Séance 5 – Régression linéaire multiple

Pierre Pora

Rappel des séances précédentes

- **Régression linéaire simple** : une façon de lier une variable à une autre
- Plein d'interprétations possibles
  - Meilleure approximation linéaire
  - Comparaisons deux à deux
  - Comparaisons de moyennes
  - Décomposition en composantes orthogonales

L'objet de la séance

- Est-il possible d'**étendre cette idée à plusieurs variables** ?
- Quelles sont les interprétations que l'on peut conserver dans ce cas ?
- Quel est le lien entre **régression simple** et **régression multiple** ?

L'objet de la séance

- Encore une séance vraiment importante
- Et potentiellement assez dense
  - Si nécessaire on étendra sur la séance suivante
- Avec les résultats de la séance précédente + ceux-là, vous devez être capables d'**interpréter sans difficulté les coefficients** de n'importe quelle régression par les MCO
- En dehors évidemment de leur interprétation causale
- Mais vous saurez qui est comparé à qui

L'objet de la séance

- Comme la séance précédente, on s'abstrait complètement de la distinction entre les **quantités estimées** (relatives à la population) et leur **estimation** à partir d'un échantillon de taille finie
- On fait tout le temps comme si on travaillait dans la population entière et que cela ne posait aucune difficulté
- On verra comment intervient l'échantillonnage, et surtout comment **quantifier l'incertitude** la séance suivante

Toujours les mêmes données...

```
library(AER)

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

Le concept existe visiblement pour R

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

regression_multiple <- lm(wage ~ female + education,
                          data = CPS1985)

regression_multiple$coefficients
```

```
(Intercept)      female   education
  0.2178312  -2.1240567   0.7512834
```

Le concept existe visiblement pour R

- R estime trois coefficients
- Mais que représentent-ils exactement ?
- Comment sont-ils construits ?
- Comment les interpréter ?

Toujours la même idée

- Il existe un unique vecteur β et une unique variable aléatoire réelle ϵ tels que :
  - $\texttt{wage} = X'\beta + \epsilon$
    - avec X v.a. de dimension 3, $X = (1 \ \texttt{female} \ \texttt{education})'$
  - $\mathbb{E}[X\epsilon] = 0$

Clarification du formalisme

- **Petit calcul matriciel** :
  $X'\beta = \beta_1 + \beta_2 \texttt{female} + \beta_3 \texttt{education}$
- La première composante de $X\epsilon$ est... ϵ
  - Donc $\mathbb{E}[X\epsilon] = 0 \Rightarrow \mathbb{E}[\epsilon] = 0$
- Les autres composantes de $X\epsilon$ sont les produits $\texttt{female}\,\epsilon$ et $\texttt{education}\,\epsilon$
  - On sait déjà que $\mathbb{E}[\epsilon] = 0$
  - Donc $\mathbb{E}[X\epsilon] = 0 \Rightarrow \mathcal{C}(\texttt{female}, \epsilon) = 0$ et $\mathcal{C}(\texttt{education}, \epsilon) = 0$

Une petite vérification s'impose

```
CPS1985[,
        salaire_predit :=
          regression_multiple$coefficients[
            "(Intercept)"] +
          regression_multiple$coefficients[
            "female"] *
          female +
          regression_multiple$coefficients[
            "education"] *
          education]

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

```
[1] TRUE
```

Une petite vérification s'impose

```
CPS1985[,
        residu :=
          wage - salaire_predit]

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

```
[1] TRUE
```

Une petite vérification s'impose

```
all.equal(mean(CPS1985$residu),
          0)
```

```
[1] TRUE
```

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

```
[1] TRUE
```

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

```
[1] TRUE
```

Comment ça marche ?

- Petit détour un petit peu plus mathématisé
- Supposons qu'on a dispose d'un tel β
  - Par linéarité de l'espérance $\mathbb{E}[X\epsilon] = \mathbb{E}[X\texttt{wage}] - \mathbb{E}[XX']\beta$
  - Donc $\mathbb{E}[XX']\beta = \mathbb{E}[X\texttt{wage}]$
  - Si $\mathbb{E}[XX']$ est inversible on est bons !
    - $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[X\texttt{wage}]$
- Si $\mathbb{E}[XX']$ est **inversible**, alors on peut montrer qu'un tel β et le résidu qui s'en déduisent conviennent

Une condition nécessaire

- Tout ça ne marche que sous la **condition d'inversibilité** de $\mathbb{E}[XX']$!
  - On parle parfois de **condition de rang** (c'est la même chose)
  - Qu'est-ce que ça veut dire en pratique ?

