

Modélisation et identification causale

Séance 1 – Anatomie des moindres carrés ordinaires

Pierre Pora

2022-10-18

Avant de commencer

Calendrier

- ▶ Programme de l'enseignement
 - ▶ 20 oct. **Rappels sur les régressions linéaires** (P. Pora)
 - ▶ 10 nov. **Expériences naturelles etc.** partie 1 (O. Godechot)
 - ▶ 17 nov. **Expériences naturelles etc.** partie 2 (O. Godechot)
 - ▶ 1er déc. **Contrôler les facteurs confondants** partie 1 (P. Pora)
 - ▶ 8 déc. **Contrôler les facteurs confondants** partie 2 (P. Pora)
 - ▶ 15 déc. **Variables instrumentales** partie 1 (P. Pora)
 - ▶ 5 janv. **Variables instrumentales** partie 2 (P. Pora)
 - ▶ 19 janv. **Panels et effets fixes** (O. Godechot)

Evaluation

- ▶ **Mini-mémoire** de quelques pages appliquant une des méthodes proposées sur un exemple empirique
- ▶ En groupe ou pas
- ▶ Propositions de sujets + possibilité de proposer son propre exemple

Enseignement appliqué

- ▶ L'objectif est autant de se familiariser avec les **présupposés théoriques** de certaines méthodes . . .
- ▶ . . . que d'apprendre à les **mettre en pratique**
- ▶ Applications sur des données (pas trop difficilement) accessibles en ligne
- ▶ **Logiciel R + RStudio**
- ▶ **Packages : AER, lfe**
- ▶ Mise en pratique **dès aujourd'hui** → ordinateur allumé + R lancé + packages installés

Disponibilité des supports etc.

- ▶ **Dossier partagé** : supports de cours, applications etc.
- ▶ Sur mon site (*en construction*) : version plus écrite des éléments que je couvre + tout le code des exemples présentés + des détails additionnels non-présentés
→ pierreporra.github.io/inference_causale

Des ressources utiles ?

- ▶ De nombreuses références synthétiques sont disponibles sur les différentes méthodes introduites dans cet enseignement
- ▶ Trois points de précaution :
 - ▶ Niveaux de détail, de technicité et d'abstraction
 - ▶ Traitement plus ou moins actuel de certains sujets
 - ▶ Plus ou moins adapté à la pratique ordinaire des sciences sociales
- ▶ Peu de ressources vraiment satisfaisantes en français (d'où le site)

Des ressources utiles ?

- ▶ Manuels ou quasi-manuels, du plus accessible au plus difficile : Huntington-Klein (2021), Morgan and Winship (2014), Cunningham (2021), Glymour, Pearl, and Jewell (2016), Angrist and Pischke (2008), Imbens and Rubin (2015), Pearl (2009)
- ▶ Tous n'insistent pas sur les mêmes aspects et peuvent diverger sur certaines questions
- ▶ Huntington-Klein (2021) et Cunningham (2021) sont disponibles en ligne avec de nombreux exemples codés
- ▶ Vraiment plus techniques mais aussi vraiment bien faites : les slides du cours d'économétrie appliquée en Ph.D program à Yale sur le site de Paul Goldsmith-Pinkham

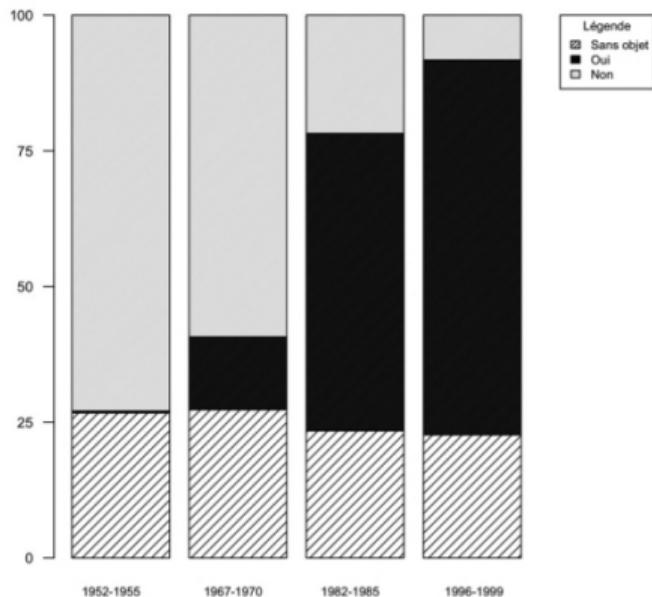
Introduction : pourquoi, comment parler des régressions linéaires ?

Contenu et motivation de la séance

Contenu de la séance

- ▶ Retour sur les **régressions linéaires** par les **moindres carrés ordinaires**
 - ▶ Exclues : toutes les techniques de régression utilisant des modèles non-linéaires (e.g. probit, logit etc.)
 - ▶ Exclues aussi : toutes les techniques de régression linéaire en plusieurs étapes (→ séance sur les variables instrumentales)
- ▶ Les **moindres carrés ordinaires** :
 - ▶ une approche avec une **longue histoire** (début XIXème, lien avec l'astronomie)
 - ▶ qui essaime dans **presque tous les champs scientifiques**

Et en sciences sociales ? Près de 70% des articles publiés à la fin du XXème siècle dans l'*American Sociological Review* font appel à des techniques de régression (Ollion (2011))



Analyse de données recueillies par Martin et Yeung (2003)

Un peu d'ontologie pour préciser l'objet

Trois strates pour deux façons de parler de régressions linéaires

- ▶ **Strate 1** : les **raisons** qui conduisent les agents du monde social à prendre certaines décisions plutôt que d'autres
- ▶ **Strate 2** : ce que l'on peut **observer** dans les données quantitatives que l'on peut collecter
- ▶ **Strate 3** : les **quantités synthétiques** que l'on peut construire pour résumer l'information contenue dans les données dont on dispose

Utiliser les régressions linéaires pour passer d'une strate à celles qui la précèdent logiquement

- ▶ **Strate 3 → Strate 2** : les quantités synthétiques que l'on construit sont-elles une façon intéressante de résumer l'information contenue dans les données que l'on a collectées ?
 - ▶ Quel est le lien entre les quantités renvoyées par les **régressions linéaires** (e.g. les **coefficients**) et les données brutes ?
- ▶ **Strate 3 → Strate 1** : ces quantités synthétiques renseignent-elles sur la façon dont les agents du monde social prennent leurs décisions ?
 - ▶ Les quantités obtenues grâce aux régressions linéaires renseignent-elles sur les **effets** de certaines entités du monde sociale sur d'autres ?

Quels outils conceptuels pour répondre à ce questions ?

Difficile de répondre à la question 2 sans :

1. Avoir bien répondu à la question 1 → questions usuelles de **statistiques** et de **probabilités**
2. Avoir un modèle implicite de la façon dont les agents du monde social agissent, pour pouvoir interpréter ces quantités → outils d'**inférence causale**, et **théorie de l'objet sociologique** que l'on étudie

L'objet de cette séance : revenir sur le passage des données collectées aux quantités synthétiques renvoyées par les régressions linéaires → l'interprétation causale ou comportementale de ces quantités sera abordée dans les **séances ultérieures**.

Un (très bref) rappel de statistique

Préalable de statistique : distinguer quantités estimées, estimateurs et estimation

De nouveau trois strates ! Mais un peu décalées par rapport au premier découpage

- ▶ Strate 1 : des quantités d'intérêt, **impossibles à observer directement** au niveau où elles nous intéressent → **quantités estimées**
 - ▶ Le taux de chômage au sens du BIT dans la population active française toute entière
- ▶ Strate 2 : des **procédures** qui permettent de passer de données observées sur un échantillon à un ou des indicateurs synthétiques → **estimateurs**
 - ▶ La part d'individus actifs qui sont au chômage au sens du BIT dans l'enquête Emploi, avec les poids appropriés
- ▶ Strate 3 : les **résultats** de ces procédures, appliquées à un échantillon donné → **estimations**
 - ▶ 7.4%, le taux de chômage au T2 2022 publié par l'Insee et estimé à partir de l'EEC

Le résultat très classique autour duquel tout tourne

- ▶ Lorsque la **taille d'échantillon** devient suffisamment grande, la **moyenne dans toute une population** est bien approximée par la **moyenne dans un échantillon aléatoire** de taille finie beaucoup plus petit qu'elle
- ▶ Et on connaît la **distribution de l'écart** entre la moyenne dans l'échantillon et la moyenne dans la population
- ▶ Cela justifie d'utiliser la moyenne dans l'échantillon comme un estimateur de la moyenne dans la population
 - ▶ Base de tout ce qui a trait à la **statistique d'enquête** !
- ▶ Point culturel : ces résultats sont connus comme la **loi des grands nombres** et le **théorème central limite**
- ▶ Très utile : presque toutes les quantités qui nous intéressent peuvent (avec quelques efforts) s'écrire comme des moyennes

Disséquer les moindres carrés ordinaires

Anatomie des moindres carrés ordinaires

- ▶ Pour bien comprendre les **régressions linéaires**, il faut comprendre à quelles **quantités estimées** elles renvoient
- ▶ On va considérer de nombreux exemples *en faisant abstraction* de la distinction entre quantités estimées, estimateurs et estimations → faire comme si la moyenne dans l'échantillon et la moyenne dans la population étaient la même chose
 - ▶ Pas de discussion de la **précision**, de la **significativité** etc. pour commencer
- ▶ L'objectif : se convaincre qu'il n'y a rien d'autre derrière tout ça que des **comparaisons de moyennes**
- ▶ Comment s'en rendre compte : décomposer plusieurs exemples simples sur des données d'enquête

Le Current Population Survey

- ▶ Les données que l'on va principalement utiliser aujourd'hui
- ▶ **Enquête étatstunienne** conduite par le *Bureau of Census*
- ▶ Equivalent approximatif de l'**Enquête Emploi** en France
- ▶ Extrait de 534 enquêtés de la vague de mai 1985 que l'on peut récupérer grâce au
package AER

Récupérer les données du CPS

```
#On charge les données du CPS 1985  
data("CPS1985")  
CPS<-data.table(CPS1985)
```

Note : j'utilise essentiellement le package `data.table` pour les manipulations de données et calculs à la main.

- ▶ Avantages : gros gains de temps sur des données volumineuses + syntaxe pauvre en fonctions (→ facile à mémoriser selon moi)
- ▶ N'hésitez pas à avoir recours au `tidyverse` si vous vous sentez plus à l'aise

Explorer rapidement les données du CPS

#On regarde rapidement ce à quoi cela ressemble

```
head(CPS)
```

```
##      wage education experience age ethnicity region gender occupation
## 1:  5.10          8            21   35 hispanic other female    worker
## 2:  4.95          9            42   57    cauc other female    worker
## 3:  6.67         12            1   19    cauc other male     worker
## 4:  4.00         12            4   22    cauc other male     worker
## 5:  7.50         12            17  35    cauc other male     worker
## 6: 13.07         13            9   28    cauc other male     worker
##              sector union married
## 1: manufacturing   no    yes
## 2: manufacturing   no    yes
## 3: manufacturing   no     no
## 4:        other     no     no
## 5:        other     no    yes
## 6:        other    yes     no
```

Principe général des moindres carrés ordinaires

Premier exemple : régresser le salaire (wage) sur l'éducation (education)

- ▶ **Salaire horaire** mesuré en \$ par heure et éducation en **années passées dans le système scolaire**
- ▶ Utiliser la fonction préimplémentée de R `lm`
- ▶ **Quelles quantités et objets intéressants cette régression permet-elle de construire ?**
- ▶ Quelles sont les **propriétés** et les **relations** qui lient ces objets ?

La régressions linéaire renvoie un jeu de coefficients, associé à la constante et la variable indépendante

```
#On effectue la régression du salaire horaire sur l'éducation  
reg_wage_educ<-lm(wage ~ education,  
                     data=CPS)
```

```
#On inspecte les coefficients  
reg_wage_educ$coefficients
```

```
## (Intercept)    education  
##   -0.7459797    0.7504608
```

La régression linéaire renvoie, pour chaque individu, une valeur prédictée du salaire

#On regarde les valeurs du salaire prédit

```
head(reg_wage_educ$fitted.values)
```

```
##      1      2      3      4      5      6  
## 5.257706 6.008167 8.259549 8.259549 8.259549 9.010010
```

#Combien y en a-t-il ?

```
length(reg_wage_educ$fitted.values)
```

```
## [1] 534
```

La régression linéaire renvoie, pour chaque individu, un terme résiduel

#On regarde les termes résiduels

```
head(reg_wage_educ$residuals)
```

```
##          1          2          3          4          5          6  
## -0.1577063 -1.0581671 -1.5895493 -4.2595493 -0.7595493  4.0599899
```

#Combien y en a-t-il ?

```
length(reg_wage_educ$residuals)
```

```
## [1] 534
```

Le terme résiduel est la différence entre la valeur prédictée et la valeur réalisée du salaire

#On crée les variables qui correspondent à la décomposition, c'est-à-dire les
valeurs prédictées du salaire d'une part, les termes résiduels d'autre part
CPS\$predi_wage<-reg_wage_educ\$fitted.values
CPS\$resid_wage<-reg_wage_educ\$residuals

#On peut vérifier que la somme des deux variables est bien égale au salaire
all.equal(CPS\$predi_wage+CPS\$resid_wage,
 CPS\$wage)

[1] TRUE

La valeur prédictée du salaire est une fonction affine de l'éducation que l'on peut calculer grâce aux coefficients de la régression

#On peut vérifier que la valeur prédictée du salaire est bien une fonction affine
de l'éducation, qui peut s'écrire à l'aide des coefficients de la régression

```
all.equal(CPS$predi_wage,  
          reg_wage_educ$coefficients["(Intercept)"]+  
          CPS$education*  
          reg_wage_educ$coefficients["education"])
```

```
## [1] TRUE
```

