MESURER LES DIFFÉRENCES SALARIALES ENTRE DEUX GROUPES

Rapport méthodologique et paquet R "decr"

Journées suisses de la statistique, Zurich, 27 Août 2018

Sandro Petrillo (sandro.petrillo@ti.ch)
Oscar Gonzalez (oscar.gonzalez@ti.ch)

Repubblica e Cantone Ticino
Ufficio di statistica

Introduction

Quand on parle de salaires, on compare souvent les salaires de deux groupes d'individus, par exemple les hommes et les femmes ou suisses et étrangers. La comparaison directe entre les statistiques salariales de deux groupes peut cacher certains pièges. En effet, les travailleurs en comparaison peuvent être hétérogènes en termes de caractéristiques personnelles et de lieu de travail (profils de formation, postes hiérarchiques dans l'entreprise, branches économiques, etc.). Si ces différences ne sont pas suffisamment prises en compte, il existe un risque important d'interprétation erronée de l'ampleur et de la nature des disparités mesurées. À cet égard, nous avons développé un rapport méthodologique qui formalise une technique **non paramétrique** pour décomposer les différences observées entre les statistiques salariales de deux groupes d'individus en deux composantes:

- une attribuable au fait que les deux groupes ont des caractéristiques différentes (appelée **"partie expliquée"**);
- l'autre qui reflète les différences dans la structure salariale des deux groupes et n'est pas attribuable aux divergences dans la répartition des caractéristiques observées chez les travailleurs (appelée **"partie inexpliquée"**).

Caractéristiques principales de la méthode:

- Non paramétrique
- Utilisation des poids d'échantillonage (pas présente dans d'autres méthodes)
- Applicable à toutes les statistiques qui dépendent de la distribution salariale
- Prise en compte explicite du **support commun**
- Paquet R "decr"

Étapes de l'analyse

- On observe deux groupes d'individus, A et B, dont on veut examiner les différences entre des statistiques salariales.
- Sur la base de certaines caractéristiques (par exemple la branche économique, la position hiérarchique, le niveau de formation, ...), on établit d'abord le **support commun**, c'est-à-dire les combinaisons de caractéristiques où l'on observe des individus des deux groupes.
- Dans le **support commun**, on estime un **facteur de repondération** pour chaque individu du groupe A ($\widehat{\Psi}_A(\mathbf{x}_i) = \frac{\widehat{f}_{X_B|\mathcal{S}}(\mathbf{x}_i)}{\widehat{f}_{X_A|\mathcal{S}}(\mathbf{x}_i)}$).
- Les **facteurs de repondération** (multipliés par les **poids d'échantillonage**) balancent la distribution des caractéristiques des individus du groupe A dans le support commun à celle des individus du groupe B. Les facteurs de repondération permettent donc d'estimer une **distribution des salaires "contrefactuelle"**, qui représente la distribution des salaires du groupe A comme s'ils avaient la même distribution des caractéristiques du groupe B.
- La distribution contrefactuelle des salaires permet l'estimation d'une **statistique salariale contrefactuelle** (moyenne, quantiles, ...), qui est la quantité fondamentale afin de **décomposer la différence des statistiques salariales des deux groupes**, dans le support commun, en une **partie expliquée** (par les différentes distributions des caractéristiques des deux groupes) et une **partie inexpliquée**.

Distribution contrefactuelle des salaires

La distribution contrefactuelle des salaires "mélange" la structure des salaires du groupe A avec les caractéristiques du groupe B.

$$\begin{aligned} F_{Y_A^C|\mathcal{S}}(y) &= \int F_{Y_A|X_A,\mathcal{S}}(y|X) \cdot dF_{X_B|\mathcal{S}}(X) \\ &= \int F_{Y_A|X_A,\mathcal{S}}(y|X) \cdot \Psi_A(X) \cdot dF_{X_A|\mathcal{S}}(X), \end{aligned}$$

où $\Psi_A(X) = dF_{X_B|\mathcal{S}}(X)/dF_{X_A|\mathcal{S}}(X)$ est un facteur de repondération.

Décomposition dans le support commun ($\mathcal{S}$)

$$\begin{aligned} \Delta^{\nu}_{O|\mathcal{S}} &= \nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y_B|D_B,\mathcal{S}}) = \\ &= \left(\nu(F_{Y_A|D_A,\mathcal{S}}) - \nu(F_{Y_A^C|D_B,\mathcal{S}})\right) + \left(\nu(F_{Y_A^C|D_B,\mathcal{S}}) - \nu(F_{Y_B|D_B,\mathcal{S}})\right) = \\ &= \Delta^{\nu}_X + \Delta^{\nu}_S, \end{aligned} \tag{1}$$

où Δ^{ν}_X est la partie expliquée par le différentes distributions des caractéristiques des deux groupes dans le support commun et Δ^{ν}_S est la partie inexpliquée.

