Analyse statistique - Séance 3

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2024

Programme de la séance

La régression linéaire multiple

Les effets d'interaction

Le biais de variable omise

Modèle simple et multiple

La *régression linéaire simple* est une technique économétrique qui permet de résumer une relation entre deux variables et de s'interroger sur sa significativité statistique. Le nuage de points y est résumé par une droite, appelée droite de régression.
Ce modèle tente de lier les variations d'une *variable expliquée* (ou dépendante) à celles d'une variable *explicative* (ou indépendante, covariable, régresseur).
On dit qu'un modèle de régression est *multiple* plutôt que simple lorsqu'on étend l'explication à plusieurs variables explicatives.

Modèle simple et multiple

La *régression linéaire simple* est une technique économétrique qui permet de résumer une relation entre deux variables et de s'interroger sur sa significativité statistique. Le nuage de points y est résumé par une droite, appelée droite de régression.
Ce modèle tente de lier les variations d'une *variable expliquée*(ou dépendante) à celles d'une variable *explicative* (ou indépendante, covariable, régresseur).
On dit qu'un modèle de régression est *multiple* plutôt que simple lorsqu'on étend l'explication à plusieurs variables explicatives.

Le salaire en fonction de l'âge, une relation linéaire ?

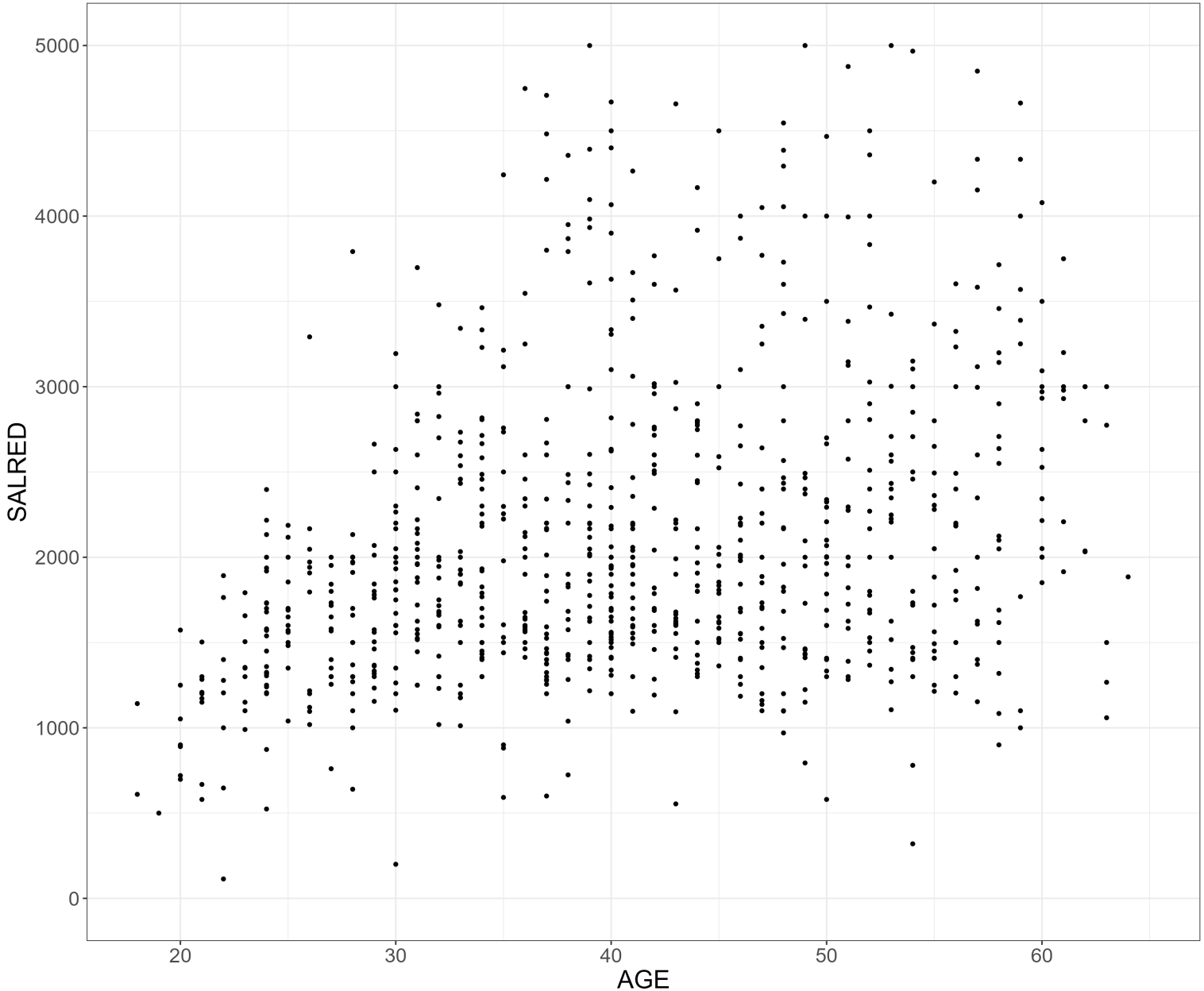


Figure: Nuage de points du salaire en fonction de l'âge

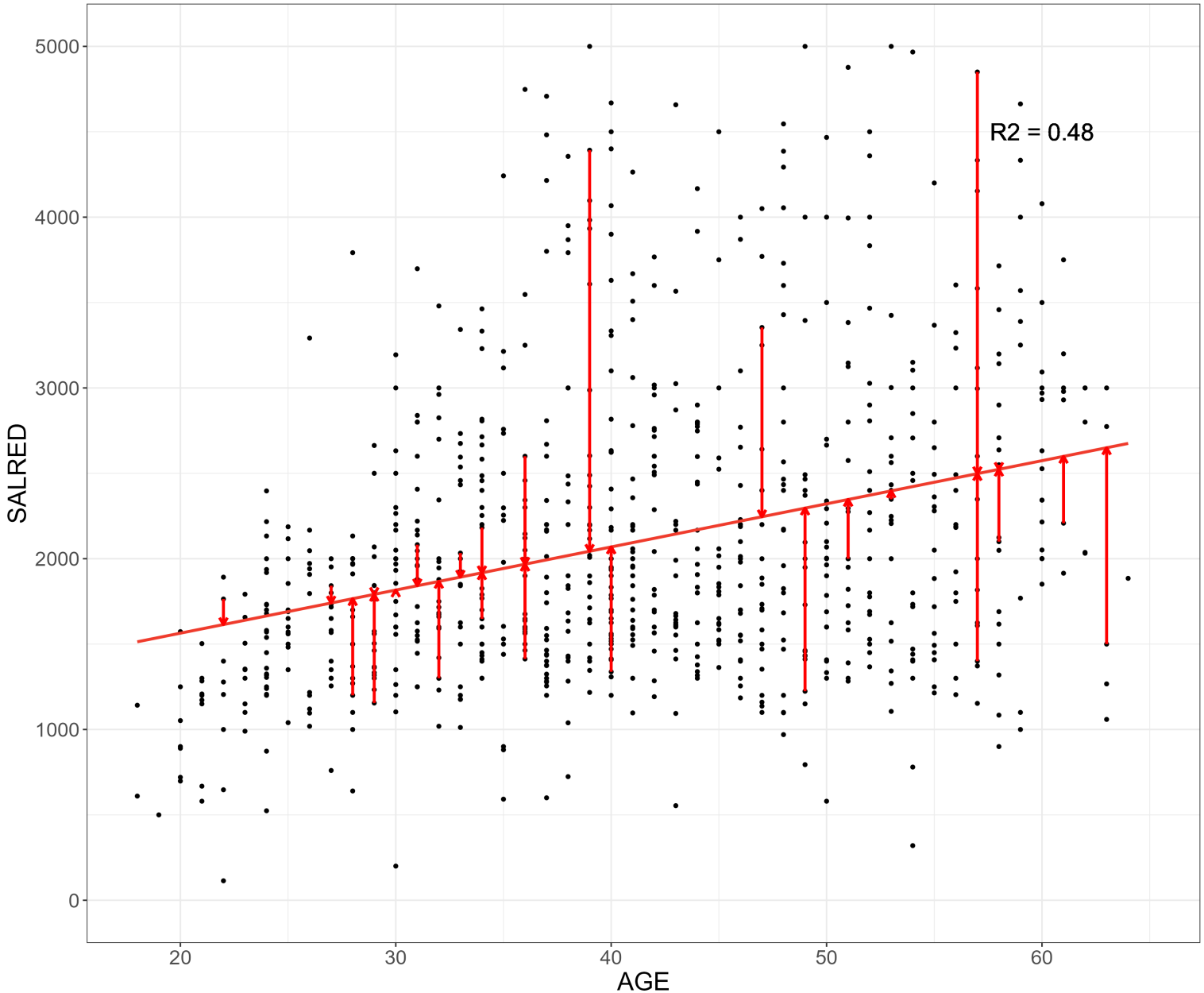
Le salaire en fonction de l'âge

Le modèle s'écrit sous la forme d'une équation applicable à tout individu statistique i :

$$salaire_i = \beta_0 + \beta_1 \times age_i + \varepsilon_i \quad (1)$$

Estimer un modèle consister à déterminer la valeur des paramètres β_0 et β_1 de manière à maximiser l'ajustement du modèle aux données. Cela revient à chercher β_0 et β_1 tels que, à partir de l'âge d'un individu, on soit en mesure de déterminer son salaire en se trompant en moyenne le moins possible.
La méthode des moindres carrés ordinaires maximise l'ajustement en minimisant la somme des termes résiduels ε_i. R_2 correspond à la part de la variance expliquée par notre modèle, c'est une mesure de sa qualité.

