



L'impact causal des habitudes sur le revenu et la croissance économique



Introduction

Afin de respecter les conditions d'utilisation des données de l'IPUMS, je me dois de citer les **auteurs** suivants qui ont collecté la database.

Licence d'utilisation des données.

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS, 2024. <https://doi.org/10.18128/D010.V15.0>

Licence d'utilisation personnelle :

Recherche sur les Nhis sur le lien entre habitudes de vie et revenus © 2025 par Grégoire Fuchs sous licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Abstract :

Dans nos sociétés modernes, les habitudes de vie, qu'elles soient adoptées individuellement ou encouragées par l'entourage, jouent un rôle fondamental dans le **bien-être** des individus. Parmi ces habitudes, celles liées à la santé – comme une **alimentation équilibrée**, une activité physique régulière ou encore une **meilleure gestion du stress** – ne se limitent pas à des bénéfices personnels. Elles peuvent également avoir **des répercussions économiques**, tant sur le plan individuel que collectif.

En effet, une meilleure santé peut améliorer la productivité, réduire l'absentéisme et favoriser une insertion plus durable sur le marché du travail.

Ainsi, l'objet de mon analyse sera d'examiner dans quelle mesure ces habitudes de vie, **souvent choisies en premier lieu au niveau individuel, alors** même qu'elles relèvent **d'une dynamique collective(changements sociaux, culturels, mentalités)**, peuvent contribuer à **l'augmentation du revenu des individus** et, par extension, à la croissance économique.

Ensuite, il est **clair que les habitudes de vie** façonnent la vie des individus. Elles créent des actions **qui , répétés, peuvent avoir des impacts** souvent sous-estimés de la part des individus.

Nombre de pages : 6 pages /2, sans tableau/annexe. Mille excuses.

Problématique : Dans quelle mesure les choix de vie des individus, adoptés à titre **individuel** peuvent-ils avoir un effet causal sur **leur revenu**, voire , **par extension, sur la croissance économique d'un pays ?**

Nous développerons **cette problématique en plusieurs parties** : revues de littérature, le cadre statistique.

Table des matières :

L'impact causal des habitudes de vie sur le revenu.....	1
I.Revue de littérature.....	3
1. Une manipulation des données manquantes sur un grand dataset.....	4
1.L'utilisation de la médiane pour les valeurs manquantes.....	5
2.Réencodage de variables afin d'avoir une justesse statistique.....	5
3.Comparaison du R2 obtenu de l'OLS Pooled avec les différentes méthodes d'imputations de revenus.....	5
2.Respect des hypothèses de Gauss Markov afin d'atteindre une optique causale.....	5
1. Respect de l'hypothèse de normalité des variables.....	5
2. La recherche de l'hypothèse d'exogénéité $E(e/X)=0$: la régression multiple.....	5
3.Le respect du caractère qualitatif/quantitatif de nos données : la linéarité	5
4. Vérification du respect de l'hypothèse de multicollinéarité parfaite ou forte.....	6
5. Homoscédasticité et autocorrélation : inférence statistiques.....	6
3.Transformation du modèle avec non-linéarités.....	6
4. Choix d'une méthode de modélisation d'inférence causale.....	6
III.Significativité statistique et signification économique.....	7
1.Tests d'effets fixes et temporels.....	7
3.Résultats de la modélisation finale.....	7
1. Significativité des variables.....	9
2. Interprétation économique des variables.....	9
IV. Conclusion et discussion.....	10
V.Bibliographie.....	11

I.Revue de littérature

Enfin, il semble important de souligner **les avancées de la recherche sur la régression sur le log revenu, autant que celle, prenant une importance plus macroéconomique de la croissance économique.**

D'emblée, il convient de souligner l'importance de la **littérature économétrique** sur les enjeux des données de l'IPUMS, qui est la plus grande base de données individuelles au monde. Ces bases de données offrent une précieuse source d'informations pour l'analyse des dynamiques économiques et sociales. Cependant, **Kennickell (1971)** souligne déjà **la réticence d'une partie des répondants à déclarer eux même leurs revenus pour ne pas subir une divulgation.** Ainsi, l'IPUMS a imputé les valeurs **de revenus avec 5 méthodes différentes** afin de résoudre ce problème. Ils conseillent d'ailleurs sur le site internet de **tester les cinq**, et j'ai donc choisi le modèle avec le R2 le plus élevé.

