Devoir Maison

Pratiques de la Recherche en Économie, CPES3

16 Décembre 2024

Consignes

- Le DM est à rendre le 28 Janvier avant 23h59
- Le rendu doit comprendre un fichier pdf qui comprend vos réponses (au format nom_prenom.pdf) et un script `R` au format (nom_prenom.R ou nom_prenom.Rmd)
- Vous pouvez le faire seul/seule ou en binôme
- **Veillez à commenter votre code**
- La base de données à utiliser est `eec_t1_2017_simulated_wage.rds`, téléchargeable sur le Moodle

Données

Quelques mots sur l'Enquête Emploi en Continu (EEC)

L'Enquête Emploi en Continu est la seule source fournissant une mesure des concepts d'activité, chômage, emploi et inactivité tels qu'ils sont définis par le Bureau international du travail (BIT). Elle comporte par ailleurs des informations très nombreuses sur les caractéristiques des personnes (sexe, âge, diplôme, expérience, etc.), les conditions d'emploi (profession, type de contrat, temps de travail, ancienneté dans l'emploi, sous-emploi, etc.), les situations de non-emploi (méthodes de recherche d'emploi, études, retraite, etc.).

L'enquête Emploi est produite selon un calendrier trimestriel et sa collecte se déroule en continu tout au long de l'année. L'échantillon est constitué de logements. Une fois qu'un logement a été tiré, ses occupants seront enquêtés six trimestres consécutifs.

Données mises à disposition

Les données de l'Enquête Emploi disponibles en open source[1] contiennent uniquement un sous-ensemble de variables (125 variables disponibles contre 722 dans sa version complète accessible après habilitation par le Comité du Secret Statistique, via le site de l'Adisp - Progedo). **Parmi les variables manquantes figure le salaire**.

Dans le cadre de ce devoir, le **salaire net mensuel** des individus a donc été simulé de manière à respecter une distribution réaliste des salaires (c'est à dire avec une moyenne et une dispersion cohérentes avec les données complètes de l'EEC). Il s'agit de la variable `wage`. Par ailleurs, on se restreint aux observations du premier trimestre 2017.

Le dictionnaire des codes (pour toutes les variables sauf `wage`) est disponible dans le fichier *eec17_dictionnaire.pdf*.

Objectif de ce DM

L'objectif de ce DM est de quantifier l'écart de salaire moyen entre hommes et femmes et d'en identifier les principaux déterminants.

[1] https://www.data.gouv.fr/fr/datasets/activite-emploi-et-chomage-enquete-emploi-en-continu/

Partie 1: Traitement de la variable de salaire

Cette première partie vise à examiner et préparer la variable de salaire mensuel net pour l'analyse de l'écart de salaire H/F ultérieure. L'objectif est de comprendre la distribution des salaires, d'identifier et de traiter les éventuels problèmes engendrés par l'utilisation de données brutes.

Question 1 (1pt):

Représenter graphiquement la distribution du salaire net mensuel des individus sous la forme d'un histogramme, en ajoutant un titre et le nom des axes. Commentez.

Question 2 (2pts):

Quel peut être le problème induit par l'utilisation de ces données "brutes" de salaire dans un modèle économétrique? Comment peut-on traiter les données en conséquence?

Question 3 (1pt):

Calculez les percentiles du salaire mensuel net (*hint: utiliser la fonction* `quantile()`). Stockez la valeur du 99ème percentile dans `q_99`. Indiquez la valeur du 99ème percentile et interprétez.

Question 4 (1pt):

- Créez la variable `wage_winsor` qui correspond au salaire mensuel net, `wage`, où les 1% des valeurs les plus élevées sont remplacées par la valeur du 99ème percentile du salaire mensuel net
- Représentez graphiquement la distribution de `wage_winsor`. Commentez.

Question 5 (1pt):

- Créez la variable `log_wage_winsor`, le logarithme de `wage_winsor`
- Représentez graphiquement la distribution de `log_wage_winsor`. Commentez.

Partie 2: Statistiques Descriptives

Cette partie a pour objectif d'explorer les données de l'enquête emploi afin de mieux appréhender les écarts entre femmes et hommes en termes de niveau de diplôme, d'activité, d'occupation et de rémunération.

Partie A: Diplôme

Question 6 (1pt):

Représenter graphiquement la distribution du niveau de diplôme selon la classe d'âge. Conservez uniquement les individus âgés de 30 ans ou plus (*hint 1: utiliser la variable* `AGE5`; *hint 2: on pourra par exemple faire un barplot*). Veillez à ce que la légende indique le libellé du niveau de diplôme et non la modalité correspondante. Commentez.

