Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Décomposition des inégalités salariales selon le sexe par calage : application à l'Enquête suisse sur la structure des salaires

par Mihaela-Catalina Anastasiade et Yves Tillé

Date de diffusion : le 21 décembre 2017

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 231-254
Statistique Canada, nº 12-001-X au catalogue

Décomposition des inégalités salariales selon le sexe par calage : application à l'Enquête suisse sur la structure des salaires

Mihaela-Catalina Anastasiade et Yves Tillé[1]

Résumé

L'article propose une nouvelle approche de décomposition de l'écart salarial entre les hommes et les femmes fondée sur une procédure de calage. Cette approche généralise deux méthodes de décomposition courantes, qui sont réexprimées en se servant des poids de sondage. La première est la méthode de Blinder-Oaxaca et la seconde est une méthode de repondération proposée par DiNardo, Fortin et Lemieux. La nouvelle approche offre un système de pondération qui nous permet d'estimer des paramètres d'intérêt tels que les quantiles. Une application aux données de l'Enquête suisse sur la structure des salaires illustre l'intérêt de cette approche.

Mots-clés : Blinder-Oaxaca; discrimination salariale selon le sexe; quantiles; repondération; salaires.

1 Introduction

La discrimination salariale peut être fondée sur différents critères, dont le sexe, la race ou la religion. La discrimination salariale selon le sexe a lieu quand un homme et une femme sont rémunérés différemment pour un travail qui requiert les mêmes qualifications ou qui implique une productivité identique (voir, par exemple, Neumark, 1988; Gardeazabal et Ugidos, 2005). Puisqu'il faut quantifier la discrimination pour en évaluer l'importance, le sujet a suscité l'intérêt des statisticiens. La technique originale proposée par Blinder (1973) et Oaxaca (1973) consiste à estimer la part de l'écart entre les salaires moyens des hommes et des femmes qui est due à la discrimination. Cependant, en général, la répartition des hommes et des femmes entre les emplois n'est pas uniforme (voir, par exemple Bielby et Baron, 1986). Si les membres d'un de ces deux groupes, habituellement les femmes, sont concentrés dans les emplois faiblement rémunérés, l'écart entre les salaires moyens pourrait ne pas être très pertinent. Donc, au lieu d'analyser le niveau de discrimination dans les salaires moyens, il serait peut-être intéressant de voir si la discrimination se produit uniformément dans tous les types d'emplois. Une bibliographie détaillée des différents articles statistiques consacrés à l'estimation de la discrimination peut être consultée dans Fortin, Lemieux et Firpo (2011).

Bien que les méthodes de décomposition offertes dans la littérature soient nombreuses, deux d'entre elles seulement seront discutées dans le présent article. Ces deux méthodes ne sont pas présentées dans leur forme originale, mais plutôt en tenant compte des poids de sondage. Il s'agit de la méthode de Blinder-Oaxaca (ci-après nommée BO) et la méthode semi-paramétrique élaborée par DiNardo, Fortin et Lemieux (1996) (ci-après nommée DFL). La méthode BO originale analyse la différence entre les salaires moyens des hommes et les salaires moyens des femmes. Cependant, elle ne permet pas d'analyser l'écart salarial pour d'autres paramètres, comme les quantiles. La méthode DFL originale résout ce problème. Son point de départ est un modèle logistique dans lequel, pour chaque observation, la probabilité d'être un homme ou une femme est modélisée comme une fonction des caractéristiques observées. Le ratio de ces probabilités est utilisé pour

1. Mihaela-Catalina Anastasiade et Yves Tillé, Institut de statistique, Université de Neuchâtel, 51 Avenue de Bellevaux, 2000 Neuchâtel, Suisse. Courriel : mihaela.anastasiade@unine.ch; yves.tille@unine.ch.

construire un facteur de repondération. L'objectif de la méthode est de rapprocher la distribution des caractéristiques des femmes de la distribution des caractéristiques des hommes. En disposant de distributions similaires des caractéristiques, il est possible d'obtenir une estimation du niveau de discrimination pour d'autres paramètres que la moyenne. Cependant, la variance du facteur de repondération peut être grande dans les cas où une ou plusieurs caractéristiques sont de bons prédicteurs du sexe. En outre, la distribution repondérée des caractéristiques des femmes peut ne pas concorder avec la distribution des caractéristiques des hommes. Nous abordons les problèmes associés aux deux méthodes par une approche de calage. L'idée qui sous-tend le calage est la même que celle qui sous-tend la méthode DFL. Elle consiste à rapprocher la distribution des caractéristiques des femmes de celle des hommes, afin d'estimer le niveau de discrimination tout le long de la distribution des salaires.

La présentation de l'article est la suivante. À la section 2, nous définissons la notation et à la section 3, nous réexprimons la décomposition BO en nous servant de données d'enquête. Les poids de sondage sont pris en compte afin de corriger l'écart entre l'échantillon et la population d'intérêt. Par conséquent, la décomposition sera nommée « BO pondérée ». Nous présentons aussi le concept clé de la distribution contrefactuelle des salaires des femmes. Elle est définie comme étant la distribution des salaires des femmes si celles-ci avaient les mêmes caractéristiques que les hommes. Ensuite, nous discutons de l'utilisation de la distribution contrefactuelle des salaires dans la décomposition de l'écart salarial. À la section 4, nous développons la méthode DFL, de nouveau en utilisant les poids de sondage. Puisque la méthode originale n'inclut pas ces poids, nous nommons cette nouvelle méthode « DFL pondérée ». Ensuite, à la section 5, nous proposons une nouvelle approche pour calculer la distribution contrefactuelle des salaires, en utilisant la méthode de calage (Deville et Särndal, 1992). Nous discutons de deux cas particuliers de calage, à savoir le calage linéaire et le calage par raking ratio (ou ratissage croisé). Le premier cas produit le même résultat que la méthode BO pondérée pour les salaires moyens. Le deuxième cas présente une approche similaire à la méthode DFL pondérée, mais sans émettre l'hypothèse d'un modèle logistique. Autrement dit, la technique proposée peut être considérée comme une généralisation des deux méthodes susmentionnées. La section 6 comprend un aperçu de l'ensemble de données utilisé, ainsi que les statistiques descriptives sur les salaires observés. Nous donnons aussi une brève description du modèle utilisé et des résultats obtenus en appliquant les méthodes discutées. Enfin, à la section 7, nous résumons les conclusions et à l'annexe B, nous présentons le calcul de la variance du salaire contrefactuel.

2 Problème et notation

La question d'intérêt est l'estimation des écarts salariaux entre les hommes et les femmes, plus précisément, quelle part de ces écarts est attribuable à la discrimination. Supposons qu'il existe une population finie U de taille N pouvant être divisée en deux sous-populations, les femmes (F) et les hommes (M), et que nous désignons par $U_h, h \in \{F, M\}$, de taille N_h. En outre, nous tirons de U un échantillon aléatoire S qui contient à la fois des hommes et des femmes. L'échantillon S est sélectionné selon un plan d'échantillonnage $p(s) = \Pr(S = s)$ pour tout $s \subset U$, où

$$p(s) \geq 0 \quad \text{et} \quad \sum_{s \subset U} p(s) = 1.$$

L'échantillon S peut être divisé en deux sous-échantillons, $S_h, h \in \{F, M\}$, de femmes et d'hommes, tels que $S = \cup S_h$. La variable d'intérêt, désignée par y, est ici le logarithme du salaire. Les totaux de la variable d'intérêt dans les deux sous-populations sont donnés par

$$Y_h = \sum_{k \in U_h} y_k, h \in \{F, M\},$$

où y_k est le logarithme du salaire du k^{e} individu. Puisque l'on n'observe pas toutes les unités des sous-populations, les totaux peuvent être estimés par

$$\hat{Y}_h = \sum_{k \in S_h} d_k y_k, h \in \{F, M\},$$

où d_k est un poids de sondage affecté à la k^{e} unité de l'échantillon. Les poids de sondage sont obtenus après plusieurs traitements statistiques (par exemple, ajustement pour la non-réponse).

Les moyennes de population des logarithmes des salaires sont données par

$$\bar{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, h \in \{F, M\},$$

et peuvent être estimées par

$$\hat{\bar{Y}}_h = \frac{\sum_{k \in S_h} d_k y_k}{\sum_{k \in S_h} d_k}, h \in \{F, M\}.$$

En outre, supposons que, pour chaque k^{e} individu dans l'un ou l'autre des deux sous-échantillons, il existe un vecteur de p variables auxiliaires désigné par

$$\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kj}, \ldots, x_{kp}\right)^\top \in \mathbb{R}^p.$$

Ce vecteur est supposé connu pour chaque unité sélectionnée dans l'échantillon. Les variables auxiliaires contiennent certaines caractéristiques de l'individu, par exemple l'âge, le niveau d'études ou le niveau d'ancienneté. Il peut exister des variables qualitatives ou quantitatives, de sorte que x_{kj} peut être une variable catégorielle ou une quantité. En outre, supposons que la première variable auxiliaire est une constante, c'est-à-dire $x_{k1} = 1,$ pour tout $k \in U$.