Le terme résiduel est nul en moyenne et n'est pas corrélé à l'éducation

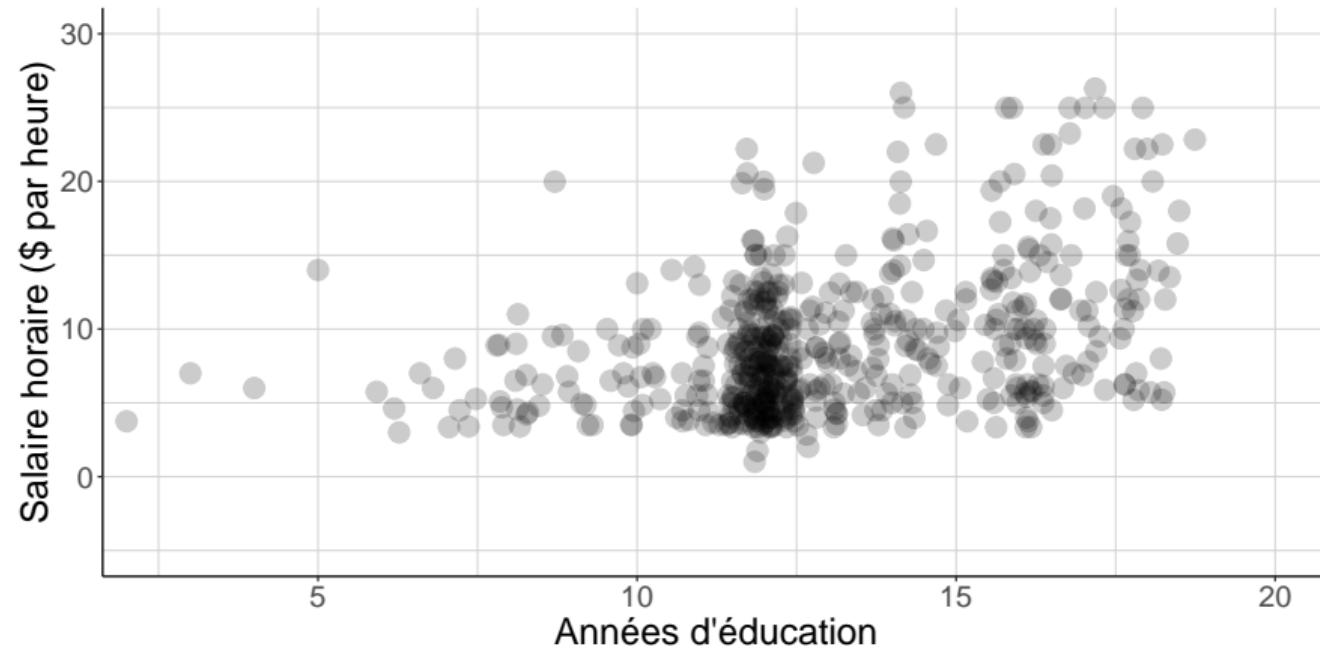
#On peut vérifier que la moyenne des résidus est nulle
all.equal(mean(CPS\$resid_wage),
 0)

[1] TRUE

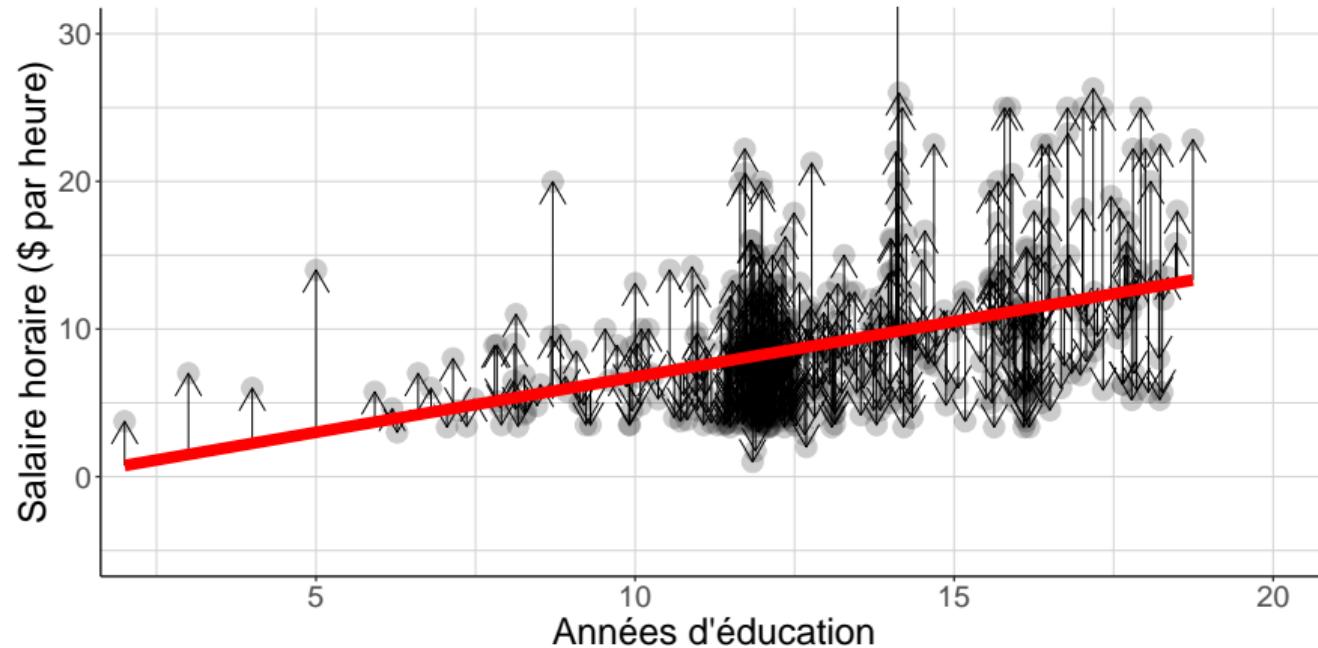
#On peut aussi vérifier que les résidus ne sont pas corrélés à l'éducation
all.equal(cov(CPS\$resid_wage,
 CPS\$education),
 0)

[1] TRUE

La régression linéaire décompose le salaire en une fonction affine de l'éducation et un terme résiduel, nul en moyenne et non-corrélu à l'éducation



La régression linéaire décompose le salaire en une fonction affine de l'éducation et un terme résiduel, nul en moyenne et non-corrélu à l'éducation



Une définition générale

La régression linéaire par les moindres carrés ordinaires d'une **variable aléatoire réelle** Y sur une **variable aléatoire X possiblement multidimensionnelle** définies sur la **même expérience aléatoire** est la décomposition de Y en la somme d'une **fonction affine de X** et d'un **terme résiduel** d'espérance nulle et non-corrélat à X .

- ▶ Rappel : une fonction affine de l'éducation
 - ▶ une fonction telle que si on considère deux individus différents, la différence entre les deux dépend linéairement de la différence entre leurs niveaux d'éducation
 - ▶ ou encore : la somme d'une constante et d'un terme linéaire en éducation

À retenir

- ▶ Cette décomposition **existe et est unique** dès lors que la matrice de variance-covariance de X est inversible (explicitation à suivre)
- ▶ Cela **ne dépend pas** de la structure causale du problème
 - ▶ La régression est bien définie si X est une cause de Y , mais aussi si Y est une cause de X , ou n'importe quelle autre chaîne causale
- ▶ Cela **ne dépend pas** non plus de la distribution de Y
 - ▶ Y peut être une variable continue, mais aussi dichotomique, discrète etc. sans rien changer au résultat

Problèmes de colinéarité

Un autre exemple : salaire horaire (wage) et sexe (gender)

- ▶ En partant des données du CPS, définir deux variables dichotomiques `male` et `female` (à valeur dans $\{0, 1\}$), l'une qui vaut 1 si l'individu est un homme et 0 sinon, l'autre qui vaut 1 si l'individu est une femme et 0 sinon (partir de la variable `gender`)
- ▶ Peut-on régresser le salaire horaire sur ces variables ?
- ▶ Pourquoi ?

Ce qu'il se passe si on essaie : on ne récupère pas le bon nombre de coefficients

```
#On crée deux variables dichotomiques indiquant le sexe déclaré dans
#l'enquête
CPS[,  
  c("female",  
    "male"):=list(as.numeric(gender=="female"),  
                  as.numeric(gender=="male"))]  
  
reg_wage_two_genders<-lm(wage ~male + female,  
                           data=CPS)  
  
reg_wage_two_genders$coefficients  
  
## (Intercept)      male      female  
##    7.878857    2.116056        NA
```

Pourquoi rencontre-t-on ce problème ?

- ▶ Les deux variables définissent des **catégories**
 - ▶ **mutuellement exclusives**
 - ▶ **qui recouvrent tout le champ des possibles**
- ▶ Cela implique `male + female == 1`
- ▶ On dit que ces variables sont **colinéaires** : l'une peut s'écrire comme combinaison linéaire de l'autre et de la constante
- ▶ La matrice de variance-covariance des variables indépendantes n'est pas inversible
- ▶ **On enfreint l'hypothèse sous laquelle il y a unicité de la solution** au problème des moindres carrés ordinaires

Les deux variables sont colinéaires et la matrice de variance-covariance n'est pas inversible

```
#On peut vérifier que la somme des deux variables est toujours égale à 1
all.equal(CPS$male+
          CPS$female,
          rep(1,
              nrow(CPS)))

## [1] TRUE

#On peut voir que la matrice de variance-covariance des deux variables n'est
# pas inversible
cov(CPS[,c("male",
           "female")])

##               male     female
## male    0.2487685 -0.2487685
## female -0.2487685  0.2487685
```

Il n'y a pas unicité au problème des moindres carrés ordinaires

```
#On estime le salaire moyen des femmes et le salaire moyen des hommes
mean_wages<-CPS[,  
                    list(mean_wage=mean(wage)),  
                    by=c("gender")]  
  
#On crée les valeurs prédites deux choix de coefficients différents  
CPS$valpred_model1<-  
  0+#constante  
  mean_wages[gender=="female"]$mean_wage*CPS$female+#coeff female  
  mean_wages[gender=="male"]$mean_wage*CPS$male#coeff male  
CPS$valpred_model2<-  
  mean_wages[gender=="female"]$mean_wage+#constante  
  0*CPS$female+#coeff female  
  (mean_wages[gender=="male"]$mean_wage-  
   mean_wages[gender=="female"]$mean_wage)*CPS$male#coeff male
```

Il n'y a pas unicité au problème des moindres carrés ordinaires

#On vérifie que l'on remplit bien toutes les contraintes que l'on veut
all(

#Nullité en moyenne des résidus

```
all.equal(mean(CPS$wage-CPS$valpred_model1),  
         0),
```

```
all.equal(mean(CPS$wage-CPS$valpred_model2),  
         0),
```

#Les résidus ne sont pas corrélés aux variables indépendantes

```
all.equal(cov(CPS$wage-CPS$valpred_model1, CPS$male),  
         0),
```

```
all.equal(cov(CPS$wage-CPS$valpred_model1, CPS$female),  
         0),
```

```
all.equal(cov(CPS$wage-CPS$valpred_model2, CPS$male),  
         0),
```

```
all.equal(cov(CPS$wage-CPS$valpred_model2, CPS$female),  
         0))
```

```
## [1] TRUE
```

En fait les valeurs prédites sont les mêmes pour les deux jeux de paramètres proposés

```
###En fait les valeurs prédites sont les mêmes  
all.equal(CPS$valpred_model1,  
          CPS$valpred_model2)
```

```
## [1] TRUE
```

Quelles solutions aux problèmes de colinéarité ?

- ▶ Il faut ajouter des **contraintes supplémentaires** sur les coefficients pour avoir l'unicité
- ▶ En général on en fixe un (ou plus !) à 0 ce qui revient à **omettre un (ou des) variable(s) indépendante(s)**
- ▶ Pour les variables dichotomiques, cela revient à choisir un **groupe de référence**
- ▶ Ce choix est en général anodin pour le comportement des estimateurs mais **change l'interprétation des coefficients**

Une solution possible : que valent les coefficients dans ce cas ?

```
#On régresse le salaire sur les deux indicatrices de sexe sans constante
noconstant_reg<-lm(wage~female + male - 1,
                     data=CPS)
noconstant_reg

##
## Call:
## lm(formula = wage ~ female + male - 1, data = CPS)
##
## Coefficients:
## female     male
## 7.879     9.995
```

Une solution possible : que valent les coefficients dans ce cas ?

```
#On compare les résultats aux salaires moyens par sexe  
all.equal(as.numeric(noconstant_reg$coefficients["female"]),  
         mean(CPS[female==1]$wage))
```

```
## [1] TRUE
```

```
all.equal(as.numeric(noconstant_reg$coefficients["male"]),  
         mean(CPS[male==1]$wage))
```

```
## [1] TRUE
```

Et dans ce cas ?

```
#On régresse le salaire sur la constante et l'indicatrice d'être un homme
omitfemale_reg<-lm(wage~male,
                      data=CPS)
omitfemale_reg

##
## Call:
## lm(formula = wage ~ male, data = CPS)
##
## Coefficients:
## (Intercept)      male
##           7.879       2.116
```

Et dans ce cas ?

```
#On compare les résultats aux salaires moyens par sexe  
all.equal(as.numeric(omitfemale_reg$coefficients["(Intercept)"]),  
         mean(CPS[female==1]$wage))
```

```
## [1] TRUE
```

```
all.equal(as.numeric(omitfemale_reg$coefficients["male"]),  
         mean(CPS[male==1]$wage)-mean(CPS[female==1]$wage))
```

```
## [1] TRUE
```

Pourquoi le nom de “moindres carrés ordinaires” ?

Comment interpréter la valeur prédictive ?

- ▶ Le terme résiduel – la différence entre salaire prédict et salaire réalisé – est **nul en moyenne**
- ▶ Sa **variance** quantifie la fréquence à laquelle il prend des valeurs éloignées de 0
- ▶ Sa variance est la moyenne sur toute la population du carré de l'écart entre salaire prédict et salaire réalisé $\mathbb{E}[(\hat{Y} - Y)^2]$
- ▶ Que peut-on dire à son propos ?