Exemple : composition sociale des lycées et réussite au bac

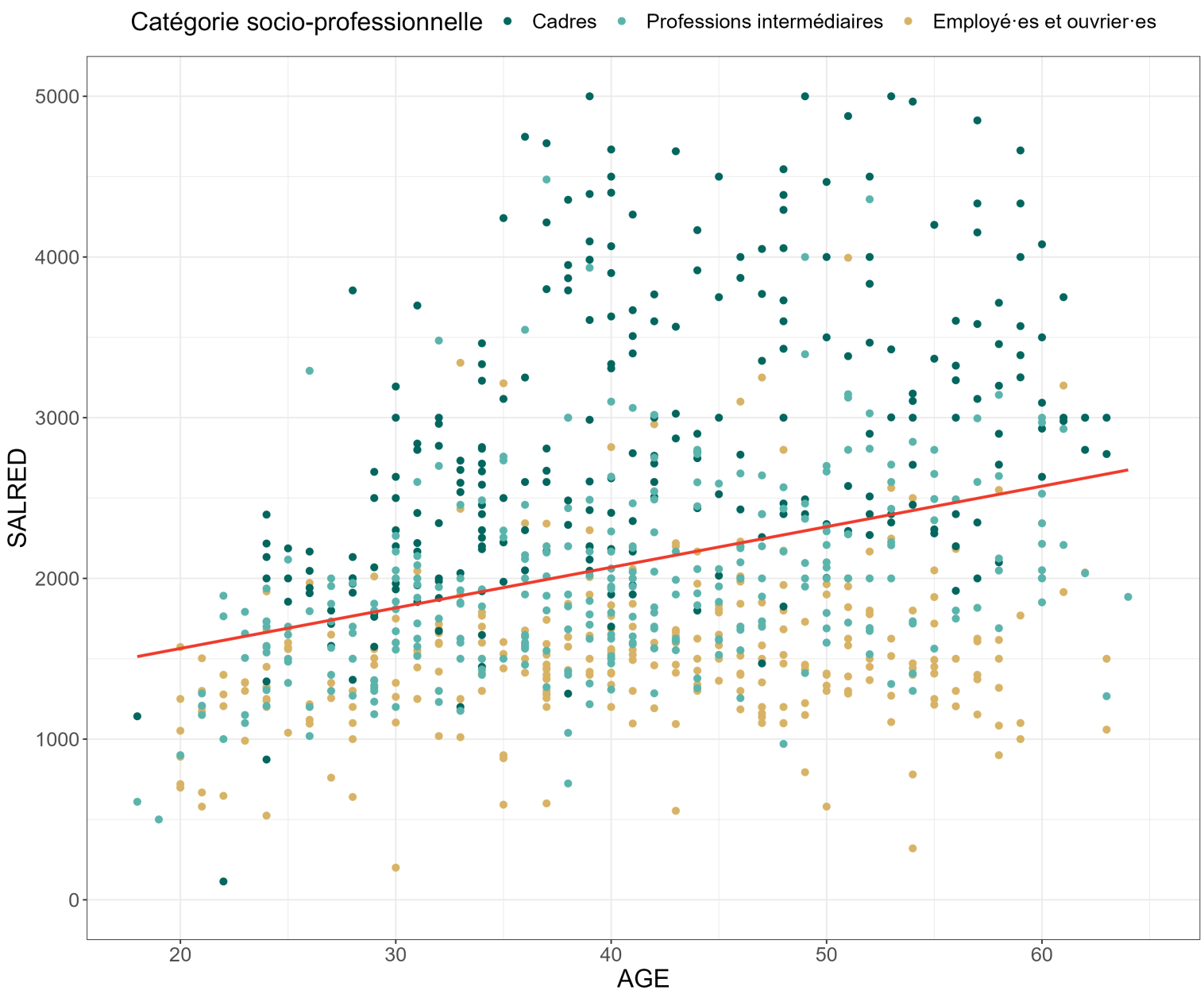


Lire un modèle de régression

	Dependent variable:
	Salaire
AGE	25.25***
	(2.61)
Constant	1,058.98***
	(110.38)
Observations	861
R^2	0.10
Adjusted R^2	0.10
Residual Std. Error	818.82 (df = 859)
Note:	*p<0.1; **p<0.05; ***p<0.01

- La constante intercept vaut 1060. Le modèle permet d'estimer qu'en moyenne, le salaire à 0 an est de 1060 euros.
- Le coefficient pour l'âge est significatif et il vaut 25,3. Le modèle permet d'estimer qu'une année d'âge augmente en moyenne le salaire de 25,3 euros.
- À 18 ans, le salaire estimé par le modèle est de $1060 + 25,3 \times 18 = 1515$ euros.
- À 60 ans, le salaire estimé par le modèle est de $1060 + 25,3 \times 60 = 2578$ euros.

Un modèle adapté ?



Les variables indicatrices

Une première façon de complexifier le modèle consiste à introduire des *variables indicatrices* : elles permettent une augmentation *de niveau* en modifiant la constante de la variable expliquée.
Ainsi, la catégorie socio-professionnelle peut-être exprimée par un codage disjonctif de 3 variables : la variable $cadre_i$ qui vaut 1 si l'individu est cadre et 0 sinon; la variable $prof\_inter_i$ qui vaut 1 si l'individu est profession intermédiaire et 0 sinon; la variable emp_ouv_i qui vaut 1 si l'individu est employé·e ou ouvrier·e et 0 sinon.

L'écriture du modèle

Dans l'idée, on aurait envie de modéliser :

$$salaire_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times cadre_i + \beta_3 \times prof\_inter_i + \beta_4 \times emp\_ouv_i + \varepsilon_i \quad (2)$$

Qu'on écrit en réalité avec une catégorie en moins, puisque le coefficient d'une catégorie parmi les trois devient la référence et correspond à β_0.

$$salaire_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times prof\_inter_i + \beta_3 \times emp\_ouv_i + \varepsilon_i \quad (3)$$

