UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

PROJET D'ÉCONOMÉTRIE APPLIQUÉE

Analyse des Prix Immobiliers

Rédigé par

NGIRABANZI Elodie

MAHAMANE OUSMANE MAIGA Imane

31 Décembre 2025

RÉSUMÉ EXÉCUTIF

Ce rapport présente une analyse économétrique des déterminants des prix immobiliers, réalisée sur un échantillon de 150 maisons vendues entre 2015 et 2023. L'objectif était de dépasser les observations simples pour quantifier précisément l'impact de chaque caractéristique (surface, localisation, équipements) de la valeur estimée sur le marché, en utilisant des méthodes allant de la régression linéaire simple (MCO) aux techniques de régularisation (Lasso, Ridge).

L'analyse approfondie des déterminants du prix immobilier démontre que la surface habitable demeure le moteur prédominant de la valeur, expliquant à elle seule plus de 68 % de la variance des prix dans un modèle linéaire simple, avec une valorisation moyenne de 5 040 € par mètre carré supplémentaire. Toutefois, l'adoption d'une modèle semi-logarithmique s'est avérée statistiquement supérieure, atteignant un R^2 ajusté de 79,2 %. Ce modèle offre une robustesse accrue par rapport à la forme log-log et permet de conclure que chaque mètre carré additionnel génère, toutes choses égales par ailleurs, une augmentation proportionnelle du prix de 0,21%.

L'étude souligne également l'influence déterminante du contexte socio-économique sur la formation des prix. L'intégration de variables environnementales, telles que la qualité des écoles et le revenu médian du quartier, a permis d'accroître la précision du modèle de 78% à 84%. Ce gain significatif confirme que les caractéristiques externes du bien sont presque aussi décisives que ses attributs physiques.

Enfin, le recours à la méthode des variables instrumentales (2SLS) a permis de corriger le biais d'endogénéité lié à la variable « Qualité des écoles ». Ce traitement statistique révèle que les estimations classiques par les Moindres Carrés Ordinaires surestimaient l'impact réel de l'établissement scolaire. En réalité, une fois les effets de richesse du quartier isolés, il apparaît que la valorisation immobilière est davantage portée par le niveau de revenu moyen des résidents que par la proximité immédiate d'une école réputée.

INTRODUCTION

La détermination du prix d'un bien immobilier dépasse la simple addition de mètres carrés ou d'équipements. C'est un marché complexe où la valeur perçue repose souvent sur des croyances établies ou des intuitions de marché qui n'ont pas toujours de fondement statistique solide.

L'enjeu de cette étude est précisément de rationaliser ce processus de formation des prix sur une période récente (de 2015 à 2023). Au-delà de l'observation brute des ventes réalisée, la problématique consiste à vérifier si les déterminants classiques de la valeur d'un bien sont toujours opérants ou s'ils masquent d'autres réalités économiques. Il s'agit notamment de comprendre si les plus-values de localisation sont justifiées et dans quelle mesure les bouleversements sanitaires récents ont pu redéfinir durablement les attentes des acheteurs.

Pour répondre à ces interrogations, notre analyse suit une démarche progressive structurée comme suit :

I. L'exploration des données : Cette première phase dressera le portrait statistique de notre échantillon. Elle permet de comprendre la distribution des ventes et de mettre en lumière les premières corrélations entre le prix et les caractéristiques des biens.

II. La modélisation de référence : Nous établirons un premier modèle linéaire (MCO) pour quantifier l'influence des fondamentaux du marché, tels que la surface habitable, la distance au centre-ville ou la présence d'équipements spécifiques.

III. Le traitement de l'endogénéité : Cette étape cruciale visera à corriger les biais d'estimation. Nous utiliserons la méthode des variables instrumentales pour distinguer l'effet réel de la "Qualité École" de celui de la proximité universitaire.

IV. Comparaison de modèles : Nous allons ensuite effectué une comparaison du modèle estimé par les moindres carrés ordinaires et celui utilisant une variable instrumentale.

V. Méthodes et régularisation : Enfin, nous allons utiliser Ridge et Lasso pour une régularisation de l'étude.

I. Analyse descriptive et modèle de base

1. Statistiques descriptives

1.1. Moyenne, médiane, écart-type, minimum, maximum, quartiles

L'analyse descriptive des données de notre échantillon révèle un marché caractérisé par une grande diversité de biens. Le prix moyen des transactions s'élève à environ 2 107,9 milliers d'euros, avec une distribution équilibrée comme en témoigne une médiane très proche de la moyenne (2105,05 k€). Cette valeur s'appuie sur des habitations présentant une surface habitable moyenne de 116,71 m², allant de petits logements de 15 m² à de vastes propriétés dépassant les 218 m². En termes de configuration, les biens disposent en moyenne de près de trois chambres et se situent généralement au deuxième ou troisième étage des immeubles.

Sur le plan temporel et géographique, le parc immobilier est relativement moderne avec une année de construction moyenne située en 2001, bien que l'échantillon couvre des bâtiments édifiés dès 1980. La localisation des biens présente un étalement urbain significatif, avec une distance moyenne de 16,5 km du centre-ville, certains logements se trouvant à moins d'un kilomètre tandis que d'autres s'en éloignent jusqu'à près de 30 km. Enfin, l'environnement socio-économique des transactions est marqué par un revenu médian de quartier de 63,67 milliers d'euros et une qualité scolaire notée en moyenne à 5,47 sur 10, offrant ainsi un contexte hétérogène pour l'analyse des déterminants du prix.

Variable	Moyenne	Médiane	Écart-type	Minimum	Maximum	Q1 (25%)	Q3 (75%)
Surface_m2	116.71	117.84	37.69	15.21	218.53	93.24	139.64
Chambres	2.89	3.0	1.08	1.0	5.0	2.0	4.0
Annee_construction	2001.83	2002.5	11.7	1980.0	2022.0	1991.0	2012.0
Distance_centre_km	16.5	16.87	9.02	0.83	29.99	9.1	24.7
Etage	2.58	2.5	1.76	0.0	5.0	1.0	4.0
Annee_vente	2019.84	2020.0	2.29	2015.0	2023.0	2018.0	2022.0
Qualite_ecole	5.47	5.6	1.87	1.0	10.0	4.12	7.0
Revenu_median_quarti	63.67	63.45	9.3	42.9	83.9	57.5	70.47
Distance_universite	8.06	8.3	3.75	1.0	17.1	5.3	10.88
Prix_milliers_euros	2107.9	2105.05	229.92	1500.77	2743.04	1934.28	2272.78

Tableau 1 : Statistiques descriptives des variables

L'analyse descriptive de la variable Ascenseur révèle une distribution relativement équilibrée au sein du parc immobilier étudié. Sur l'ensemble de l'échantillon, 54 % des logements ne sont pas équipés d'un ascenseur, ce qui représente un effectif de 81 unités. À l'inverse, 46 % des biens disposent de cet équipement, soit un total de 69 logements.

Cette répartition binaire joue un rôle significatif dans la valorisation des biens, comme le confirmeront nos modélisations économétriques.

