

Introduction.....	3
Données.....	3
1. Sources et Description de la Base de Données.....	3
a. Nature et Unités des Variables.....	4
b. Préparation et Traitement des Données avec Python.....	4
c. Tableau Synthétique des Variables.....	5
2. statistiques descriptives et représentations graphiques.....	5
Statistiques descriptives.....	5
Analyse.....	6
Représentations graphiques.....	6
Illustration.....	6
Analyse.....	6
Conclusion.....	7
Illustrations.....	7
Analyse.....	7
1. Relation entre l'IPS et le Niveau de vie Commune (Graphique de gauche).....	7
2. Relation entre l'IPS et les Effectifs (Graphique de droite).....	8
Présentation des modèles.....	8
1. Modèle Niveau-Niveau.....	9
a. Présentation du modèle.....	9
b. Interprétation des paramètres.....	9
c. Tests de significativité.....	10
d. Respect des hypothèses.....	11
2. Modèle Niveau-Niveau avec terme quadratique.....	12
a. Présentation du modèle.....	12
b. Intuition sur les signes et Interprétation des paramètres.....	12
c. Tests de significativité.....	13
d. Respect des hypothèses.....	14
3. Meilleur Modèle :.....	14
Conclusion.....	15

Introduction

En partant d'une recherche dans le domaine de l'économie de l'éducation, nous avons choisi la question suivante : "Dans quelle mesure les caractéristiques socio-économiques des élèves et les ressources des écoles influencent-elles les performances scolaires des élèves ?

Cependant, étant donné que la variable dépendante est l'IPS et qu'elle ne permet pas une mesure directe des performances scolaires, il nous est paru plus pertinent d'examiner les facteurs qui influencent le niveau socio-économique des élèves. Nous avons donc opté pour une solution qui permet l'étude de l'impact des caractéristiques des écoles et du niveau de vie des communes sur l'IPS .

Nous nous sommes donc posé la question : "Comment les caractéristiques des écoles et le niveau de vie des communes influencent l'Indice de Position Sociale (IPS) ?".

L'IPS est un indicateur clé pour analyser les inégalités sociales et éducatives . Avec des valeurs comprises entre 45 et 185. L' IPS mesure les conditions socio-économiques et culturelles des élèves en se basant sur les catégories socioprofessionnelles des parents .

Sa moyenne est autour de 104.7, cela signifie que globalement le milieu socio-économique des élèves est légèrement au-dessus de la valeur de référence nationale, souvent fixée autour de 100. Un IPS supérieur à 100 indique un milieu socio-économique plus favorisé, tandis qu'un IPS plus bas signifie un environnement moins aisé.

En appliquant les concepts et méthodes vus en cours , nous avons pu explorer les interactions entre l'environnement scolaire, les contextes socio-économiques locaux, et leur impact sur l'IPS.

L'ajout de la variable niveaux de vie des communes a permis d'élargir notre analyse, afin d'intégrer une perspective plus complète sur les facteurs qui contribuent à ces inégalités.

L'objectif étant de mieux comprendre ces relations pour identifier des leviers d'action visant à réduire les disparités de l'IPS .

Données

1. Sources et Description de la Base de Données

Pour cette analyse empirique, nous avons travaillé avec une base de données en coupe transversale contenant **28 125 observations** et couvrant **5 variables principales**. Cette base regroupe des données à la fois sur les établissements scolaires et sur leur contexte socio-économique. Les données proviennent de deux sources principales :

Base principale : Données sur les établissements scolaires (Secteur, Effectifs, IPS, type d'établissement) obtenues à partir des données ouvertes du ministère de l'Éducation nationale.

Base complémentaire : Niveau de vie des communes, obtenu depuis les statistiques de l'INSEE.

La fusion des deux bases a été réalisée à l'aide de Python et Jupyter Notebook, en utilisant le **code INSEE des communes** comme clé de jointure.

Ces données peuvent être consultées directement sur les portails des données ouvertes des institutions respectives, comme le ministère de l'Éducation ou l'INSEE.

a. Nature et Unités des Variables

La base est composée de **variables qualitatives et quantitatives** :

- **Secteur (0 = Public, 1 = Privé) :** Variable qualitative transformée en indicatrice binaire pour identifier le type d'établissement.
- **type_Établissement (0 = Élémentaire, 1 = Primaire) :** Variable également transformée en indicatrice binaire.
- **Effectifs :** Nombre d'élèves dans chaque établissement, une variable quantitative discrète.
- **IPS (Indice de Position Sociale) :** Une variable quantitative continue qui mesure le contexte socio-économique des élèves. Sa moyenne nationale est souvent fixée autour de 100.
- **Niveau de vie commune :** Indicateur quantitatif continu mesurant le niveau de vie moyen des habitants d'une commune en milliers d'euros.

Les unités de mesure sont claires :

- Les **Effectifs** sont comptés en nombre d'élèves.
- L'**IPS** est une valeur composite sans unité, mais interprétée comme un indice.
- Le **Niveau de vie Commune** est mesuré en milliers d'euros.

b. Préparation et Traitement des Données avec Python

Plusieurs étapes ont été réalisées à l'aide de Python pour rendre la base exploitable :

Transformation des Variables Qualitatives :

- La variable **Secteur** (initialement sous forme de texte "Public" ou "Privé") a été convertie en une variable binaire : 0 pour Public et 1 pour Privé.
- De même, **type_Établissement** (texte : "Élémentaire" ou "Primaire") a été transformé en binaire : 0 pour Élémentaire et 1 pour Primaire.