Ensuite, il apparaît ensuite crucial, et opportun, de mettre en avant la littérature économétrique. **Gary Becker (1962)** est le premier à avoir défini le concept de capital humain, le considérant non seulement comme un stock de connaissances, mais aussi comme un ensemble de capacités au sens large, incluant la santé, qui contribuent à accroître la productivité individuelle. Dans cette continuité, **Jacob Mincer** a développé la première modélisation économétrique établissant un lien entre le capital humain, l'expérience et le revenu.

Son modèle, bien que perfectible, affichait un coefficient de détermination (R^2) de **0,15**, marquant ainsi une avancée majeure dans l'analyse de **l'impact du capital humain sur le revenu**. Grossman souligne dès le départ que les individus ajustent leurs investissements pour atteindre un niveau optimal de **capital santé**, considéré comme une composante du **capital humain**. Cet investissement présente un double avantage : **un effet direct** sur le bien-être et **un effet indirect** qui favorise une meilleure productivité au travail, une augmentation des revenus et, in fine, une **consommation plus élevée**. Par cette analyse, Grossman établit donc un lien clair entre santé et croissance économique.

Dans la continuité de ces travaux, **Aghion (2010)** propose une synthèse des modèles de croissance, s'inscrivant dans la lignée de **Lucas**, qui met en avant le rôle de l'accumulation du **capital santé** (noté H) sur la croissance économique (Y). Dans cette analyse, on pourrait d'ailleurs approximer Y par le revenu W des individus. Dans ces modèles, au niveau théorique, il modélise une équation assez simple, $Y = A \cdot H$, où A représente un paramètre de productivité, **qui joue un rôle non négligeable dans la production**, en combinaison avec le stock H de capital humain. Au niveau économétrique, Aghion s'inspire de **Nelson et Phelps**, qui insistent sur l'effet de la **capacité cognitive**, améliorée par une meilleure santé, sur l'accroissement de la **productivité future**. Ainsi, il modélise en tentant de **capter autant l'effet "physique" que "cognitif" du capital santé sur la croissance économique**, avec un R^2 de 0,70.

Ainsi, même en variable instrumentale, **pourraient se révéler biaisées**, notamment par **l'éducation**, qui ne semble pas être **inclus dans ces modèles**. Ainsi, la **santé** pourrait tout de même capter **une partie** de l'information **liée au "diplôme"**, ce qui ne sera pas le cas de cette analyse, grâce à **l'inclusion d'une variable "Education"**. Au final, **il pourrait exister un biais dans la valeur de l'effet dans ce type d'analyse**, basée sur l'espérance de vie, et le PIB du pays. Or, la régression **Within**, dans mon analyse, va permet de **contrôler de nombreux inobservables**, en supprimant l'effet fixe ; ce qui permettra donc **un progrès notable**.

II. Le cadre statistique.

1. Une manipulation des données manquantes sur un grand dataset

Cet article **propose une analyse sur une base de données de 169 variables**, dont j'ai tiré une échantillon, couvrant certaines années entre 1980 et 2010, **venant d'études effectuées par le NHIS**, avec un échantillon représentatif, sur les habitudes de vie (heures de sommeil, santé,

type et quantités de légumes) autant que leurs revenus. En réponse à la difficulté de **charger la base de données (environ 1,6 go)**, qui sous Pandas ou datatable, **est chargée dans RAM**, je l'ai chargée avec **Dask sur le disque dur, ce qui permet l'échantillonnage aléatoire**. Cette échantillonnage aléatoire a permis alors d'avoir une **base de données plus petite, d'environ 60.000 individus**. Cette technique permet de conserver **l'ensemble des variables pour des recherches plus précises**.

1.L'utilisation de la médiane pour les valeurs manquantes

Toutefois, en raison du grand nombre **d'années et du décalage d'années**, et donc de la complexité que cela peut engendrer dans la gestion des valeurs manquantes, **il a été décidé de remplacer par la médiane**, car cela permet de conserver plus facilement la distribution de la variable, **plus que la moyenne** .

