Estimation de la variance de la moyenne contrefactuelle des salaires

Valérie Schürch Todeschini [1] & Alina Matei [2]

[1] *Université de Neuchâtel, Suisse, valerie.schurch@unine.ch*
[2] *Université de Neuchâtel, Suisse, alina.matei@unine.ch*

Résumé. L'estimation d'une "discrimination" salariale entre les hommes et les femmes ne peut pas être donnée directement en comparant les deux salaires respectifs puisqu'il faut tenir compte des différences de caractéristiques des deux groupes. C'est pourquoi, on utilise la notion de distribution contrefactuelle. La moyenne contrefactuelle des salaires $\overline{Y}_{F|H}$ représente le salaire moyen des femmes si ces dernières avaient les mêmes caractéristiques que les hommes mais les paramètres modélisant leurs salaires resteraient inchangés. Anastasiade & Tillé (2017) ont proposé une méthode basée sur le calage permettant d'estimer cette moyenne contrefactuelle etune approximation de la variance de cet estimateur en utilisant la méthode introduite par Graf (2011). Toutefois, la méthode d'estimation de la variance n'avait pas été évaluée numériquement. Cette méthode d'estimation de la variance pour l'estimateur de $\overline{Y}_{F|H}$ est évaluée à l'aide de simulations de Monte-Carlo. Des données artificielles et des données réelles provenant de l'Office Fédéral de la Statistique, Suisse sont utilisées.

Mots-clés. estimation de la variance, moyenne contrefactuelle, calage.

Abstract. The estimation of wage 'discrimination' between men and women cannot be given directly by comparing the two respective wages, since the differences in the wage structure of the two groups must be taken into account. This is why we use the notion of counterfactual distribution. The counterfactual mean wage $Y_{F|H}$ represents the mean wage of women if they had the some characteristics of men but the parameters modelling their wages would remain unchanged. Anastasiade & Tillé (2017) proposed a calibration-based method to estimate this counterfactual mean and an approximation of the variance of this estimator using the method introduced by Graf (2011). However, the variance estimation method had not been evaluated numerically. This variance estimation method for the estimator of $Y_{F|H}$ is evaluated using Monte-Carlo simulations. Artificial data and real data from the Swiss Federal Statistical Office are used.

Keywords. variance estimation, counterfactual mean, calibration.

1 Introduction et notations

Nous travaillons dans le contexte de la discrimination salariale entre femmes et hommes. Supposons que la population est U, de taille $\text{card}(U) = N$ et que pour chaque individu $k \in U$, on dispose d'informations relatives non seulement à son sexe mais également d'autres

informations auxiliaires sous la forme d'un vecteur $\mathbf{x}_k$ de longueur $p+1$, dont l'ensemble forme la matrice $\mathbf{X} = (\mathbf{x}_k)_{k\in U}$. La première colonne de $\mathbf{X}$ ne contient que des éléments égaux à 1.

La connaissance du sexe de chaque individu permet de partitionner l'univers U en deux sous-populations disjointes U_F et U_H, chacune des parties correspondant à l'ensemble des femmes, respectivement des hommes de U. Soit $S \subseteq U$ l'échantillon sélectionné selon un plan de sondage sans remise qui engendre les probabilités d'inclusion d'ordre 1 et 2, notées par π_k et $\pi_{k\ell}, k, \ell \in U$, respectivement. On a également que $S = S_F \cup S_H$, où $S_g \subseteq U_g$, $g \in \{F, H\}$ est l'échantillon des femmes respectivement des hommes et $S_F \cap S_H = \emptyset$.

Supposons que la variable d'intérêt y soit le logarithme du salaire (avec y_k pour de l'individu k) et que notre intérêt se porte sur l'estimation d'une moyenne. Nous utiliserons ainsi l'estimateur de Hajèk

$$\widehat{\overline{Y}} = \frac{\sum\limits_{k\in S} d_k\, y_k}{\sum\limits_{k\in S} d_k},$$

où $d_k = 1/\pi_k$. Nous noterons également $\widehat{\overline{Y}}_g$, avec $g \in \{F, H\}$ les estimateurs de type Hajèk des moyennes des logarithmes des salaires des femmes, respectivement des hommes, $\overline{Y}_g, g \in \{F, H\}$.

2 Décomposition de Blinder et Oaxaca

L'économiste Ronald Oaxaca, intéressé par l'étude des discriminations salariales entre les femmes et les hommes, proposa en 1973 un coefficient de discrimination (Oaxaca, 1973). Simultanément mais indépendamment, Blinder (1973) proposa une décomposition de la différence Δ entre les moyennes des salaires (en échelle logarithmique) des groupes des deux sexes sous la forme

$$\Delta = \overline{Y}_H - \overline{Y}_F = \left(\overline{X}_H - \overline{X}_F\right)^T \boldsymbol{\beta}_F + \overline{X}_H^T \left(\boldsymbol{\beta}_H - \boldsymbol{\beta}_F\right), \tag{1}$$

cette dernière tenant compte des moyennes des caractéristiques $\overline{X}_F$ et $\overline{X}_H$ dans U_F, respectivement U_H, ainsi que des coefficients $\boldsymbol{\beta}_F$ et $\boldsymbol{\beta}_H$ correspondant aux régressions dans les deux sous-groupes et tels que

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta}_g + \epsilon_k,$$

avec $\epsilon_k \sim \mathcal{N}(0, \sigma_g^2)$, indépendantes, pour $k \in U_g$, $g \in \{F, H\}$.

