)

tau= 0.1 tau= 0.2 tau= 0.3 tau= 0.4 tau= 0.5 tau= 0.6 tau= 0.7 tau= 0.8
    -0.008      0.001      0.011     -0.003      0.024      0.010      0.003      0.045
tau= 0.9
    0.072

```

L'approximation locale qui est implicitement effectuée par les régressions RIF apparaît variable selon les quantiles considérés. Les écarts de salaire totaux reconstitués ci-dessous sont à comparer aux écarts effectifs présentés en application 5.

```

#Reconstitution des écarts de salaire totaux
par somme des écarts expliqués et inexpliqués
round(rif.expl+rif.inexpl,3)

tau= 0.1 tau= 0.2 tau= 0.3 tau= 0.4 tau= 0.5 tau= 0.6 tau= 0.7 tau= 0.8
    0.034      0.050      0.063      0.060      0.094      0.083      0.082      0.126
tau= 0.9
    0.164

```

4.3 Les méthodes d'estimation de la distribution conditionnelle - Régressions sur fonction de répartition

Repartons de la distribution contrefactuelle $F_{Y_B|X_A}$, qui correspond à la distribution conditionnelle des salaires du groupe B appliquée aux caractéristiques des individus du groupe A . Une façon naturelle de l'obtenir est d'estimer directement la distribution $F_{Y_B|X}$ des Y conditionnelle aux X parmi les individus du groupe B , et de l'appliquer aux caractéristiques du groupe A (X_A).

La fonction de distribution (ou fonction de répartition) peut être résumée par un ensemble de probabilités de se situer en-dessous d'un certain seuil : pour reprendre l'exemple des salaires, on souhaiterait modéliser le fait d'avoir un salaire inférieur à 1500 € par mois, d'avoir un salaire inférieur à 2000 € par mois, etc., en raisonnant à chaque fois à caractéristiques données. Le problème revient ainsi à réaliser un ensemble d'estimations sur des indicatrices (se situer au-dessus ou en-dessous d'un seuil donné), ce qui peut être fait de façon très classique par un modèle logit, un modèle probit, ou même un modèle de probabilité linéaire. Plus le nombre de seuils considérés sera grand, plus l'estimation de la fonction de répartition sera fine. C'est sur cette idée que repose la méthode de régression sur distribution proposée par [Chernozhukov et al. \(2013\)](#).

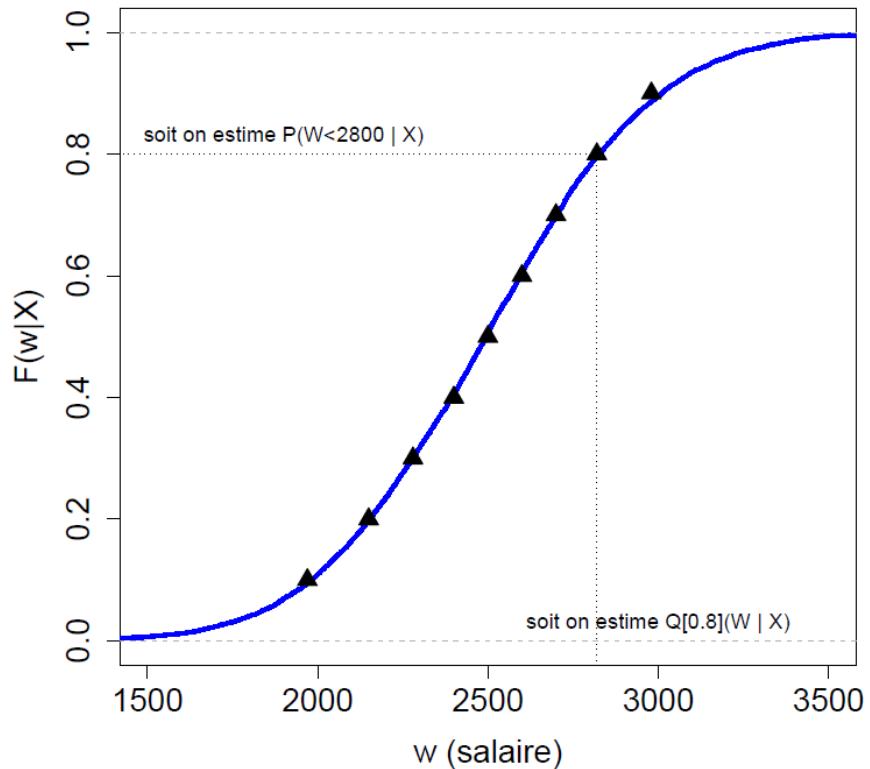
En pratique, la méthode de [Chernozhukov et al. \(2013\)](#) consiste donc en trois étapes, pour chaque seuil $y \in [\min(Y), \max(Y)]$:

- (1) On estime au sein du groupe B un logit, un probit ou une régression linéaire sur $\mathbf{1}\{Y_i \leq y\}$, avec $P(Y < y|X) = G(X\beta(y))$, selon que $G(\cdot)$ est la fonction de répartition de la loi logistique, celle de la loi normale, ou la fonction identité.
- (2) On utilise les coefficients estimés à l'étape (1) pour calculer la probabilité prédictive $G(X_i\hat{\beta}_B(y))$ pour chaque individu i du groupe A .
- (3) On calcule la moyenne de ces probabilités prédictives sur l'ensemble du groupe A pour finalement obtenir $\hat{F}_{Y_B|X_A}(y) = \frac{1}{N_A} \sum_{i \in A} G(X_i\hat{\beta}_B(y))$, c'est-à-dire la probabilité de se situer au-dessous du seuil y qu'auraient les individus du groupe A si leurs caractéristiques étaient valorisées de la même façon que pour les B .

On obtient ainsi un ensemble de probabilités contrefactuelles qui permettent de reconstituer la fonction de distribution souhaitée. Il est toutefois possible que la fonction de distribution estimée ne soit pas monotone : pour y_1 et y_2 proches avec $y_1 < y_2$, rien ne garantit que $\hat{F}_{Y_B|X_A}(y_1) < \hat{F}_{Y_B|X_A}(y_2)$. Il est alors nécessaire d'utiliser une procédure de lissage pour s'assurer qu'elle pourra bien être inversée en quantiles.

Cette méthode est intensive en calculs : elle demande d'effectuer un grand nombre de régressions lorsqu'on souhaite inverser la distribution et revenir à l'écart entre quantiles. On peut toutefois souhaiter simplement décomposer l'écart de probabilité de se trouver au-dessus ou en-dessous d'un certain niveau de salaire, ce qui est même parfois plus parlant (par exemple lorsqu'on considère un seuil de pauvreté en termes absolus ; ou lorsqu'on souhaite définir une population de "hauts revenus" comparable dans les deux groupes). On a alors besoin d'effectuer la procédure uniquement pour le y d'intérêt (ce qui correspond à la méthode de Fairlie, voir section 3.2). En cela, il est plus simple d'estimer des distances "verticales" que des distances "horizontales", qui contiennent toutes les deux la même information. C'est ce qu'illustre la figure 5 : il est équivalent de raisonner sur les probabilités de se situer de part et d'autre d'un certain seuil qui se lisent en ordonnée (par exemple, de gagner moins 2 800 €), ou sur les quantiles correspondant qu'on lit en abscisse (le seuil de salaire en-deçà duquel se situent 80 % des individus). Notons que comme pour Fairlie, si on modélise cette probabilité d'être au-delà d'un seuil par un modèle de probabilité linéaire, on retombe sur une décomposition de type Oaxaca-Blinder.

Figure 5 – Distances verticales ou horizontales : illustration avec une fonction de répartition conditionnelle



Application 7 : Décomposition par les régressions sur fonction de répartition

On reprend le même cas d'usage que dans l'application 6, en considérant les écarts de distribution de log salaire entre descendants et non descendants d'immigrés maghrébins. On utilise la fonction `counterfactual` du package `Counterfactual` pour faire une décomposition à la Chernozhukov et al. (2013).