Problème de colinéarité

- La condition d'inversibilité équivaut à dire que toutes les composantes de X sont **linéairement indépendantes** (au sens de l'algèbre linéaire !)
  - Equivaut aussi à l'inversibilité de la matrice de variance-covariance de $(\texttt{female education})'$
  - Il n'y a pas de combinaison linéaire constante de `female` et `education`
  - Pas très compliqué à montrer à la main

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

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

- Il y a bien une combinaison linéaire constante de `male` et `female`

```
all.equal(
  CPS1985$male + CPS1985$female,
  rep(1, times = nrow(CPS1985))
)
```

```
[1] TRUE
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

```
matrice_XXprime <-
  (1 / nrow(CPS1985)) *
  t(cbind(rep(1, nrow(CPS1985)),
          CPS1985$female,
          CPS1985$male)) %*%
  cbind(rep(1,
            nrow(CPS1985)),
        CPS1985$female,
        CPS1985$male)
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

```
matrice_XXprime
```

```
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.4588015 0.5411985
[2,] 0.4588015 0.4588015 0.0000000
[3,] 0.5411985 0.0000000 0.5411985
```

- La première colonne (ligne) est simplement la somme des deux autres !
- Le déterminant est nul

```
all.equal(det(matrice_XXprime),
          0)
```

```
[1] TRUE
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

- Deux façons *a priori* différentes d'approximer le salaire comme une fonction affine de `male` et `female`
  - Correspondent à deux vecteurs de coefficients différents !
    - Dans le premier cas $(0 \ \texttt{sal\_moy\_f} \ \texttt{sal\_moy\_m})'$
    - Dans le second $(\texttt{sal\_moy\_f} \ 0 \ \texttt{sal\_moy\_m} - \texttt{sal\_moy\_f})'$

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

```
sal_moy_f <- mean(CPS1985[female == 1]$wage)
sal_moy_m <- mean(CPS1985[male == 1]$wage)

CPS1985[,
        c("salaire_predit_mf_1",
          "salaire_predit_mf_2") :=
          list(sal_moy_f * female +
                 sal_moy_m * male,
               sal_moy_f +
                 male * (sal_moy_m - sal_moy_f))]
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

- En fait la valeur prédite du salaire est exactement la même dans les deux cas !

```
all.equal(as.numeric(CPS1985$salaire_predit_mf_1),
          as.numeric(CPS1985$salaire_predit_mf_2))
```

```
[1] TRUE
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

- On vérifie que ce sont bien deux constructions admissibles : conditon d'orthogonalité

```
CPS1985[,residu_mf := wage - salaire_predit_mf_1]

CPS1985[,
        lapply(X = .SD,
               FUN = function(var)
                 mean(var * residu_mf)),
        .SDcols = c("male",
                    "female")]
```

```
             male          female
            <num>           <num>
1: -7.843614e-17 -8.172067e-17
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

- La condition d'orthogonalité suffit toujours à définir le résidu
  - Et donc aussi la valeur prédite
  - Ce n'est pas là qu'est le problème
- Le sous-espace engendré par la v.a. constante, `female` et `male` est de dimension 2 et pas 3
  - On ne peut pas définir de façon unique 3 coefficients solutions du problème

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

- Comment R gère-t-il le problème ?

```
regression_mf <-
  lm(wage ~ male + female,
     data = CPS1985)

regression_mf$coefficients
```

```
(Intercept)        male      female
   7.878857    2.116056          NA
```

Problème de colinéarité : un exemple trivial (mais utile pour comprendre)

- Une valeur manquante qui permet de revenir à un problème dont la solution est unique
  - C'est une solution au problème
    - Plus généralement, il faut **rajouter une contrainte linéaire** sur les coefficients
  - Si l'intérêt porte sur les coefficients plutôt que sur les valeurs prédites, il est préférable de le gérer soi-même pour choisir le(s) coefficient(s) manquant(s) et avoir l'interprétation que l'on souhaite
  - Ou la contrainte pertinente