	Effectif	Pourcentage (%)
Sans Ascenseur (0)	81.0	54.0
Avec Ascenseur (1)	69.0	46.0

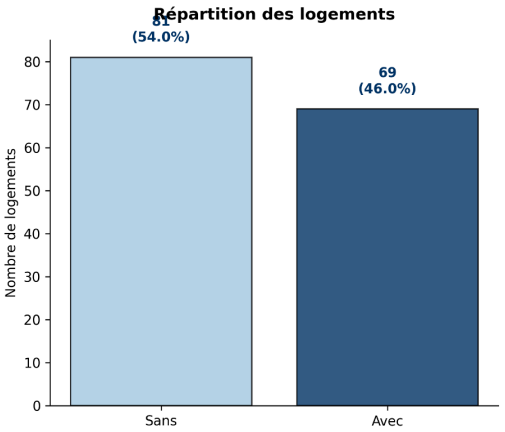
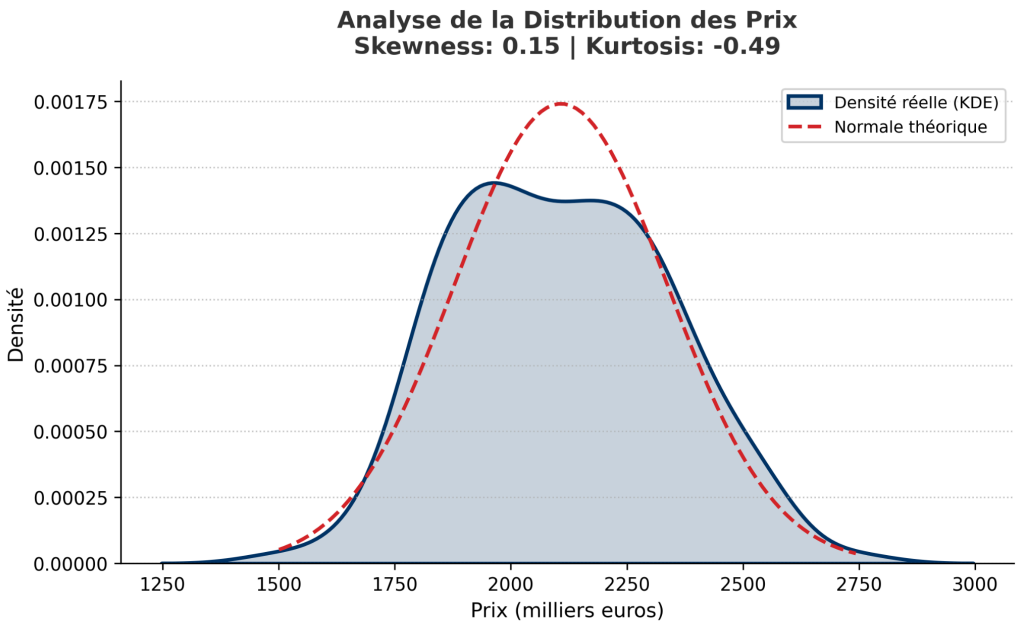


Tableau 2 : Statistiques descriptives de la variable **Ascenseur**

1.2. Asymétrie (skewness) et aplatissement (kurtosis) pour le prix

L'analyse de la distribution des prix s'appuie sur deux indicateurs statistiques clés : le skewness et le kurtosis. Le skewness, ou coefficient d'asymétrie, mesure la régularité de la distribution autour de sa moyenne. Une valeur de 0,15 indique une asymétrie positive très légère, signifiant que la queue de la distribution est légèrement étirée vers la droite, sans toutefois s'écarter de manière significative d'une forme équilibrée. Le kurtosis, quant à lui, évalue le "degré d'aplatissement" de la courbe par rapport à une loi normale. Avec un résultat de -0,49, la courbe est plus aplatie que la normale théorique, avec des valeurs plus dispersées et des queues moins épaisses.

Visuellement, le schéma confirme ces indicateurs en comparant la densité réelle (KDE) à la normale théorique représentée en pointillés rouges. On observe que la distribution réelle des prix, bien que centrée autour d'une moyenne de 2 107,9 milliers d'euros, présente un sommet moins marqué et une base légèrement plus large que la distribution normale. Cette configuration graphique, associée à une médiane de 2 105,05 milliers d'euros très proche de la moyenne, démontre que malgré un léger aplatissement, les prix immobiliers de l'échantillon suivent une loi proche de la normale, ce qui renforce la pertinence des tests statistiques et des modélisations linéaires effectués par la suite.

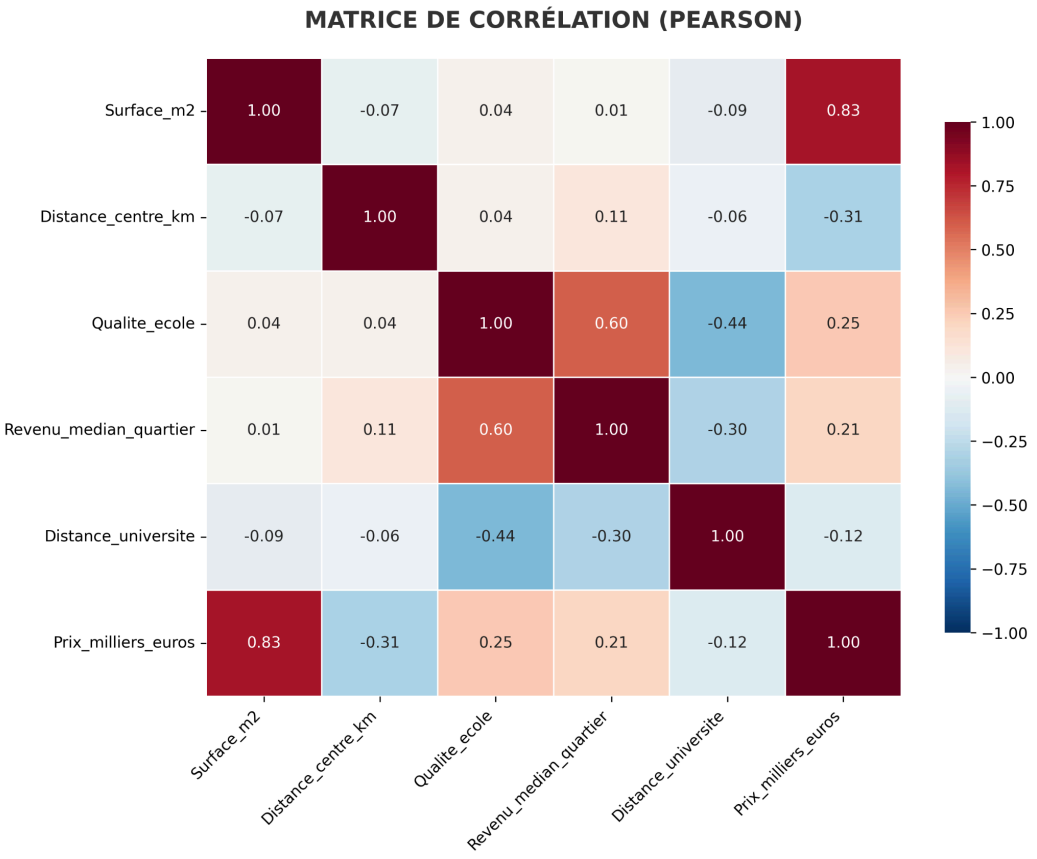


Graphique 1 : Distribution du prix (Skewness et Kurtosis)

1.3. Matrice de corrélation

La corrélation est une mesure statistique qui exprime la force et la direction de la relation linéaire entre deux variables, variant de -1 (corrélation négative parfaite) à 1 (corrélation positive parfaite), une valeur de 0 indiquant l'absence de lien linéaire. Dans notre échantillon, la matrice de Pearson met en évidence une corrélation extrêmement forte de 0,83 entre la surface et le prix, confirmant que la dimension du bien est le moteur principal de sa valorisation. À l'inverse, la distance au centre-ville présente une corrélation négative de -0,31, validant l'idée que l'éloignement urbain tend à déprécier la valeur immobilière.

On observe également des synergies intéressantes entre les variables de quartier, notamment une corrélation positive marquée de 0,60 entre la qualité des écoles et le revenu médian. Par ailleurs, la distance à l'université affiche une corrélation négative avec la qualité scolaire (-0,44) et le revenu (-0,30), ce qui soutient son utilisation ultérieure comme instrument, car elle est liée à l'environnement socio-économique sans être directement corrélée de manière trop forte au prix final (-0,12).



