Séance 4 – Régression linéaire simple

Pierre Pora

Introduction

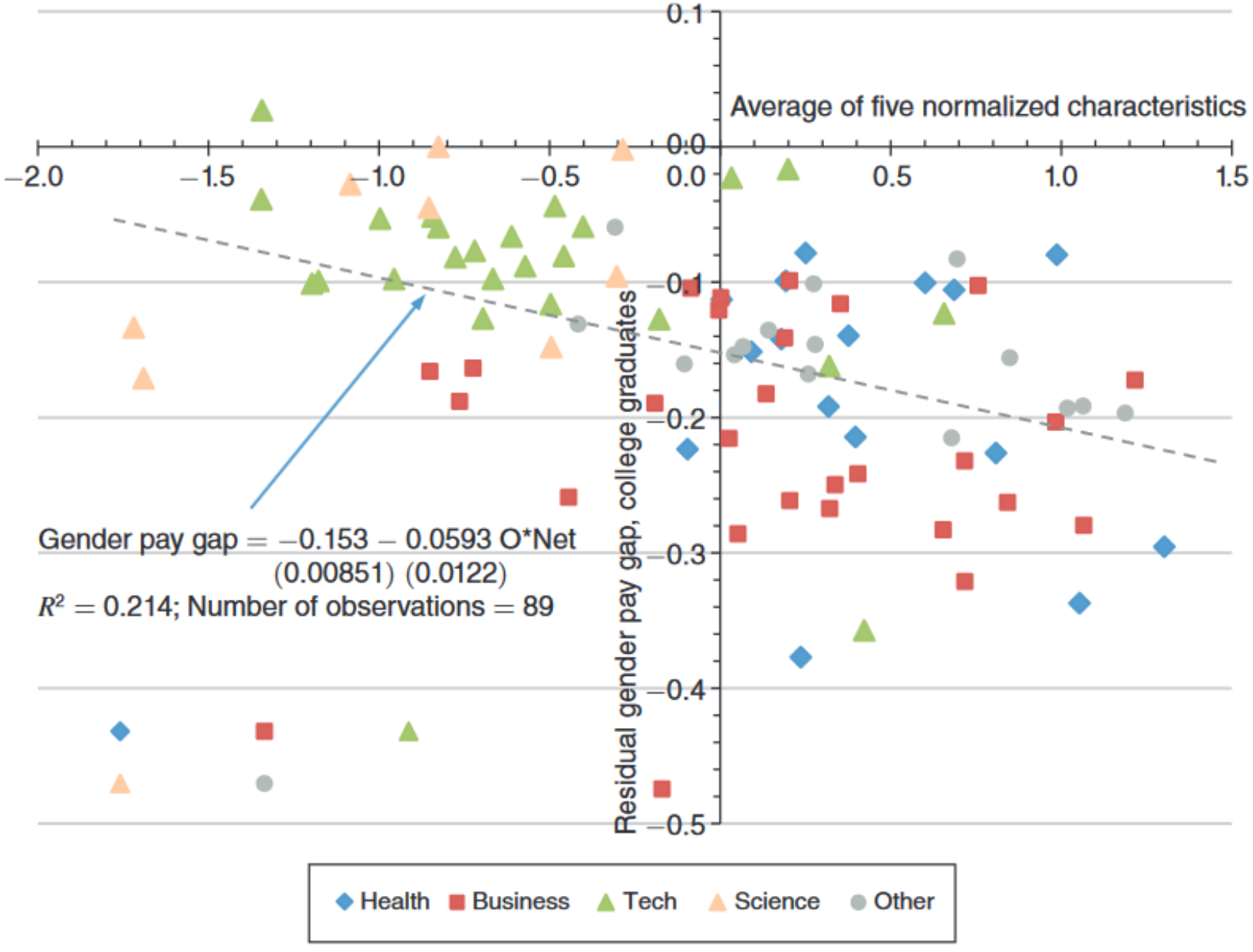
Rappel des épisodes précédents

- Pour mettre à l'épreuve leurs théories, les économistes ont en général recours à des **données quantitatives** de nature variée
- On dispose d'un **formalisme probabiliste** qui permet notamment de justifier que des données portant sur un nombre fini d'individus sont informatives sur une population beaucoup plus grande

Rappel des épisodes précédents

- Concepts permettant de commencer à quantifier le lien entre deux variables :
  - **indépendance**
  - **espérance conditionnelle** / indépendance en moyenne
  - **covariance** et **corrélation**

L'objet de la séance : comment construit-on cette droite ?



L'objet de la séance

- Comme son titre l'indique : **régression linéaire simple**
  - Par les **moindres carrés ordinaires** (MCO)
- La séance la plus importante du semestre
  - Pas si difficile d'étendre aux régressions multiples une fois qu'on a en tête la bonne mécanique dans le cas simple

L'objet de la séance

- Séance qui risque d'être longue
  - Au pire on finira la semaine prochaine
- Expose de **nombreuses façons équivalentes de voir les MCO dans le cadre de la régression simple**
  - Construire son intuition de l'objet
    - La façon intuitive de penser à cette approche n'est pas la même d'une personne à l'autre !
  - Passer d'un point de vue à l'autre est souvent utile pour réfléchir
    - Une forme d'agilité à acquérir progressivement

Sont-ce des statistiques descriptives ? Oui !

- Ce n'est rien d'autre que q'une **comparaison de moyennes** !
  - Plein de façons de le voir $\rightarrow$ c'est ce qu'on va regarder ensemble
- Il n'y a aucune raison de penser que parce qu'on fait une régression, cela cesse d'être descriptif !
- Si cela cesse d'être descriptif, c'est parce que l'on veut connecter les quantités relatives à la régression à une **(proto-)théorie économique**
  - Interprétation causale
  - Identification d'un paramètre économique d'intérêt (e.g. élasticité)

Sont-ce des statistiques descriptives ? Oui !

- Ce qu'il y a dans la régression en tant que telle n'a rien à voir avec cette couche d'interprétation qu'on rajoute par-dessus
- Cette couche est parfois très bonne, parfois affreuse, et il y a de bonnes recettes en la matière
- Mais **la régression en tant que telle n'a rien à voir là-dedans**
  - Il faut aborder ça aussi simplement que n'importe quelle façon de résumer l'information contenue dans des données quantitatives

Partir d'un exemple simple

```
library(AER)

data("CPS1985")
```

Une première définition

Partir d'un exemple simple

- Echantillon issu du CPS américain
- `education` représente le temps passé dans le système scolaire
- `wage` représente le salaire horaire

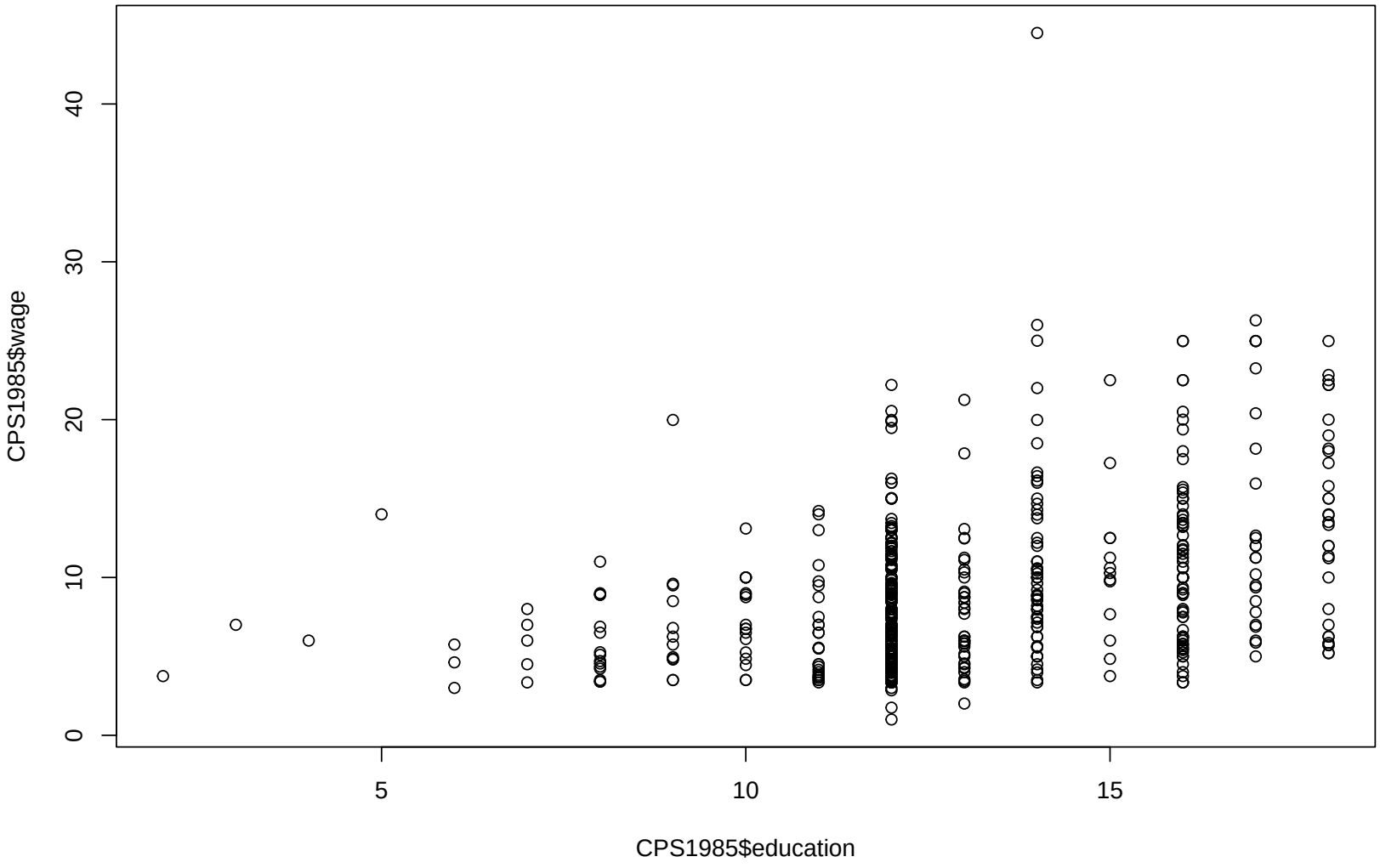
Partir d'un exemple simple

```
help(CPS1985)
```