Cas particulier des moyennes

$$\begin{aligned} \widehat{\Delta}^{\mu}_O &= \widehat{\overline{Y}}_A - \widehat{\overline{Y}}_B = \\ &= \underbrace{\widehat{p}_{\mathcal{S}^C|A} \cdot \left(\widehat{\overline{Y}}_{A|\mathcal{S}^C} - \widehat{\overline{Y}}_{A|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_A \text{ (due aux individus hors support)}} + \underbrace{\left(\widehat{\overline{Y}}_{A|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_{O|\mathcal{S}}=\widehat{\Delta}^{\mu}_X+\widehat{\Delta}^{\mu}_S} + \underbrace{\widehat{p}_{\mathcal{S}^C|B} \cdot \left(\widehat{\overline{Y}}_{B|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}^C}\right)}_{\widehat{\Delta}^{\mu}_B \text{ (due aux individus hors support)}}. \end{aligned} \tag{2}$$

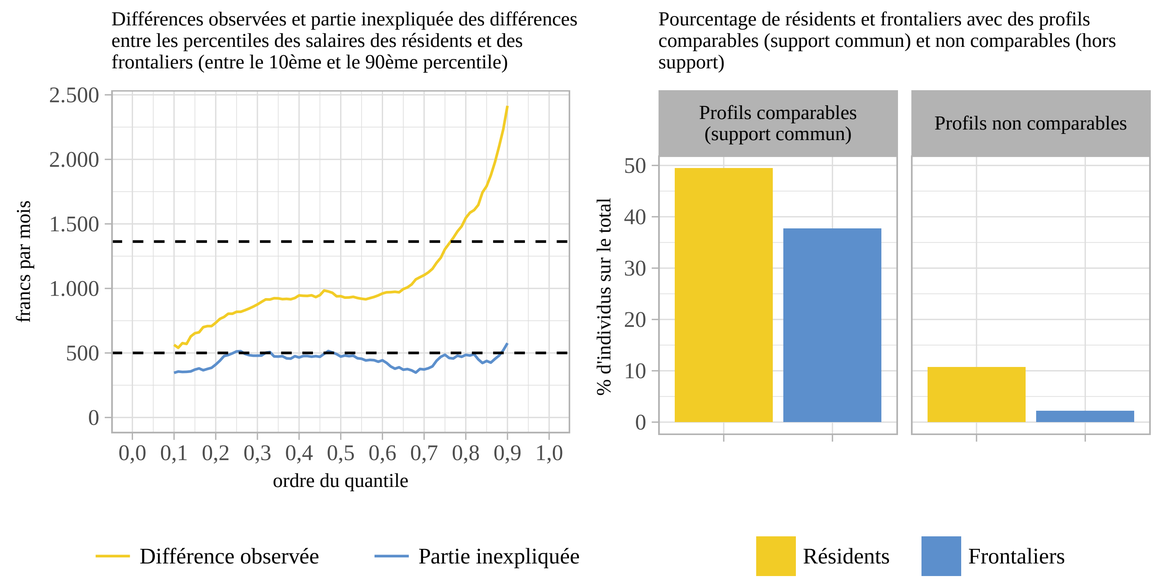
$$\widehat{\Delta}^{\mu}_{O|\mathcal{S}} = \underbrace{\left(\widehat{\overline{Y}}_{A|\mathcal{S}} - \widehat{\overline{Y}}^{C}_{A|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_X \text{ (expliquée)}} + \underbrace{\left(\widehat{\overline{Y}}^{C}_{A|\mathcal{S}} - \widehat{\overline{Y}}_{B|\mathcal{S}}\right)}_{\widehat{\Delta}^{\mu}_S \text{ (non expliquée)}}. \tag{3}$$