Les totaux de ces variables auxiliaires au niveau de la sous-population sont donnés par

$$\mathbf{X}_h = \sum_{k \in U_h} x_k, h \in \{F, M\}.$$

En utilisant les poids d_k définis plus haut, ces deux totaux peuvent être estimés par

$$\hat{\mathbf{X}}_h = \sum_{k \in S_h} d_k \mathbf{x}_k, h \in \{F, M\}.$$

Les vecteurs des valeurs moyennes peuvent être estimés de manière analogue. Les valeurs moyennes au niveau des sous-populations sont données par

$$\bar{\mathbf{X}}_h = \frac{1}{N_h} \sum_{k \in U_h} \mathbf{x}_k, h \in \{F, M\},$$

et estimées par

$$\hat{\bar{\mathbf{X}}}_h = \frac{\sum_{k\in S_h} d_k \mathbf{x}_k}{\sum_{k\in S_h} d_k}, h \in \{F, M\}. \tag{2.1}$$

3 La décomposition BO pondérée

3.1 La décomposition

Partant des conditions établies à la section 2, nous résumons les constatations de Blinder (1973) et d'Oaxaca (1973) dans le contexte de la théorie du sondage, à savoir en utilisant les poids de sondage. Supposons que, dans chaque échantillon, une relation linéaire entre les p caractéristiques disponibles et le logarithme du salaire est appropriée. Une régression est effectuée séparément dans chaque sous-population $U_h, h = \{M, F\}$. Au niveau de la sous-population, les valeurs des coefficients de régression sont données par

$$\boldsymbol{\beta}_h = \left(\sum_{k\in U_h} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k\in U_h} \mathbf{x}_k y_k.$$

Elles peuvent être estimées d'après l'échantillon par

$$\hat{\boldsymbol{\beta}}_h = \left(\sum_{k\in S_h} d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k\in S_h} d_k \mathbf{x}_k y_k, \tag{3.1}$$

où d_k désigne les poids de sondage. Les coefficients de régression $\hat{\boldsymbol{\beta}}_h$, que nous appelons structure salariale de groupe ou rendement des caractéristiques, représentent la contribution de chaque caractéristique au salaire.

Résultat 1 *Une condition suffisante pour obtenir les égalités suivantes*

$$\bar{Y}_h = \bar{\mathbf{X}}_h^\top \boldsymbol{\beta}_h \quad et \quad \hat{\bar{Y}}_h = \hat{\bar{\mathbf{X}}}_h^\top \hat{\boldsymbol{\beta}}_h$$

est qu'il existe un vecteur $\boldsymbol{\varsigma} \in \mathbb{R}^p$, *tel que* $\boldsymbol{\varsigma}^\top \mathbf{x}_k = 1$, *pour tout* $k \in U_h$.

Puisque nous supposons que $x_{k1} = 1$ pour tout $k \in U$, avec $\boldsymbol{\varsigma}^\top = (1\ 0 \ldots 0)$, l'égalité est toujours vérifiée. La preuve du résultat 1 figure à l'annexe A. En regroupant le résultat susmentionné et les équations (2.1) et (3.1), l'écart moyen entre les salaires des deux groupes peut s'écrire

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right). \tag{3.2}$$

L'écart entre les salaires moyens des groupes contient deux éléments, à savoir une partie expliquée, également appelée *effet de composition* $\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F$ et une partie inexpliquée, ou *effet de structure* $\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right)$. Le premier effet englobe les différences de caractéristiques entre les deux groupes. Le second représente la différence de rendement des caractéristiques entre les deux groupes, soit la partie qui n'est pas attribuable à des facteurs objectifs (Oaxaca, 1973; Blinder, 1973). On l'obtient en utilisant les caractéristiques comme mesure de substitution de la productivité. L'estimation de l'*effet de structure* est l'élément central du présent article. L'équation (3.2) contient les mêmes éléments que celle proposée par Oaxaca (1973) et Blinder (1973). La méthodologie appliquée pour estimer les valeurs moyennes et les

coefficients diffère de la technique de régression classique. La méthode BO utilise les coefficients de régression estimés obtenus par les moindres carrés ordinaires (MCO) et les vecteurs des valeurs moyennes des variables explicatives observées. L'approche proposée tient compte des poids de sondage. Cependant, la méthode BO pondérée est la même que la méthode BO originale si les poids de sondage valent tous 1.

3.2 Une note sur l'effet de structure

Les deux éléments de l'équation 3.2 ont des noms différents dans les divers textes publiés. Le premier effet, pour lequel nous avons retenu le nom d'*effet de composition* est également nommé *effet de dotation*. Le second, que nous appelons *effet de structure* est également trouvé dans la littérature sous les noms de *résidu inexpliqué, effet du prix, effet du sexe, effet calculé* ou *traitement inégal* (Weichselbaumer et Winter-Ebmer, 2006). En utilisant la méthode BO, l'effet de structure est une estimation du niveau de discrimination. Toutefois, la discrimination est un phénomène complexe qui pourrait ne pas être toujours entièrement observé. Les variables inobservables, le biais de sélection ou certains mécanismes du marché du travail peuvent contribuer à l'accroissement de la part inexpliquée de l'écart salarial. En outre, Weichselbaumer et Winter-Ebmer (2005) notent deux problèmes possibles concernant le modèle choisi. Premièrement, si les caractéristiques choisies dans le modèle linéaire sont elles-mêmes sujettes à la discrimination, alors l'effet de structure résultant sera surestimé. Deuxièmement, si les caractéristiques ne représentent pas une mesure appropriée de la productivité, de nouveau, l'effet de structure pourrait être sous- ou surestimé. Weichselbaumer et Winter-Ebmer (2006) mettent en garde au sujet de la légitimité des caractéristiques en tant qu'indicateurs de la productivité, puisque les salaires pourraient aussi être déterminés par le pouvoir de négociation, les différences de rémunération ou les salaires basés sur le rendement. Cependant, pour simplifier, dans la suite, nous supposerons que ce genre de problème n'existe pas et que l'effet de structure estimé est le résultat de la discrimination sur le marché du travail. En outre, nous n'examinons pas le biais de sélection de l'échantillon ni d'autres mécanismes qui sous-tendent la distribution des hommes et des femmes dans certains emplois.

3.3 La distribution contrefactuelle des salaires

En général, la distribution contrefactuelle des salaires est une distribution artificielle obtenue en utilisant les caractéristiques d'un groupe pour estimer les salaires d'un autre groupe (voir, par exemple, Bourguignon, Ferreira et Leite, 2002). Des exemples de distributions contrefactuelles sont donnés dans DiNardo et coll. (1996) ou DiNardo (2002). Le terme $\hat{\bar{\mathbf{X}}}_M \hat{\boldsymbol{\beta}}_F$ qui figure dans l'équation (3.2) est appelé salaire moyen contrefactuel des femmes. Il est interprété comme étant le salaire moyen estimé des femmes si celles-ci avaient les mêmes caractéristiques moyennes que les hommes et que leur rendement des caractéristiques demeurait inchangé. La distribution contrefactuelle des salaires s'obtient en utilisant les caractéristiques des hommes $(\mathbf{X}_M)$ et la structure des salaires des femmes $(\boldsymbol{\beta}_F)$. Pour ce qui est de l'interprétation, il s'agit de la distribution des salaires des femmes, si celles-ci avaient les mêmes caractéristiques que les hommes.

En utilisant le résultat 1 de la section précédente, le salaire moyen contrefactuel des femmes est égal à

$$\bar{Y}_{F|M} = \bar{\mathbf{X}}_M^\top \boldsymbol{\beta}_F,$$

et est estimé d'après l'échantillon par

$$\hat{\bar{Y}}_{F|M} = \hat{\bar{\mathbf{X}}}_M^\top \hat{\boldsymbol{\beta}}_F,$$

où $\hat{\bar{\mathbf{X}}}_M$ est estimé dans l'équation (2.1) et $\hat{\boldsymbol{\beta}}_F$ représente les coefficients estimés au moyen de l'équation (3.1). Selon cette notation, la décomposition BO donnée en (3.2) se réexprime sous la forme

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right). \quad (3.3)$$

3.4 Utilisation de la distribution contrefactuelle pour estimer les effets de composition et de structure

Construire le salaire moyen contrefactuel permet d'estimer les deux effets qui constituent l'écart salarial au niveau moyen. Partant de l'équation (3.3), l'effet de composition est égal à

$$\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right).$$

L'effet de composition peut être interprété comme la différence entre ce que les femmes pourraient gagner, en moyenne, si elles avaient les caractéristiques des hommes et ce qu'elles gagnent effectivement. Donc, il reflète l'inégalité due aux différences de caractéristiques. L'effet de structure dans l'équation (3.3) est égal à

$$\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right).$$

L'effet de structure est la différence entre le salaire moyen réel des hommes et ce que les femmes gagneraient si elles avaient les caractéristiques moyennes des hommes et la structure salariale propre à ceux-ci. Les équations susmentionnées expriment les effets de composition et de structure aux niveaux moyens, puisqu'il s'agit de la limite de la méthode BO. À la section suivante, nous présentons une méthode qui permet de construire la distribution contrefactuelle complète. Cela, à son tour, donne la capacité d'estimer les effets de composition et de structure tout le long de la distribution des salaires.

