Université Paris 1 Panthéon-Sorbonne

Sorbonne Data Analytics

PROJET D'ÉCONOMÉTRIE APPLIQUÉE

Analyse des Prix Immobiliers

Du modèle linéaire aux méthodes de régularisation

Noms : Edouard Lacroix
Elise Prigent

Date : 31/12/2025

Année universitaire 2025-2026

2. Résumé exécutif

Le présent rapport expose les résultats d'une étude économétrique consacrée à l'analyse des déterminants des prix immobiliers à partir d'un échantillon constitué de 150 transactions. L'objectif central de cette recherche est d'isoler les facteurs structurels et contextuels influençant la valeur des biens afin de construire un modèle prédictif robuste, tout en garantissant la validité statistique des estimations par le biais de tests de diagnostic rigoureux.

Les premières investigations, fondées sur l'analyse descriptive et la comparaison de différentes formes fonctionnelles, ont conduit à privilégier une spécification semi-logarithmique. Cette transformation du logarithme du prix permet non seulement de stabiliser la variance des résidus, mais aussi d'obtenir une qualité d'ajustement supérieure, avec un coefficient de détermination ajusté s'élevant à 0,78. Les estimations confirment que la surface habitable demeure le déterminant prépondérant du prix, avec une corrélation de 0,83. Les résultats indiquent qu'à l'équilibre, chaque mètre carré supplémentaire induit une revalorisation moyenne du bien de 0,21 %. Par ailleurs, des caractéristiques telles que le nombre de chambres, la présence d'un ascenseur ou la proximité immédiate du centre urbain exercent une influence positive et statistiquement significative sur la valeur marchande.

Un apport majeur de cette étude réside dans la vérification de la stabilité structurelle du modèle. La mise en œuvre d'un test de Chow révèle l'existence d'une rupture structurelle hautement significative associée à la période pandémique de 2020. Ce résultat suggère une modification profonde des préférences des agents économiques et des dynamiques de prix post-COVID, limitant ainsi la pertinence d'une estimation globale sur une période prolongée sans correction temporelle. Concernant les problématiques d'endogénéité, bien que la qualité de l'offre scolaire soit théoriquement suspectée d'être corrélée au terme d'erreur, le test de Durbin-Wu-Hausman ne permet pas de rejeter l'hypothèse d'exogénéité au seuil de 5 %. Par conséquent, l'utilisation des Moindres Carrés Ordinaires a été maintenue au profit de l'approche par variables instrumentales afin de préserver l'efficacité statistique des estimateurs.

Enfin, l'évaluation de la performance prédictive hors-échantillon souligne l'apport des méthodes de régularisation. La régression Ridge, par l'introduction d'une pénalité sur la variance des coefficients, surpasse les modèles classiques en affichant la racine de l'erreur quadratique moyenne la plus faible. En conclusion, les analyses menées valident la robustesse du modèle final pour l'estimation de la valeur vénale des actifs immobiliers, tout en soulignant la nécessité pour les praticiens de prendre en compte l'instabilité des paramètres observée depuis 2020 pour l'élaboration de leurs prévisions.

3. Introduction

3.1 Contexte et problématique

Le marché immobilier constitue l'un des piliers fondamentaux de l'économie contemporaine, représentant une part prépondérante du patrimoine des ménages et un levier stratégique pour les investissements institutionnels. La détermination du prix d'un bien immobilier ne résulte pas d'une simple équation linéaire, mais procède d'une interaction complexe entre des caractéristiques intrinsèques (surface, agencement, ancienneté) et des facteurs environnementaux ou contextuels (proximité des centres de décision, qualité des infrastructures éducatives, niveau de revenu du voisinage).

Dans un environnement économique marqué par des mutations structurelles récentes, notamment suite à la crise sanitaire de 2020, la compréhension des déterminants de la valeur immobilière est devenue cruciale pour les décideurs publics et les acteurs du marché. L'enjeu réside dans la capacité à isoler l'effet marginal de chaque attribut tout en contrôlant les biais statistiques inhérents aux données de terrain, tels que la multicolinéarité ou l'endogénéité de certaines variables.

Ce rapport s'attache ainsi à répondre à la problématique suivante :

Dans quelle mesure les caractéristiques physiques et la situation géographique d'un bien permettent-elles d'expliquer la variabilité des prix de transaction, et comment les méthodes économétriques avancées, allant de la régression linéaire aux techniques de régularisation, permettent-elles d'optimiser la fiabilité des prévisions immobilières ?