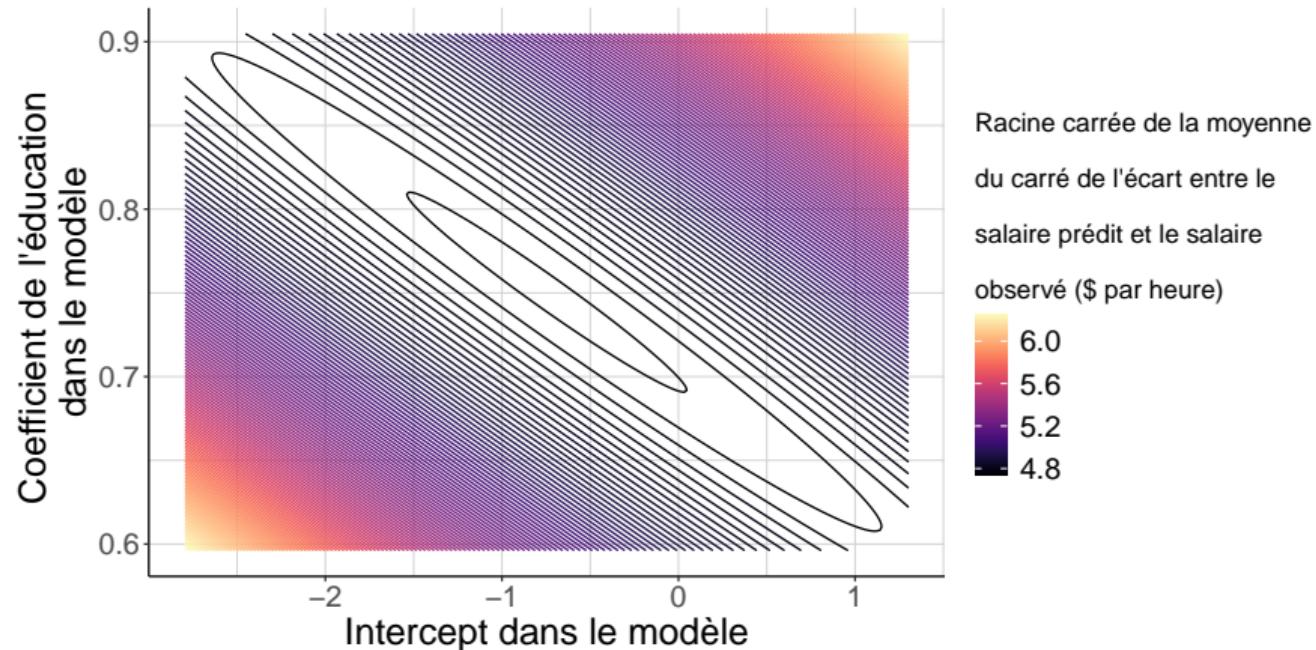
Comparer à toutes les façons de prédire le salaire comme une fonction affine de l'éducation

```
#On crée une fonction qui pour tout jeu de réels a et b calcule la
# moyenne du carré de l'écart entre le salaire horaire et le salaire
# prédit
ecart_carre<-function(a,b){

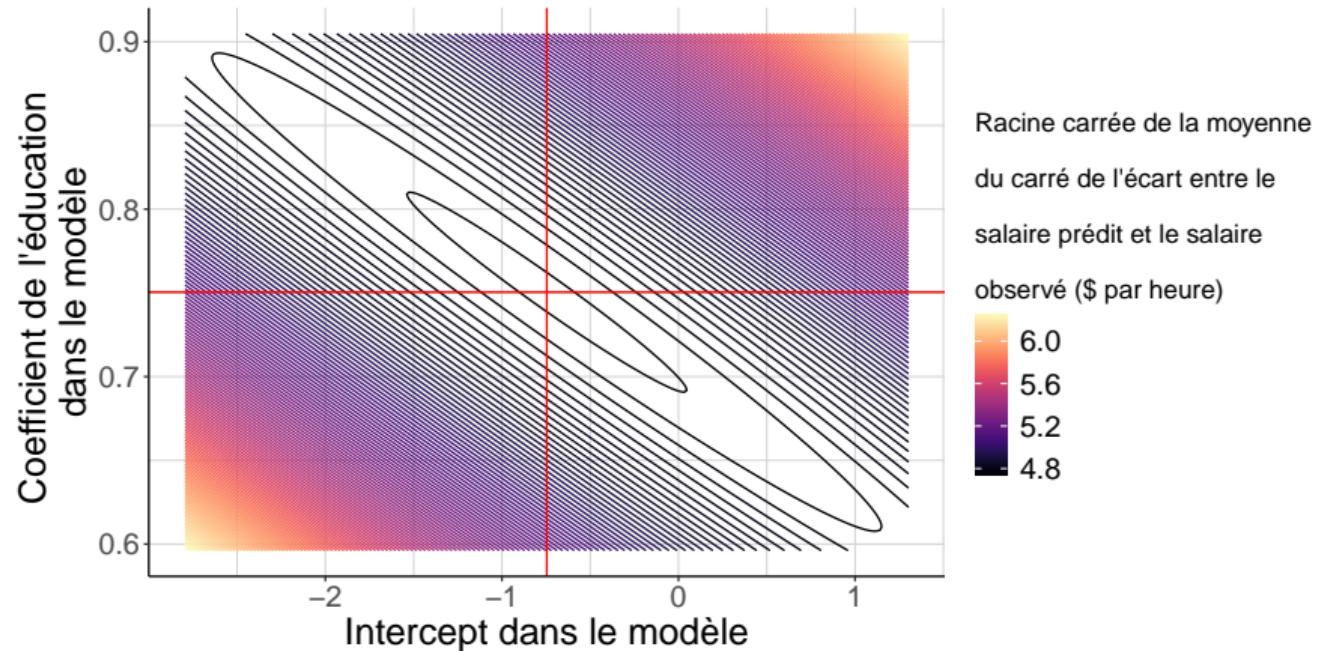
  mean((CPS$wage-a-b*CPS$education)^2)

}
```

Le salaire prédit par les moindres carrés ordinaires est la fonction affine de l'éducation qui minimise la moyenne du carré de l'écart entre salaire prédict et salaire réalisé



Le salaire prédict par les moindres carrés ordinaires est la fonction affine de l'éducation qui minimise la moyenne du carré de l'écart entre salaire prédict et salaire réalisé

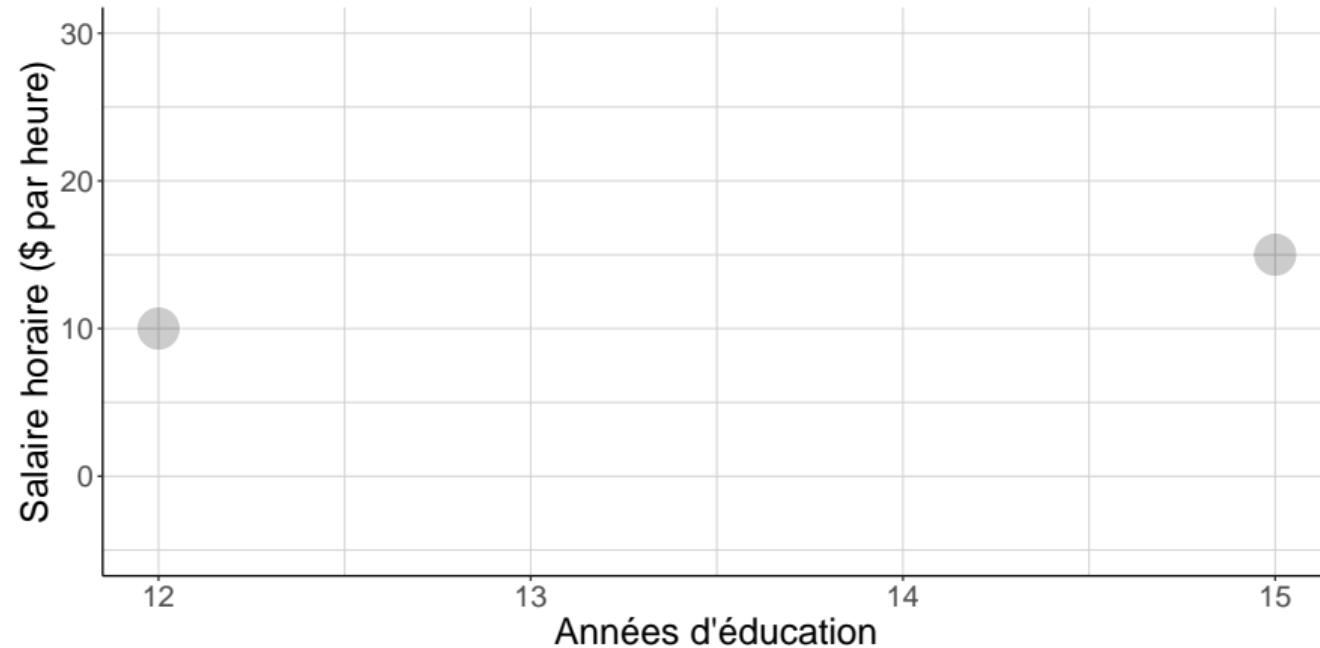


La régression linéaire simple est une moyenne de comparaisons
deux à deux

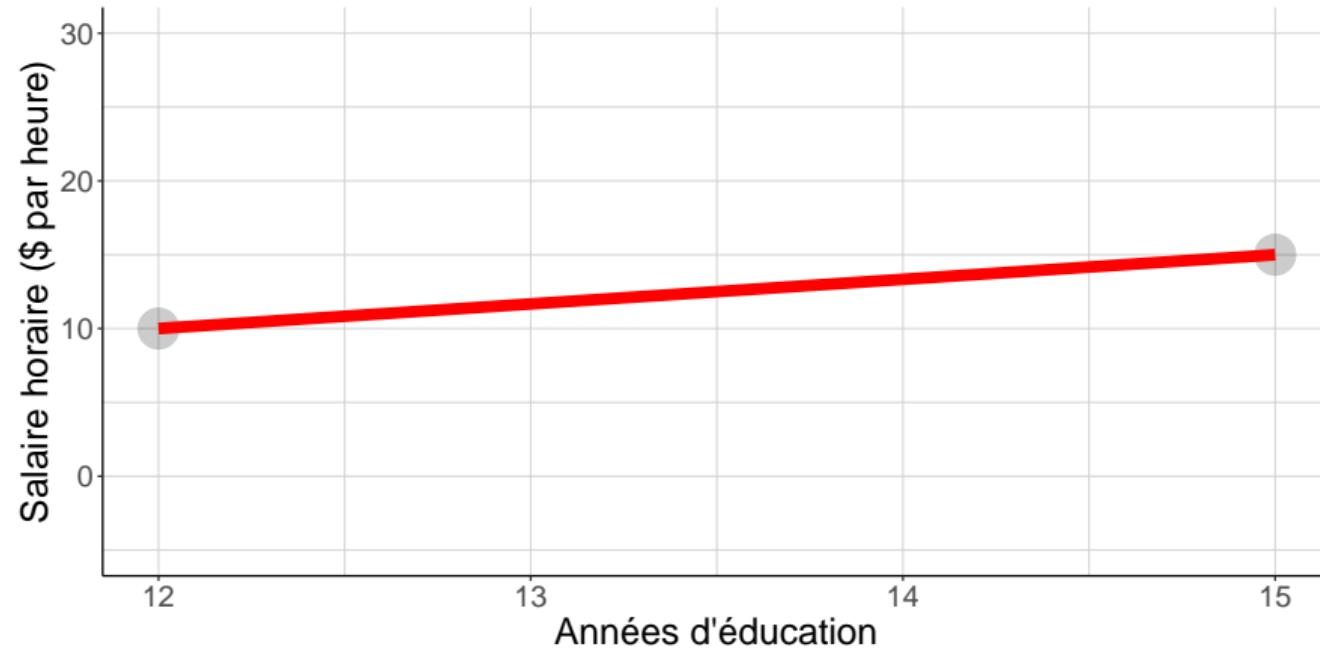
Comment interpréter le coefficient sur la variable d'éducation ?

- ▶ On revient à la régression du salaire horaire (`wage`) sur l'éducation (`education`)
- ▶ Que quantifie le coefficient qui porte sur la variable d'éducation ?
- ▶ Rappel : il n'y a ici aucune notion de causalité → interdiction d'utiliser le mot "effet"

Dans un monde fictif dans lequel il n'y a que deux niveaux d'éducation, et un salaire horaire par niveau



Dans un monde fictif dans lequel il n'y a que deux niveaux d'éducation, et un salaire horaire par niveau



Généraliser cette interprétation au monde réel ?

- ▶ De **nombreux niveaux d'éducation** différents
- ▶ Pour chacun d'entre eux de **nombreuses valeurs du salaire** possibles
- ▶ Comment s'en sortir ?

Considérer chaque couple de salariés séparément et faire la même comparaison

#On crée tous les couples de points possibles

```
CPS$key<-1  
CPS_couple<-merge(CPS[,c("key","wage","education")],  
                    CPS[,c("key","wage","education")],  
                    by="key",  
                    allow.cartesian=TRUE)
```

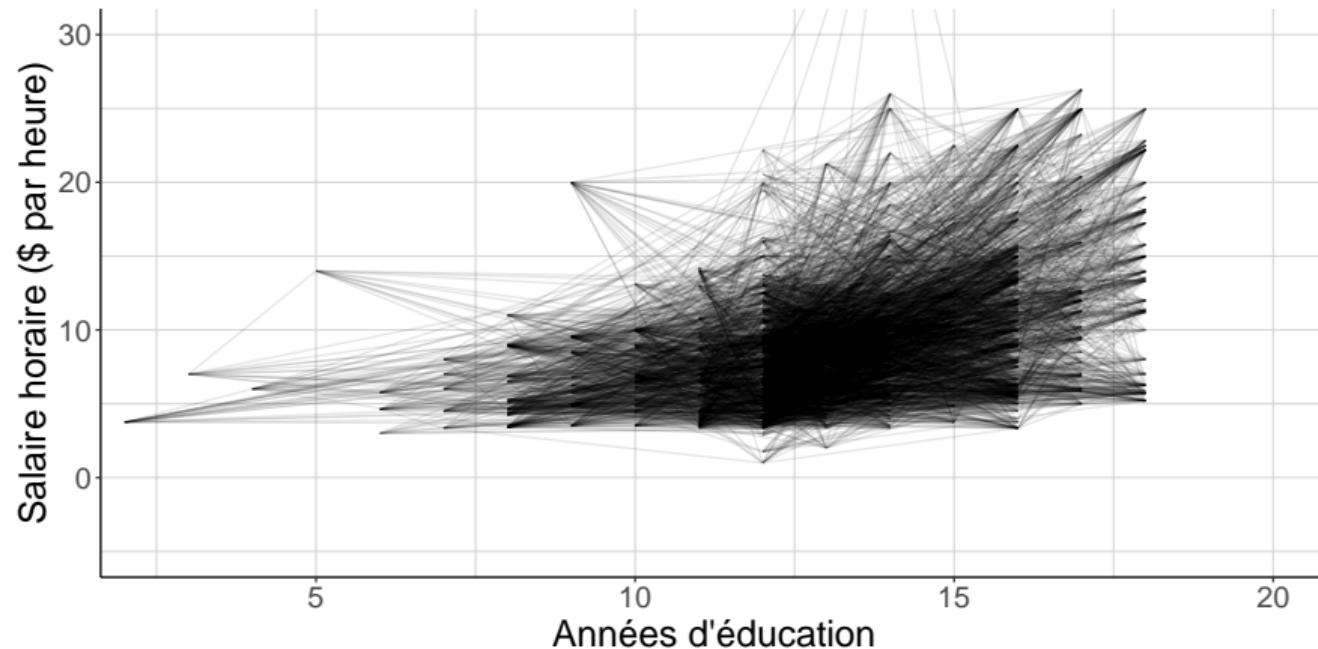
#On calcule la pente pour chacun d'entre eux lorsque le niveau d'éducation
diffère

```
CPS_couple[education.x!=education.y,  
           pente:=(wage.x-wage.y)/(education.x-education.y)]
```