Fusion des Bases :

- Une jointure a été effectuée entre les deux bases de données (base principale et base INSEE) en utilisant le **Code INSEE des communes** comme clé de fusion.
- Les différences de format (par exemple, les zéros initiaux manquants dans les codes) ont été corrigées à l'aide de Python pour garantir une correspondance parfaite.

Nettoyage des Données :

- Les observations contenant des données manquantes ou incohérentes (par exemple, des valeurs manquantes pour le Niveau de vie Commune) ont été identifiées et supprimées.
- La structure finale des données a été vérifiée pour garantir la qualité de l'analyse.

c. Tableau Synthétique des Variables

Variable	Type	Description	Unité
Secteur	Qualitative (binaire)	0 = Public, 1 = Privé	-
type_Etablissement	Qualitative (binaire)	0 = Élémentaire, 1 = Primaire	-
Effectifs	Quantitative (discrète)	Nombre d'élèves	Élèves
IPS	Quantitative (continue)	Indice socio-économique	Indice
Niveau de vie Commune	Quantitative (continue)	Niveau de vie moyen en milliers d'euros	Milliers d'euros

1. Statistiques descriptives des variables

2. statistiques descriptives et représentations graphiques

Statistiques descriptives

	Secteur	Effectifs	IPS	type_Etablissement	Niveau de vie Commune
count	28125.000000	28125.000000	28125.000000	28125.000000	28125.000000
mean	0.142222	125.014293	104.750706	0.593991	20.296103
std	0.349284	82.318881	15.274468	0.491095	3.227229
min	0.000000	25.000000	60.700000	0.000000	11.761000
25%	0.000000	61.000000	95.200000	0.000000	18.323000
50%	0.000000	105.000000	104.400000	1.000000	19.755000
75%	0.000000	167.000000	114.200000	1.000000	21.677000
max	1.000000	755.000000	160.600000	1.000000	46.251000

2. Statistiques Descriptives des Variables de l'Étude, par *Python*

Analyse

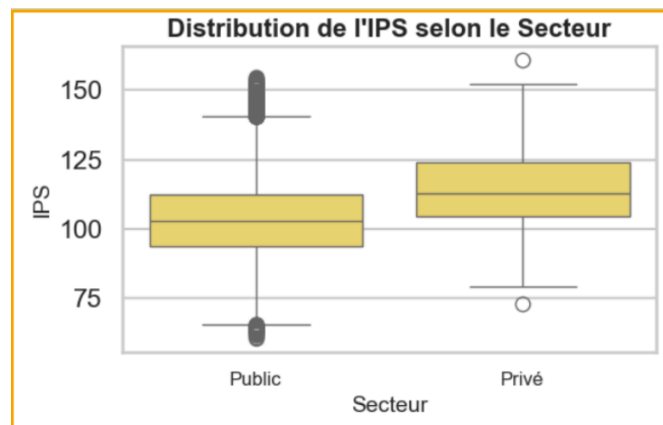
Ce tableau offre un bon aperçu de nos données. Nous remarquons que seulement 14 % des établissements sont privés, ce qui montre la prédominance du secteur public. En termes de taille, les établissements ont en moyenne 125 élèves, mais nous observons une grande variation, avec des effectifs allant de 25 à 755 élèves.

L'IPS moyen est de 104,75, ce qui est légèrement supérieur à la norme nationale. Toutefois, il existe une forte disparité, avec des établissements situés dans des zones très défavorisées (IPS de 60,7) et d'autres dans des zones très favorisées (IPS de 160,6).

Pour ce qui est du niveau de vie des communes, la moyenne est de 20,29 (en milliers d'euros). La répartition est relativement homogène autour de cette valeur, mais nous observons également des écarts notables entre les communes les plus défavorisées et les plus favorisées. ([voir 2](#))

Représentations graphiques

Illustration



3. Distribution de l'IPS selon le secteur, par Python

Analyse

Ce graphique illustre la **distribution de l'IPS en fonction du secteur** (public ou privé). Voici ce que nous pouvons en dire :

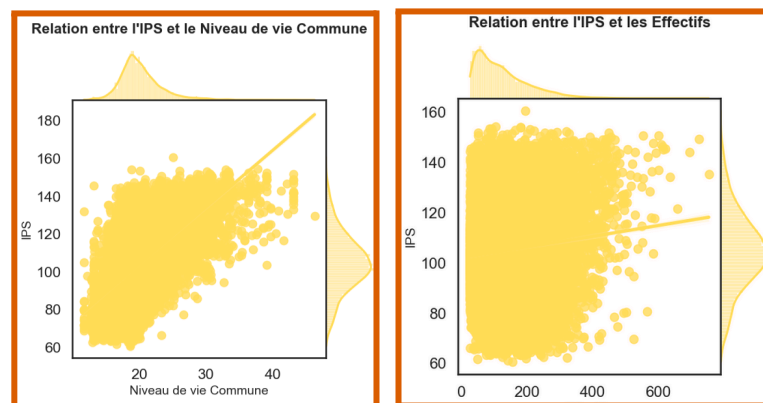
- **Secteur public :**
 - La médiane de l'IPS est proche de 100, ce qui correspond à la moyenne nationale.
 - La distribution est beaucoup plus étendue que dans le secteur privé, avec des établissements dans des zones très défavorisées (valeurs inférieures à 75) et d'autres dans des zones plus privilégiées (valeurs au-delà de 150).
- **Secteur privé :**
 - La médiane est plus élevée (autour de 125), indiquant que les établissements privés sont généralement situés dans des zones socio-économiques plus favorisées.