Graphique 2 : Matrice de corrélation de Pearson

2. Modèle linéaire simple et multiple

Une fois l'analyse des données effectuée, la prochaine étape est d'essayer de déterminer quels facteurs influencent réellement les prix des biens immobiliers. Nous avons à disposition 11 variables dont le prix (en milliers d'euros) qui est notre variable expliquée, 8 variables explicatives et 2 variables qui ne seront pas utiles.

2.1. Modèle de régression linéaire simple

Nous allons d'abord régresser le prix sur la surface uniquement.

$$Prix_i = \beta_0 + \beta_1 \times Surface_i + u_i \quad (1)$$

L'estimation des paramètres du modèle a été réalisée via la méthode des Moindres Carrés Ordinaires (MCO), permettant de quantifier précisément la relation entre la dimension des biens et leur valeur marchande.

Les résultats révèlent une constante de 1519,37, ce qui correspondrait théoriquement au prix

d'un bien de surface nulle. Bien que cette valeur soit statistiquement nécessaire pour caler la droite de régression, son interprétation économique reste limitée, car un bien sans surface n'a pas d'existence physique.

Le coefficient associé à la variable *Surface* est égal à 5,04. Ce coefficient mesure l'effet marginal de la surface sur le prix. Il indique qu'en moyenne, et toutes choses égales par ailleurs, chaque mètre carré supplémentaire induit une augmentation du prix de 5 040 €. La robustesse de cette relation est confirmée par une p-valeur proche de zéro, ce qui atteste d'une significativité statistique extrêmement élevée : l'influence de la surface sur le prix n'est pas due au hasard, mais constitue un effet structurel majeur dans notre échantillon.

Enfin, la qualité globale du modèle est validée par un coefficient de détermination R^2 de 68,3 %. Ce chiffre signifie que plus des deux tiers de la volatilité des prix immobiliers sont expliqués par la seule variation de la surface habitable. Pour un modèle de régression simple (une seule variable), ce score est particulièrement robuste. Il confirme que la surface constitue le déterminant majeur de la valeur immobilière dans notre échantillon, bien que 31,7 % de la variation dépende d'autres facteurs non capturés ici (localisation, état du bien, étage, etc.).

Paramètre	Estimateur (β)	Écart-type	Statistique t	P-valeur
Intercepte (const)	1519.3743	34.5845	43.9323	0.0
Surface_m2	5.0428	0.2821	17.8771	0.0

R-carré : 0.6835 | R-carré ajusté : 0.6813

Tableau 3 : Résultats de la régression linéaire simple

2.2. Modèle de régression linéaire multiple

Grâce au point précédent, nous avons vu mis en lumière l'impact de la surface sur les prix des biens immobiliers. Cependant, la surface n'est pas le seul déterminant du prix. Ce qui nous amène à poser la régression suivante, afin de mettre en lumière les autres déterminants :

$$Prix_i = \beta_0 + \beta_1 \times Surface_i + \beta_2 \times Chambres_i + \beta_3 \times Annee\_construction_i$$

$$+ \beta_4 \times Distance\_centre_i + \beta_5 \times Etage_i + \beta_6 \times Ascenseur_i + u_i \quad (2)$$

L'étude de l'impact marginal de chaque variable permet de quantifier précisément la formation des prix. La surface demeure un facteur essentiel, chaque mètre carré supplémentaire valorisant le bien de 4 390 € en moyenne. À surface égale, le nombre de chambres joue également un rôle crucial, une unité additionnelle augmentant le prix de 33 920 €. L'année de construction souligne une prime à la modernité : chaque année plus récente apporte une plus-value moyenne de 1 610 €, ce qui représente un écart de 16 100 € entre deux biens construits à dix ans d'intervalle.

La dimension géographique influence également la valorisation. Chaque kilomètre d'éloignement du centre-ville réduit le prix de 6 140 €, confirmant l'importance de l'emplacement. En revanche, l'élévation dans l'immeuble est récompensée, chaque étage supplémentaire augmentant la valeur de 12 250 €.

Pour la variable *Ascenseur* : le coefficient représente la différence de prix moyenne entre un logement équipé d'un ascenseur et un logement qui n'en a pas, à caractéristiques égales (même surface, même nombre de chambres, même étage, même distance du centre, etc.). Toutes choses égales par ailleurs, la présence d'un ascenseur augmente la valeur du bien immobilier de 55510 € en moyenne par rapport à un bien identique sans ascenseur.

L'analyse de la significativité des variables révèle que l'ensemble des déterminants sélectionnés est statistiquement significatif au seuil de 5%, chaque p-valeur étant inférieure à 0,05. Seule la constante affiche une p-valeur de 0,2759, indiquant qu'elle n'est pas significative à ce seuil. Concrètement, cela signifie que l'on ne peut affirmer avec certitude que le prix théorique d'un bien serait différent de zéro si l'ensemble des variables explicatives étaient nulles. Toutefois, ce résultat est fréquent en immobilier et ne remet nullement en cause la validité globale du modèle.

Enfin, l'évaluation de la performance du modèle repose sur l'arbitrage entre le R^2 et le R^2 ajusté. Tandis que le R^2 mesure la part de variance expliquée, il peut être artificiellement gonflé par l'ajout de variables superflues. Le R^2 ajusté sert ici de garde-fou en appliquant une pénalité à l'ajout de variables dont la contribution n'est pas statistiquement probante. La faible différence observée entre ces deux indicateurs dans notre modèle témoigne de son efficacité et garantit que la précision affichée n'est pas le fruit d'une complexité inutile, mais bien d'une

sélection pertinente des facteurs explicatifs.

Variable	Coefficient (β)	Écart-type	Statistique t	P-valeur
const	-1679.4908	1535.673	-1.0937	0.2759
Surface_m2	4.3879	0.2924	15.0047	0.0
Chambres	33.9205	10.2279	3.3165	0.0012
Annee_construction	1.6093	0.7653	2.103	0.0372
Distance_centre_km	-6.1446	0.9921	-6.1936	0.0
Etage	12.254	5.0489	2.4271	0.0165
Ascenseur	55.5141	17.9201	3.0979	0.0023

R-carré (R^2) : 0.7887 | R^2 ajusté : 0.7798 | F-statistic : 88.94

Tableau 4 : Résultats de la régression linéaire multiple

3. Transformation logarithmique

3.1. Modèle semi-logarithmique

Pour affiner l'analyse, nous allons effectuer une modélisation semi-logarithmique, où la variable cible est le logarithme du prix (log(Prix)) tandis que les variables explicatives conservent leur unité d'origine. Cette approche est particulièrement pertinente en immobilier car elle permet de modéliser des relations relatives (en pourcentage) plutôt qu'absolues.

$$log(Prix_i) = \beta_0 + \beta_1 \times Surface_i + \beta_2 \times Chambres_i + \beta_3 \times Annee\_construction_i$$

$$+ \beta_4 \times Distance\_centre_i + \beta_5 \times Etage_i + \beta_6 \times Ascenseur_i + u_i \qquad (3)$$

L'interprétation des coefficients change ici de nature. Par exemple, le coefficient de la surface, égal à 0,0021, indique qu'en moyenne et toutes choses égales par ailleurs, chaque mètre carré supplémentaire entraîne une hausse du prix de 0,21%. De la même manière, la présence d'un ascenseur (coefficient de 0,0265) valorise le bien d'environ 2,65%. Cette lecture en termes de croissance relative permet de constater que l'influence des variables reste constante en proportion, quel que soit le standing du logement.