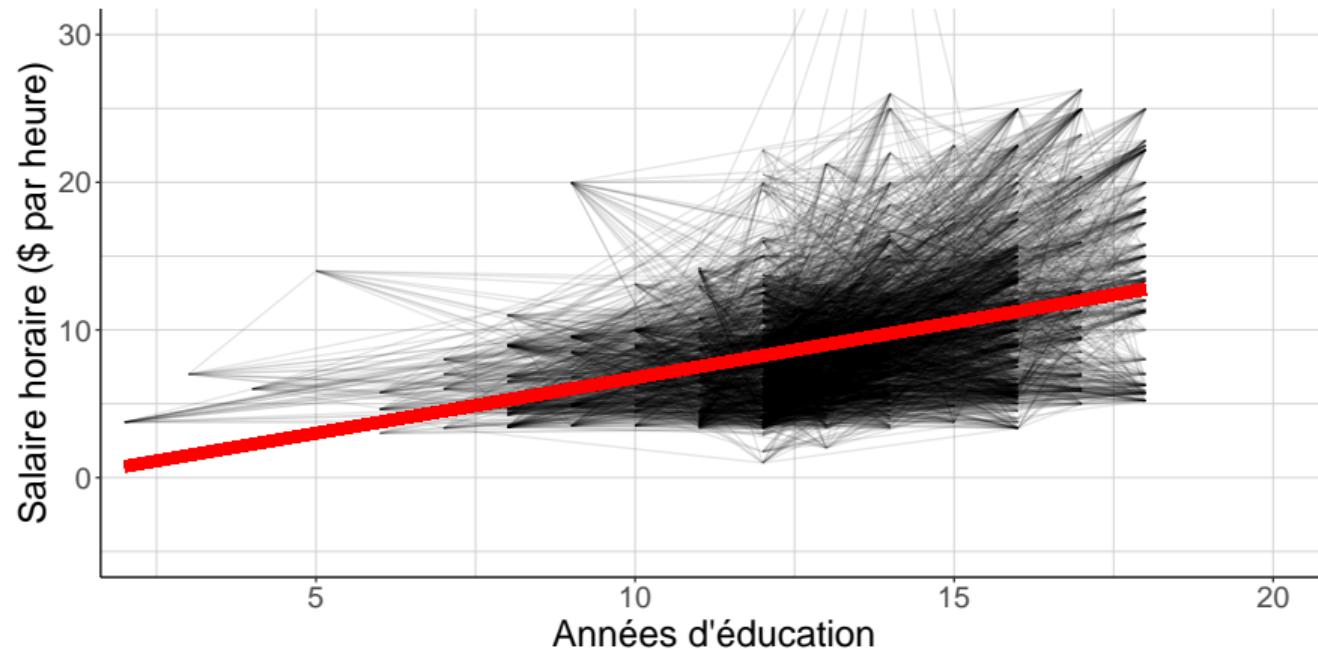
Il faut juste faire attention à la façon dont on pondère

```
#On calcule le carré de l'écart d'éducation
CPS_couple[,  
           carre_ecart_educ:=(education.x-education.y)^2]  
  
#On estime enfin la moyenne pondérée des pentes
pente_moyenne<-CPS_couple[education.x!=education.y,  
                           list(sum(pente*carre_ecart_educ)/  
                               sum(carre_ecart_educ))]  
  
#On peut enfin comparer les résultats des deux approches
all.equal(as.numeric(reg_wage_educ$coefficients["education"]),
          as.numeric(pente_moyenne))  
  
## [1] TRUE
```

Le coefficient sur la variable d'éducation est la moyenne des pentes calculées pour toutes les comparaisons deux-à-deux possibles de salariés, en donnant davantage de poids aux couples de salariés dont les niveaux d'éducation diffèrent le plus



Le coefficient sur la variable d'éducation est la moyenne des pentes calculées pour toutes les comparaisons deux-à-deux possibles de salariés, en donnant davantage de poids aux couples de salariés dont les niveaux d'éducation diffèrent le plus



Comparaisons avec l'espérance conditionnelle

Qu'est-ce que l'espérance conditionnelle ?

- ▶ Concept un peu fondamental de probabilité
- ▶ Schématiquement :
 - ▶ L'espérance $\mathbb{E}[Y]$ est la **moyenne de Y dans toute la population** (\neq l'échantillon observé)
 - ▶ L'espérance conditionnelle $\mathbb{E}[Y | X = x]$ est la moyenne de Y *dans la sous-population des individus tels que $X = x$*

Quelle information utilisent les moindres carrés ordinaires ?

Les régressions linéaires font-elles appel à une autre information que l'espérance conditionnelle ?

- ▶ Calculer pour chaque niveau d'éducation (`education`) possible le salaire (`wage`) moyen des salariés ayant ce niveau d'éducation
- ▶ Créer une variable `mean_wage_educ` qui associe à chaque salarié le salaire moyen des salariés ayant le même niveau d'éducation que lui
- ▶ Régresser cette variable de salaire moyen sur le niveau d'éducation. Quelle est la valeur du coefficient sur la variable d'éducation ?

La régression linéaire par les moindres carrés ordinaires ne dépend que de l'espérance conditionnelle

```
#On crée une variable égale à la moyenne du salaire pour chaque niveau
# d'éducation
mean_wage_educ<-CPS[,  
                      list(mean_wage=mean(wage)),  
                      by="education"]  
  
CPS<-merge(CPS,  
            mean_wage_educ,  
            by=c("education"))
```

La régression linéaire par les moindres carrés ordinaires ne dépend que de l'espérance conditionnelle

#On estime la régression de la moyenne conditionnelle du salaire horaire sur # l'éducation

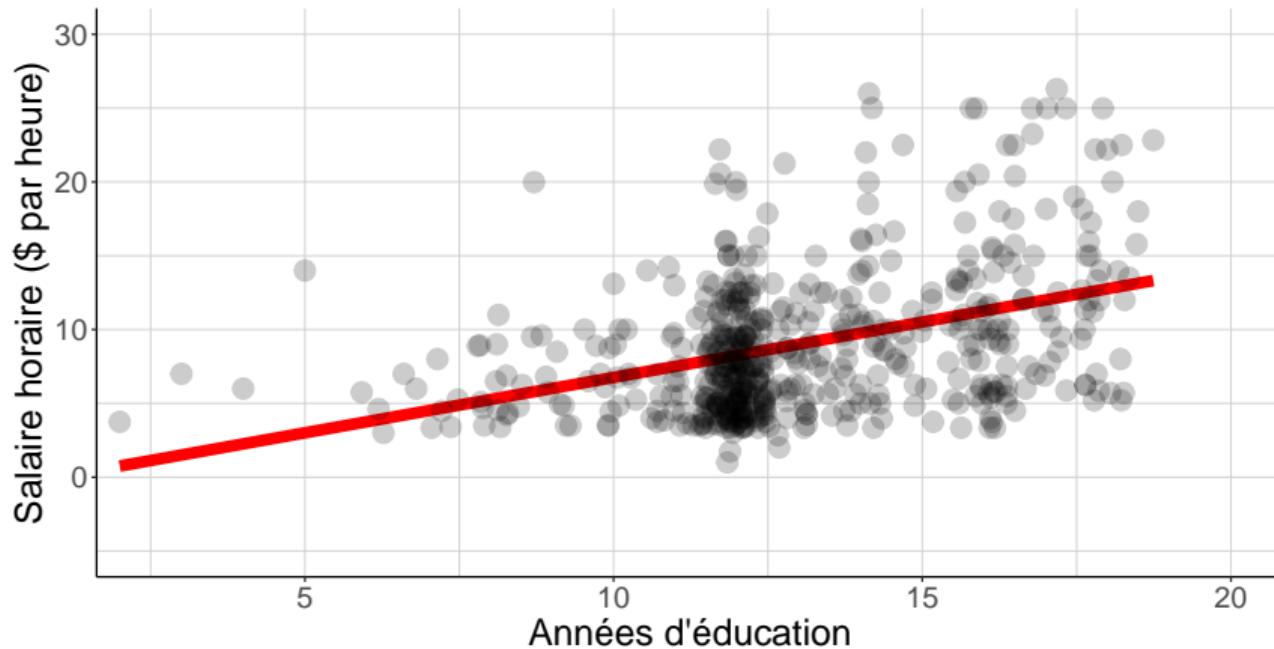
```
reg_wage_mean_educ<-lm(mean_wage~education,  
                         data=CPS)
```

#On vérifie que les deux jeux de coefficients sont bien égaux

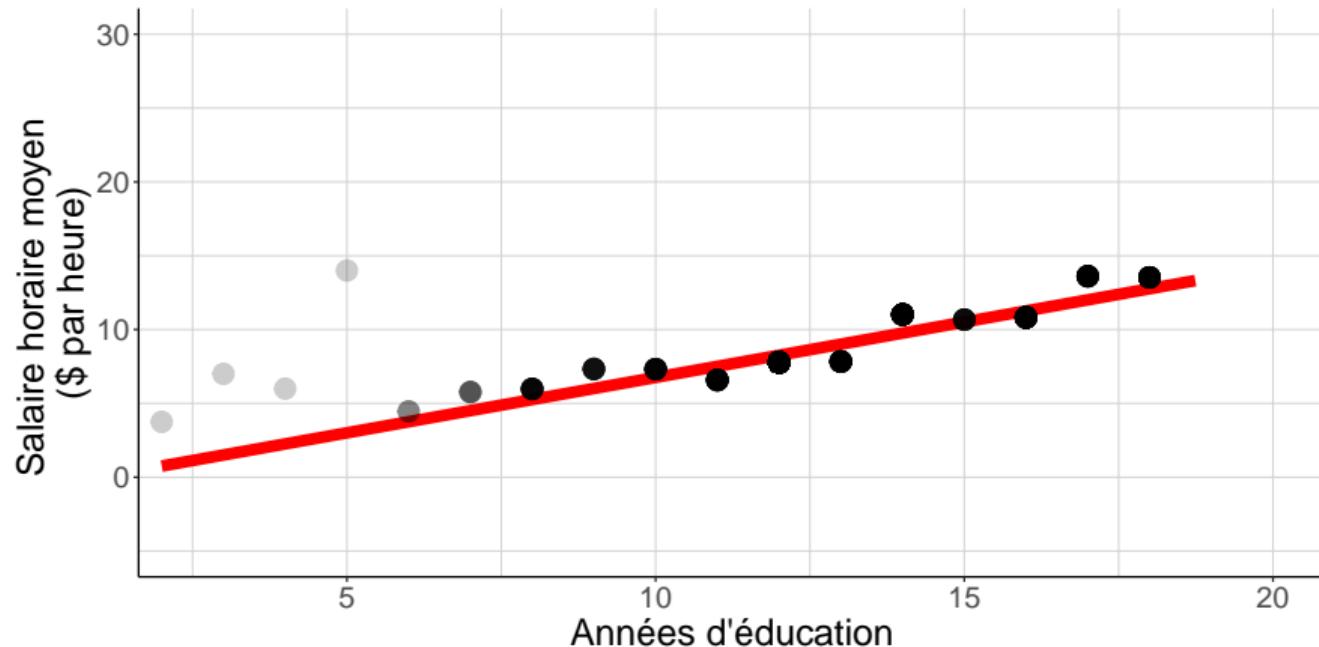
```
all.equal(reg_wage_educ$coefficients,  
          reg_wage_mean_educ$coefficients)
```

```
## [1] TRUE
```

La régression linéaire par les moindres carrés ordinaires ne dépend que de l'espérance conditionnelle



La régression linéaire par les moindres carrés ordinaires ne dépend que de l'espérance conditionnelle



Retrouver l'espérance conditionnelle à partir des moindres carrés ordinaires

Retour sur les variables dichotomiques

- ▶ Créer 4 variables dichotomiques croisant l'appartenance à un sexe d'une part (*gender*), et le fait d'avoir un niveau d'éducation inférieur ou supérieur à 16 années passées dans le système scolaire d'autre part (*education*)
- ▶ Régresser le salaire horaire sur ces 4 variables (en omettant la constante pour éviter les problèmes de colinéarité)
- ▶ Quelle est la valeur des différents coefficients ?

La régression linéaire par les moindres carrés ordinaires coïncide avec l'espérance conditionnelle pour la régression saturée

La régression linéaire par les moindres carrés ordinaires coïncide avec l'espérance conditionnelle pour la régression saturée

```
#On régresse le salaire sur ces 4 indicatrices (en omettant la constante)
reg_sat<-lm(wage~women_college +
              women_noncollege +
              men_college +
              men_noncollege -
              1,
              data=CPS)
```

La régression linéaire par les moindres carrés ordinaires coïncide avec l'espérance conditionnelle pour la régression saturée

Une autre façon de définir ces groupes : par des interactions

Plus facile pour faire le découpage que l'on a fait ici :

	female	college	college*female
Femmes diplômées	1	1	1
Femmes non-diplômées	1	0	0
Hommes diplômés	0	1	0
Hommes non-diplômés	0	0	0

Attention : on change l'interprétation des coefficients

Quelle est la valeur des coefficients dans ce cas ?

```
#On crée aussi les variables représentant les sexes et le niveau d'éducation  
# scolaire  
CPS$college<-as.numeric(CPS$education>=16)  
CPS$female<-as.numeric(CPS$gender=="female")  
  
#On peut comparer avec une régression spécifiée en interagissant les  
# variables  
reg_interact<-lm(wage~college + female + college*female,  
                  data=CPS)
```

La constante est la moyenne dans le groupe de référence

```
#On vérifie que la constante est égale au salaire moyen des hommes
# non-diplômés
all.equal(as.numeric(reg_interact$coefficients["(Intercept)"]),
          mean(CPS[gender=="male" & college==0]$wage))

## [1] TRUE
```

Les coefficients sur les termes simples sont égaux à des différences de moyenne entre certains groupes et le groupe de référence

```
all(  
  #Le coefficient sur la variable college est égal à la différence entre le  
  # salaire moyen des hommes diplômés et non-diplômés  
  all.equal(as.numeric(reg_interact$coefficients["college"]),  
           mean(CPS[gender=="male" & college==1]$wage)-  
           mean(CPS[gender=="male" & college==0]$wage)),  
  #Le coefficient sur la variable female est égal à la différence entre le  
  # salaire moyen des femmes non-diplômées et le salaire moyen des hommes  
  # non-diplômés  
  all.equal(as.numeric(reg_interact$coefficients["female"]),  
           mean(CPS[gender=="female" & college==0]$wage)-  
           mean(CPS[gender=="male" & college==0]$wage)))  
  
## [1] TRUE
```

Le coefficient sur le terme interagi est une différence de différences entre groupes

```
#Le coefficient sur le terme college*female est la différence entre le
#salaire moyen des femmes diplômées et non-diplômées, moins la différence
#entre le salaire moyen des hommes diplômés et non-diplômés
all.equal(as.numeric(reg_interact$coefficients["college:female"]),
          (mean(CPS[gender=="female" & college==1]$wage)-
           mean(CPS[gender=="female" & college==0]$wage))-
          (mean(CPS[gender=="male" & college==1]$wage)-
           mean(CPS[gender=="male" & college==0]$wage)))
## [1] TRUE
```

Passer aux choses sérieuses : régresser sur plusieurs variables

Lier les moindres carrés ordinaires aux contrastes

Un premier exemple : salaire horaire (wage) et sexe (gender), de nouveau !

- ▶ Toujours à partir des données du CPS
- ▶ Régresser le salaire horaire sur la variable indicatrice d'être une femme (`female`) et la variable de profession (`occupation`)
- ▶ Comment interpréter le coefficient sur la variable de sexe ?
 - ▶ En particulier comment se compare-t-il à l'écart de salaire entre femmes et hommes à l'intérieur de chaque *occupation* ?

