# Séance 6 – Estimation et inférence pour les moindres carrés ordinaires

Pierre Pora

# Introduction

# Rappel des séances précédentes

- **Régression linéaire par les MCO** : un seul objet, plein de points de vue
  - Approximation de la variable dépendante
  - Projection orthogonale
  - Résumé de l'information contenue dans une matrice de variance-covariance
  - Comparaisons de moyennes
  - Comparaisons deux-à-deux

# Rappel des séances précédentes

- Jusque là on n'a considéré que l'interprétation des quantités estimées, relatives à la **population** toute entière
  - Ou plutôt on a fait comme si c'était ce que l'on récupérait en manipulant des données sous R

- Ce n'est évidemment pas le cas !
  - On regarde des **échantillons finis**

  - Les quantités estimées ne sont pas directement connues

# L'objet de la séance

- Qu'est-ce que cela change que l'on regarde non pas la population entière mais un échantillon ?
  - Peut-on acquérir une information utile sur ces quantités non-observées à partir d'un échantillon ?
  - Comment apprécier la qualité de cette information ?
- On va se concentrer sur la façon dont on **estime les MCO à partir d'un échantillon**, et le comportement de l'estimateur lorsque l'échantillon devient grand
- L'utilisation de cet estimateur pour **tester des hypothèses** fera l'objet de la séance suivante

# L'objet de la séance

- Tout tient au fait que finalement on ne fait rien d'autre que de **comparer entre elles des moyennes**... On a déjà tous les outils nécessaires pour cela :
  - Peut-on acquérir une information utile sur ces quantités non-observées à partir d'un échantillon ?
    - Oui $\rightarrow$ **loi des grands nombres**
  - Comment apprécier la qualité de cette information ?
    - En revenant au **TCL et à ses conséquences**

# Estimateur des moindres carrés ordinaires

# Quantités estimées et estimateur

- On considère une régression du type $Y = X'\beta + \epsilon$ avec la condition d'orthogonalité $\mathbb{E}[X\epsilon] = 0$
- On sait exprimer le vecteur que l'on veut estimer en termes d'**espérances certaines variables aléatoires** :
  $\beta = \mathbb{E}[X'X]^{-1}\mathbb{E}[XY]$
  - Ce sont des moyennes prises sur toute la population
  - On n'a accès qu'à un échantillon de taille beaucoup plus petite que cette population !

# Quantités estimées et estimateur

- On va construire un **estimateur** c'est-à-dire une procédure qui permet de passer des données portant sur le petit échantillon à un vecteur $\widehat{\beta}$ dont on espère qu'il donne une idée raisonnable de la valeur de $\beta$
  - En réalité $\widehat{\beta}$ n'est pas vraiment un vecteur : **c'est une v.a. qui prend ses valeurs dans l'espace dans lequel vit** $\beta$
  - A la fin on n'aura accès qu'à une réalisation de cette v.a. qui sera, elle, un vecteur qui vit dans le même espace que $\beta$
  - Ce sera notre **estimation**

# Un peu de formalisme

- Jusque là on avait considéré une seule **expérience aléatoire**, et les v.a. associées, qui représentait schématiquement le tirage d'un seul individu dans la population d'intérêt
  - Ces v.a. représentent la façon dont les grandeurs individuelles qui nous intéressent se distribuent dans la population toute entière
  - Cela ne permet pas de représenter ce qu'est l'échantillonnage !
  - Il n'y a pas de notion de nombre d'observations !

# Un peu de formalisme

- On va plutôt considérer qu'on fait $n$ **tirages indépendants dans une population infinie**, et que ces tirages nous permettent de définir les v.a. (éventuellement multidmensionnelles) réelles $X_i$ et $Y_i$ pour $i \in [\![1, n]\!]$
  - Les v.a. $(X_i, Y_i)$ sont **indépendantes** et **identiquement distribuées**
    - Elles ont la même loi !
  - Supposer que la population est infinie permet de ne pas se poser la question du tirage avec / sans remise
    - Cela revient au même puisque comme la population est infinie, une fois qu'on a attrapé un individu, que l'on le remette ou pas il y en a toujours autant à l'endroit où on l'a pris

# Un peu de formalisme

- Comme les $(X, Y_i)$ ont toutes la même loi, et que l'espérance d'une v.a. ne dépend que de sa loi, pour une fonction (mesurable) $\phi$ quelconque, pour tout $i$ et $j$ dans $[\![1, n]\!]$ $\mathbb{E}[\phi(X_i, Y_i)] = \mathbb{E}[\phi(X_j, Y_j))$
  - Ce sont juste les moyennes prises dans toute la population
  - On peut omettre les indices pour une notation plus légère

# Estimateur des moindres carrés ordinaaires

- On veut estimer $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$
- On va simplement **remplacer les espérances par les moyennes empiriques correspondantes**
  - $\widehat{\beta} = \left\{\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right\} = \left\{\sum_{i=1}^{n} X_i X_i'\right\}^{-1}\left\{\sum_{i=1}^{n} X_i Y_i\right\}$
  - Point de vocabulaire : remplacer les espérances par les moyennes = estimateur *plug-in*
- **Estimateur des moindres carrés ordinaires**
  - C'est bien une v.a. : dépend du tirage de tout l'échantillon

# Estimateur des moindres carrés ordinaire : notation matricielle

- Les coefficients de la matrice que l'on veut inverser sont chaque fois la somme sur tous les individus $i$ des produits $X_i^k X_i^l$ avec les exposants qui indexent les composantes de $X$
  - On peut les voir comme un produit matriciel : $(X_1^k ... X_n^k)(X_1^l ... X_n^l)'$

- On peut représenter le tirage de tout l'échantillon et les variables indépendantes par une v.a. $\mathbf{X}$ à valeurs dans l'espace des matrices de taille $n \times d$

$$\mathbf{X} := \underbrace{\begin{pmatrix} X_1^1 & X_1^2 & \dots & X_1^{d-1} & X_i^d \\ X_2^1 & X_2^2 & \dots & X_2^{d-1} & X_2^d \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n-1}^1 & X_{n-1}^2 & \dots & X_{n-1}^{d-1} & X_{n-1}^d \\ X_n^1 & X_n^2 & \dots & X_n^{d-1} & X_n^d \end{pmatrix}}_{d \text{ variables}} \quad n \text{ observations}$$

# Estimateur des moindres carrés ordinaires : notation matricielle

- Ce n'est pas une notation très artificielle !
- Si on considère l'estimation de la régression de `wage` sur `education` et `experience` dans le CPS c'est juste

| | `rep(1, nrow(CPS1985))` | `education` | `experience` |
|---|---|---|---|
| 1 | 1 | 8 | 21 |
| 2 | 1 | 9 | 42 |
| 3 | 1 | 12 | 1 |
| 4 | 1 | 12 | 4 |
| 5 | 1 | 12 | 17 |
| 6 | 1 | 13 | 9 |

# Estimateur des moindres carrés ordinaires : notation matricielle

- Dans $\widehat{\beta} = \left\{\sum_{i=1}^{n} X_i X_i'\right\}^{-1} \left\{\sum_{i=1}^{n} X_i Y_i\right\}$ les coefficients de la matrice que l'on veut inverser ne sont rien d'autre que le produit de la transposée de la $k$-ème colonne de $\mathbf{X}$ par la $l$-ème colonne de $\mathbf{X}$
- En définitive, avec cette notation matricielle : $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ avec $\mathbf{Y} = (Y_1 \ldots Y_n)'$

# Estimateur des moindres carrés ordinaires : notation matricielle

- $\mathbf{X}'\mathbf{X}$ prend ses valeurs dans l'espace des matrices carrées de dimension $d$
  - Produit d'une matrice de dimension $n \times d$ par une matrice de dimension $d \times n$
- $\mathbf{X}'\mathbf{Y}$ vit dans l'espace des matrices colonnes de taille $d$
  - Produit d'une matrice de taille $n \times d$ par une colonne de taille $n$
- Donc le produit de l'inverse du premier par le second vit bien dans l'espace des matrices colonnes de taille $d$

# Une petite vérification

```
#On estime la régression du salaire horaire
# sur l'éducation et l'expérience
regression <-lm(wage ~ education + experience,
                data=CPS1985)

regression$coefficients
```

```
(Intercept)   education  experience
 -4.9044823   0.9259646   0.1051316
```

# Une petite vérification

```
#On construit la matrice X
matrice_covariables <-
  as.matrix(cbind(rep(1,
                      nrow(CPS1985)),
                  CPS1985[,
                          c("education",
                            "experience")]))

colnames(matrice_covariables)[1] <- "constante"
```

# Une petite vérification

```
#On peut vérifier que X ressemble bien à la matrice
# que l'on pense
head(matrice_covariables)
```

```
  constante education experience
1         1         8         21
2         1         9         42
3         1        12          1
4         1        12          4
5         1        12         17
6         1        13          9
```

# Une petite vérification

```
# On calcule X'X : c'est bien une matrice 3*3
XprimeX <-
  t(matrice_covariables) %*%
  matrice_covariables
XprimeX
```

| | constante | education | experience |
|---|---|---|---|
| constante | 534 | 6952 | 9517 |
| education | 6952 | 94152 | 117813 |
| experience | 9517 | 117813 | 251299 |

# Une petite vérification

```
# On calcule X'Y : c'est bien un vecteur de dimension 3
# (ou une matrice de taille 3*1)
XprimeY <-
  t(matrice_covariables) %*%
  as.matrix(CPS1985$wage)
XprimeY
```

```
                [,1]
constante    4818.85
education   65471.33
experience 88834.18
```

# Une petite vérification

```
#On multiple à gauche X'Y par l'inverse de X'X
estimateurMCO <-
  solve(XprimeX) %*% XprimeY

#On vérifie que l'on retombe bien sur les coefficients
# estimés par lm
all.equal(
  as.numeric(estimateurMCO),
  as.numeric(regression$coefficients))
```

```
[1] TRUE
```

# Convergence de l'estimateur des moindres carrés ordinaires

- L'estimateur s'écrit $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- La **loi faible des grands nombres** assure que :
  - $\frac{1}{n}\mathbf{X}'\mathbf{X}$ converge en probabilité vers $\mathbb{E}[XX']$
  - $\frac{1}{n}\mathbf{X}'\mathbf{Y}$ converge en probabilité vers $\mathbb{E}[XY]$
- En définitive $\widehat{\beta}$ **converge en probabilité vers** $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$

# Conclusion partielle

- Pour le dire autrement : pourvu que l'on tire un **échantillon suffisamment grand**, on peut rendre aussi faible que l'on veut la probabilité que le résultat de cette estimation s'éloigne de la quantité que l'on veut estimer davantage que **n'importe quel seuil arbitraire**
- Ce que l'on attrape avec cet échantillon est informatif de ce qui se passe dans la population toute entière
  - Petit devant la taille de la population
  - Pas devant 1 puisque c'est un résultat asymptotique !