Partir d'un exemple simple

```
plot(CPS1985$education, CPS1985$wage)
```

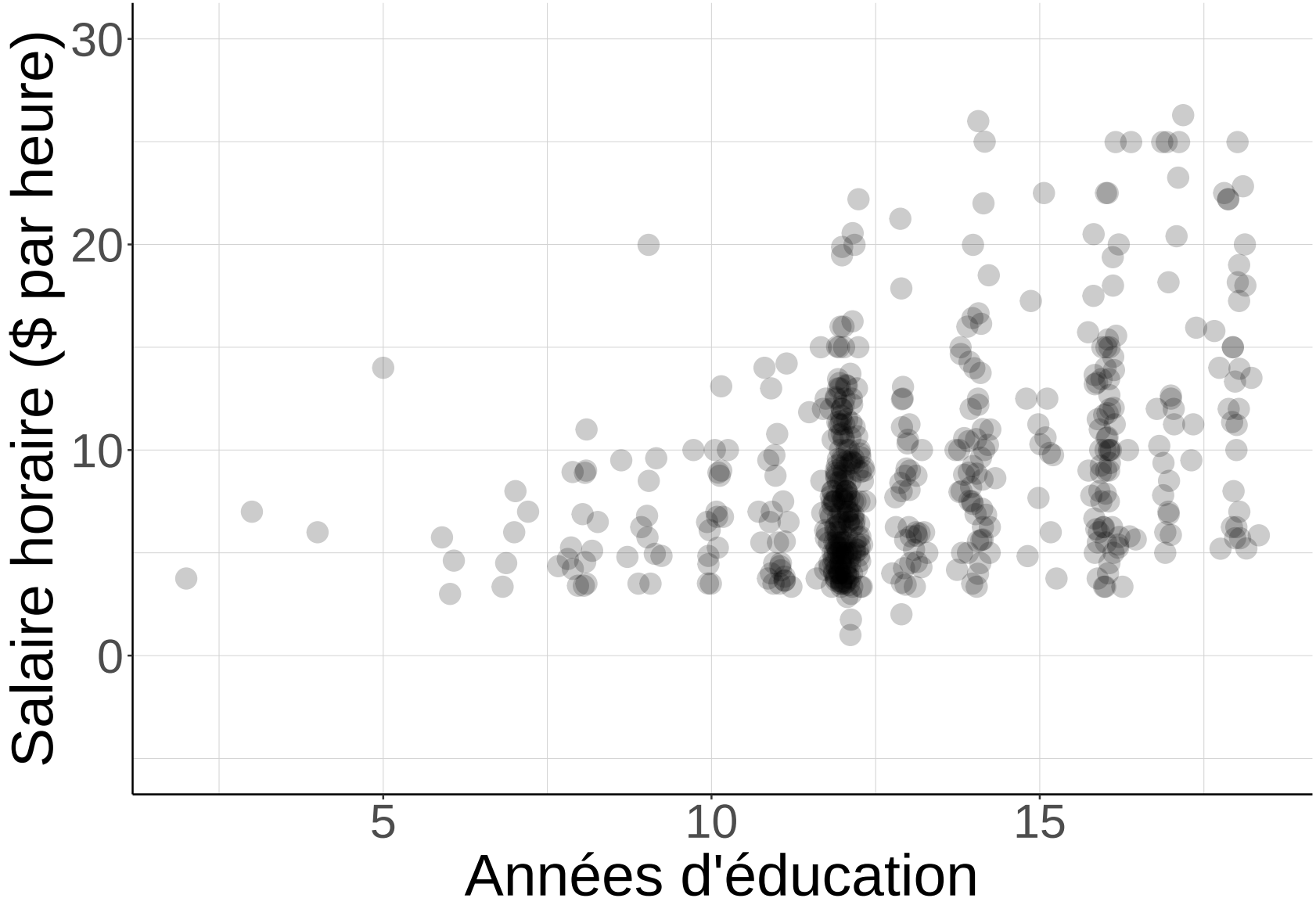
Partir d'un exemple simple



Aparté : réalisation de graphique

- `plot` permet de réaliser des graphiques très basiques
- Le package `ggplot2` est le **package le plus utilisé de R pour réaliser des graphiques**
  - Permet de réaliser des choses un peu peu sophistiquées que `plot`
  - Nombreuses options pour personnaliser
  - Apprentissage = investissement rentable $\rightarrow$ faire des figures est une des façons les plus efficaces de communiquer avec d'autres économistes !

Partir d'un exemple simple



Partir d'un exemple simple

- Remarque : à ce stade, **on ignore complètement le fait qu'on travaille sur un échantillon**
- On va faire comme si on travaillait déjà sur la population entière, et donc s'abstraire de la distinction entre les quantités que l'on veut estimer, qui portent sur toute la population, et leur estimation à partir du petit échantillon disponible dans le CPS
- Ce sera le cas pour toute la séance, et aussi toute la séance suivante
- Les questions liées à l'**échantillonage** et à la **précision** arriveront à la séance 6

Partir d'un exemple simple

- On voudrait construire **une droite qui résume ce nuage de points**
- A ce stade pas de discussion de l'interprétation théorique / causale de cette droite !
- Tout bonnement une droite = 2 paramètres
  - Ordonnée à l'origine
  - Et surtout : pente !
  - Façon efficace de **synthétiser l'information contenue dans les données**

Partir d'un exemple simple

- On voudrait construire **une droite qui résume ce nuage de points**
- Comment procéder ?

Partir d'un exemple simple

- Comment procéder ?
  - Utiliser tous les couples de points
  - Comparer les salaires moyens de niveaux d'éducation adjacents
  - Minimiser la distance entre le nuage et la droite
  - ...
- En fait une **régression linéaire simple par les moindres carrés ordinaires** fait tout cela !
  - Ca ne se voit pas forcément dans la définition

Le théorème qui justifie tout

- Le salaire `wage` peut se **décomposer de façon unique** en la somme de :
  - une fonction affine de l'éducation `education`, que l'on peut noter $\widehat{\texttt{wage}} = \alpha + \beta\,\texttt{education}$
  - une terme résiduel ϵ d'espérance nulle et non-corrélé à l'éducation

Pourquoi c'est vrai ?

- Supposons que l'on dispose d'un tel ϵ
- Alors $\mathcal{C}(\epsilon, \texttt{education}) = 0$
  - Mais $\epsilon = \texttt{wage} - \alpha - \beta\,\texttt{education}$
  - Donc $\mathcal{C}(\texttt{wage}, \texttt{education}) - \beta\mathcal{C}(\texttt{education}, \texttt{education}) = 0$
  - On en tire $\beta = \frac{\mathcal{C}(\texttt{wage},\texttt{education})}{\mathcal{V}(\texttt{education})}$
  - Et en utilisant $\mathbb{E}[\epsilon] = 0$ on récupère finalement $\alpha = \mathbb{E}[\texttt{wage}] + \frac{\mathcal{C}(\texttt{wage},\texttt{education})}{\mathcal{V}(\texttt{education})}\mathbb{E}[\texttt{education}]$
  - En définitive :
    - $\widehat{\texttt{wage}} = \mathbb{E}[\texttt{wage}] + \frac{\mathcal{C}(\texttt{wage},\texttt{education})}{\mathcal{V}(\texttt{education})}\{\texttt{education} - \mathbb{E}[\texttt{education}]\}$
    - $\epsilon = \texttt{wage} - \widehat{\texttt{wage}}$

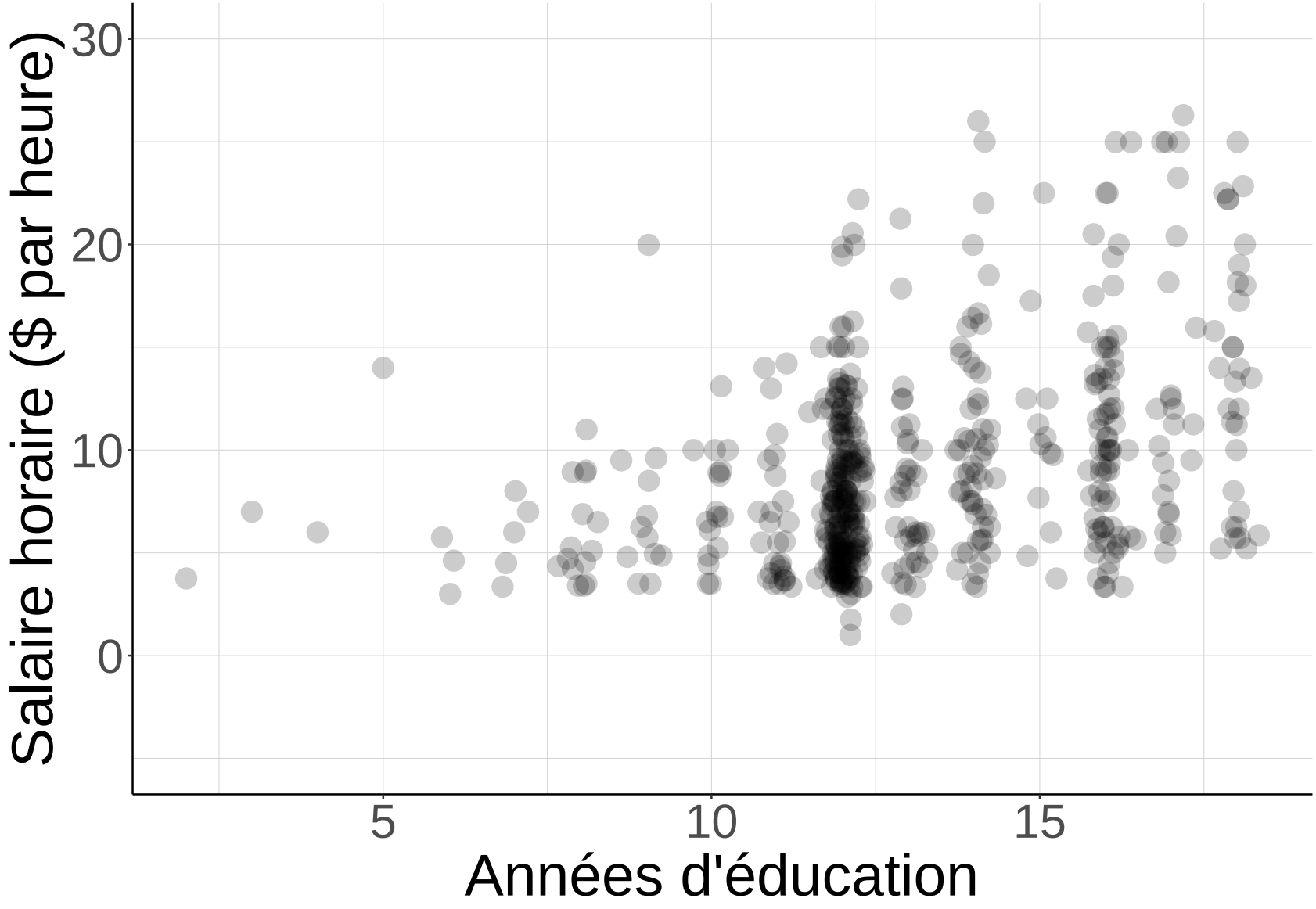
Pourquoi c'est vrai ?