On effectue la régression

```
#On régresse le salaire sur l'indicatrice d'être une femme et l'occupation (une
# sorte d'analogie de la CS)
reg_wage_gender_occupation<-lm(wage~female + occupation,
                                 data=CPS)
```

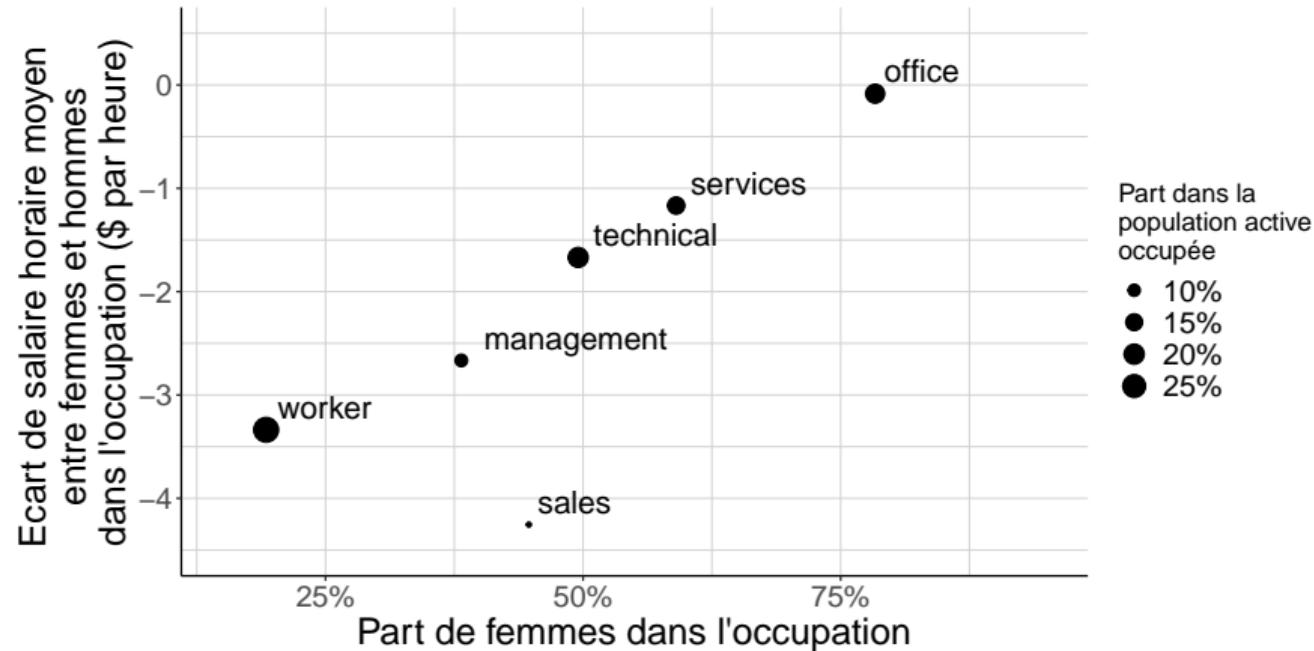
Le coefficient sur la variable de sexe est une moyenne pondérée d'écart de salaire moyen entre femmes et hommes à l'intérieur de chaque *occupation*

```
#On calcule pour chaque occupation :  
# 1. sa taille dans la population  
# 2. la part de femmes dans cette occupation  
# 3. l'écart de salaire moyen entre femmes et hommes dans cette occupation  
aggreg_CPS_occupation<-CPS[,  
                                list(size=sum(female+male),  
                                     share_female=sum(female)/sum(female+male),  
                                     var_female=sum(female)/sum(female+male)*  
                                         (1-sum(female)/sum(female+male)),  
                                     gender_gap=sum(female*wage)/sum(female)-  
                                         sum((male)*wage)/sum(male)),  
                                by="occupation"]
```

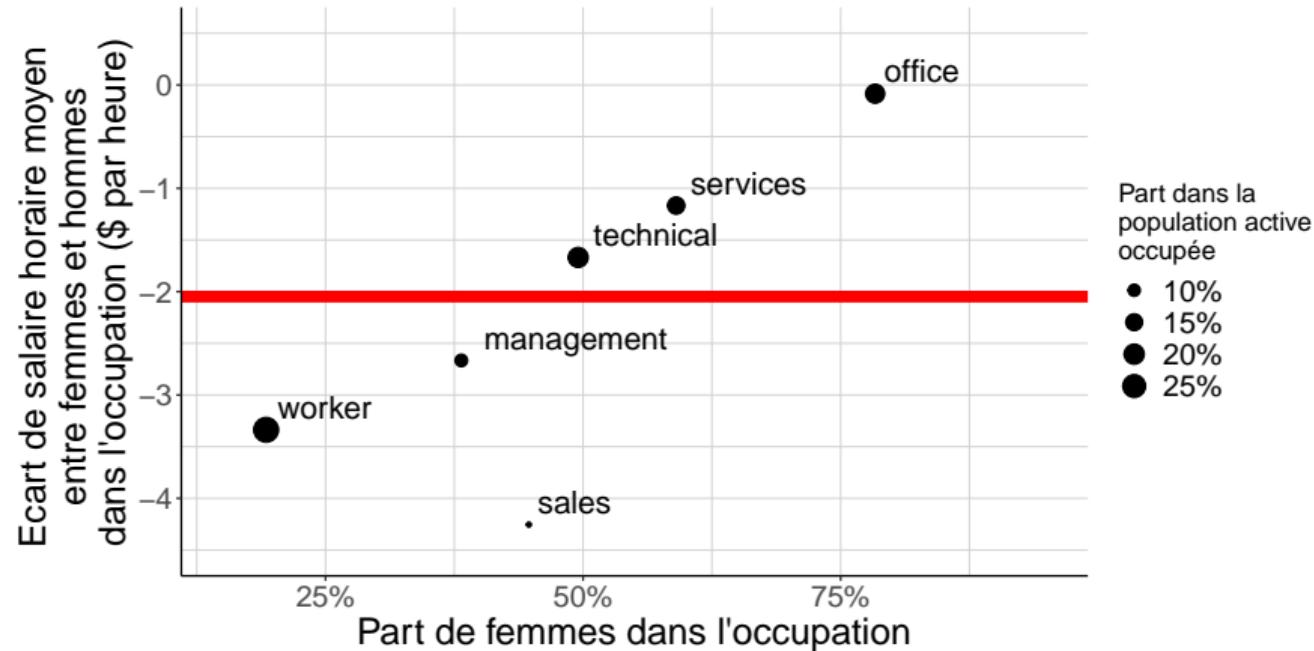
Le coefficient sur la variable de sexe est une moyenne pondérée d'écart de salaire moyen entre femmes et hommes à l'intérieur de chaque *occupation*

```
#On agrège les écarts de salaire dans chaque occupation en un seul écart
# moyen avec des poids proportionnels à la taille de chaque occupation et à
# la variance var_female = part de femmes * (1- part de femmes)
aggreg_gender_gap<-aggreg_CPS_occupation[,  
                                (aggreg_gender_gap=  
                                 sum(size*var_female*gender_gap)/  
                                 sum(size*var_female))]  
  
#On compare finalement cette quantité au coefficient de la régression
# linéaire
all.equal(aggreg_gender_gap,
          as.numeric(reg_wage_gender_occupation$coefficients["female"]))  
  
## [1] TRUE
```

Le coefficient sur la variable de sexe est une moyenne pondérée d'écart de salaire moyen entre femmes et hommes à l'intérieur de chaque *occupation*



Le coefficient sur la variable de sexe est une moyenne pondérée d'écart de salaire moyen entre femmes et hommes à l'intérieur de chaque *occupation*



Comment comprendre les poids ?

- ▶ **Si une profession est vide**, alors l'écart de salaire spécifique à cette profession n'est **pas défini**
 - ▶ On veut un poids nul pour les professions vides
 - ▶ En fait le poids est proportionnel à la taille de chaque profession (en nombre de salariés)
- ▶ **Si une profession est exclusivement masculine ou exclusivement féminine**, alors l'écart de salaire spécifique à cette profession n'est **pas défini**
 - ▶ On veut un poids nul pour les professions exclusivement féminines ou exclusivement masculines
 - ▶ En fait le poids est proportionnel au produit de la part des femmes par la part des hommes dans les salariés de la profession
 - ▶ Maximal pour les professions 50-50

Lier régression linéaire multiple et régression linéaire simple

Un nouvel exemple : salaire horaire (wage), éducation (education) et géographie (region)

- ▶ Toujours sur les données du CPS
- ▶ Régresser le salaire horaire (wage) sur l'éducation (education) et la variable qui indique si le salarié vit dans un Etat du Nord ou du Sud des Etats-Unis (region)
- ▶ Comment interpréter le coefficient sur la variable d'éducation ?
 - ▶ En particulier comment se compare-t-il aux coefficients que l'on obtiendrait en faisant la régression séparément pour chaque région ?

On effectue la régression

```
#On régresse le salaire sur l'éducation et l'indicatrice de région  
reg_full_spec<-lm(wage~education + region,  
                     data=CPS)
```

On effectue les régressions pour chaque région séparément

```
#On régresse le salaire sur l'éducation dans chaque groupe défini par la
#région et on récupère le coefficient correspondant
reg_educ_region<-lapply(levels(CPS$region),
                         function(x){
                           lm(wage~education,
                               data=CPS[region==x])$coefficients["education"]
                         })
reg_educ_region<-data.table(reg_educ_region)
reg_educ_region$region<-levels(CPS$region)
```

Le coefficient sur la variable d'éducation est une moyenne pondérée des coefficients des régressions spécifiques à chaque région

```
CPS_region<-CPS[,  
  list(taille_region=sum(as.numeric(wage>=0)),  
       variance=var(education)*  
         #Attention ici il faut corriger car le dénominateur  
         # utilisé pour le calcul de la variance est n-1  
         (sum(as.numeric(wage>=0))-1)/  
         sum(as.numeric(wage>=0))),  
       by="region"]  
CPS_region<-merge(CPS_region,  
  reg_educ_region,  
  by="region")
```

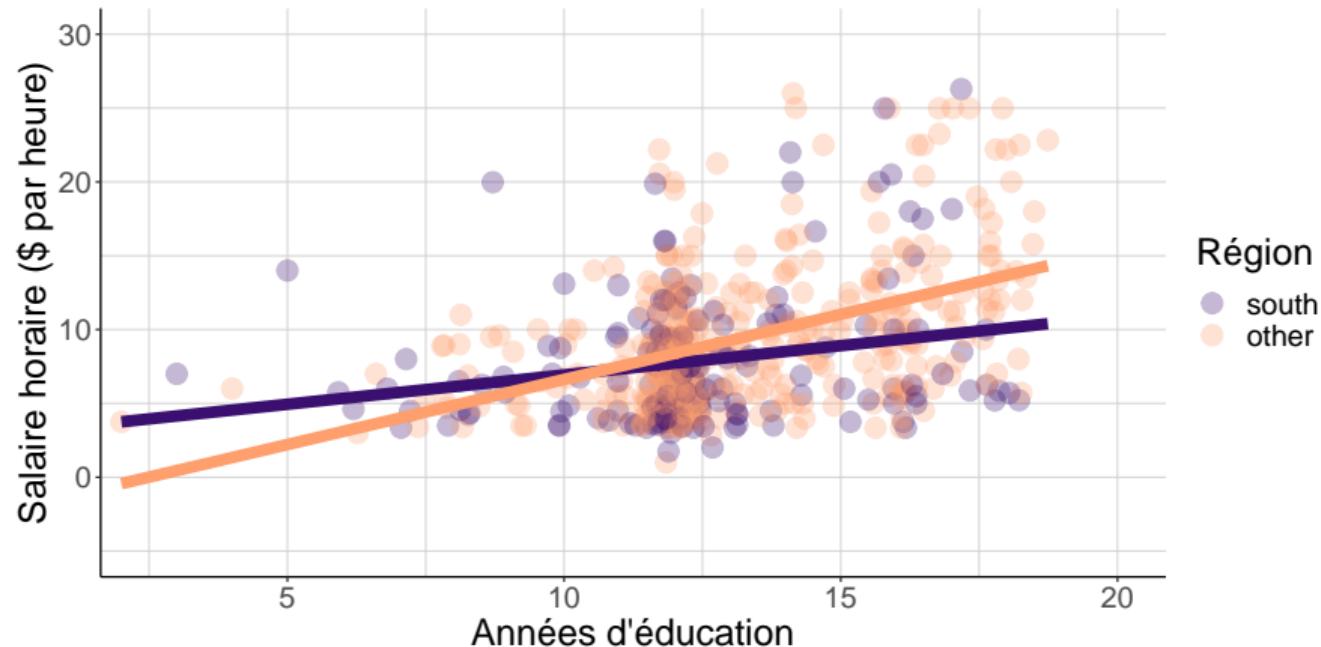
Le coefficient sur la variable d'éducation est une moyenne pondérée des coefficients des régressions spécifiques à chaque région

```
aggreg_coefficient<-CPS_region[,
  list(aggreg_coefficient=
    sum(taille_region*
        variance*
        as.numeric(reg_educ_region))/(
    sum(taille_region*
        variance))]

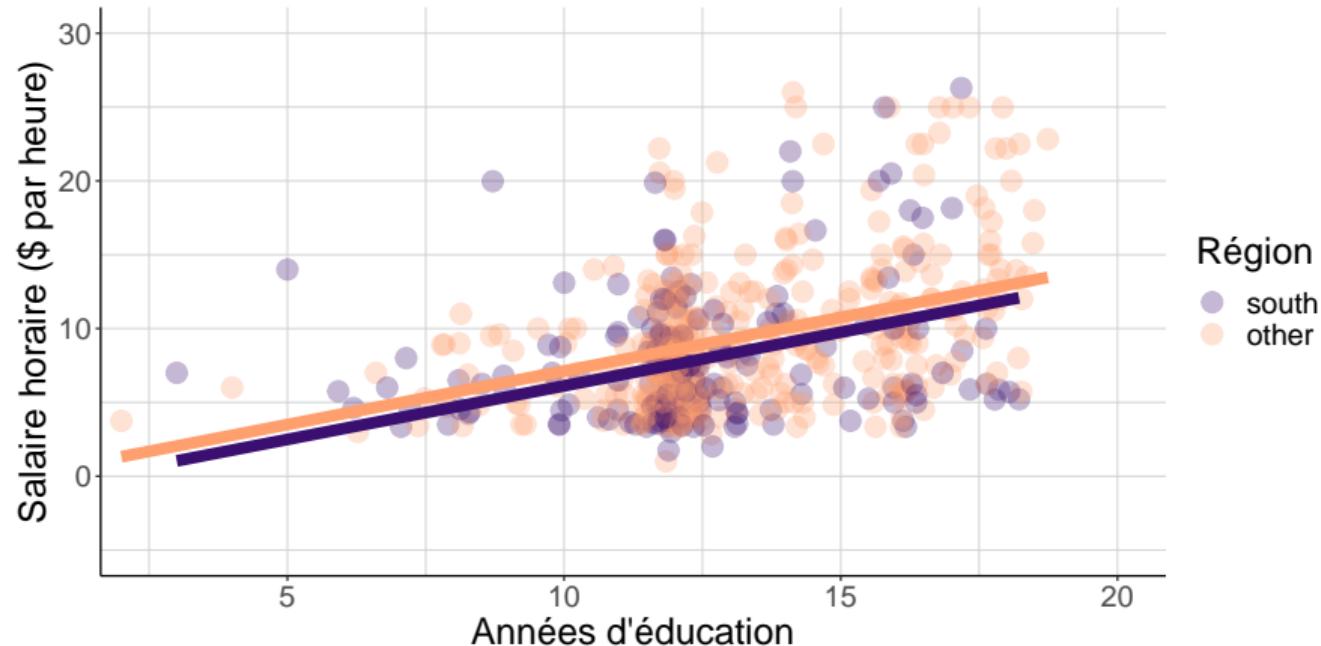
#On compare les deux coefficients
all.equal(as.numeric(aggreg_coefficient$aggreg_coefficient),
         as.numeric(reg_full_spec$coefficients["education"]))

## [1] TRUE
```

Le coefficient sur la variable d'éducation est une moyenne pondérée des coefficients des régressions spécifiques à chaque région



Le coefficient sur la variable d'éducation est une moyenne pondérée des coefficients des régressions spécifiques à chaque région



Interpréter ce résultat

- ▶ Dans chaque région le coefficient de l'éducation dans la régression du salaire horaire sur l'éducation spécifique à cette région résulte de multiples comparaisons entre salariés de cette région
- ▶ On fait donc :
 - ▶ Toutes les comparaisons possibles **entre salariés du Nord**
 - ▶ Toutes les comparaisons possibles **entre salariés du Sud**
 - ▶ **Mais on s'interdit de comparer des salariés du Nord et des salariés du Sud ensemble**

Comment comprendre les poids ?