- La distribution est beaucoup plus resserrée, montrant moins de variabilité par rapport au secteur public. Cela suggère une homogénéité plus marquée dans le contexte socio-économique des établissements privés. ([voir 3](#))

Conclusion

Ce box plot met en évidence des disparités importantes entre le secteur public et le secteur privé. Le secteur public couvre une plus grande diversité de contextes socio-économiques, tandis que le secteur privé est davantage concentré dans des zones privilégiées. Cela reflète bien les écarts structurels entre les deux types d'établissements.

Illustrations



4. Les relations de l'IPS avec les variables quantitatives, par Python

Analyse

1. Relation entre l'IPS et le Niveau de vie Commune (Graphique de gauche)

- **Observation générale :**
 - Le graphique montre une **relation positive globale** entre le niveau de vie des communes et l'IPS.
 - Les points tendent à suivre une tendance ascendante, confirmée par la ligne de régression, ce qui suggère que plus le niveau de vie de la commune est élevé, plus l'IPS est important.
- **Interprétation :**
 - Cette relation est attendue, car un niveau de vie élevé dans une commune est souvent associé à un environnement socio-économique favorable, ce qui peut influencer positivement l'IPS des établissements scolaires.
 - Bien que la tendance soit linéaire, il y a une certaine dispersion des points, surtout pour les valeurs de niveau de vie élevées. Cela pourrait indiquer que d'autres facteurs influencent également l'IPS. ([voir 4](#))

2. Relation entre l'IPS et les Effectifs (Graphique de droite)

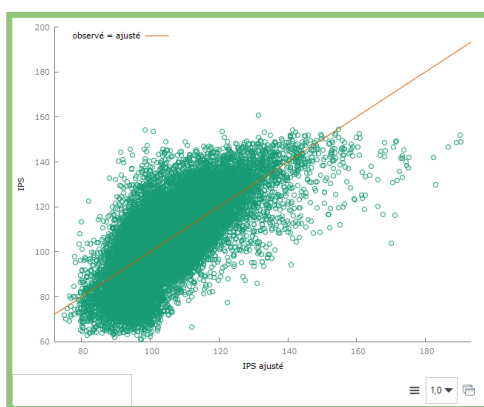
- **Observation générale :**
 - La relation entre l'IPS et les effectifs est **moins évidente**. La ligne de régression montre une très légère tendance positive, mais les points sont largement dispersés.
 - Il n'y a pas de relation linéaire forte visible entre les effectifs et l'IPS.
- **Interprétation :**
 - La faible corrélation pourrait indiquer que les effectifs, en tant que variable seule, n'expliquent pas directement l'IPS.
 - La dispersion importante pourrait aussi suggérer une relation non linéaire. Par exemple, il est possible que l'effet des effectifs varie en fonction d'un seuil, ce qui expliquerait la nécessité d'un terme quadratique pour capturer cette dynamique.

Présentation des modèles

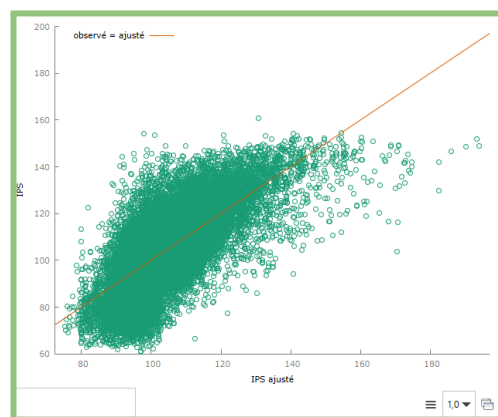
Nous allons mettre en place un modèle niveau-niveau pour commencer et nous modifions petit à petit notre modèle pour trouver le plus pertinent, celui qui expliquera au mieux les variations de notre IPS. Nous avons décidé de développer deux des modèles que nous avons mis en place ci-dessous.

Intuition graphique :

Les graphiques des deux modèles, niv-niv avec et sans Effectifs², montrent une bonne correspondance entre les valeurs observées de l'IPS et les valeurs ajustées. Dans les deux cas, les points se dispersent de manière relativement aléatoire autour de la ligne de référence, observé = ajusté. Cela indique que les deux modèles capturent correctement la relation entre les variables explicatives et l'IPS. Cette répartition aléatoire des points suggère que l'hypothèse H1, selon laquelle la moyenne des perturbations est nulle, est respectée dans les deux modèles. Aucun des graphiques ne montre de biais systématique ou de tendance particulière dans les résidus, ce qui valide la spécification des modèles à ce niveau.