- Réciproquement, on peut vérifier que cela définit bien $\widehat{\texttt{wage}}$ et ϵ, et que dans ce cas, ϵ est de moyenne nulle et non-corrélé à `education`
- On voit que tout cela ne peut fonctionner qu'à condition que $\mathcal{V}(\texttt{education}) \neq 0$

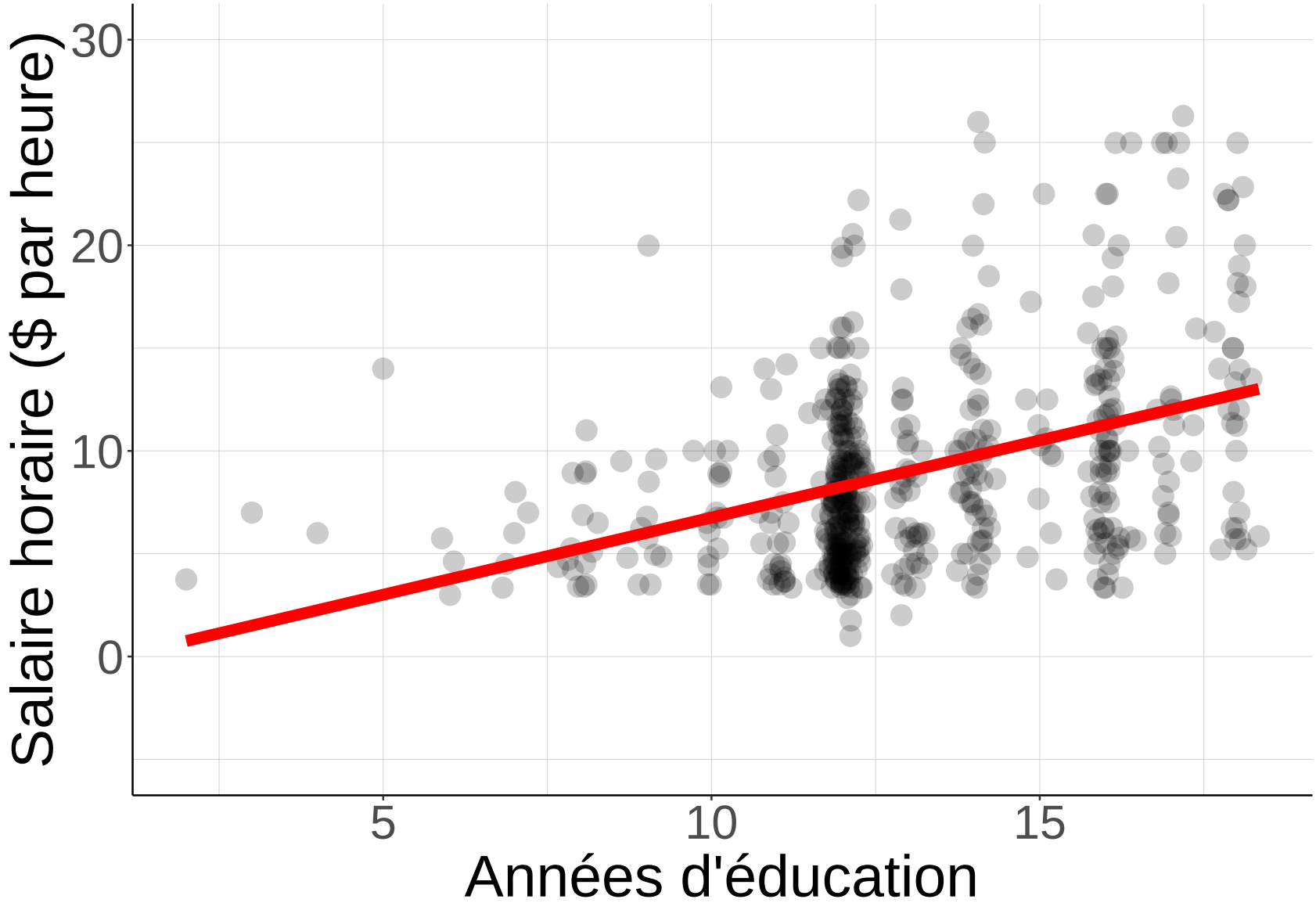
Pourquoi c'est vrai ?

- Si ce n'était pas le cas, alors `education` serait une constante : toute la population recevrait le même niveau d'éducation
  - Visuellement tous les points sont sur une droite verticale
  - La question devient beaucoup moins intéressante
  - On dit que **les coefficients ne sont pas identifiés**
  - Différent de la question de l'**identification des effets causaux** !
- C'est la **seule condition nécessaire** pour que tout cela fonctionne
  - Aucun besoin d'hypothèses supplémentaires !

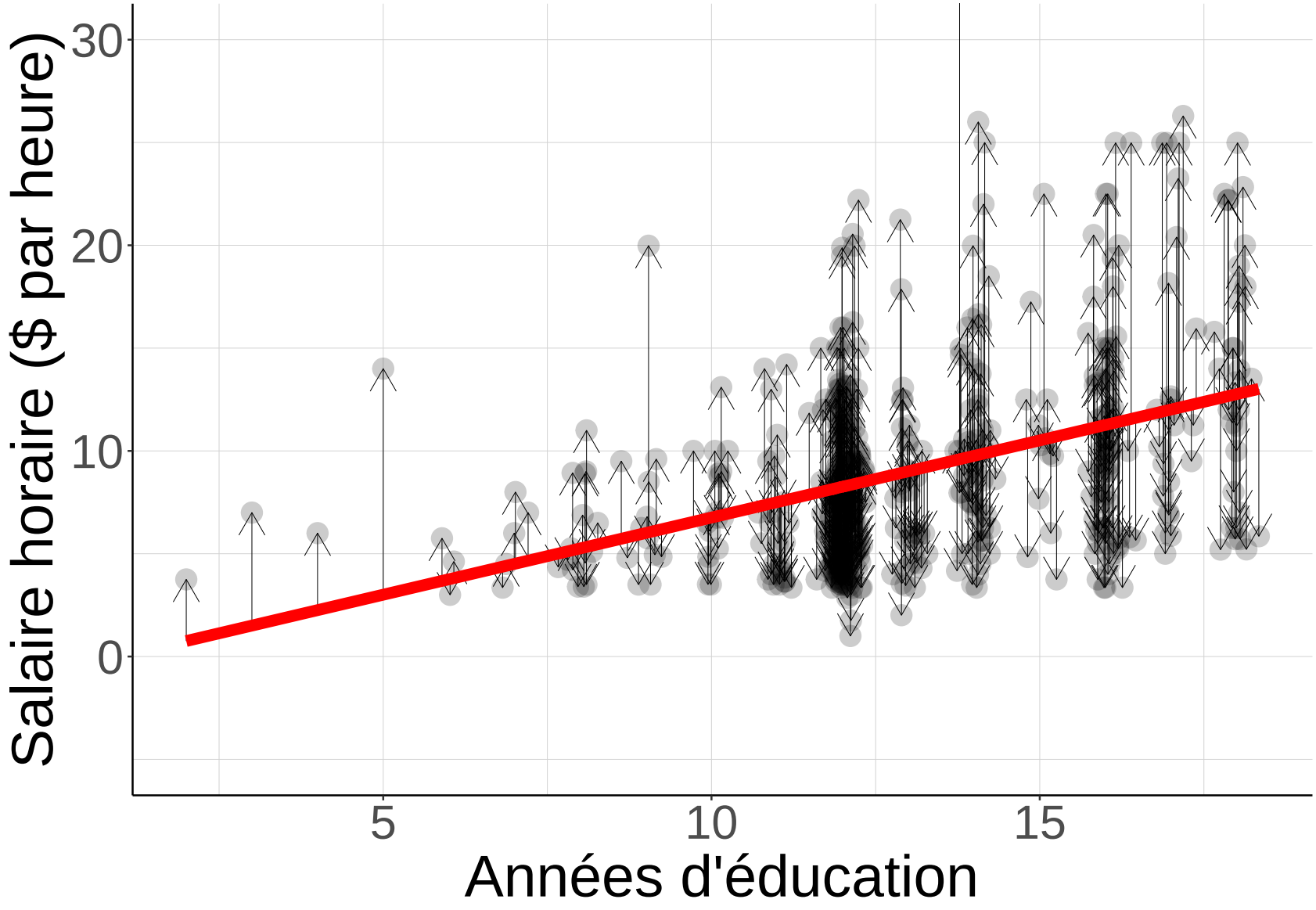
A quoi ça ressemble ?



A quoi ça ressemble ?



A quoi ça ressemble ?



Un peu de vocabulaire

- Dans la **régression** précédente :
  - On dit que `wage` est la **variable expliquée** / la **variable dépendante** / la **variable endogène** /
  - `education` est la **variable explicative** / la **variable indépendante** / la **variable exogène** / le **régresseur** / la **covariable**
  - α et β sont les **coefficients**
    - α est parfois appelé ***intercept***
  - $\widehat{\texttt{wage}}$ est la **valeur prédite**
  - ϵ est le **résidu** / le **choc idiosyncratique** / le **terme d'erreur**

Un peu de vocabulaire

- Beaucoup de ces termes sont en lien avec l'**interprétation causale** de la régression
  - En tant que tels ils peuvent un peu induire en erreur !
  - On fait abstraction ici de cette interprétation causale !
    - Je préfère donc dire variable dépendante / variable indépendante-régresseur-covariable / résidu
- Les autres termes sont plus là pour vous permettre de comprendre de quoi on parle quand on les utilise

En pratique avec R

```
regression <- lm(formula = wage ~ education,
                 data = CPS1985)

regression$coefficients
```

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

Une vérification bienvenue

```
beta_a_la_main <-
  cov(CPS1985$wage,
      CPS1985$education) /
  var(CPS1985$education)

alpha_a_la_main <-
  mean(CPS1985$wage) -
  beta_a_la_main * mean(CPS1985$education)

beta_a_la_main
```

```
[1] 0.7504608
```

```
alpha_a_la_main
```

```
[1] -0.7459797
```

Une vérification bienvenue

```
all.equal(
  as.numeric(regression$coefficients["education"]),
  beta_a_la_main)
```

```
[1] TRUE
```

```
all.equal(
  as.numeric(regression$coefficients["(Intercept)"]),
  alpha_a_la_main)
```

```
[1] TRUE
```

Une vérification bienvenue

```
CPS1985$salaire_predit <-
  alpha_a_la_main +
  beta_a_la_main * CPS1985$education

CPS1985$residu <-
  CPS1985$wage -
  CPS1985$salaire_predit

all.equal(as.numeric(regression$fitted.values),
          CPS1985$salaire_predit)
```

```
[1] TRUE
```

```
all.equal(as.numeric(regression$residuals),
          CPS1985$residu)
```

```
[1] TRUE
```

Une vérification bienvenue

```
cov(CPS1985$residu,
    CPS1985$education)
```

```
[1] -1.822192e-16
```

```
all.equal(cov(CPS1985$residu,
              CPS1985$education),
          0)
```

```
[1] TRUE
```

Aparté : interprétation causale

Une remarque sur l'interprétation causale (et on n'en reparle plus !)

- Si l'éducation est assignée aux individus indépendamment de tous les autres déterminants (observés ou non !) du salaire
  - Par exemple par une loterie, ou une expérience aléatoire contrôlée
- Et si les rendements de l'éducation sont constants
  - Entre individus
  - Quelle que soit la variation envisagée

Une remarque sur l'interprétation causale (et on n'en reparle plus !)

- Alors ϵ tel que $\texttt{wage} = \alpha + \beta\texttt{education} + \epsilon$, $\mathbb{E}[\epsilon] = 0$ et $\mathcal{C}(\epsilon, \texttt{education})$ peut s'interpréter comme l'ensemble des déterminants (observés ou non) du salaire autres que l'éducation
  - En particulier $\mathcal{C}(\epsilon, \texttt{education}) = 0$ est garanti par l'assignation aléatoire
- Dans ce cas β se lit comme l'effet d'une année d'éducation supplémentaire sur le salaire
  - On se connecte à la littérature sur les rendements de l'éducation

Une remarque sur l'interprétation causale (et on n'en reparle plus !)

- Dans cette situation, on dit souvent que la régression identifie les rendements de l'éducation / l'effet causal de l'éducation etc.
  - C'est différent de savoir si le coefficient de la régression est identifié !
  - La régression est légitime même en dehors de cette situation
    - Tant qu'on ne se met pas à dire que β a une interprétation causale comme rendement
  - L'hypothèse d'exogénéité n'a d'intérêt que si on veut avoir cette interprétation !
    - Bonne raison de s'y intéresser mais pas la seule !

Une remarque sur l'interprétation causale (et on n'en reparle plus !)