4 La méthode DFL pondérée

4.1 La méthode

La méthode proposée par DiNardo et coll. (1996) fait appel à une fonction de repondération au moyen de laquelle la distribution des caractéristiques des femmes est rendue similaire à la distribution des caractéristiques des hommes. La distribution repondérée est la distribution contrefactuelle des caractéristiques des femmes. La méthode DFL est présentée en utilisant les poids de sondage, afin de tenir compte du plan de sondage.

La fonction de repondération est égale à

$$\psi(\mathbf{x}_k) = \frac{\Pr(D_{Mk} = 1 | \mathbf{x}_k) / \Pr(D_{Mk} = 1)}{\Pr(D_{Mk} = 0 | \mathbf{x}_k) / \Pr(D_{Mk} = 0)},$$

où $D_{Mk} = 1$ si l'individu k est un homme et $D_{Mk} = 0$ sinon, et $\mathbf{x}_k$ est le vecteur des caractéristiques observées pour l'individu k. Manifestement, $\Pr(D_{Mk} = 1 | \mathbf{x}_k)$ et $\Pr(D_{Mk} = 0 | \mathbf{x}_k)$ doivent être estimées. Pour ce type d'estimation, DiNardo et coll. (1996) ont proposé d'utiliser un modèle logit ou probit. En se servant de l'information provenant de l'échantillon,

$$\hat{\psi}(\mathbf{x}_k) = \frac{\widehat{\Pr}(D_{Mk} = 1 | \mathbf{x}_k) \big/ \widehat{\Pr}(D_{Mk} = 1)}{\widehat{\Pr}(D_{Mk} = 0 | \mathbf{x}_k) \big/ \widehat{\Pr}(D_{Mk} = 0)}. \tag{4.1}$$

En utilisant le facteur de repondération, la moyenne contrefactuelle des salaires des femmes est estimée par

$$\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}, \tag{4.2}$$

et les moyennes contrefactuelles des caractéristiques des femmes sont estimées par

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}. \tag{4.3}$$

Le facteur de repondération estimé défini en l'équation (4.1) sera égal à

$$\hat{\psi}(\mathbf{x}_k) = \hat{a}\ \exp\left(\mathbf{x}_k^{\top} \hat{\boldsymbol{\gamma}}\right),$$

où $\hat{\boldsymbol{\gamma}}$ est l'estimation de $\boldsymbol{\gamma}$ d'après l'échantillon en utilisant la vraisemblance empirique et $\hat{a}$ est le ratio des proportions estimées de femmes et d'hommes. Il est donné par :

$$\hat{a} = \frac{\widehat{\Pr}(D_{Mk} = 0)}{\widehat{\Pr}(D_{Mk} = 1)} = \frac{\sum_{k \in S_F} d_k}{\sum_{k \in S_M} d_k}.$$

Puisque la méthode DFL est présentée en tenant compte des poids de sondage, le facteur de repondération ψ_k sera multiplié par $d_k, k \in S_F$. Nous appellerons ce facteur résultant « facteur DFL pondéré ». La moyenne contrefactuelle des salaires estimée des femmes peut être réexprimée sous la forme

$$\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right)}. \tag{4.4}$$

Les moyennes contrefactuelles des caractéristiques des femmes sont estimées par

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right)}.$$

En utilisant le facteur de repondération, les coefficients contrefactuels dans l'échantillon de femmes sont donnés par

$$\boldsymbol{\beta}_F^{\mathrm{DFL}} = \left(\sum_{k \in U_F} \psi(\mathbf{x}_k)\, \mathbf{x}_k \mathbf{x}_k^{\top} \right)^{-1} \sum_{k \in U_F} \psi(\mathbf{x}_k)\, \mathbf{x}_k y_k,$$

et estimés par

$$\hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} = \left(\sum_{k \in S_F} d_k \hat{\psi}\left(\mathbf{x}_k\right) \mathbf{x}_k \mathbf{x}_k^{\top} \right)^{-1} \sum_{k \in S_F} d_k \hat{\psi}\left(\mathbf{x}_k\right) \mathbf{x}_k y_k . \tag{4.5}$$

Les coefficients susmentionnés doivent être calculés, parce que, sous la même condition que dans le résultat 1, la moyenne contrefactuelle des salaires des femmes définie dans (4.2) est donnée par

$$\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} = \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} .$$

La formule de la décomposition BO peut maintenant être exprimée sous la forme

$$\begin{aligned} \hat{\bar{Y}}_M - \hat{\bar{Y}}_F &= \left(\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} - \hat{\bar{Y}}_F \right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} \right) \\ &= \left(\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} - \hat{\bar{\mathbf{X}}}_F^{\top} \hat{\boldsymbol{\beta}}_F \right) + \left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} \right), \end{aligned} \tag{4.6}$$

où $\hat{\boldsymbol{\beta}}_M$ et $\hat{\boldsymbol{\beta}}_F$ sont définis dans (3.1). Le premier terme de l'équation (4.6) est l'effet de composition et le second, l'effet de structure.

4.2 Décomposition plus poussée de l'effet de structure

Comme le soulignent Fortin et coll. (2011), l'objectif du facteur de repondération DFL est de rendre la distribution des caractéristiques des femmes identique à celle des hommes. Cela implique que les moyennes des variables auxiliaires dans les deux groupes doivent être égales. Or, cela n'est pas le cas pour la méthode DFL. En effet,

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} \neq \hat{\bar{\mathbf{X}}}_M \tag{4.7}$$

(voir, par exemple, Fortin et coll. 2011; Donzé, 2013). Le facteur de repondération n'arrive donc pas à faire concorder parfaitement les deux distributions.

L'effet de structure dans l'équation (4.6) peut être subdivisé en les éléments suivants

$$\left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} \right) = \hat{\bar{\mathbf{X}}}_M^{\top} \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} \right) + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} \right) \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}}, \tag{4.8}$$

où $\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}}$ et $\hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}}$ sont définis dans les équations (4.3) et (4.5), respectivement (Fortin et coll. 2011). Le premier terme du deuxième membre de l'équation (4.8) est l'*effet pur* et le deuxième, l'*effet résiduel* ou *erreur totale de repondération* (Fortin et coll. 2011). L'effet pur est la part inexpliquée réelle de l'écart salarial. L'effet résiduel contient l'inadéquation du modèle, autrement dit, ce que le facteur de repondération n'a pas réussi à faire concorder entre les distributions des caractéristiques des hommes et des femmes. Cette méthode permet de construire une distribution contrefactuelle des salaires. Cette nouvelle distribution peut alors être comparée aux distributions observées des salaires des femmes et des hommes. L'inconvénient de la méthode est qu'il peut arriver qu'au moins une caractéristique soit un bon prédicteur du sexe (par exemple, le secteur économique). Cette situation implique que $\Pr\left(D_{Mk} = 1 \middle| \mathbf{x}_k\right)$ peut s'approcher de 1 et que le facteur de repondération prendra une grande valeur (Fortin et coll. 2011). Cela mène de toute évidence à une grande variance de ce facteur, comme nous le montrerons à la section 8.

5 L'approche du calage

5.1 La méthode de calage

La méthode de calage a été introduite par Deville et Särndal (1992). L'idée qui sous-tend la technique est d'utiliser l'information sur certaines variables auxiliaires disponibles au niveau de la population pour estimer une fonction d'une variable d'intérêt. Habituellement, les variables auxiliaires et la variable d'intérêt sont corrélées. Les estimations résultantes sont convergentes et efficaces.

Sous l'hypothèse que l'on dispose des poids de sondage d_k et que l'on connaît les totaux des données auxiliaires au niveau de la population exprimés par

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k ,$$

de nouveaux poids $w_k , k \in S$ doivent être construits, de manière que la contrainte (ou équation de calage) qui suit soit respectée

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k . \tag{5.1}$$

Les poids sont déterminés en résolvant en $\boldsymbol{\lambda}$ les équations de calage qui deviennent

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in S} d_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k ,$$

où $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right)$ est la fonction de calage. L'estimation par calage résultante de Y est

$$\hat{Y} = \sum_{k \in S} d_k y_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right). \tag{5.2}$$

Dans la partie qui suit, nous utiliserons le cas linéaire, où la pseudo fonction de distance est la distance du khi carré et la fonction de calage est donnée par $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = 1 + \mathbf{x}_k^\top \boldsymbol{\lambda}.$ Dans le second cas, nous appliquerons la méthode du raking ratio, qui utilise la pseudo-distance d'entropie et où la fonction de calage est donnée par $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = \exp \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right)$.