- ▶ **Dans une région vide**, le coefficient de la régression du salaire horaire sur l'éducation n'est **pas défini**
 - ▶ On veut mettre un poids nul sur une telle région
 - ▶ En fait le poids est proportionnel à la taille de la région (en nombre de salariés)
- ▶ **Dans une région où tout le monde aurait le même niveau d'éducation**, le coefficient n'est **pas défini** (l'éducation est colinéaire à la constante)
 - ▶ On veut mettre un poids nul sur une telle région
 - ▶ En fait le poids est proportionnel à la variance de l'éducation dans cette région
- ▶ En fait ce sont les **mêmes poids** que dans l'exemple précédent

Le domaine de validité de ces résultats

- ▶ Les deux résultats précédents valent dans la mesure où on fait une régression du style :

$$Y = \beta X + \sum_{i=1}^d \gamma_i G_i + \epsilon$$

- ▶ X la variable indépendante qui nous intéresse particulièrement
- ▶ les G_i des indicatrices de groupe qui recouvrent tout l'espace des possibles
 $\sum_{i=1}^n G_i = 1$
- ▶ Quand on sort de ce cadre les choses peuvent devenir **un peu moins claires**
 - ▶ Discussion à venir lors des séances sur le contrôle des facteurs confondants

Augmenter peu à peu le nombre de variables indépendantes

Les moindres carrés ordinaires sont en un sens compatibles avec l'itération

- ▶ Régresser le salaire horaire (`wage`) sur l'éducation (`education`) et l'âge (`age`)
- ▶ Régresser le salaire horaire sur l'éducation, l'âge sur l'éducation, et les résidus de la première régression sur les résidus de la seconde.
- ▶ Quelle est la valeur du coefficient de la dernière régression ?

Les deux approches permettent d'estimer le même coefficient : celui de l'âge dans la régression du salaire horaire sur l'âge et l'éducation

#On régresse le salaire sur les variables d'âge et d'éducation

```
reg_fullmodel<-lm(wage~education + age,  
                    data=CPS)
```

#On régresse d'abord le salaire uniquement sur l'éducation

```
reg_wageeduc<-lm(wage~education,  
                   data=CPS)
```

#On régresse également l'éducation sur l'âge

```
reg_ageduc<-lm(age~education,  
                 data=CPS)
```

Les deux approches permettent d'estimer le même coefficient : celui de l'âge dans la régression du salaire horaire sur l'âge et l'éducation

#On récupère les résidus de chacune de ces régressions

```
CPS$reg_wageeduc_resid<-reg_wageeduc$residuals
```

```
CPS$reg_ageduc_resid<-reg_ageduc$residuals
```

#On régresse enfin les résidus de la deuxième régression sur ceux de la

troisième

```
reg_FWL<-lm(reg_wageeduc_resid~reg_ageduc_resid,  
               data=CPS)
```

#On compare enfin les coefficients de la première et de la dernière

régression

```
all.equal(as.numeric(reg_fullmodel$coefficients["age"]),  
         as.numeric(reg_FWL$coefficients["reg_ageduc_resid"]))
```

```
## [1] TRUE
```

Itérer les régressions

- ▶ Ce résultat se généralise facilement à un **nombre quelconque** de variables indépendantes
 - ▶ Régresser la variable dépendante et un lot de p variables indépendantes séparément sur un autre lot de q variables indépendantes
 - ▶ Puis les résidus de la première régression sur les p résidus des autres régressions
- ▶ On l'appelle souvent **théorème de Frisch-Waugh-Lovell**
- ▶ Rarement utilisé directement en pratique mais souvent très **utile pour comprendre la mécanique** qui se cache derrière les moindres carrés ordinaires
 - ▶ Une application : la séance sur les effets fixes dans les données de panel

Quelques remarques sur la décomposition de la variance

Notion de coefficient de détermination

Quelle est la variance du salaire horaire ?

- ▶ On repart de l'exemple initial en régressant le salaire horaire (`wage`) sur l'éducation (`education`)
- ▶ Peut-on écrire la variance du salaire horaire en utilisant les objets que l'on récupère à partir de cette régression ?

La variance du salaire horaire est la somme de la variance du salaire prédit et de celle des résidus

#On vérifie que la variance du salaire est bien égale à la somme de la variance
du salaire prédict et de la variance des résidus
all.equal(var(CPS\$wage),
 var(reg_wage_educ\$fitted.values)+var(reg_wage_educ\$residuals))
[1] TRUE

Le coefficient de détermination est le rapport de la variance du salaire prédit et de la variance du salaire réalisé

- ▶ Toujours **compris entre 0 et 1**
- ▶ Souvent noté (et appelé) R^2
- ▶ Dans le cas de la régression linéaire simple : **le carré du coefficient de corrélation** entre variable dépendante et variable indépendante

Le coefficient de détermination est le rapport de la variance du salaire prédit et de la variance du salaire réalisé

```
all(  
  #On vérifie que le R2 estimé par R est bien le rapport de la variance du  
  #salaire prédit et de la variance du salaire réalisé  
  all.equal(summary(reg_wage_educ)$r.squared,  
            var(reg_wage_educ$fitted.values)/var(CPS$wage)),  
  #On vérifie enfin que le coefficient de détermination est bien égal au  
  #carré de la corrélation entre le salaire et l'éducation  
  all.equal(summary(reg_wage_educ)$r.squared,  
            cor(CPS$wage,  
                 CPS$education)^2))  
  
## [1] TRUE
```

Interprétations du coefficient de détermination

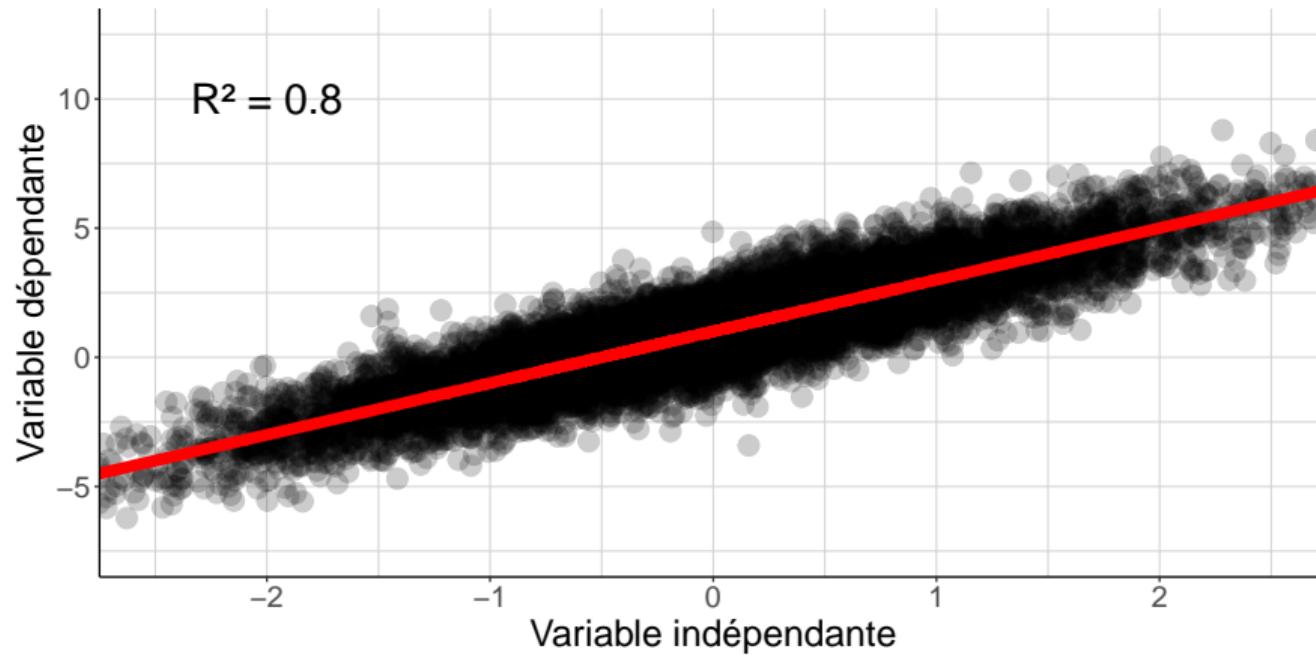
Le coefficient de détermination n'a en général aucune signification causale

- ▶ On dit parfois que le coefficient de détermination est la **part de la variance du salaire horaire expliquée par l'éducation**
- ▶ Ce concept de part expliquée n'a **aucune signification causale** ! C'est une mesure de corrélation
- ▶ C'est **purement comptable** :
 - ▶ 15% de la variance du salaire horaire correspondent à des différences de salaire entre salariés ayant un niveau d'éducation différent
 - ▶ Les 85% restant correspondent à des différences entre salariés ayant le même niveau d'éducation

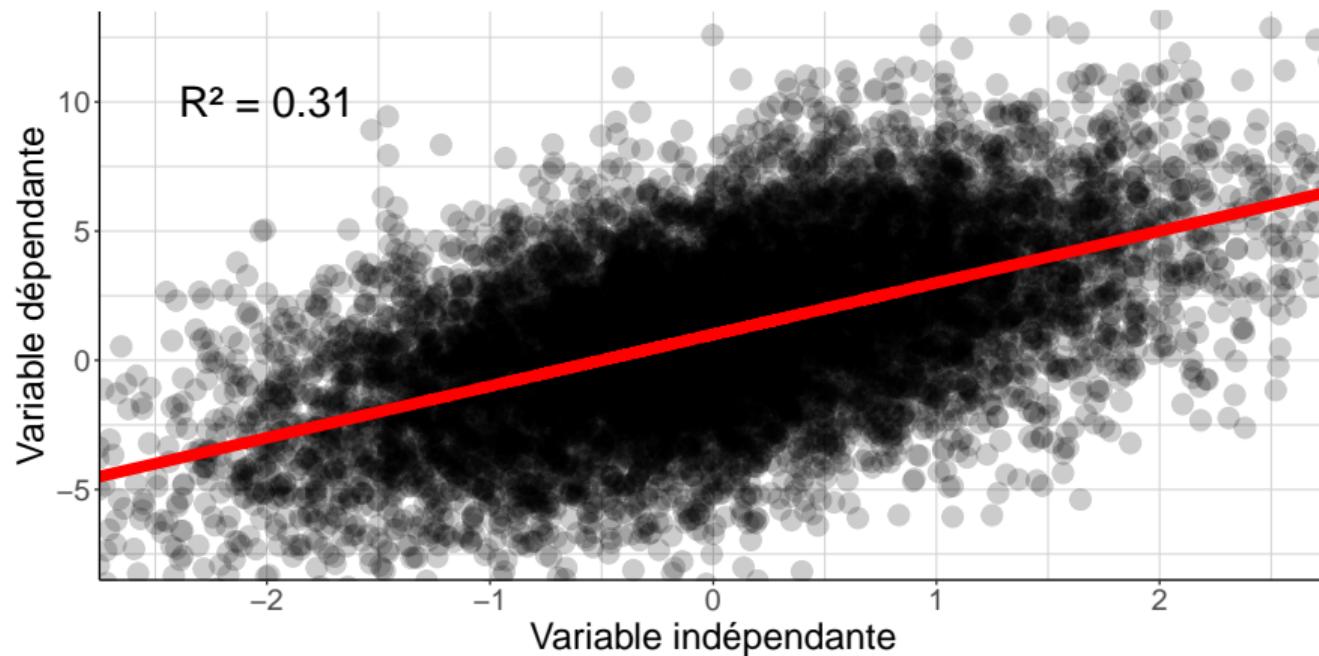
Une mesure de qualité ?

- ▶ Les moindres carrés ordinaires cherchent la meilleure approximation du salaire parmi toutes les fonctions affines de l'éducation
- ▶ En ce sens précis, le coefficient de détermination mesure la qualité, c'est-à-dire la **distance entre le salaire prédit et le salaire réalisé**
 - ▶ La raison est que $R^2 = \frac{\mathcal{V}(\hat{Y})}{\mathcal{V}(Y)} = 1 - \frac{\mathcal{V}(\epsilon)}{\mathcal{V}(Y)}$
 - ▶ Une mesure de la qualité qui n'a d'intérêt **que si on s'intéresse à la valeur prédictive** et pas par exemple aux coefficients

Une mesure de qualité ?



Une mesure de qualité ?