2 phénomènes: - *massification scolaire*: - *stratification scolaire*:

Question 7 (1pt):

- Créez une nouvelle variable `at_least_bac` qui vaut 1 si l'individu a un diplôme au moins égal au Bac, 0 sinon.

- Comparez graphiquement la proportion d'individus ayant un diplôme au moins égal au Bac (`at_least_bac`) entre les hommes et les femmes pour chaque tranche d'âge. Représentez vos résultats graphiquement. Veillez à conserver uniquement les individus de 30 ans ou plus. Commentez.

Partie B: Participation au marché du travail

Question 8 (1pt):

Calculez la distribution de la variable `ACTEU`, par genre, par tranche d'âge, et par genre × tranche d'âge. Proposez une représentation graphique. Que peut-on dire de la participation au marché du travail des femmes?

Question 9 (2pts):

Selon vous, quels sont les facteurs qui déterminent la participation au marché du travail des femmes? Expliquez pourquoi.

Partie C: Rémunération

Question 10 (1pt):

Créez la variable *quotité* qui vaut (*hint: utilisez les variables* `TPPRED` *et* `TXTPPRED`):
- 1 si l'individu travaille à temps complet
- 2 si l'individu travaille plus de 80%
- 3 si l'individu travaille à 80%
- 4 si l'individu travaille à temps partiel entre 50 et 80%
- 5 si l'individu travaille à mi-temps (50%)
- 6 si l'individu travaille moins d'un mi-temps

Question 11 (1pt):

Étudiez la distribution de la quotité de temps de travail (`quotité`) des individus en fonction de la présence d'au moins un enfant dans le ménage (`ENFRED`), sur l'échantillon global, puis séparément sur celui des hommes et des femmes. Commentez.

Question 12 (1pt):

Étudiez la répartition par genre dans chacune des CSP (variable `CSP`) en ne conservant que les individus âgés d'au moins 30 ans et dont la CSP est bien définie et non nulle. Commentez.

Question 13 (1pt):

Créez un dataframe qui comprend, pour chaque CSP, la moyenne du log du salaire mensuel net winsorisé et la proportion de femmes en vous restreignant aux individus de 30 ans ou plus et aux CSP connues et non nulles. Qu'observez-vous?

Partie 3: Analyse Économétrique

Cette analyse économétrique explore les déterminants du salaire horaire et évalue l'effet de leur prise en compte sur l'écart salarial moyen entre femmes et hommes.

Question 14 (0.5pt):

- Créez la variable `femme` qui vaut 1 si l'individu est une femme, 0 sinon. Assurez-vous que la variable soit de type `factor`
- Créez la variable `CSP_r` qui est égale au premier chiffre de la CSP (chiffre des dizaines) si la CSP est différente de 23 (Chefs d'entreprises de 10 salariés ou plus); si la CSP est égale à 23, alors attribuer à `CSP_r` la valeur 3 (i.e. on considère que leur situation se rapproche davantage de celle des Cadres et Professions intermédiaires que de celle des Artisans et Commerçants).

Question 15 (0.5pt):

Créez le dataframe `data_reg` à partir de `data` qui comprend uniquement les observations:
- des individus âgés de 30 ans ou plus
- qui sont en CDD ou CDI
- qui appartiennent à une `CSP_r` différente de 0
- qui appartiennent à une `CSP` comprend au moins 30 individus
- des individus qui travaillent au moins une heure par semaine

Utilisez ce dataframe pour la suite du devoir.

Question 16 (1pt):

Calculez l'écart du salaire moyen entre hommes et femmes (en niveau et en pourcentage, en utilisant le salaire mensuel net winsorisé).

```
gap_perc = (mean(data_reg$wage_winsor[data_reg$femme == 1], na.rm = T)-mean(data_reg$wage_winsor[data_r

gap_euros = (mean(data_reg$wage_winsor[data_reg$femme == 1], na.rm = T) - mean(data_reg$wage_winsor[data

mean(data_reg$wage_winsor[data_reg$SEXE==1], na.rm = T)
```

```
## [1] 2517.062
```

```
mean(data_reg$wage_winsor[data_reg$SEXE==2], na.rm = T)
```

```
## [1] 1935.359
```

```
mean(data_reg$log_wage_winsor[data_reg$femme == 1], na.rm = T)-mean(data_reg$log_wage_winsor[data_reg$f
```

```
## [1] -0.291054
```

Question 17 (2pt):