5.2 Calage des caractéristiques des femmes sur les caractéristiques des hommes

Supposons que, pour toutes les unités de l'échantillon, il existe un poids de sondage donné d_k. Dans le présent contexte, les variables auxiliaires utilisées dans le processus de calage sont certaines caractéristiques mesurées pour chaque individu. L'objectif est de « détourner » la technique de calage afin de calculer un système de pondération qui ajuste les totaux des variables auxiliaires des femmes sur les totaux des hommes. La variable d'intérêt est le logarithme du salaire.

Dans l'échantillon de femmes, de nouveaux poids w_k proches de d_k sont calculés, de manière que $\sum_{k \in S_F} G \left(w_k , d_k \right)$ soit minimisée. L'équation de calage qui suit est satisfaite

$$\sum_{k \in S_F} w_k \mathbf{x}_k = \hat{\hat{\mathbf{X}}}_M , \tag{5.3}$$

où le vecteur $\hat{\hat{\mathbf{X}}}_M$ contient les totaux des caractéristiques des hommes ajustés sur le total des poids des femmes divisé par le total des poids des hommes.

$$\hat{\tilde{\mathbf{X}}}_M = \frac{\sum_{k\in S_F} d_k}{\sum_{k\in S_M} d_k} \sum_{k\in S_M} d_k \mathbf{x}_k .$$

La division de l'équation de calage (5.3) par $\sum_{k\in S_F} d_k$ donne

$$\frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} d_k} = \frac{\hat{\tilde{\mathbf{X}}}_M}{\sum_{k\in S_F} d_k} = \hat{\tilde{\mathbf{X}}}_M . \tag{5.4}$$

Donc, avec les nouveaux poids w_k, les nouvelles moyennes des caractéristiques des femmes sont égales à celles des hommes. Une autre égalité intéressante est

$$\sum_{k\in S_F} w_k = \sum_{k\in S_F} d_k , \tag{5.5}$$

qui est vérifiée parce que $x_{k1} = 1, k \in S_M$ et le calage est effectué dessus. Si

$$\hat{\tilde{\tilde{\mathbf{X}}}}_M = \frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} w_k},$$

en regroupant les équations (5.4) et (5.5), cela signifie que

$$\hat{\tilde{\tilde{\mathbf{X}}}}_M = \hat{\tilde{\mathbf{X}}}_M . \tag{5.6}$$

L'estimateur de la moyenne contrefactuelle des salaires des femmes est donc

$$\hat{\tilde{\tilde{Y}}}_{F|M} = \frac{\sum_{k\in S_F} w_k y_k}{\sum_{k\in S_F} d_k} = \frac{\sum_{k\in S_F} w_k y_k}{\sum_{k\in S_F} w_k} .$$

5.3 Calage linéaire

Résultat 2 *La moyenne contrefactuelle des salaires des femmes obtenue en utilisant le calage linéaire est égale à la moyenne contrefactuelle des salaires obtenue en utilisant la méthode BO pondérée, c'est-à-dire* $\hat{\tilde{\tilde{Y}}}_{F|M} = \hat{\tilde{\mathbf{X}}}^\top \hat{\boldsymbol{\beta}}_F$.

Preuve

Afin de déterminer le vecteur $\boldsymbol{\lambda}$ dans le cas où l'on utilise la pseudo-distance du khi carré, l'équation qui suit doit être résolue

$$\begin{aligned}\hat{\tilde{\tilde{\mathbf{X}}}}_M &= \sum_{k\in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k\in S_F} d_k \mathbf{x}_k \left(1 + \mathbf{x}_k^\top \boldsymbol{\lambda}\right) \\ &= \sum_{k\in S_F} d_k \mathbf{x}_k + \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right) \boldsymbol{\lambda}.\end{aligned}$$

Donc,

$$\boldsymbol{\lambda} = \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\hat{\tilde{\tilde{\mathbf{X}}}}_M - \sum_{k\in S_F} d_k \mathbf{x}_k\right) = \mathbf{T}^{-1}\left(\hat{\tilde{\tilde{\mathbf{X}}}}_M - \hat{\mathbf{X}}_F\right),$$

où

$$\mathbf{T} = \sum_{k \in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top .$$

Donc,

$$w_k = d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = d_k \left\{1 + \mathbf{x}_k^\top \mathbf{T}^{-1}\left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)\right\}.$$

En utilisant le résultat de l'équation précédente, le numérateur de l'expression (5.2) devient

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k \\ &= \sum_{k \in S_F} d_k y_k + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k , \end{aligned} \tag{5.7}$$

où $\hat{Y}_{F|M}^{\mathrm{LC}}$ désigne le total des logarithmes des salaires dans l'échantillon de femmes, quand le total est construit en utilisant la pseudo-distance du khi carré. Soit

$$\hat{\boldsymbol{\beta}}_F = \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k .$$

Le vecteur $\hat{\boldsymbol{\beta}}_F$ a déjà été défini de la même façon dans l'équation (3.1) pour la méthode BO pondérée. Nous réécrivons l'équation (5.7) sous la forme

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k y_k + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{Y}_F + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{\bar{\mathbf{X}}}_M^\top \hat{\boldsymbol{\beta}}_F , \end{aligned} \tag{5.8}$$

parce que, sous la condition du résultat 1, $\hat{\mathbf{X}}_F^\top \hat{\boldsymbol{\beta}}_F = \hat{Y}_F$. En divisant (5.8) par $\sum_{k \in S_F} w_k$, nous obtenons le résultat 2.

Si l'on utilise la pseudo-distance du khi carré, les poids résultants n'ont pas de bornes. Cela signifie que les poids de calage pourraient être négatifs. Même si cette approche de calage donne les mêmes résultats que la méthode BO pour les salaires moyens, nous recommandons d'utiliser une approche qui donne des poids non négatifs.

5.4 Calage par raking ratio

La deuxième approche de calage s'appuie sur la pseudo-distance d'entropie. Elle est également appelée calage par « raking ratio ». Si l'on utilise la pseudo-distance d'entropie, l'équation (5.3) devient

$$\hat{\bar{\mathbf{X}}}_M = \sum_{k \in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k \in S_F} d_k \mathbf{x}_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right). \tag{5.9}$$

Ce système d'équations résultant ne peut pas être résolu analytiquement. Cependant, on peut trouver la valeur de $\boldsymbol{\lambda}$ au moyen de l'algorithme de Newton-Raphson.

L'équation (5.2) peut maintenant s'écrire

$$\hat{Y}_{F|M}^{\mathrm{RRC}} = \sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^{\top}\boldsymbol{\lambda}\right) y_k,$$

où $\hat{Y}_{F|M}^{\mathrm{RRC}}$ désigne le total du logarithme du salaire dans l'échantillon de femmes, quand le total est construit en utilisant le calage par raking ratio. La moyenne contrefactuelle des salaires des femmes s'écrit

$$\hat{\bar{Y}}_{F|M}^{\mathrm{RRC}} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^{\top}\boldsymbol{\lambda}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^{\top}\boldsymbol{\lambda}\right)}.$$

L'équation qui précède est très similaire à l'équation (4.4). La seule différence tient à l'estimation des paramètres $\boldsymbol{\lambda}$ et $\boldsymbol{\gamma}$. Le vecteur $\boldsymbol{\lambda}$ contient les multiplicateurs de Lagrange résolvant l'équation 5.9 sous la contrainte (5.1), tandis que l'on trouve le vecteur $\boldsymbol{\gamma}$ par la méthode du maximum de vraisemblance.

Après le calcul des poids de calage w_k définis en (5.3) et en utilisant l'information qui figure dans l'équation (5.6), il résulte que

$$\hat{\bar{\mathbf{X}}}_M = \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{RRC}} = \frac{\sum_{k \in S_F} \mathbf{x}_k w_k}{\sum_{k \in S_F} w_k},$$

qui assure que la part résiduelle de l'effet de structure définie dans l'équation (4.8) sera nulle. Il s'agit d'une solution au problème présenté à la section 4.3. Cette approche de calage remédie aussi au problème des poids négatifs que l'on peut obtenir en utilisant la pseudo-distance du khi carré.

6 Application à l'Enquête suisse sur la structure des salaires

6.1 Description des données

L'ensemble de données utilisé contient des renseignements recueillis en 2008 par l'Office fédéral de la statistique suisse au moyen d'une enquête appelée Enquête suisse sur la structure des salaires. Un questionnaire a été envoyé à des organisations publiques et privées des secteurs secondaire et tertiaire afin de recueillir des renseignements sur des aspects particuliers. Ces aspects comprennent la taille de l'organisation, les types de contrats d'emploi et la rémunération des employés au sein de l'organisation. Le questionnaire a été rempli par un membre autorisé de l'organisation et non par des employés. Cela améliore la fiabilité des données et les rend moins susceptibles aux approximations. Les analyses qui suivent ont été limitées au secteur privé. Les observations valides incluses dans les analyses correspondaient aux individus sans valeurs manquantes, qui avaient travaillé plus d'une heure par semaine et pour lesquels la différence entre l'âge et les années d'expérience de travail était égale ou supérieure à 15 (d'après les lois suisses sur l'emploi, cela représente l'âge minimum légal pour pouvoir travailler). Donc, 29 048 cas ont été exclus de l'ensemble de données original. L'ensemble de données final comptait 647 139 hommes et 435 507 femmes. L'ensemble de données fourni par l'Office fédéral de la statistique contenait aussi les poids de sondage, de sorte qu'aucun traitement ni calcul de ces poids n'a été effectué dans la présente application.

