L'estimation de la discrimination salariale en Suisse au travers d'une approche par calage

Mihaela-Catalina Anastasiade [1] & Yves Tillé[1]

[1] *Institut de Statistique, Université de Neuchâtel. Avenue de Bellevaux 51, 2000 Neuchâtel, Suisse*

Résumé. En Suisse, la discrimination salariale à l'encontre des femmes existe encore même si elle est interdite par la loi. Plusieurs méthodes ont été proposées pour mesurer les niveaux de discrimination. Trois méthodes sont discutées dans cet article dont deux sont déjà bien établies dans la littérature : la méthode de Blinder et Oaxaca et celle de DiNardo, Fortin et Lemieux. La méthode de Blinder et Oaxaca fut la première à être proposée. Plusieurs autres techniques ont ensuite été suggérées. Ces méthodes consistent à établir un modèle entre les salaires et les caractéristiques des salariés comme l'âge, l'ancienneté ou le niveau de formation. La méthode de Blinder et Oaxaca permet de mesurer la discrimination salariale en décomposant les différences entre les salaires moyens en une partie explicable par les caractéristiques des individus et une partie non explicable due à la discrimination. La méthode de DiNardo, Fortin et Lemieux est basée sur une technique de repondération des caractéristiques d'un groupe afin de les rendre identiques à ceux de l'autre groupe. Avec cette méthode, on peut mesurer la discrimination non seulement pour des moyennes dans chaque groupe, mais aussi d'autres paramètres comme des quantiles. Cependant, cette méthode ne réussit pas à rendre les distributions des caractéristiques parfaitement identiques. La troisième méthode que nous proposons est basée sur une approche par calage. Cette approche résout les problèmes rencontrés par les deux premières méthodes. En utilisant le calage, on s'assure que les distributions des caractéristiques sont identiques et on parvient aussi à estimer la discrimination salariale non seulement pour des moyennes, mais aussi pour des quantiles.

Mots-clés. Blinder-Oaxaca, calage, discrimination salariale, repondération

1 La discrimination salariale en Suisse

En Suisse, la discrimination au travail en raison du sexe est interdite par la loi depuis 1996. Cette loi couvre tout type de discrimination, par exemple celle liée à la formation ou à la rémunération. Dans cet article, on traite la discrimination salariale, définie comme l'attribution d'une rémunération différente pour un travail identique. La discrimination salariale à l'encontre des femmes existe encore dans le secteur privé et dans le secteur public. Dans les derniers rapports de l'Office fédéral de la statistique, en 2012, la différence entre les salaires moyens des hommes et ceux des femmes s'élève à 21.3% dans le secteur privé et à 16.5% dans le secteur public. Cette différence ne s'explique évidemment pas seulement par une discrimination.

2 L'estimation de la discrimination salariale

Comme la discrimination salariale persiste malgré la législation, de nombreuses méthodes ont été proposées afin de la mesurer. Comme les travailleurs ont des caractéristiques différentes et qu'il est difficile de trouver des individus avec des profils identiques, il est normal d'observer des différences de salaire. Le problème consiste donc à établir quelle proportion des différences salariales est justifiable et quelle est la part due à la discrimination.

2.1 La méthode de Blinder-Oaxaca

La première méthode a été proposée par Blinder (1973) et Oaxaca (1973). Elle traite le problème des profils différents des individus, en séparant les différences des salaires en deux parties: une partie expliquée par les différentes caractéristiques et une partie non-expliquée par des facteurs objectifs. Les deux échantillons sont notés, pour les hommes et femmes, $S_h, h \in \{H, F\}$. Dans chaque groupe, on mesure pour chaque individu p caractéristiques, par exemple, l'âge et le niveau de formation. Pour chaque individu $k \in S_h$, on considère le vecteur $\boldsymbol{x}_k$ des p valeurs prises par ces caractéristiques où

$$\boldsymbol{x}_k = (1 \dots x_{kp})^\top, k \in S_h.$$

Cette méthode est basée sur un modèle linéaire entre le logarithme du salaire et ces caractéristiques. Ainsi, on a

$$\Delta = \widehat{\overline{Y}}_M - \widehat{\overline{Y}}_F = (\widehat{\overline{\boldsymbol{X}}}_M - \widehat{\overline{\boldsymbol{X}}}_F)^\top \widehat{\boldsymbol{\beta}}_F + \widehat{\overline{\boldsymbol{X}}}_M^\top (\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_F), \tag{1}$$

où $\widehat{\overline{Y}}_h$ est la moyenne estimée du logarithme du salaire dans le groupe h. De plus, les $\widehat{\boldsymbol{\beta}}_h$ sont les coefficients de régression estimés dans chaque groupe et les $\widehat{\overline{\boldsymbol{X}}}_h$ sont les vecteurs estimés des moyennes des caractéristiques mesurées dans le groupe h. Les coefficients de régression représentent la contribution de chaque caractéristique au salaire. Tous les éléments dans l'Équation (1) sont calculés en tenant compte des poids de sondage, notés par d_k, $k \in S_h$. Ainsi,