# Inférence

# Pour continuer

- On sait que $\widehat{\beta}$ peut être rendu aussi proche de $\beta$ que l'on veut **avec un échantillon assez grand**
- Mais ce résultat ne dit rien de plus !
- On sait que l'on peut finir par être très proche, mais on ne sait pas dire **à quelle vitesse on se rapproche**
- Et donc face au résultat de notre estimation si l'on est raisonnablement près ou bien encore assez loin

# Pour continuer

- L'objectif est maintenant de se donner un moyen de quantifier la vitesse à laquelle $\hat{\beta} - \beta$ se concentre vers 0
  - Le fait que ça finisse par **se concentrer vers 0** avec un échantillon suffisamment grand est garanti par le résultat précédent

- Pas beaucoup de suspense : on a un estimateur qui s'écrit en fonction de moyennes empiriques dont on se demande à quelle vitesse il se rapproche d'une quantité estimée qui s'écrit en fonction des espérances correspondantes...
  - **On va bien sûr utiliser le TCL**

# Comportement asymptotique de l'estimateur

- Avec $\epsilon$ matrice colonne de taille $n$ où l'on empile tous les $\epsilon_i$ :

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) \\ &= \beta + (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}'\epsilon\end{aligned}$$

# Comportement asymptotique de l'estimateur

- Par conséquent :

$$\sqrt{n}\left(\hat{\beta}-\beta\right)=\left(\frac{1}{n}\mathbf{X}\mathbf{X}'\right)^{-1}\sqrt{n}\left(\frac{1}{n}\mathbf{X}'\epsilon\right)$$

- **Théorème de Slustky** :
  - Le premier terme converge en probabilité vers $\mathbb{E}[XX']$
  - Le second terme converge en loi vers $\mathcal{N}(0,\mathcal{V}(X\epsilon))$ (car $\mathbb{E}[X\epsilon]=0$ par définition de $\epsilon$)
  - Donc **le produit converge en loi vers** $\mathcal{N}(0,\mathbb{E}[XX']^{-1}\mathbb{E}[XX'\epsilon^2]\mathbb{E}[XX']^{-1})$

# Comportement asymptotique de l'estimateur

- Dans la limite d'un très grand échantillon, on connaît donc la distribution asymptotique de $\hat{\beta} - \beta$
  - C'est-à-dire a vitesse à laquelle $\hat{\beta}$ vers $\beta$
  - On s'y atendait à cause du TCL et de Bienaymé-Tchebychev : c'est en $\frac{1}{\sqrt{n}}$
- La précision est principalement déterminée par la taille de l'échantillon
- Expression de la matrice de variance-covariance asymptotique : $\frac{1}{n}\mathbb{E}[XX']^{-1}\mathbb{E}[XX'\epsilon^2]\mathbb{E}[XX']^{-1}$
  - Terme central qui contient du $\epsilon^2 \rightarrow$ pour atteindre une précision donnée, il faut un plus grand échantillon lorsque les résidus sont très dispersés
    - Donc quand le $R^2$ est très petit

# Estimation de la matrice de variance-covariance asymptotique

- La matrice $\mathbb{E}[XX']^{-1}\mathbb{E}[XX'\epsilon^2]\mathbb{E}[XX']^{-1}$ n'est pas connue *a priori* !
- On peut néanmoins estimer les contreparties empiriques de tous les termes :
  - $(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}$ pour estimer $\mathbb{E}[XX']^{-1}$
  - $\frac{1}{n}\mathbf{X}\hat{\Sigma}\mathbf{X}'$ pour estimer $\mathbb{E}[XX'\epsilon^2]$
    - $\hat{\Sigma} = \mathrm{diag}(\hat{\epsilon_1}^2, ..., \hat{\epsilon_n}^2)$ est la matrice carrée diagonale dans laquelle on met le carré tous les résidus estimés $(\hat{\epsilon_i} = Y_i - X_i'\hat{\beta} \neq Y_i - X_i'\beta = \epsilon_i)$

# Estimation de la matrice de variance-covariance asymptotique

- Toutes ces contreparties convergent en probabilité vers les espérances que l'on ne connaît pas directement
- Donc le produit converge vers la matrice de variance-covariance asymptotique que l'on voulait
- Petit point de vocabulaire : $(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}\hat{\Sigma}\mathbf{X}'(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}$
  - estimateur *sandwich*
  - estimateur de White-Huber
  - matrice de variance-covariance robuste (à l'hétéroscédasticité)

# Une petite vérification

```
#On construit la matrice diagonale Sigma
Sigma <-
  diag(x=regression$residuals^2,
       nrow = nrow(CPS1985),
       ncol = nrow(CPS1985))
```

# Une petite vérification

```
#On estime le terme central du sandwich : X'SigmaX
XprimeSigmaX <-
  1/nrow(CPS1985)*
  t(matrice_covariables)%*%Sigma%*%matrice_covariables

#On estime la matrice de variance-covariance
estimated_vcovHC <-
  1/(nrow(CPS1985))*
  solve(1/nrow(CPS1985)*XprimeX)%*%
  XprimeSigmaX%*%
  solve(1/nrow(CPS1985)*XprimeX)
estimated_vcovHC
```

```
             constante     education    experience
constante   1.56901241 -0.1053153446 -0.0137215057
education  -0.10531534  0.0077070415  0.0006313856
experience -0.01372151  0.0006313856  0.0003223285
```

# Une petite vérification

```
#On vérifie que la matrice de variance-covariance
# robuste à l'hétéroscédasticité estimée par R est
# égale à celle que l'on vient d'estimer
all.equal(as.numeric(estimated_vcovHC),
          as.numeric(vcovHC(regression,
                            type="HC0")))
```

```
[1] TRUE
```

# Hétéroscédasticité et homoscédasticité

# Hétéroscédasticité vs. homoscédasticité

- Pour faire cette estimation de la matrice de variance-covariance, on n'a fait aucune hypothèse supplémentaire !
  - (OK l'existence des espérances que l'on cherche à estimer)
- C'est donc la façon la plus générale de procéder !

# Hétéroscédasticité vs. homoscédasticité

- Sous des hypothèses plus restrictives, on pourrait se simplifier la vie
- Hypothèse d'**homoscédasticité** : $\mathbb{E}[\epsilon^2 \mid X] = \sigma^2$
- Sous cette hypothèse la matrice de variance-covariance asymptotique se simplifie en $\frac{1}{n}\sigma^2\mathbb{E}[XX']^{-1}$
  - On peut l'estimer par $\frac{1}{n}\left(\sum_{i=1}^{n}\hat{\epsilon_i}^2\right)(\mathbf{X}'\mathbf{X})^{-1}$

# Hétéroscédasticité vs. homoscédasticité

- Hypothèse historiquement importante
  - Moindres carrés ordinaires apparus pour régler des problèmes de mesure imprécise en astronomie
  - Si on a une interprétation réaliste de $\epsilon$ comme l'erreur de mesure due à l'imprécision de l'appareil que l'on utilise pour toutes les mesures, supposer que celle-ci est fondamentalement la même pour toute les mesures
- C'est en général l'**hypothèse faite par défaut par les logiciels de calcul statistique**

# Une petite vérification

```
#On estime sigma2 comme la moyenne empirique du
# carré des résidus
estimated_sigma2<-mean(regression$residuals^2)
```

# Une petite vérification

```r
#L'estimateur de la matrice de variance-covariance
# n(X'X)-1 multiplié par l'estimateur de sigma2 multiplié
# par 1/n
estimated_vcov <- 1/(nrow(CPS1985) - nrow(XprimeX)) *
  #On divise par n-d et pas par n pour
  #avoir de meilleures propriétés à distance finie
  solve((1/nrow(CPS1985)) * XprimeX) *
  estimated_sigma2

estimated_vcov
```

```
               constante     education    experience
constante    1.48577566 -0.0950681766 -0.0116986534
education   -0.09506818  0.0066265277  0.0004937255
experience -0.01169865  0.0004937255  0.0002957551
```

# Une petite vérification

```
#On vérifie l'égalité entre cet estimateur et
# la matrice de variance-covariance
# estimée par lm
all.equal(as.numeric(estimated_vcov),
          as.numeric(vcov(regression)))
```

```
[1] TRUE
```

# Aparté : estimateur non-biaisé de la variance

- **Estimateur naïf de la variance** $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- L'espérance de cet estimateur n'est pas la variance de $X$ !
  - Petit calcul : son espérance est $\frac{n-1}{n}\mathcal{V}(X)$
- Pour un **estimateur sans biais** : $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

# Aparté : estimateur non-biaisé de la variance

- Cela vient de ce que pour estimer la variance de $X$ il faut estimer auparavant l'espérance de $X$
  - Sans ce **degré de liberté**, l'estimateur naïf serait sans biais
- Le même genre de considération intervient ici
  - Pour faire le calcul précédent il faut avoir déjà estimé $\beta$, c'est-à-dire $d$ paramètres, et l'estimateur $\hat{\beta}$ intervient implicitement dans l'expression
  - On annule le biais en divisant par $n - d$ plutôt que par $n$
- Asymptotiquement, cela revient au même parce que $d$ est fixe ($n - d \overset{n \to \infty}{\sim} n$)

# Quelques conséquences additionnelles de l'hypothèse d'homoscédasticité

- L'hypothèse d'homoscédasticité est en fait souvent convoquée quand on veut discuter de l'**efficacité des estimateurs**, et en particulier de l'estimateur des MCO
- C'est dans ce contexte qu'il faut comprendre l'intérêt de cette hypothèse
- Mais cela rend aussi les discussions de ces questions un peu éloignées de la pratique empirique...
- Je tente un aperçu des **résultats classiques**

# Estimateurs efficaces

- Qu'est-ce qu'un estimateur efficace ? Une comparaison pour se donner une idée
  - On veut estimer la régression de `wage` sur `education` à partir du CPS, et on se propose deux estimateurs possibles
    - Le premier qui est l'estimateur des MCO
    - Le second qui est l'estimateur des MCO, appliqué à un sous-échantillon du CPS : on garde aléatoirement un enquêté sur deux
- Les deux estimateurs convergent en probabilité vers le coefficient de la régression dans la population toute entière
- On voit pour autant que les deux n'utilisent pas l'information disponible de façon aussi efficace l'un que l'autre...