- Et si les effets sont hétérogènes dans la population, peut-on dire que β représente un effet moyen ?
- La seule façon de le savoir c'est de s'intéresser à la construction de β sans se poser la question de l'interprétation causale
  - Motivation pour cette construction de l'enseignement
  - C'est ce que l'on va faire dans toute la suite de la séance

La suite

- On s'est donné un moyen de construire une droite qui résume en quelque sorte un nuage de points
- Le moyen utilisé pour le faire n'est pas parfaitement évident *a priori*
- En fait cette façon de faire est (à peu près) équivalente à **plein d'autres façons** peut-être plus évidentes dont on aurait pu procéder

Une construction élémentaire

Faire des comparaisons deux à deux

- **Par deux points différents passe une droite et une seule**
- Dans notre nuage de points on a finalement pleins de couples de points disponibles !
- Donc plein de droites que l'on pourrait construire de cette façon très simple

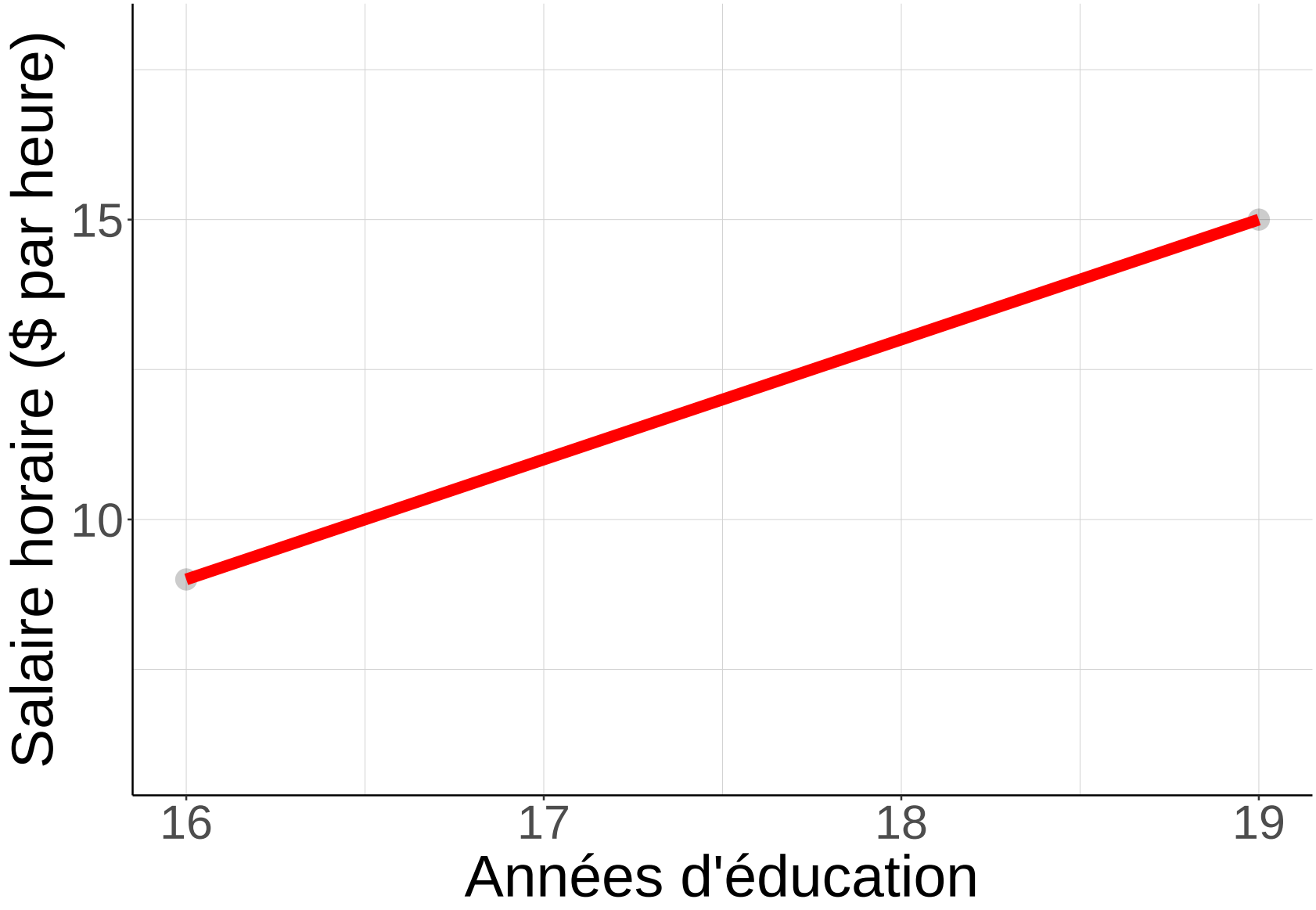
Un cas fictif pour vérifier que cela peut fonctionner

- Un monde dans lequel il n'y aurait que deux positions possibles, disons
  - `education` $= 13$ et `wage` $= 9$
  - `education` $= 16$ et `wage` $= 15$
- La droite qui passe par ses deux points a pour pente 2
- Et la droite de régression ?

Un cas fictif pour vérifier que cela peut fonctionner

- Dans ce cas, en revenant à la décomposition entre valeur prédite et résidu
  - Le résidu est nécessairement constant !
    - S'il était par exemple plus grand pour le groupe plus éduqué alors la corrélation avec l'éducation ne serait pas nulle
    - Comme il est aussi de moyenne nulle, il est nul pour tout le monde
  - La droite de régression passe donc par les deux points
  - C'est donc la même droite qu'avant et sa pente est 2 !

Un cas fictif pour vérifier que cela peut fonctionner



Et dans le cas général ?

- Dans le nuage de points, on peut considérer tous les couples de points différents possibles sans restriction
- Par chacun de ces couples passe une unique droite
  - Cela nous donne une pente pour chaque couple
- Comment en tirer une seule pente qui vaille pour toute la population ?

Et dans le cas général ?

- On prend simplement la **moyenne de ces pentes**
  - Il faut choisir une **pondération appropriée** pour cette moyenne

- Avec un petit calcul, on peut montrer que la pente de la droite de régression est :
  - la **moyenne de la pente de tous les couples de points** représentant des individus dont le niveau d'éducation diffère (sinon la pente pour ce couple est infinie)

  - avec des **poids proportionnels au carré de l'écart** d'éducation entre les individus concernés

Une petite vérification

```
#On crée tous les couples de points possibles
CPS1985$key<-1
CPS_couple<-
  merge(CPS1985[,c("key","wage","education")],
        CPS1985[,c("key","wage","education")],
        by="key",
        allow.cartesian=TRUE)

CPS_couple <- data.table(CPS_couple)
```

Une petite vérification

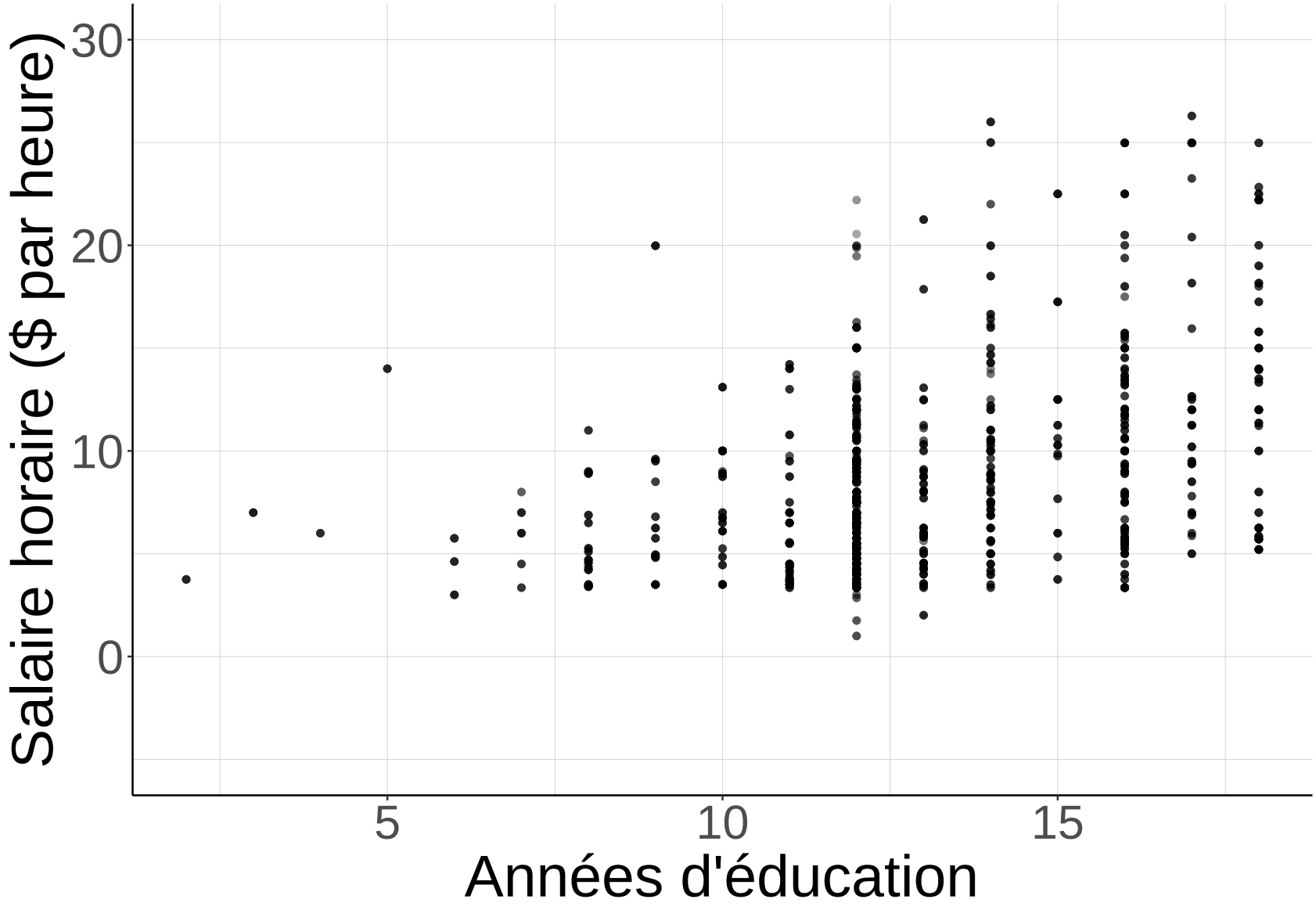
```
#On calcule la pente pour chacun d'entre eux lorsque
# le niveau d'éducation diffère
CPS_couple[education.x != education.y,
           pente :=
             (wage.x-wage.y)/
             (education.x-education.y)]

#On calcule le carré de l'écart d'éducation
CPS_couple[,
           carre_ecart_educ :=
             (education.x-education.y)^2]
```