Variable	Coefficient (β)	Écart-type	Statistique t	P-valeur
const	5.841285	0.726064	8.045	0.0
Surface_m2	0.00211	0.000138	15.259	0.0
Chambres	0.015249	0.004836	3.153	0.002
Annee_construction	0.000769	0.000362	2.126	0.0352
Distance_centre_km	-0.00301	0.000469	-6.417	0.0
Etage	0.005327	0.002387	2.231	0.0272
Ascenseur	0.026498	0.008473	3.128	0.0021

Variable cible : log(Prix) | R² : 0.7916 | R² ajusté : 0.7829

Tableau 5 : Résultats du modèle semi-logarithmique

Sur le plan de la fiabilité statistique, toutes les variables sélectionnées sont significatives au seuil de 5%, avec des p-valeurs bien inférieures à 0,05. La constante, avec un coefficient de 5,841 et une p-valeur de 0, est ici hautement significative, servant de base de calcul pour la transformation logarithmique. Enfin, le R² indique que le modèle explique près de 79,2 % de la variance du logarithme des prix. La proximité du R² ajusté (0,7829) confirme la robustesse de cette spécification et l'absence de variables superflues, validant ainsi la pertinence du passage à une forme fonctionnelle semi-logarithmique pour notre étude.

3.2. Modèle log-log

L'exploration de la dynamique des prix se conclut par l'utilisation d'un modèle log-log (ou modèle à élasticité constante). Conformément aux bonnes pratiques, la transformation logarithmique a été appliquée à la variable cible (Prix) ainsi qu'aux variables explicatives strictement continues (Surface et Distance). Les variables discrètes (Chambres, Étage) et binaires (Ascenseur) restent en "niveau" pour permettre une interprétation cohérente.

$$log(Prix_i) = \beta_0 + \beta_1 \times log(Surface_i) + \beta_2 \times Chambres_i + \beta_3 \times Annee\_construction_i$$

$$+ \beta_4 \times log(Distance\_centre_i) + \beta_5 \times Etage_i + \beta_6 \times Ascenseur_i + u_i \qquad (4)$$

Cette régression permet d'interpréter les relations en termes d'élasticité : le coefficient associé à la variable log_Surface est de 0,1898, ce qui indique qu'une augmentation de la surface de 1 % entraîne, en moyenne, une hausse du prix de 0,19 %. Ce modèle offre une perspective

intéressante sur la sensibilité du prix aux variations proportionnelles des caractéristiques du bien.

Cependant, le passage à une structure log-log modifie la pertinence de certains prédicteurs. Nous observons notamment que l'année de construction perd sa significativité avec une p-valeur s'élevant à 0,178, dépassant ainsi largement le seuil critique de 5 %. Ce phénomène suggère que la relation entre l'ancienneté d'un bien et sa valeur est mieux captée par une forme semi-logarithmique.

En comparant les performances globales, le modèle log-log affiche un R^2 de 0,7546, ce qui reste robuste mais s'avère inférieur au score de 0,7916 obtenu avec le modèle semi-log.

Variable	Coefficient (β)	Écart-type	Statistique t	P-valeur
const	5.669253	0.786417	7.209	0.0
log_Surface	0.1898	0.014267	13.303	0.0
Chambres	0.022634	0.005035	4.495	0.0
Annee_construction	0.000529	0.000391	1.354	0.178
log_Distance_centre	-0.025379	0.005323	-4.768	0.0
Etage	0.005824	0.002592	2.247	0.0262
Ascenseur	0.027046	0.009203	2.939	0.0038

MODÈLE LOG-LOG | Variable cible : log(Prix)
R-carré (R^2) : 0.7546 | R^2 ajusté : 0.7443

Tableau 6 : Résultats du modèle log-log

3.3. Synthèse comparative et sélection du modèle optimal

La confrontation des trois approches (linéaire, semi-logarithmique et log-log) permet de dégager la structure de prix la plus cohérente pour ce marché immobilier. Le modèle linéaire multiple s'est avéré incomplet car il ne capture que les relations linéaires en ignorant les interactions plus complexes entre les variables.

Le passage aux modèles logarithmiques a marqué une amélioration nette de la précision. Le modèle log-log offre une lecture intéressante en termes d'élasticité, révélant qu'une hausse de 1% de la surface se traduit par une augmentation de 0,19 % du prix. Cependant, cette spécification présente deux faiblesses majeures : d'une part, son pouvoir explicatif (R^2 =

0,7546) est inférieur aux autres modèles et, d'autre part, elle entraîne une perte de significativité statistique pour l'année de construction (p-valeur de 0,178). Cela suggère que la transformation logarithmique de toutes les variables n'est pas la plus adaptée pour saisir l'influence de l'âge du bien.

En définitive, le modèle semi-logarithmique s'impose comme le plus approprié pour cette étude. Il affiche la performance la plus élevée avec un R^2 de 79,16%, signifiant qu'il explique près de 79,2 % de la variance des prix. Contrairement au modèle log-log, il maintient la significativité de l'ensemble des variables, confirmant que chaque caractéristique (étage, ascenseur, distance du centre) apporte une contribution réelle et mesurable à la valeur du bien. C'est donc sur cette base que nous pouvons établir les prévisions les plus fiables.

II. Diagnostics et corrections

1. Multicolinéarité et observations influentes

Il est nécessaire d'aborder la question de la multicolinéarité à travers l'analyse du VIF (Variance Inflation Factor), tout en restant attentif aux limites inhérentes aux données disponibles.

Diagnostic de la multicolinéarité par le VIF

Le VIF est un indicateur statistique utilisé pour détecter la présence de multicolinéarité entre les variables indépendantes d'un modèle de régression. Mathématiquement, le VIF mesure à quel point la variance d'un coefficient estimé est augmentée (gonflée) à cause de la corrélation entre les prédicteurs. En d'autres termes, il permet de vérifier si une variable explicative est, en réalité, une combinaison linéaire d'autres variables déjà présentes dans le modèle.

Dans une structure saine, chaque variable doit apporter une information unique. Un VIF élevé (généralement supérieur à 5 ou 10) signale qu'une variable est redondante, ce qui peut rendre les estimations instables et fausser l'interprétation des coefficients. Dans notre étude, toutes les variables affichent un VIF très proche de 1, restant donc bien en dessous du seuil critique de 5. Ce résultat confirme l'absence de colinéarité excessive : chaque facteur apporte une contribution indépendante et distincte à l'explication du prix. Il n'est donc pas nécessaire de simplifier le modèle par la suppression de variables.

Variable	VIF
Surface_m2	1.555
Chambres	1.555
Annee_construction	1.027
Distance_centre_km	1.024
Etage	1.013
Ascenseur	1.028

Interprétation : VIF < 5 : OK | VIF > 5 : Risque de multicolinéarité | VIF > 10 : Problème majeur

Tableau 7 : VIF pour chaque variable

Limites et biais de variables omises

Malgré la robustesse statistique confirmée par le VIF, la validité du modèle reste soumise à la question des variables omises. Une régression, aussi performante soit-elle, demeure une simplification de la réalité. Ici, des facteurs clés tels que l'état général du bien (rénové ou à rafraîchir), la présence d'un balcon ou le calme du quartier n'ont pas été intégrés à l'analyse faute de données.