Une première interprétation

- Au vu de ce qu'on a dit, en revenant sur l'exemple de départ
  - $\widehat{\texttt{wage}} = X'\beta$ est le **projeté orthogonal** de `wage` sur le sous-espace engendré par les régresseurs
  - β correspond à l'écriture de $\widehat{\texttt{wage}}$ comme combinaison linéaire des régresseurs, vus comme une **base de ce sous-espace**
    - La condition d'inversibilité de $\mathbb{E}[XX']$ dit seulement que **le nombre de régresseurs doit être égal à la dimension de ce sous-espace**
    - Remarque : cette base n'est pas nécessairement orthogonale, et *a fortiori* orthonormée !!
    - Ce n'est le cas que lorsque la corrélation entre variables indépendantes est nulle
- Interprétation qui peut paraître ésotérique mais est la plus générale

Une seconde interprétation

- Comme pour le cas de la régression simple, la structure préhilbertienne fournit une interprétation équivalente
- $\widehat{\texttt{wage}} = X'\beta$ est la façon d'approximer wage qui minimise la distance quadratique $\sqrt{\mathbb{E}[\epsilon^2]}$
  - Si on considère une façon alternative de construire l'approximation :
    - $\mathbb{E}[\tilde{\epsilon}^2] = \mathbb{E}[\{\epsilon + \tilde{\epsilon} - \epsilon\}^2]$
    - Donc $\mathbb{E}[\tilde{\epsilon}^2] = \mathbb{E}[\epsilon^2] + \mathbb{E}[\{\tilde{\epsilon} - \epsilon\}^2] + 2\mathbb{E}[\epsilon\{\tilde{\epsilon} - \epsilon\}]$
    - Mais $\tilde{\epsilon} - \epsilon$ est linéaire en X donc le dernier terme est nul
    - *In fine* $\mathbb{E}[\tilde{\epsilon}^2] = \mathbb{E}[\epsilon^2] + \mathbb{E}[\{\tilde{\epsilon} - \epsilon\}^2] \geq \mathbb{E}[\epsilon^2]$

Une conséquence

- Comme pour la régression linéaire, cela fournit déjà une comparaison avec l'espérance conditionnelle $\mathbb{E}[\texttt{wage} \mid X]$
- $\widehat{\texttt{wage}}$ est la **meilleure approximation de** $\texttt{wage}$ **qui s'écrive comme une forme linéaire de** X
- $\mathbb{E}[\texttt{wage} \mid X]$ est la **meilleure approximation de** $\texttt{wage}$ **qui s'écrive comme une fonction (mesurable, pas forcément linéaire) de** X
- Ici on a bien "meilleure" dans le même sens !
  - Distance en norme quadratique ou projection orthogonale (c'est équivalent)

Une conséquence

- Comme on parle de “meilleure” dans le même sens
- Et comme les formes linéaires sont des fonctions parmi d’autres
- Dans le “meilleur” des cas $\widehat{\texttt{wage}} = \mathbb{E}[\texttt{wage} \mid X]$
- Et **la régression linéaire ne dépend que de l’espérance conditionnelle**
  - Découle du fait que c’est pareil de projeter orthogonalement wage sur le sous-espace des fonctions mesurables de X, puis de projeter le projeté sur le sous-espace des formes linéaires de X qui est inclus dans le second
  - On pouvait déjà voir ça simplement en voyant que ça ne dépend que de $\mathbb{E}[X\texttt{wage}]$

Conclusion partielle

- Ce dont on dispose à ce stade
  - Une définition du problème
  - Une solution qui demande (en gros) de savoir inverser une matrice pour calculer les coefficients
  - Deux **interprétation géométriques** : $\widehat{\texttt{wage}}$ comme "bonne" approximation de `wage` par une combinaison linéaire de régresseurs
- La suite :
  - Construire des interprétation plus simples des coefficients
    - Qui marchent dans certains cas
  - Regarder une façon pratique de construire les coefficients en partant d'une régression simple

Un cas trivial (mais important)

- Comment gérer un régresseur qui est une variable qualitative ?
  - Par exemple le secteur `sector`

```
table(CPS1985$sector)
```

```
manufacturing  construction         other
           99            24           411
```

Un cas trivial (mais important)

- On définit une variable indicatrice pour chaque niveau possible de la variable qualitative
- Quelle difficulté cela pose-t-il ?