Une petite vérification

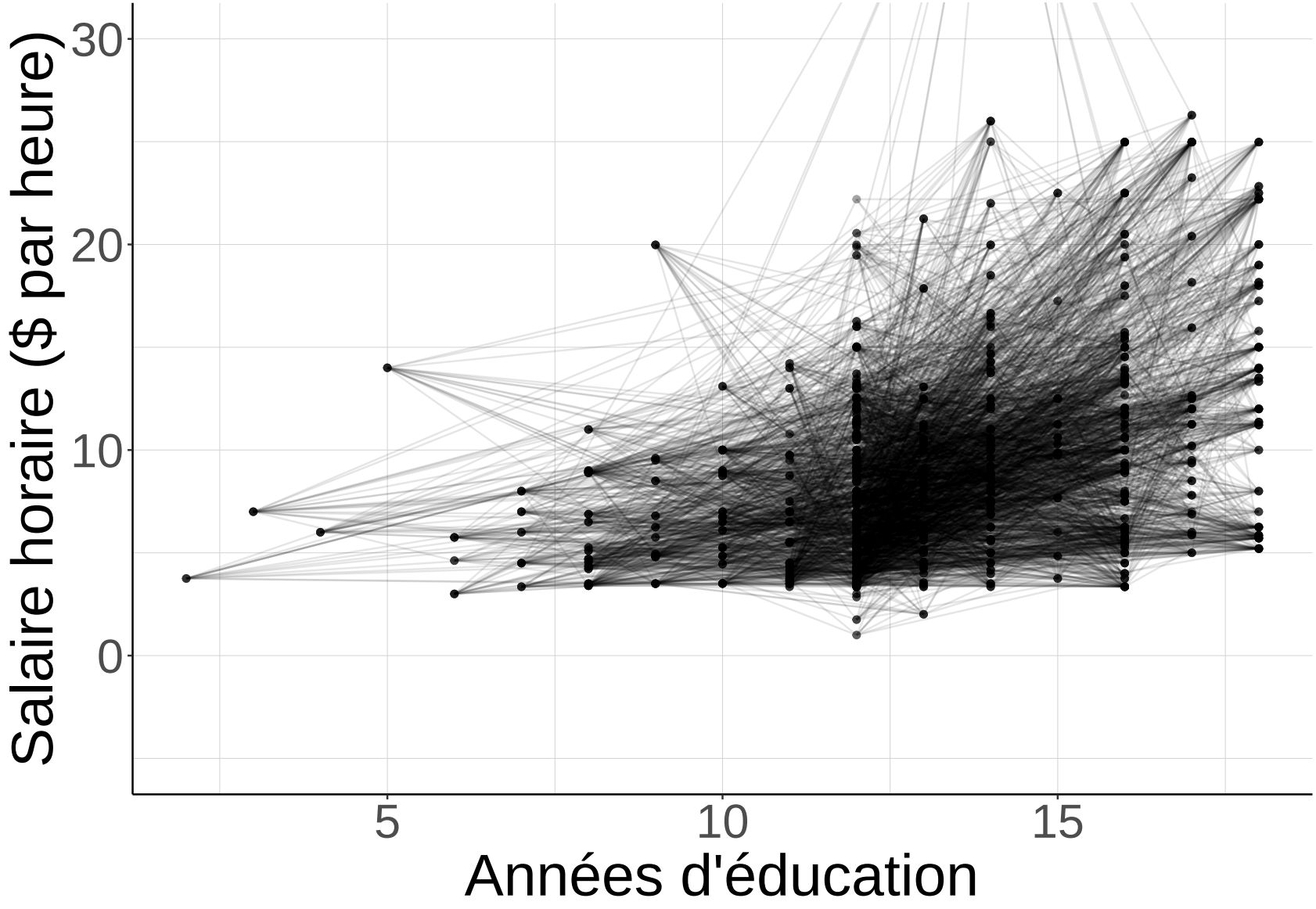
```
#On estime enfin la moyenne pondérée des pentes
pente_moyenne<-
  CPS_couple[,
             list(sum(pente*carre_ecart_educ,
                      na.rm=TRUE)/
                    sum(carre_ecart_educ))]

#On peut enfin comparer les résultats des deux
# approches
all.equal(
  as.numeric(regression$coefficients["education"]),
  as.numeric(pente_moyenne))
```

```
[1] TRUE
```

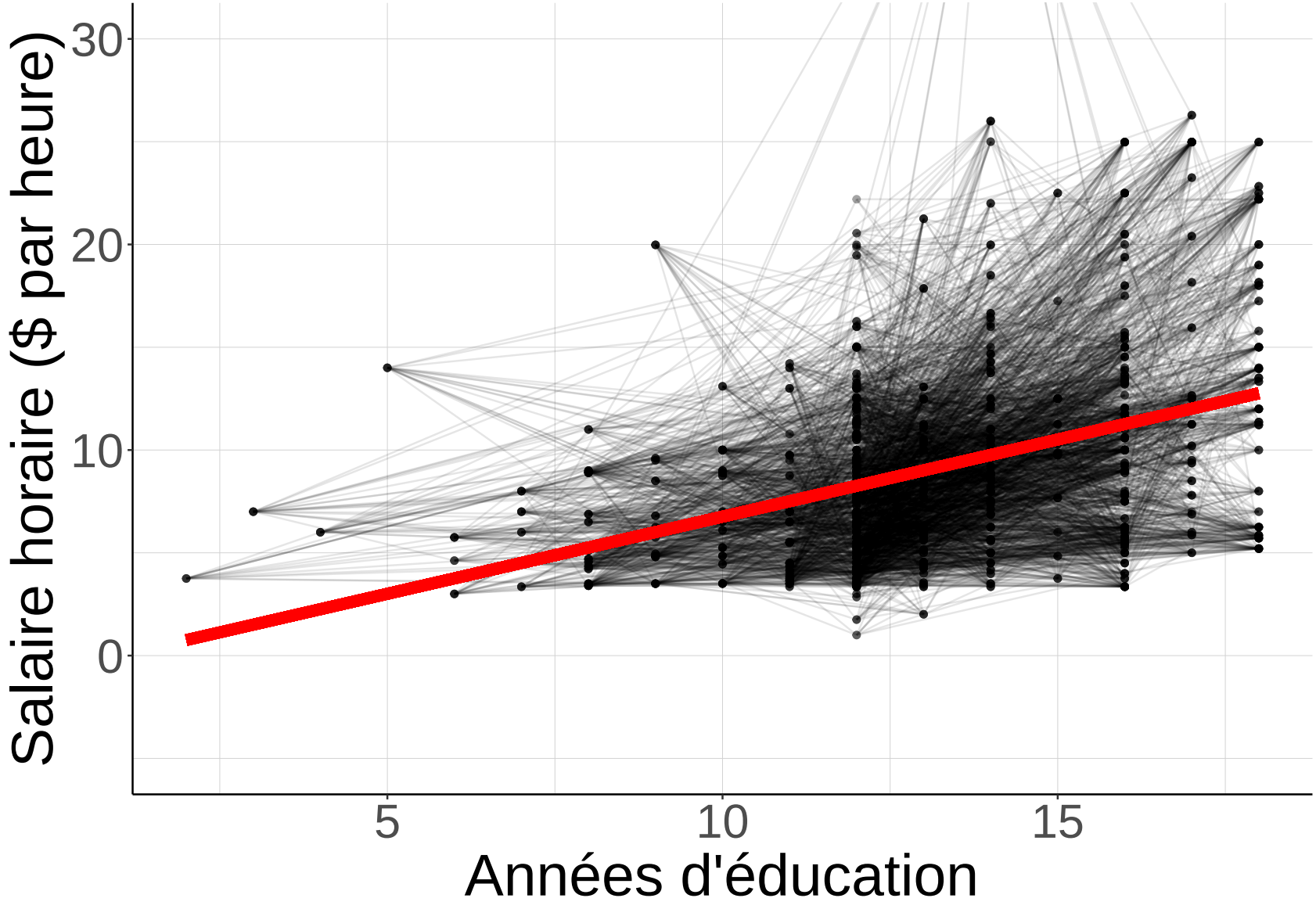
Pour visualiser



Pour visualiser



Pour visualiser



Comparaisons avec l'espérance conditionnelle

Régression et espérance conditionnelle

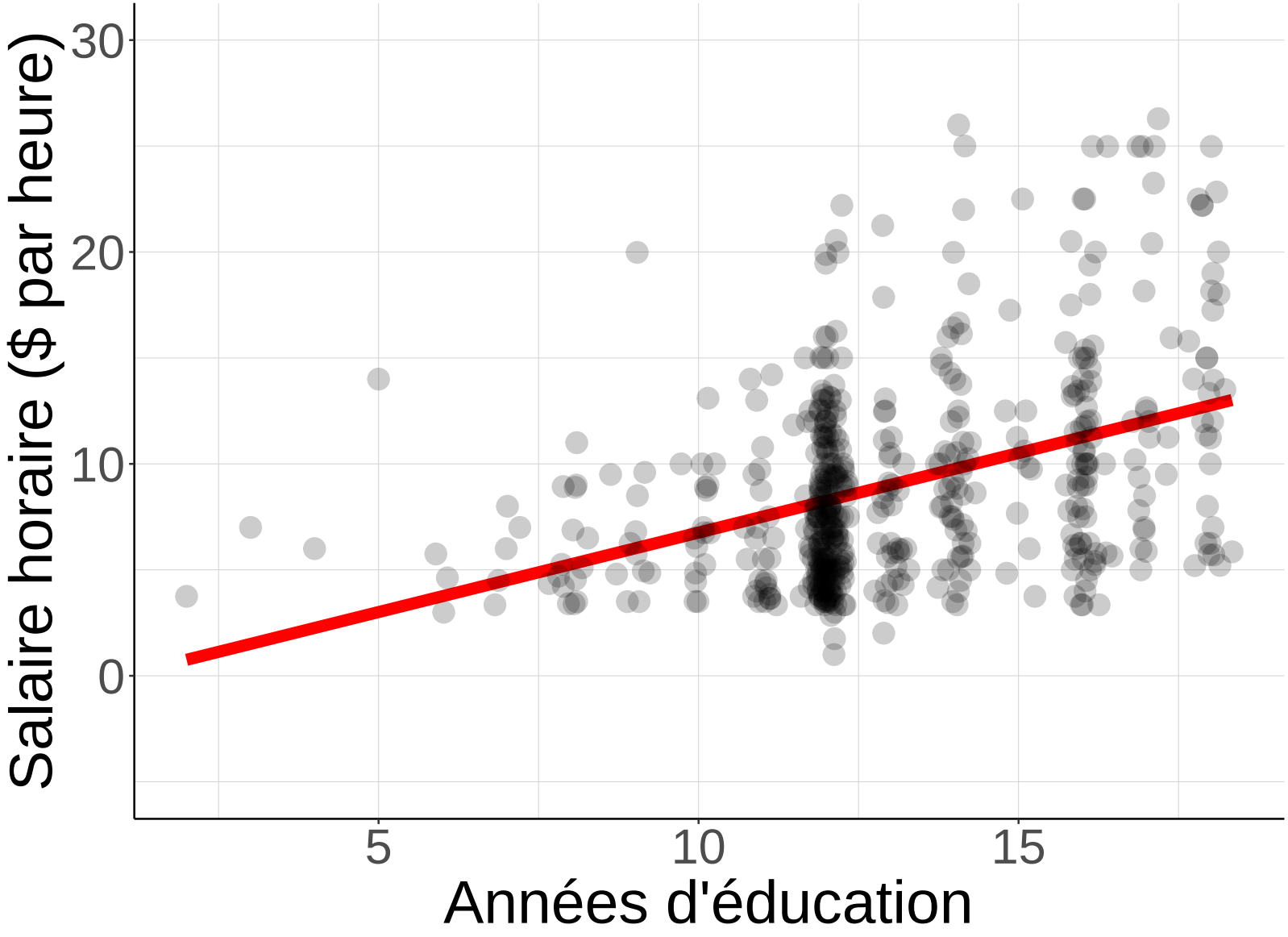
- Les coefficients ne dépendent que de l'**espérance conditionnelle**
  - C'est-à-dire ici du salaire moyen pour chaque niveau d'éducation
  - (et de la distribution de l'éducation dans la population !)
- Facile à montrer en revenant à l'expression de β en fonction de la covariance
  - Et en revenant à la définition de la covariance