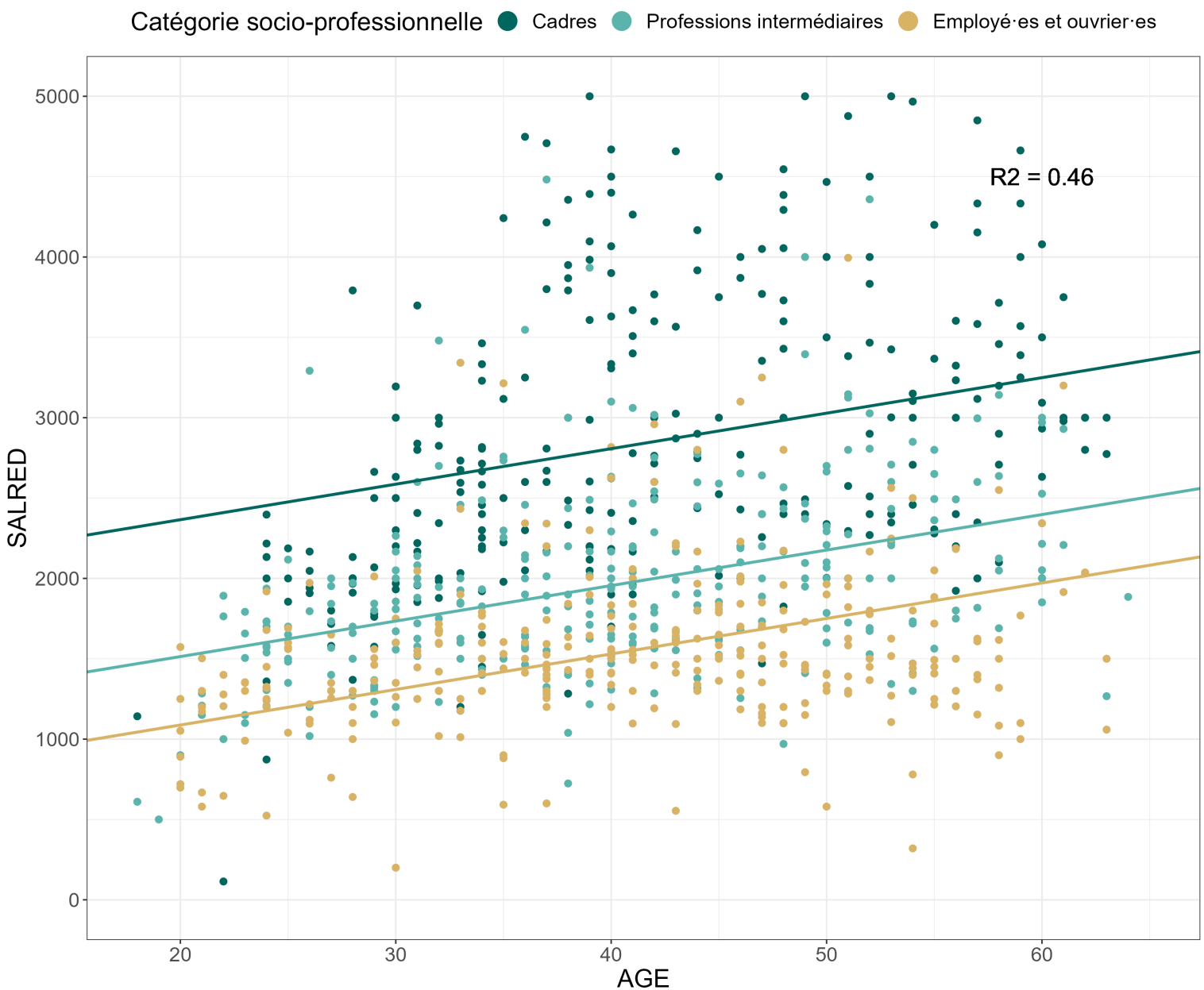
Le salaire d'un·e cadre est estimé par :

$$salaire_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_i \quad (4)$$

Quand celui d'un·e employé·e ou d'un·e ouvrier·e est estimé par :

$$salaire_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times 0 + \beta_3 \times emp\_ouv_i + \varepsilon_i \quad (5)$$