```
library(Counterfactual)
```

Le paramètre `group` permet d'indiquer quels sont les groupes comparés. On utilise des fonctions logit, c'est-à-dire qu'on modélise en 500 points de la distribution (paramètre `nreg`) la probabilité de se situer à gauche ou à droite de ce point. En indiquant `noboot=TRUE` on renonce à bootstraper l'analyse pour gagner du temps (le calcul peut s'avérer très long).

```
#régressions sur distribution
cvfm<-counterfactual(logsal ~ exp + exp2 + as.factor(ddipl),
                       data= base,
                       group=ref,
                       treatment=TRUE,
                       decomposition=TRUE,
                       method="logit",
                       nreg=200,
                       noboot=TRUE)
```

Une fois les 500 modèles estimés, on peut récupérer simplement les écarts totaux, les effets de composition et les écarts inexpliqués obtenus à chaque décile de la distribution des salaires. Ils sont calculés en appliquant la fonction conditionnelle estimée chez les non descendants aux caractéristiques des descendants.

```
#Ecarts totaux
#(calculés comme écarts expliqués + inexpliqués)
t(round(cvfm$total_effect,3))

 [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
[1,] 0.042 0.056 0.072 0.084 0.099 0.092 0.095 0.11 0.164

#Dont effet de composition
t(round(cvfm$composition_effect,3))

 [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
[1,] 0.037 0.035 0.045 0.057 0.051 0.053 0.044 0.065 0.094
```

```
#Dont écart inexpliqué
t(round(cvfm$structural_effect,3))