5.1. Relation entre les valeurs observées et ajustées de l'IPS pour le modèle niv-niv avec Effectifs², avec Gretl



5.2. Relation entre les valeurs observées et ajustées de l'IPS pour le modèle niv-niv sans Effectifs², avec Gretl

Étant donné que nous travaillons avec une base de données de nature coupe transversale, nous avons testé l'hétéroscédasticité pour chaque modèle à l'aide des tests de White et de Breusch-Pagan, mentionnés respectivement dans [MCO1](#) et [MCO2](#). Les résultats obtenus montrent que, dans les deux modèles, les p valeurs associées à ces tests sont inférieures à 0,05. Cela signifie que l'hypothèse nulle d'homoscédasticité est rejetée, confirmant ainsi que les perturbations sont hétéroscédastiques. Pour remédier à ce problème et garantir la robustesse des résultats, nous avons corrigé les modèles en utilisant des écarts-types robustes.

Tous les R2 utilisés sont les R2 ajustés pour permettre de comparer les différents modèles entre eux.

1. Modèle Niveau-Niveau

a. Présentation du modèle

Équation à estimer

$$IPS = \beta_0 + \beta_1 \text{Secteur} + \beta_2 \text{Effectifs} + \beta_3 \text{TypeEtablissement} + \beta_4 \text{NiveauDeVieCommune} + e_i$$

Modèle 2: MCO, utilisant les observations 1-28125					
Variable dépendante: IPS					
Écarts-types robustes (hétéroscédasticité), variante HC1					
	coefficient	éc. type	t de Student	p. critique	
const	38,0924	0,578405	65,86	0,0000	***
Secteur	11,9556	0,196037	60,99	0,0000	***
Effectifs	0,0148484	0,000912128	16,28	2,60e-59	***
type_Etablissement	1,94668	0,147880	13,16	1,85e-39	***
NiveauDeVieCommune	3,05208	0,0281077	108,6	0,0000	***
Moyenne var. dép.	104,7507	Éc. type var. dép.	15,27447		
Somme carrés résidus	3266769	Éc. type régression	10,77833		
R2	0,502138	R2 ajusté	0,502067		
F(4, 28120)	4050,440	P. critique (F)	0,000000		
Log de vraisemblance	-106773,4	Critère d'Akaike	213556,8		
Critère de Schwarz	213598,0	Hannan-Quinn	213570,1		
Test de Breusch-Pagan pour l'hétéroscédasticité -					
Hypothèse nulle: homoscédasticité					
Statistique de test: LM = 630,401					
avec p. critique = P(Khi-deux(4) > 630,401) = 4,07476e-135					
Test de White pour l'hétéroscédasticité -					
Hypothèse nulle: homoscédasticité					
Statistique de test: LM = 1851,31					
avec p. critique = P(Khi-deux(12) > 1851,31) = 0					

6. Estimation par les MCO du modèle Niveau Niveau sans terme quadratique, avec *Gretl*

Équation estimée

$$\hat{IPS} = 38,09 + 11,96 \text{Secteur} + 0,01 \text{Effectifs} + 1,95 \text{TypeEtablissement} + 3,05 \text{NiveauDeVieCommune}$$

$$N = 28125 \quad R^2 = 0.502$$

b. Interprétation des paramètres

Nous sommes en modèle niveau-niveau, donc on se place en effet marginal pour tous les paramètres.

- ❖ $\hat{\beta}_0$: la valeur de l'IPS est de 38.09 lorsqu'il n'est pas influencé par le secteur ni par l'effectif, ni par le type d'établissement et le niveau de vie de la commune.

- ❖ $\hat{\beta}_1$: Si une école passe de secteur public à privé, l'IPS augmente de 11.96 unités. Toutes choses égales par ailleurs.
- ❖ $\hat{\beta}_2$: Une augmentation d'un élève dans l'effectif entraîne une augmentation de l'IPS de 0.015 unité. Toutes choses égales par ailleurs.
- ❖ $\hat{\beta}_3$: Si l'établissement est de type primaire, l'IPS augmente de 1.95 unité. Toutes choses égales par ailleurs.
- ❖ $\hat{\beta}_4$: Une augmentation d'une unité du niveau de vie moyen de la commune entraîne une augmentation de l'IPS de 3.05 unités. Toutes choses égales par ailleurs.

c. Tests de significativité

Test de Student : seuil = 5%

Hypothèse : $H_0 : B=0$ vs $H_1 : B \neq 0$

❖ Pour B_0 :

$$t(\hat{\beta}_0) = 38.09/0.58 = 65.86 \text{ et } t_{th}(28120) = 1.96 < t(\hat{\beta}_0)$$

Décision : on rejette H_0 , B_0 est significatif.

❖ Pour B_1 :

$$t(\hat{\beta}_1) = 11.96/0.2 = 60.99 \text{ et } t_{th}(28120) = 1.96 < t(\hat{\beta}_1)$$

Décision : on rejette H_0 , B_1 est significatif.

❖ Pour B_2 :

$$t(\hat{\beta}_2) = 0.015/0.0009 = 16.28 \text{ et } t_{th}(28120) = 1.96 < t(\hat{\beta}_2)$$

Décision : on rejette H_0 , B_2 est significatif.