2.Réencodage de variables afin d'avoir une justesse statistique.

Ensuite, face à des codages souvent inadéquats pour les **valeurs manquantes** (999 pour NA, ...), j'ai décidé de remplacer **ces valeurs par l'appellation NA officielle**, ce qui permet bien de les remplacer par la médiane correctement.

Statistiquement, et surtout économiquement, **cela permet de ne pas confondre une valeur très élevée (999) , avec la réalité : une valeur manquante**. Cela permet de trouver quelques **premières statistiques** : les personnes ont tendance à manger 2 fruits et légumes par jour, à avoir 6 heures de sommeil et à effectuer 6 minutes d'exercices physiques vigoureux par semaine.

3.Comparaison du R2 obtenu de l'OLS Pooled avec les différentes méthodes d'imputations de revenus.

Ainsi, le NHIS proposent 5 types d'imputations pour le revenu afin de corriger des valeurs **aberrantes** : EARNIMP1,EARNIMP2,EARNIMP4....Ici, **EARNIMP4** a conduit à un R2 **plus élevé**, et sera donc utilisé tout le long de cette analyse.

2.Respect des hypothèses de Gauss Markov afin d'atteindre une optique causale.

1. Respect de l'hypothèse de normalité des variables.

Comme observé en annexe 1 dans les statistiques récapitulatives **des différentes variables** , ces variables ont tendance à **ne pas suivre une distribution normale** à cause d'une **forte dispersion des variables**. **Asymptotiquement**, au vu de la taille importante de l'échantillon (60.000 individus), il apparaît évident que l'on va **considérer que les variables suivent une loi normale** .

2. La recherche de l'hypothèse d'exogénéité $E(e/X)=0$: la régression multiple.

Ensuite, il est important de rappeler **que la question de recherche implique la recherche d'effets causaux**. Or , tenter d'obtenir un **effet causal**, signifie notamment la mise en place de variables omises, et donc de **contrôles efficaces de nos variables**. Plm nous indique d'ailleurs que $e=0$.

D'emblée, il est essentiel de distinguer les **variables de contrôle, qui vont être notamment l'éducation et le sexe et l'âge**. Ces trois variables **influencent autant la santé que le revenu de l'autre côté** ; car l'éducation/l'âge/le sexe a tendance à **être corrélé** autant au revenu qu'à santé. Ainsi, **omettre ces variables** ne permettrait pas l'analyse de nos **coefficients comme étant "causaux"**, c'est-à-dire **exogènes et facilement identifiables**.

3. Le respect du caractère qualitatif/quantitatif de nos données : la linéarité .

Etant donné que **la variable** de revenu originelle est une variable discrète (1:5000 euros, 2:10.000euros, il apparaît donc **intéressant de vouloir discréditer, et de créer un logit ordonné de panel**. Toutefois, comme l'objet de cours était de traiter des variables continues, la variable EARNIMP4 a été transformée en variables **continues en prenant la médiane: 2500, 7500...**c'est aussi une limite de cette analyse. Cela permet de **pouvoir respecter l'hypothèse de linéarité, et de permettre un respect de cette hypothèse clé** .

4. Vérification du respect de l'hypothèse de multicollinéarité parfaite ou forte.

```
< vif2 = vif2_model(fit)  
> print(vif2)
```

VEGENO	ORANGYR	FRUTNO	BREAKFAST
1.071753	1.007620	1.070302	4.043319
SNACKS	HRSLEEP	VIG10DMIN	AGE
4.023209	1.062955	1.096746	1.877553
I (AGE^2)	SEX	EDUC_13	EDUC_14
1.844127	1.056431	1.011676	1.088849
EDUC_3eme	EDUC_4eme	EDUC_5eme	EDUC_Bac2technique
1.008310	1.011084	1.003492	1.031609

Annexe : exemples de VIF, variance Inflation ratio proche de 1

Comme les VIFs sont tous inférieurs à 1, on peut percevoir notamment que l'hypothèse d'absence de multicollinéarité forte ou parfaite, est vérifiée.

Ensuite, le VIF indiquait tout de même un VIF=13, pour l'éducation **en tant que variable linéaire (identique en carré)**. Par conséquent, j'ai décidé de tenter de mettre en place des dummies pour l'éducation, afin de **diminuer fortement les problèmes de multicollinéarité soulevée** par le VIF.