Cette différence Δ est composée de deux parties : la première, appelée *effet de composition*, est explicable puisqu'elle est liée à la différence de caractéristiques des individus. La deuxième $\overline{X}_H^T \left(\boldsymbol{\beta}_H - \boldsymbol{\beta}_F\right)$ n'est pas imputable à des facteurs objectifs et sera appelée *effet structurel*.

Au niveau des échantillons, cette décomposition reste applicable ; il faut toutefois exiger que la première composante de chaque variable auxiliaire $\mathbf{x}_k$ soit égale à 1 afin d'assurer les identités $\overline{Y}_g = \overline{X}_g^T \boldsymbol{\beta}_g$ et $\widehat{\overline{Y}}_g = \widehat{\overline{X}}_g^T \widehat{\boldsymbol{\beta}}_g$, avec $\widehat{\overline{X}}_g = \sum_{k\in S_g} d_k \mathbf{x}_k / \sum_{k\in S_g} d_k, g \in \{F, H\}$ (voir résultat 1, Anastasiade et Tillé, 2017).

On définit la moyenne contrefactuelle des salaires comme

$$\overline{Y}_{F|H} = \overline{X}_H^T \boldsymbol{\beta}_F.$$

Elle représente le salaire moyen des femmes si ces dernières avaient les mêmes caractéristiques que les hommes mais leurs paramètres $\boldsymbol{\beta}_F$ resteraient inchangés et sera estimée par

$$\widehat{\overline{Y}}_{F|H}^{BO} = \widehat{\overline{X}}_H^T \widehat{\boldsymbol{\beta}}_F, \tag{2}$$

où $\widehat{\boldsymbol{\beta}}_F$ est l'estimateur de $\boldsymbol{\beta}_F$ calculé sur S_F, par la méthode des moindres carrés ponderés.

DiNardo et al. (1996) ont proposé une méthode par repondération pour estimer $\overline{Y}_{F|H}$, qui peut être également appliquée pour estimer les quantiles de la distribution contrefactuelle des salaires.

3 Utilisation du calage pour estimer $\overline{Y}_{F|H}$

Anastasiade & Tillé (2017) ont proposé une méthode basée sur calage comme une alternative à la méthode de DiNardo et al. (1996). Comme proposé dans Anastasiade & Tillé (2017), posons

$$\widehat{\widetilde{X}}_H = \frac{\sum\limits_{k\in S_H} d_k \mathbf{x}_k}{\sum\limits_{k\in S_H} d_k} \sum_{k\in S_F} d_k. \tag{3}$$

Ainsi, $\widehat{\widetilde{X}}_H$ correspond à la moyenne des variables auxiliaires des hommes ajustée aux femmes. Supposons que l'on applique la méthode de calage avec pour équations de calage

$$\widehat{\widetilde{X}}_H = \sum_{k\in S_F} w_k \mathbf{x}_k, \tag{4}$$

et sous la condition que les poids w_k soient les plus proches possible des d_k, c'est-à-dire minimisant $\sum\limits_{k\in S_F} G(w_k, d_k)$ pour une pseudo-distance $G(.,.)$. Les deux pseudo-distances utilisées seront celles dites de l'entropie et du khi-carré, donnant lieu à des fonctions de calage linéaire et exponentielle (voir Table 1).

Pseudo-distance	$G(w_k, d_k)$	$F(u)$
khi-carré	$\frac{(w_k-d_k)^2}{2d_k}$	$1+u$
entropie	$w_k \log(\frac{w_k}{d_k}) - w_k + d_k$	$\exp(u)$

TABLE 1 – Pseudo-distances et fonctions de calage utilisées.

Suivant l'équation de calage (4), on obtient l'estimation de la moyenne contrefactuelle

$$\widehat{\overline{Y}}_{F|H} = \frac{\sum\limits_{k\in S_F} w_k\, y_k}{\sum\limits_{k\in S_F} w_k}. \tag{5}$$

Notons que l'application du calage linéaire pour obtenir les poids w_k en (5) produit le même estimateur que celui définie en (2).

4 Estimation de la variance de $\widehat{\overline{Y}}_{F|H}$

Graf (2011) a démontré que l'estimation de la variance d'une statistique $Q(.)$, deux fois différentiable suivant I_k, avec $\mathbf{I}(S) = (I_1, ..., I_k, ..., I_N)^T$ le vecteur formé des variables indicatrices du présent échantillon S ($I_k = 1$ si $k \in S$ et 0, sinon) est donné par

$$\widehat{\text{var}}(Q(S)) \approx \sum_{j,k\in S} \widehat{Q'}_j\, \widehat{Q'}_k \frac{\pi_{jk} - \pi_j\pi_k}{\pi_{jk}}, \tag{6}$$

où $\widehat{Q'}_j = \frac{\partial Q(S)}{\partial I_j}\Big|_{\mathbf{I}(S)}$. Considérant que $Q(S) = \widehat{\overline{Y}}_{F|H}$ et calculant les dérivées partielles de $\widehat{\overline{Y}}_{F|H}$ par rapport aux indicatrices des deux sous-échantillons S_F et S_H, il devient possible d'utiliser la formule (6) et ainsi estimer la variance de l'estimateur.