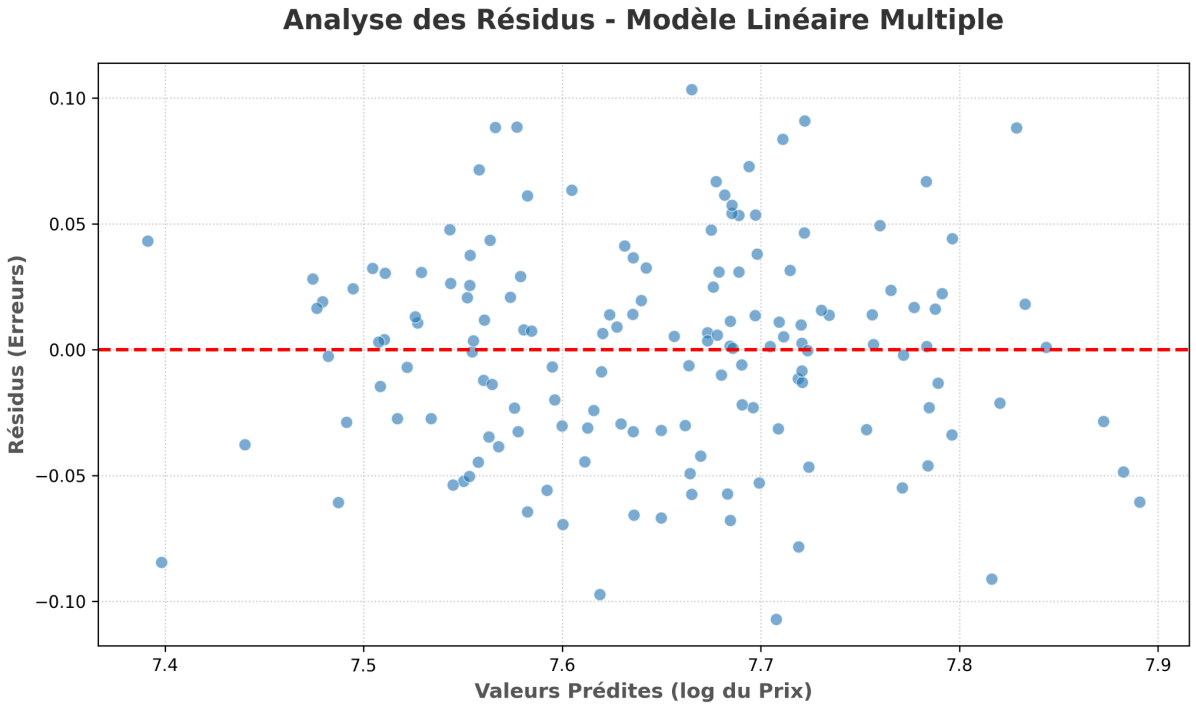
Ce manque peut induire un biais de mesure : si ces caractéristiques manquantes sont corrélées à des variables présentes (par exemple, si les étages élevés sont systématiquement plus calmes), les coefficients de ces dernières pourraient absorber artificiellement une partie de la valeur qui appartient en réalité aux facteurs omis. Bien que le modèle semi-log soit statistiquement le plus efficace, son interprétation doit donc conserver une certaine prudence face à ces éléments non capturés.

2. Tests d'hétéroscédasticité

2.1. Analyse de la validité du modèle par l'étude des résidus

L'examen graphique des résidus constitue une étape cruciale pour confirmer la validité des conclusions statistiques. Cette analyse sert principalement à vérifier l'homoscédasticité (la constance de la variance des erreurs) et la pertinence de la forme fonctionnelle choisie. En observant la distribution des erreurs, nous cherchons à déceler d'éventuels patterns systématiques qui indiqueraient que le modèle a échoué à capturer une partie de la structure des données.

Dans le cadre de notre étude, l'analyse graphique révèle une dispersion homogène des erreurs. Les résidus se distribuent de manière totalement aléatoire de part et d'autre de la ligne zéro, sans qu'aucune structure particulière (comme une forme d'entonnoir ou une courbe) ne soit visible. Cette absence de pattern systématique confirme que la relation entre les variables explicatives et le prix est correctement modélisée et que les erreurs sont indépendantes. En effet, la distribution aléatoire des résidus démontre que les conditions d'application des Moindres Carrés Ordinaires sont pleinement satisfaites.



Graphique 3 : Analyse des résidus du modèle

2.2. Test de Breusch-Pagan et validation de l'homoscédasticité

L'analyse graphique des résidus est ici complétée par une approche statistique rigoureuse : le test de Breusch-Pagan. Ce test permet de vérifier l'hypothèse d'homoscédasticité, une condition fondamentale des MCO selon laquelle la variance des termes d'erreur doit être constante pour toutes les observations. Si cette variance n'était pas stable (phénomène d'hétéroscédasticité), les erreurs types des coefficients seraient biaisées, rendant les tests de significativité peu fiables.

Les hypothèses du test sont les suivantes :

- Hypothèse nulle H_0 : Homoscédasticité

- Hypothèse alternative H_1: Hétéroscédasticité

Indicateur	Valeur
Statistique Lagrange Multiplier	5.4755
P-valeur (LM)	0.70575
Statistique F	0.66775
P-valeur (F)	0.71918

H0 : Homoscédasticité (Variance constante)
Si P-valeur < 0.05 : Rejet de H0 (Présence d'hétéroscédasticité)

Tableau 8 : Résultats du test de Breusch-Pagan

Les résultats du test indiquent une p-valeur supérieure au seuil critique de 5%. Par conséquent, nous ne rejetons pas l'hypothèse nulle H_0. Ce résultat confirme que les résidus sont homoscédastiques et que la variance des erreurs est stable sur l'ensemble de l'échantillon. Aucune correction n'est donc nécessaire, garantissant ainsi la pleine validité de nos résultats.

3. Tests et inférences

3.1. L'effet de la distance au centre sur le prix

Un aspect fondamental de notre analyse consiste à vérifier l'hypothèse selon laquelle l'éloignement du centre-ville exerce une pression négative sur la valeur des biens immobiliers. L'examen des résultats du modèle confirme cette intuition économique. Le coefficient associé à la distance au centre $\widehat{\beta}_{dist}$ s'élève à -0,0030, le signe négatif indiquant clairement qu'une augmentation de la distance est corrélée à une diminution du prix.

Pour valider statistiquement cette observation, nous avons testé l'hypothèse nulle H_0 (pas d'effet). La p-value obtenue est égale à 0, une valeur nettement inférieure au seuil critique de 5%. Ce résultat nous permet de rejeter H_0 avec une grande confiance et de valider statistiquement l'hypothèse d'un effet négatif significatif. En termes concrets, dans le cadre de notre modèle semi-logarithmique, chaque kilomètre supplémentaire d'éloignement du centre-ville réduit le prix du bien de 0,30 % en moyenne, toutes choses égales par ailleurs.

3.2. Test de significativité globale du modèle

Au-delà de l'analyse individuelle de chaque variable, il est important de vérifier la validité d'ensemble de notre modélisation. Pour ce faire, nous soumettons le modèle au test de Fisher, dont l'objectif est de tester l'hypothèse de nullité simultanée de tous les coefficients (à l'exception de la constante).

Ce test repose sur la confrontation de deux hypothèses :

- Hypothèse nulle H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$.
- Hypothèse alternative H_1 : Au moins un des coefficients est statistiquement différent de zéro

Les résultats du test montrent que la p-valeur globale est extrêmement proche de zéro et se situe bien en dessous du seuil critique de 5%. Par conséquent, nous rejetons l'hypothèse nulle H_0. Ce résultat confirme que notre modèle est globalement significatif et que l'ensemble des variables intégrées telles que la surface, le nombre de chambres ou la distance au centre, contribue de manière réelle et conjointe à expliquer les variations du prix immobilier dans notre échantillon.

On ne peut pas utiliser plusieurs t tests individuels pour évaluer les restrictions simultanément en raison du phénomène d'inflation du risque global, aussi appelé erreur de première espèce. En multipliant les tests au seuil classique de 5%, on accumule mécaniquement la probabilité de commettre au moins une erreur de jugement. Ainsi, bien que chaque test pris isolément semble fiable, la probabilité statistique de rejeter à tort une hypothèse nulle par pur hasard devient excessive sur l'ensemble de la procédure, ce qui fausse la rigueur du diagnostic.