Dans les tableaux qui suivent, les valeurs exprimées en francs suisses sont données entre parenthèses. Cependant, les chiffres sont représentés graphiquement en utilisant les logarithmes des salaires. Les valeurs sont obtenues en tenant compte des poids de sondage.

Le tableau 6.1 contient la médiane et la moyenne salariale pour l'échantillon complet, ainsi que pour les femmes et pour les hommes.

Tableau 6.1
Moyenne et médiane des salaires calculées pour l'ensemble de données complet, les femmes et les hommes, en francs suisses

	Moyenne	Médiane
Ensemble de données complet	6 977	5 905
Femmes	5 843	5 220
Hommes	7 725	6 346

Les valeurs de la moyenne ainsi que de la médiane des salaires des hommes sont supérieures aux valeurs pour l'ensemble de données complet, tandis que les valeurs pour les femmes sont inférieures. Le tableau 6.2 montre la répartition des hommes et des femmes dans les emplois faiblement et fortement rémunérés. Les quantiles pondérés des salaires pour l'ensemble de données complet sont calculés à la première ligne. Les deux lignes suivantes montrent les proportions cumulées de femmes et d'hommes qui gagnent moins que la valeur du quantile.

Tableau 6.2
Quantiles pondérés du logarithme du salaire et proportions de femmes et d'hommes qui gagnent moins que la valeur qui représente un quantile particulier du salaire calculé pour l'ensemble de données complet (les valeurs en francs suisses sont données entre parenthèses)

	Quantile										
	1 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	99 %
Logarithme du salaire	7,89 (2 683)	8,27 (3 897)	8,39 (4 412)	8,50 (4 905)	8,59 (5 400)	8,68 (5 905)	8,78 (6 488)	8,89 (7 233)	9,03 (8 380)	9,27 (10 667)	10,09 (24 202)
Proportion cumulée de femmes	0,02	0,17	0,32	0,43	0,53	0,63	0,72	0,81	0,89	0,96	1
Proportion cumulée d'hommes	0,006	0,06	0,12	0,21	0,31	0,42	0,52	0,63	0,74	0,86	0,99

Alors que 43 % de femmes ont un salaire inférieur à 4 905 CHF (comparativement à seulement 21 % d'hommes), seulement 11 % de femmes ont un salaire compris entre 8 380 CHF et 24 202 CHF (comparativement à 25 % d'hommes). En outre, 63 % de femmes gagnent moins que la valeur médiane des salaires pour l'ensemble complet de données, comparativement à seulement 42 % d'hommes. Les mécanismes possibles à l'origine de cette répartition doivent être étudiés. Néanmoins, ce n'est pas l'objet du présent article. Afin de mieux comprendre la distribution des salaires dans chaque échantillon, le tableau 6.3 montre les quantiles pondérés des logarithmes des salaires des femmes et des hommes, ainsi que l'écart entre eux. Une valeur surprenante de l'écart entre les salaires s'observe au quantile d'ordre 1 %. En principe, ces emplois sont d'un type qui ne requiert ni de grandes qualifications ni un niveau d'études élevé. Alors que seulement 0,6 % d'hommes occupent ce genre d'emplois (voir le tableau 6.3), ils gagnent plus

que les 2 % de femmes ayant des emplois comparables. La figure 6.1 montre les données du tableau 6.3 qui suit sous forme graphique.

Tableau 6.3
Salaires des femmes et des hommes et écart entre les salaires des hommes et des femmes, exprimés en logarithmes (les valeurs en francs suisses sont données entre parenthèses)

	Quantile										
	1 %	**10 %**	**20 %**	**30 %**	**40 %**	**50 %**	**60 %**	**70 %**	**80 %**	**90 %**	**99 %**
Femmes	7,80	8,19	8,30	8,38	8,47	8,56	8,66	8,76	8,88	9,06	9,67
	(2 432)	(3 602)	(4 005)	(4 344)	(4 756)	(5 220)	(5 743)	(6 353)	(7 154)	(8 577)	(15 761)
Hommes	8,01	8,36	8,49	8,58	8,67	8,76	8,86	8,98	9,14	9,38	10,26
	(3 000)	(4 259)	(4 850)	(5 344)	(5 820)	(6 346)	(7 012)	(7 908)	(9 291)	(11 905)	(28 571)
Écart	0,21	0,17	0,19	0,21	0,20	0,20	0,20	0,22	0,26	0,33	0,59
	(568)	(657)	(845)	(1 000)	(1 064)	(1 126)	(1 269)	(1 555)	(2 137)	(3 328)	(12 810)

La distance entre les deux ensembles de points augmente vers les quantiles de niveau plus élevé, ce qui signifie que l'écart entre les salaires devient plus important. Il reste à établir quelle part de ces écarts n'est pas attribuable à des différences de caractéristiques entre les hommes et les femmes. En guise de preuve graphique finale des inégalités salariales, la figure 6.2 montre les distributions des logarithmes des salaires des femmes et des hommes.

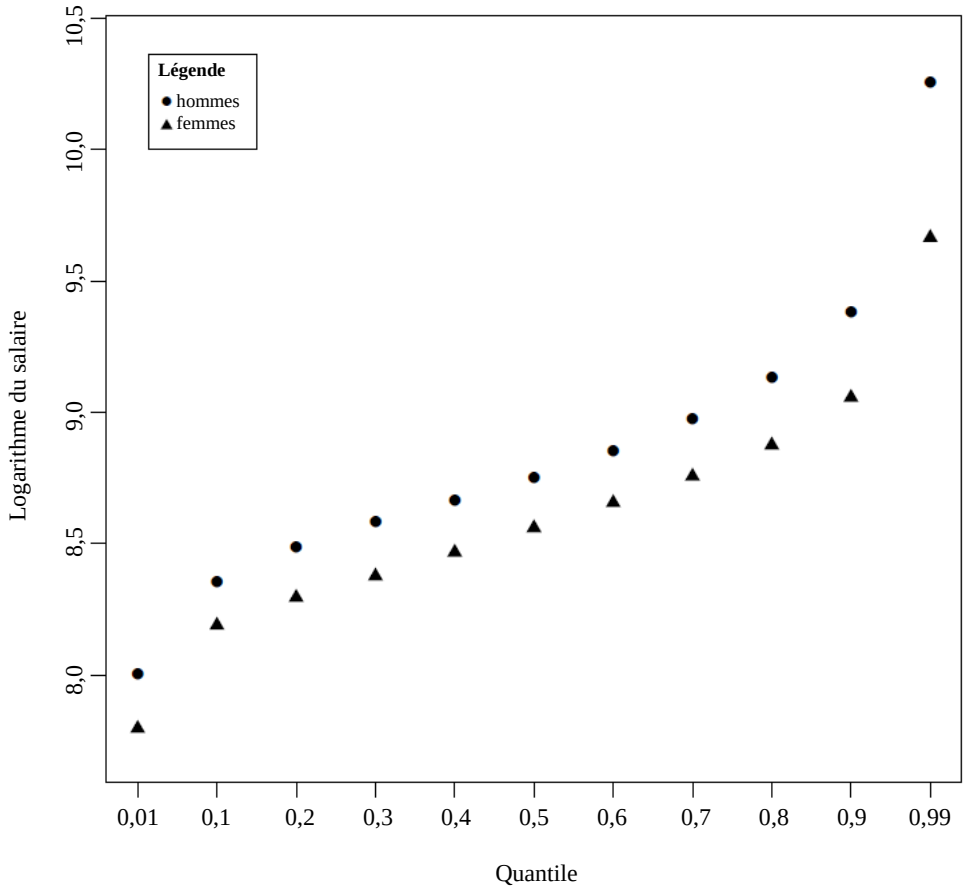

Figure 6.1 Quantiles pondérés du logarithme des salaires des femmes et des hommes.