Anastasiade & Tillé (2017) ont réalisé ce calcul en supposant que l'échantillonnage était effectué de manière indépendante dans les deux sous-échantillons S_F et S_H. Les dérivées partielles en (6) peuvent être exprimées comme

$$\frac{\partial \widehat{\overline{Y}}_{F|H}}{\partial I_j} = \begin{cases} \frac{d_j}{\sum\limits_{k\in S_F} d_k}\{F(\mathbf{x}_j^T\boldsymbol{\lambda})y_j - \widehat{\overline{Y}}_{F|H} + \widehat{\mathbf{B}}_F(\widehat{\overline{X}}_H - F(\mathbf{x}_j^T\boldsymbol{\lambda})\mathbf{x}_j)\} & \text{si } j \in S_F, \\ \widehat{\mathbf{B}}_F \cdot \frac{d_j}{\sum\limits_{l\in S_H} d_l} \cdot \big(\mathbf{x}_j - \widehat{\overline{X}}_H\big) & \text{si } j \in S_H, \end{cases} \tag{7}$$

où $\widehat{\mathbf{B}}_F = \big(\sum\limits_{k\in S_F} d_k\, F'(\mathbf{x}_k^T\boldsymbol{\lambda})\mathbf{x}_k^T y_k\big) \cdot \big(\sum\limits_{k\in S_F} d_k\, F'(\mathbf{x}_k^T\boldsymbol{\lambda})\mathbf{x}_k^T\mathbf{x}_k\big)^{-1}$, $F(.)$ et $F'(.)$ correspondent à la fonction de calage choisie (voir Table 1) ainsi que sa dérivée, et $\boldsymbol{\lambda}$ est le vecteur des multiplicateurs de Lagrange.

Il ne reste plus qu'à introduire ces expressions dans l'estimation de la variance (6) pour conclure que

$$\widehat{\text{var}}(\widehat{\overline{Y}}_{F|H}) \approx \sum_{j,k\in S_F} \frac{\partial \widehat{\overline{Y}}_{F|H}}{\partial I_j} \frac{\partial \widehat{\overline{Y}}_{F|H}}{\partial I_k} \frac{\pi_{jk} - \pi_j\pi_k}{\pi_{jk}} + \sum_{j,k\in S_H} \frac{\partial \widehat{\overline{Y}}_{F|H}}{\partial I_j} \frac{\partial \widehat{\overline{Y}}_{F|H}}{\partial I_k} \frac{\pi_{jk} - \pi_j\pi_k}{\pi_{jk}} \tag{8}$$

où les dérivées partielle $\frac{\partial \widehat{\overline{Y}}_{F|H}}{\partial I_j}$ sont données sous (7).

Toutefois, Anastasiade & Tillé (2017) n'ont pas fourni des résultats de l'application de l'expression (8). La section 5 présente des résultats qui évaluent cet estimateur.

5 Résultats

5.1 Méthode utilisée

Afin d'évaluer la qualité de l'estimateur proposé sous (8) et en utilisant les coefficients donnés sous (7), des simulations de Monte-Carlo sont effectuées en implémentant dans `R` les différentes formules. Deux séries de 5000 simulations sont effectuées afin d'estimer dans un premier temps la variance (qui est inconnue), notée par $\text{var}(\widehat{\overline{Y}}_{F|H})$, puis dans un deuxième temps, $\widehat{\overline{Y}}_{F|H}$ en utilisant la méthode basée sur le calage ainsi que l'estimation $\widehat{\text{var}}(\widehat{\overline{Y}}_{F|H})$ donnée par la formule d'approximation. Quatre indicateurs de performance usuels de type Monte-Carlo - le biais estimé $\widehat{B}_{\widehat{\overline{Y}}} = \widehat{\overline{Y}}_{F|H} - \overline{Y}_{F|H}$, le biais relatif absolue estimé de la variance $\widehat{BR}_{\text{var}} = \frac{|\text{var}(\widehat{\overline{Y}}_{F|H}) - \widehat{\text{var}}(\widehat{\overline{Y}}_{F|H})|}{\text{var}(\widehat{\overline{Y}}_{F|H})}$, le coefficient de variation estimé $\widehat{\text{CV}} = \frac{\sqrt{\widehat{\text{var}}(\widehat{\overline{Y}}_{F|H})}}{\widehat{\overline{Y}}_{F|H}}$ et l'erreur quadratique moyenne estimée ($\widehat{\text{EQM}}$) de l'estimateur de la variance - sont calculés. Puisque la distribution de $\widehat{\overline{Y}}_{F|H}$ est considérée normale, un cinquième est ajouté, à savoir la proportion d'intervalles de confiance à 95% contenant $\overline{Y}_{F|H}$, de type :

$$\text{IC}_{0.95}(\widehat{\overline{Y}}_{F|H}) = [\widehat{\overline{Y}}_{F|H} - 1.96\sqrt{\widehat{\text{var}}(\widehat{\overline{Y}}_{F|H})}; \widehat{\overline{Y}}_{F|H} + 1.96\sqrt{\widehat{\text{var}}(\widehat{\overline{Y}}_{F|H})}].$$