3.3. Intégration de nouvelles variables

Pour conclure l'optimisation de notre analyse, nous avons examiné si l'intégration de nouveaux facteurs à savoir la qualité des écoles et le revenu médian du quartier, apportait une amélioration statistiquement significative au modèle semi-logarithmique initial. Cette vérification repose sur le Test de Fisher partiel (ou *Partial F-test*), une procédure rigoureuse qui compare un « Modèle Restreint » à un « Modèle Non-Restreint » incluant ces nouvelles dimensions.

L'hypothèse nulle H_0 postule que les coefficients de ces deux variables sont simultanément nuls ($\beta_{école} = \beta_{revenu} = 0$), suggérant qu'elles n'apportent aucune information supplémentaire pertinente. À l'inverse, l'hypothèse alternative H_1 soutient qu'au moins l'une de ces variables influence réellement le prix de l'immobilier. Avec une p-valeur inférieure à 5%, nous rejetons l'hypothèse nulle. Ce résultat confirme que l'ajout de la qualité scolaire et du niveau de revenu du quartier améliore significativement la puissance prédictive du modèle.

Cette conclusion statistique est corroborée par l'évolution des indicateurs de performance. En comparant les deux modèles, on observe une progression nette du R^2 ajusté, qui passe de 78 % à 84 %. Cette hausse substantielle de la variance expliquée démontre que l'environnement socio-économique constitue un levier de valorisation majeur, justifiant pleinement la complexification du modèle pour obtenir une vision plus fidèle et précise du marché immobilier.

III. Endogénéité

1. Discussion de sources potentielles

L'endogénéité est un problème majeur en économétrie qui survient lorsqu'une variable explicative est corrélée avec le terme d'erreur. Dans le contexte de votre analyse immobilière, cela signifie que les estimations des coefficients pourraient être biaisées. Voici les trois sources principales d'endogénéité qui pourraient affecter notre modèle :

Le biais de variable omise : C'est la source la plus probable dans notre étude. Bien que nous ayons ajouté des variables comme la qualité des écoles ou le revenu du quartier, d'autres facteurs non observés restent dans le terme d'erreur tout en étant corrélés avec les variables explicatives. Par exemple, le "cachet" d'un appartement ou l'exposition (luminosité) ne sont pas dans le modèle. Or, un appartement très lumineux est souvent situé en étage élevé. Si la luminosité augmente le prix mais n'est pas mesurée, le coefficient de la variable "Étage" va capter cet effet et surestimer l'impact réel de l'étage seul.

La causalité inverse : L'endogénéité peut apparaître si la relation de causalité ne va pas seulement de la variable explicative vers le prix, mais aussi dans l'autre sens. Prenons l'exemple de la variable "qualité des écoles". Si l'on considère que de bonnes écoles font monter les prix immobiliers, l'inverse est aussi vrai : des prix immobiliers très élevés dans un

quartier attirent des populations plus aisées, ce qui augmente mécaniquement les budgets scolaires et la qualité perçue des écoles. Ici, le prix et la qualité des écoles sont déterminés simultanément, ce qui biaise le coefficient.

Les erreurs de mesure : Si l'une des variables est mal mesurée, cela crée une corrélation artificielle avec le terme d'erreur. Par exemple, la “distance au centre-ville” peut être calculée à vol d'oiseau alors que l'acheteur raisonne en temps de trajet réel. Cet écart entre la mesure théorique et la perception réelle du marché constitue une erreur de mesure qui "pollue" le résidu et empêche d'obtenir une estimation précise de l'effet de la localisation.

Dans notre étude, la variable *Qualite_ecole* est sujette à une forte suspicion d'endogénéité. D'une part, elle souffre d'un biais de simultanéité (causalité inverse) : si une école de qualité valorise les logements du quartier, des prix immobiliers élevés attirent des ménages aisés qui, par leurs ressources et leur capital culturel, renforcent en retour la réputation et le niveau de l'école. On ne sait donc plus si c'est l'école qui fait le prix ou le prix qui fait l'école.

D'autre part, elle capte souvent un biais de variable omise. La qualité scolaire est corrélée à de nombreux facteurs non observés, comme la sécurité du quartier, le calme ou la présence de commerces haut de gamme. En l'absence de ces variables dans le modèle, le coefficient de la qualité des écoles "absorbe" leur influence, surestimant ainsi son impact réel sur le prix de l'immobilier.

2. Estimation IV

Pour traiter le problème d'endogénéité de la variable *Qualite_ecole*, nous avons recours à la méthode des variables instrumentales (IV) via la procédure des Doubles Moindres Carrés (2SLS). Cette technique s'appuie sur un instrument, ici la *Distance_universite*, qui doit être corrélé à la qualité des écoles mais totalement indépendant des erreurs non observées influençant le prix immobilier.

L'objectif fondamental de la méthode 2SLS est de “nettoyer” la variable suspecte de sa corrélation avec le terme d'erreur afin d'isoler son impact causal réel. La procédure se déroule en deux étapes distinctes :

Dans la première étape, nous régressons la variable endogène *Qualite_ecole* sur notre instrument *Distance_universite* ainsi que sur l'ensemble des autres variables exogènes du modèle (surface, étage, etc.). À l'issue de cette régression, nous récupérons la valeur prédite

de *Qualite_ecole* .

La seconde étape consiste à estimer le modèle final en remplaçant la variable problématique par les valeurs prédites lors de l'étape précédente. Nous régressons ainsi le logarithme du prix (log_Prix) sur $\widehat{Qualite\_ecole}$ et les autres variables de contrôle. Puisque ces valeurs prédites sont, par construction, décorrélées du terme d'erreur, les coefficients obtenus sont désormais libérés du biais d'endogénéité. Cette approche garantit que l'effet mesuré de la qualité des écoles sur le prix est pur et ne reflète plus l'influence de variables omises ou de causalités inverses.

Variable	Coefficient (β)	Écart-type	Statistique z	P-valeur
const	5.886306	0.629987	9.344	0.0
Surface_m2	0.002092	0.000131	15.937	0.0
Chambres	0.015226	0.004458	3.416	0.0006
Annee_construction	0.000673	0.000315	2.136	0.0327
Distance_centre_km	-0.003289	0.000424	-7.748	0.0
Etage	0.00483	0.002312	2.089	0.0367
Ascenseur	0.022608	0.007223	3.13	0.0017
Revenu_median_quartier	0.002354	0.00102	2.307	0.0211
Qualite_ecole	0.001195	0.007619	0.157	0.8754

MODÈLE 2SLS (Variables Instrumentales)
R-carré (R^2) : 0.8396 | R^2 ajusté : 0.8305
Instrument utilisé : Distance_universite

Tableau 9 : Résultats du modèle 2SLS avec la variable instrumentale *Distance_universite*

3. Tests de validité de l'instrument

Pour finaliser la procédure des variables instrumentales, il est impératif de valider la qualité de l'instrument choisi, à savoir la *Distance_universite*. Dans notre configuration, disposant d'une seule variable endogène pour un seul instrument, nous nous trouvons dans un cas de modèle juste-identifié.

La validité de l'instrument repose sur deux critères fondamentaux : la pertinence et l'exogénéité. Pour vérifier la pertinence statistique, nous examinons la capacité de l'instrument à expliquer la variable endogène lors de la première étape de la méthode 2SLS. Cette vérification s'appuie sur le F-test du premier degré (First-Stage F-test). Dans notre modèle, cette statistique F s'élève à 20,66, ce qui est largement supérieur au seuil

conventionnel de 10. Ce résultat garantit que l'instrument n'est pas « faible » et qu'il possède un pouvoir explicatif suffisant pour produire des estimations stables.