 [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
[1,] 0.005 0.021 0.027 0.027 0.047 0.039 0.051 0.045 0.07
```

On peut remarquer qu'on obtient des décompositions un peu différentes de celles obtenues précédemment par repondération ou avec les RIF, en particulier avec un rôle plus réduit de l'effet de composition, en particulier dans le milieu de la distribution.

Encadré 6 : Les méthodes d'estimation de la distribution conditionnelle - Estimation de distributions contrefactuelles par régressions quantiles

Il existe d'autres manières que celle de [Chernozhukov et al. \(2013\)](#) de construire une distribution contrefactuelle du groupe A en mimant la distribution conditionnelle des B. [Machado and Mata \(2005\)](#) et [Melly \(2005\)](#) simulent ainsi, pour chaque individu du groupe A de caractéristiques données, un ensemble de valeurs prédites par régressions quantiles balayant la distribution conditionnelle du groupe B.

- (1) On tire un ensemble de valeurs $\tau_1, \tau_2, \dots, \tau_S$ entre 0 et 1 (chez [Machado and Mata \(2005\)](#), la grille est définie à l'avance).
- (2) En chacun des quantiles d'ordre τ_s :
 - (2a) On estime une régression quantile linéaire au sein du groupe B (cela permet d'estimer une fonction de rendement des caractéristiques au sein de ce groupe, en S points de la distribution).
 - (2b) Les rendements estimés chez les B permettent de prédire un Y_{As}^C contrefactuel à partir des X de chaque individu du groupe A.

En parcourant chaque quantile de la distribution conditionnelle des B et en intégrant sur les X des A, on retrouve la distribution contrefactuelle d'intérêt. Celle-ci est directement constituée par l'ensemble des prédictions Y_{As}^C pour chaque individu du groupe A, et pour chaque $s = 1 \dots S$.

Cette méthode connaît certaines limites : elle impose de faire la simulation en un grand nombre de points, même dans le cas où l'on ne s'intéresse qu'à la décomposition en un seul quantile de la distribution. La procédure est ainsi très intensive en calculs (même si la version de [Melly \(2005\)](#) consistant à tirer à chaque itération un ensemble de X dans l'échantillon des A l'est un peu moins). De plus, la spécification linéaire peut être restrictive dans certaines applications, notamment dans le cas où la distribution des Y présente

des points de masse (cela peut être le cas pour une distribution de salaire en présence d'un salaire minimum).

Pour ces méthodes qui s'appuient sur l'estimation de la distribution conditionnelle, la décomposition détaillée est possible mais là encore elle ne peut être à la fois additive et indépendante à l'ordre des variables.

5 Le traitement de la sélection dans la décomposition des inégalités

On a évoqué en section 2.1 que lorsque les groupes A et B sont soumis à des mécanismes de sélection (typiquement en emploi, ou dans certains statuts d'emploi) différents après contrôle des caractéristiques observables, l'hypothèse d'indépendance conditionnelle pouvait être violée et l'interprétation de la décomposition délicate. Ce point est d'autant plus problématique qu'il se retrouve souvent lorsque l'on souhaite étudier des phénomènes de discrimination. Notamment, pour interpréter un écart de salaire comme résultant d'une discrimination salariale, il faut être certain que conditionnellement aux caractéristiques observables, les inobservables des deux groupes soient en moyenne les mêmes. Or typiquement s'il y a discrimination, il est fort probable que celle-ci joue aussi sur le fait d'avoir ou non un emploi. Les barrières à l'emploi seront en toute logique plus hautes pour le groupe potentiellement discriminé que pour le groupe majoritaire, ce qui conduira à ce que les inobservables du groupe minoritaire soient en moyenne plus favorables que celles du groupe majoritaire et à une sous-estimation *a priori* de la discrimination salariale. Pour autant, si en plus d'une potentielle discrimination les différences dans les mécanismes de sélection entre les deux groupes proviennent d'autres facteurs comme les normes sociales par exemple, alors le sens de l'effet est ambigu : il n'y a plus de sous-estimation systématique. Pour quelques exemples de résultats en économie du travail, on se reporterà à l'encadré 7.

Encadré 7 : Marché du travail et sélection

A titre d'exemple, [Neal \(2004\)](#) et [Mulligan and Rubinstein \(2008\)](#) documentent aux Etats-Unis une sous-représentation dans l'emploi des femmes provenant des groupes les plus favorisés, les plus diplômés, phénomène qu'ils relient à des normes sociales. Ce phénomène tendrait à ce que l'écart de salaire entre hommes et femmes si les femmes diplômées étaient aussi représentées que les hommes dans l'emploi soit plus faible que l'écart observé. Dans tous les cas, il n'y a pas d'interprétation causale possible dès que les mécanismes de sélection ne sont pas les mêmes dans les deux groupes. La prise en compte de ces

mécanismes peut conduire à des résultats différents dans les décompositions, de même que leurs évolutions au cours du temps. L'article de [Mulligan and Rubinstein \(2008\)](#) illustre bien comment un mécanisme de sélection qui évolue dans le temps peut conduire à une tendance temporelle trompeuse sur l'écart de salaire entre les hommes et les femmes. Il conclut que la baisse de l'écart de salaire entre les hommes et les femmes observée aux Etat-Unis entre la fin des années 1970 et celle des années 1990, s'explique entièrement par l'évolution de la sélection dans l'emploi des femmes qui vient modifier la composition de la population féminine en emploi : l'écart de salaire corrigé de cette sélection est au contraire stable sur la période. [Blau and Kahn \(2006\)](#) évoquent également l'évolution des mécanismes de sélection sur le marché du travail pour comprendre le ralentissement dans la réduction des écarts de salaire entre hommes et femmes entre les années 1980 et les années 1990.

La question de la sélection est abondamment étudiée dans la littérature. Celle-ci constitue un domaine encore en cours d'investigation : il n'y a pas de méthode unique pour la traiter car elle requiert des hypothèses supplémentaires, parfois différentes, que le chargé d'étude sera prêt ou non à supposer en fonction du cas de décomposition qu'il étudie et des données qu'il a à disposition. Cette dernière partie du document de travail ne préconisera donc pas une méthode en particulier ni ne donnera une liste exhaustive des approches possibles. L'objectif est plutôt de suggérer les types de méthodes les plus adaptés selon les données à disposition et les hypothèses supplémentaires requises.

5.1 Sélection exogène, sélection endogène

L'hypothèse déterminante dans le choix du traitement de la sélection à effectuer porte sur le caractère endogène ou non du processus de sélection. Afin de mieux comprendre ce point, écrivons un modèle de sélection pour chacun des deux groupes $g = A, B$. Pour simplifier la lecture, on va considérer que l'on s'intéresse à des écarts de salaires pour des groupes qui ne connaissent pas les mêmes mécanismes de sélection dans l'emploi, même si les cas d'application sont plus généraux. On modélise ici la sélection en fonction des mêmes caractéristiques observables que celles incluses dans la décomposition des écarts de salaire.

$$\begin{cases} \text{Equation de sélection : } E = \mathbf{1}\{\sum_{k=0}^K X_k \gamma_{gk} + u_g \geq 0\} \\ \text{Equation de salaire potentiel : } Y^* = \sum_{k=0}^K X_k \beta_{gk} + v_g \end{cases}$$

où $Y = Y^* \times E$, le salaire n'est observé que lorsque $E = 1$, l'individu est en emploi. On peut distinguer deux cas. La sélection est dite “exogène” (sélection sur observables) lorsque le terme d'erreur de l'équation de salaire potentiel (v_g) et celui de l'équation de sélection (u_g) sont indépendants (conditionnellement à X), elle est dite “endogène” (sélection sur inobservables) sinon. Il faudra dans ce cas tenir compte de la corrélation entre les deux termes d'erreur.

La sélection exogène est dite aussi “ignorable” en ce sens que des rendements estimés sur les seuls individus en emploi valent aussi en population générale. En ce sens, les méthodes de décomposition présentées jusqu’ici autorisent qu’il existe par exemple un processus de sélection plus strict en fonction du diplôme et de l’âge dans l’un des deux groupes (accès à l’emploi qui demande d’être plus qualifié ou expérimenté dans l’un des deux groupes). Ce n’est pas vrai en cas de sélection endogène, ou “non ignorable”.

5.1.1 Décrire une sélection exogène

On désigne par E_g le processus de sélection dans l’emploi sur observables au sein du groupe g et par $F_{Y_B|X_A, E_g}$ la distribution contrefactuelle appliquant la structure de salaires des B aux caractéristiques observables des A, selon le processus de sélection du groupe g . L’écart observé de salaire entre les individus B et A en emploi en terme de statistique ν , soit Δ_O^ν , se décompose comme :

$$\Delta_O^\nu = \underbrace{[\nu(F_{Y_B}) - \nu(F_{Y_B|X_A, E_B})]}_{\Delta_X^\nu} + \underbrace{[\nu(F_{Y_B|X_A, E_B}) - \nu(F_{Y_B|X_A, E_A})]}_{\Delta_E^\nu} + \underbrace{[\nu(F_{Y_B|X_A, E_A}) - \nu(F_{Y_A})]}_{\Delta_S^\nu}$$

L’écart inexpliqué Δ_S^ν correspond au même écart que dans la décomposition ne tenant pas compte de la sélection. La distribution contrefactuelle $F_{Y_B|X_A, E_B}$ est introduite afin de scinder en deux parties l’effet de composition initial Δ_X^ν en une partie liée à l’effet en termes de salaire des différences de composition entre les deux groupes à processus de sélection donné ($\Delta_{X'}^\nu$), et une partie liée aux différences de processus de sélection sur observables Δ_E^ν . En effet, l’effet de composition initial ne distingue pas d’emblée ce qui relève de différences de caractéristiques observables et ce qui relève de différences induites par des processus de sélection sur observables différents. Par exemple, si les descendants d’immigrés sont moins diplômés que les descendants de natifs mais qu’une partie importante des moins diplômés d’entre eux n’est pas en emploi, on pourra conclure à tort que les différences d’éducation entre les deux groupes n’ont que peu d’impact visible en termes d’écarts de salaire, alors même que les différences initiales de niveau d’éducation entre les deux groupes conduiraient à des écarts de composition importants s’ils n’étaient pas gommés (en partie) par le processus de sélection.

Repondération Chiquiar and Hanson proposent une adaptation de la méthode de répondération de DiNardo et al. (1996) permettant de décrire la part de l’effet de composition liée à la sélection selon les observables. Dans le facteur Ψ_{DFL} introduit plus haut qui conduit, par répondération de la population des individus B en emploi, au contrefactuel $F_{Y_B|X_A, E_A}$, et donc à l’effet de composition $\Delta_X^\nu = \Delta_E^\nu + \Delta_{X'}^\nu$, ils isolent la partie correspondant à l’effet des différences de processus de sélection entre les deux populations du reste, ce qui leur permet

ainsi de reconstruire Δ_E^ν par repondération aussi, et $\Delta_{X'}^\nu$ par différence.

Plus précisement, $\Psi_{DFL}(X)$, calculé sur les seuls individus A et B en emploi, peut aussi se réécrire en fonction du ratio des probabilités d'emploi des individus du groupe A et B et du ratio des probabilités d'appartenance au groupe A ou B en population générale, en appliquant deux fois la règle de Bayes :

$$\begin{aligned}\Psi_{DFL}(X) &= \frac{P(X|g = A, E = 1)}{P(X|g = B, E = 1)} \\ &= \underbrace{\frac{P(E = 1|g = A, X)}{P(E = 1|g = B, X)}}_{\Psi_E} \times \underbrace{\frac{P(g = A|X)}{P(g = B|X)}}_{\Psi_{X'}} \times \frac{P(E = 1, g = B)}{P(E = 1, g = A)}\end{aligned}$$

ainsi Ψ_E rend compte des écarts de probabilité d'emploi entre les deux populations, $\Psi_{X'}$ des écarts de distribution des caractéristiques observables entre les groupes A et B en population générale et le dernier terme est une constante. Chiquiar and Hanson proposent alors de repondérer les salaires des B en emploi par $\Psi_{X'}$ pour obtenir $F_{Y_B|X_A, E_B}$ et en déduire $\Delta_{X'}^\nu$. En pratique, $\Psi_{X'}$ s'obtient à partir d'une estimation logistique chez tous les individus du groupe A ou B (poids $w2$ dans l'exemple numérique), alors que Ψ_{DFL} s'obtient à partir de la même estimation logistique sur les seuls individus en emploi (poids $w1$ dans l'exemple numérique).

Application 8 : Correction de la sélection sur observables par repondération

On repart de la base de l'enquête Emploi en continu comportant tous les individus actifs et on considère le fait d'être en emploi, à temps plein et avec un salaire observé dans les données ($sel=1$). On calcule les deux facteurs de repondération présentés plus haut :

```
sel<-base$tppred=="1"&base$acteu=="1"&!is.na(base$logsal)

#Facteur de pondération chez les personnes vérifiant sel=1
logit<-glm(magh ~ exp + exp2 + as.factor(ddipl),
            family=binomial (link='logit'), data=base[which(sel),])
p<-predict(logit,type='response',newdata=base)
w1<-ifelse(ref,p/(1-p), 1)

#Le deuxième facteur de repondération est calculé en population générale
logit.pg<-glm(magh ~ exp + exp2 + as.factor(ddipl),
               family=binomial (link='logit'), data=base)
p.pg<-predict(logit.pg,type='response',newdata=base)
w2<-ifelse(ref,p.pg/(1-p.pg), 1)
```

Comme pour les méthodes de repondération classiques, on utilise ces jeux de poids pour calculer les statistiques de la distribution contrefactuelle qui nous intéresse.