# Estimateurs efficaces

- Intuitivement un estimateur est plus efficace qu'un autre si pour un même processus générateur des données, le **risque quadratique** ou la variance de l'un est plus petit que l'autre
  - Remarque : pour les estimateurs sans biais variance = risque quadratique
- Dans l'exemple précédent, on a artificiellement divisé par deux la taille d'échantillon utilisée par le second estimateur
- Sa variance est donc deux fois plus grande
- Il est donc moins efficace que l'autre

# Théorème de Gauss-Markov

- Sous des **hypothèses fortes** :
  - La régression linéaire s'identifie à l'espérance conditionnelle : $\mathbb{E}[\epsilon \mid X] = 0$
  - Homoscédasticité : $\mathcal{V}(\epsilon \mid X) = \sigma^2$
- L'estimateur des MCO est l'estimateur linéaire non-biaisé de variance minimale
  - Estimateur linéaire : $\hat{\beta}$ s'écrit comme $\sum_{i=1}^{n} w_i Y_i$ où les poids $w_i$ dépendent des régresseurs
  - Non-biaisé : $\mathbb{E}[\hat{\beta}] = \beta$
  - De variance minimale : si on prend un autre estimateur linéaire non-biaisé de $\beta$, il sera toujours moins précis

# Théorème de Gauss-Markov

- Remarque : ce n'est **pas un résultat asymptotique**
- Les hypothèses sont beaucoup plus fortes que celles dont on a besoin pour
  - Donner un sens aux coefficients
  - Connaître le comportement asymptotique de l'estimateur

# Un autre résultat sur l'efficacité des MCO

- Sous d'**autres hypothèses fortes** :
  - $\epsilon \perp\!\!\!\perp X$
  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- L'estimateur des MCO correspond à l'**estimateur du maximum de vraisemblance**
  - Preuve facile : la minimisation de la vraisemblance implique la version empirique de la condition d'orthogonalité
- En tant qu'estimateur du maximum de vraisemblance il embarque une **propriété d'efficacité asymptotique**
  - Il atteint la borne de Cramér-Rao

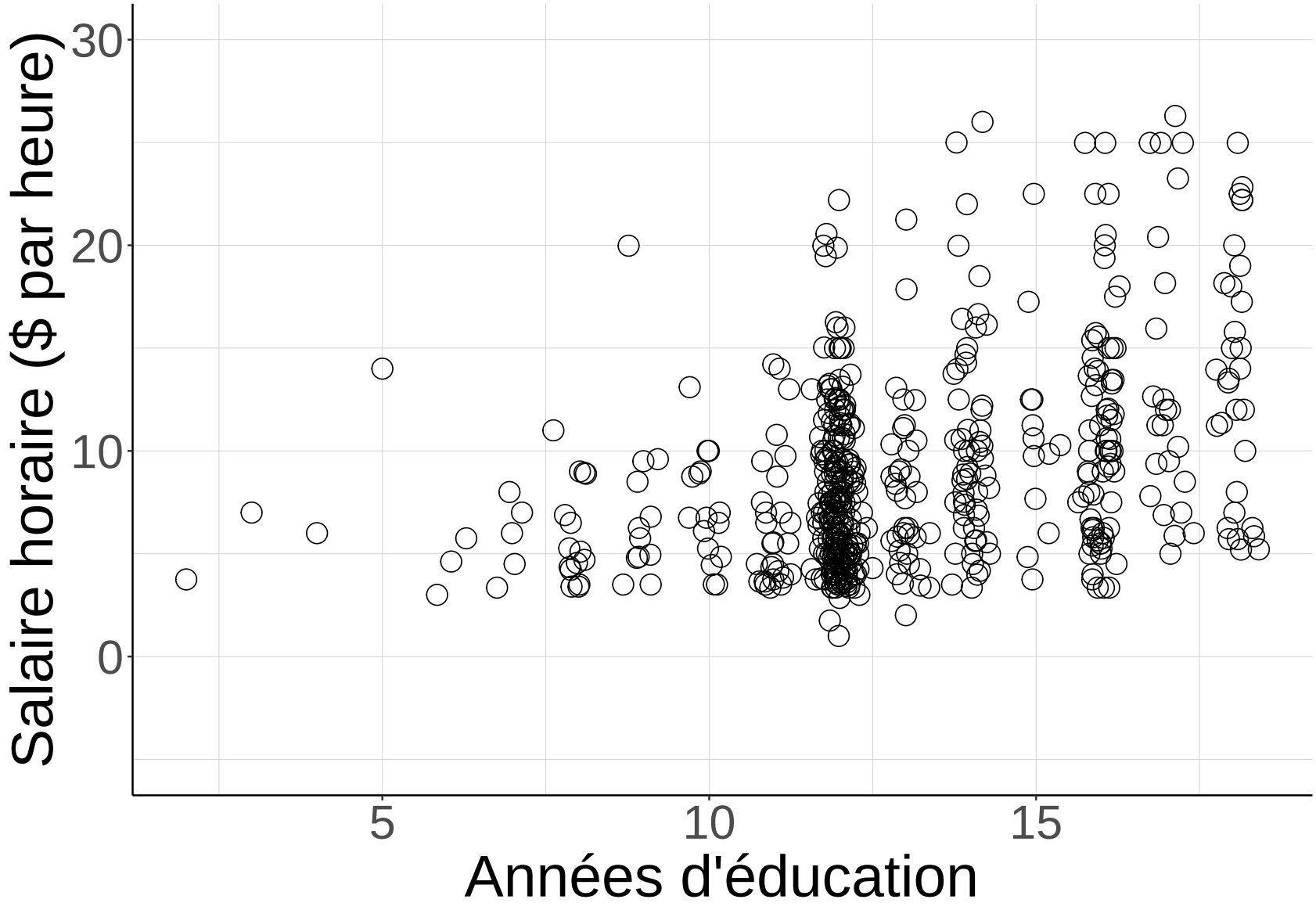
# Une autre résultat sur l'efficacité des MCO

- Ce n'est pas le même résultat qu'avant !
  - Résultat asymptotique
  - N'implique pas l'absence de biais
- Ici encore les hypothèses sont considérablement plus fortes que celles dont on avait besoin jusque là...

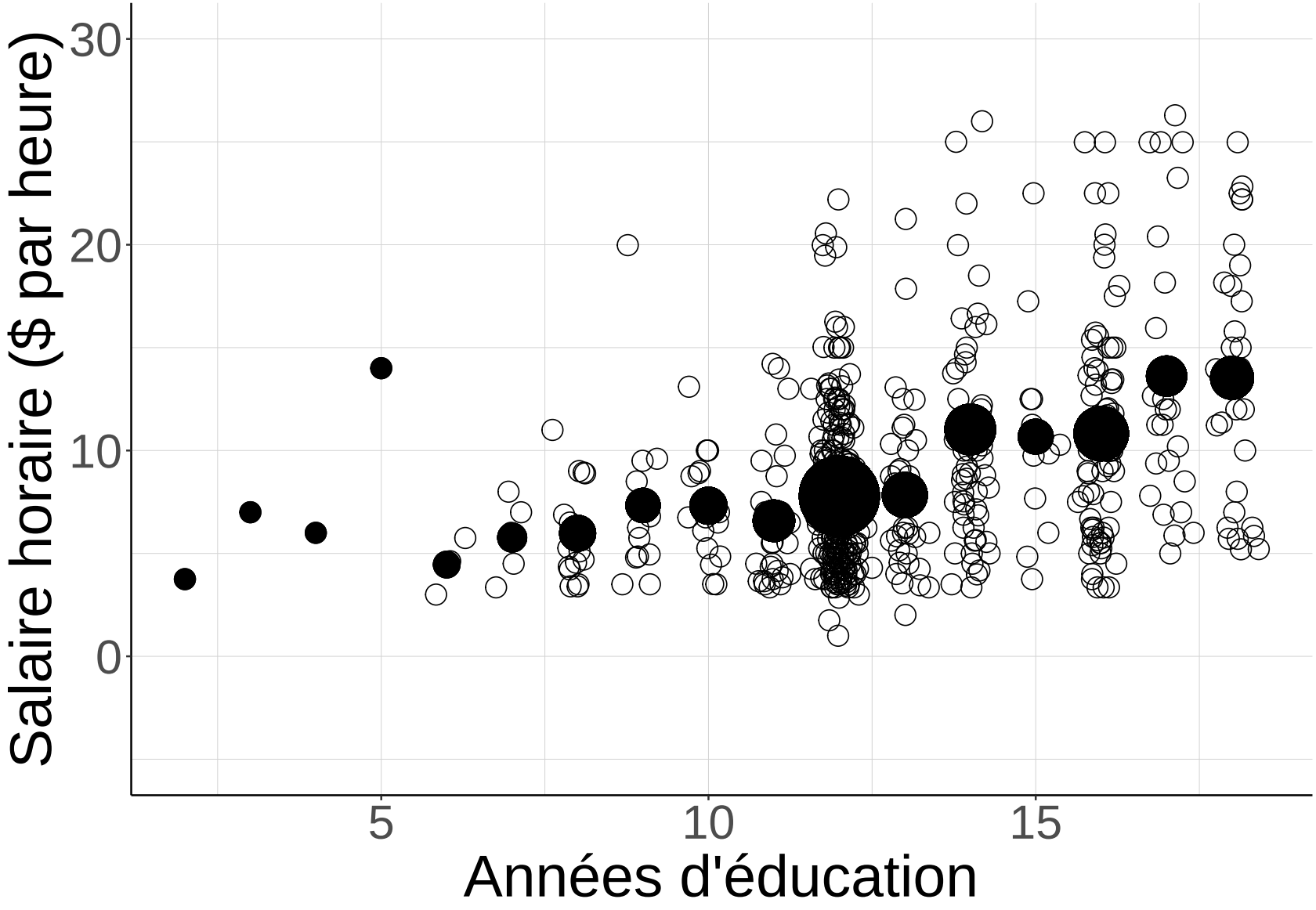
# L'homoscédasticité est-elle possible ?