5. Homoscédasticité et autocorrélation : inférence statistiques.

Au vu des tests de Breusch Pagan et Woolridge qui concluent à la violation de ces hypothèses, il apparaît évident qu'il est nécessaire **d'effectuer une correction à la White pour l'inférence statistique**.

3. Transformation du modèle avec non-linéarités.

Modele : Formula :

```
> formula <- EARNIMP4 ~ ORANGYR+ VEGENO + BREAKFAST + HRSLEEP +  
VIG10DMIN + EDUC(dummies) + SEX(dummies) + AGE +AGE^2
```

Outil utilisé : le **R2 du Between** devient plus élevé, passant de **0,90 à 0,98**.

4. Choix d'une méthode de modélisation d'inférence causale.

D'emblée, il apparaît évident qu'**une équation sur le revenu, va connaître une causalité inverse entre le revenu et l'éducation**.

Cette modélisation est la seule qui permet d'isoler un effet causal, car elle contrôle des inobservables (intelligence, milieux sociaux). .
 A contrario, le **Between** permettrait de conclure, à une plus grande échelle, en prenant en compte la variabilité inter-individuelle.

III.Significativité statistique et signification économique

1.Tests d'effets fixes et temporels.

Tests	Hausman	Fischer “twoways”	Fischer “individual effect”	Fischer “ time effect”	Breusch Pagan	Woolridge
P Value	10^-16	10^-16	10^-16	10^-16	0,2.10^-16	2.10^-16
Conclusion sur la validité statistique.	H0 rejetée :absence de corrélation entre les variables X et les résidus.. H1: est rejetée : seul le modèle Within est convergent.	H0 rejetée : rejeté :MCO est insuffisant H1: les effets fixes Within individuels et temporels sont nécessaires pour un estimateur plus efficace (ici dans le cas within) et plus convergent .	H0 : MCO et Within équivalent. H1 : les effets Within individuels sont nécessaires	H0 H1: les effets Within temporels sont nécessaires.	L'hypothèse H0 d'homoscédasticité est rejetée . H1 validée : hétéroscédasticité : il faut des erreurs robustes à la white.	Présence de corrélation des résidus .

En conclusion, seul le modèle Within semble être convergent, car le test de Hausman conclut à la convergence de l'effet fixe. Or Woolridge souligne que le Between est un cas particulier du random .En conséquence, au vu du test effectué, et afin de bénéficier également du contrôle des effets fixes, le modèle Within apparaît être la modélisation la plus adéquate.

Or, les tests de Fisher montrent également en évidence l'importance des effets fixes temporels autant qu'individuels. Par conséquent, il apparaît que la **régression la plus causale** sera la modélisation **Within**.

3.Résultats de la modélisation finale

Modèles de Panel : dependant variable LOGEARNIMP4

=====				
=====				

1	ols	2) Within	(3) Between	(4) Random

VEGENEno	-0.001** (0.0003)	-0,01*** (0.0003)	0.012*** (0.003)	-0.001** (0.0003)
ORANGYR	-0.00000*** (0.00000)	-0.00000*** (0.00000)	0.00000 (0.00000)	-0.00000*** (0.00000)
BREAKFAST	0.011*** (0.001)	0.010*** (0.002)	0.001 (0.004)	0.011*** (0.001)
HRSLEEP	-0.001*** (0.0001)	-0.001*** (0.0001)	0.0003 (0.001)	-0.001*** (0.0001)
VIG10DMIN	0.0002*** (0.00002)	0.0002*** (0.00002)	-0.004*** (0.001)	0.0002*** (0.00002)
AGE	0.002*** (0.00003)	0.002*** (0.00003)	0.003 (0.002)	0.002*** (0.00003)
I (AGE2)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	-0.00004*** (0.00001)	-0.00000*** (0.00000)
SEX	-0.107*** (0.001)	-0.108*** (0.001)	0.008 (0.012)	-0.107*** (0.001)
EDUC_Bachelor	0.184*** (0.002)	0.330*** (0.004)	3.455*** (0.652)	0.184*** (0.002)
EDUC_Doctorat	0.471*** (0.007)	0.616*** (0.008)	-14.078*** (3.398)	0.470*** (0.007)
EDUC_Master	0.291*** (0.003)	0.432*** (0.005)	-2.288* (1.189)	0.291*** (0.003)
Constant	10.367***		10.221***	10.367***