Introduire des différences de niveau



Lire un modèle de régression linéaire multiple

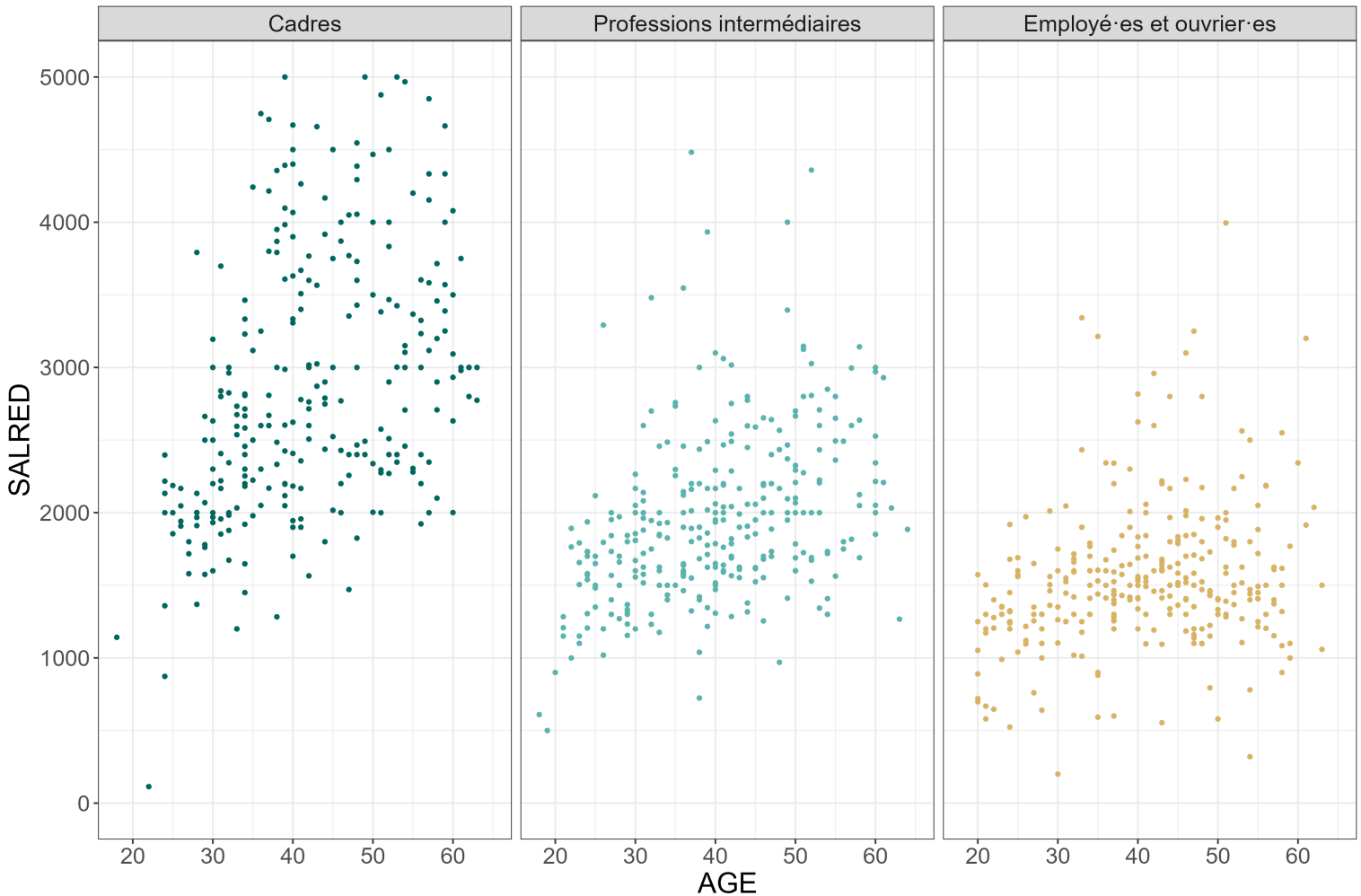
	Dependent variable:
	Salaire
AGE	22.09***
	(2.02)
CSEProfessions intermédiaires	−851.82***
	(53.41)
CSEEmployé·es et ouvrier·es	−1,277.59***
	(53.39)
Constant	1,923.44***
	(93.41)
Observations	861
R^2	0.46
Adjusted R^2	0.46
Residual Std. Error	631.45 (df = 857)
Note:	*p<0.1; **p<0.05; ***p<0.01

Lire un modèle de régression linéaire multiple

- La constante vaut 1923. En moyenne, le salaire d'un cadre (catégorie de référence) âgé de 0 ans est de 1923 euros.
- Le coefficient de l'âge est significatif et il vaut 22,1. En moyenne et **indépendamment de la catégorie socio-professionnelle**, une année en plus augmente le salaire de 22,1 euros.
- Le coefficient des professions intermédiaires est significatif et vaut -852. En moyenne, et **indépendamment de l'âge**, passer de la catégorie cadre à la catégorie profession intermédiaire fait baisser le salaire de 852 euros.

Modéliser les différences de pente

Ajuster le modèle seulement en changeant le wxpniveau des droites peut s'avérer insuffisant. Est-ce que les employé·es et les ouvrier·es bénéficient réellement de la même progression salariale au cours de leur carrière que les cadres et les professions intermédiaires ?



Complexifier le modèle - les effets d'interaction

Les *effets d'interaction* ont pour objectif de modifier – en plus de la constante – la pente des droites, donc l'intensité de la relation entre la variable expliquée et la variable explicative. Chaque catégorie se voit attribué deux coefficients propres :

$$\begin{aligned} salaire_i = &\beta_0 + \beta_1 \times age_i + \beta_2 \times prof\_inter_i + \beta_3 \times emp\_ouv_i + \\ &\beta_4 \times (age_i \times prof\_inter_i) + \beta_5 \times (age_i \times emp\_ouv_i) + \varepsilon_i \end{aligned} \tag{6}$$