- Hypothèse très forte qui ne dit pas seulement que la dispersion de la variable dépendante est (à peu près) constante
- $\mathbb{E}[\epsilon^2 \mid X] = \mathbb{E}[(Y - X'\beta)^2 \mid X]$, mais aussi $\mathbb{E}[\epsilon^2 \mid X] = \mathbb{E}[\epsilon \mid X]^2 + \mathcal{V}(\epsilon \mid X)$
  - Comme la valeur prédite est la même pour les individus qui ont la même valeur de $X$ cela dit en partie quelque chose sur la **dispersion de la variable dépendante** $(\mathcal{V}(Y \mid X) = \mathcal{V}(\epsilon \mid X))$
  - Mais aussi quelque chose sur **la distance entre $\hat{Y}$ et $Y$ à ce niveau-là**
    - Le cas le plus favorable est que la valeur prédite s'identifie à l'espérance conditionnelle : alors $\mathbb{E}[\epsilon \mid X] = 0$
    - C'était déjà une hypothèse nécessaire pour les deux résultats d'optimalité précédents

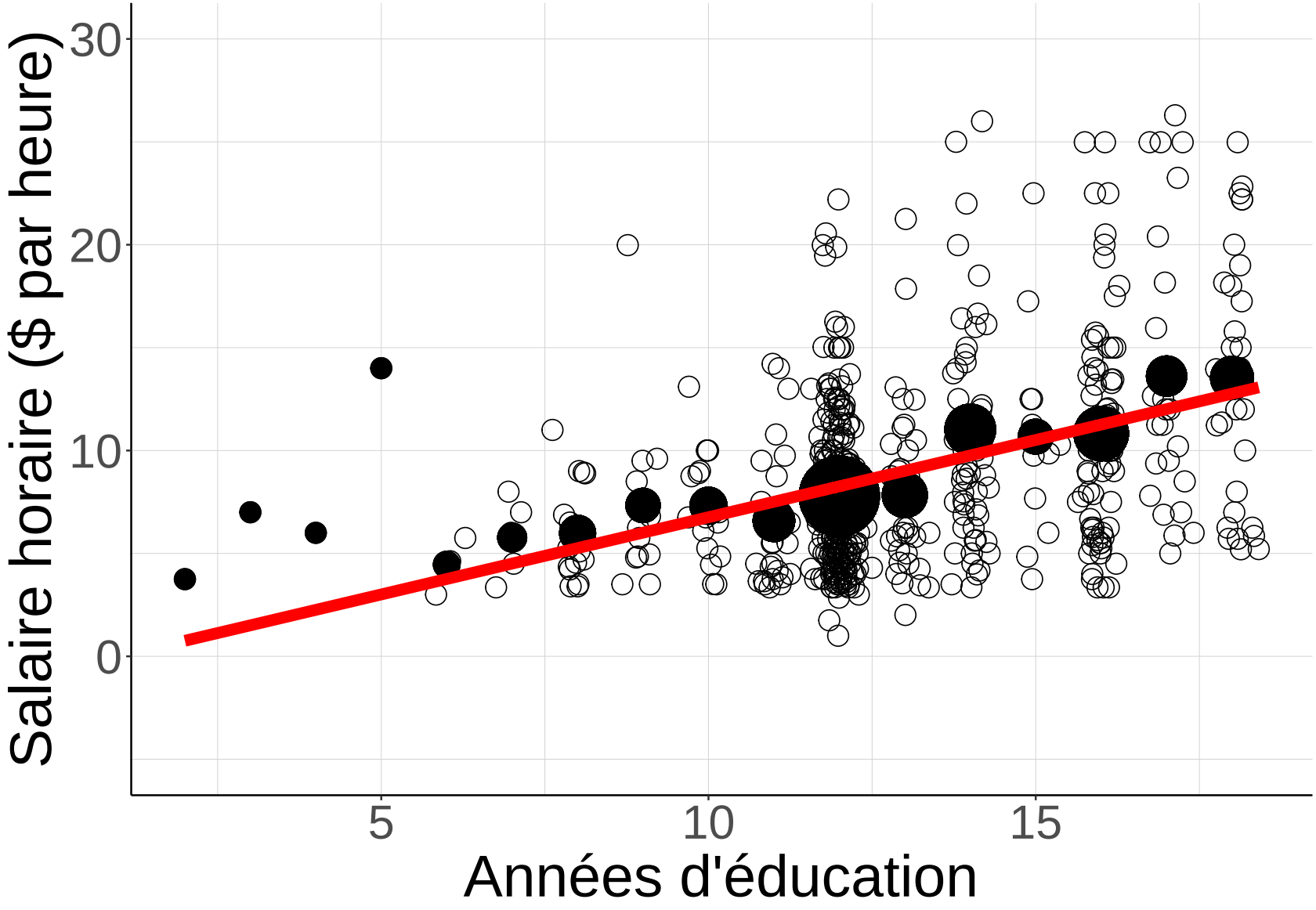
# Un exemple empirique



# Un exemple empirique



# Un exemple empirique



# L'homoscédasticité est-elle possible ?

- On a bien des cas pour lesquels mécaniquement l'espérance conditionnelle et les valeurs prédites par la régression linéaire coïncident
  - La **régression saturée**

- Mais il y a des cas courants pour lesquels l'hypothèse de variance constante est non seulement peu plausible, mais impossible !
  - **Variable dépendante binaire** :
    $\mathcal{V}(Y \mid X) = \mathbb{E}[Y \mid X]\{1 - \mathbb{E}[Y \mid X]\}$
  - L'hypothèse d'homoscédasticité revient ici à dire que $\mathbb{E}[Y \mid X]$ est essentiellement constante...

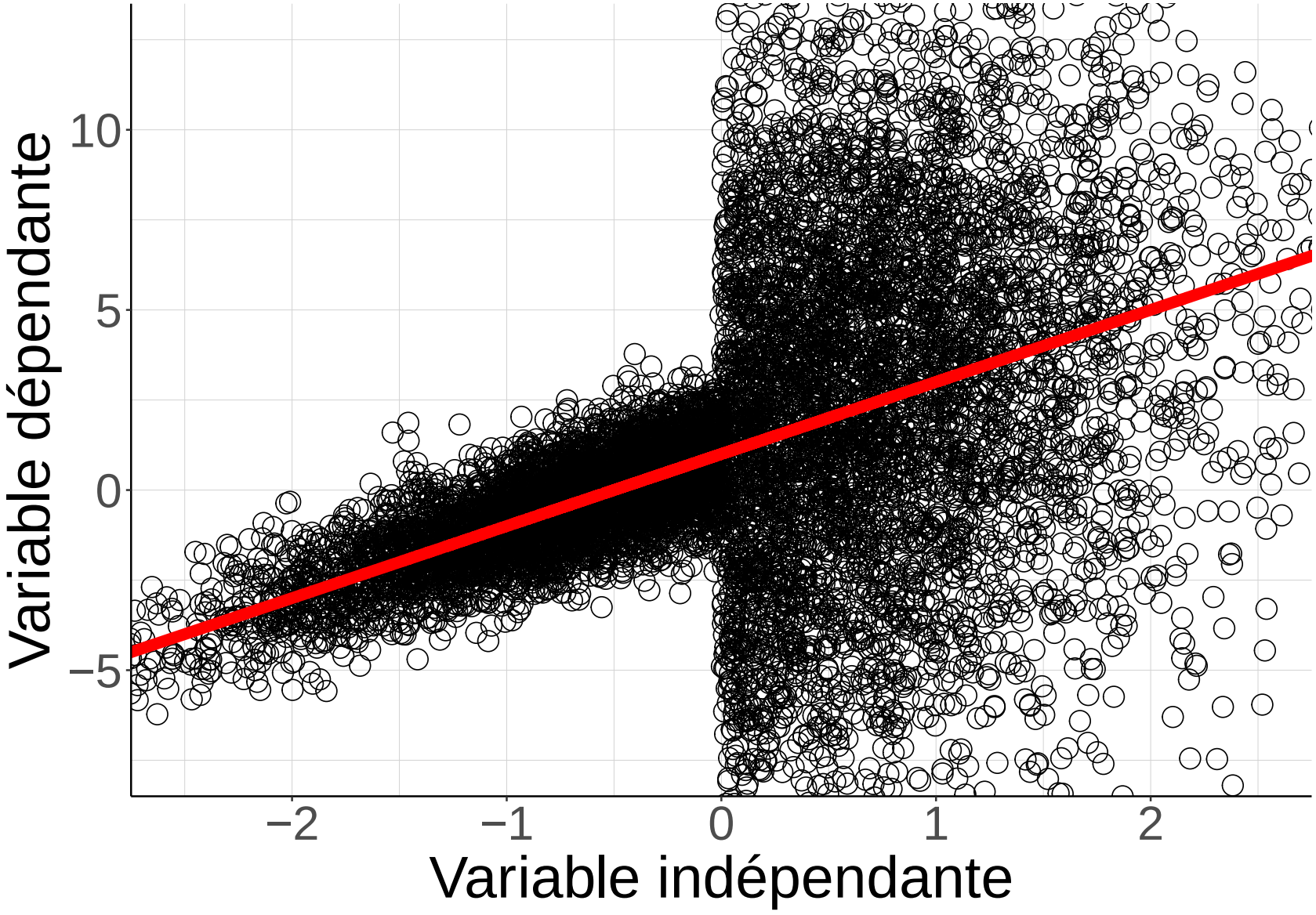
# L'homoscédasticité est-elle possible ?