Deux types d'échantillonnages sont utilisés à savoir l'échantillonnage simple sans remise et la méthode de Midzuno (1951) qui utilise des probabilités inégales. Pour la méthode de Midzuno, les probabilités d'inclusion d'ordre 1 sont proportionnelles à une variable qui est indiquée dans les tables données en Annexe. Le calage linéaire et le calage par ratissage (raking ratio) sont employés.

Deux types des données sont utilisées : des données artificielles et des données réelles issues de l'Office Fédéral de la Statistique en Suisse. Pour les données artificielles, les y_k sont générés suivant des modèles de type

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta}_g + \epsilon_k,$$

avec $\epsilon_k \sim \mathcal{N}(0, \sigma_g^2)$ indépendantes, pour $k \in U_g$, $g \in \{F, H\}$. Les paramètres $\boldsymbol{\beta}_g$, les distributions de $\mathbf{x}_k, k \in U_F$ et de $\mathbf{x}_k, k \in U_H$, les tailles de sous-populations et des sous-échantillons utilisées sont données dans les tables en Annexe.

Les données réelles proviennent de l'enquête sur la structure des salaires 2012, réalisée par l'Office Fédéral de la Statistique (OFS) auprès des entreprises en Suisse. Nous utilisons des données provenant du secteur 'Télécommunications' (secteur publique uniquement). La variable y est le logarithme du salaire horaire des employés. Les variables auxiliaires utilisées sont : l'âge, l'âge au carré, le nombre d'année de service, le bonus (un bonus peut être accordé) et le logarithme du 13ème salaire (en Suisse, un 13ème salaire peut être accordé en fin d'année ; sa valeur est en général égale à la moyenne des salaires mensuels de l'année). La dernière variable n'est pas utilisée en pratique, mais elle est employée ici pour évaluer l'estimateur, étant donné qu'elle est très bien corélée avec y. Les données ont été pondérées avec des poids fournis par l'OFS et considérées comme la population d'où des échantillons ont

été tirés. Notons que la distribution des poids est très asymétrique (le coefficient d'asymétrie est 8.71 et la médiane est 1.05). Pour assurer que les supports des variables auxiliaires de U_H soient les mêmes que ceux pour U_F, une partie des observations de U_H ont été éliminées. Au final, on a $N_F = 2073$ et $N_H = 2091$. Au niveau U_F, les corrélations suivantes entre la variable y et les variables auxiliaires (âge, âge au carré, nombre d'année de service, bonus et logarithme du 13ème salaire) ont été calculées : 0.64, 0.39, 0.04, 0.06, 0.97. Pour la méthode de Midzuno, les probabilités π_k sont proportionnelles à la variable âge. Les Tables 5 et 6 donnent les résultats pour les données provenant de l'OFS, avec le calage de type raking ratio et des échantillons de tailles $n_F = n_H = 100$ et $n_F = n_H = 300$ et pour chaque type d'échantillonnage utilisé. Le calage linéaire, donnant trop des poids négatifs, a été abandonné pour cet exemple.

5.2 Analyse des résultats

Les différentes simulations proposées montrent qu'en général les caractéristiques intrinsèques des individus ne modifient pas l'estimation de la variance proposée par l'expression (8). Toutefois, on remarque qu'une plus grande variance des caractéristiques en U_F (voir en Annexe, Tableau 2) augumentent les valeurs $\widehat{BR}_{\text{var}}$, $\widehat{\text{CV}}$ et $\widehat{\text{EQM}}_{\text{var}}$ pour les données artificielles.

Le niveau de corrélation entre les variables auxiliaires et la variable d'intérêt dans le sous-ensemble U_F des femmes est primordial, celui dans le sous-ensemble des hommes U_M n'ayant aucune influence. En considérant une seule variable auxiliaire ($p = 1$), l'effet est remarquable ; toute diminution du niveau de corrélation $\text{cor}(X_F, \mathbf{y}_F)$ entraîne directement une surestimation de la variance donnée par l'approximation utilisée. Cette dernière est même massive puisque, non seulement le pourcentage d'intervalles de confiance contenant $\overline{Y}_{F|H}$ mais également le biais relatif de la variance sont supérieurs à 100%. Notons également que cette situation est indépendante du choix de la méthode d'échantillonnage et de la méthode de calage (voir en Annexe, Table 3). Remarquons de plus que l'échantillonnage simple sans remise donne de meilleurs résultats pour tous les indicateurs que le plan de Midzuno.