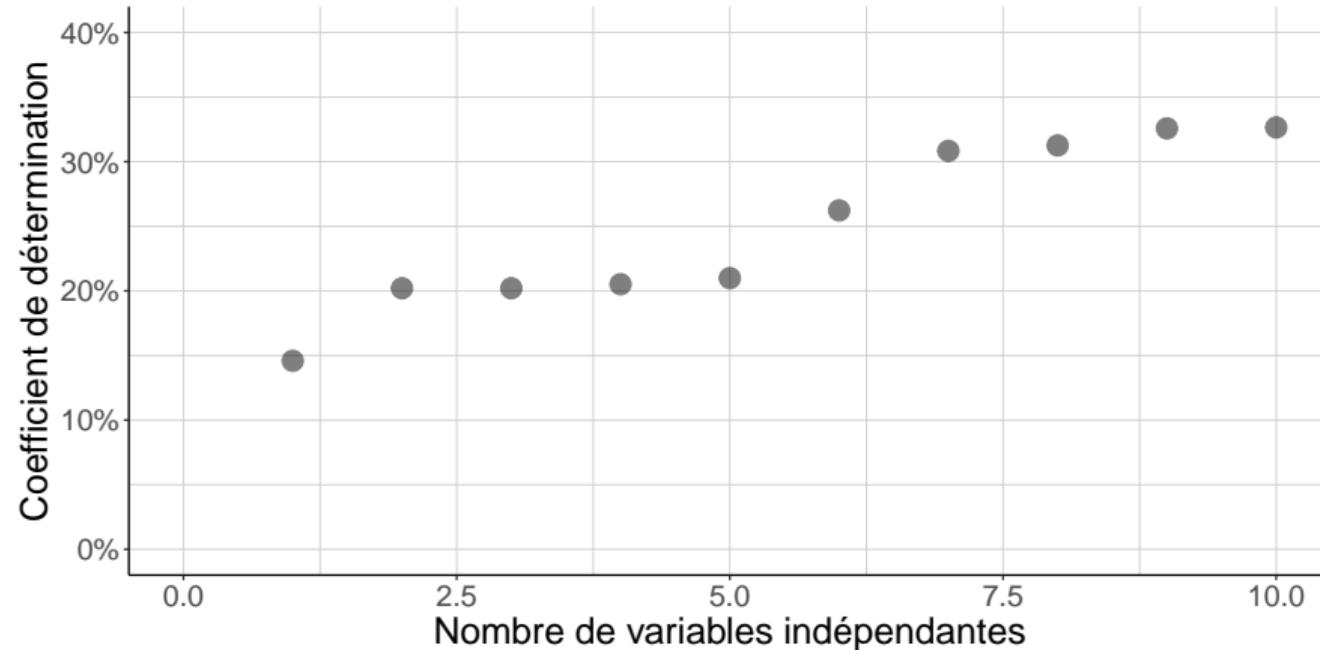
Problèmes posés par l'interprétation du coefficient de détermination

En sciences sociales le coefficient de détermination est souvent faible

- ▶ Pas un problème si on ne s'intéresse pas particulièrement à la valeur prédictive
- ▶ Exemple avec le salaire horaire (wage) comme variable dépendante

```
reg_coeff_determ<-function(nbrcov){  
  reg<-lm(as.formula(paste0("wage ~ ",  
                        paste0(colnames(CPS)[2:(1+nbrcov)],  
                               "#sélectionne les nbrcov premières  
                               # variables indépendantes  
                               collapse="+"))),  
          data=CPS)  
  
  summary(reg)$r.squared#on récupère le R² de la régression  
}  
  
#On applique cette fonction à un nombre croissant de variables  
iter_coeff_determ<-lapply(1:(ncol(CPS)-1), reg_coeff_determ)
```

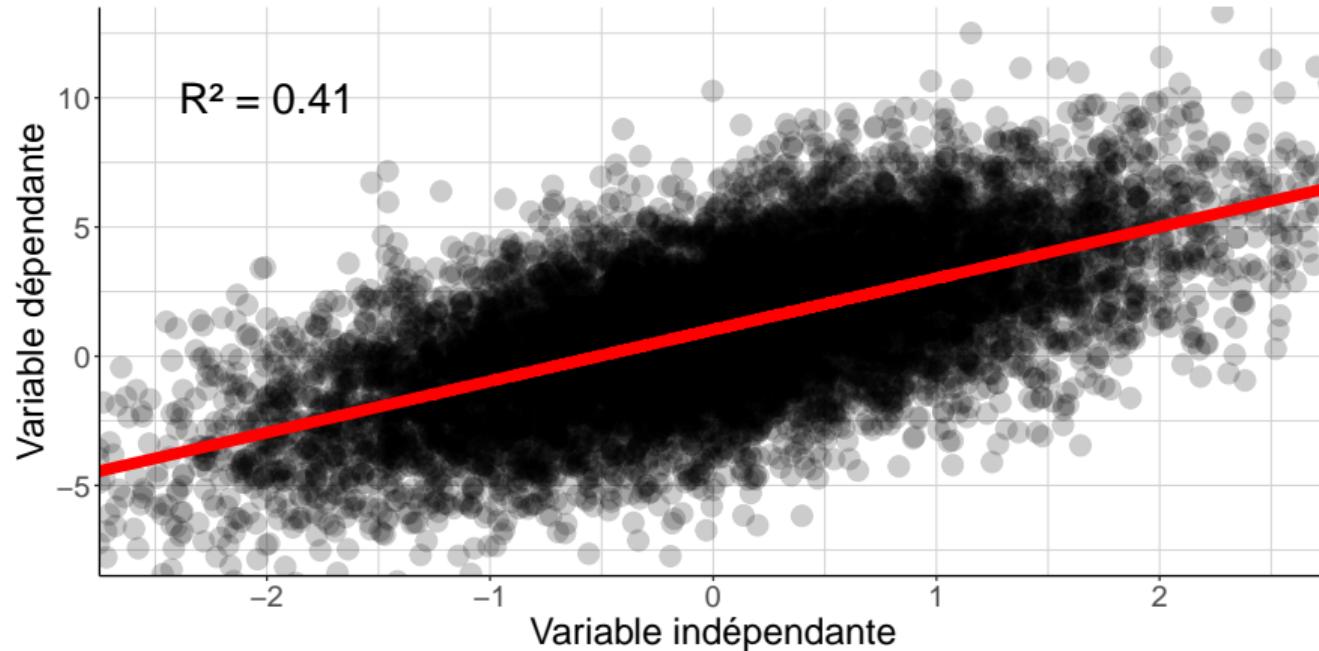
Même en mettant toutes les variables indépendantes disponibles, le coefficient de détermination est au mieux 33%



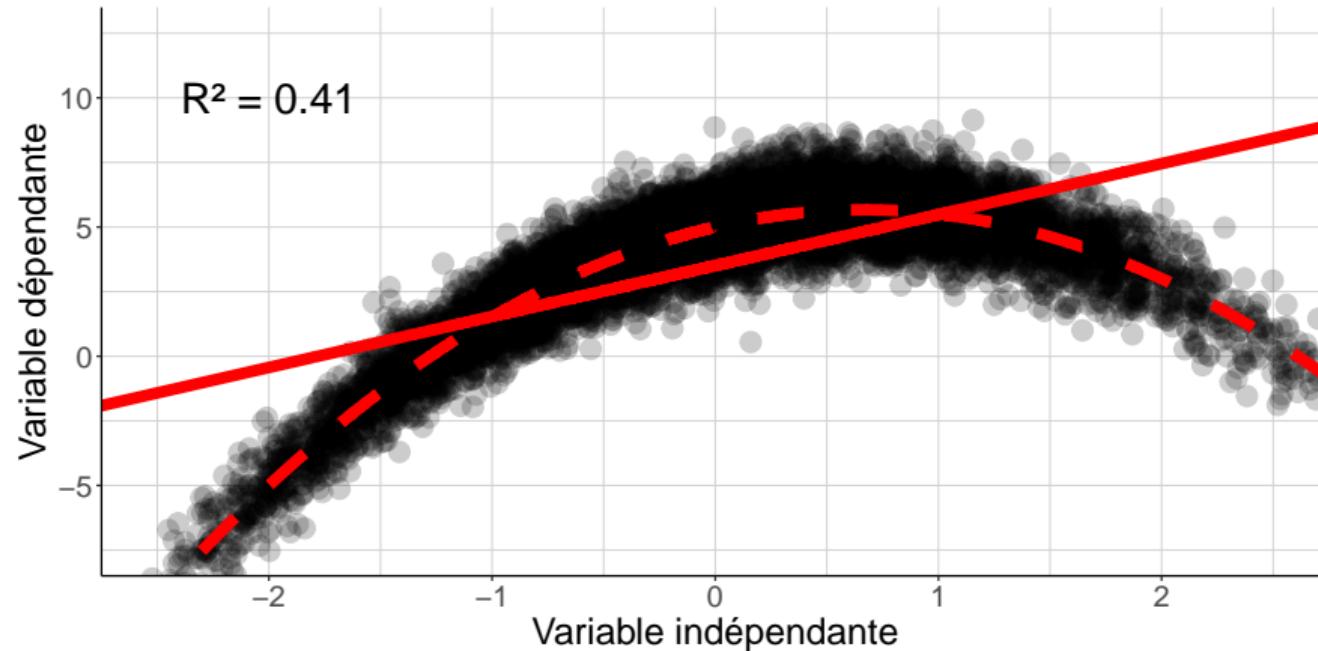
L'usage du coefficient de détermination pour apprécier l'intérêt de considérer des non-linéarités n'est pas évident

- ▶ Certes si $R^2 = 1$ alors l'espérance conditionnelle $Y = \mathbb{E}[Y | X]$ s'écrit comme une fonction affine de X et on a fini
 - ▶ Mais c'est juste que $\epsilon = 0$ dans ce cas
- ▶ Dans le cas général, une valeur faible de coefficient de détermination peut signifier
 - ▶ ou bien que l'espérance conditionnelle présente des non-linéarités → considérer d'autres façons d'approximer Y permettrait d'améliorer la prédiction
 - ▶ ou bien rien de tout cela

L'usage du coefficient de détermination pour apprécier l'intérêt de considérer des non-linéarités n'est pas évident



L'usage du coefficient de détermination pour apprécier l'intérêt de considérer des non-linéarités n'est pas évident



Estimation et inférence

Considérations générales

Des quantités estimées à l'estimateur des moindres carrés ordinaires

- ▶ Tout ce qui précède caractérise les **quantités estimées** par les MCO (\sim des différences de moyennes *prises dans toute la population*)
 - ▶ On a fait comme si estimer les coefficients à partir du CPS et parler des coefficients définis sur toute la population des salariés étatsuniens était la même chose
- ▶ On sait passer de la définition de ces quantités estimées...
 - ▶ la décomposition de la variable dépendante en une fonction affine etc.
- ▶ ... à l'expression de ces quantités en fonction de moyennes **prises dans toute la population**
- ▶ Il n'y a plus qu'à passer de ces moyennes dans toute la population à des moyennes prises **dans l'échantillon** → l'**estimateur** des MCO

Théorie asymptotique de l'estimateur des moindres carrés ordinaires

- ▶ Tout ce qu'il y a à savoir :
 - ▶ L'estimateur s'écrit en fonction de moyennes prises dans l'échantillon (\neq ces moyennes prises dans la population tout entière)
 - ▶ Quand **le nombre d'observations devient suffisamment grand** :
 - ▶ Ces moyennes dans l'échantillon sont une bonne approximation des moyennes dans la population
 - ▶ Et donc **l'estimateur approxime correctement la quantité estimée**
 - ▶ Et l'écart entre les deux s'identifie à des **fluctuations aléatoires dont on connaît la distribution**

Une quantification de l'incertitude sur les coefficients

- ▶ L'écart entre l'estimateur et les quantités estimées s'identifie à des **fluctuations aléatoires dont on connaît la distribution**
- ▶ La **variance de ces fluctuations** est le produit de deux termes :
 - ▶ l'**inverse du nombre d'observations**
 - ▶ un terme qui s'écrit en fonction de moyennes prises dans toute la population
→ discuté plus loin

Rôle du nombre d'observations

- ▶ La longueur des intervalles de confiance est proportionnelle à l'**écart-type** des fluctuations aléatoires liées au passage de la population entière à l'échantillon
 - ▶ et donc à la **racine carrée** de la variance
- ▶ Pour diviser cette longueur par deux, il faut donc multiplier par 4 la taille d'échantillon
 - ▶ Pour diviser par 10 cette longueur il faut un échantillon 100 fois plus gros
- ▶ Si on double la taille d'échantillon on réduit de 30% la longueur des intervalles de confiance

Homoscédasticité et hétéroscléasticité

Deux hypothèses pour estimer l'incertitude sur les coefficients

- ▶ Le deuxième terme qui apparaît dans la variance de l'estimateur a une expression générale qui dépend de très peu d'hypothèses
- ▶ La quasi-totalité des logiciels statistiques fait par défaut une hypothèse supplémentaire : **homoscédasticité**
 - ▶ La variance du terme résiduel est la même pour tous les niveaux possibles des variables indépendantes
- ▶ Cette hypothèse a pour intérêt de **simplifier le calcul**
- ▶ On sait quand même faire le calcul général lorsque cette hypothèse ne vaut pas : situation d'**hétéroscéasticité**

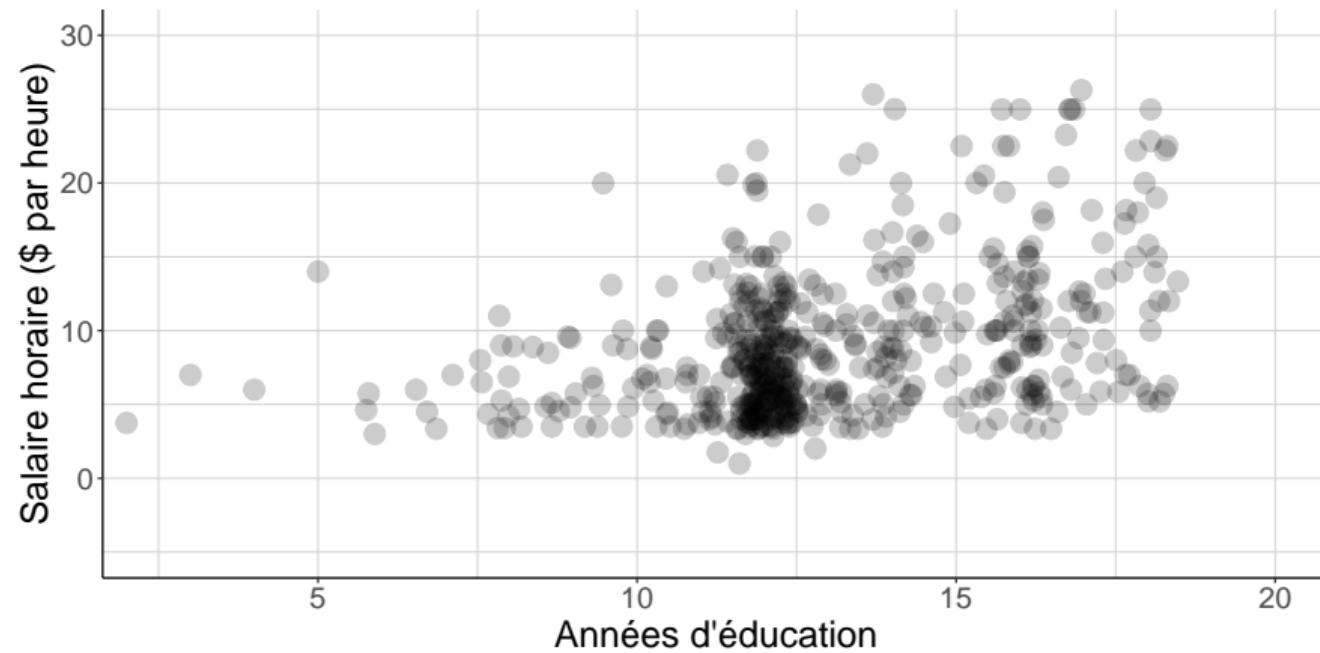
L'hypothèse d'homoscédasticité est-elle raisonnable ?

- ▶ Suppose que la moyenne du carré du terme résiduel est la même pour tous les niveaux possibles des variables indépendantes
- ▶ Impose deux choses :
 - ▶ La **dispersion de la variable dépendante** (sa variance) est la même pour tous les niveaux des variables indépendantes
 - ▶ L'**espérance conditionnelle de la variable dépendante** est une **fonction affine des variables indépendantes**
 - ▶ La linéarité n'est plus *juste* une approximation utile
- ▶ On a rarement envie d'y croire
 - ▶ Si la variable dépendante est dichotomique elle revient à dire que $\mathbb{P}(Y = 1 | X = x)$ ne dépend pas de x !