Une petite vérification (encore)

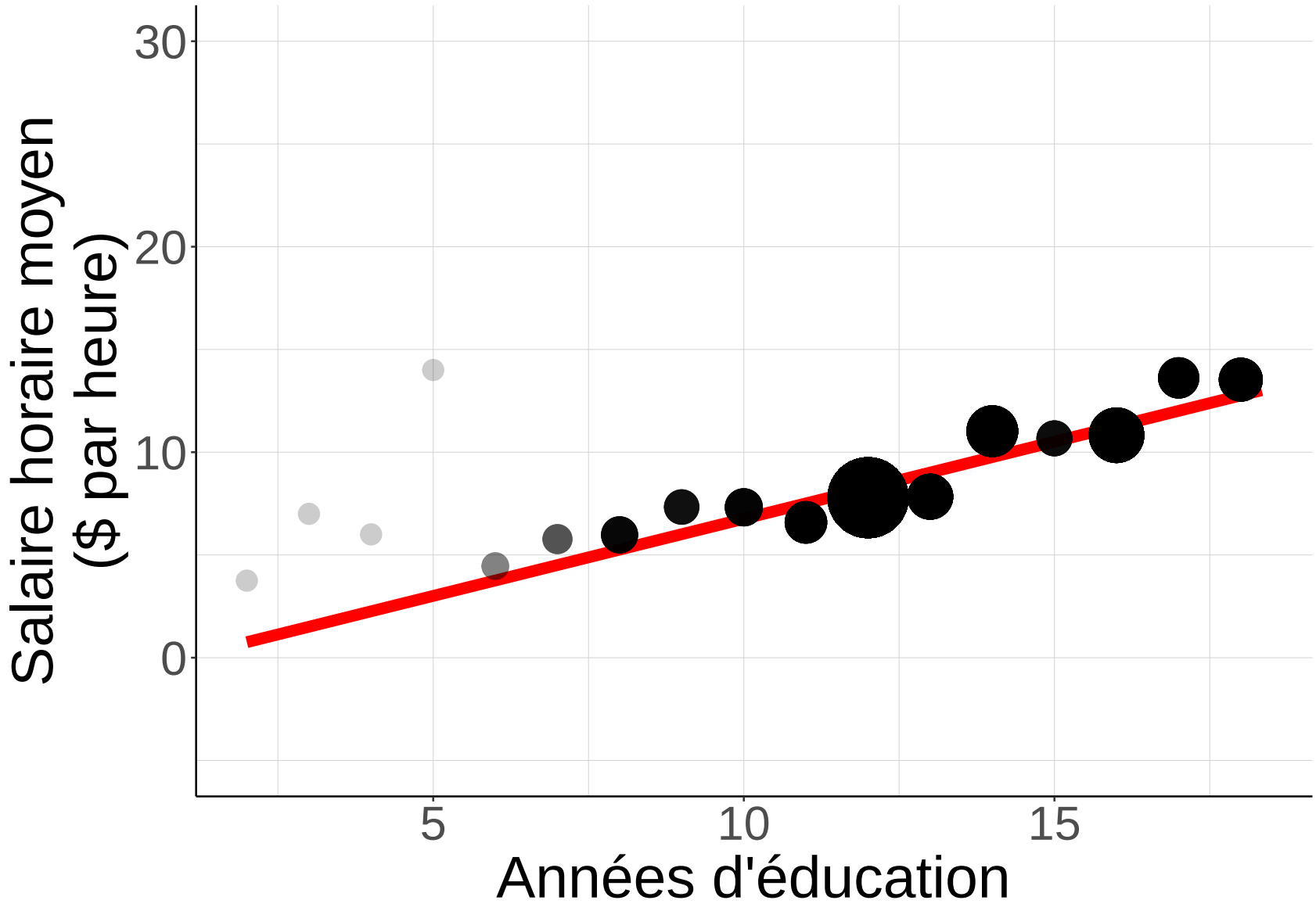
```
CPS1985 <- data.table(CPS1985)

CPS1985[,average_wage := mean(wage),
        by = c("education")]

regression_salaire_moyen <-
  lm(formula = average_wage ~ education,
     data = CPS1985)

all.equal(regression_salaire_moyen$coefficients,
          regression$coefficients)
```

```
[1] TRUE
```

Pour visualiser



Pour visualiser



Une interprétation (légèrement) plus poussée

- On peut montrer que β peut s'écrire comme une **moyenne de différences de salaire moyen entre niveaux d'éducation adjacents**
- Les poids ne sont pas particulièrement faciles à interpréter
- Et la preuve est un peu calculatoire
- Mais c'est rassurant sur l'interprétation des coefficients

Un cas presque trivial (mais important en pratique !)

- Si on prend comme régresseur une variable indicatrice d'appartenir à une certaine sous-population
  - Exemple : variable qui vaut 1 pour les femmes 0 pour les hommes

```
CPS1985[,female := as.numeric(gender == "female")]
```

- Quel coefficient récupère-t-on ?

Un cas presque trivial (mais important en pratique !)

- Le coefficient est seulement la **différence entre cette sous-population et le reste de la population**
  - Dans l'exemple : différence de salaire entre femmes et hommes

Un cas presque trivial (mais important en pratique !)

```
regression_femmes_hommes <-
  lm(wage ~ female,
     data = CPS1985)

regression_femmes_hommes$coefficients["female"]
```

```
   female
-2.116056
```

```
all.equal(
  as.numeric(regression_femmes_hommes$coefficients["female"
  mean(CPS1985[female == 1]$wage) -
    mean(CPS1985[female == 0]$wage))
```

```
[1] TRUE
```

Un problème d'optimisation

Une dernière interprétation

- Choisir la **droite la plus proche du nuage de points**
- Bonne idée mais comment **définir la distance** ?
- Une solution possible :
  - **Le carré de la différence entre le salaire réalisé et le salaire prédit** par la droite au même niveau d'éducation (ou sa racine carrée)
  - Attention : ce n'est pas (le carré de) la distance euclidienne d'un point à la droite
  - Une autre possibilité existe : prendre la valeur absolue de cette distance
    - Définit une autre droite $\rightarrow$ régression quantile au niveau de la médiane $\rightarrow$ hors de nos considérations ce semestre

Une dernière interprétation

- Parmi toutes les droites qui permettent de définir la décomposition $\widehat{\texttt{wage}} + \epsilon$
  - Sans contrainte sur ϵ !
- On veut donc regarder celle qui est la plus proche du nuage de points en ce sens-là
- Cela revient à minimiser $\mathbb{E}[\epsilon^2]$
- C'est cette vision comme un **problème d'optimisation** qui justifie le nom de **"moindres carrés ordinaires"**

Une dernière interprétation

- Cette minimisation implique $\mathbb{E}[\epsilon] = 0$
  - Sinon on peut toujours considérer $\epsilon_0 = \epsilon - \mathbb{E}[\epsilon]$
    - Evidemment de moyenne nulle
  - $\mathbb{E}[\epsilon_0^2] = \mathcal{V}(\epsilon) \leq \mathbb{E}[\epsilon^2]$

Une dernière interprétation

- Cette minimisation implique aussi $\mathcal{C}(\epsilon, \texttt{education}) = 0$
- Pour le voir, en notant ϵ_0 le résidu issu de la contrainte $\mathcal{C}(\epsilon_0, \texttt{education}) = 0$
- $\mathbb{E}[\epsilon^2] = \mathbb{E}[\{\epsilon_0 + \epsilon - \epsilon_0\}^2] = \mathbb{E}[\epsilon_0^2] + \mathbb{E}[\{\epsilon - \epsilon_0\}^2] + 2\mathbb{E}[\epsilon_0\{\epsilon - \epsilon_0\}]$
  - Mais $\epsilon - \epsilon_0$ est une fonction affine de `education`
  - Donc la corrélation avec ϵ_0 est nulle
- *In fine* $\mathbb{E}[\epsilon^2] = \mathbb{E}[\epsilon_0^2] + \mathbb{E}[\{\epsilon - \epsilon_0\}^2] \geq \mathbb{E}[\epsilon_0^2]$
- Et l'égalité est réalisée quand $\epsilon = \epsilon_0$

Un regard géométrique

Inteprétation géométrique et lien avec l'espérance conditionnelle (encore !)

- Pour une façon pédante de le dire :
  - $\widehat{\texttt{wage}}$ est le **projeté orthogonal** de `wage` sur le sous-espace des fonctions affines de `education`
    - Pour le produit scalaire $(X \mid Y) = \mathbb{E}[XY]$
  - C'est cette structure préhilbertienne qui assure le passage entre la **condition d'orthogonalité** $\mathcal{C}(\epsilon, \texttt{education}) = 0$ et la **minimisation de la distance**

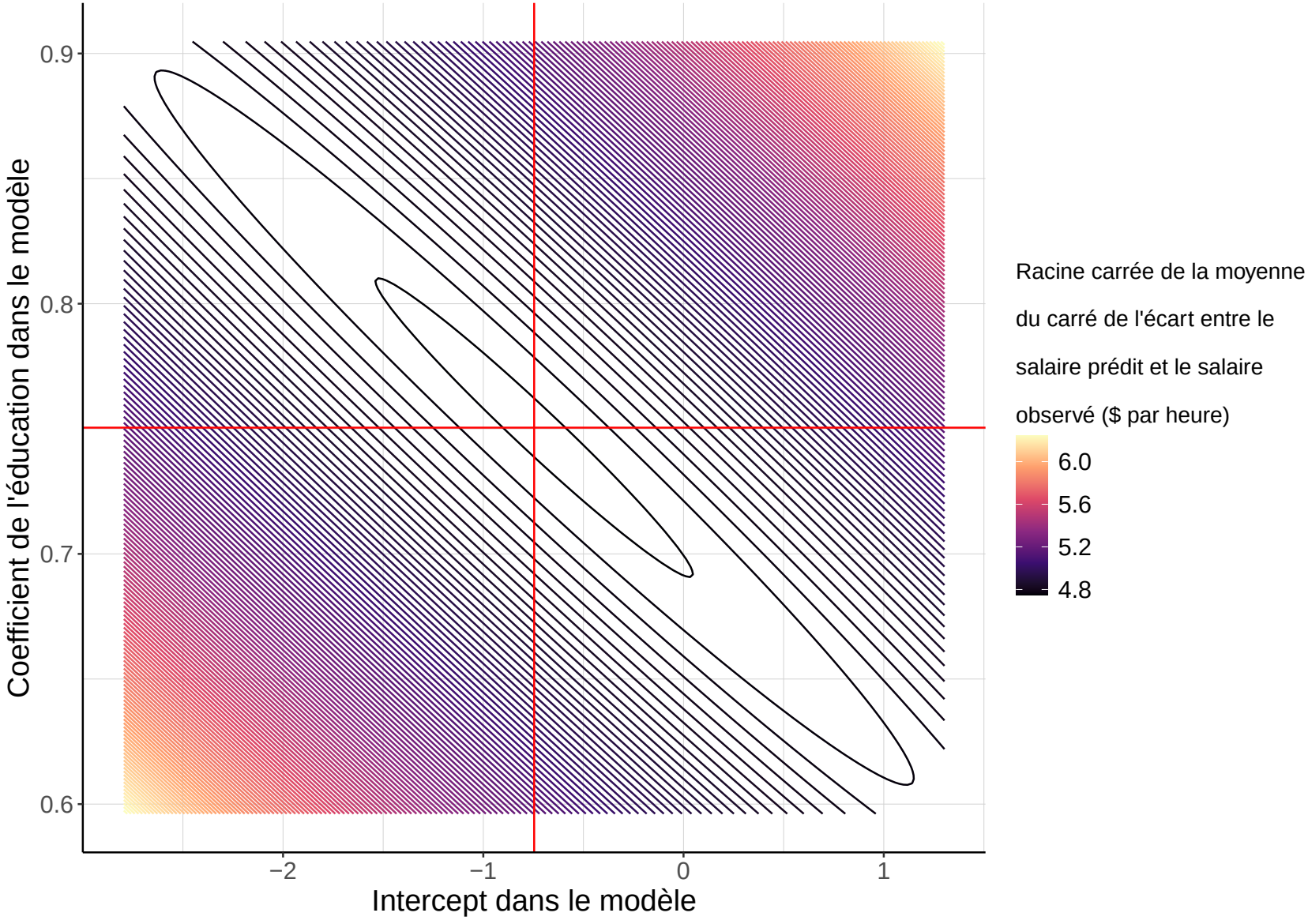
Interprétation géométrique et lien avec l'espérance conditionnelle (encore !)