	(0.006)		(0.100)		(0.006)

Observations	784,919	784,919	61		784,919
R2	0.090	0.083	0.998		0.336
Adjusted R2	0.090	0.058	0.996		0.336
F Statistic	2,573.094*** (df = 30; 784888) 2,293.848*** (df = 30;				
764829) 542.436*** (df = 30; 30) 77,224.770***					
=====					
=====					
Note:					
*p<0.1; **p<0.05; ***p<0.01					

Annexe : tableau de comparaison **Stargazer du pooled OLS Within twoways, Between temporel et GLS twoways (certaines dummies sont enlevées)**

Ainsi, ce résultat ne **contient pas les autres variables éducation à cause d'un manque de place. Between** semble avoir des coefficients relativement proche du Within.

1. **Significativité des variables.**

En voyant le tableau situé en annexe effectué avec la correction de White, il apparaît évident que la variable **VEGENO**, est statistiquement significative, mais au **seuil de 10%**, ce qui est faible. Cette absence de significativité pourrait **s'expliquer** par un nombre trop élevé de **valeurs manquantes** dans notre échantillon, mais aussi par une **multicolinéarité**.

Ainsi, tirer des conclusions sur ces variables, n'apparaît pas être opportun, étant donné que la significativité statistique est plus souple que la significativité au sens économique. **A contrario, l'ensemble des autres variables sont statistiquement significatives au seuil de 1%.**

2. **Interprétation économique des variables**

Ici, étant **donné que l'échantillon n'est pas normalisé**, on peut conclure à des effets marginaux, **sous contrôle des effets fixes**, pour le modèle Within. De plus, on peut souligner que les coefficients de Between, qui permet d'avoir une approche plus macroéconomique, **ne semblent pas éloignés**. Ainsi, l'effet des politiques de santé ou de la qualité de vie aurait pu avoir un impact plus important qu'on ne le pense.

Ensuite, VIG10DMIN met en avant qu'une minute de sport vigoureux par **semaine** (battements cardiaques élevés) pourraient en moyenne conduire à **0,02% de plus** de salaire, à âge, éducation, et d'autres variables et des effets fixes/temporels constants, pour un individu donné. **Si une personne augmente son temps par 1minute par jour, elle pourrait alors faire augmenter son salaire à 0,14%.** En faisant **7 minutes de sport par jour** durant l'année, le salaire augmenterait

alors de 1% de plus en moyenne, toutes choses étant égales par ailleurs. On pourrait expliquer ce chiffre, **causal en within twoways**, étant donné les nombreuses variables mises en places (dont Licence Master), par les effets bénéfiques du sport sur l'humeur et sur plein d'autres éléments, comme sur la productivité de l'individu.

Enfin, il semble **qu'augmenter sa fréquence de petit déjeuner** en le prenant en continu durant 2 mois (car +1=toujours pour le NHIS), **pourrait conduire à une augmentation de 1% du salaire**. On peut l'expliquer par l'augmentation de l'énergie matinale, qui créerait une forte augmentation du salaire. Ainsi, prendre un petit-déjeuner durant l'année peut conduire à une augmentation non négligeable du salaire.

En comparaison, le fait de **faire un master** entraîne une augmentation de $(1,541-1) \times 100 = 54,1\%$ **du salaire total d'une personne s'étant arrêtée à la 6eme, environ 25% par rapport à un bac**. Et manger des légumes et des fruits, ne semble pas changer le revenu des personnes, mais la variable VEGENO est trop peu significative pour en tirer des conclusions .

Ainsi, cela signifie **qu'une heure de sommeil** de plus par jour, pourrait conduire à une **diminution de 0,1% du salaire de la personne, à éducations, âge, sexe , constant**. Cela pourrait s'expliquer par le fait , que les individus peuvent dormir plus, sans forcément avoir une bonne qualité de sommeil. Ensuite, une explication scientifique serait que dormir trop...pourrait paradoxalement nous faire sentir plus fatigué.