❖ Pour B_3 :

$$t(\hat{\beta}_3) = 1.94/0.148 = 13.16 \text{ et } t_{th}(28120) = 1.96 < t(\hat{\beta}_3)$$

Décision : on rejette H_0 , B_3 est significatif.

❖ Pour B_4 :

$$t(\hat{\beta}_4) = 3.05/0.028 = 108.6 \text{ et } t_{th}(28120) = 1.96 < t(\hat{\beta}_4)$$

Décision : on rejette H_0 , B_4 est significatif.

Test de Fisher : seuil = 5%

Hypothèse : $H_0 : B_k=0$ vs $H_1 : B_k \neq 0$ $k=\{1,2,3,4\}$

Stat du test : p-value = 0.000 < 0.05 donc le modèle est globalement significatif.

Intervalle de confiance 95% :

95% confidence intervals $t(28120, 0,025) = 1,960$			
	coefficient	basse	high
const	38,0924	36,9587	39,2261
Secteur	11,9556	11,5713	12,3398
Effectifs	0,0148484	0,0130606	0,0166362
type_Etablissement	1,94668	1,65683	2,23654
NiveaudevieCommune	3,05208	2,99699	3,10718

7. Intervalle de confiance à 95%, avec Gretl

En effet, le 0 n'appartient à aucun de nos intervalles de confiance, nous avons donc 95% de chance que nos paramètres se trouvent dans leurs intervalles.

Conclusion

En termes d'intuition, nous pensions que le paramètre d'Effectifs aurait une répercussion négative sur l'IPS. Nous allons donc essayer d'ajouter un paramètre à notre modèle. Nous allons ajouter les Effectifs au carré. Un seul élève n'a pas forcément d'impact, la question est à partir de combien d'élèves il peut y avoir une répercussion négative.

Nous avons un R^2 ajusté à environ 0.502, Cela signifie qu'environ 50,2 % de la variation de l'IPS est expliquée par ce modèle, ce qui en fait un outil pertinent que nous l'utiliserons pour comparer nos modèles.

d. Respect des hypothèses

La spécification du modèle

Nous avons évalué la spécification du modèle en utilisant le test RESET de Ramsey. Les p-valeurs extrêmement faibles dans tous les cas montrent que l'hypothèse nulle, indiquant une spécification correcte, est rejetée. Cela implique que le modèle est mal spécifié.

Tests de multicollinéarité

Pour tester l'hypothèse H5 qui correspond à l'absence de multicollinéarité dans ce modèle, nous avons utilisé les facteurs d'inflation de variance (VIF) et le diagnostic de colinéarité de Belsley-Kuh-Welsch (BKW). Les VIF sont faibles, tous inférieurs à 10, indiquant une absence de colinéarité problématique. L'analyse BKW montre qu'aucun indice de condition n'est supérieur à 30. Cela confirme l'absence de multicollinéarité sévère. Un indice se situe entre 10 et 30, reflétant une colinéarité modérée, principalement liée à NiveauDeVieCommune et const. Globalement, la multicollinéarité est maîtrisée, garantissant des estimations robustes.

```
Robust RESET test for specification (carrés et cubes)
Hypothèse nulle: la spécification est adéquate
Statistique de test: F = 411,870181,
avec p. critique = P(F(2,28118) > 411,87) = 4,98e-177

Robust RESET test for specification (carrés et cubes)
Hypothèse nulle: la spécification est adéquate
Statistique de test: F = 411,870181,
avec p. critique = P(F(2,28118) > 411,87) = 4,98e-177

Robust RESET test for specification (cubes seulement)
Hypothèse nulle: la spécification est adéquate
Statistique de test: F = 873,139136,
avec p. critique = P(F(1,28119) > 873,139) = 5,27e-189
```

8. Test RESET de Ramsey, avec Gretl

```
Facteurs d'inflation de variance
Valeur minimale possible = 1.0
Des valeurs > 10.0 peuvent indiquer un problème de colinéarité

Secteur      1,108
Effectifs    1,141
type_Etablissement 1,229
NiveauDeVieCommune 1,010

VIF(j) = 1/(1 - R(j)^2), où R(j) est un coefficient de corrélation multiple
entre la variable j et les autres variables indépendantes

Diagnostic de colinéarité Belsley-Kuh-Welsch (BKW):

proportions de la variance

lambda      cond      const      Secteur Effectifs type_Eta~ NiveauDe~
3,632      1,000      0,001      0,016      0,015      0,016      0,001
0,815      2,111      0,000      0,836      0,020      0,003      0,001
0,426      2,919      0,000      0,102      0,263      0,349      0,000
0,120      5,405      0,020      0,040      0,702      0,626      0,022
0,006      24,288      0,979      0,006      0,000      0,006      0,977

lambda = valeurs propres de la matrice de covariance inverse (smallest is 0,00615745)
cond = indice de condition
note: sur une colonne, la somme des proportions de variance est égale à 1,0

Selon BKW, il y a dépendance quasi-linéaire « forte » si cond >= 30,
et « moyennement forte » si cond varie de 10 à 30. Les estimateurs dont
la variance est principalement associée à des valeurs de cond problématiques
sont eux aussi à considérer comme problématiques.

Nombre d'indices de condition >= 30: 0
Nombre d'indices de condition >= 10: 1
Proportions de variance >= 0.5 associées à cond >= 10:

const NiveauDe~
0,979      0,977
```