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

et

$$\widehat{\overline{\boldsymbol{X}}}_h = \frac{\sum_{k \in S_h} d_k \boldsymbol{x}_k}{\sum_{k \in S_h} d_k}.$$

On voit dans l'Équation (1) que la différence entre les salaires moyens peut être décomposée entre une différence due aux caractéristiques des individus, qui est la partie expliquée, et une différence entre les coefficients de régression, qui est la partie non-expliquée. Si les coefficients de régression des deux groupes sont égaux, cela signifie que dans chaque groupe, les caractéristiques ont les mêmes contributions au salaire. En réalité, ce n'est pas le cas.

Cette méthode implique une hypothèse de linéarité entre les variables, ce qui est évidemment discutable. De plus, cette méthode se concentre sur des comparaisons de moyennes, ce qui est relativement restrictif. Les deux restrictions sont levées par la méthode de DiNardo, Fortin et Lemieux.

2.2 La méthode de DiNardo, Fortin et Lemieux

DiNardo, Fortin et Lemieux (1996) ont identifié les limitations de la méthode de Blinder-Oaxaca et les ont abordées par une méthode de repondération. Ainsi, on définit le facteur de pondération estimé par

$$\widehat{\psi}(\boldsymbol{x}_k) = \frac{\widehat{\Pr}(D_{Mk} = 1 \mid \boldsymbol{x}_k)/\widehat{\Pr}(D_{Mk} = 1)}{\widehat{\Pr}(D_{Mk} = 0 \mid \boldsymbol{x}_k)/\widehat{\Pr}(D_{Mk} = 0)}. \tag{2}$$

où $D_{Mk} = 1$ si l'individu k est un homme and $D_{Mk} = 0$ sinon.

Le vecteur des moyennes des caractéristiques des femmes et le vecteur des coefficients de régression sont repondérés par

$$\widehat{\overline{\boldsymbol{X}}}_{F|M}^{DFL} = \frac{\sum_{k \in S_F} \widehat{\psi}(\boldsymbol{x}_k) d_k \boldsymbol{x}_k}{\sum_{k \in S_F} \widehat{\psi}(\boldsymbol{x}_k) d_k}.$$

et

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

En utilisant la repondération, on peut réécrire équation (1) par

$$\widehat{\overline{Y}}_M - \widehat{\overline{Y}}_F = \left(\widehat{\overline{\boldsymbol{X}}}_{F|M}^{DFL^\top} \widehat{\boldsymbol{\beta}}_F^{DFL} - \widehat{\overline{\boldsymbol{X}}}_F^\top \widehat{\boldsymbol{\beta}}_F \right) + \left(\widehat{\overline{\boldsymbol{X}}}_M^\top \widehat{\boldsymbol{\beta}}_M - \widehat{\overline{\boldsymbol{X}}}_{F|M}^{DFL^\top} \widehat{\boldsymbol{\beta}}_F^{DFL} \right),$$

où le premier élément est la partie expliquée et le deuxième est la partie non-expliquée. En utilisant le facteur de repondération, on peut non seulement mesurer la partie non-expliquée de la différence entre les salaires moyens, mais aussi pour d'autres paramètres, comme des quantiles. L'analyse des quantiles est en effet intéressante pour comprendre si le niveau de discrimination dépend du montant du salaire. Il existe encore deux problèmes avec cette méthode. La première est le fait qu'en pratique, le facteur de repondération n'ajuste pas parfaitement la distribution des caractéristiques des femmes sur celle des hommes. Ainsi, la partie non-expliquée estimée contient une partie résiduelle qui résulte de ce manque d'ajustement. Le deuxième problème est la dispersion élevée du facteur de repondération. Cette dispersion est la conséquence de l'utilisation d'un modèle logistique, où l'on calcule la probabilité d'être un homme étant donné un vecteur observé des caractéristiques. Si au moins une variable prédit bien le sexe, la probabilité s'approche de 1, donc le rapport dans l'Équation (2) est très grand. Les deux problèmes sont résolus par la méthode que nous proposons.