```
#Le premier contrefactuel obtenu pour les individus sel=1
dfl.Fc<-wtd.quantile(base$logsal[ref&sel],
                        weights=w1[ref&sel], probs=seq(0.1,0.9,0.1))

#Le deuxième contrefactuel a partir des poids en population générale
dfl.Fc.pg<-wtd.quantile(base$logsal[ref&sel],
                           weights=w2[ref&sel], probs=seq(0.1,0.9,0.1))

#La distribution dans le groupe de référence
dfl.FB<-wtd.quantile(base$logsal[ref&sel],
                       probs=seq(0.1,0.9,0.1))

#Et dans le groupe minoritaire
dfl.FA<-wtd.quantile(base$logsal[!ref&sel],
                       probs=seq(0.1,0.9,0.1))
```

On peut alors récupérer, pour chaque décile, la part de l'écart lié aux caractéristiques observables à *processus de sélection donné*; l'écart lié à la différence des mécanismes de sélection entre les deux groupes; et l'écart inexpliqué.

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

  10%   20%   30%   40%   50%   60%   70%   80%   90%
0.036 0.055 0.069 0.080 0.094 0.093 0.095 0.110 0.164

#Ecart lié aux caractéristiques observables (à processus de sélection donné)
round(dfl.FB-dfl.Fc.pg,3)

  10%   20%   30%   40%   50%   60%   70%   80%   90%
0.019 0.055 0.069 0.078 0.091 0.093 0.095 0.089 0.115

#Ecart lié à la différence dans les processus de sélection (sur observables)
round(dfl.Fc.pg-dfl.Fc,3)

  10%   20%   30%   40%   50%   60%   70%   80%   90%
-0.010 0.000 -0.011 -0.018 -0.022 -0.011 -0.004 -0.020 -0.029
```

```
#Ecart inexpliqué
round(dfl.Fc-dfl.FA,3)

 10%   20%   30%   40%   50%   60%   70%   80%   90%
0.026 0.000 0.011 0.020 0.025 0.011 0.004 0.041 0.079
```

Imputation Une autre approche visant à corriger la sélection sur caractéristiques observables consiste à imputer des outcomes potentiels aux individus pour lesquels on n'observe pas la variable intérêt, ici le salaire. Ce type d'approche s'applique à n'importe quelle statistique de salaire, moyenne ou autre, mais l'est plus fréquemment pour la médiane car dans ce cas la valeur précise imputée du salaire n'importe pas, seule compte sa position par rapport à la médiane des salaires. Comme article représentatif de cette approche, [Neal \(2004\)](#) s'intéresse à l'écart de salaire entre les femmes blanches et noires. Il propose plusieurs scenarios pour imputer des salaires aux femmes blanches et noires qui ne participent pas au marché du travail et analyse comment cela change l'écart de salaire à la médiane. Il suppose par exemple que les femmes noires peu éduquées avec un niveau de vie faible qui n'ont pas travaillé depuis un certain temps, ont manqué d'opportunités professionnelles - il leur affecte alors un salaire potentiel similaire, voire plus bas que celui observé en moyenne pour les femmes blanches de mêmes caractéristiques. Il suppose que les femmes blanches les plus éduquées, de niveau de vie élevé (dont le revenu de l'époux est élevé) qui ne travaillent pas, ont au contraire choisi de ne pas travailler - il leur affecte alors un salaire potentiel très élevé.

Les approches par imputation s'appliquent aussi dans le cas des données de panel. Dans ce cas, il est même possible de tenir compte d'un certain type de sélection endogène, une sélection sur inobservables fixes dans le temps. Cette approche sera détaillée en section [5.2](#), après la présentation de l'approche classique de traitement de la sélection sur inobservables à partir d'une fonction de contrôle.

5.1.2 Traiter la sélection sur inobservables avec une fonction de contrôle

Si les caractéristiques observables ne suffisent pas à rendre compte des différences dans les processus de sélection dans l'emploi – par exemple, ce ne sont pas seulement les descendants d'immigrés les plus diplômés qui sont en emploi, mais ce sont également, parmi les plus diplômés, les plus motivés – alors on est confronté à un problème de sélection endogène (différenciée entre les deux groupes). L'approche classique pour corriger les salaires moyens de la sélection consiste à ajouter une fonction de contrôle (*control function*) comme régresseur de l'équation de salaire ([Fortin et al., 2011](#)). Cette fonction de contrôle doit faire intervenir un (ou des) instrument(s), c'est-à-dire une caractéristique observée corrélée avec la partici-

pation mais pas avec le niveau de salaire (condition d'exclusion classique dans les modèles de sélection). Sans instrument valide, l'identification des modèles de sélection repose sur la forme distributionnelle supposée des erreurs, ce qui s'avère rarement vérifié en pratique (l'hypothèse de normalité des résidus requise est forte, voir par exemple Wooldridge (2010), page 563). L'effet différencié de cette fonction de contrôle entre les deux groupes est intégré à la décomposition. Dans la plupart des applications, elle correspond à l'inverse du ratio de Mills obtenu en régressant un modèle probit sur la sélection. C'est l'approche paramétrique inspirée de Heckman (1990) et détaillée ci-dessous. La fonction de contrôle peut aussi s'estimer de manière semi-paramétrique (voir Vella 1998).

L'approche inspirée de Heckman (1990) revient à considérer de façon conjointe une équation de sélection/participation et une équation de salaire, dont les erreurs sont corrélées. Tout comme l'équation de salaire, l'équation de sélection n'est pas régie par les mêmes paramètres dans chacun des groupes puisque la sélectivité est propre à chaque groupe. Neuman and Oaxaca (2004) (voir aussi Aeberhardt et al. 2010a) considèrent ainsi le modèle de sélection et de salaire suivant :

$$\begin{cases} \text{Equation de sélection : } E = \mathbf{1}\{\sum_{k=0}^K H_k \gamma_{gk} + u_g \geq 0\}. \\ \text{Equation de salaire potentiel : } Y^* = \sum_{k=0}^K X_k \beta_{gk} + v_g \end{cases}$$

où $Y = E \times Y^*$ le salaire, n'est observé que si l'individu est sélectionné / est en emploi $E = 1$, $X \subset H$, mais H contient aussi un instrument (corrélé avec l'équation de sélection mais n'influant pas sur le salaire). Les erreurs des deux équations v et u - sélection et salaire - sont supposées suivre une loi normale bivariée de paramètres $\begin{pmatrix} 0 & 0 \\ 0 & \rho_g \sigma_{vg} \end{pmatrix}, \begin{pmatrix} 1 & \rho_g \sigma_{vg} \\ \rho_g \sigma_{vg} & \sigma_{vg}^2 \end{pmatrix}$, g désignant l'appartenance au groupe A ou B. La décomposition de la différence de salaire entre les deux groupes en emploi est alors :

$$\begin{aligned} \mathbb{E}(Y_{iA}|E_{iA} = 1) - \mathbb{E}(Y_{iB}|E_{iB} = 1) &= (\mathbb{E}_B(X'_i|E_{iB} = 1) - \mathbb{E}_A(X'_i|E_{iA} = 1)) \beta_B \\ &\quad + \mathbb{E}_B(X'_i|E_{iB} = 1)(\beta_B - \beta_A) \\ &\quad + \rho_B \sigma_{vB} \mathbb{E}_B(\lambda_{iB}|E_{iB} = 1) - \rho_A \sigma_{vA} \mathbb{E}_A(\lambda_{iA}|E_{iA} = 1). \end{aligned}$$

Soit,

$$\overline{Y_B} - \overline{Y_A} = \underbrace{\sum_{k=1}^K (\overline{X_{Bk}} - \overline{X_{Ak}}) \hat{\beta}_{Bk}}_{\text{Expliqué}} + \underbrace{(\hat{\beta}_{B0} - \hat{\beta}_{A0}) + \sum_{k=1}^K \overline{X_{Ak}} (\hat{\beta}_{Bk} - \hat{\beta}_{Ak})}_{\text{Inexpliqué}} + \underbrace{\hat{\theta}_B \hat{\lambda}_B - \hat{\theta}_A \hat{\lambda}_A}_{\text{Sélection}}, \quad (9)$$

avec $\theta_g = \rho_g \sigma_{vg}$ et $\lambda_g = \frac{\phi\left(\sum_{k=0}^K H_{gk} \gamma_{gk}\right)}{\Phi\left(\sum_{k=0}^K H_{gk} \gamma_{gk}\right)}$, $\phi(\cdot)$ faisant référence à la densité d'une loi normale standard, et $\Phi(\cdot)$ à sa fonction de répartition.