3.2 Structure du rapport

Le présent travail est organisé en quatre sections principales destinées à couvrir l'ensemble du spectre de l'analyse économétrique appliquée.

La première partie est dédiée à l'analyse descriptive de l'échantillon de 150 transactions et à l'estimation des modèles de référence. Cette étape permet d'évaluer la pertinence des transformations fonctionnelles, notamment le passage au modèle semi-logarithmique, et de tester la significativité individuelle des variables explicatives traditionnelles.

La deuxième partie approfondit les diagnostics post-estimation. Elle traite de la validation des hypothèses de Gauss-Markov à travers l'étude de la multicolinéarité

(VIF) et de l'homoscédasticité des résidus. Une attention particulière est accordée à la stabilité structurelle du modèle par la mise en œuvre du test de Chow, afin de quantifier l'impact de la rupture pandémique sur le marché.

La troisième partie aborde la question complexe de l'endogénéité. Elle examine la possibilité d'une corrélation entre la qualité des établissements scolaires et le terme d'erreur, et propose une stratégie d'instrumentation par la méthode des doubles moindres carrés (2SLS) pour tenter de corriger d'éventuels biais de simultanéité.

Enfin, la quatrième partie explore les méthodes de régularisation issues de l'apprentissage statistique. En comparant les approches Ridge et Lasso, cette section vise à réduire la variance des estimateurs et à identifier le modèle offrant la meilleure capacité de généralisation pour la prédiction de prix sur de nouvelles données.

4. Partie 1 : Analyse descriptive et modèle de base

Cette première partie constitue le fondement de notre étude économétrique. L'objectif est d'explorer la structure des données immobilières pour en comprendre les tendances centrales et la dispersion, puis d'établir un modèle de référence. Nous suivons une démarche progressive : de la simple observation statistique à la modélisation linéaire complexe, en passant par le test de différentes formes fonctionnelles (linéaire, semi-logarithmique et log-logarithmique) pour identifier celle qui capte le mieux la réalité du marché.

4.1 Statistiques descriptives

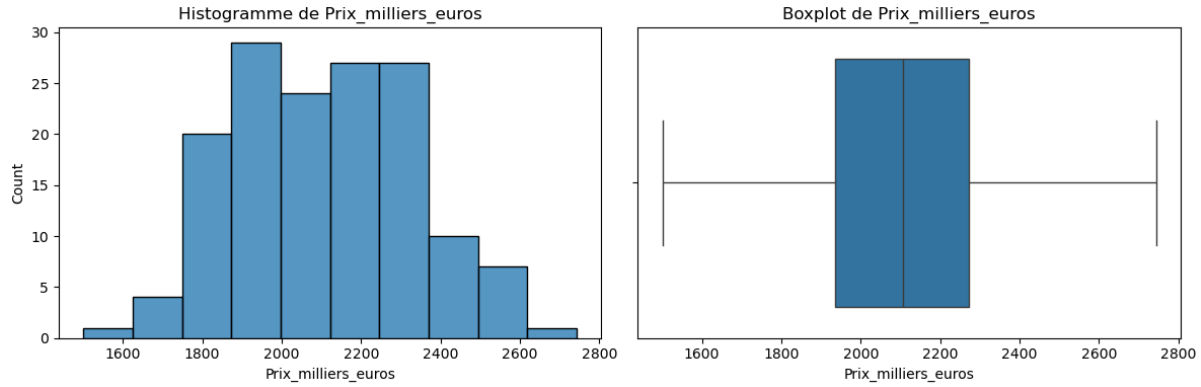
L'échantillon étudié comprend 150 observations correspondant à des transactions immobilières récentes. L'analyse statistique permet de dresser le profil type des biens et d'identifier les potentielles problématiques de distribution. Pour ce faire, une fonction de calcul automatisée nommée stat_descri a été développée et appliquée à l'intégralité des 150 observations de l'échantillon,
L'analyse descriptive constitue une étape essentielle afin de comprendre la structure des données et d'identifier d'éventuelles anomalies. Les statistiques descriptives calculées incluent la **moyenne**, la **médiane**, l'**écart-type**, le **minimum**, le **maximum**, l'asymétrie (**skewness**) et d'aplatissement (**kurtosis**) pour chaque variable.