2.3 L'approche par calage

Notre méthode se base sur les techniques de calage développées par Deville et Särndal (1992). En général, dans le cadre d'un sondage, quand l'on dispose des totaux de quelques variables au niveau de la population, on essaye de les utiliser pour améliorer les estimateurs d'une variable d'intérêt qui est corrélée avec ces variables. Ainsi, le calage ajuste des poids de sondage d_k et établit de nouveaux poids w_k qui satisfont l'égalité

$$\sum_{k \in S} w_k \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k. \tag{3}$$

La méthode consiste à minimiser l'expression

$$\sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k}, \tag{4}$$

où $G_k(w_k, d_k)$, est une pseudo-distance convexe et positive, et q_k est un coefficient lié à l'importance de l'unité k dans l'échantillon. En minimisant l'Équation (4) sous la contrainte (3), on trouve que

$$w_k = d_k F_k(\boldsymbol{x}_k^\top \boldsymbol{\lambda}), \tag{5}$$

où $F_k(\boldsymbol{x}_k^\top \boldsymbol{\lambda})$ est la fonction de calage. L'expression $d_k F_k(.)$ est l'inverse de $g_k(., w_k)/q_k, k \in S$. Cette fonction est égale à

$$g_k(w_k, d_k) = \frac{\partial G_k(w_k, d_k)}{\partial w_k}.$$

Le vecteur $\boldsymbol{\lambda}$ contient les multiplicateurs de Lagrange. Il existe plusieurs manières de construire ces poids, en fonction de la pseudo-distance $G_k(w_k, d_k)$ que l'on choisit entre les poids de sondage d_k et les poids de calage w_k. Nous utilisons seulement deux distances présentées dans le Tableau 1.

Notre approche consiste à caler les caractéristiques des femmes sur celles des hommes, pour rendre les deux distributions parfaitement identiques. Ainsi, on cherche de nouveaux poids $w_k, k \in S_F$, tels que

$$\widehat{\widehat{\boldsymbol{X}}}{}^{C}_{F|M} = \widehat{\widehat{\boldsymbol{X}}}_M,$$

Table 1: Pseudo-distances pour calage

Fonction de pseudo-distance	$G_k(w_k, d_k)$	$g_k(w_k, d_k)$	$F(\boldsymbol{x}_k^\top \boldsymbol{\lambda})$
Chi-carré	$\dfrac{(w_k - d_k)^2}{2d_k}$	$\dfrac{w_k}{d_k} - 1$	$1 + q_k \boldsymbol{x}_k^\top \boldsymbol{\lambda}$
Entropie	$w_k \log \dfrac{w_k}{d_k} + d_k - w_k$	$\log \dfrac{w_k}{d_k}$	$\exp(q_k \boldsymbol{x}_k^\top \boldsymbol{\lambda})$

où

$$\widehat{\overline{\boldsymbol{X}}}^C_{F|M} = \frac{\sum_{k \in S_F} w_k \boldsymbol{x}_k}{\sum_{k \in S_F} w_k}.$$

En construisant ces nouveaux poids, il est possible d'établir une distribution des salaires des femmes, en supposant qu'elles aient les mêmes caractéristiques des hommes. Cette technique présente plusieurs avantages. En premier, cette méthode se détache de l'hypothèse de linéarité entre le salaire et les caractéristiques proposée par Blinder et Oaxaca. Elle contourne aussi le modèle logistique, qui est le point de départ de la méthode de DiNardo, Fortin et Lemieux. Le calage avec la distance du chi-deux donne des résultats identiques à la méthode de Blinder-Oaxaca. L'utilisation de la distance basée sur l'entropie se rapproche de la méthode de DiNardo, Fortin et Lemieux mais élimine le problème de la partie résiduelle que celle-ci présente. Ainsi, on peut regarder le calage comme une généralisation des deux méthodes.

Avec cette approche, on peut comparer les salaires des hommes et des femmes dans les deux secteurs, privé et public, et estimer la partie non-expliquée de la différence de salaire pour divers quantiles. On aboutit ainsi à faire une comparaison entre les deux secteurs et on peut établir si le secteur public est plus équitable. Il est aussi possible d'examiner plus précisément le niveau de discrimination selon les secteurs et le niveau de rémunération.

Bibliographie

[1] Blinder, A. S. (1973), Wage discrimination: Reduced Form and Structural Estimates, *Journal of Human Resources*, 8(4):436-455.

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

[3] DiNardo, J., N. M. Fortin, and T. Lemieux (1996), Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach, *Econometrica*, 64(5), 1001-44.

[4] Oaxaca, R. (1973), Male-Female Wage Differentials in Urban Labor Markets, *International Economic Review*, 14(3):693-709.

[5] Office fédéral de la statistique (2015), Inégalité salariale entre les sexes: différences marquées selon les branches économiques, *http://www.ebg.admin.ch/org/?lang=en*. Date

d'accès: le 12 février 2016.