Un cas trivial (mais important)

- Problème de colinéarité
  - Il faut omettre un niveau
  - ou bien la constante
- Les valeurs prédites s'identifient à l'**espérance conditionnelle**
- Les coefficients se lisent comme les **différences entre le le salaire moyen du secteur concerné et le secteur omis**
  - Et l'intercept est le salaire de ce secteur omis
- Ou bien simplement les salaires moyens si c'est la constante qu'on a omise
- On parle de régression saturée

Une petite vérification (toujours)

```
regression_sect <- lm(wage ~ sector,
                      data = CPS1985)

t(regression_sect$coefficients)
```

```
     (Intercept) sectorconstruction sectorother
[1,]    9.604444          -0.3836111  -0.7316707
```

Une petite vérification (toujours)

```
sal_moy <- CPS1985[,
                   list(sal_moy = mean(wage)),
                   by = c("sector")]

ecarts_moyens <-
  sal_moy[,
          lapply(X = levels(CPS1985$sector),
                 FUN = function(sect)
                   sum(sal_moy *
                         (as.numeric(sector ==
                                       sect) -
                            as.numeric(
                              sector ==
                                "manufacturing"))))]
```

Une petite vérification (toujours)

```
all.equal(
  as.numeric(regression_sect$coefficients[
    names(regression_sect$coefficients) !=
      "(Intercept)"]),
  as.numeric(ecarts_moyens)[2:3])
```

```
[1] TRUE
```

Une remarque rapide sur la régression saturée

- Elle n'apparaît pas toujours explicitement sous la forme d'une variable qualitative avec tous ses niveaux
- On peut aussi utiliser des interactions
  - Les coefficients se lisent comme des différences entre différences etc. (selon le niveau de l'interaction)

Un exemple rapide

```
CPS1985[,
        c("college",
          "female"):=
          list(as.numeric(education>=16),
               as.numeric(gender=="female"))]

reg_interact<-
  lm(wage~college + female + college*female,
     data=CPS1985)

t(reg_interact$coefficients)
```

```
     (Intercept)  college    female college:female
[1,]    9.093955 3.773582 -2.160816      0.2913502
```

Un exemple rapide

```
#On vérifie que la constante est égale au salaire moyen
# des hommes non-diplômés
all.equal(
  as.numeric(reg_interact$coefficients["(Intercept)"]),
  mean(CPS1985[gender=="male"
               & college==0]$wage))
```

```
[1] TRUE
```

Un exemple rapide

```
#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$coefficients["college"]),
  mean(CPS1985[gender=="male"
               & college==1]$wage)-
    mean(CPS1985[gender=="male"
                 & college==0]$wage))
```

```
[1] TRUE
```

Un exemple rapide

```
#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$coefficients["female"]),
          mean(CPS1985[gender=="female"
                       & college==0]$wage)-
            mean(CPS1985[gender=="male"
                         & college==0]$wage))
```

```
[1] TRUE
```

Un exemple rapide

```
#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$coefficients["college:female"]),
  (mean(CPS1985[gender=="female"
                & college==1]$wage)-
     mean(CPS1985[gender=="female"
                  & college==0]$wage))-
    (mean(CPS1985[gender=="male"
                  & college==1]$wage)-
       mean(CPS1985[gender=="male"
                    & college==0]$wage)))
```

```
[1] TRUE
```

Construire la régression multiple à partir de régressions simples

- Comment calculer le coefficient sur `education` dans la régression multiple suivante, en n'utilisant que des régressions simples ?

```
regression_multiple <-
  lm(wage ~ education + experience,
     data = CPS1985)

regression_multiple$coefficients
```

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

Construire la régression multiple à partir de régressions simples

- **Théorème de Frisch-Waugh-Lovell** :
  - Régresser d'abord `wage` sur `experience`
  - Et `education` sur `experience`
  - Et finalement le résidu de la première régression sur celui de la seconde

Une petite vérification

```
regression1 <- lm(wage ~ experience,
                  data = CPS1985)
residu1 <- regression1$residuals

regression2 <- lm(education ~ experience,
                  data = CPS1985)
residu2 <- regression2$residu

regression_FWL <- lm(residu1 ~ residu2)
regression_FWL$coefficients
```

```
 (Intercept)      residu2
7.815299e-16 9.259646e-01
```

Une petite vérification

```
all.equal(
  as.numeric(regression_multiple$coefficients["education"]),
  as.numeric(regression_FWL$coefficients["residu2"]))
```

```
[1] TRUE
```

Une conséquence intéressante

- Comment interpréter le coefficient sur `female` dans la régression suivante ?

```
regression_gender_occ <-
  lm(wage ~ female + occupation,
     data = CPS1985)

regression_gender_occ$coefficients
```

```
          (Intercept)                female  occupationtechni
            8.8203894           -2.0483583            4.1414
  occupationservices      occupationoffice       occupationsa
           -1.0736478            0.2070872            -0.311
occupationmanagement
            4.6657110
```