- En un sens l'**espérance conditionnelle** fonctionne presque pareil !
  - C'est le projeté orthogonal de `wage` sur le sous-espace des fonctions (mesurables) de `education`
  - Pour la même raison, c'est le produit de la minimisation de $\mathbb{E}[\epsilon^2]$ sous la contrainte $\texttt{wage} = f(\texttt{education}) + \epsilon$

Interprétation géométrique et lien avec l'espérance conditionnelle (encore !)

- On peut **itérer les projections** :
  - Les fonctions affines de `education` sont en particulier des fonctions mesurables de `education`
  - Projeter `wage` pour récupérer $\mathbb{E}[\texttt{wage} \mid \texttt{education}]$ puis projeter $\mathbb{E}[\texttt{wage} \mid \texttt{education}]$ sur le sous-espace des fonctions affines
  - C'est pareil que projeter directement `wage` sur le sous-espace des fonctions affines pour récupérer $\widehat{\texttt{wage}}$
- L'équivalent de ces opérations dans l'espace à 3 dimensions :
  - Le projeté orthogonal du point (x, y, z) sur l'axe des abcisses est $(x, 0, 0)$
  - C'est aussi le projeté sur cette droite du projeté orthogonal du point initial (x, y, z) sur le plan engendré par l'axe des abcisses et l'axe des ordonnées $(x, y, 0)$

Une petite vérification



Prolonger l'interprétation géométrique : coefficient de détermination

Une petite conséquence : décomposition de la variance

- Par définition $\mathcal{V}(\texttt{wage}) = \mathcal{V}(\widehat{\texttt{wage}} + \epsilon)$
- Mais $\mathcal{C}(\epsilon, \texttt{education}) = 0$ et $\widehat{\texttt{wage}}$ est une fonction affine de `education`
- Donc $\mathcal{V}(\texttt{wage}) = \mathcal{V}(\widehat{\texttt{wage}}) + \mathcal{V}(\epsilon)$
- Ce n'est qu'une version du **théorème de Pythagore** !

Une petite vérification

```
#On vérifie que la variance du salaire est bien égale
# à la somme de la variance du salaire prédit et de
# la variance des résidus
all.equal(var(CPS1985$wage),
          var(regression$fitted.values) +
            var(regression$residuals))
```

```
[1] TRUE
```

Coefficient de détermination

- Permet de motiver la définition du **coefficient de détermination** $R^2 = \frac{\mathcal{V}(\widehat{\texttt{wage}})}{\mathcal{V}(\texttt{wage})}$
  - S'interprète comme une part, comprise entre 0 et 1

- On peut dire que c'est la **part de la variance de** `wage` expliquée par `education`
  - A condition de se souvenir que ce concept d'explication n'est pas causal !

  - Sauf si bien sûr on est dans un cas où on défend l'interprétation causale de la régression
    - On pourrait plutôt parler de variance *inter* et *intra*

- Dans le cas unidimensionnel, $R^2 = \rho^2_{\texttt{wage},\texttt{education}}$

Une dernière vérification

```
#On vérifie que le R² estimé par R est bien le rapport
# de la variance du salaire prédit et de la variance
# du salaire observé
all.equal(summary(regression)$r.squared,
          var(regression$fitted.values)/
            var(CPS1985$wage))
```

```
[1] TRUE
```

```
#On vérifie que le coefficient de détermination est
# bien égal au carré de la corrélation entre le salaire
# et l'éducation
all.equal(summary(regression)$r.squared,
          cor(CPS1985$wage,
              CPS1985$education)^2)
```

```
[1] TRUE
```

Une interprétation courante du coefficient de détermination

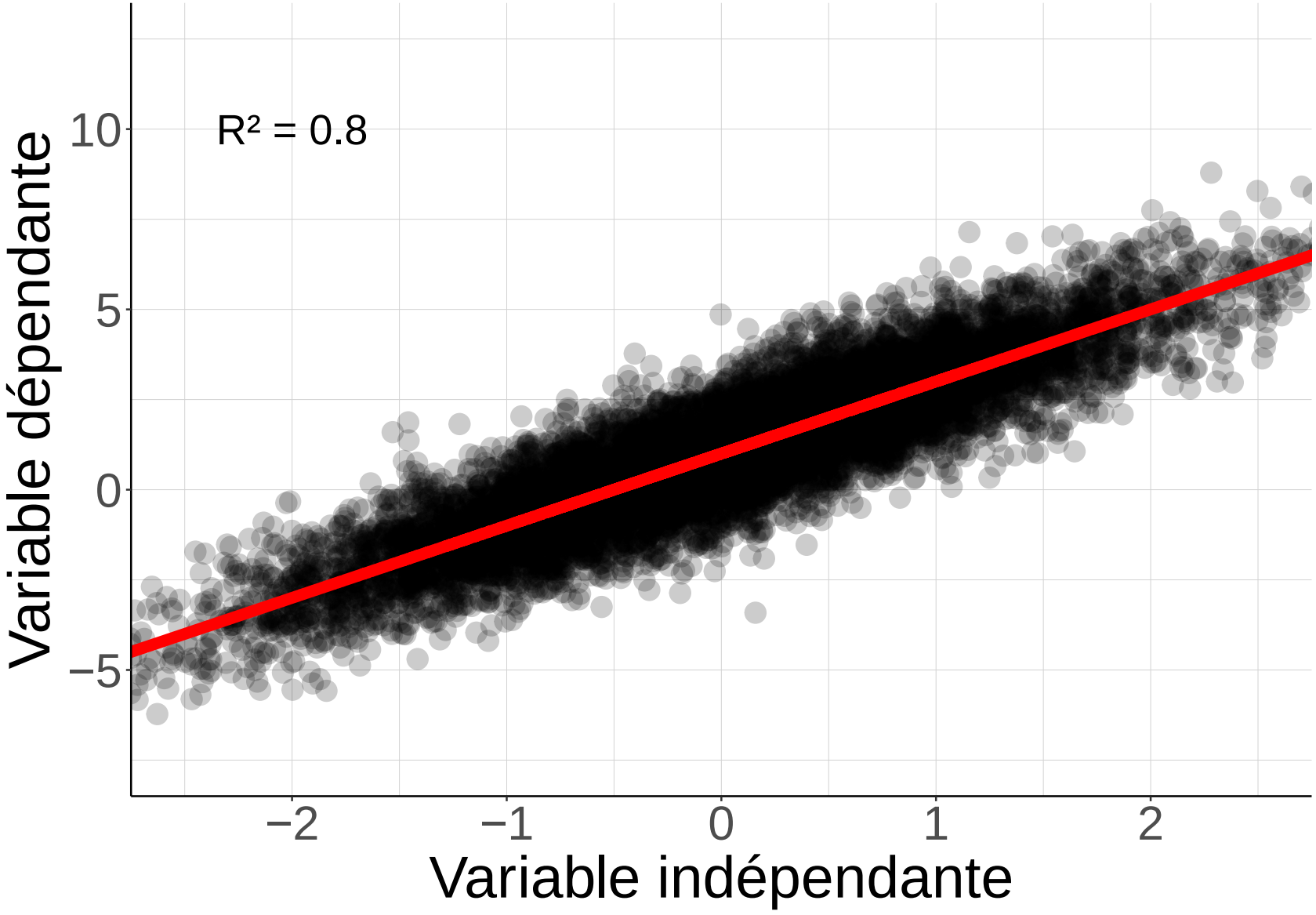
- Quantifie la **“qualité” du modèle de régression**
  - Attention aux ambiguités sur ce que l’on entend par qualité !

- Ici la “qualité” porte sur la *prédiction*
  - La prédiction est de bonne qualité si elle est souvent proche de la vraie valeur

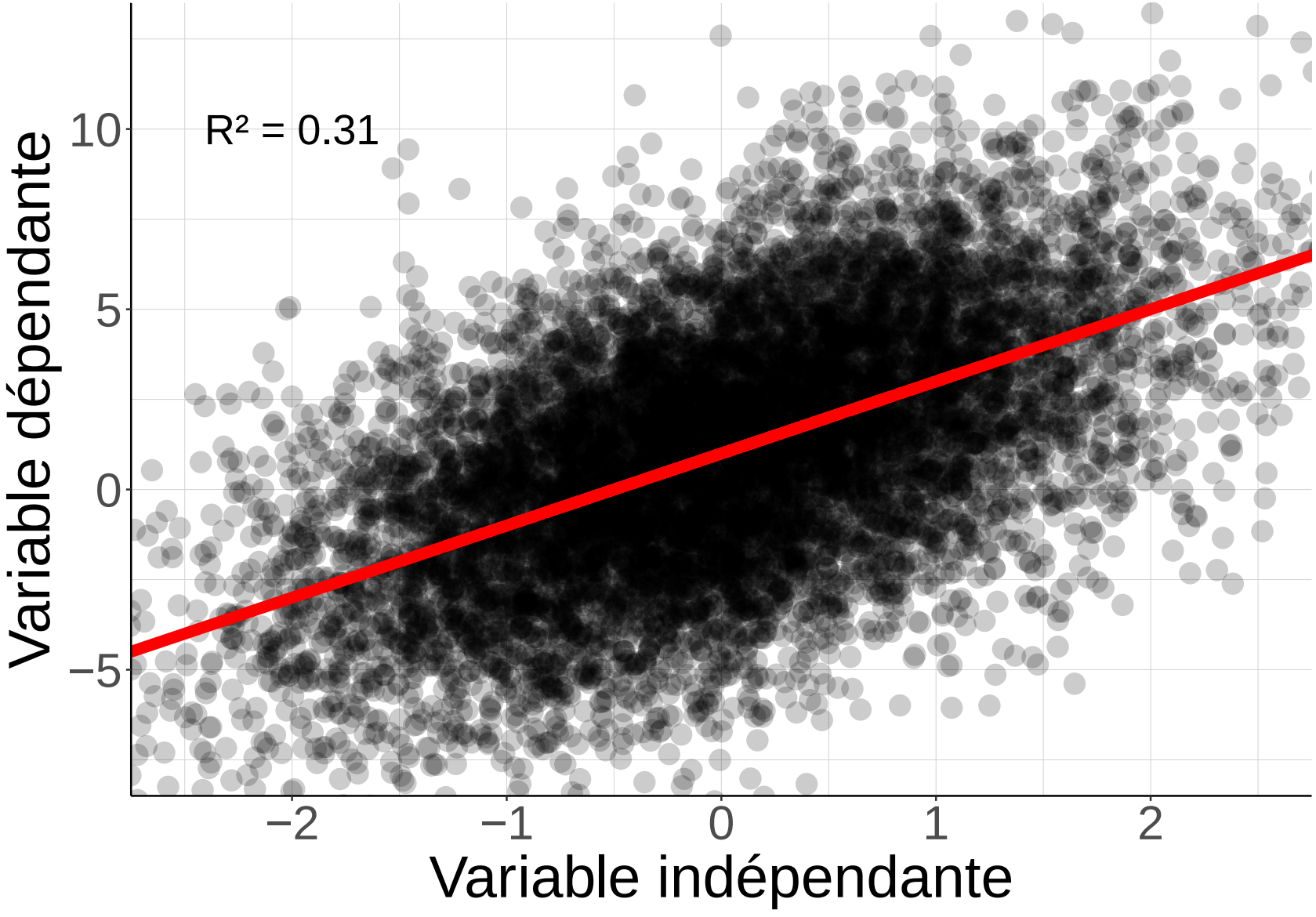
  - **Cela n’a rien à voir avec la qualité des coefficients**
    - La qualité des coefficients c’est d’être estimé précisément
      $\rightarrow$ l’objet des séances suivantes