- Au total **il n'est presque jamais crédible de faire l'hypothèse d'homoscédasticité**
- Pour cette raison, on s'abstient presque toujours !
- Il faut considérer la matrice de variance-covariance générale $\frac{1}{n}\mathbb{E}[XX']^{-1}\mathbb{E}[XX'\epsilon^2]\mathbb{E}[XX']^{-1}$ et un de ses estimateurs
- Ca ne remet pas en cause la convergence asymptotique vers $\beta$ et la normalité asymptotique de l'estimateur !

# Que faire dans le cas hétéroscédastique ?

- En cas d'hétéroscédasticité (c'est-à-dire toujours !), on connaît le comportement asymptotique de l'estimateurs des MCO
  - On sait donc en estimer la précision !
- **On n'a plus de garanties d'optimalité**, qu'elles viennent de Gauss-Markov ou du maximum de vraisemblance
  - On peut imaginer de meilleurs estimateurs

# Pour comprendre : le cas homogène et hétéroscédastique



# Pour comprendre : le cas homogène et hétéroscédastique

- Intuitivement, si on suppose que la droite de régression et l'espérance conditionnelle sont le même objet, alors on parvient beaucoup mieux à l'estimer avec les observations de gauche qu'avec celles de droite
- L'estimateur des MCO donne le même poids aux deux parties du nuage
- Un estimateur plus efficace donnerait plus de poids à gauche, moins de poids à droite
- C'est (en gros) ce que fait l'estimateur des moindres carrés généralisés
  - On donne un poids inversement proportionnel à la variance conditionnelle des résidus

# Une petite vérification

```
regression_mco <-
  lm(y ~x, data = simuldat)

regression_mcg <-
  nlme::gls(y ~x,
            weights = nlme::varConstProp(form = ~ as.numer
            data = simuldat)

sqrt(vcovHC(regression_mco)["x","x"])
```

```
[1] 0.03707384
```

```
sqrt(vcov(regression_mcg)["x","x"])
```

```
[1] 0.02020353
```

# Peut-on vraiment utiliser cette idée ?

- **On ne connaît pas la variance conditionnelle des résidus** avant d'avoir estimé la régression elle-même...
  - La solution est de procéder en deux étapes
    - Utiliser l'estimateur des MCO pour récupérer les résidus
    - Réestimer la régression en pondérant par l'inverse de la moyenne locale du carré des régresseurs
- C'est moins bien que les moindres carrés généralisés !
  - Asymptotiquement on sait quand même que ça va se comporter pareil
  - A distance finie il n'est pas garanti que ce soit aussi bien, et ça peut en fait être moins bien que les MCO...

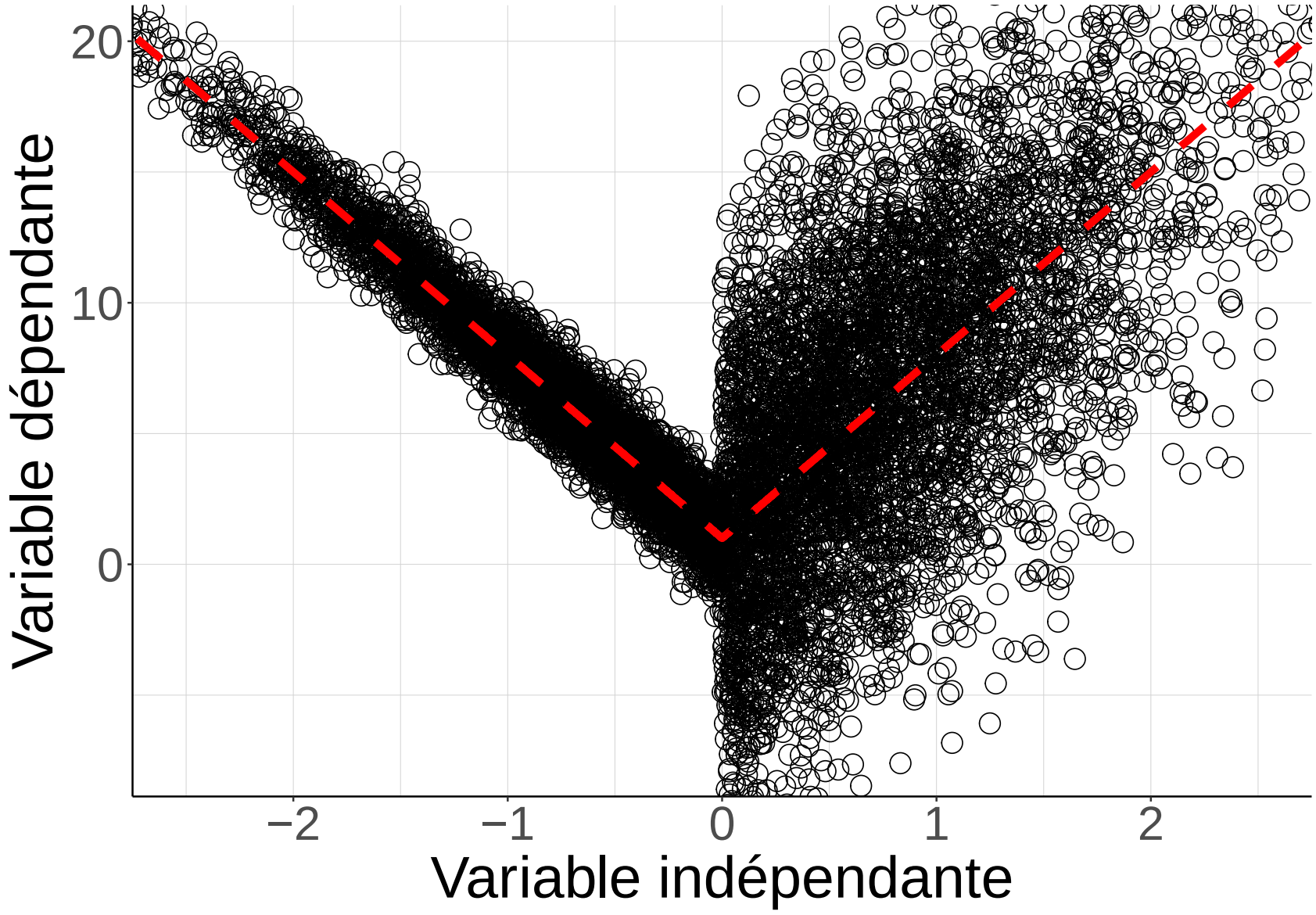
# Peut-on vraiment utiliser cette idée ?

- Plus fondamentalement cette idée ne fonctionne que si ce qui se passe là où les résidus sont moins dispersés et ce qui se passe là où les résidus sont plus dispersés est la même chose
- Si ce n'est pas le cas, on ne converge plus vers les coefficients définis par les MCO dans la population
- Mais vers d'autres coefficients
  - Il faut revoir toute notre interprétation
  - Schématiquement on garde l'idée qu'on fait des comparaisons de moyennes mais tous les poids abordés aux séances précédentes sont modifiés
  - Abstraitement on change le produit scalaire et donc la distance définie sur l'espace des v.a.

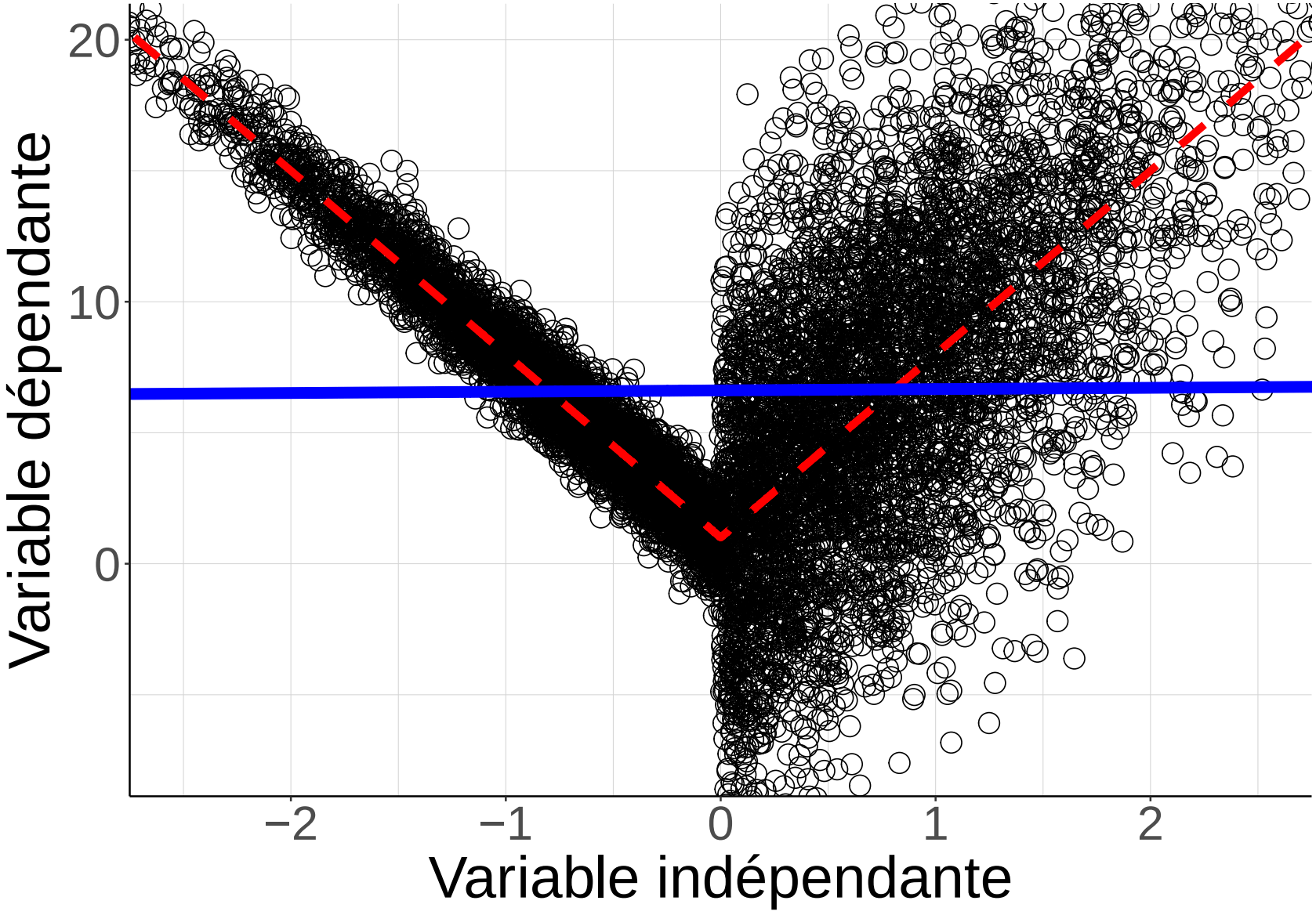
# Peut-on vraiment utiliser cette idée ?