9. Les facteurs d'inflation de variance (VIF) et le diagnostic de colinéarité de Belsley-Kuh-Welsch (BKW), avec Gretl

2. Modèle Niveau-Niveau avec terme quadratique

a. Présentation du modèle

Équation à estimer

$$IPS = \beta_0 + \beta_1 \text{Secteur} + \beta_2 \text{Effectifs} + \beta_3 \text{TypeEtablissement} + \beta_4 \text{NiveauDeVieCommune} + \beta_5 \text{Effectifs}^2 + e_i$$

Modèle 4: MCO, utilisant les observations 1-28125
Variable dépendante: IPS

	coefficient	éc. type	t de Student	p. critique	
const	39,6471	0,457870	86,59	0,0000	***
Secteur	11,8593	0,193708	61,22	0,0000	***
Effectifs	-0,00555742	0,00233871	-2,376	0,0175	**
type_Etablissement	1,79090	0,145827	12,28	1,41e-34	***
NiveauDeVieCommune	3,04464	0,0199970	152,3	0,0000	***
sq_Effectifs	5,59521e-05	5,99243e-06	9,337	1,06e-20	***
Moyenne var. dép.	104,7507	Éc. type var. dép.	15,27447		
Somme carrés résidus	3256672	Éc. type régression	10,76185		
R2	0,503677	R2 ajusté	0,503588		
F(5, 28119)	5707,118	P. critique (F)	0,000000		
Log de vraisemblance	-106729,9	Critère d'Akaike	213471,7		
Critère de Schwarz	213521,2	Hannan-Quinn	213487,7		
Test de Breusch-Pagan pour l'hétéroscédasticité -					
Hypothèse nulle: homoscedasticité					
Statistique de test: LM = 691,347					
avec p. critique = P(Khi-deux(5) > 691,347) = 3,64916e-147					
Test de White pour l'hétéroscédasticité -					
Hypothèse nulle: homoscedasticité					
Statistique de test: LM = 1821,35					
avec p. critique = P(Khi-deux(17) > 1821,35) = 0					

10. Estimation par les MCO du modèle Niveau Niveau avec terme quadratique, avec *Gretl*

Équation estimée

$$\hat{IPS} = 39,65 + 11,86 \text{Secteur} - 0,006 \text{Effectifs} + 1,79 \text{TypeEtablissement} + 3,04 \text{NiveauDeVieCommune} + 0,000056 \text{Effectifs}^2$$

N = 28 125

R² = 0.504

b. Intuition sur les signes et Interprétation des paramètres

Intuition sur les signes des paramètres

Les paramètres du modèle représentent tous des effets marginaux. Nous nous attendions à ce que la majorité des paramètres soient positifs, à l'exception de l'effectif. En effet, IPS est toujours positif, et un établissement privé augmente IPS, car il est souvent associé à de meilleures conditions matérielles et éducatives. De plus, le niveau de vie de la commune et le

type d'établissement ont également un effet positif, puisqu'ils reflètent des environnements socio-économiques favorables qui influencent positivement l'IPS.

Concernant l'effectif, nous nous attendions à ce qu'il soit négatif, car une augmentation du nombre d'élèves peut diluer les ressources disponibles par élève, ce qui peut avoir un effet négatif sur l'IPS.

Cependant, pour Effectifs², nous nous attendions initialement à ce qu'il ait un coefficient négatif, parce qu'un effectif très important pourrait théoriquement entraîner des déséconomies d'échelle, notamment une dilution des ressources disponibles par élève, ce qui renforcerait l'effet négatif sur l'IPS.

test unilatéral :

$H_0: B_2 + 2 \cdot B_5 \geq 0$ et $H_1: B_2 + 2 \cdot B_5 < 0$ avec $t(B_2 + 2 \cdot B_5) = -2,15$, on a $1,645 < 2,15$ donc rejet H_0

Il y a donc une relation négative de l'effectif sur l'IPS, mais à partir de quel seuil ?

On fait $B_2 / 2 \cdot B_5 = -49,66$, il n'y a pas de moitié d'élèves donc à partir de 50 élèves en plus dans les effectifs il y a une répercussion négative.

Interprétation des paramètres

- ❖ $\hat{\beta}_0$: La valeur initiale de l'IPS est de 39,65, lorsqu'il n'est influencé par aucune des variables explicatives du modèle, à savoir le secteur, l'effectif, le type d'établissement, le niveau de vie de la commune ou Effectifs².
- ❖ $\hat{\beta}_1$: Si une école passe du secteur public au secteur privé, IPS augmente de 11,86 unités, toutes choses égales par ailleurs. Cela signifie que le secteur privé est associé à une meilleure situation d'IPS par rapport au secteur public.
- ❖ $\hat{\beta}_2$: Une augmentation d'un élève dans les effectifs entraîne une diminution de l'IPS de 0,006 unité, toutes choses égales par ailleurs. Cela suggère qu'un effectif plus important soit légèrement associé à une diminution de l'IPS.
- ❖ $\hat{\beta}_3$: Si l'établissement est de type primaire, l'IPS augmente de 1,79 unité, toutes choses égales par ailleurs.
- ❖ $\hat{\beta}_4$: Une augmentation d'une unité du niveau de vie moyen de la commune entraîne une augmentation de l'IPS de 3,04 unités, toutes choses égales par ailleurs. Cela montre une forte corrélation positive entre le niveau de vie de la commune et l'IPS.
- ❖ $\hat{\beta}_5$: Pour chaque augmentation d'une unité d'Effectifs², l'IPS augmente de 0,000056 unité, toutes choses égales par ailleurs. Cela reflète un effet quadratique légèrement croissant des effectifs sur l'IPS, indiquant que l'impact des effectifs n'est pas linéaire et peut devenir plus positif avec des effectifs très élevés.