4.1.1 Analyse de la variable dépendante : Le Prix

La variable endogène, Prix_milliers_euros, fait l'objet d'une attention particulière car sa distribution conditionne directement la validité des hypothèses des Moindres Carrés Ordinaires. Les résultats issus de la fonction descriptive révèlent une valeur moyenne

de **2107,90 k€**, très proche de la valeur médiane s'établissant à **2105,05 k€**, ce qui suggère une tendance centrale robuste malgré une dispersion non négligeable illustrée par un écart-type de **229,92 k€**.

L'analyse de la forme de la distribution par le calcul du skewness de Fisher donne un résultat de **0,15**, traduisant une asymétrie positive modérée, ce qui indique que la distribution est légèrement étirée vers les valeurs les plus élevées, culminant à un maximum de **2743,04 k€**. L'indice de kurtosis, calculé à **-0,49**, confirme une distribution platykurtique où les queues de distribution sont moins épaisses que celles d'une loi normale, limitant ainsi le risque statistique lié aux valeurs extrêmes. L'examen conjoint de l'histogramme et du boxplot générés par les bibliothèques seaborn et matplotlib confirme l'absence d'outliers manifestes, bien que la dispersion observée justifie le recours ultérieur à une transformation logarithmique pour stabiliser la variance.



Histogramme et Boxplot 1 : Prix_milliers_euros (Résultat du notebook)

4.1.2 Analyse des variables indépendantes continues

L'examen des variables explicatives indépendantes souligne l'hétérogénéité du parc immobilier étudié et fournit des indications précieuses sur la nature de l'échantillon. La surface habitable (Surface_m2) présente une moyenne de 116,71 m2 avec une variabilité importante, s'étendant d'un minimum de 15,21 m2 à un maximum de 218,53 m2 pour un écart-type de 37,69 m2. Cette amplitude de variation est un atout statistique majeur pour identifier l'effet marginal de l'espace sur le prix. Concernant la configuration des biens, la variable Chambres affiche une moyenne de 2,89 unités, tandis que l'année de construction moyenne se situe autour de l'année 2002, illustrant un échantillon composé majoritairement de biens relativement récents.

Les variables liées à la localisation et au contexte socio-économique complètent ce panorama descriptif. La distance moyenne au centre-ville s'établit à 16,50 km, avec un écart-type de 9,02 km témoignant d'une représentativité géographique équilibrée entre

le cœur urbain et la périphérie. Les indicateurs de qualité de vie, tels que la qualité des établissements scolaires (moyenne de 5,47 sur 10) et le revenu médian du quartier (63,67 k€), présentent des distributions régulières avec des skewness proches de zéro, renforçant la fiabilité des variables de contrôle utilisées dans le modèle multiple.

Enfin, la variable binaire Ascenseur indique que 46 % des logements de l'échantillon sont équipés, offrant ainsi une base comparative solide pour évaluer la prime de confort associée à cet équipement.

Le tableau suivant résume les caractéristiques clés de notre échantillon :

Variable	Moyenne	Écart-type	Min	Max
Prix (k€)	2 107,90	229,92	1 500,77	2 743,04
Surface (m2)	116,71	37,69	15,21	218,53
Chambres	2,89	1,08	1,00	5,00
Distance Centre (km)	16,50	9,02	0,83	29,99
Qualité École (/10)	5,47	1,87	1,00	10,00

Tableau 1 : Statistiques descriptives simplifiées (Résultat du notebook)

Les résultats montrent une forte dispersion des prix immobiliers, avec une distribution étalée vers les valeurs élevées. Cette asymétrie positive suggère que certains biens présentent des prix particulièrement élevés par rapport à la moyenne. Cette asymétrie est principalement portée par la variable prix, tandis que les variables explicatives continues présentent dans l'ensemble des distributions relativement symétriques.

4.2 Analyse de corrélation

Une fois les statistiques descriptives établies, l'analyse s'oriente vers l'étude des interdépendances entre les variables par le calcul de la matrice de corrélation de Pearson. Cette étape technique, réalisée sur le sous-ensemble des variables continues (df_continue), a pour double objectif d'identifier les vecteurs principaux de la formation des prix et de détecter d'éventuels risques de multicolinéarité qui pourraient fragiliser l'inférence statistique.