IV. Conclusion et discussion

En conclusion, il apparaît évident **que le modèle Within est le modèle le plus efficace, et le plus approprié pour notre analyse économétrique**. Il permet de conclure, en contrôlant des effets fixes, et donc en concluant de **manière causale** sur les coefficients. **De plus**, il s'agissait de la régression qui possédait le plus de variables statistiquement significatives.

De plus, la proximité des coefficients **avec ceux de la régression Between**, nous permet d'étendre ces résultats au niveau macroéconomique, avec un effet sur la croissance économique.

Toutefois, cette analyse **possède de nombreuses limites , notamment dû aux données utilisées :**

- les valeurs manquantes ont été remplacées par la médiane, on pourrait tester d'autres méthodes **comme le Knn ou le MICE**, en parallélisant dès le départ sur l'ensemble de la base de données.
- résumer les variables restantes : **tester une ACP pour les variables restantes**, permettant d'augmenter la significativité de la variable de nourriture.

- La variable breakfast devra être mise en dummies, ce qui permettra une lecture moins soudaine. Il est possible aussi que son effet soit sous -estimé, car 0 correspond plutôt à une valeur manquante dans cette base de données....
- extension du nombre de variables : **faire des analyses plus précises**, comme le montre la variable **ORANGYR qui certes est négative**, on pourrait ainsi tester SALADSNO pour les salades, SWEET pour les sucreries....et se faire plaisir !

Il reste important de souligner **que cette analyse nous permet de conclure sur des variables peu exploitées dans la littérature économique**, pouvant conduire à la création d'un nouveau courant, associant essentiellement des **impacts bénéfiques sur la croissance économique** et la mise en place **d'habitudes saines**. Néanmoins, cela reste sous réserve d'une **meilleure gestion des valeurs manquantes**, et sûrement d'une sélection de données plus robustes.

Ainsi, **ce mémoire permet** également de remettre en question l'analyse classique de la régression de l'éducation sur salaire, en incluant **des variables liées à la santé**, et donc aux choix pertinents de l'individu .

V.Bibliographie

FSEG, Revue de littérature de Grégoire Fuchs, "Le potentiel impact causal entre la santé et les revenus " , L3,Licence d'économie, Magistère.

Philippe Aghion (Harvard), Xavier Jaravel (Harvard). Torsten Persson (Stockholm), Dorothee Rouzet (OECD). NBER Summer Institute: July 21, 2014.

Aghion, P., Howitt, P. et Martin, F. (2010) . Le bénéfice de la santé Un apport des théories de la croissance endogène. Revue de l'OFCE, n° 112(1), 87-108. <https://doi.org/10.3917/reof.112.0087>.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5, Part 2), 9–49. <https://doi.org/10.1086/258724>

Grossman, Michael. (1972). *On the Concept of Health Capital and the Demand for Health*. *Journal of Political Economy*, **80**(2), 223–255.

VI. Annexe

Annexe 1: Tableau statistiques de la distribution des variables

EDUC	
Lyceeadulte	: 74756
14	: 56980
Bachelor	: 56816
Master	: 24069

3rd Qu.:9996	3rd Qu.:0.0000
HRSLEEP	BREAKFAST
Min. : 0.000	Min. :0.00000
1st Qu.: 6.000	1st Qu.:0.00000
Median : 6.000	Median :0.00000
Mean : 5.794	Mean :0.05553
3rd Qu.: 6.000	3rd Qu.:0.00000

MOD10DMIN	VIG10DMIN
Min. : 0.00	Min. : 0.000
1st Qu.: 0.00	1st Qu.: 0.000
Median : 0.00	Median : 0.000
Mean : 7.04	Mean : 5.381
3rd Qu.: 0.00	3rd Qu.: 0.000

FRUTNO	VEGENO
Min. : 0.000	Min. : 0.000
1st Qu.: 2.000	1st Qu.: 2.000
Median : 2.000	Median : 2.000
Mean : 2.061	Mean : 2.069
3rd Qu.: 2.000	3rd Qu.: 2.000

Tableau récapitulatif de quelques statistiques descriptives.: HRSLEEP avec une moyenne de 6 heures ; breakfast (0 est très important).