Cependant, une fois arrivé à la décomposition (9), il n'est pas évident de savoir à quelle partie, expliquée ou inexpliquée, rattacher la différence des effets de sélection ($\hat{\theta}_B \hat{\lambda}_B - \hat{\theta}_A \hat{\lambda}_A$). Ici, il est choisi de seulement isoler ce terme (voir l'application). Le lecteur pourra se référer à [Neuman and Oaxaca \(2004\)](#) et [Neuman and Oaxaca \(2005\)](#) qui cherchent à décomposer ce terme, en recourant à des hypothèses ou des conventions supplémentaires parfois conduisant à plusieurs décompositions, ainsi qu'à [Aeberhardt et al. \(2010a,b\)](#) qui proposent une alternative qui ne requiert pas d'estimer de paramètres au sein du groupe minoritaire - on pourra s'y référer dès qu'on compare des groupes de taille très différente afin de gagner en précision.

L'exemple numérique ci-dessous donne les éléments pour retrouver la décomposition (9). Ici, à la façon de [Mroz \(1987\)](#) ou [Hyslop \(1999\)](#), on utilise la composition familiale (typiquement, le nombre d'enfants) comme variable instrumentale, en supposant que celle-ci joue sur la décision de participation mais n'a pas d'impact direct sur le niveau du salaire.

Application 9 : Correction de la sélection à l'aide d'une fonction de contrôle

On charge le package `sampleSelection` qui propose une boîte à outils pour l'estimation de modèles de sélection, dont le modèle d'Heckman. On veut contrôler de la sélection en emploi à taux plein avec salaire effectivement observé dans les données (la variable est la même que dans l'application 8) dans la comparaison des salaires entre hommes et femmes.

```
library(sampleSelection)
#taux d activite chez les hommes et les femmes
prop.table(table(base$sel, base$sexe), margin=2)

  1      2
0 0.268 0.414
1 0.732 0.586
```

73 % des hommes sont en emploi à taux plein avec un salaire observé, contre 59 % des femmes. On va utiliser comme instruments le nombre d'enfants âgés de moins de 3 ans dans le ménage et le statut matrimonial de l'individu. Avec la fonction `selection` on peut estimer le modèle à la Heckman en 2 étapes ou par maximum de vraisemblance. La méthode à 2 étapes permet de récupérer un estimateur associé à l'inverse du ratio de Mills (mais dans ce cas, on n'a pas les éléments de la matrice de variance-covariance pour σ et ρ).

```

#Estimation du modèle Heckman en 2 étapes, par exemple chez les femmes
base_femmes <- filter(base, sexe==2)
selection1A <- selection(selection = sel ~ exp + exp2 + as.factor(ancentr4)
                           + as.factor(ddipl) + as.factor(nbenfa3)
                           + as.factor(matri), outcome = logsal ~ exp + exp2
                           + as.factor(ddipl) + as.factor(ancentr4),
                           data = base_femmes, method = "2step")
selection1A$coefficients

  (Intercept)           exp           exp2
-0.623713      -0.024922      0.000361
as.factor(ancentr4)1 as.factor(ancentr4)2 as.factor(ancentr4)3
1.266105       1.448127       1.599688
as.factor(ancentr4)4 as.factor(ddipl)3 as.factor(ddipl)4
1.845301       -0.106833      -0.159751
as.factor(ddipl)5 as.factor(ddipl)6 as.factor(ddipl)7
-0.193139       -0.248390      -0.363967
as.factor(nbenfa3)1 as.factor(nbenfa3)2 as.factor(nbenfa3)3
-0.332363       -0.572790      -0.522045
as.factor(nbenfa3)4 as.factor(matri)2 as.factor(matri)3
-0.922225       -0.291307      -0.269150
as.factor(matri)4 (Intercept)           exp
-0.009377       7.228191       0.013287
exp2            as.factor(ddipl)3 as.factor(ddipl)4
-0.000145       -0.212409      -0.389859
as.factor(ddipl)5 as.factor(ddipl)6 as.factor(ddipl)7
-0.519521       -0.512748      -0.665901
as.factor(ancentr4)1 as.factor(ancentr4)2 as.factor(ancentr4)3
0.147291        0.238954      0.289123
as.factor(ancentr4)4           invMillsRatio sigma
0.385394        0.043569      0.321535
rho
0.135504

```

En procédant à la main, on peut cependant facilement récupérer la composante des écarts de salaire liée à des différences de mécanismes de sélection entre les deux groupes. On estime le modèle probit, puis on calcule l'inverse du ratio de Mills qu'on intègre ensuite au modèle de salaire (chez les participants).

```

#On estime un probit pour la participation
seleqn1_A <- glm(sel ~ exp + exp2 + as.factor(ancentr4) + as.factor(ddipl)
+ as.factor(nbenfa3) + as.factor(matri),
family=binomial(link="probit"), data=base_femmes)

#On calcule l'inverse du ratio de Mills (ratio de la densite et de
#la fdr de la loi normale standard)
base_femmes$IMR <- dnorm(seleqn1_A$linear.predictors)/
pnorm(seleqn1_A$linear.predictors)

#On introduit l'inverse du ratio de Mills dans l'équation de salaire
#(estimée chez les participants)
outeqn1_A <- lm(logsal ~ exp + exp2 + as.factor(ddipl) + as.factor(ancentr4)
+ IMR, data=filter(base_femmes, sel==1))
coeff_IMR_A <- outeqn1_A$coefficients[
which(names(outeqn1_A$coefficients)=="IMR")]
coeff_IMR_A

IMR
0.0436

#idem chez les hommes
base_hommes <- filter(base, sexe==1)
seleqn1_B <- glm(sel ~ exp + exp2 + as.factor(ancentr4) + as.factor(ddipl)
+ as.factor(nbenfa3) + as.factor(matri),
family=binomial(link="probit"), data=base_hommes)
base_hommes$IMR <- dnorm(seleqn1_B$linear.predictors)/
pnorm(seleqn1_B$linear.predictors)
outeqn1_B <- lm(logsal ~ exp + exp2 + as.factor(ddipl) + as.factor(ancentr4)
+ IMR, data=filter(base_hommes, sel==1))
coeff_IMR_B <- outeqn1_B$coefficients[
which(names(outeqn1_B$coefficients)=="IMR")]
coeff_IMR_B

IMR
-0.828

```

De là, on peut calculer la contribution des termes de correction de la sélection à l'écart total de salaire observé entre les deux groupes :

```

#Ecart de salaire entre hommes et femmes sans prise en compte de la sélection
brut <- mean(base_hommes$logsal, na.rm=TRUE)-
mean(base_femmes$logsal, na.rm=TRUE)
brut

[1] 0.285

#Contribution des termes de correction de la sélection à l'écart de salaire
#observée entre H et F
effet_select <- coeff_IMR_B*mean(base_hommes$IMR, na.rm=TRUE)-
coeff_IMR_A*mean(base_femmes$IMR, na.rm=TRUE)
effet_select

IMR
-0.419

#Ecart de salaire entre hommes et femmes purgé des effets de la sélection
net <- brut - effet_select
net

IMR
0.705

```

On trouve une contribution négative très forte de la sélection aux écarts de salaire, autrement dit, les écarts de salaire entre hommes et femmes seraient nettement plus élevés si les femmes ne faisaient pas face à des barrières à l'entrée dans l'emploi à temps plein aussi élevées. Ainsi, l'écart de salaire entre hommes et femmes purgé des effets de sélection s'élèverait à environ 70 % d'après notre modélisation, au lieu de 29 % observé.