- Une question ouverte (je n'ai pas la réponse) :
  - Dans le cas hétérogène et hétéroscédastique, y a-t-il un estimateur optimal des coefficients définis par les MCO ?

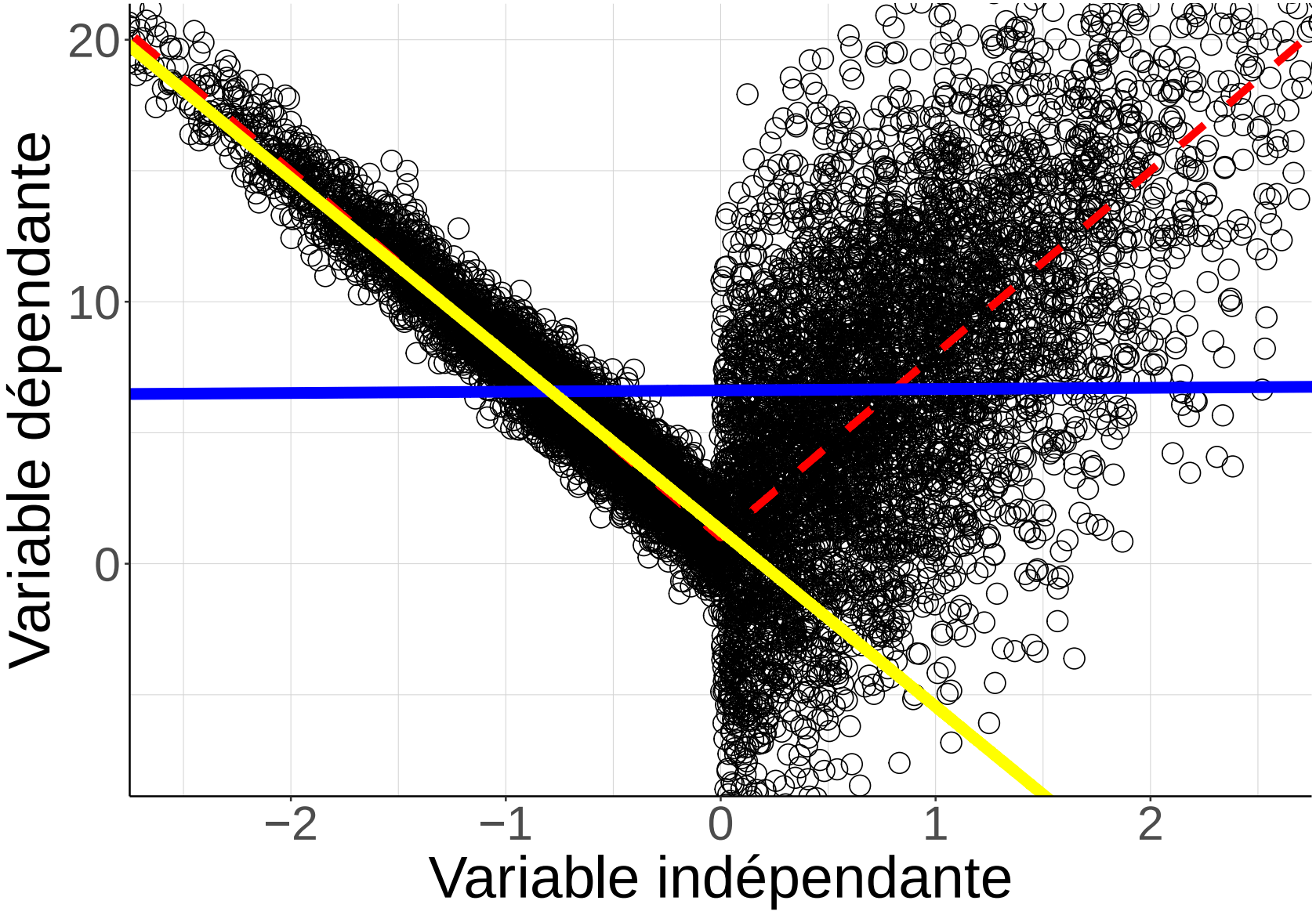
# Pour comprendre : le cas hétérogène et hétéroscédastique



# Pour comprendre : le cas hétérogène et hétéroscédastique



# Pour comprendre : le cas hétérogène et hétéroscédastique



# Ce qu'il faut en retenir

- ** L'hétéroscédasticité est le cas standard, et c'est l'hypothèse à retenir dans tout ce qui a trait à l'inférence**
- Lorsque la régression linéaire ne s'identifie pas à l'espérance conditionnelle, ce qui est le cas le plus fréquent, les estimateurs optimaux dans le cas hétéroscédastique ne convergent pas vers les quantités définies par le problème des moindres carrés ordinaires
  - Cela change l'interprétation d'une façon qui n'est pas très confortable
- Cela justifie approximativement la **pratique empirique la plus courante aujourd'hui en économie** :
  - Utiliser l'estimateur des MCO qui converge vers les quantités que l'on a définies dans les séances précédentes
  - Ne pas faire l'hypothèse d'homoscédasticité lorsque l'on veut discuter de sa précision

# En pratique avec R : homoscédasticité par défaut !

```r
#On estime la régression du salaire horaire
# sur l'éducation et l'âge
regression <-lm(wage ~ education + experience,
                data=CPS1985)

coeftest(regression)
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -4.904482   1.218924 -4.0236 6.564e-05 ***
education    0.925965   0.081403 11.3750 < 2.2e-16 ***
experience   0.105132   0.017198  6.1132 1.893e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

# En pratique ave R : quelques manipulations pour récupérer des écarts-types robustes à l'hétéroscédasticité

```
coeftest(regression,
         vcov. = vcovHC)
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -4.904482   1.267195 -3.8703 0.0001222 ***
education    0.925965   0.088813 10.4260 < 2.2e-16 ***
experience   0.105132   0.018117  5.8030 1.121e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

# En pratique avec R

- Dans le cas présent, utiliser la matrice de variance-covariance robuste conduit à des **écarts-types légèrement plus grands**
  - C'est en général le cas
  - L'ajustement peut-être beaucoup plus dramatique que ça
- Dans le tableau précédent, les écarts-types ne sont rien d'autre que **les racines carrées des éléments diagonaux de la matrice de variance-covariance**
- On discutera des autres colonnes la séance suivante

# Une petite vérification

```
test_regression <-
  coeftest(regression,
           vcov. = vcovHC)

matrice_vcov <-
  vcovHC(regression)

all.equal(
  test_regression[,"Std. Error"],
  sqrt(diag(matrice_vcov))
)
```

```
[1] TRUE
```

# Que faire des données groupées ?

# Est-ce seulement la taille d'échantillon qui compte ?

- Une expérience de pensée : supposons que l'on veuille faire une enquête par **sondage aléatoire** sur la population des collégien-ne-s français-es, et que l'on a les moyens de faire passer environ 500 questionnaires.
- Vaut-il mieux :
  - Tirer indépendamment 500 collégien-ne-s dans la population d'intérêt ?
  - Tirer un collège et interroger les 500 collégien-nes-s qui y sont scolarisé-e-s ?
- Si l'on n'a pas de problème de non-réponse, dans les deux cas on peut estimer sans biais les caractéristiques moyennes
  - La probabilité de réponse de chaque collégien-ne-s français-e est la même !

# Est-ce seulement la taille d'échantillon qui compte ?

- La **taille d'échantillon** est la même dans les deux cas !
- Et pourtant intuitivement la précision n'est pas du tout la même
- C'est parce que les élèves d'un même collège se ressemblent plus que les élèves d'établissements différents
- En tirant un seul collège, on dégrade la précision parce que l'écart de chaque collégien-ne-s à la moyenne a des chances d'aller dans le même sens que celui des autres élèves de son établissement
- Cette idée se généralise à la précision de l'estimateur des MCO qui n'est après tout qu'une comparaison de moyennes

# Pour un exemple moins ridicule

- Retour sur un article discuté à la séance 2
  - Dans l'enquête Emploi en continu, le tirage est fait par grappes d'une vingtaine de logements situés les uns à côté des autres
  - Avantage pratique pour les enquêteurs : réduit la distance à parcourir et donc le coût de l'enquête
  - Mais on retombe potentiellement sur la même difficulté : les ménages qui vivent les uns à côté des autres ont tendance à se ressembler !

# Un détour par la théorie des sondages

- On se demander par exemple quel est le nombre de collégien-ne-s qui ont une chambre à eux dans leur logement
  - Pour le-a collégien-ne $i$, $x_i$ est l'indicatrice que c'est le cas
  - Ici ce n'est pas une variable aléatoire
    - L'aléa viendra du tirage
  - Le total que l'on veut estimer $\tau = \sum_{i \in \mathcal{I}} x_i$
- $S_i$ est la v.a. qui indique si $i$ est retenu dans l'échantillon
- $S_i = \tilde{S}_{c(i)}$ où $c \in \mathcal{C}$ est le collège dans lequel $i$ est scoalarisé-e

# Un détour par la théorie des sondages

- Un estimateur sans biais de $\tau$ est $\hat{\tau} = \sum_{i \in \mathcal{J}} \frac{S_i}{\mathbb{E}[S_i]} x_i$
- Quelle est sa variance ?