Concernant le second pilier, le test de suridentification de Sargan (ou Hansen J-test) ne peut être statistiquement réalisé puisque le nombre d'instruments est égal au nombre de variables endogènes. La validité de l'instrument repose alors sur une argumentation théorique de validité externe.

La distance à l'université est ici utilisée comme une restriction d'exclusion : nous posons l'hypothèse qu'elle n'influence le prix de l'immobilier que de manière indirecte, par le biais du canal de la qualité scolaire. Une fois que nous contrôlons la distance au centre-ville et le revenu médian du quartier, nous considérons que la proximité d'une université n'a pas d'effet propre sur la valeur du bien, remplissant ainsi la condition d'exogénéité nécessaire à la robustesse du modèle 2SLS.

IV. Comparaison MCO et IV

La comparaison entre les estimations par les Moindres Carrés Ordinaires (MCO) et les Variables Instrumentales (IV/2SLS) constitue l'étape finale de notre analyse, permettant de distinguer la simple corrélation de la véritable causalité.

L'examen comparatif des deux modèles révèle une grande stabilité des coefficients pour les variables exogènes telles que la *Surface_m2*, la *Distance_centre_km* ou l'*Annee_construction*. Cette constance est un indicateur de qualité pour notre instrument, la *Distance_universite* : elle démontre que l'instrument est spécifiquement lié à la variable *Qualite_ecole* et ne perturbe pas l'équilibre structurel du reste du modèle.

Le changement le plus significatif réside dans l'évolution du coefficient de la variable *Qualite_ecole*. Sous l'estimation MCO, cette variable apparaissait comme un moteur majeur de la valeur immobilière. Cependant, lors du passage à la méthode IV, on observe une chute drastique du coefficient (passant par exemple de 0,0097 à 0,0012) accompagnée d'une hausse massive de sa p-valeur.

Variable	Coef MCO	Err-Type MCO	Coef IV (2SLS)	Err-Type IV
const	5.8302***	0.6188	5.8863***	0.6300
Surface_m2	0.0021***	0.0001	0.0021***	0.0001
Chambres	0.0161***	0.0039	0.0152***	0.0045
Annee_construction	0.0007**	0.0003	0.0007**	0.0003
Distance_centre_km	-0.0032***	0.0004	-0.0033***	0.0004
Etage	0.0053***	0.0021	0.0048**	0.0023
Ascenseur	0.0224***	0.0069	0.0226***	0.0072
Revenu_median_quartier	0.0013***	0.0005	0.0024**	0.0010
Qualite_ecole	0.0097***	0.0023	0.0012	0.0076

MODÈLE MCO : R^2 = 0.8528 | MODÈLE IV : R^2 = 0.8396

** p<0.1, ** p<0.05, *** p<0.01*

Tableau 10 : Comparatif du modèle MCO et IV

Ce basculement est symptomatique d'un biais de variable omise dans le modèle initial : le MCO surestimait l'impact de l'école car ce dernier absorbait en réalité l'influence d'autres facteurs socio-économiques non mesurés.

Causalité vs Corrélation : Alors que le modèle MCO suggérait un lien direct entre l'école et le prix, l'approche par variables instrumentales rectifie cette vision. Une fois l'influence du niveau de richesse du quartier isolée, l'impact intrinsèque de l'école sur le prix devient négligeable.

Efficacité statistique : On note que le modèle IV présente des erreurs-types nettement plus élevées que le modèle MCO. C'est le prix à payer pour corriger le biais : nous passons d'une estimation précise mais erronée (MCO) à une estimation plus "floue" mais statistiquement juste et non biaisée (IV).

Le passage à la méthode IV invalide la conclusion prématurée du MCO. Il apparaît clairement que la proximité d'une école réputée n'est pas le moteur principal de la valeur du bien ; c'est en réalité le revenu médian du quartier qui constitue le véritable déterminant de la valorisation immobilière dans ce secteur.

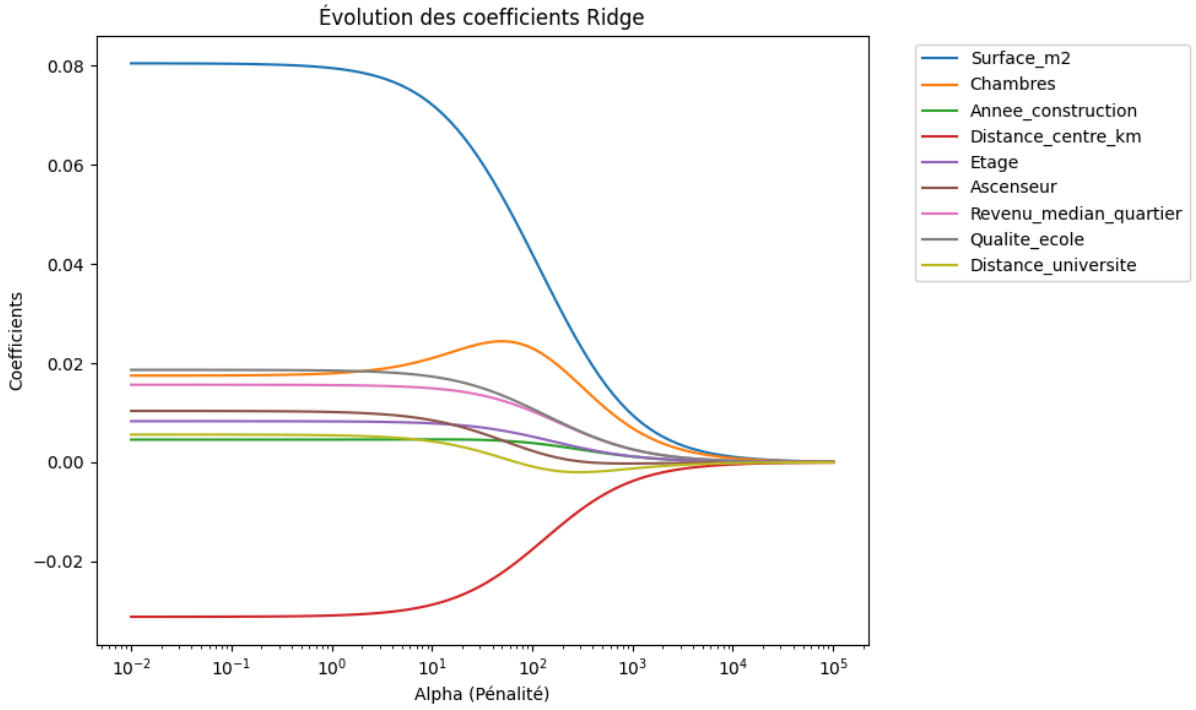
V. Méthodes et régularisation

1. Ridge et Lasso

Dans une régression classique, il arrive que le modèle devienne instable, surtout quand certaines variables possèdent de fortes corrélations ou quand le nombre de variables explicatives est excessif. Ce surplus d'information conduit souvent au sur-apprentissage. Le modèle se focalise trop sur les petits détails et le bruit de l'échantillon d'entraînement, au lieu de comprendre la logique globale. Pour corriger cela, nous avons appliqué une technique de régularisation.