c. Tests de significativité

95% confidence intervals t(28119, 0,025) = 1,960			
	coefficient	basse	high
const	39,6471	38,7496	40,5445
Secteur	11,8593	11,4796	12,2390
Effectifs	-0,00555742	-0,0101414	-0,000973427
type_Etablissement	1,79090	1,50507	2,07673
NiveaudevieCommune	3,04464	3,00545	3,08384
sq_Effectifs	5,59521e-05	4,42067e-05	6,76976e-05

11. Intervalle de confiance à 95%, avec Gretl

Nous avons testé la significativité de chaque paramètre du modèle. Tous les paramètres sont significatifs au seuil de 5 %, car leurs p-valeurs sont inférieures à 0,05. Cela signifie que chacune des variables explicatives augmente significativement l'IPS. De plus, il y a 95 % de chance que les paramètres estimés soient inclus dans leurs intervalles de confiance respectifs.

Par ailleurs, le test de Fisher a également été réalisé pour évaluer la significativité globale du modèle. Avec une p-valeur inférieure à 0,05, ce test confirme que le modèle est globalement significatif au seuil de 5 %.

Conclusion

Nous avons un R^2 ajusté d'environ 0,504. Cela signifie qu'environ 50,4 % de la variation de l'IPS est expliquée par ce modèle avec le terme quadratique. Cela indique une légère amélioration par rapport au modèle sans ce terme, renforçant sa pertinence pour analyser l'impact des variables explicatives sur l'IPS.

d. Respect des hypothèses

La spécification du modèle

Nous avons testé si le modèle est correctement spécifié à l'aide du test RESET de Ramsey. Les p-valeurs très faibles indiquent que l'hypothèse nulle, spécification adéquate, est rejetée dans tous les cas. Cela signifie que le modèle est mal spécifié.

Tests de multicollinéarité

Nous avons testé l'hypothèse H5 concernant l'absence de multicollinéarité, en utilisant les VIF et le diagnostic de Belsley-Kuh-Welsch (BKW). Les résultats montrent que les VIF des variables sont généralement faibles, avec des valeurs inférieures à 10, à l'exception des variables Effectifs (VIF = 9) et Effectifs² (VIF = 8,685), indiquant une colinéarité modérée entre ces deux variables. L'analyse BKW, basée sur les valeurs propres de la matrice des variables explicatives, révèle qu'aucun indice de condition n'est supérieur à 30, ce qui confirme l'absence de multicollinéarité sévère. Cependant, deux indices se situent entre 10 et 30, reflétant une multicollinéarité modérée principalement due aux variables Effectifs, NiveaudevieCommune et Effectifs². Ces résultats suggèrent que la multicollinéarité est maîtrisée, mais une attention particulière doit être portée aux variables fortement corrélées, notamment Effectifs et Effectifs², pour garantir la stabilité des estimations.

```
Robust RESET test for specification (carrés et cubes)
Hypothèse nulle: la spécification est adéquate
Statistique de test: F = 447,190447,
avec p. critique = P(F(2,28117) > 447,19) = 6,49e-192

Robust RESET test for specification (carrés et cubes)
Hypothèse nulle: la spécification est adéquate
Statistique de test: F = 447,190447,
avec p. critique = P(F(2,28117) > 447,19) = 6,49e-192

Robust RESET test for specification (cubes seulement)
Hypothèse nulle: la spécification est adéquate
Statistique de test: F = 858,754007,
avec p. critique = P(F(1,28118) > 858,754) = 5,7e-186
```

12. Test RESET de Ramsey, avec Gretl

```
Facteurs d'inflation de variance
Valeur minimale possible = 1.0
Des valeurs > 10.0 peuvent indiquer un problème de colinéarité

Secteur      1,112
Effectifs    9,000
type_Etablissement 1,245
NiveaudevieCommune 1,011
sq_Effectifs 8,685

VIF(j) = 1/(1 - R(j)^2), où R(j) est un coefficient de corrélation multiple
entre la variable j et les autres variables indépendantes

Diagnostic de colinéarité Belsley-Kuh-Welsch (BKW):

proportions de la variance

lambda      cond      const      Secteur Effectifs type_Eta~ Niveaude~ sq_Effec~
4,007      1,000      0,001      0,012      0,002      0,012      0,001      0,003
0,948      2,076      0,000      0,304      0,005      0,069      0,000      0,024
0,712      2,396      0,002      0,579      0,001      0,053      0,003      0,018
0,217      4,344      0,013      0,102      0,000      0,796      0,018      0,050
0,024      12,964      0,000      0,000      0,007      0,045      0,175      0,737
0,011      19,251      0,983      0,002      0,186      0,025      0,799      0,169

lambda = valeurs propres de la matrice de covariance inverse (smallest is 0,0110289)
cond = indice de condition
note: sur une colonne, la somme des proportions de variance est égale à 1,0

Selon BKW, il y a dépendance quasi-linéaire « forte » si cond >= 30,
et « moyennement forte » si cond varie de 10 à 30. Les estimateurs dont
la variance est principalement associée à des valeurs de cond problématiques
sont eux aussi à considérer comme problématiques.

Nombre d'indices de condition >= 30: 0
Nombre d'indices de condition >= 10: 2
Proportions de variance >= 0.5 associées à cond >= 10:

const Effectifs Niveaude~ sq_Effec~
0,984 0,993 0,978 0,906
```