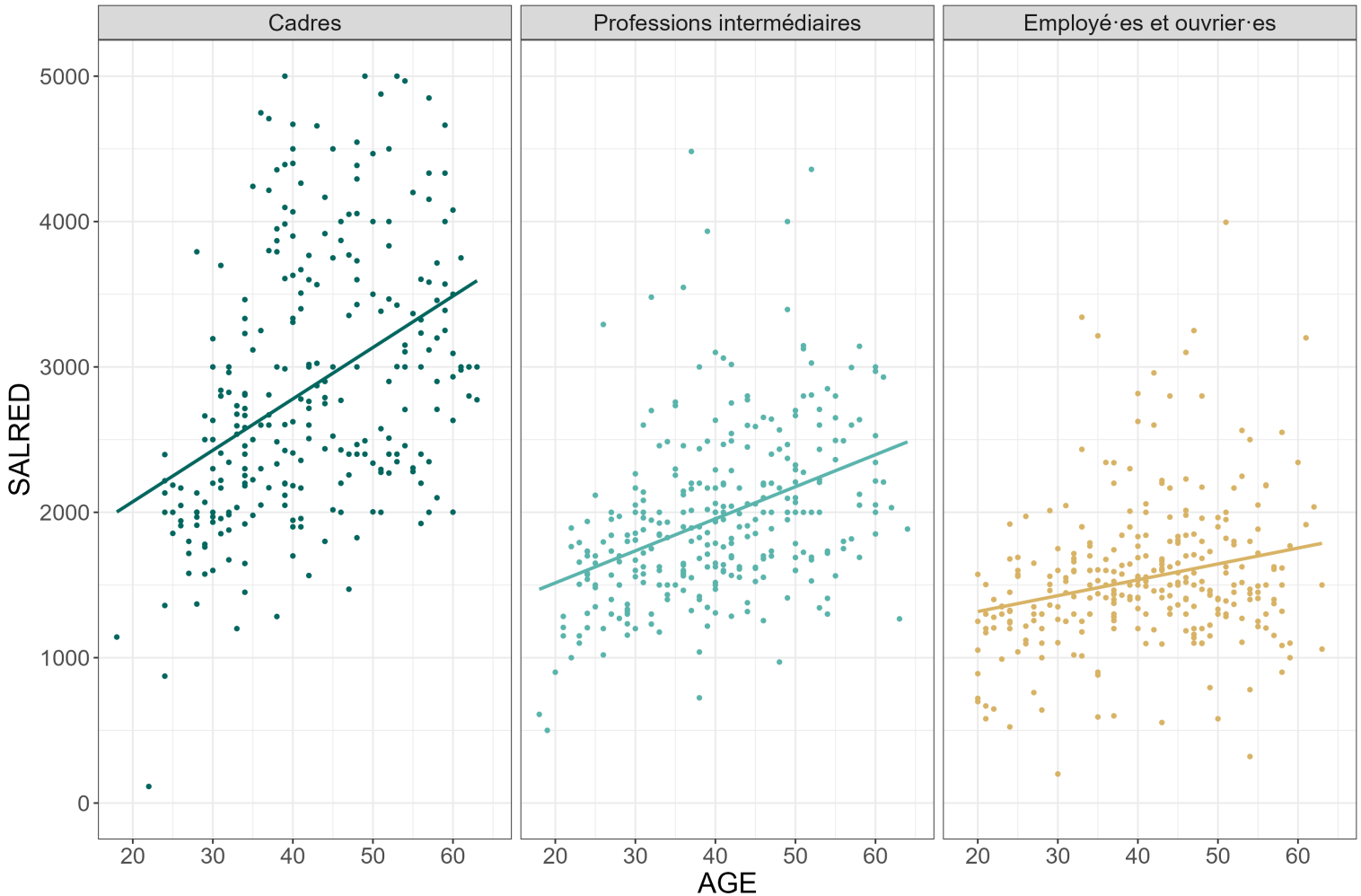
En le décomposant, on a l'effet de niveau de la catégorie socioprofessionnelle :

$$\beta_0 + \beta_2 \times prof\_inter_i + \beta_3 \times emp\_ouv_i \tag{7}$$

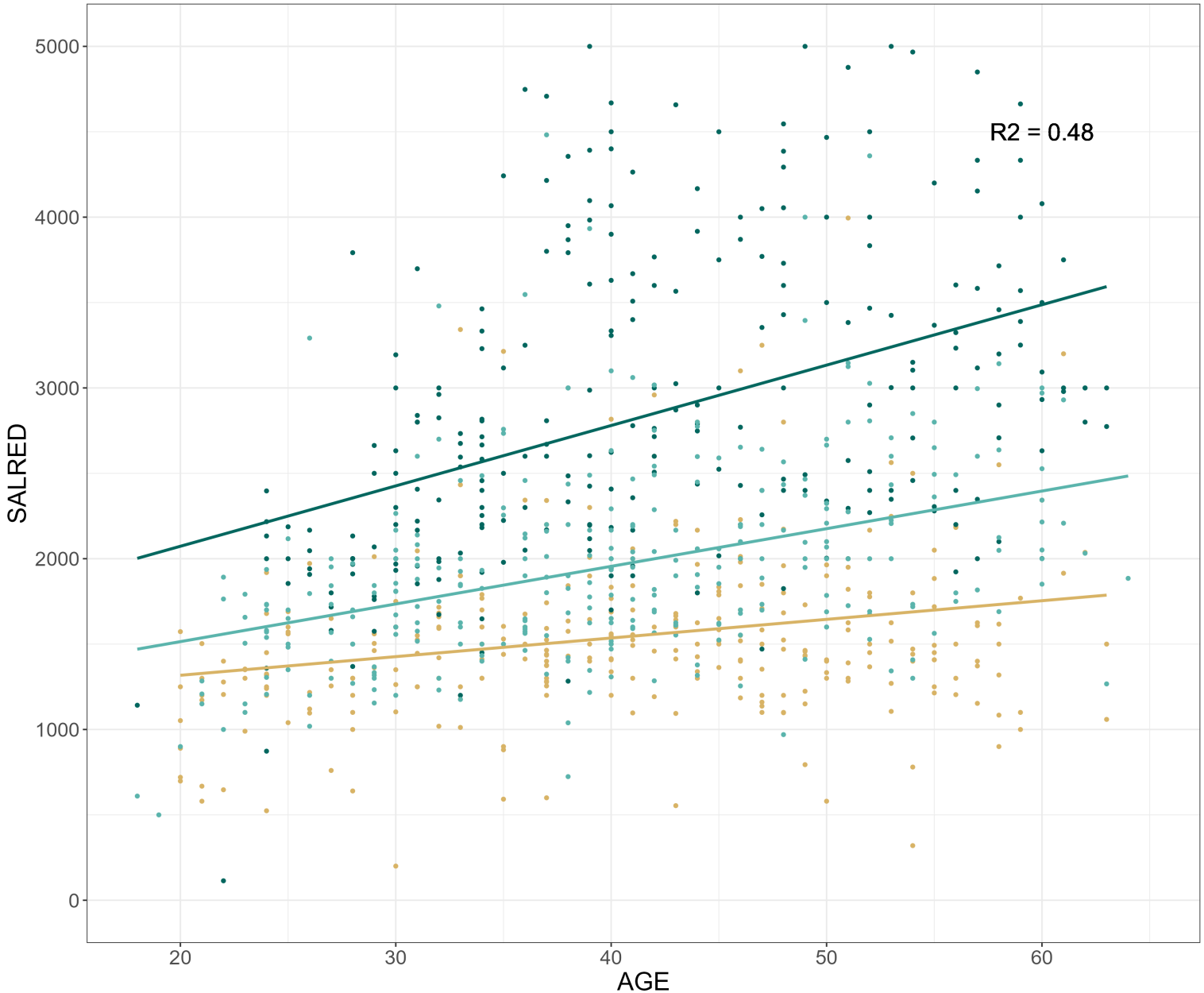
Auquel on additionne l'effet de l'âge, **en fonction de la catégorie socio-professionnelle** :