Estimez le modèle $log(\text{Salaire mensuel Winsorisé}) = \alpha + \beta\text{Femme} + \varepsilon$. Que représentent α et β? Peut-on dire que β est causal?

```
reg = lm(log_wage_winsor ~ femme, data = data_reg)
summary(reg)
```

```
##
## Call:
## lm(formula = log_wage_winsor ~ femme, data = data_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9714 -0.3139 -0.0437  0.3306  1.5147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  7.697032   0.004390 1753.41   <2e-16 ***
## femme1      -0.291054   0.006227  -46.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5538 on 31638 degrees of freedom
##   (3723 observations deleted due to missingness)
## Multiple R-squared:  0.06459,    Adjusted R-squared:  0.06456
## F-statistic:  2185 on 1 and 31638 DF,  p-value: < 2.2e-16
```

Question 18 (1.5pt):

- Créez la variable salaire (winsorisé) **horaire** et son log, que vous nommerez `hwage_winsor` et `log_hwage_winsor` (*hint: utiliser la variable* ***HHCE****, le nombre d'heures travaillées en moyenne* ***par semaine*** *dans l'emploi principal*)
- Quel est maintenant l'écart de salaire moyen entre femmes et hommes? Comment se compare-t-il à l'écart de salaire calculé en question 16 et comment expliquez-vous cette différence?

```
data_reg$hwage_winsor = data_reg$wage_winsor/(data_reg$HHCE*151.67/35)
data_reg$log_hwage_winsor = log(data_reg$hwage_winsor)

gap_h_perc = (mean(data_reg$hwage_winsor[data_reg$femme == 1], na.rm = T) - mean(data_reg$hwage_winsor[d

mean(data_reg$log_hwage_winsor[data_reg$SEXE==1], na.rm = T)
```

```
## [1] 2.601855
```

```
mean(data_reg$log_hwage_winsor[data_reg$SEXE==2], na.rm = T)
```

```
## [1] 2.457945
```

Question 19 (2pt):

Proposez un modèle de régression linéaire qui permet d'atténuer les problèmes d'identification soulevés à la question 17. Commentez les résultats. En particulier, comment varie β? Qu'est-ce que cela indique de l'écart de salaire moyen entre hommes et femmes calculé en question 17?

Question 20 (2pt):

Selon vous, peut-on dire que β estimé à la question 19 est causal? Expliquez.

Partie 4: Décomposition d'Oaxaca-Blinder

La méthode de décomposition d'Oaxaca-Blinder permet de mesurer la part de l'écart entre le salaire moyen des femmes et celui des hommes qui est due à des différences dans les caractéristiques moyennes et celle due à des différences d'effets de ces caractéristiques sur le salaire moyen (dit autrement à des différences dans les coefficients de régression). On considère deux groupes, celui des femmes F et celui des hommes H. L'écart de salaire moyen entre les deux groupes s'écrit:

$$\Delta\overline{Y} = \overline{Y}_H - \overline{Y}_F \tag{1}$$

où $\overline{Y}_i$, $i \in \{H, F\}$ est le salaire horaire moyen du groupe i.

Décomposition du *gender wage gap* en deux parties

Cette décomposition permet d'écrire l'écart moyen de salaire H/F comme la somme d'une partie expliquée par des différences de caractéristiques et d'une partie inexpliquée.

$$\Delta\overline{Y} = \underbrace{\left(\overline{X}'_H - \overline{X}'_F\right)' \hat{\beta}_H}_{\text{Expliquée}} + \underbrace{\overline{X}'_F \left(\hat{\beta}_F - \hat{\beta}_H\right)}_{\text{Inexpliquée}} \tag{2}$$

Question 21 (1pt):

Décrivez ce que mesure la seconde partie de l'équation (2)? Pourquoi parle t-on de composante inexpliquée de l'écart de salaire hommes/femmes?

Question 22 (2pts):

Installez et chargez le package `oaxaca`[2]. Utilisez la fonction `oaxaca` pour réaliser une décomposition d'Oaxaca-Blinder en deux parties. Que peut-on conclure sur les écarts de rémunération salariale entre femmes et hommes? Vous pourrez vous référer à la documentation du package disponible en ligne et vous aider d'un graphique pour interpréter les résultats.

```
library(oaxaca)
```

```
##
## Please cite as:
##  Hlavac, Marek (2022). oaxaca: Blinder-Oaxaca Decomposition in R.
##  R package version 0.1.5. https://CRAN.R-project.org/package=oaxaca
```

```
oaxaca = oaxaca(log_hwage_winsor ~ -1 + as.factor(AGE5) + as.factor(CSP_r) + as.factor(DIP11) + as.facto
```

```
## oaxaca: oaxaca() performing analysis. Please wait.
##
## Bootstrapping standard errors:
## 1 / 100 (1%)
## 10 / 100 (10%)
## 20 / 100 (20%)
## 30 / 100 (30%)
## 40 / 100 (40%)
## 50 / 100 (50%)
## 60 / 100 (60%)
## 70 / 100 (70%)
## 80 / 100 (80%)
## 90 / 100 (90%)
## 100 / 100 (100%)
```

```
oaxaca$y
```