13. les VIF et le diagnostic de Belsley-Kuh-Welsch (BKW), avec Gretl

3. Meilleur Modèle :

Nous avons mis en place différents modèles pour trouver le meilleur, celui qui expliquerait au mieux les variations de l'IPS.

Après plusieurs essais :

- Le modèle niveau-niveau sans l'effectif et remplacer par l'effectif au carré (R^2 : 50.35%)
- IPS en log (R^2 : 47.60 %)
- IPS en log avec effectif au carré (R^2 : 47.78%)
- Toujours l'IPS en log et les variables quantitatives en log (R^2 : 49.55 %)

Nous avons trouvé que notre meilleur modèle était le modèle niveau-niveau avec le terme quadratique. Ce modèle représente seulement 50% de la variation de l'IPS mais c'est le meilleur modèle que nous avons trouvé avec nos variables . ([ici](#))

Conclusion

Pour conclure, nous avons mis en place différents modèles afin de comprendre les relations entre nos quatre variables explicatives : *Secteur*, *Effectifs*, *Type_Etablissement* et *NiveauDeVieCommune*, et leur influence sur l'IPS.

Nous avons commencé par le modèle le plus simple, de type niveau-niveau, pour structurer nos tests et analyses. Cependant, nous avons constaté que le paramètre "Effectifs" ne correspondait pas à notre intuition initiale. Pour remédier à ce problème, nous avons introduit un terme quadratique Effectifs^2 dans notre modèle. Ce choix est justifié par l'hypothèse qu'un effet non linéaire des Effectifs pourrait mieux refléter leur influence sur l'IPS. Bien que l'ajout de ce terme n'ait amélioré que légèrement le R^2 ajusté, il a permis une meilleure interprétation de l'effet des Effectifs, en capturant notamment des déséconomies d'échelle potentielles.


Par la suite, nous avons testé des modèles alternatifs avec des paramètres en log afin d'améliorer le R^2 ajusté et de mieux comprendre les effets des variables. Malheureusement, aucun de ces modèles n'a surpassé le modèle niveau-niveau avec le terme quadratique en termes de qualité d'ajustement et d'interprétation. Ainsi, nous concluons que ce modèle est le plus adapté pour répondre à notre question, même s'il explique seulement 50 % de la variation de l'IPS. Cela souligne la complexité de l'IPS et l'importance de compléter le modèle avec d'autres variables explicatives.

Pour aller plus loin, nous suggérons d'ajouter des variables explicatives supplémentaires, comme le nombre d'élèves par enseignant, le salaire des parents ou encore la situation familiale. Ces ajouts permettraient d'améliorer la spécification du modèle et d'expliquer davantage les variations de l'IPS. De plus, en ajoutant le calcul de l'IPS sur plusieurs années (*JPanel*) on pourrait offrir une meilleure compréhension des dynamiques temporelles et de l'évolution de l'IPS au fil du temps. Cela pourrait nous permettre en tant qu'économistes de fournir des conclusions sur les disparités sociales et éducatives des communes, ce qui leur permettrait de mettre en place des politiques et des stratégies adaptées.

Pour répondre à notre question, nous pouvons donc dire que les caractéristiques des écoles et le niveau de vie de la commune expliquent seulement la moitié des variations de l'IPS. Mais quelles sont les variables qui influencent l'autre moitié ?

Bibliographie et sources :

❖ Sources de données

- *Indices de position sociale des écoles (à partir de 2022)*. (2024, 30 septembre).
https://data.education.gouv.fr/explore/dataset/fr-en-ips-ecoles-ap2022/table/?disjunctive.academie&disjunctive.code_du_departement&disjunctive.departement&disjunctive.uai&disjunctive.code_insee_de_la_commune&disjunctive.nom_de_la_commune&disjunctive.secteur
-  Niveau_de_vie_2013_a_la_commune-Global_Map_Solution.xlsx

❖ Pour aller plus loin

- Morin, V. (2023, 18 janvier). L'IPS, un outil statistique particulièrement fin pour mesurer le profil social. *Le Monde.fr*.
https://www.lemonde.fr/societe/article/2022/12/17/l-ips-un-outil-statistique-particulierement-fin-pour-mesurer-le-profil-social_6154850_3224.html

❖ Logiciels :

- *Gretl* : Logiciel pour l'analyse économétrique.
- *Jupyter Notebook* : Environnement interactif utilisé pour l'analyse de données, la visualisation et l'exécution de scripts *Python*.