La situation est similaire avec $p = 2$ ou $p = 3$ (résultats présentés pour $p = 2$ uniquement) mais de manière moins flagrante. Une diminution de la corrélation moyenne au niveau de U_F entraîne une augmentation de la valeur estimée de la variance par la formule d'approximation (8), visible tant sur $\widehat{BR}_{\text{var}}$ que sur le pourcentage d'intervalles de confiance contenant $\overline{Y}_{F|H}$ (voir Annexe, Table 4).

On peut toutefois distinguer quelques différences entre les deux méthodes d'échantillonage. Avec le plan simple sans remise, les valeurs obtenues pour les différents indicateurs (en prenant de fortes corrélations dans l'ensemble des femmes) sont bonnes. Une diminution de la corrélation péjore les résultats. En particuliers, en considérant 2 ou 3 variables auxiliaires, mais dont la rémunération ne portait réellement que sur une (resp. deux) variables, l'évaluation est largement moins bonne que sa valeur correspondante à une (resp. deux) variable(s) pour l'échantillonnage simple sans remise.

Avec la méthode de Midzuno, la surestimation de la variance est très marquée si les $\pi_k, k \in U_F$ sont calculés à l'aide d'une variable auxiliaire peu corélée avec y. En choisissant la

variable la mieux corrélée pour le calcul des probabilités d'inclusion d'ordre 1, l'effet semble s'atténuer (voir Annexe, Table 4).

La modification de la taille des échantillons n'a qu'une portée limitée sur les résultats et ne semble pas prédominante pour les deux méthodes de calage. Un échantillon plus grand des femmes permet toutefois d'obtenir de meilleurs résultats sur la valeur de $\widehat{BR}_{\text{var}}$ (à comparer en Annexe, Table 4, les deux dernières simulations). Pour le plan de Midzuno, les valeurs de $\widehat{BR}_{\text{var}}$ et $\in \text{IC}_{0.95}$ sont visiblement impactées par la corrélation entre y et la variable x_{k,α_F} utilisée pour calculer $\pi_k, k \in U_F$ (voir en Annexe, Table 4). Nous avons aussi constaté que les valeurs obtenues pour la variance estimée dans l'échantillon des femmes, pour les données artificielles, est très faible en regard de la variance obtenue dans l'ensemble des hommes, et la contribution à la valeur globale de la variance estimée sur S_F par la formule (8) est minime.

Pour les données réelles, le comportement semble similaire. La surestimation de la variance est importante en présence des variables auxiliaires partiellement corrélées à la variable d'intérêt. En ne conservant que la variable très bien corrélée à y (le logarithme du 13ème salaire), l'estimation utilisant le plan de Midzuno est très bonne. Toutefois, c'est un cas idéal. En augmentant la taille des échantillons, la valeur du $\widehat{\text{CV}}$ diminue (voir en Annexe, Table 6 par rapport à la Table 5).

En conclusion, sur nos simulations, la méthode proposée par Anastasiade & Tillé (2017) donne de bons résultats si les variables y_F et $\mathbf{X}_F$ sont bien corrélées positivement dans l'échantillon des femmes et si on opte pour un échantillonnage simple sans remise. Pour la méthode de Midzuno, il serait judicieux d'utiliser des probabilités π_k proportionnelles à x_{k,α_F} où x_{α_F} est la variable la mieux corrélée à y_F et qui est utilisée également dans le calcul de la moyenne contrefactuelle. En présence d'une corrélation plus faible entre y_F et $\mathbf{X}_F$, l'estimateur de la variance surestime le paramètre, indifféremment de la méthode d'échantillonnage utilisée.

Annexe

Calage linéaire et raking ratio pour différentes caractéristiques ($p = 1$)												
		distribution $x_{k,2}$ femmes	distribution $x_{k,2}$ hommes	$\overline{Y}_{F\|H}$	$\widehat{\overline{Y}}_{F\|H}$	$\widehat{B}_{\widehat{\overline{Y}}}$	$\mathrm{var}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\mathrm{var}}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{BR}_{\mathrm{var}}$ [%]	% $\in \mathrm{IC}_{0.95}$	$\widehat{\mathrm{CV}}$ [%]	$\widehat{\mathrm{EQM}}_{\mathrm{var}}$
Plan simple	linéaire	$\mathcal{N}(4.5, 1)$	$\mathcal{N}(4.0, 1)$	11.199	11.209	0.009	0.055	0.066	20.90	95.00	2.30	0.055
		$\mathcal{N}(5.0, 1)$	$\mathcal{N}(4.0, 1)$	11.201	11.210	0.010	0.055	0.067	21.98	95.00	2.30	0.055
		$\mathcal{N}(4.75, 2)$	$\mathcal{N}(4.0, 2)$	11.250	11.271	0.020	0.221	0.262	18.64	95.00	4.54	0.221
	raking	$\mathcal{N}(4.5, 1)$	$\mathcal{N}(4.0, 1)$	11.199	11.157	-0.042	0.068	0.067	1.60	96.00	2.31	0.069
		$\mathcal{N}(5.0, 1)$	$\mathcal{N}(4.0, 1)$	11.201	11.159	-0.042	0.068	0.067	1.38	96.00	2.32	0.070
		$\mathcal{N}(4.75, 2)$	$\mathcal{N}(4.0, 2)$	11.250	11.167	-0.083	0.270	0.263	2.72	96.00	4.59	0.277
Midzuno	linéaire	$\mathcal{N}(4.5, 1)$	$\mathcal{N}(4.0, 1)$	11.199	11.190	-0.010	0.078	0.15	90.40	99.50	3.45	0.078
		$\mathcal{N}(5.0, 1)$	$\mathcal{N}(4.0, 1)$	11.201	11.190	-0.010	0.079	0.135	71.84	100.00	3.28	0.079
		$\mathcal{N}(4.75, 2)$	$\mathcal{N}(4.0, 2)$	11.250	11.511	0.261	0.422	0.801	89.85	97.00	7.77	0.490
	raking	$\mathcal{N}(4.5, 1)$	$\mathcal{N}(4.0, 1)$	11.199	11.168	-0.031	0.069	0.144	108.76	100.00	3.39	0.070
		$\mathcal{N}(5.0, 1)$	$\mathcal{N}(4.0, 1)$	11.201	11.169	-0.032	0.069	0.135	94.89	99.50	3.29	0.070
		$\mathcal{N}(4.75, 2)$	$\mathcal{N}(4.0, 2)$	11.250	11.621	0.371	0.296	0.808	173.30	96.50	7.73	0.433