$$\beta_1 \times age_i + \beta_4 \times (age_i \times prof\_inter_i) + \beta_5 \times (age_i \times emp\_ouv_i) \tag{8}$$

Une progression salariale différenciée



Une progression salariale différenciée



Lire un modèle de régression linéaire avec un effet d'interaction

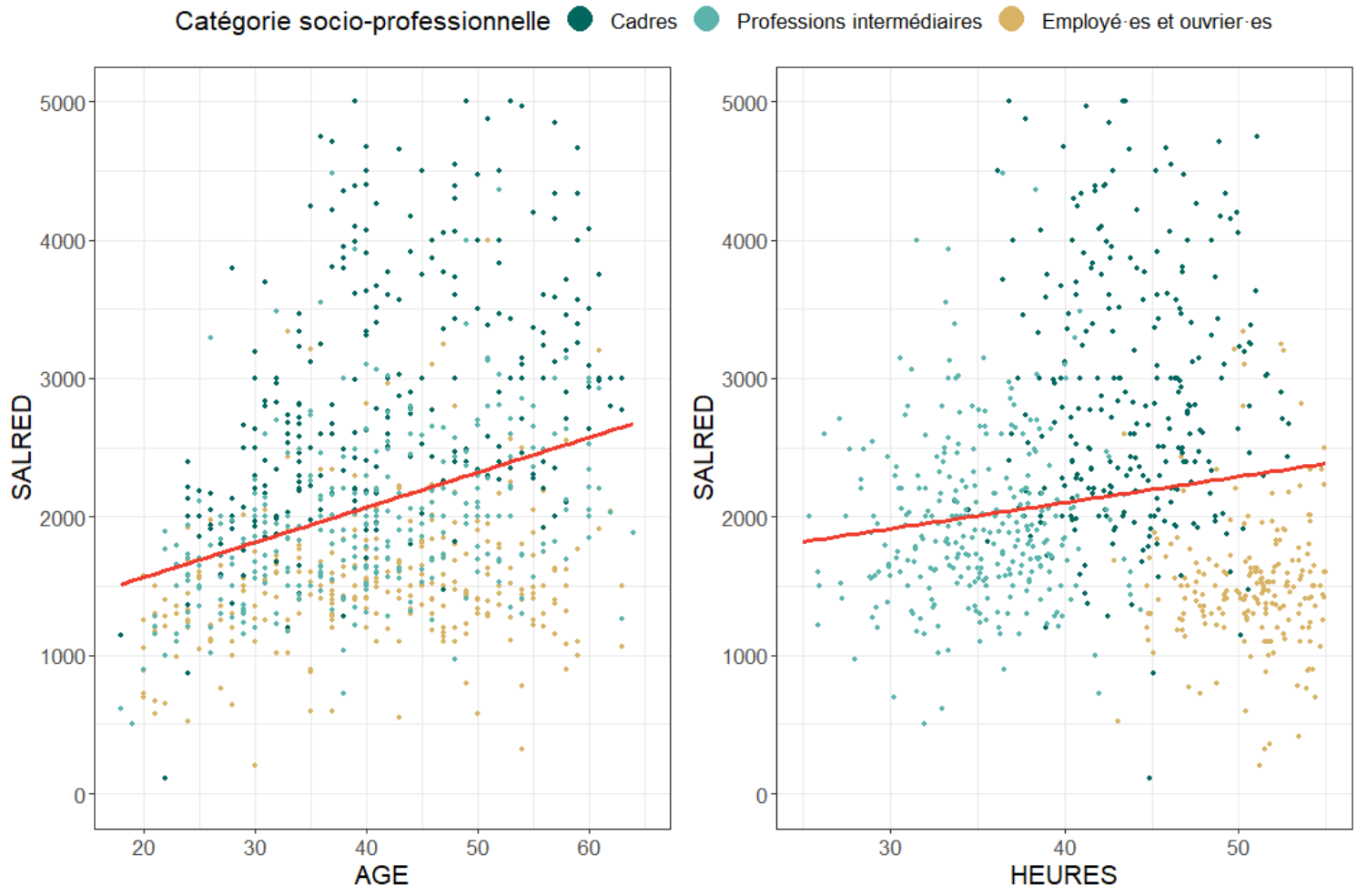
	Dependent variable:
	Salaire
AGE	35.36***
	(3.62)
CSEProfessions intermédiaires	−292.15
	(211.94)
CSEEmployé·es et ouvrier·es	−266.61
	(210.25)
AGE:CSEProfessions intermédiaires	−13.31***
	(4.99)
AGE:CSEEmployé·es et ouvrier·es	−24.44***
	(4.92)
Constant	1,365.73***
	(157.00)
Observations	861
R^2	0.48
Adjusted R^2	0.48
Residual Std. Error	623.26 (df = 855)
Note:	*p<0.1; **p≤0.05; ***p≤0.01