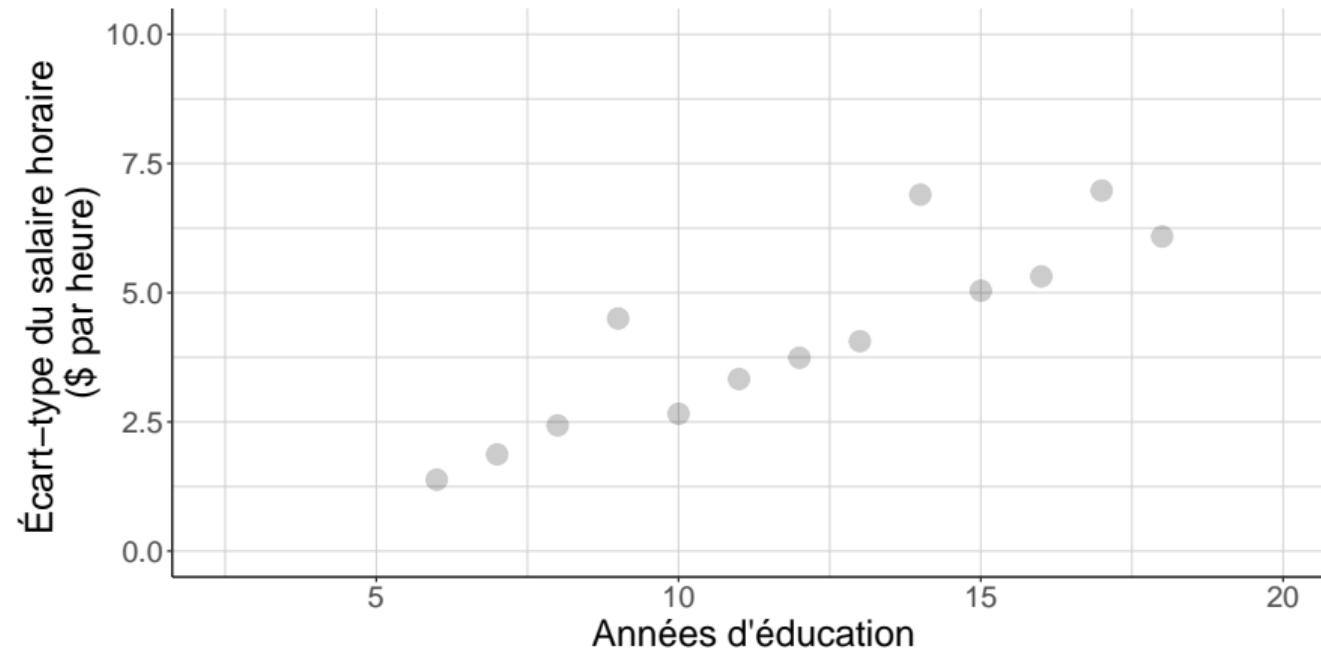
Peut-on relâcher l'hypothèse d'homoscédasticité ?

- ▶ C'est assez facile en pratique : presque tous les logiciels statistiques le permettent
- ▶ On parle souvent de matrice de variance-covariance **robuste** ou d'**estimateur sandwich** de la matrice de variance-covariance
- ▶ En pratique cela peut faire une différence → l'hypothèse d'homoscédasticité conduit à **sous-estimer** l'incertitude

Relâcher l'hypothèse d'homoscédasticité sur un exemple empirique



Relâcher l'hypothèse d'homoscédasticité sur un exemple empirique



L'hypothèse d'homoscédasticité conduit à sous-estimer l'écart-type sur le coefficient de la variable d'éducation de 7%

```
#Les écarts-types et les intervalles de confiance sous l'hypothèse
# d'homoscédasticité

coeftest(reg_wage_educ)

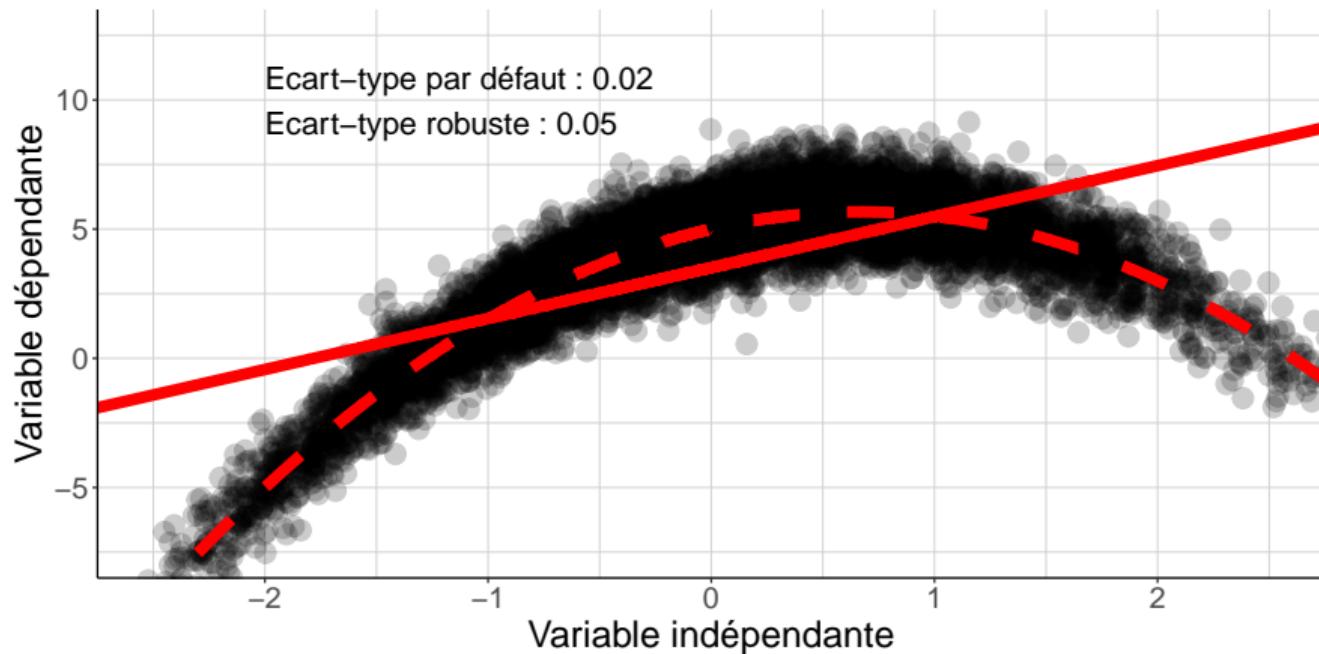
## 
## t test of coefficients:
## 
##           Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.745980  1.045454 -0.7135  0.4758  
## education    0.750461  0.078734  9.5316  <2e-16 *** 
## ---                                                 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'hypothèse d'homoscédasticité conduit à sous-estimer l'écart-type sur le coefficient de la variable d'éducation de 7%

```
#Les écarts-types et les intervalles de confiance sans l'hypothèse
# d'homoscédasticité et donc en présence d'hétéroscédasticité)
coeftest(reg_wage_educ,
         vcov=vcovHC(reg_wage_educ))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.745980   1.048221 -0.7117   0.477
## education    0.750461   0.084905  8.8389 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Un autre exemple avec des données simulées

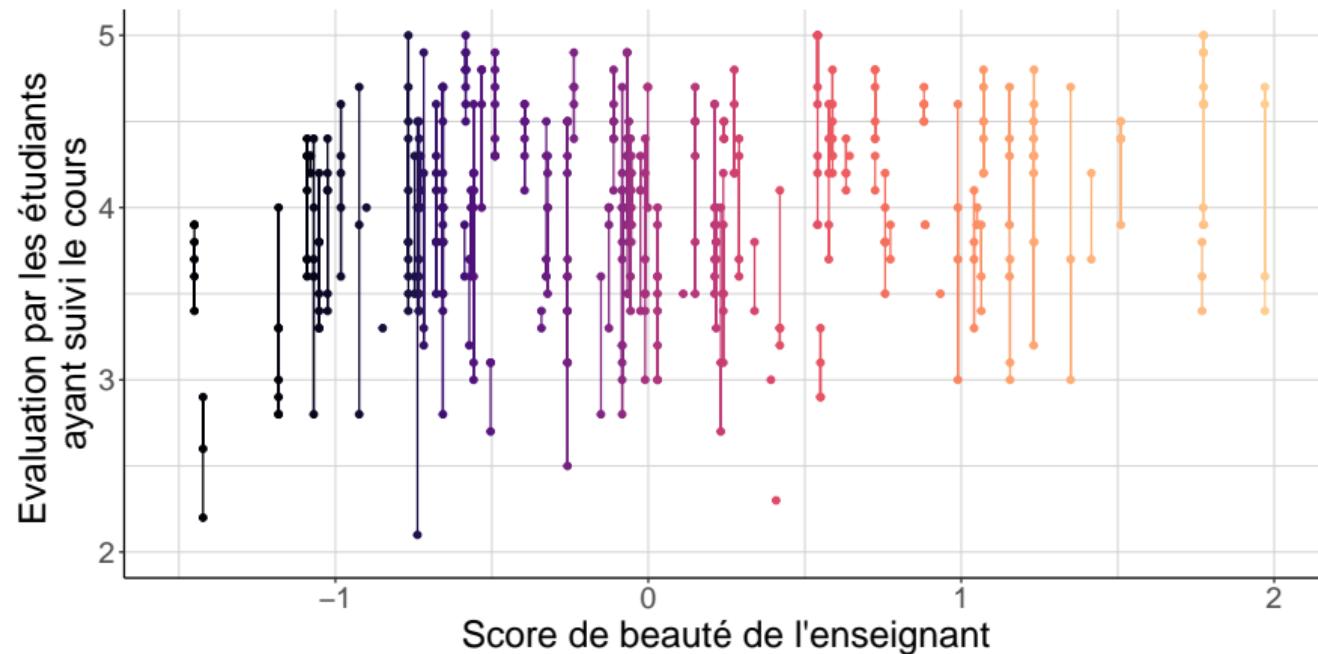


Données en *clusters*

Relâcher l'hypothèse d'indépendance

- ▶ Tout ce qui précède repose sur une hypothèse fondamentale : **chaque observation est supposée indépendante des autres**
- ▶ Cette hypothèse est contestable notamment si :
 - ▶ les variables dépendantes sont **assignées par groupe d'individus** → s'il y a des erreurs sur une, alors il y a la même erreur sur les autres individus du même groupe
 - ▶ l'échantillon vient d'un **sondage par grappe**
- ▶ Dans la littérature, ces groupes et grappes sont souvent appelés **clusters**

Un exemple empirique : Hamermesh and Parker (2005)



Le problème des sondages en grappe

- ▶ Sous l'hypothèse d'indépendance, **l'incertitude décroît en $\frac{1}{\sqrt{N}}$**
- ▶ Si l'on voulait apprendre quelque chose sur les écarts de réussite scolaire entre filles et garçons, commet-on la même erreur d'échantillonage en :
 1. Interrogeant 1000 élèves chacun dans un collège différent
 2. Tirant deux établissements de 500 élèves et en interrogeant tous les élèves de ces établissements ?
- ▶ **Sous l'hypothèse d'indépendance, oui !** Même nombre d'individus \implies même incertitude

Prendre en compte les *clusters* dans le calcul de l'incertitude

- ▶ On retrouve des résultats similaires à ceux obtenus sous l'hypothèse d'indépendance en supposant que **les *clusters* sont indépendants**
- ▶ L'incertitude décroît en $\frac{1}{\sqrt{N_c}}$ où N_c est le **nombre de *clusters***
- ▶ Ne pas prendre en compte les *clusters* conduit à **sous-estimer** l'incertitude

Retour sur Hamermesh and Parker (2005)

```
#On réplique Hamermesh and Parker, 2005, Table 3
reg_eval<-lm(eval ~ beauty + gender + minority + native +
              tenure + division + credits,
              weights = students,
              data = TeachingRatings)

#On considère d'abord l'inférence robuste à l'hétéroscédasticité
head(coeftest(reg_eval,
               vcov=vcovHC(reg_eval)))
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	4.2231419	0.06618241	63.810643	3.987527e-229
## beauty	0.2748052	0.03615980	7.599744	1.708200e-13
## genderfemale	-0.2389934	0.05876721	-4.066782	5.617470e-05
## minorityyes	-0.2489367	0.09461651	-2.631007	8.802170e-03
## nativeño	-0.2527135	0.10340875	-2.443831	1.491110e-02
## tenureyes	-0.1359225	0.06207263	-2.189734	2.905061e-02

Prendre en compte le bon niveau d'assignation du score de beauté augmente l'incertitude

```
#On considère l'inférence une fois prise en compte cette structure
# de dépendance dans les données
head(coeftest(reg_eval,
               vcov = vcovCL,
               cluster =
                 TeachingRatings$prof))
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	4.2231419	0.09745585	43.333898	1.330171e-163
## beauty	0.2748052	0.05872643	4.679413	3.800489e-06
## genderfemale	-0.2389934	0.08512409	-2.807589	5.205914e-03
## minorityyes	-0.2489367	0.11168058	-2.229006	2.630112e-02
## nativeño	-0.2527135	0.13399357	-1.886012	5.993003e-02
## tenureyes	-0.1359225	0.09422615	-1.442514	1.498454e-01

Pour aller plus loin

- ▶ Les questions liées à l'inférence sont **souvent difficiles** → littérature récente et assez technique qui approfondit ou discute ces questions apparemment basiques
- ▶ Ici tout est présenté comme si le problème était de passer du petit échantillon à la population
 - ▶ Des auteurs importants proposent d'autres façons de caractériser et estimer l'incertitude **notamment quand on travaille sur des données exhaustives** (Abadie et al. (2020))
 - ▶ Approche que l'on adopte souvent quand on fait des RCT
- ▶ Pas toujours évident de savoir **à quel niveau considérer les clusters**
 - ▶ Questions de justification (Abadie et al. (2022))
 - ▶ En considérer plusieurs à la fois (Davezies, D'Haultfœuille, and Guyonvarch (2021))

Bibliographie

Bibliographie I

- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1) : 265–96.
<https://doi.org/https://doi.org/10.3982/ECTA12675>.
- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2022. "When Should You Adjust Standard Errors for Clustering?*" *The Quarterly Journal of Economics*, October. <https://doi.org/10.1093/qje/qjac038>.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. Princeton University Press.
- Cunningham, Scott. 2021. *Causal Inference*. Yale University Press.
- Davezies, Laurent, Xavier D'Haultfœuille, and Yannick Guyonvarch. 2021. "Empirical process results for exchangeable arrays." *The Annals of Statistics* 49 (2) : 845–62.
<https://doi.org/10.1214/20-AOS1981>.
- Glymour, Madelyn, Judea Pearl, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics : A Primer*. John Wiley & Sons.

Bibliographie II

- Hamermesh, Daniel, and Amy Parker. 2005. "Beauty in the Classroom : Instructors' Pulchritude and Putative Pedagogical Productivity." *Economics of Education Review* 24 (4) : 369–76.
<https://EconPapers.repec.org/RePEc:eee:ecoedu:v:24:y:2005:i:4:p:369-376>.
- Huntington-Klein, Nick. 2021. *The Effect. An Introduction to Research Design and Causality*. Chapman ; Hall/CRC.
- Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference : Methods and Principles for Social Research*. 2nd ed. Analytical Methods for Social Research. Cambridge University Press.
- Ollion, Étienne. 2011. "De La Sociologie En Amérique. Éléments Pour Une Sociologie de La Sociologie étasunienne Contemporaine." *Sociologie* 2 (3) : 277–94.
- Pearl, Judea. 2009. *Causality*. Cambridge University Press.