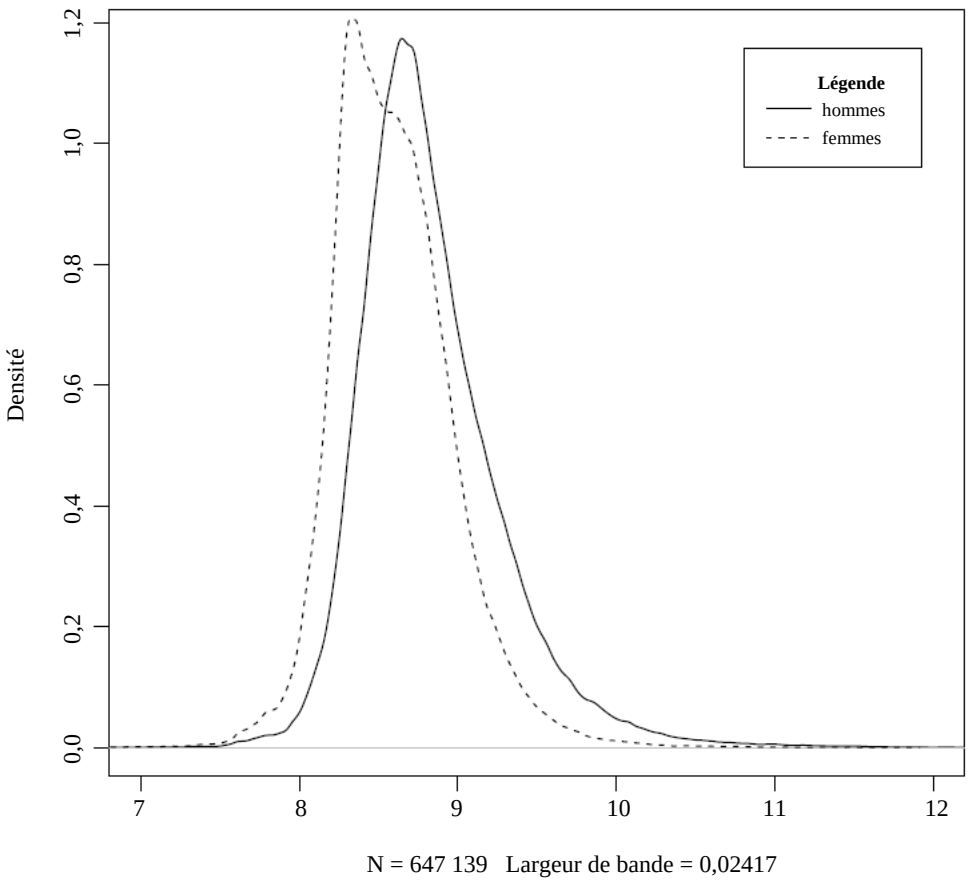

Figure 6.2 Densités estimées des logarithmes des salaires des hommes et des femmes.

6.2 Le modèle

Le modèle de régression contient huit variables explicatives :

- niveau d'études : variable nominale comprenant 9 catégories indiquant le niveau d'études le plus élevé atteint;
- nombre d'années de service dans l'emploi courant (variable de substitution pour l'expérience de travail);
- qualifications requises : variable ordinale comprenant 4 niveaux indiquant le degré de qualification requis pour le poste;
- région de l'institution : variable nominale comptant 7 catégories;
- secteur économique : variable nominale comptant 10 catégories;
- degré d'activité : le taux d'activité de l'employé (si la valeur est 1, l'employé travaille à temps plein);
- âge : l'âge réel;

- carré de l'âge : le carré de l'âge est également inclus, parce qu'il a été constaté que le salaire augmente jusqu'à un certain âge, puis diminue par après (voir, par exemple, Williams, 2010).

Le modèle a été choisi parmi un certain nombre de modèles contenant plusieurs variables en utilisant le critère AIC minimal. La variable dépendante est le logarithme du salaire standardisé. Par salaire standardisé d'une personne, nous entendons le salaire calculé pour cette personne si elle travaillait à temps plein. Cette variable est fournie par l'Office fédéral de la statistique suisse dans l'ensemble de données, de sorte qu'aucun calcul n'a été effectué par les auteurs.

6.3 Poids et distributions contrefactuelles

La présente section comprend uniquement les résultats exprimés en logarithmes. Lorsqu'on utilise la méthode BO, l'écart entre les salaires moyens des hommes et des femmes est de 0,23, dont 0,09 seulement représentent la part expliquée et 0,14 représentent la part inexpliquée. Nous comparons les résultats obtenus par les méthodes présentées plus haut. La méthode de calage au moyen de la pseudo-distance du khi carré est appelée « linéaire », le calage au moyen de la divergence de Kullback-Leibler est appelé « raking ratio » et la méthode proposée par DiNardo et coll. (1996), ajustée pour tenir compte des poids de sondage est appelée « DFL ». Premièrement, le tableau 6.4 montre les valeurs minimale et maximale des poids, ainsi que les écarts-types, obtenus en utilisant le calage linéaire, le calage par raking ratio et la méthode DFL pondérée.

Tableau 6.4
Poids minimal et maximal et écart-type

Méthode	Minimal	Maximal	Écart-type
Linéaire	-39,06	319,8	4,97
Raking ratio	0,0011	904,7	6,79
DFL pondérée	0,0022	804,4	6,16

Le cas linéaire donne les mêmes résultats que la méthode BO pondérée. Cependant, comme le montre le tableau 6.4, ce cas particulier produit des poids négatifs. Le nombre de ces poids négatifs était de 69 553 (14,59 %). L'option du raking ratio donne toujours des poids positifs, mais l'écart-type des poids est plus grand. Le facteur DFL pondéré possède un plus petit écart-type que les poids obtenus par la méthode de calage par raking ratio. Il existe 1 319 cas où la probabilité conditionnelle d'être un homme est plus grande que 0,98. Initialement, le facteur DFL est multiplié par le ratio entre la somme des poids de sondage des femmes et la somme des poids de sondage des hommes. Puisque $\hat{a}$ est plus petit que 1, le facteur de repondération diminuera. Si, d'autre part, $\hat{a}$ est plus grand que 1 (par exemple, pour les secteurs tels que le secteur public), le facteur de repondération pourrait être plus grand. Le tableau 6.5 montre l'effet de structure estimé aux niveaux moyens des salaires. Les deux approches de calage donnent des effets de structure et de composition égaux. L'utilisation du facteur de repondération DFL résulte en un effet de structure légèrement plus faible et un effet de composition plus grand que les deux autres méthodes.

Tableau 6.5
Effets de composition et de structure estimés dans la différence entre les moyennes

Méthode	Effet de composition	Effet de structure	Total
Linéaire	0,09	0,14	0,23
Raking ratio	0,09	0,14	0,23
DFL pondérée	0,10	0,13	0,23

Étant donné que des poids négatifs sont obtenus dans le premier cas de calage, la densité estimée correspondante ne peut pas être représentée graphiquement. Seules les distributions contrefactuelles des salaires des femmes obtenues en utilisant le raking ratio et le facteur de repondération DFL sont construites. Elles sont présentées à la figure 6.3.

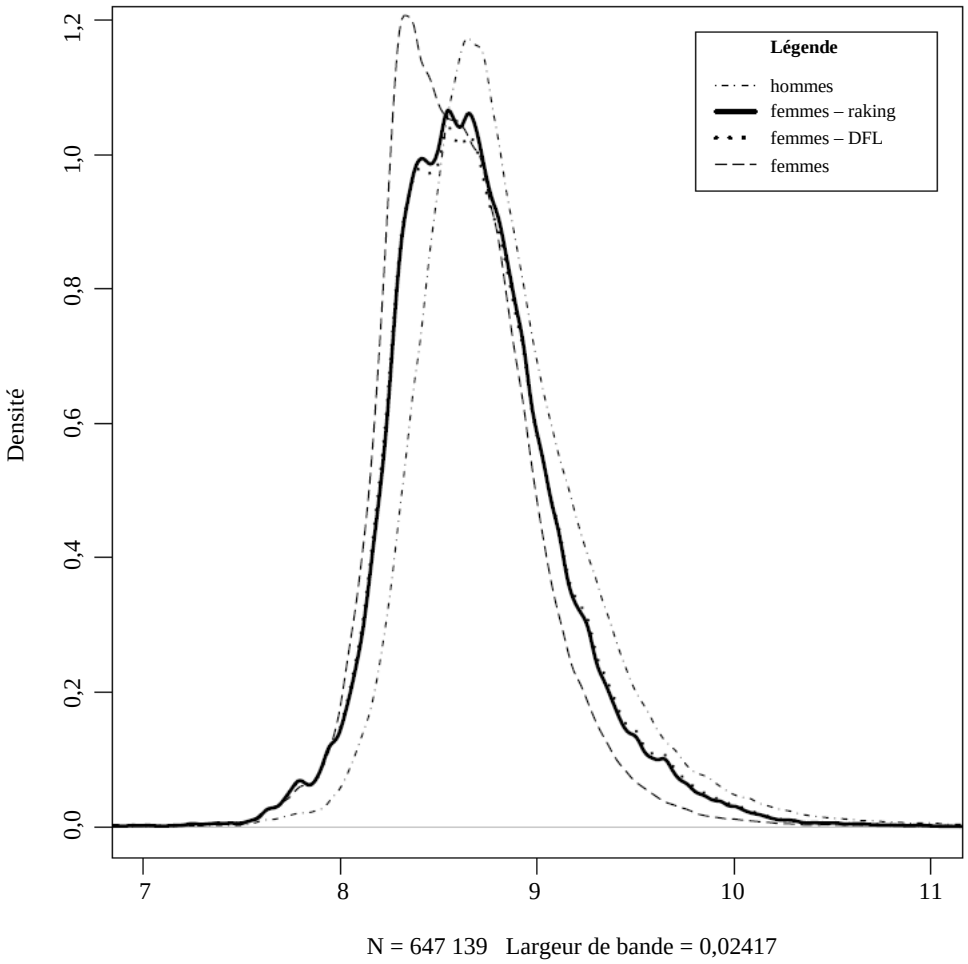

Figure 6.3 Densités estimées du logarithme du salaire des femmes et des hommes et distributions contrefactuelles du logarithme du salaire des femmes construites en utilisant le raking ratio et le facteur DFL repondéré, respectivement.