TABLE 2 – Données artificielles. Comparaison des performances de $\widehat{\mathrm{var}}(\widehat{\overline{Y}}_{F|H})$ entre calage linéaire et raking ratio, $p = 1$, plan simple sans remise et méthode de Midzuno. Pour toutes les simulations, les modèles utilisés sont définis par $\boldsymbol{\beta}_F = (1.15, 2.5)^T$ and $\boldsymbol{\beta}_H = (2, 3.1)^T$, ceux-ci permettant des corrélations entre (X_F, y_F) et (X_H, y_H) dans les deux sous-populations proches de 1. Les distributions des caractéristiques $x_{k,2}$ des femmes ainsi que celles des hommes sont données dans les deux premières colonnes de la table. Pour la méthode de Midzuno, les probabilités π_k sont proportionnelles à $x_{k,2}$, $k \in U_F$ ou $k \in U_H$, $N_F = N_H = 5000$, $n_F = n_H = 100$.

Effet des différents degrés de corrélation en U_F ($p = 1$)											
		$\text{cor}(X_F, y_F)$	$\overline{Y}_{F\|H}$	$\widehat{\overline{Y}}_{F\|H}$	$\widehat{B}_{\widehat{\overline{Y}}}$	$\text{var}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\text{var}}(\widehat{\overline{Y}}_{F\|M})$	$\widehat{BR}_{\text{var}}$ [%]	$\% \in \text{IC}_{0.95}$	$\widehat{\text{CV}}$ [%]	$\widehat{\text{EQM}}_{\text{var}}$
Plan simple	lin.	0.7	1.550	1.550	<1e-3	<1e-3	0.002	849.37	100	2.91	<1e-3
		0.5	1.389	1.390	<1e-3	0.0002	0.002	1081.44	100	3.07	<1e-3
	rak.	0.7	1.550	1.551	<1e-3	<1e-3	0.002	1059.31	100	3.16	<1e-3
		0.5	1.389	1.390	<1e-3	<1e-3	0.002	1373.40	100	3.35	<1e-3
Midzuno	lin.	0.7	1.550	1.550	<1e-3	<1e-3	0.185	73539.66	100	27.71	<1e-3
		0.5	1.389	1.390	<1e-3	<1e-3	0.187	101410.2	100	31.14	<1e-3
	rak.	0.7	1.550	1.550	<1e-3	<1e-3	0.182	68471.86	100	27.55	<1e-3
		0.5	1.389	1.389	<1e-3	<1e-3	0.185	93275.60	100	30.96	<1e-3

TABLE 3 – Données artificielles. Effets de différents degrés de corrélation entre X_F et y_F ($p = 1$), plan simple sans remise et méthode de Midzuno. Pour toutes les simulations, les caractéristiques sont données par $x_{k,2} \sim \mathcal{N}(4.5, 1)$ si k est une femme et $x_{k,2} \sim \mathcal{N}(4, 1)$ si k est un homme. Les modèles sont modifiés afin d'avoir différents niveaux de corrélation entre X_F et $\mathbf{y}_F$, en conservant $\text{cor}(X_H, y_H)$ proche de 1. Pour la méthode de Midzuno, les probabilités π_k sont proportionelles à $x_{k,2}$, $k \in U_F$ ou $k \in U_H$, $N_F = N_H = 5000$, $n_F = n_H = 100$.