Lire un modèle de régression linéaire avec un effet d'interaction

- Le coefficient de l'âge est significatif et vaut 35,4. En moyenne chez les cadres (catégorie de référence), une année de plus augmente de 35,4 euros le salaire.
- Le coefficient d'interaction entre l'âge et la catégorie profession intermédiaire est significatif et vaut 13,3. En moyenne, passer de la catégorie cadre à celle de profession intermédiaire fait baisser de 13,3 euros l'augmentation salariale gagnée chaque année. Ainsi, en moyenne chez les professions intermédiaires, une année de plus augmente de $35,4 - 13,3 = 22,1$ euros le salaire.
- Les coefficients des deux catégories socio-professionnelles ne sont pas significatifs. En moyenne, indépendamment de l'âge et en tenant compte de la progression salariale propre à chaque catégorie socio-professionnelle, passer de la catégorie cadre à la catégorie profession intermédiaire ne fait pas significativement baisser le salaire.
- Autrement dit, il existe bien des inégalités salariales entre catégories professionnelles mais celles-ci se constituent au fil de la carrière, quand les individus deviennent plus âgés.

Introduire des variables de contrôle

Le premier modèle de régression linéaire multiple présenté permettait d'estimer, indépendamment de l'âge, l'augmentation moyenne du salaire selon qu'on appartienne à telle ou telle catégorie socio-professionnelle.
Or, on peut s'interroger sur le rapport de causalité : les cadres sont-ils mieux payés uniquement parce que les professions qu'ils exercent sont mieux reconnues et rétribuées, ou est-ce que d'autres facteurs expliquent ces écarts salariaux ?


Catégorie socio-professionnelle
Cadres
Professions intermédiaires
Employé·es et ouvrier·es
SALRED
AGE
HEURES
0
1000
2000
3000
4000
5000
20
30
40
50
60

Le biais de variable omise

Un faux rapport de causalité

On parle *d'erreur ou de biais de la variable omise* lorsqu'on rend compte de la variance d'une variable expliquée par une *variable explicative donnée* et qu'on ne tient pas compte d'une *troisième variable susceptible d'expliquer la variance*, elle-même *corrélée* avec la variable explicative.
L'erreur de la variable manquante vient *biaiser l'estimation de l'effet causal* : on attribue à la variable explicative des variations de la variable expliquée qui sont en réalité dues à la variable manquante.

Un faux rapport de causalité

La régression simple du salaire par l'âge ne fait que comparer les salaires de personnes plus ou moins âgées. Pour que cette comparaison soit instructive, il faut être sûr qu'on a comparé des personnes comparables. « Comparer du comparable », c'est respecter le critère d'analyse toutes choses égales par ailleurs : les personnes comparées ne doivent différer au départ que par leur âge. On a déjà modifié notre modèle pour mesurer l'effet de la catégorie socio-professionnelle. Or, pour interpréter ces effets de manière pertinente, il faut s'assurer qu'on ne confond pas ce qui relève de la position des catégories sur le marché du travail avec ce qui relève du temps travaillé.

Complexifier le modèle - les variables quantitatives

Le salaire est fonction du temps de travail. Or, les catégories socio-professionnelles les mieux rémunérées sont aussi celles qui travaillent le plus.
Il y a donc un biais liée à une troisième variable omise, le temps de travail, qui nous induisait en erreur : on prenait pour l'effet de la catégorie socio-professionnelle ce qui est en réalité l'effet de l'inégalité en moyenne du temps de travail.
On met en place un nouveau modèle avec une variable quantitative du temps de travail mensuel :

$$salaire_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times prof\_inter_i + \beta_3 \times emp\_ouv_i + \beta_4 \times heures_i + \varepsilon_i \quad (9)$$

Comparer deux modèles

Dans cette situation simulée où les cadres travailleraient en moyenne plus longtemps que les professions intermédiaires et les employé·es et ouvrier·es, on obtient :

Résultats de deux modèles de régression du salaire

	Modèle 1			Modèle 2		
Caractéristique	**Beta**	**95% IC[1]**	**p-valeur**	**Beta**	**95% IC[1]**	**p-valeur**
AGE	22	18 – 26	<0,001	22	18 – 26	<0,001
CSE						
Cadres	—	—		—	—	
Professions intermédiaires	-852	-957 – -747	<0,001	-829	-970 – -687	<0,001
Employé·es et ouvrier·es	-1 278	-1 382 – -1 173	<0,001	-1 301	-1 445 – -1 158	<0,001
HEURES				2,6	-8,1 – 13	0,6

[1] IC = intervalle de confiance