    - Ou à la limite d’avoir une bonne interprétation (causale)
      $\rightarrow$ dépend du contexte, et surtout le R^2 qui est une mesure de corrélation ne répondra jamais à la question !

Une interprétation courante du coefficient de détermination



Une interprétation courante du coefficient de détermination



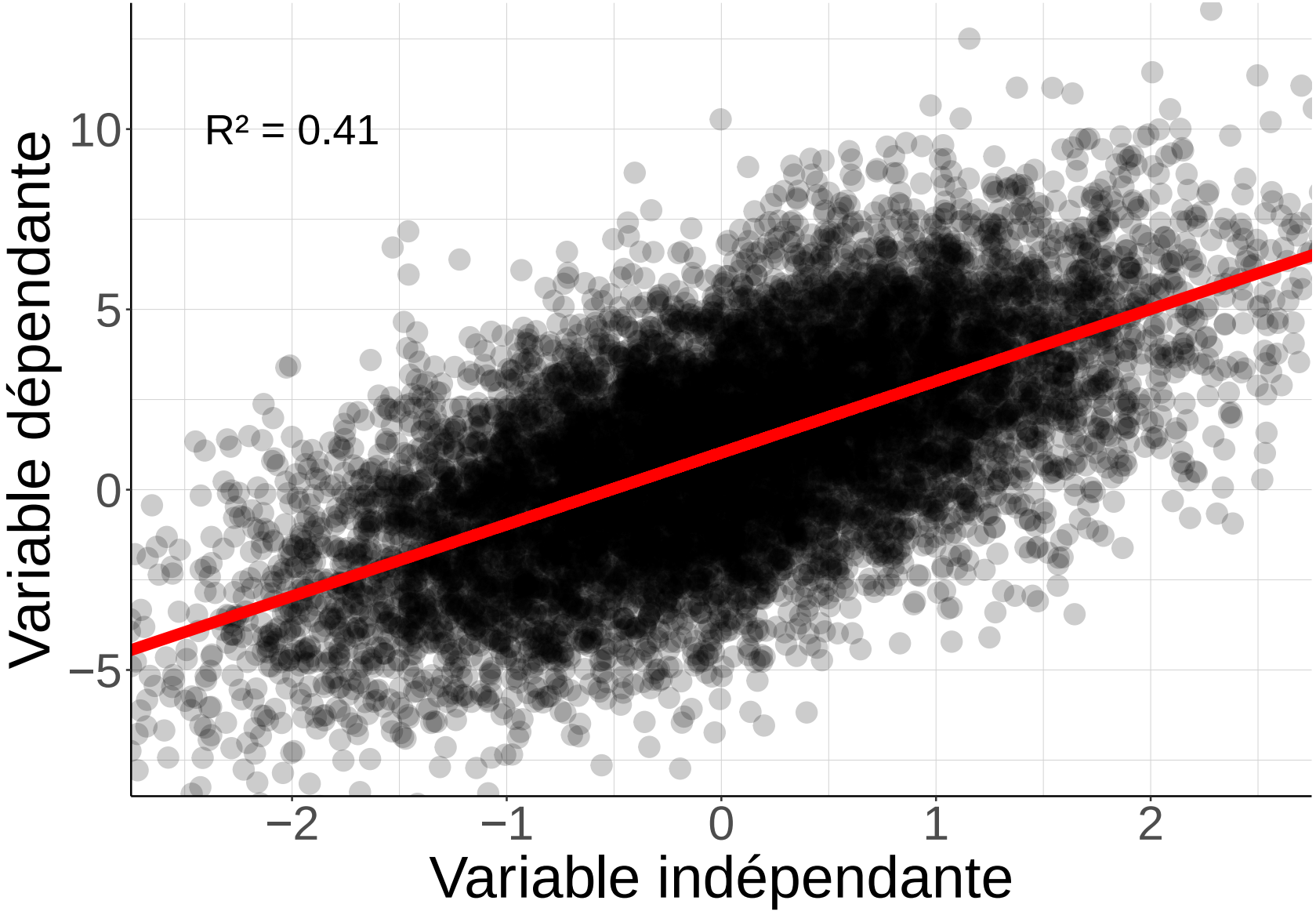
Difficultés de l'interprétation

- Retour à l'interprétation géométrique : R^2 quantifie la qualité de l'approximation de la variable dépendante par la valeur prédite
  - La valeur prédite par la régression linéaire est la meilleure parmi celles qui s'écrivent comme fonction affine de la variable indépendante
  - L'espérance conditionnelle est la meilleure parmi celles qui s'écrivent comme fonction (mesurable) de la variable indépendante

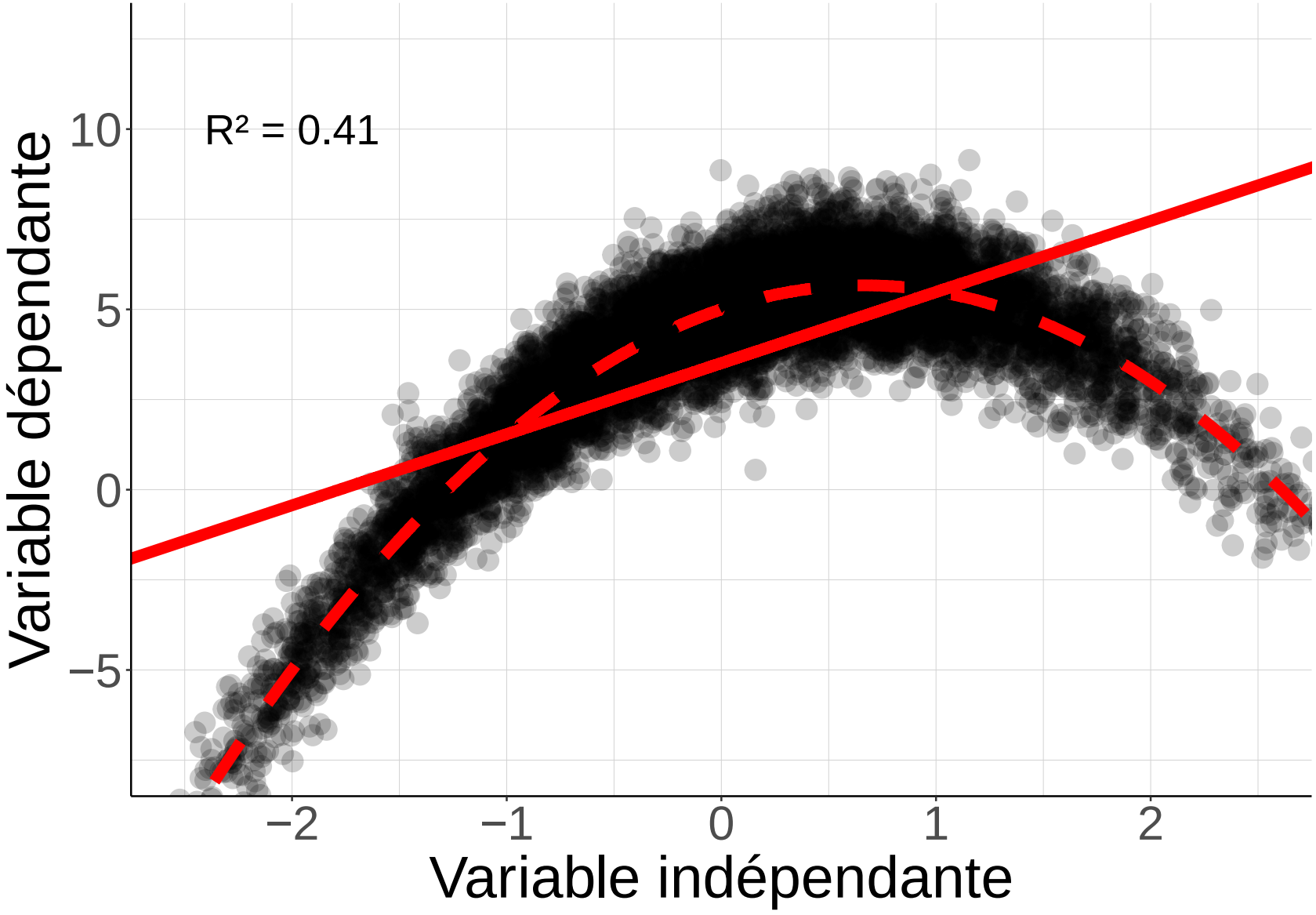
Difficultés de l'interprétation

- On a donc **deux cas de figures possibles** pour expliquer un faible R^2 :
  - La valeur prédite par la régression linéaire et l'espérance conditionnelle sont très proches
    - Les vraies valeurs sont très dispersées à un niveau donné de la variable indépendante
    - On ne pourra jamais faire mieux en continuant à regarder des fonctions de la variable dépendante !
  - La valeur prédite par la régression linéaire et l'espérance conditionnelle diffèrent beaucoup
    - Les vraies valeurs ne sont pas forcément très dispersées à un niveau donné de la variable indépendante, mais l'espérance conditionnelle peut être assez loin de la prédiction linéaire
    - On peut faire mieux en regardant des fonctions plus flexibles de la variable dépendante

Difficultés de l'interprétation



Difficultés de l'interprétation



Pour finir

Conclusion

- On a un **outil pour passer, dans le cas de deux variables, d'un nuage de points à une droite**
- Plein d'interprétations possibles de la construction de cette droite, équivalentes les unes aux autres
  - Condition d'orthogonalité
  - Minimisation de la distance au nuage
  - Comparaisons deux à deux
  - Petites incrémentations du régresseur
- Au total il n'y a rien d'autre que des **comparaisons de moyennes** !

La suite

- Considérer l'**extension à plusieurs régresseurs**
- Quels sont les résultats qui se transfèrent, quels sont les résultats qui ne passent pas ?
- Comment construire de façon simple et intuitive les **régressions multiples** ?