Annexe 2 : Matrice de corrélation.

Correlation:

	(Intr)	BREAKF	HRSLEE	VIG10D	EDUC	VEGENO	SEX
BREAKFAST	0.026						
HRSLEEP	-0.024	-0.232					
VIG10DMIN	-0.085	0.026	-0.051				
EDUC	-0.024	0.045	0.001	-0.104			
VEGENO	-0.864	-0.031	0.037	0.096	0.070		
SEX	-0.525	-0.001	-0.013	0.010	-0.071	0.061	
I (AGE^2)	-0.033	0.008	-0.014	0.009	-0.017	0.007	-0.021

Twoways effects Within Model

Call:

plm(formula = formula, data = pdata, effect = "twoways", model = "within"

Unbalanced Panel: n = 20000, T = 1-61, N = 784919

Annexe 3: régression avec l'ensemble des variables

Residuals:

<u>Min.</u>	<u>1st Qu.</u>	<u>Median</u>	<u>3rd Qu.</u>	<u>Max.</u>
-3.220204	-0.073845	0.010247	0.119646	2.073910

Coefficients:

	<u>Estimate</u>	<u>Std. Error</u>	<u>t-value</u>	<u>Pr(> t)</u>	
<u>VEGENO</u>	-1.1837e-03	3.1102e-04	-3.8059	0.0001413	***
<u>ORANGYR</u>	-1.6168e-06	6.2734e-07	-2.5773	0.0099583	**
<u>BREAKFAST</u>	1.0135e-02	1.6441e-03	6.1646	7.071e-10	***
<u>HRSLEEP</u>	-7.3019e-04	8.1033e-05	-9.0110	< 2.2e-16	***
<u>VIG10DMIN</u>	1.7310e-04	2.4456e-05	7.0783	1.461e-12	***
<u>AGE</u>	2.1670e-03	3.3253e-05	65.1677	< 2.2e-16	***
<u>I (AGE^2)</u>	-2.1463e-06	5.5245e-08	-38.8497	< 2.2e-16	***
<u>SEX</u>	-1.0785e-01	1.0986e-03	-98.1667	< 2.2e-16	***
<u>EDUC_13</u>	-1.1002e-01	6.1352e-03	-17.9323	< 2.2e-16	***
<u>EDUC_14</u>	2.9979e-02	4.1124e-03	7.2899	3.105e-13	***
<u>EDUC_3eme</u>	-1.5231e-01	6.8431e-03	-22.2570	< 2.2e-16	***
<u>EDUC_4eme</u>	-7.7104e-02	7.0658e-03	-10.9123	< 2.2e-16	***
<u>EDUC_5eme</u>	-8.9108e-02	1.0526e-02	-8.4654	< 2.2e-16	***
<u>EDUC_Bac2technique</u>	1.4731e-01	4.9135e-03	29.9804	< 2.2e-16	***
<u>EDUC_Bac2univ</u>	1.7229e-01	5.3251e-03	32.3544	< 2.2e-16	***
<u>EDUC_Bachelor</u>	3.3040e-01	4.1049e-03	80.4889	< 2.2e-16	***
<u>EDUC_Bacrate</u>	-5.5630e-02	6.6397e-03	-8.3784	< 2.2e-16	***
<u>EDUC_CE1</u>	-2.2336e-01	1.7878e-02	-12.4931	< 2.2e-16	***
<u>EDUC_CE2</u>	-1.6537e-01	1.3289e-02	-12.4442	< 2.2e-16	***
<u>EDUC_CM1</u>	-8.4211e-02	1.3803e-02	-6.1010	1.055e-09	***
<u>EDUC_CM2</u>	-1.1858e-01	1.2458e-02	-9.5185	< 2.2e-16	***
<u>EDUC_CP</u>	-1.6985e-01	2.3641e-02	-7.1844	6.758e-13	***
<u>EDUC_Doctorat</u>	6.1559e-01	8.1498e-03	75.5347	< 2.2e-16	***
<u>EDUC_Doctoratpro</u>	1.3118e-01	1.4348e-02	9.1425	< 2.2e-16	***
<u>EDUC_EcolePro</u>	6.8751e-01	8.1159e-03	84.7115	< 2.2e-16	***
<u>EDUC_Lyceeadulte</u>	2.1790e-02	3.9939e-03	5.4557	4.879e-08	***
<u>EDUC_Master</u>	4.3197e-01	4.6754e-03	92.3936	< 2.2e-16	***
<u>EDUC_Maternelle</u>	-1.2549e-01	1.2959e-02	-9.6836	< 2.2e-16	***
<u>EDUC_Premiere</u>	-1.7763e-01	6.3033e-03	-28.1798	< 2.2e-16	***
<u>EDUC_Seconde</u>	-1.2880e-01	6.5794e-03	-19.5764	< 2.2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 168080