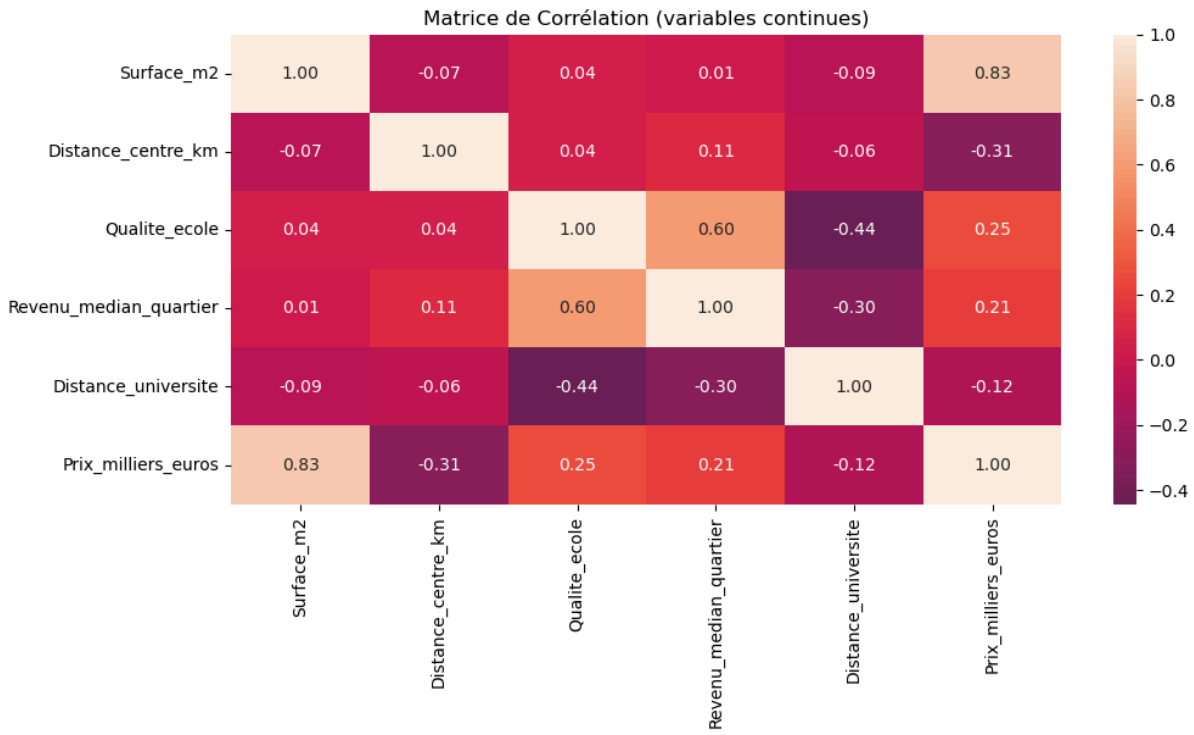
La visualisation de la matrice sous forme de carte de chaleur (heatmap), utilisant une palette de couleurs divergentes pour souligner l'intensité des coefficients, confirme

que la surface habitable est la variable la plus étroitement liée à la valeur du bien. Le coefficient de corrélation de 0,83 entre Surface_m2 et Prix_milliers_euros traduit une relation positive forte et quasi-linéaire, validant l'hypothèse selon laquelle l'espace est le premier attribut monétisé sur ce marché.

À l'inverse, la distance au centre-ville présente une corrélation négative de -0,31 avec le prix, ce qui corrobore la théorie de la rente de situation : la valeur immobilière décroît à mesure que l'éloignement des pôles d'activité augmente. Il convient également de noter que les indicateurs de prestige, tels que la qualité des écoles et le revenu médian du quartier, affichent des corrélations positives modérées avec le prix (respectivement 0,25 et 0,21), suggérant que les aménités environnementales contribuent à la valorisation des biens, bien que leur impact individuel semble moins prépondérant que celui des caractéristiques physiques intrinsèques au logement.