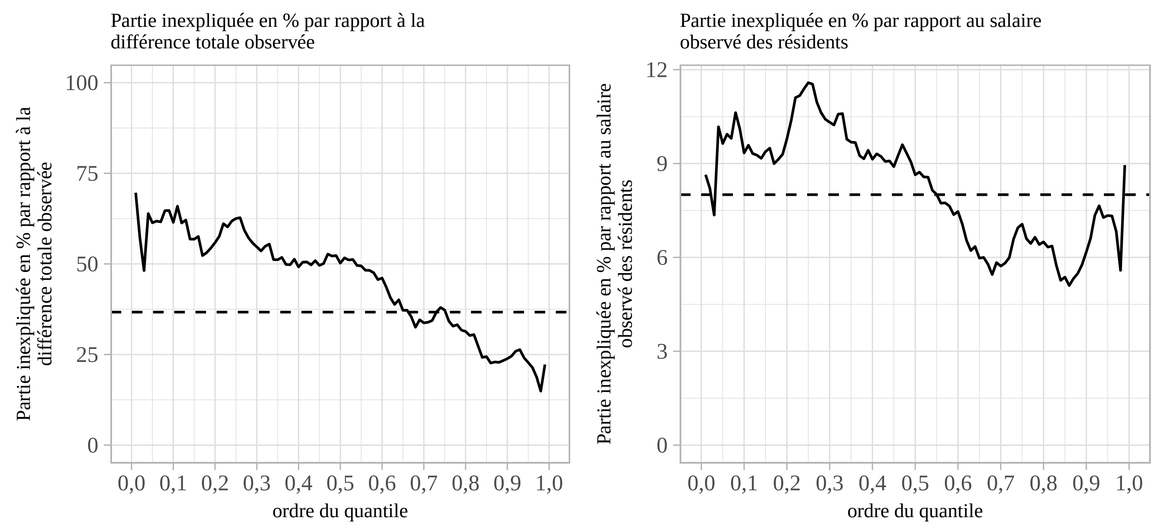
Application / Exemple

- En 2014, au Tessin, dans le secteur privé, différences entre résidents (A) et frontaliers (B) (Enquête suisse sur la structure des salaires, variable MBLS)
- Salaires moyens observés: Résidents: 6.249 francs/mois; Frontaliers: 4.886 francs/mois
- On observe donc une différence de 1.363 francs/mois
- Si on considère la section économique (17), la position dans la profession (4), le degré de formation (3/4), les classes d'âge de 10 ans (6) et les classes d'années de services dans la même entreprise (5), il y a 3.092 combinaisons ("profils", "strates") où on observe au moins un salarié.
- Dans 1.404 strates il existe au moins un individu de chacun des deux groupes (support commun). Les individus dans le support commun représentent le $87,2\%$ des effectifs de l'échantillon.
- Décomposition de la différence des salaires moyens des résidents (A) et des frontaliers (B):

$\widehat{\Delta}^{\mu}_O$	$\widehat{\Delta}^{\mu}_A$	$\widehat{\Delta}^{\mu}_X$	$\widehat{\Delta}^{\mu}_S$	$\widehat{\Delta}^{\mu}_B$
1.363	130	775	500	-42

- La différence totale observée de 1.363 francs par mois ($\widehat{\Delta}^{\mu}_O$) est décomposée en:
  - 130 francs ($\widehat{\Delta}^{\mu}_A$) **expliquables** par le fait qu'il y a des résidents hors support qui ont un salaire moyen supérieur aux résidents du support commun,
  - 775 francs ($\widehat{\Delta}^{\mu}_X$) **expliquables** par le fait que dans les strates comparables (support commun) les deux groupes se distribuent différemment,
  - -42 francs ($\widehat{\Delta}^{\mu}_B$) **expliquables** par le fait qu'il y a des frontaliers hors support qui ont un salaire moyen supérieur aux frontaliers du support commun,
  - 500 francs ($\widehat{\Delta}^{\mu}_S$) **inexpliquables** par les différentes caractéristiques considérées des deux groupes. En d'autres termes, cette composante est la différence "à parité de caractéristiques".





Exemple de code R avec le paquet "decr"

```r
# install.packages("devtools")
devtools::install_github("gibonet/decr")
library(decr)
data(invented_wages)
str(invented_wages)

example(nopodec)  # différence entre les salaires moyens
example(dec_)     # différence entre les quantiles d'un niveau
example(dec_all_) # plusieurs quantiles en une fois
```

Principales références bibliographiques

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." Econometrica 64 (5):1001–44.
Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011. "Causal Inference Without Balance Checking: Coarsened Exact Matching." Political Analysis 20:1–24.
Nopo, Hugo. 2008. "Matching as a Tool to Decompose Wage Gaps." Review of Economics and Statistics 90 (2):290–99.