Une conséquence intéressante

- C'est une **moyenne sur toutes les professions** (`occupation`) de l'**écart de salaire moyen entre femmes et hommes spécifique à chaque profession**
- Avec des poids proportionnels à :
  - La **part de chaque profession** dans l'emploi salarié
  - Un terme nul dans les professions exclusivement masculines ou exclusivement féminines, et maximal dans celles avec 50% de chaque sexe
    - C'est la **variance conditionnelle** de `female` : $\mathcal{V}(\texttt{female} \mid \texttt{occupation}) = \mathbb{E}[\texttt{female} \mid \texttt{occupation}]\{1 - \mathbb{E}[\texttt{female} \mid \texttt{occupation}]\}$

Une conséquence intéressante

- Ce résultat vaut parce que l'on est **saturé en `occupation`**
  - Le résultat général vaut pour $Y = \alpha + \beta D + X'\gamma + \epsilon$ lorsque la régression linéaire de Y sur X s'identifie à l'espérance conditionnelle
    - C'est en particulier le cas si X correspond à une partition de la population par des variables indicatrices

Une petite vérification (encore !)

```
ecarts_par_occ <-
  CPS1985[,
          list(ecart_fh =
                 sum(wage * female) /
                 sum(female) -
                 sum(wage * (1 - female)) /
                 sum(1 - female),
               part_occ = .N,
               part_f = mean(female),
               var_f = mean(female) *
                 (1 - mean(female))),
          by = c("occupation")]
```

Une petite vérification (encore !)

```
ecart_agrege_reg <-
  ecarts_par_occ[,
                 sum(ecart_fh *
                       part_occ *
                       var_f) /
                   sum(part_occ *
                         var_f)]

ecart_agrege_reg
```