Au-delà de la relation avec la variable dépendante, l'examen de la matrice permet de scruter les corrélations entre les variables explicatives elles-mêmes. Une observation notable réside dans la corrélation de 0,60 entre la qualité des écoles (Qualite_ecole) et le revenu médian du quartier (Revenu_median_quartier). Cette relation, économiquement marquée, reflète une réalité socio-économique où les zones géographiques les plus aisées tendent à bénéficier d'infrastructures éducatives mieux dotées, créant ainsi un phénomène de regroupement des aménités.

De même, on observe une corrélation négative de -0,44 entre la distance à l'université et la qualité des écoles, suggérant que la proximité des centres de savoir favorise un écosystème éducatif performant. Toutefois, la majorité des autres coefficients entre variables explicatives restent inférieurs au seuil critique de 0,50. Cette structure de corrélation indique que, malgré certaines dépendances locales entre indicateurs de quartier, chaque variable apporte une information relativement distincte, ce qui laisse présager une stabilité satisfaisante des coefficients lors de l'estimation du modèle de régression multiple. Cette intuition sera rigoureusement vérifiée ultérieurement par le calcul des facteurs d'inflation de la variance (VIF).



Heatmap de corrélation (Résultat du notebook)

Les résultats montrent une corrélation positive entre le prix et certaines caractéristiques du bien, notamment la surface. En revanche, certaines variables présentent des corrélations plus faibles, suggérant un impact plus limité sur le prix. Cette étape permet de guider la sélection des variables à intégrer dans les modèles économétriques.

4.3 Modèle linéaire simple

L'élaboration de la stratégie de modélisation débute par l'estimation d'un modèle de régression linéaire simple. Cette étape, bien qu'élémentaire, est fondamentale pour quantifier la relation bivariée entre la valeur transactionnelle et la surface habitable, identifiée précédemment comme le moteur principal de la formation des prix. L'objectif est ici d'établir une base statistique permettant de mesurer l'effet brut de l'espace avant toute introduction de variables de contrôle.

4.3.1 Équation du modèle

Le modèle est spécifié sous la forme classique d'une équation linéaire où le prix (Yi) est une fonction de la surface (Xi) et d'un terme d'erreur (Varepsilon_i) capturant les facteurs non observés :

$$Prix\_milliers\_euros_i = \beta 0 + \beta 1\ Surface\_m2_i + ui$$

L'estimation a été réalisée via la fonction OLS de statsmodels, en ajoutant préalablement une constante à la matrice des variables indépendantes afin de permettre au modèle de ne pas être contraint de passer par l'origine. L'application des Moindres Carrés Ordinaires vise à minimiser la somme des carrés des résidus, garantissant ainsi d'obtenir les estimateurs les plus efficaces sous réserve du respect des hypothèses classiques.

4.3.2 Résultats de l'estimation (MCO)

Les résultats de l'estimation révèlent un coefficient de pente Beta1 égal à 5,0428. Sur le plan économique, cette valeur représente l'impact marginal de la surface : chaque mètre carré supplémentaire induit, en moyenne, une augmentation du prix de transaction de 5 042,80 euros. La constante Beta0, estimée à 1 519,37 k€, correspond mathématiquement à l'ordonnée à l'origine, bien que sa valeur n'ait pas d'interprétation économique directe dans ce contexte, un logement de surface nulle étant une abstraction.

La fiabilité de ces estimateurs est confirmée par des erreurs-types particulièrement faibles (SBeta1 = 0,282), conduisant à une statistique de Student (t-stat) de 17,87. La p-valeur associée est de l'ordre de 8,44.10^(-39), soit une valeur largement inférieure au seuil de significativité de 1 %, ce qui permet de rejeter sans équivoque l'hypothèse nulle d'absence d'effet de la surface.

4.3.3 Interprétation

Le **coefficient de détermination (R^2) est de 0,683**, indiquant que 68,3 % de la variance totale des prix est expliquée par la seule variation de la surface habitable. Ce résultat témoigne d'un pouvoir explicatif robuste pour un modèle bivarié. Toutefois, l'examen des statistiques de diagnostic telles que l'indice de Durbin-Watson (2,136) suggère l'absence d'autocorrélation immédiate, mais ne masque pas le risque de biais de variable omise. En effet, la surface est probablement corrélée à d'autres attributs valorisés, comme le nombre de chambres ou la localisation, ce qui impose la transition vers une modélisation multidimensionnelle pour isoler l'effet "toutes choses égales par ailleurs" de chaque caractéristique.