Effet des corrélations en U_F ($p = 2$)											
modèle (femmes)	plan		$\overline{Y}_{F\|H}$	$\widehat{\overline{Y}}_{F\|H}$	$\widehat{B}_{\widehat{\overline{Y}}}$	$\text{var}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\text{var}}(\widehat{\overline{Y}}_{F\|M})$	$\widehat{BR}_{\text{var}}$ [%]	$\% \in \text{IC}_{0.95}$	$\widehat{\text{CV}}$ [%]	$\widehat{\text{EQM}}_{\text{var}}$
$\boldsymbol{\beta}_F = (1.3, 2.0, 2.9)^T$	simple			22.409	<1e-3	0.106	0.129	21.96	95.00	1.60	0.106
	Midzuno										
	α_F	α_H	22.410								
$\text{cor}(X_2, y_F) = 0.55$	2	2		22.411	0.001	0.132	0.474	258.05	99.95	3.07	0.132
$\text{cor}(X_3, y_F) = 0.83$	2	3		22.424	0.015	0.129	0.468	262.84	99.90	3.05	0.129
$\overline{\text{cor}}(X_F, y_F) = 0.69$	3	2		22.408	-0.002	0.137	0.349	154.48	99.76	2.64	0.137
$n_F = n_M = 100$	3	3		22.415	0.006	0.137	0.346	153.56	99.75	2.62	0.137
$\boldsymbol{\beta}_F = (1.3, 5.0, 1.0)^T$	simple			25.926	0.018	0.224	0.309	38.11	95.00	2.15	0.224
	Midzuno										
	α_F	α_H	25.908								
$\text{cor}(X_2, y_F) = 0.98$	2	2		25.910	0.002	0.276	0.424	53.81	98.12	2.51	0.276
$\text{cor}(X_3, y_F) = 0.18$	2	3		25.917	0.010	0.266	0.420	57.67	98.28	2.50	0.267
$\overline{\text{cor}}(X_F, y_F) = 0.58$	3	2		25.914	0.006	0.287	0.963	235.10	99.81	3.79	0.287
$n_F = n_M = 100$	3	3		25.911	0.003	0.268	0.957	257.19	99.92	3.78	0.268
$\boldsymbol{\beta}_F = (1.3, 5.0, 1.0)^T$	simple			25.847	-0.061	0.232	0.239	3.31	94.50	1.89	0.235
	Midzuno										
	α_F	α_H	25.908								
$\text{cor}(X_2, y_F) = 0.98$	2	2		25.922	0.014	0.282	0.298	5.66	95.42	2.11	0.283
$\text{cor}(X_3, y_F) = 0.18$	2	3		25.912	0.005	0.268	0.294	9.99	95.79	2.09	0.268
$\overline{\text{cor}}(X_F, y_F) = 0.58$	3	2		25.92	0.017	0.288	0.579	101.27	99.27	2.93	0.288
$n_F = 200$, $n_M = 100$	3	3		25.897	-0.010	0.261	0.572	118.73	99.57	2.92	0.261

TABLE 4 – Données artificielles. Effet des corrélation entre $\mathbf{X}_F$ et y_F, $p = 2$ ou $p = 3$, plan simple sans remise et méthode de Midzuno. Le modèle utilisé correspond à $x_{k,2} \sim \mathcal{N}(4.5, 1)$ et $x_{k,3} \sim \mathcal{N}(5, 1)$ si k est une femme et $x_{k,2} \sim \mathcal{N}(4, 1)$ et $x_{k,3} \sim \mathcal{N}(4.5, 1)$ si k est un homme. Le salaire est modifié afin d'obtenir différents niveaux de corrélation entre $\mathbf{X}_F$ et y_F, tout en conservant fixe $\boldsymbol{\beta}_H = (2.0, 3.3, 3.4)^T$, ceci permettant une corrélation moyenne $\overline{\text{cor}}(\mathbf{X}_H,\ y_H) = 0.72$. Pour la méthode de Midzuno, les probabilités π_k sont proportionnelles à x_{k,α_F}, $k \in U_F$ et à x_{k,α_H}, $k \in U_H$, respectivement ; $N_F = N_H = 5000$, $n_F = n_H = 100$ pour les trois premières simulations et $n_F = 200$ et $n_M = 100$ pour la dernière.

Plan	$\overline{Y}_{F\|H}$	$\widehat{\overline{Y}}_{F\|H}$	$\widehat{B}_{\widehat{\overline{Y}}}$	$\mathrm{var}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\mathrm{var}}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\mathrm{BR}}_{\mathrm{var}}$ [%]	$\% \in \mathrm{IC}_{0.95}$	$\widehat{\mathrm{CV}}$ [%]	$\widehat{\mathrm{EQM}}_{\mathrm{var}}$
simple sans remise									
$\mathbf{X}_F = (1, x_2, x_3, x_4, x_5, x_6)^T$	4.310	4.315	0.004	0.021	0.313	1362.335	100	12.966	0.021
$\mathbf{X}_F = (1, x_2, x_3, x_6)^T$	4.292	4.288	-0.003	0.021	0.221	942.362	100	10.970	0.021
$\mathbf{X}_F = (1, x_2, x_6)^T$	4.333	4.334	0.0005	0.023	0.065	183.017	98.89	5.921	0.023
$\mathbf{X}_F = (1, x_6)^T$	4.332	4.329	-0.003	0.023	0.021	5.559	88.90	3.412	0.023
Midzuno									
$\mathbf{X}_F = (1, x_2, x_3, x_4, x_5, x_6)^T$	4.310	4.329	0.019	0.008	0.154	1813.658	100	9.075	0.008
$\mathbf{X}_F = (1, x_2, x_3, x_6)^T$	4.292	4.309	0.016	0.008	0.128	1470.498	100	8.311	0.008
$\mathbf{X}_F = (1, x_2, x_6)^T$	4.333	4.341	0.007	0.007	0.028	297.738	99.97	3.906	0.007
$\mathbf{X}_F = (1, x_6)^T$	4.332	4.341	0.009	0.008	0.007	5.873	94.11	2.027	0.008