La figure 6.3 montre que les deux distributions contrefactuelles des salaires sont très proches l'une de l'autre autour des queues. Cependant, vers le milieu, les deux méthodes ne donnent pas les mêmes résultats. Comme nous l'avons mentionné plus haut, l'utilisation des méthodes de repondération DFL et de calage permet d'estimer les effets de composition et de structure non seulement aux niveaux moyens, mais aussi tout le long de la distribution. Le tableau 6.6 donne les effets de structure et de composition estimés des écarts salariaux entre les hommes et les femmes calculés en utilisant les trois méthodes pour certains quantiles.

Tableau 6.6
Effets de composition et de structure estimés des écarts salariaux pour certains quantiles

Quantile	Méthode	Effet de composition (%)	Effet de structure (%)	Total
1 %	Linéaire	0,01 (3 %)	0,20 (97 %)	0,21
	Raking	-0,01 (-3,5 %)	0,22 (103,5 %)	0,21
	DFL pondérée	-0,01 (-3,4 %)	0,22 (103,4 %)	0,21
10 %	Linéaire	0,05 (28,8 %)	0,12 (71,2 %)	0,17
	Raking	0,04 (22,4 %)	0,13 (77,6 %)	0,17
	DFL pondérée	0,03 (19,4 %)	0,14 (80,6 %)	0,17
20 %	Linéaire	0,07 (34,2 %)	0,13 (65,8 %)	0,20
	Raking	0,06 (29,7 %)	0,13 (70,3 %)	0,19
	DFL pondérée	0,05 (28,2 %)	0,14 (71,8 %)	0,19
50 %	Linéaire	0,09 (46,3 %)	0,10 (53,7 %)	0,19
	Raking	0,09 (44,7 %)	0,11 (55,3 %)	0,20
	DFL pondérée	0,09 (45,7 %)	0,11 (54,3 %)	0,20
80 %	Linéaire	0,11 (43,9 %)	0,15 (56,1 %)	0,26
	Raking	0,12 (46,5 %)	0,14 (53,5 %)	0,26
	DFL pondérée	0,13 (50,8 %)	0,13 (49,2 %)	0,26
90 %	Linéaire	0,15 (46,0 %)	0,18 (54,0 %)	0,33
	Raking	0,17 (51,6 %)	0,16 (48,4 %)	0,33
	DFL pondérée	0,19 (58,0 %)	0,14 (42,0 %)	0,33
99 %	Linéaire	0,24 (40,0 %)	0,36 (60,0 %)	0,60
	Raking	0,27 (45,3 %)	0,33 (54,7 %)	0,60
	DFL pondérée	0,29 (49,4 %)	0,30 (50,6 %)	0,59

La proportion de l'effet de structure dans l'écart salarial entier entre les hommes et les femmes diminue à mesure qu'augmente l'ordre du quantile. Cela signifie qu'une plus grande part de l'écart salarial peut être expliquée par des différences de caractéristiques des groupes pour les emplois à salaires élevés que pour les emplois à faibles salaires. Le raking ratio et le facteur de repondération DFL donnent des résultats similaires jusqu'au quantile d'ordre 90 %. Au premier percentile, l'effet de composition estimé est négatif, ce qui signifie qu'à ce point, les différences de salaires sont dues uniquement à la discrimination.

La figure 6.4 montre les quantiles pondérés des logarithmes du salaire des hommes et du salaire des femmes, et contraste les distributions contrefactuelles obtenues au moyen du calage par raking ratio et du facteur de repondération DFL. Comme le calage linéaire donnait des poids négatifs, le graphique correspondant n'est pas présenté.

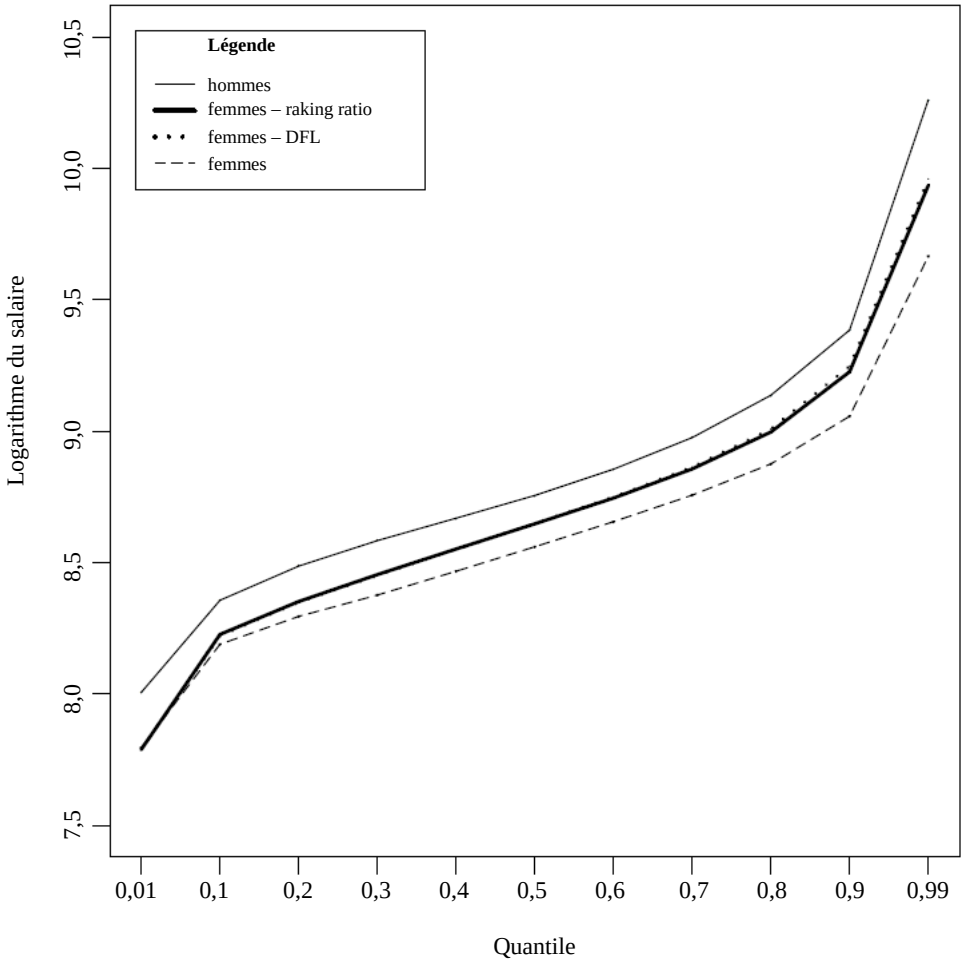

Figure 6.4 Quantiles pondérés des logarithmes des salaires des femmes et des hommes, et quantiles pondérés des distributions contrefactuelles du logarithme du salaire des femmes construites en utilisant le calage par raking ratio et le facteur DFL pondéré.

6.4 Décomposition plus poussée de l'effet de structure

Un modèle logistique de la probabilité d'être un homme donne des valeurs estimées comprises entre 0,002 et 0,99. C'est pour les variables « années d'occupation du poste courant », « âge » et « carré de l'âge » que les écarts entre les valeurs moyennes des hommes et les valeurs repondérées des femmes, calculées en utilisant le facteur de repondération, sont les plus grands. Dans l'équation (4.8), l'effet de structure est composé de l'effet pur et l'effet résiduel. En utilisant le facteur de repondération DFL, l'effet résiduel vaut -0,00474. Par contre, en utilisant l'une ou l'autre des techniques de calage, cet effet est nul dans les deux cas. En outre, l'approche de calage permet de contourner le calcul des coefficients de régression contrefactuels, parce que la technique assure l'égalité entre les moyennes $\hat{\bar{\mathbf{X}}}_M$ et $\hat{\bar{\mathbf{X}}}_{F|M}$. Donc, le calage représente une généralisation de la méthode du facteur de repondération DFL, car il permet une estimation plus précise de l'effet de structure, vu que la valeur résultante n'inclut que l'effet pur.