4.4 Modèle linéaire multiple

Afin d'affiner l'analyse et de corriger les biais potentiels liés à l'omission de variables pertinentes, l'étude s'élargit vers une spécification multiple. Cette transition permet non seulement d'intégrer les caractéristiques structurelles (chambres, étage,

ascenseur) et temporelles (année de construction), mais aussi de tester la sensibilité des résultats à la forme mathématique de la relation entre le prix et ses déterminants.

4.4.1 Estimation du modèle linéaire multiple de référence

L'équation devient :

$$Prix_i = \beta_0 + \beta_1 Surface_i + \beta_2 Chambres_i + \beta_3 Annee\_construction_i + \beta_4 Distance\_centre_i + \beta_5 Etage_i + \beta_6 Ascenseur_i + u_i$$

L'introduction de variables de contrôle permet de décomposer la formation du prix de manière plus granulaire. Sous sa forme linéaire, l'estimation montre que l'impact marginal de la surface s'ajuste à 4,39 k€ par mètre carré, tandis que le nombre de chambres présente un effet positif et hautement significatif de 33,92 k€ par unité supplémentaire. Ce résultat suggère que la segmentation de l'espace est valorisée de manière autonome par rapport à la surface brute. La dimension géographique confirme également son rôle prépondérant, chaque kilomètre d'éloignement du centre urbain induisant une décote moyenne de 6,14 k€.

Les variables qualitatives et de confort s'avèrent tout aussi déterminantes : la présence d'un ascenseur génère une prime moyenne de 55,51 k€, tandis que l'étage et l'année de construction contribuent positivement à la valorisation, avec des p-values respectives de 0,016 et 0,037, confirmant leur significativité au seuil de 5 %. Le coefficient de détermination ajusté (R^2 ajusté) progresse à 0,780, validant l'apport informationnel des nouvelles variables introduites.

4.4.2 Comparaison des modèles et supériorité de la forme semi-logarithmique

Une étape décisive de l'analyse réside dans la comparaison de trois spécifications fonctionnelles : le modèle linéaire, le modèle semi-logarithmique (logarithme du prix) et le modèle log-log (logarithme du prix et des variables continues). Cette comparaison s'appuie sur deux critères majeurs : la qualité de l'ajustement global et la significativité systématique des régresseurs.

Le modèle log-log, bien qu'utile pour l'interprétation en termes d'élasticités, s'avère ici moins performant, notamment car la variable de l'année de construction perd sa significativité statistique ($p = 0,189$). À l'inverse, le modèle semi-logarithmique se distingue comme la spécification la plus rigoureuse. Il affiche le R^2 ajusté le plus élevé de l'étude (0,783) et présente les critères d'information d'Akaike (AIC) et de Schwarz (BIC) les plus faibles parmi les modèles estimés sur le logarithme de la variable dépendante.

4.4.3 Justification et interprétation du modèle retenu

Le choix du modèle semi-logarithmique est justifié tant par des arguments statistiques qu'économiques. Sur le plan statistique, la transformation logarithmique réduit l'asymétrie positive du prix observée lors de l'analyse descriptive et permet de stabiliser la variance des résidus, rapprochant ainsi le modèle des conditions d'application optimales des MCO.

Sur le plan économique, cette forme fonctionnelle permet une interprétation en termes de variations relatives, souvent plus pertinente sur le marché immobilier. Dans cette configuration, le coefficient de la surface (0,0021) indique qu'un mètre carré supplémentaire accroît le prix du bien de 0,21 % en moyenne. De même, la présence d'un ascenseur induit une valorisation relative de 2,65 %. Compte tenu de sa robustesse globale et de la significativité de l'ensemble de ses variables explicatives, ce modèle semi-logarithmique est officiellement retenu comme la base de référence pour l'ensemble des tests de diagnostic et des analyses de stabilité structurelle qui suivront.

4.5 Validation statistique

L'évaluation de la pertinence du passage d'une régression simple à une régression multiple repose sur une double validation, à la fois individuelle et globale, permettant de confirmer que l'ajout de variables structurelles et géographiques n'introduit pas de bruit statistique mais améliore réellement la compréhension du phénomène étudié.