# Un détour par la théorie des sondages

- Si on ne passait pas par le collège pour faire le tirage : le tirage de deux collégien-ne-s différent-e-s est indépendant
  - $\mathbb{E}[\hat{\tau}^2] = \sum_{i,j\in\mathcal{J}} \frac{\mathbb{E}[S_i S_j]}{\mathbb{E}[S_i]\mathbb{E}[S_j]} x_i x_j$
  - Si $i \neq j$ alors $\mathbb{E}[S_i S_j] = \mathbb{E}[S_i]\mathbb{E}[S_j]$
  - Sinon $\mathbb{E}[S_i S_j] = \mathbb{E}[S_i]$
  - Donc $\mathbb{E}[\hat{\tau}^2] = \sum_{i\neq j} x_i xj + \sum_{i\in\mathcal{J}} \frac{1}{\mathbb{E}[S_i]} x_i^2$
  - Ou encore $\mathbb{E}[\hat{\tau}^2] = \left\{\sum_{i\in\mathcal{J}} x_i\right\}^2 + \sum_{i\in\mathcal{J}} \left(\frac{1}{\mathbb{E}[S_i]} - 1\right) x_i^2$
- Si le tirage est uniforme $\mathbb{E}[S_i] = p = \frac{n}{N}$ alors $\mathcal{V}(\hat{\tau}) = \frac{1-p}{p} \sum_i x_i^2$
  - Pour $p \ll 1$ on a bien une variance en $\frac{1}{n}$

# Un détour par la théorie des sondages

- On passe par le collège pour faire le tirage : le tirage de deux collégien-ne-s différent-e-s n'est pas indépendant !
  - $\mathbb{E}[\hat{\tau}^2] = \sum_{i,j \in \mathcal{I}} \frac{\mathbb{E}[\tilde{S}_{c(i)} \tilde{S}_{c(j)}]}{\mathbb{E}[\tilde{S}_{c(i)}]\mathbb{E}[\tilde{S}_{c(j)}]} x_i x_j$
  - Si $c(i) \neq c(j)$ alors $\mathbb{E}[\tilde{S}_{c(i)} \tilde{S}_{c(i]} = \mathbb{E}[\tilde{S}_{c(i)}]\mathbb{E}[\tilde{S}_{c(i)}]$
  - Sinon $\mathbb{E}[\tilde{S}_{c(i)} \tilde{S}_{c(i)}] = \mathbb{E}[\tilde{S}_{c(i)}]$
  - Donc $\mathbb{E}[\hat{\tau}^2] = \sum_{c(i) \neq c(j)} x_i x_j + \sum_{c \in \mathcal{C}} \frac{1}{\mathbb{E}[\tilde{S}_c]} x_i x_j$
  - Ou encore $\mathbb{E}[\hat{\tau}^2] = \left\{ \sum_{i \in \mathcal{I}} x_i \right\}^2 + \sum_{c \in \mathcal{C}} \left( \frac{1}{\mathbb{E}[\tilde{S}_c]} - 1 \right) \sum_{c(i)=c} \sum_{c(j)=c} x_i x_j$

# Un détour par la théorie des sondages

- Si le tirage est uniforme $\mathbb{E}[\tilde{S}_c] = \tilde{p} = \frac{\tilde{n}}{\tilde{N}}$ alors $\mathcal{V}(\hat{\tau}) = \frac{\tilde{p}-1}{\tilde{p}} \sum_{c \in \mathcal{C}} \sum_{c(i)=c} \sum_{c(j)=c} x_i x_j$
  - Pour $\tilde{p} \ll 1$ on a bien une variance en $\frac{1}{\tilde{n}}$
  - La variance dépend d'un terme $\sum_{c \in \mathcal{C}} \sum_{c(i)=c} \sum_{c(j)=c} x_i x_j$ qui quantifie combien les élèves d'un même collège se ressemblent (et aussi bien sûr la dispersion de $x_i$)

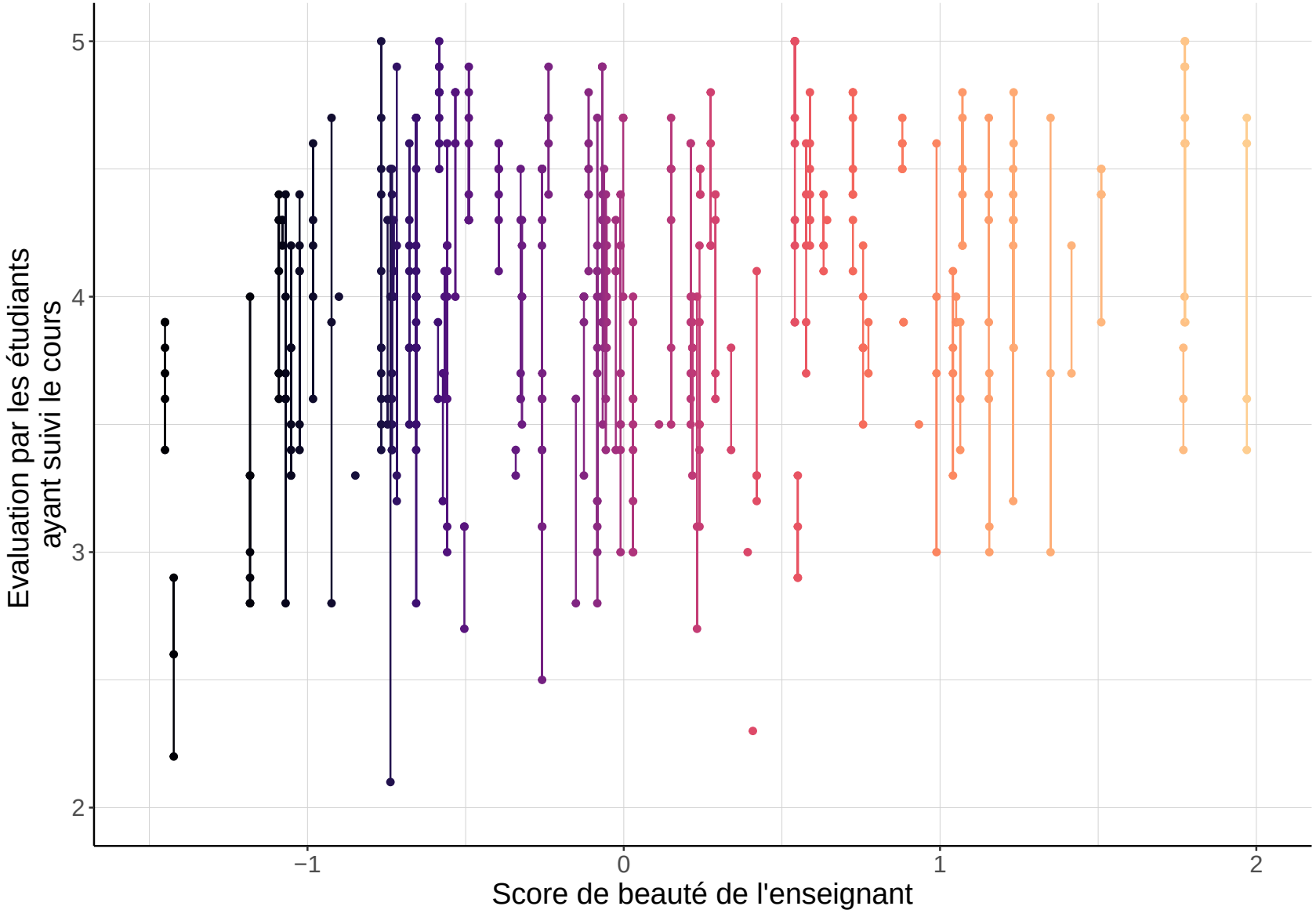
# Retour aux régressions linéaires

- On a jusque là étudié le comportement asymptotique de l'estimateur des MCO en supposant les **observations indépendantes les unes des autres**
- C'est une **approximation** utile
- Mais on doit parfois prendre en compte des violations de cette hypothèse
  - Soit parce que le **plan de sondage** qui a permis de générer les données ne se fait pas au niveau des individus de la population d'intérêt
    - Ce n'est que l'extension du cas que l'on vient de considérer
  - Soit parce que les variables que l'on regarde sont assignées à un niveau plus élevé

# Des régresseurs assignés à un niveau plus élevé ?

- Hamermesh D. et Parker A., 2005. “Beauty in the classroom : instructors’ pulchritude and putative pedagogical productivity”, *Economics of Education Review*, 24(4) :369-376.

# Des régresseurs assignés à un niveau plus élevé ?



# Des régresseurs assignés à un niveau plus élevé ?

- On veut régresser les scores attribués par les étudiant-e-s à leurs enseignant-e-s sur une mesure de la beauté de ces enseignant-e-s
- On a un score par enseignant-e-s mais les enseignant-e-s sont évalué-e-s pour chacun de leurs enseignements
- Si, par exemple, on mesure mal ce score pour une enseignant-e, alors cela se répercute sur toutes les obsevations relatives à cet-te enseignant-e
  - L'hypothèse d'indépendance cesse d'être crédible

# Quelle solution à ce problème ?

- Comme dans le cas plus simple examiné avant, on n'est pas sans recours !
- Avec de petites modifications on a toujours des résultats de **normalité asymptotique** et on sait toujours estimer la matrice de variance-covariance asymptotique sous une forme que l'on sait estimers
- Il faut quand même un peu de travail !