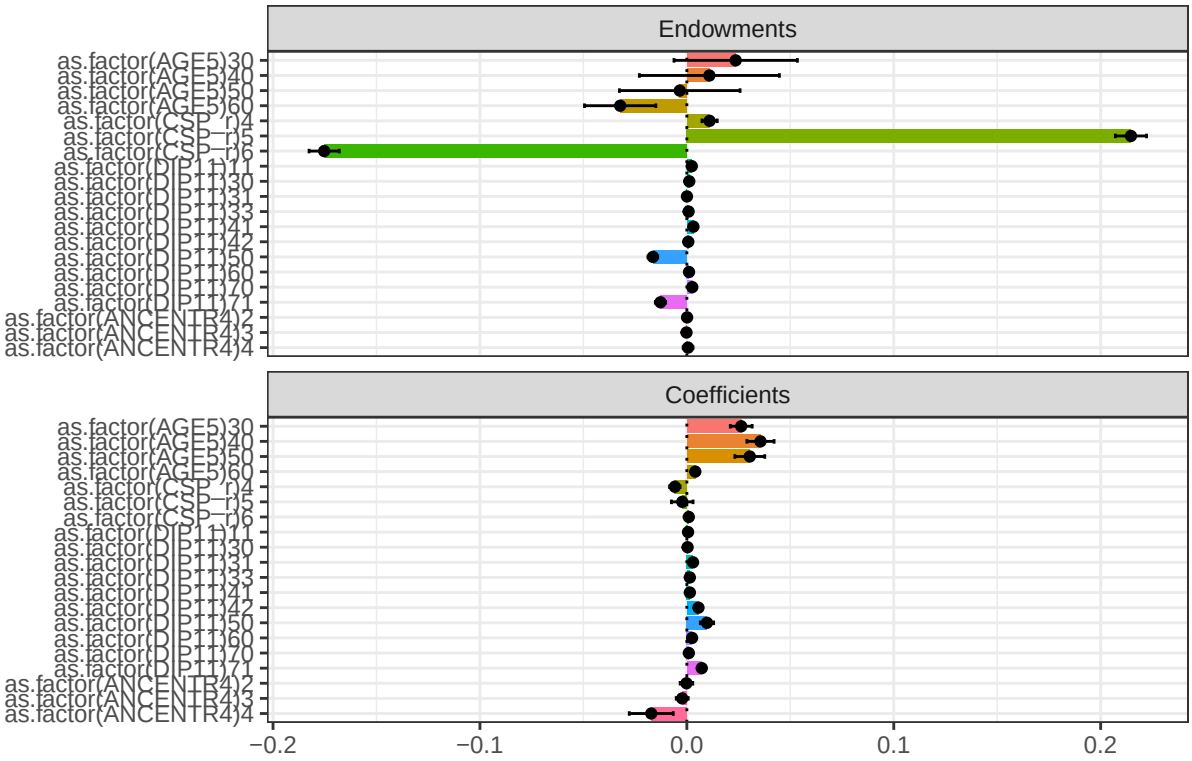
[2] https://cran.r-project.org/web/packages/oaxaca/vignettes/oaxaca.pdf

```
## $y.A
## [1] 2.601855
##
## $y.B
## [1] 2.457945
##
## $y.diff
## [1] 0.1439094
```

```r
# $y.A
# [1] 2.620656
#
# $y.B
# [1] 2.476212
#
# $y.diff
# [1] 0.1444448

# Meaning: average log of hourly wage is 2,62 for males and the average for women is 2,48, with the 0,1

oaxaca$threefold$overall
```

```
##   coef(endowments)     se(endowments) coef(coefficients)   se(coefficients)
##        0.032658982        0.003910376        0.102638926        0.002220693
##  coef(interaction)    se(interaction)
##        0.008611506        0.002245588
```

```r
# coef(endowments)     se(endowments) coef(coefficients)   se(coefficients)  coef(interaction)    se(in
#      0.030508055        0.004997755        0.102151373        0.004053094        0.011785402        0

# Interpretation: of the overall 0,144 difference, approximately 0,030 can be attributed to group diffe

plot(oaxaca, components = c("endowments", "coefficients"))
```




```
# Significant part of the F/H wage gap is driven by group differences in the proportion of individuals
```

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