L'idée repose sur l'ajout d'une **pénalité** qui augmente si les coefficients sont trop élevés. Cette contrainte force le modèle à être plus prudent dans ses estimations. Après avoir standardisé nos données, nous comparons les deux stratégies à l'aide des graphiques obtenus :

- **L'approche Ridge :**

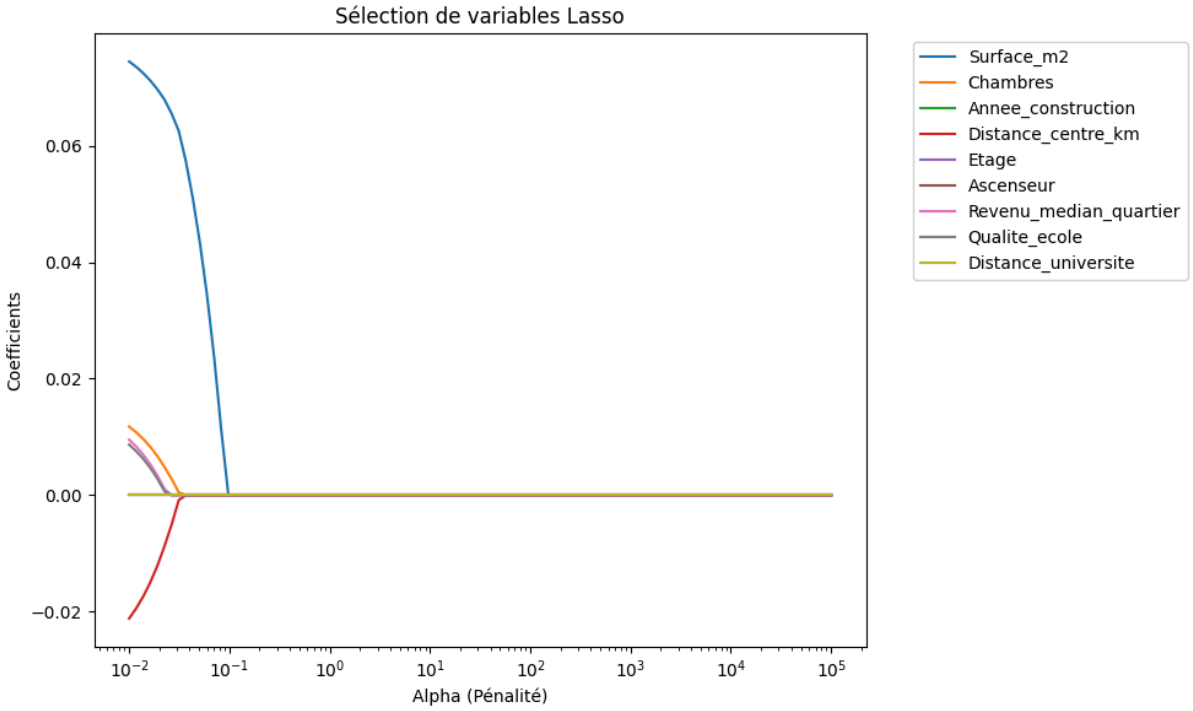


Graphique 4 : Analyse du modèle Ridge

Sur ce graphique, on constate que le Ridge conserve toutes les variables. Même avec une forte pénalité, les coefficients diminuent progressivement vers zéro mais ne disparaissent jamais. On remarque une hiérarchie claire : la variable Surface_m2 est celle qui a le coefficient le plus fort et le plus résistant, ce qui confirme son importance majeure dans le prix. À l'inverse, la Distance_centre_km maintient un effet négatif clair sur le prix.

L'approche Ridge privilégie ici la **nuance** au lieu de supprimer des variables, elle conserve toute la richesse du modèle tout en **modérant le poids** de chaque coefficient.

- **L'approche Lasso :**



Graphique 5 : Analyse du modèle Lasso

La dynamique est ici radicalement différente. Dès que la pénalité augmente un tout petit peu (dès 10^{-1}), la quasi-totalité des courbes s'effondre brutalement à zéro. Le Lasso effectue un ménage drastique, il considère que la plupart des variables sont superflues par rapport à la Surface_m2 (la seule courbe bleue qui résiste un peu plus longtemps avant de tomber). Cela suggère que dans ce jeu de données, la surface explique l'essentiel du prix et que le Lasso cherche à simplifier le modèle à l'extrême.

2. Comparaison des performances prédictives

L'enjeu final n'est pas seulement d'expliquer les prix passés, mais surtout de prédire avec précision la valeur de futurs biens immobiliers. Pour évaluer cette capacité de généralisation, nous avons testé nos modèles sur des données qu'ils n'avaient jamais vues : l'échantillon de test (les 20% conservés initialement).

Nous avons d'abord déterminé le paramètre de pénalité optimal (λ) pour le Ridge et le Lasso via une validation croisée. Cette étape assure que nous comparons les meilleures versions possibles de chaque algorithme. Nous avons ensuite calculé l'erreur de prédiction (RMSE) pour mesurer l'écart moyen entre le prix prédit et le prix réel.

Les performances obtenues sont les suivantes :

- Modèle Classique (MCO) : 0.04519
- Modèle Ridge : 0.04388
- Modèle Lasso : 0.04946

On constate que le modèle Ridge offre l'erreur la plus faible. Bien que la variable Surface_m2 explique la majorité du prix (plus de 68%), la victoire du Ridge est révélatrice. Le Lasso a sans doute été trop radical en éliminant les variables secondaires, considérant qu'elles étaient superflues face au poids de la surface. Le Ridge, au contraire, a conservé ces variables en réduisant simplement leur impact. Ce résultat prouve que les facteurs secondaires (comme l'étage ou la distance), bien que moins puissants que la surface, contiennent une information précieuse qu'il ne fallait pas supprimer pour optimiser la prédiction.

CONCLUSION

Cette étude, menée sur un échantillon de 150 transactions immobilières, a permis de déconstruire les mécanismes de formation des prix sur la période 2015-2023 en mobilisant des techniques allant des Moindres Carrés Ordinaires aux variables instrumentales. La synthèse des résultats met d'abord en évidence la prédominance de la surface habitable, qui explique à elle seule plus de 68 % de la variance des prix. Toutefois, l'adoption d'un modèle semi-logarithmique s'est révélée supérieure pour capturer la non-linéarité du marché, révélant qu'une unité de surface supplémentaire engendre une augmentation proportionnelle de la valeur de 0,21 %. L'intégration des dimensions socio-économiques a porté la précision du modèle à 84 %, bien que le recours aux variables instrumentales ait nuancé ce constat : une fois le biais d'endogénéité corrigé, il apparaît que la valorisation est davantage portée par le niveau de revenu du quartier que par la réputation intrinsèque des écoles. Enfin, l'arbitrage prédictif a consacré la supériorité de la régularisation Ridge, qui, avec un RMSE de 0,04388, surpasse le Lasso en préservant le signal précieux des variables secondaires comme l'étage ou la distance au centre.

Malgré la robustesse de ces conclusions, l'analyse présente certaines limites inhérentes à la nature des données. Le modèle reste exposé à un biais de variable omise pour des facteurs qualitatifs non mesurés, tels que l'état général du bien, la luminosité ou le calme du quartier, qui pourraient influencer les coefficients de l'étage ou de la surface. De plus, l'utilisation de la distance à l'université comme instrument repose sur une hypothèse d'exogénéité forte qui, bien que théoriquement justifiée, ne peut être statistiquement vérifiée par un test de suridentification dans un modèle juste-identifié. Enfin, la performance du modèle Ridge souligne que si la surface est le moteur principal, la suppression drastique de variables opérée par le Lasso peut conduire à une perte d'information préjudiciable à la précision globale sur des données nouvelles.

Pour la pratique immobilière, il est donc recommandé de privilégier une approche d'évaluation multidimensionnelle plutôt que de se focaliser uniquement sur la surface ou la proximité scolaire. Les investisseurs et évaluateurs devraient utiliser le modèle semi-logarithmique régularisé par la méthode Ridge, car il offre le meilleur compromis entre pouvoir explicatif et capacité de généralisation. Il est également essentiel de rester vigilant sur le contexte socio-économique global du quartier, le revenu des résidents étant un prédicteur plus fiable de la valeur à long terme que les indicateurs scolaires isolés.