Residual Sum of Squares: 154200

R-Squared: 0.082548
Adj. R-Squared: 0.05845
F-statistic: 2293.85 on 30 and 764829 DF, p-value: < 2.22e-16

Annexe 4: correction à la White des écarts types, inférence robuste.

```
coeftest(fe_modelLIN, vcov = vcovHC(fe_modelLIN, type = "HC0"))
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
VEGENO	-1.1837e-03	5.4287e-04	-2.1805	0.0292224 *
ORANGYR	-1.6168e-06	1.3527e-07	-11.9528	< 2.2e-16 ***
BREAKFAST	1.0135e-02	7.6598e-04	13.2316	< 2.2e-16 ***
HRSLEEP	-7.3019e-04	1.1212e-04	-6.5127	7.387e-11 ***
VIG10DMIN	1.7310e-04	4.0122e-05	4.3144	1.600e-05 ***
AGE	2.1670e-03	3.1917e-05	67.8965	< 2.2e-16 ***
I(AGE^2)	-2.1463e-06	3.7236e-08	-57.6397	< 2.2e-16 ***
SEX	-1.0785e-01	1.1745e-03	-91.8221	< 2.2e-16 ***
EDUC_13	-1.1002e-01	9.4658e-03	-11.6226	< 2.2e-16 ***
EDUC_14	2.9979e-02	6.1934e-03	4.8405	1.295e-06 ***
EDUC_3eme	-1.5231e-01	1.0760e-02	-14.1547	< 2.2e-16 ***
EDUC_4eme	-7.7104e-02	1.0434e-02	-7.3895	1.476e-13 ***
EDUC_5eme	-8.9108e-02	1.5563e-02	-5.7257	1.031e-08 ***
EDUC_Bac2technique	1.4731e-01	7.3900e-03	19.9336	< 2.2e-16 ***
EDUC_Bac2univ	1.7229e-01	7.5600e-03	22.7896	< 2.2e-16 ***
EDUC_Bachelor	3.3040e-01	6.0428e-03	54.6766	< 2.2e-16 ***
EDUC_Bacrate	-5.5630e-02	1.0346e-02	-5.3770	7.577e-08 ***
EDUC_CE1	-2.2336e-01	2.8930e-02	-7.7207	1.158e-14 ***
EDUC_CE2	-1.6537e-01	2.0828e-02	-7.9396	2.031e-15 ***
EDUC_CM1	-8.4211e-02	1.9184e-02	-4.3898	1.135e-05 ***
EDUC_CM2	-1.1858e-01	1.8248e-02	-6.4982	8.134e-11 ***
EDUC_CP	-1.6985e-01	4.0563e-02	-4.1873	2.824e-05 ***
EDUC_Doctorat	6.1559e-01	1.3072e-02	47.0911	< 2.2e-16 ***
EDUC_Doctoratpro	1.3118e-01	5.7068e-03	22.9864	< 2.2e-16 ***
EDUC_EcolePro	6.8751e-01	1.2973e-02	52.9958	< 2.2e-16 ***
EDUC_Lyceeadulte	2.1790e-02	5.8858e-03	3.7020	0.0002139 ***
EDUC_Master	4.3197e-01	6.9130e-03	62.4873	< 2.2e-16 ***
EDUC_Maternelle	-1.2549e-01	2.0894e-02	-6.0062	1.900e-09 ***
EDUC_Premiere	-1.7763e-01	1.0394e-02	-17.0893	< 2.2e-16 ***
EDUC_Seconde	-1.2880e-01	1.0323e-02	-12.4776	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1