7 Conclusion

Le phénomène de discrimination présente de multiples facettes et peut être produit par de nombreux mécanismes. Néanmoins, nous n'examinons ici son estimation que d'un point de vue méthodologique. Les deux approches de calage prises en considération représentent une généralisation de deux méthodes de décomposition existantes, à savoir la méthode de Blinder (1973) et Oaxaca (1973) et la méthode semi-paramétrique de DiNardo et coll. (1996), toutes deux exprimées en utilisant les poids de sondage. Les méthodes originales peuvent aussi être obtenues si tous les poids de sondage sont considérés comme égaux à 1. Le cas linéaire donne le même résultat que la méthode BO. Cependant, puisque les poids résultants ne sont pas bornés, des valeurs négatives sont parfois observées. Tout comme la méthode DFL, l'approche de calage permet de décomposer les écarts salariaux à d'autres points que la moyenne, tels que les quantiles. Cependant, le calage par raking ratio représente une amélioration de la méthode DFL, en ce sens que l'estimation de l'effet de structure comprendra toujours un effet résiduel nul. Donc, l'effet de structure sera composé uniquement de l'effet pur. La décomposition des écarts salariaux le long des quantiles permet de conclure que, dans les emplois faiblement rémunérés, les inégalités sont dues uniquement à la discrimination. Dans le présent article, l'accent est mis sur la généralisation de deux méthodes de décomposition bien établies au moyen de l'approche de calage.

Remerciements

Les auteurs remercient l'Office fédéral de la statistique suisse de son soutien financier, et son département des salaires (LOHN) d'avoir fourni les données. Cependant, les opinions exprimées dans le présent article ne reflètent pas nécessairement celles de l'Office fédéral de la statistique suisse.

Annexe A

Preuve du résultat 1

$$\begin{aligned}\hat{\mathbf{X}}_h\hat{\boldsymbol{\beta}}_h &= \hat{\mathbf{X}}_h\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\ &= \sum_{j\in S_h} d_j\mathbf{x}_j^\top\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\ &= \left(\sum_{j\in S_h}\boldsymbol{\varsigma}^\top d_j\mathbf{x}_j\mathbf{x}_j^\top\right)\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\ &= \sum_{l\in S_h}\boldsymbol{\varsigma}^\top d_l\mathbf{x}_l y_l = \sum_{l\in S_h} d_l y_l = \hat{Y}_h.\end{aligned}$$

En divisant cette équation par $\sum_{k\in S_h} d_k$, on obtient le résultat 1.

Annexe B

B.1 Linéarisation des moyennes

Afin de calculer la variance des moyennes et des moyennes contrefactuelles, nous avons utilisé la méthode de linéarisation proposée par Graf (2011). L'auteur propose de calculer la dérivée partielle de l'estimateur par rapport à l'indicateur d'échantillon. Cette dérivée fournit la variable linéarisée qui peut être insérée dans l'estimateur de variance. Les moyennes sont définies par :

$$\hat{\bar{Y}}_F = \frac{\sum_{k\in S_F} d_k y_k}{\sum_{k\in S_F} d_k},$$

et

$$\hat{\bar{Y}}_M = \frac{\sum_{l\in S_M} d_l y_l}{\sum_{l\in S_M} d_l}.$$

Pour les deux salaires moyens, nous obtenons les variables linéarisées :

$$\frac{\partial\hat{\bar{Y}}_F}{\partial I_j} = \begin{cases}\dfrac{d_j\left(y_j - \hat{\bar{Y}}_F\right)}{\sum_{k\in S_F} d_k} & j\in S_F \\ 0 & j\in S_M\end{cases},$$

et

$$\frac{\partial\hat{\bar{Y}}_M}{\partial I_j} = \begin{cases}\dfrac{d_j\left(y_j - \hat{\bar{Y}}_M\right)}{\sum_{l\in S_M} d_l} & j\in S_M \\ 0 & j\in S_F\end{cases}.$$

B.2 Linéarisation de la moyenne contrefactuelle

Afin de calculer la moyenne contrefactuelle, nous calculons les poids v_k définis par le système

$$\mathbf{A} = \sum_{k \in S_F} v_k d_k \mathbf{x}_k = \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \sum_{l \in S_M} d_l \mathbf{x}_l = \hat{\bar{\mathbf{X}}}_M \sum_{k \in S_F} d_k,$$

avec

$$v_k = F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right).$$

Pour les variables linéarisées, nous avons considéré deux cas :

- Si $j \in S_F$

$$\frac{\partial \mathbf{A}}{\partial I_j} = v_j d_j \mathbf{x}_j + \left[\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \hat{\bar{\mathbf{X}}}_M.$$

  Donc,

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = -\mathbf{T}^{-1} d_j \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right),$$

  où

$$\mathbf{T} = \sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top.$$

- Si $j \in S_M$

$$\frac{\partial \mathbf{A}}{\partial I_j} = \left[\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l}.$$

  Donc,

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = \mathbf{T}^{-1} d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l}.$$

Puisque nous avons supposé qu'il existe un vecteur $\boldsymbol{\gamma}$ tel que $\boldsymbol{\gamma}^\top \mathbf{x}_k = 1$ pour tout $k \in U$, alors nous avons

$$\boldsymbol{\gamma}^\top \mathbf{A} = \sum_{k \in S_F} v_k d_k = \sum_{k \in S_F} d_k.$$

Considérons maintenant

$$\hat{\bar{Y}}_{F|M} = \frac{\sum_{k \in S_F} v_k d_k y_k}{\sum_{k \in S_F} v_k d_k} = \frac{\sum_{k \in S_F} v_k d_k y_k}{\sum_{k \in S_F} d_k}.$$

De nouveau, deux cas doivent être considérés :

- Si $j \in S_F$

$$\frac{\partial \hat{\bar{Y}}_{F|M}}{\partial I_j} = \frac{d_j\left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) + \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j}\left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k\right]}{\sum_{k\in S_F} d_k}$$

$$= \frac{d_j\left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) - d_j\left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{T}^{-1}\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k}$$

$$= \frac{d_j\left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F\right]}{\sum_{k\in S_F} d_k},$$

où

$$\mathbf{B}_F = \mathbf{T}^{-1}\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k.$$

- Si $j \in S_M$

$$\frac{\partial \hat{\bar{\mathbf{Y}}}_{F|M}}{\partial I_j} = \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j}\frac{\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k}$$

$$= d_j\left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l}\mathbf{T}^{-1}\frac{\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k}$$

$$= d_j\left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{1}{\sum_{l\in S_M} d_l}\mathbf{B}_F.$$

Donc, la variable linéarisée est

$$z_k = \begin{cases} \dfrac{d_j\left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F\right]}{\sum_{k\in S_F} d_k} & \text{si} \quad j \in S_F \\ \dfrac{d_j\left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F}{\sum_{l\in S_M} d_l} & \text{si} \quad j \in S_M. \end{cases}$$

La variable linéarisée doit uniquement être insérée dans l'estimateur de variance correspondant au plan de sondage. Notons que la variance de la moyenne contrefactuelle dépend de la variance calculée pour l'échantillon d'hommes en ce qui concerne la part qui est expliquée par la régression, et de la variance calculée pour l'échantillon de femmes en ce qui concerne la part qui demeure inexpliquée.

Bibliographie

Bielby, W.T., et Baron, J.N. (1986). Men and women at work: Sex segregation and statistical discrimination. *American Journal of Sociology*, 759-799.

Blinder, A.S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4), 436-455.

Bourguignon, F., Ferreira, F.H. et Leite, P.G. (2002). Beyond Oaxaca-Blinder: Accounting for differences in household income distributions across countries. *Inequality and Economic Development in Brazil*, 105.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. Document de discussion, University of Michigan.

DiNardo, J., Fortin, N.M. et Lemieux, T. (1996). Labor market Institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5), 1001-44.

Donzé, L. (2013). Erreurs de spécification dans la décomposition de l'inégalité salariale. Dans l'*International Conference Ars Conjectandi*, 1713-2013.

Fortin, N., Lemieux, T. et Firpo, S. (2011). Decomposition methods in economics. Dans *Handbook of Labor Economics*, (Éds., O. Ashenfelter et D. Card), 4, 1-102. Elsevier.

Gardeazabal, J., et Ugidos, A. (2005). Gender wage discrimination at quantiles. *Journal of Population Economics*, 18(1), 165-179.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. Dans *Compositional Data Analysis: Theory and Applications*, (Éds., V. Pawlowsky-Glahn et A. Buccianti), 114-127. Wiley, Chichester.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage Discrimination. *Journal of Human Resources*, 23(3), 279-295.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3), 693-709.

Weichselbaumer, D., et Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, 19(3), 479-511.

Weichselbaumer, D., et Winter-Ebmer, R. (2006). Rhetoric in economic research: The case of gender wage differentials. *Industrial Relations: A Journal of Economy and Society*, 45(3), 416-436.

Williams, C. (2010). Bien-être économique. Femmes au Canada : rapport statistique fondé sur le sexe, Statistique Canada.