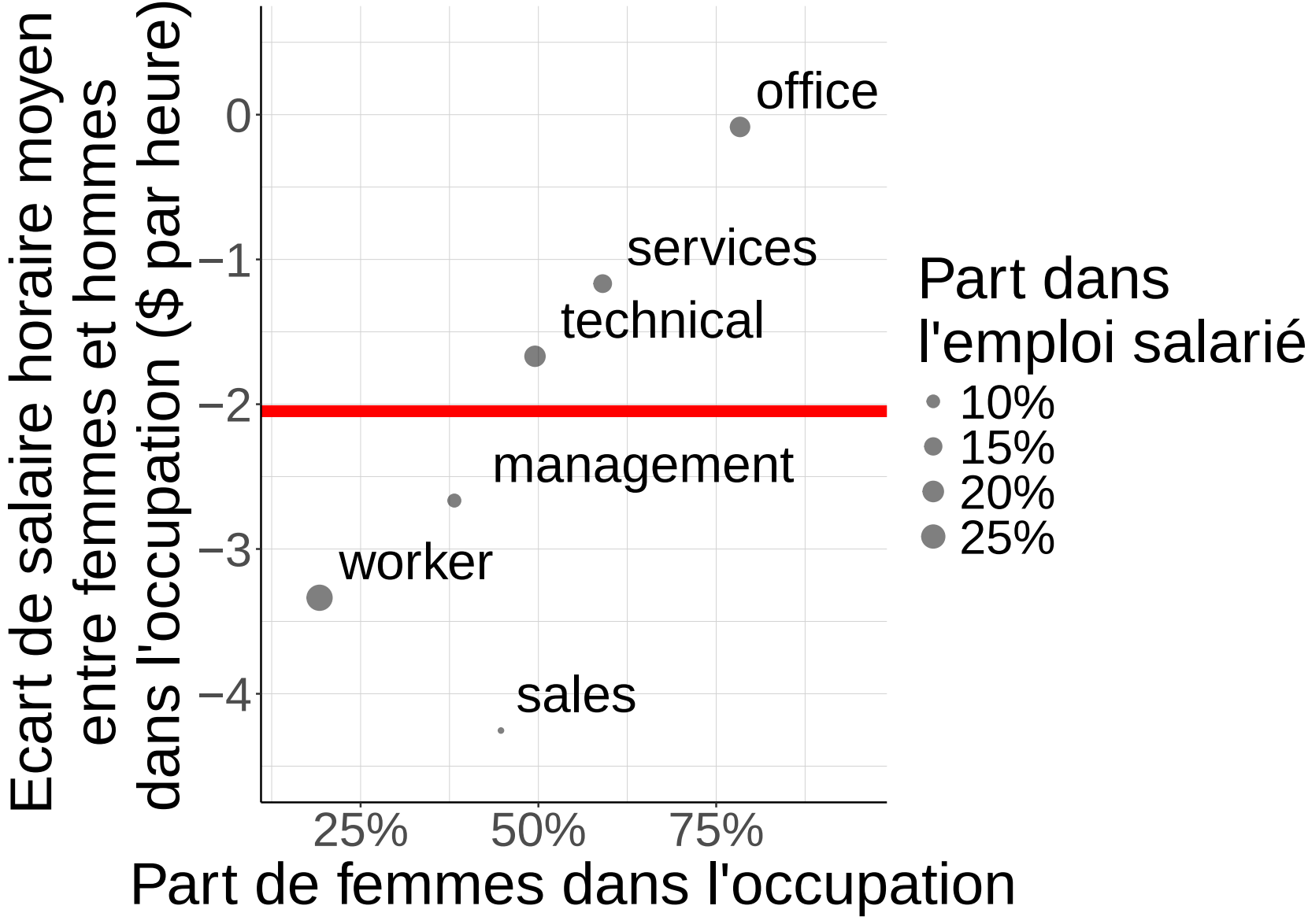
```
[1] -2.048358
```

Une petite vérification (encore !)

```
all.equal(
  as.numeric(regression_gender_occ$coefficients["female"]),
  as.numeric(ecart_agrege_reg))
```

```
[1] TRUE
```

Une visualisation possible



Une généralisation dans le cas d'un régresseur continu

- Comment interpréter le coefficient de `education` dans la régression suivante ?

```
regression_educ_reg <-
  lm(wage ~ education + region,
     data = CPS1985)

regression_educ_reg$coefficients
```

```
(Intercept)   education regionother
 -1.1393605   0.7258807   1.0077941
```

Une généralisation dans le cas d'un régresseur continu

- C'est la moyenne des coefficients sur `education` dans une série de régressions simples effectuées dans chaque groupe défini par `region`
- Avec des poids proportionnels à :
  - La taille de chaque groupe dans l'emploi salarié
  - La variance de `education` dans chaque groupe
- Même remarque que précédemment, ça marche ici parce que l'on est saturé en `region`

Une petite vérification...

```
regressions_reg <-
  CPS1985[,
          unlist(lapply(X = .SD,
                 FUN = function(variable)
                   list(coeff =
                          cov(variable, wage) /
                          var(variable),
                        variance =
                          var(variable) *
                          (.N - 1) / .N,
                        part =
                          .N)),
                 recursive = FALSE),
          .SDcols = "education",
          by = c("region")]
```

Une petite vérification (encore !)

```
regression_agreg <-
  regressions_reg[,
                  sum(education.coeff *
                      education.variance *
                      education.part) /
                    sum(education.variance *
                        education.part)]

regression_agreg
```