5.2 Traiter la sélection avec des données de panel

Quand on dispose de données de panel, la sélection endogène peut être traitée à l'aide d'effets fixes individuels corrigeant les estimateurs. Si l'hypothèse d'indépendance conditionnelle est vérifiée en introduisant des effets fixes individuels (dit autrement, si toute l'hétérogénéité individuelle est bien captée par ces effets fixes constants dans le temps), on peut retrouver des résultats convergents sans recourir à des techniques instrumentales. Comme évoqué plus haut, les données de panel peuvent aussi être mobilisées pour imputer des salaires manquants aux individus qui ne participent pas au marché du travail. C'est par exemple une des approches utilisées par [Olivetti and Petrongolo \(2008\)](#) pour corriger les écarts médians de salaire entre hommes et femmes. Pour un salarié absent une année du marché du travail, elles imputent un

salaire potentiel par le salaire observé le plus proche dans le temps de cette année manquante. Les auteures remarquent que dans un cadre où ce sont les écarts médians qui sont étudiés, cette approche revient à faire une hypothèse d'invariance des rangs seulement entre le salaire d'un individu et le salaire médian (voir aussi [Blau and Kahn 2006](#)).

5.3 Traiter la sélection sans la modéliser

Dans cette partie, on présente des approches plus agnostiques cherchant à traiter la sélection sans la modéliser en tant que telle.

Identification à l'infini [Heckman \(1990\)](#) et [Chamberlain \(1986\)](#) ont proposé de n'estimer les paramètres d'intérêt que sur le sous-groupe des individus pour lesquels la sélection est négligeable, que ceux-ci proviennent de la population A ou B. Ce sous-groupe est déterminé conditionnellement aux caractéristiques observables. En pratique, on estime deux modèles probit (un pour la population A et un pour la B) du fait d'être en emploi (sélectionné) sur des caractéristiques observables et on rassemble dans le sous-groupe des individus les plus enclins à être en emploi, tous ceux dont la probabilité prédictive d'être en emploi selon leurs caractéristiques observables dépasse un seuil élevé (proche de 1). Il y a évidemment un arbitrage entre la taille du sous-groupe et le niveau de ce seuil. On peut ensuite reconstruire les décompositions pour ce seul sous-groupe. Pour éviter d'avoir à réestimer tous les paramètres des décompositions sur ce seul (petit) groupe, Heckman conseille d'estimer les paramètres de l'équation de salaire pour lesquels la sélection ne joue pas dans une première étape et de ne se concentrer ici que sur les paramètres les plus sensibles. Cette méthode ne requiert pas de condition d'exclusion (puisque sur le sous-groupe d'estimation il n'y a pas par construction de sélection). Elle n'est cependant pas souvent généralisable à d'autres sous-populations. Elle est surtout appliquée pour valider des décompositions alternatives, voir [Mulligan and Rubinstein \(2008\)](#) qui l'appliquent dans le cadre de l'écart de salaire entre les hommes et les femmes aux Etats-Unis, en considérant les groupes de femmes et d'hommes ayant des caractéristiques démographiques associées à des probabilités de travailler à temps plein toute l'année proches de 1.

Traitement de la sélection hétérogène en s'inspirant du LATE [Machado \(2017\)](#) propose une façon souple de corriger de la sélection sans avoir à en estimer la forme, pour peu qu'un instrument binaire soit à disposition. Sa démarche s'inspire du local average treatment effect (LATE) de [Angrist et al. \(1996\)](#). Elle l'applique à l'analyse de l'écart de salaire entre les hommes et les femmes. On suppose que l'on a à disposition un instrument binaire Z , et l'on note E_1 le statut d'emploi potentiel si $Z = 1$; E_0 celui si $Z = 0$. On note Y le salaire potentiel. On va se concentrer sur le groupe des individus qui sont toujours en emploi quelle

que soit la valeur de l'instrument, ie $E_1 = E_0 = 1$ (similaires aux "always-takers" dans la littérature sur le LATE). L'identification repose sur trois hypothèses :

- AI : existence d'un instrument Z tel que (indépendance) $Z \perp (\mathcal{Y}, E_0, E_1)$, et (non trivial) $\Psi = P(E = 1|Z = 1) - P(E = 1|Z = 0) \neq 0$ et $P(E = 1|Z = 1) > 0$ quand $\Psi < 0$, et $P(E = 1|Z = 0) > 0$ quand $\Psi > 0$.
- AII : restriction d'exclusion : $Y = \mathcal{Y}$ si $E = 1$ et $E = ZE_1 + (1 - Z)E_0$
- AIII : monotonicité : soit $E_1 \leq E_0$ pour tout le monde, soit $E_1 \geq E_0$ pour tout le monde (l'instrument agit dans le même sens pour tous le monde).

L'instrument qu'utilise [Machado \(2017\)](#) est la présence d'enfants de moins de 6 ans dans le ménage, supposant donc que pour aucune femme la présence d'un enfant de moins de 6 ans n'entraîne une hausse de la probabilité d'emploi (monotonicité). Sous AI-AIII, les salaires moyens des femmes "toujours en emploi" sont identifiés. En effet, si $E_1 \leq E_0$ ($\Psi < 0$) :

$$\begin{aligned}\mathbb{E}(Y|E = 1, Z = 1) &= \mathbb{E}(\mathcal{Y}|E_1 = 1, Z = 1) && (\text{AII}) \\ &= \mathbb{E}(\mathcal{Y}|E_1 = 1) && (\text{AI}) \\ &= \mathbb{E}(\mathcal{Y}|E_0 = 1, E_1 = 1) && (\text{AIII})\end{aligned}$$

Si $E_1 \geq E_0$ ($\Psi > 0$) :

$$\mathbb{E}(Y|E = 1, Z = 0) = \mathbb{E}(\mathcal{Y}|E_1 = 1, E_0 = 1)$$

L'estimation se fait alors en trois étapes. Dans une première étape, pour chaque cellule $X = x$ on calcule pour les femmes

$$\Psi_{xtG=1} = P(E = 1|Z = 1, X = x, T = t, G = 1) - P(E = 1|Z = 0, X = x, T = t, G = 1),$$

où $G = 1$ pour les femmes et 0 pour les hommes, et T désigne la période. Dans la deuxième étape, on estime sur les hommes et les femmes par OLS l'équation :

$$Y_i = \beta_{0xt} + \beta_{1xt}G_i + \beta_{2xt}G_i * Z_i + u_i.$$

Enfin, l'écart de salaire entre les hommes et les femmes corrigé de la sélection est simplement

$$\Delta(x, t) = \beta_{1xt} + \beta_{2xt}1(\Psi_{xtG=1} < 0)$$

L'avantage principal de cette méthode est qu'on n'impose pas la forme de la sélection, qui peut être positive ou négative dépendant des X . Une fois l'écart de salaire conditionnel à $X = x$ calculé, on peut l'intégrer, sur différentes compositions de X . Pour autant, il ne décrit que le comportement des femmes restant ici en emploi qu'elles aient ou non un enfant de moins

de 6 ans, [Machado \(2017\)](#) montre qu'il s'agit du groupe le plus comparable aux hommes. La méthode peut s'adapter pour tenir compte aussi d'une sélection (potentiellement différente) sur les hommes.