# Une petite modification du cadre

- Auparavant on tirait des individus indicés par $i$ dans une population infinie
- Maintenant on change de point de vue : on tire des *clusters* dans une population infinie de *clusters* indicés par $c$
  - Les collèges dans la discussion précédente
- Et pour chaque *cluster*, l'échantillon inclut les individus $i$ qui lui appartiennent
- Pour simplifier, on suppose que tous les *clusters* font la même taille
  - Même nombre d'individus
  - Hypothèse qui simplifie l'argument mais ne change rien au résultat

# Une petite modification du cadre

- On tire indépendamment $\tilde{n}$ *clusters* (dans la discussion précédente : les collèges) disjoints $C_1, \ldots, C_{\tilde{n}}$
- Dans chaque *cluster* $c$ on peut construire la matrice $\mathbf{X}_c$ construite comme $\mathbf{X}$ dans tout ce qui précède, mais spécifique à ce *cluster*
- Et le vecteur colonne $\mathbf{Y}_c$ qui empile toutes les variables dépendantes dans le même ordre
- Les couples de v.a. $(\mathbf{X}_{C_1}, \mathbf{Y}_{C_1}), \ldots, (\mathbf{X}_{C_{\tilde{n}}}, \mathbf{Y}_{C_{\tilde{n}}})$ sont indépendants et identiquement distribués

# Une petite modification du cadre

- On construit un score pour les *clusters* :
  $S_c = \sum_{i \in c} X_i \epsilon_i = \mathbf{X}_c' \epsilon_c$
  - $S_c$ est un vecteur de dimension $d$
  - $\epsilon_c$ est construit sur le même principe que $\mathbf{X}_c$
- $\mathbb{E}[S_c] = \mathbb{E}\left[\sum_{i \in c} X_i \epsilon_i\right]$ donc la condition d'orthogonalité $\mathbb{E}[X\epsilon] = 0$ impose $\mathbb{E}[S_c] = 0$

# Une petite modification du cadre

- Petit calcul matriciel : $\mathbf{X}_c'\mathbf{X}_c = \sum_{i \in c} X_i X_i'$
- Donc $\mathbf{X}'\mathbf{X} = \sum_{k=1}^{\tilde{n}} \sum_{i \in C_k} X_i X_i' = \sum_{i=1}^{\tilde{n}} \mathbf{X}_{C_k}' \mathbf{X}_{C_k}$
- De la même façon
  $\mathbf{X}'\epsilon = \sum_{k=1}^{\tilde{n}} \sum_{i \in C_k} X_i \epsilon_i = \sum_{k=1}^{\tilde{n}} \mathbf{X}_{C_k} \epsilon_{C_k} = \sum_{i=1}^{\tilde{n}} S_{C_k}$
  - Par conséquent : $\hat{\beta} - \beta = \left( \sum_{k=1}^{\tilde{n}} \mathbf{X}_{C_k}' \mathbf{X}_{C_k} \right)^{-1} \sum_{k=1}^{\tilde{n}} S_{C_k}$

# On passe à la limite

- $\frac{1}{\tilde{n}} \sum_{k=1}^{\tilde{n}} \mathbf{X}'_{C_k} \mathbf{X}_{C_k}$ converge en probabilité vers $\mathbb{E}[\mathbf{X}'_c \mathbf{X}_c]$ en vertu de la loi faible des grands nombres
  - Attention : l'espérance ici correspond à une moyenne sur les *clusters* !
- $\sqrt{\tilde{n}} \left( \frac{1}{\tilde{n}} \sum_{k=1}^{\tilde{n}} S_{C_k} \right)$ converge en loi vers $\mathcal{N}(0, \mathcal{V}(S_c))$ en vertu du TCL
- Théorème de Slustky : $\sqrt{\tilde{n}} \left\{ \hat{\beta} - \beta \right\}$ converge en loi vers $\mathcal{N}(0, \mathbb{E}[\mathbf{X}'_c \mathbf{X}_c]^{-1} \mathcal{V}(S_c) \mathbb{E}[\mathbf{X}'_c \mathbf{X}_c]^{-1})$

# Comment estimer la matrice de variance-covariance en pratique ?

- On ne connaît pas les valeurs $S_c$, ni *a fortiori* les espérances et les variances
- On peut estimer $\mathbb{E}[\mathbf{X}_c'\mathbf{X}_c]$ par $\frac{1}{\tilde{n}}\sum_{k=1}^{\tilde{n}}\mathbf{X}_{C_k}'\mathbf{X}_{C_k}$
- Pour le terme $\mathcal{V}(S_c)$
  - On connaît les résidus estimés $\hat{\epsilon}_i = Y_i - X_i'\hat{\beta} \neq Y_i - X_i'\beta = \epsilon_i$
  - Par construction la moyenne dans l'échantillon $\frac{1}{\tilde{n}}\sum_{k=1}^{\tilde{n}}\sum_{i\in C_k} X_i\hat{\epsilon}_i$ est nulle
  - On peut estimer la matrice de variance-covariance par $\frac{1}{\tilde{n}}\sum_{k=1}^{\tilde{n}} S_{C_k}S_{C_k}' = \frac{1}{\tilde{n}}\sum_{k=1}^{\tilde{n}}\left\{\sum_{i\in C_k} X_i\hat{\epsilon}_i\right\}\left\{\sum_{i\in C_k} X_i\hat{\epsilon}_i\right\}'$

# Une petite vérification (la dernière !)

```
#On charge dans un premier temps les données nécessaires
data("TeachingRatings")
TeachingRatings <-
  data.table::data.table(TeachingRatings)
```

# Une petite vérification (la dernière !)

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

# Une petite vérification (la dernière !)

```
#On récupère les résidus estimés
residus<-reg_eval$residuals

#On récupère la matrice X
matrice_X <- reg_eval$x

#On récupère les poids
# (régression pondérée : on va en avoir besoin)
poids <- reg_eval$weights
somme_poids <- sum(poids)
```

# Une petite vérification (la dernière !)

```
XprimeX <-
  1 / somme_poids *
  t(sqrt(poids) * matrice_X) %*% (matrice_X * sqrt(poids))
```

# Une petite vérification (la dernière !)

```
#La vraie difficulté : on veut estimer le terme
# central
#Il faut : sommer toutes les variables
# indépendantes*residus estimé à  l'intérieur de
# chaque  cluster = le groupe des observations
# correspondant à un enseignant
Xepsilon<-cbind(matrice_X * residus,
                TeachingRatings[,"prof"],
                #la définition des clusters
                poids)
```

# Une petite vérification (la dernière !)

```
score_cluster<-
 Xepsilon[,
          lapply(X=.SD,
                 FUN=function(x){
                   sum(x*poids)/somme_poids
                   }),
          .SDcols=colnames(matrice_X),
          by="prof"]
```

# Une petite vérification (la dernière)

```
nb_clusters <-
  length(unique(TeachingRatings$prof))

meat <-
  1/nb_clusters*
  t(as.matrix(
    score_cluster[,
                  .SD,
                  .SDcols=
                    names(score_cluster)!=
                    "prof"]))%*%
  as.matrix(
    score_cluster[,
                  .SD,
                  .SDcols=
                    names(score_cluster)!=
                    "prof"])
```

# Une petite vérification (la dernière)

```
#Enfin on multiplie par X'X^-1 des deux côtés
vcov_sandwich_clusters <-
  nb_clusters / (nb_clusters-1) *
  1 / nb_clusters*
  solve(1 / nb_clusters * XprimeX) %*%
  meat %*%
  solve(1 / nb_clusters * XprimeX)
```

# Une petite vérification (la dernière)

```
#On vérifie que la matrice de variance-covariance
# robuste à la dépendance estimée par R est égale
# à celle que l'on vient d'estimer
all.equal(as.numeric(vcov_sandwich_clusters),
          as.numeric(vcovCL(reg_eval,
                            type="HC0",
                            cluster=TeachingRatings$prof)))
```

```
[1] TRUE
```

# Une petite vérification (la dernière)

```
#En fait l'estimateur par défaut utilise
# encore une correction supplémentaire
# qui prend en compte la dimension des
# variables indépendendantes (qui réduit
# les degrés de  liberté)
all.equal(
  (nrow(TeachingRatings) - 1) /
    (nrow(TeachingRatings) -
       ncol(matrice_X)) *
    as.numeric(vcov_sandwich_clusters),
  as.numeric(vcovCL(reg_eval,
                    cluster=TeachingRatings$prof)))
```

```
[1] TRUE
```

# Ne pas prendre en compte les violations de l'indépendance conduit à sous-estimer l'incertitude

```
#Ne pas prendre en compte cette structure de
# dépendance conduit à sous-estimer l'écart-type
#du coefficient sur le score de beauté de plus de 38%
1-
  sqrt(vcov(reg_eval)["beauty",
                      "beauty"])/
  sqrt(vcovCL(reg_eval,
              cluster=
                TeachingRatings$prof)["beauty",
                                      "beauty"])
```

```
[1] 0.5301468
```

# Ne pas prendre en compte les violations de l'indépendance conduit à sous-estimer l'incertitude

```
1-
  sqrt(vcovHC(reg_eval)["beauty",
                        "beauty"])/
  sqrt(vcovCL(reg_eval,
              cluster=
                TeachingRatings$prof)["beauty",
                                      "beauty"])
```

```
[1] 0.384267
```

# Quelques remarques finales sur l'inférence pour les données groupées

- Quel est le bon niveau pour le *clustering* ?
  - Cela dépend de la question à laquelle on cherche à répondre et des données que l'on utilise
  - Traitement de la question par Abadie et al. (2023) :
    - Si on traite une question causale : le niveau auquel le traitement est assigné $\rightarrow$ OK ici si on s'intéresse à l'effet causal de la beauté !
    - Si on utilise des données avec un plan de sondage en grappe : le niveau des grappes
- Il peut donc y avoir plusieurs niveaux pertinents !
- Remarque : si c'est la dimension causale qui motive la décision, alors le cadre retenu jusqu'ici pour l'inférence n'est pas nécessairement le plus pertinent cf Abadie et al. (2020)

# Quelques remarques finales sur l'inférence pour les données groupées

- Que faire lorsqu'il y a plusieurs niveaux pertinents ?
  - Certains logiciels proposent un calcul avec plusieurs niveaux de *clustering*
  - L'estimateur utilisé est parfois problématique
  - Davezies, D'Haultfœuille et Guyonvarch (2021) : la solution théoriquement la plus pertinente est
    1. d'estimer la matrice de variance-covariance séparément pour chaque niveau de clustering
    2. sommer ces matrices de variance-covariance pour obtenir la matrice de variance-covariance correcte

# Conclusion

- On sait à présent :
  - Interpréter les coefficients d'une régression linéaire par les moindres carrés ordinaires
  - Estimer ces coefficients à partir d'un échantillon fini
  - Estimer l'incertitude, c'est-à-dire quantifier l'écart vraisemblable entre cette estimation et la vraie valeur pour toute la population

# Conclusion

- On sait que l'estimateur se comporte pour des échantillons suffisamment grands, comme un vecteur gaussien dont on connaît la matrice de variance-covariance
- A la séance prochaine, nous verrons comment utiliser cette idée pour tester des hypothèses