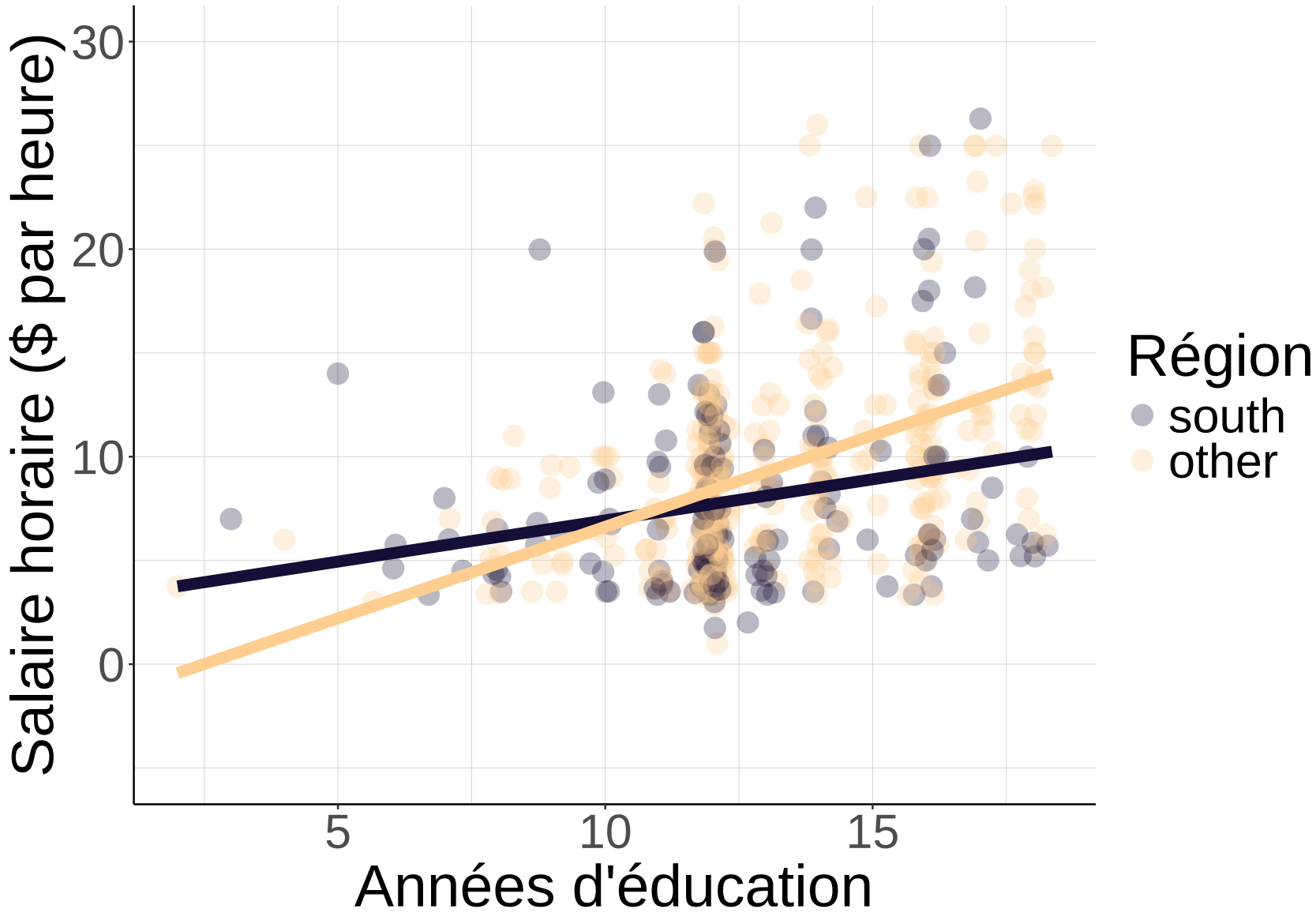
```
[1] 0.7258807
```

Une petite vérification...