Approche par bornes Cette méthode proposée initialement par [Manski \(1994\)](#) vise à trouver un intervalle encadrant les écarts qui seraient obtenus en l'absence de mécanisme de sélection différencié entre les groupes. Dans l'approche de [Manski \(1994\)](#), on note à nouveau \mathcal{Y} le salaire potentiel, Y le salaire observé (pour ceux qui travaillent) et comme précédemment E vaut 1 si l'individu travaille et 0 sinon. Le sexe est noté G et X désigne les autres caractéristiques observables des individus. L'espérance du salaire s'écrit alors :

$$\begin{aligned}\mathbb{E}(Y \mid X, G) &= \mathbb{E}(Y \mid E = 1, X, G) \times P(E = 1 \mid X, G) \\ &\quad + \underbrace{\mathbb{E}(\mathcal{Y} \mid E = 0, X, G)}_{\text{inobservé}} \times P(E = 0 \mid X, G)\end{aligned}$$

où le deuxième terme correspond à l'effet de la sélection ou de la non participation sur l'espérance de salaire. Ainsi, le salaire moyen observé est la contrepartie empirique de l'espérance de salaire conditionnelle au fait d'être actif en emploi. Dans ce cadre, l'approche de Manski consiste à considérer des bornes du salaire potentiel telles que $\underline{\mathcal{Y}} \leq \mathcal{Y} \leq \bar{\mathcal{Y}}$ pour tout \mathcal{Y} . Alors il vient :

$$\begin{aligned}&\mathbb{E}(Y \mid E = 1, X, G) \times P(E = 1 \mid X, G) + \underline{\mathcal{Y}} P(E = 0 \mid X, G) \\ &\leq \mathbb{E}(Y \mid X, G) \\ &\leq \mathbb{E}(Y \mid E = 1, X, G) \times P(E = 1 \mid X, G) + \bar{\mathcal{Y}} P(E = 0 \mid X, G).\end{aligned}$$

L'utilisation de bornes n'est cependant pas toujours informative faute de suffisamment d'informations sur les mécanismes de sélection à l'œuvre. Cet ensemble de méthodes nécessite de faire des hypothèses sur la distribution du Y (observé) dans le groupe sélectionné et celle du \mathcal{Y} (potentiel) dans le groupe non sélectionné, ce qui revient souvent à supposer le signe du mécanisme de sélection (c'est-à-dire le signe de la corrélation entre caractéristiques inobservées et probabilité d'être sélectionné).

Un exemple d'application peut être trouvé dans [Blundell et al. \(2007\)](#) qui étudient les évolutions de salaire médian des femmes et des hommes au Royaume-Uni. Ils supposent soit (i) la dominance stochastique d'ordre 1 de la distribution observée sur la distribution potentielle, soit (ii) la supériorité de la médiane du salaire observé sur celle du salaire potentiel. Ils considèrent donc une sélection positive en emploi : les femmes ont d'autant plus de chances d'être en emploi qu'elles ont une productivité inobservée élevée. Cette méthode a l'avantage de pouvoir

s'adapter à différentes statistiques ou moments de la distribution d'intérêt, quitte à adapter les hypothèses formulées. Elle s'adapte d'ailleurs mieux pour la médiane ou les quantiles que pour la moyenne, car seule la position relative par rapport à la médiane (ou aux quantiles) compte et non la valeur précise de la borne. D'autres extensions importantes reposant sur l'idée de recourir à des bornes ont été proposées par [Mulligan and Rubinstein \(2008\)](#) et [Olivetti and Petrongolo \(2008\)](#).

Encadré 8 : Tenir compte de la sélection dans les approches distributionnelles

Le traitement de la sélection dans les décompositions distributionnelles est encore une voie active de recherche. Il n'y a pas à ce jour de méthode unanimement préconisée par la communauté académique, même si l'approche proposée très récemment par [Arellano and Bonhomme \(2017b\)](#) est très prometteuse. On présente ici deux exemples de traitement sans chercher à être exhaustif ni à entrer dans des complexités calculatoires (voir [Arellano and Bonhomme \(2017a\)](#) pour une revue de littérature).

Le premier exemple suit l'approche traditionnelle de traitement de la sélection à la [Heckman \(1990\)](#). [Albrecht et al. \(2009\)](#) analysent ainsi les écarts de salaire selon le sexe en s'appuyant sur la décomposition par régressions quantiles conditionnelles de [Machado and Mata \(2005\)](#) tout en intégrant dans les régressions quantiles conditionnelles une fonction de contrôle pour capturer les effets de la sélection sur le marché du travail selon l'approche proposée par [Buchinsky \(1998\)](#). Cette fonction de contrôle joue le même rôle que le ratio de Mills dans l'approche de [Heckman \(1990\)](#). Elle dépend d'un index ($Z\gamma$, c'est-à-dire d'un score synthétique où interviennent des variables corrélées avec la participation mais pas avec le salaire), varie selon l'ordre du quantile et s'estime de manière semi-paramétrique. Et surtout, son effet s'ajoute (hypothèse d'additivité) aux effets des autres explicatives. Cette hypothèse que la sélection puisse être entièrement captée par un terme additif est cruciale pour l'identification des paramètres. Pourtant dans un modèle plus général, cette hypothèse ne tient pas. C'est notamment le cas dans le modèle de sélection sur les quantiles (*quantile selection model*) qui combine une spécification linéaire sur les quantiles de la variable outcome, dont Y^* est la variable latente, et un mécanisme gaussien de sélection :

$$\begin{aligned} Y^* &= X'\beta(U) \\ D &= \mathbf{1}_{\eta \leq Z'\gamma} = \mathbf{1}_{V \leq \Phi(Z'\gamma)}, \end{aligned}$$

où U est uniformément distribué sur $[0, 1]$ et correspond à l'ordre du quantile ; X est un

sous-ensemble de Z ; η est indépendant de Z , potentiellement corrélé à U et suit une distribution gaussienne; V qui correspond au rang de η est uniformément distribué sur $[0, 1]$.

La deuxième méthode de correction de la sélection dans les approches distributionnelles est celle d'[Arellano and Bonhomme \(2017b\)](#), qui proposent de corriger la sélection de manière non additive en recourant à une modélisation de la copule qui caractérise la dépendance entre U et V . En effet,

$$Pr(Y^* \leq x'\beta(\tau) | D = 1, Z = z) = Pr(U \leq \tau | V \leq \Phi(z'\gamma), Z = z) = G(\tau, \Phi(z'\gamma); \rho) = \tau^*(z), \quad (10)$$

où $G(\tau, \Phi(z'\gamma); \rho)$ est la copule gaussienne conditionnelle de paramètre ρ . En l'absence de sélection, on aurait $Pr(Y \leq x'\beta(\tau)) = \tau$, remplacé ici par τ^* . Le rang du τ ième quantile n'est plus τ dans l'échantillon sélectionné, mais τ^* et l'écart entre τ et τ^* dépend de l'ampleur de l'effet de la sélection. [Arellano and Bonhomme \(2017b\)](#) proposent des estimateurs corrigés de la sélection en translatant les ordres des quantiles de l'ampleur de cet écart (*rotated quantile regressions*). Leur approche s'implémente en pratique, en trois étapes : (1) estimation par probit de $Z'\gamma$; (2) estimation de ρ le paramètre de la copule; (3) estimation des *rotated quantile regressions*. Elle peut se combiner à des décompositions distributionnelles pour reconstruire des contrefactuels suivant les approches de [Machado and Mata \(2005\)](#), [Chernozhukov et al. \(2013\)](#). [Maasoumi and Wang \(2018\)](#) l'applique à l'analyse de l'écart de salaire entre femmes et hommes aux Etats-Unis.