TABLE 5 – Données réelles. Plan simple sans remise et méthode de Midzuno, $N_F = 2017, N_H = 2081, n_F = n_H = 100$. Les caractéristiques sont : $x_2 = \mathrm{age}, x_3 = \mathrm{age}^2, x_4 =$ nombre d'année de service, $x_5 =$ bonus, $x_6 = \log(\text{13ème salaire})$. Les corrélations entre les caractéristiques et la variable $y = \log(\text{salaire})$ sont respectivement 0.64, 0.39, 0.04, 0.06, 0.97. Pour la méthode de Midzuno, les probabilités π_k sont proportionnelles à $x_{k,2}$, $k \in U_F$ ou $k \in U_H$.

Plan	$\overline{Y}_{F\|H}$	$\widehat{\overline{Y}}_{F\|H}$	$\widehat{B}_{\widehat{\overline{Y}}}$	$\mathrm{var}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\mathrm{var}}(\widehat{\overline{Y}}_{F\|H})$	$\widehat{\mathrm{BR}}_{\mathrm{var}}$ [%]	$\% \in \mathrm{IC}_{0.95}$	$\widehat{\mathrm{CV}}$ [%]	$\widehat{\mathrm{EQM}}_{\mathrm{var}}$
simple sans remise									
$\mathbf{X}_F = (1, x_2, x_3, x_4, x_5, x_6)^T$	4.310	4.319	0.008	0.006	0.128	1762.006	100	8.294	0.006
$\mathbf{X}_F = (1, x_2, x_3, x_6)^T$	4.292	4.296	0.003	0.006	0.079	1080.607	100	6.542	0.006
$\mathbf{X}_F = (1, x_2, x_6)^T$	4.333	4.335	0.001	0.006	0.019	213.820	99.61	3.236	0.006
$\mathbf{X}_F = (1, x_6)^T$	4.332	4.335	0.003	0.006	0.006	1.563	93.79	1.903	0.006
Midzuno									
$\mathbf{X}_F = (1, x_2, x_3, x_4, x_5, x_6)^T$	4.310	4.327	0.017	0.002	0.072	3318.356	100	6.226	0.002
$\mathbf{X}_F = (1, x_2, x_3, x_6)^T$	4.292	4.312	0.020	0.002	0.045	1842.702	100	4.938	0.002
$\mathbf{X}_F = (1, x_2, x_6)^T$	4.333	4.340	0.006	0.001	0.008	332.697	99.97	2.116	0.001
$\mathbf{X}_F = (1, x_6)^T$	4.332	4.339	0.007	0.002	0.002	3.636	95.47	1.050	0.002

TABLE 6 – Données réelles. Plan simple sans remise et méthode de Midzuno, $N_F = 2017, N_H = 2081, n_F = n_H = 300$. Les caractéristiques sont : $x_2 = \mathrm{age}, x_3 = \mathrm{age}^2, x_4 =$ nombre d'année de service, $x_5 =$ bonus, $x_6 = \log(\text{13ème salaire})$. Les corrélations entre les caractéristiques et la variable $y = \log(\text{salaire})$ sont respectivement 0.64, 0.39, 0.04, 0.06, 0.97. Pour la méthode de Midzuno, les probabilités π_k sont proportionnelles à $x_{k,2}$, $k \in U_F$ ou $k \in U_H$.

Bibliographie

Anastasiade, M.-C. & Tillé, Y., (2017), Decomposition of gender wage inequalities through calibration : Application to the Swiss structure of earnings survey, *Survey Methodology*, Statistics Canada, Catalog No. 12-001-X, Vol.43, No. 2, pp. 211–234.

Blinder, A. S. (1973), Wage Discrimination : Reduced Form and Structural Estimates, *The Journal of Human Resources*, Vol. 8, No. 4, pp. 436–455.

DiNardo, J., Fortin, N.M. and Lemieux, T. (1996). Labor market Institutions and the distribution of wages, 1973–1992 : A semiparametric approach. *Econometrica*, Vol. 64, No. 5, pp. 1001–1044.

Graf, M. (2011), Use of Survey Weights for the Analysis of Compositional Data, chapitre 9 dans *Compositional Data Analysis : Theory and Applications*, éditeurs : Vera Pawlowsky-Glahn et Antonella Buccianti, Wiley.

Midzuno, H. (1951), On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Stat. Math.*, No. 3, pp. 99–107.

Oaxaca, R. (1973), Male-female wage differentials in urban labor markets, *International economic review*, vol.14, No. 3, pp. 693–709.