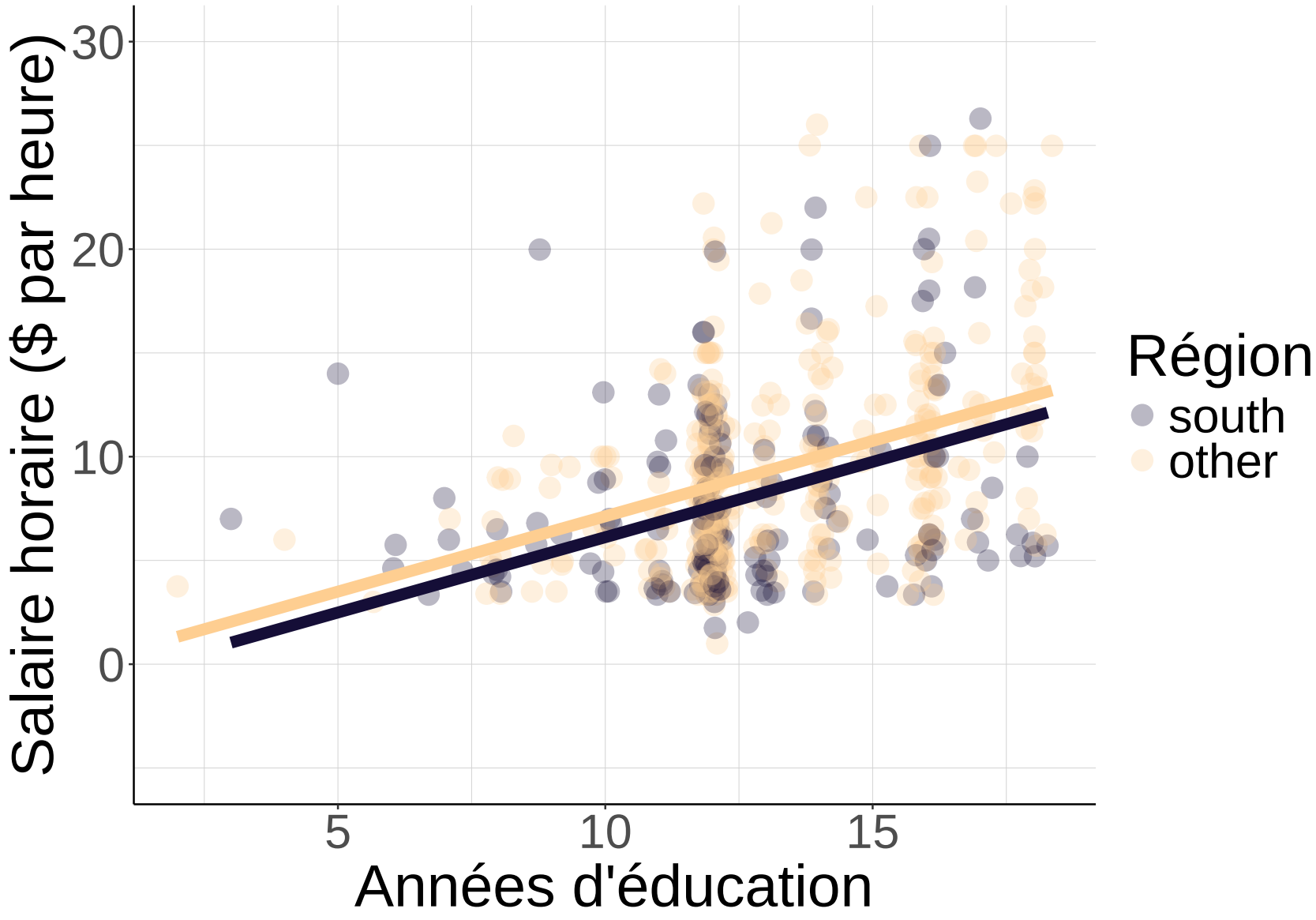
```
all.equal(
  as.numeric(regression_educ_reg$coefficients["education"]),
  as.numeric(regression_agreg))
```

```
[1] TRUE
```

Une tentative de visualisation



Une tentative de visualisation



Retour sur le coefficient de détermination

- Comme dans le cas unidimensionnel on a $\mathcal{V}(\texttt{wage}) = \mathcal{V}(\widehat{\texttt{wage}}) + \mathcal{V}(\epsilon)$
  - Théorème de Pythagore
- On peut donc définir le **coefficient de détermination** $R^2 = \frac{\mathcal{V}(\widehat{\texttt{wage}})}{\mathcal{V}(\texttt{wage})}$
  - Valeurs comprises entre 0 et 1
- **Part de la variance de `wage` que l'on peut expliquer par une combinaison linéaire des régresseurs**
  - Là encore aucune raison de donner un sens causal à ce concept d'explication !

Retour sur le coefficient de détermination

- Comment se compare le coefficient de détermination dans le cas de la régression de `wage` sur `education` à celui dans le cas de la régression de `wage` sur `education` et `experience` ?

Retour sur le coefficient de détermination

- Le sous-espace des combinaisons linéaires de 1 et `education` est inclus dans le sous-espace des combinaisons linéaires de 1, `education` et `experience`
  - Il suffit de les écrire $\beta_1 + \beta_2 \texttt{education} + 0 \cdot \texttt{experience}$
- Donc la meilleure approximation de `wage` par dans le premier ne peut jamais être strictement meilleure que la meilleure approximation dans le second
  - Au mieux ce sont les mêmes
- **Le coefficient de détermination croît avec l'inclusion de régresseurs additionnels**
  - Attention : ce n'est pas dire qu'il croît avec le *nombre* de régresseurs

Retour sur le coefficient de détermination

- Comme dans le cas unidimensionnel, le coefficient de détermination mesure la "qualité" des valeurs prédites par la régression
- **Ce n'est pas la qualité des coefficients !**
  - Ne dit rien sur la précision avec laquelle ils sont estimés
    - Intuitivement la précision dépend de la taille d'échantillon, alors que le coefficient de détermination est une quantité relative à la population !
  - Ne dit rien de l'interprétation causale / économique des coefficients
    - C'est une simple mesure de corrélation