Enfin les méthodes de bornes peuvent s'adapter à d'autres quantiles d'intérêt que la médiane ([Blundell et al., 2007](#)) et les méthodes d'identification à l'infini peuvent aussi s'appliquer pour identifier les paramètres nécessaires à reconstruire une distribution conditionnelle. Cependant, les limites de ces méthodes rapidement présentées dans le texte, à savoir la difficulté à définir le groupe non affecté par la sélection, les bornes non informatives ou recourant à des hypothèses non réalistes, s'appliquent tout autant au cas distributionnel.

Références

- John Abowd, Francis Kramarz, and David Margolis. High wage workers and high wage firms. *Econometrica*, 67(2) :251–333, 1999.
- Romain Aeberhardt, Denis Fougère, Julien Pouget, and Roland Rathelot. Wages and employment of french workers with african origin. *Journal of Population Economics*, 23(3) : 881–905, Jun 2010a.
- Romain Aeberhardt, Denis Fougère, Julien Pouget, and Roland Rathelot. L’emploi et les salaires des enfants d’immigrés. *Économie et Statistique*, 433(1) :31–46, 2010b.
- James Albrecht, Aico van Vuuren, and Susan Vroman. Counterfactual distributions with sample selection adjustments : Econometric theory and an application to the netherlands. *Labour Economics*, 16(4) :383–396, 2009.
- Joseph G. Altonji, Prashant Bharadwaj, and Fabian Lange. Changes in the Characteristics of American Youth : Implications for Adult Outcomes. *Journal of Labor Economics*, 30(4) : 783 – 828, 2012.
- J. Angrist, G. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434) :444–454, 1996.
- M. Arellano and S. Bonhomme. Sample Selection in Quantile Regression : A Survey. Working Papers 1702, CEMFI, 2017a.
- Manuel Arellano and Stéphane Bonhomme. Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, 85(1) :1–28, 1 2017b.
- David Audenaert, José Bardaji, Raphaël Lardeux, Michaël Orand, and Michaël Sicsic. La résistance des salaires depuis la grande récession s’explique-t-elle par des rigidités à la baisse ? *Insee Références, L’économie française - Comptes et dossiers*, 2014.
- Christophe Bertran. Le revenu d’activité des non-salariés : plus élevé en moyenne dans les départements du nord que dans ceux du sud. *Insee Première*, (1672), 2017.
- Francine D. Blau and Lawrence M. Kahn. The u.s. gender pay gap in the 1990s : Slowing convergence. *Industrial and Labor Relations Review*, 60(1) :45–66, 2006.
- Alan Blinder. Wage discrimination : reduced form and structural estimates. *Journal of Human resources*, (1672), 1973.
- Richard Blundell, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir. Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75(2) :323–363, 2007.

- Moshe Buchinsky. The dynamics of changes in the female wage distribution in the usa : A quantile regression approach. *Journal of Applied Econometrics*, 13(1) :1–30, 1998.
- David Card, Ana Cardoso, and Patrick Kline. Bargaining, sorting, and the gender wage gap : Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, 131(2) :633–686, 2016.
- Gary Chamberlain. Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics*, 32(2) :189 – 218, 1986.
- Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6) :2205–2268, 2013.
- Daniel Chiquiar and Gordon H. Hanson. *Journal of Political Economy*, (2) :239–281.
- Jeremiah Cotton. On the decomposition of wage differentials. *The review of economics and statistics*, pages 236–243, 1988.
- John DiNardo, Nicole Fortin, and Thomas Lemieux. Labour market institutions and the distribution of wages, 1973-1992 : a semi parametric approach. *Econometrica*, 64(5) :1001, 1996.
- Chloé Duvivier, Joseph Lanfranchi, and Mathieu Narcy. Les sources de l'écart de rémunération entre femmes et hommes dans la fonction publique. *Economie et Statistique*, 488(1) :123–150, 2016.
- Robert W. Fairlie. An extension of the blinder-oaxaca decomposition technique to logit and probit models. *Journal of economic and social measurement*, 30(4) :305–316, 2005.
- Sergio Firpo, Nicole M. Fortin, and Thomas Lemieux. Decomposing Wage Distributions using Recentered Influence Functions Regressions. mimeo, University of British Columbia, 2007.
- Sergio Firpo, Nicole M. Fortin, and Thomas Lemieux. Unconditional Quantile Regressions. *Econometrica*, 77(3) :953–973, 05 2009.
- Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. *Handbook of labor economics*, 4 :1–102, 2011.
- Javier Gardeazabal and Arantza Ugidos. More on identification in detailed wage decompositions. *Review of Economics and Statistics*, 86(4) :1034–1036, 2004.
- James Heckman. Varieties of selection bias. *American Economic Review*, 80(2) :313–18, 1990.
- Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4) :1161–1189, 07 2003.

- Marek Hlavac. Oaxaca : Blinder-oaxaca decomposition in r. 2014.
- Dean R Hyslop. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, 67(6) :1255–1294, 1999.
- Ben Jann. The blinder-oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4) :453–479, 2008.
- Frank L. Jones and Jonathan Kelley. Decomposing differences between groups a cautionary note on measuring discrimination. *Sociological Methods & Research*, 12(3) :323–343, 1984.
- Claire Kubrak. Principe et mise en oeuvre des approches comptable et économétrique. *Document de travail Insee-Direction de la Diffusion et de l'Action régionale*, H 2018/01, 2018.
- Rasmus Lentz and Dale T. Mortensen. Labor market models of worker and firm heterogeneity. *Annual Review of Economics*, 2(1) :577–602, 2010.
- E. Maasoumi and L. Wang. The gender gap in the earnings distribution. *Journal of Political Economy*, 2018. forthcoming.
- Cecilia Machado. Unobserved selection heterogeneity and the gender wage gap. *Journal of Applied Econometrics*, 32(7) :1348–1366, 2017.
- José Machado and José Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, 20(4) :445–465, 2005.
- Charles Manski. The selection problem. 1 :143–70, 1994.
- Blaise Melly. Decomposition of differences in distribution using quantile regression. *Labour economics*, 12(4) :577–590, 2005.
- Thomas A Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, pages 765–799, 1987.
- Casey B. Mulligan and Yona Rubinstein. Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, 123(3) :1061–1110, 2008.
- Derek Neal. The measured black-white wage gap among women is too small. *Journal of Political Economy*, 112(S1) :S1–S28, 2004.
- Shoshana Neuman and Ronald L. Oaxaca. Wage decompositions with selectivity-corrected wage equations : A methodological note. *Journal of Economic Inequality*, 2(1) :3–10, 2004.
- Shoshana Neuman and Ronald L. Oaxaca. Wage differentials in the 1990s in israel : endowments, discrimination, and selectivity. *International Journal of Manpower*, 26(3) :217–236, 2005.

- David Neumark. Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources*, 23(3) :279–295, 1988.
- Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, pages 693–709, 1973.
- Ronald Oaxaca and Michael Ransom. Identification in detailed wage decompositions. *Review of Economics and Statistics*, 81(1) :154–157, 1999.
- Claudia Olivetti and Barbara Petrongolo. Unequal pay or unequal employment ? a cross-country analysis of gender gaps. *Journal of Labor Economics*, 26(4) :621–654, 2008.
- Cordelia Reimers. Labor market discrimination against hispanic and black men. *The review of economics and statistics*, pages 570–579, 1983.
- Francis Vella. Estimating models with sample selection bias : a survey. *Journal of Human Resources*, pages 127–169, 1998.
- Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.
- Myeong-Su Yun. Decomposing differences in the first moment. *Economics letters*, 82(2) :275–280, 2004.
- Myeong-Su Yun. A simple solution to the identification problem in detailed wage decompositions. *Economic inquiry*, 43(4) :766–772, 2005.
- Myeong-Su Yun. Identification problem and detailed oaxaca decomposition : A general solution and inference. *Journal of economic and social measurement*, 33(1) :27–38, 2008.