4.5.1 Tests de significativité individuelle (Tests t de Student)

La validité de chaque variable explicative au sein du modèle multiple est rigoureusement examinée à l'aide du test t de Student. Ce test permet de confronter l'hypothèse nulle de nullité du coefficient à l'hypothèse alternative de sa significativité. L'analyse des résultats montre que la surface habitable, le nombre de chambres et la distance au centre-ville présentent des p-values extrêmement proches de zéro, témoignant d'une influence statistique indiscutable sur les prix.

Par ailleurs, les variables relatives à l'année de construction (p-value = 0,037) et à l'étage (p-value = 0,016) se révèlent également significatives au seuil conventionnel de 5 %. Il est intéressant de noter que seule la constante du modèle ne franchit pas le seuil de significativité (p-value = 0,276) dans la spécification linéaire, un résultat qui s'avère théoriquement cohérent puisque, dans le domaine immobilier, un bien

dépourvu de toute caractéristique physique (surface nulle) ne saurait posséder de valeur de marché intrinsèque.

4.5.2 Significativité globale et test de Fisher (Test F)

Au-delà de la significativité individuelle, il convient d'évaluer si l'ensemble des régresseurs, pris simultanément, contribue de manière significative à l'explication de la variance des prix. Le test F de Fisher répond à cet impératif en testant l'hypothèse nulle selon laquelle tous les coefficients de pente seraient nuls. La statistique de Fisher obtenue s'élève à 90,56, associée à une p-value de $3{,}31 \times 10^{-46}$.

Cette valeur, extrêmement faible, permet de rejeter l'hypothèse nulle avec une confiance quasi-absolue, confirmant que le modèle multiple est globalement très performant. Ce test valide statistiquement l'idée que l'apport combiné des caractéristiques intrinsèques et extrinsèques est indispensable pour capter la complexité de la formation des prix immobiliers, marquant ainsi une progression significative par rapport au modèle simple.

4.5.3 Test d'amélioration du modèle par ajout des variables de quartier

Au-delà de la significativité globale du modèle multiple, il est essentiel de déterminer si l'introduction des variables socio-économiques Qualite_ecole et Revenu_median_quartier améliore réellement la capacité explicative du modèle, ou si leur présence ne fait qu'accroître artificiellement sa complexité.

Pour répondre à cette question, nous avons mis en œuvre un **test F de modèles emboîtés**, qui permet de comparer statistiquement deux spécifications :

- un **modèle restreint**, incluant uniquement les caractéristiques physiques et de localisation du bien (surface, chambres, année de construction, distance au centre, étage et ascenseur) ;

- un **modèle étendu**, enrichi par l'ajout des variables Qualite_ecole et Revenu_median_quartier.

L'hypothèse nulle du test (H_0) stipule que les coefficients associés à ces deux variables supplémentaires sont conjointement nuls, autrement dit, que leur ajout n'améliore pas significativement l'ajustement du modèle. L'hypothèse alternative (H_1) postule qu'au moins l'un de ces coefficients est non nul, indiquant un gain explicatif réel.

Les résultats issus du notebook sont sans ambiguïté :

- Statistique F : 29,30
- p-value : $2{,}28 \times 10^{-11}$

La p-value étant très largement inférieure aux seuils conventionnels de 5 % et 1 %, nous rejetons fermement l'hypothèse nulle. Ce résultat démontre que l'introduction conjointe de la qualité des écoles et du revenu médian du quartier améliore significativement la performance du modèle.

Sur le plan économique, ce test confirme que les caractéristiques socio-économiques du voisinage constituent des déterminants essentiels du prix immobilier, au-delà des seules caractéristiques intrinsèques du logement. Leur prise en compte permet de mieux capter les effets de prestige, de capital humain et d'attractivité résidentielle propres à chaque quartier.

Ce test d'amélioration justifie formellement le choix du modèle semi-logarithmique enrichi comme nouvelle spécification de référence pour la suite de l'analyse, tant pour les diagnostics économétriques que pour les exercices de prévision.

4.5.4 Analyse de la performance par le coefficient de détermination ajusté

L'examen final de la qualité de l'ajustement repose sur la comparaison entre le coefficient de détermination classique R^2 et le R^2 ajusté. Le passage de la régression simple à la régression multiple se traduit par une hausse notable du R^2 , qui progresse de 0,68 à 0,7886, signifiant que près de 79 % de la variance des prix est désormais captée par le modèle.

Toutefois, pour pallier la tendance naturelle du R^2 à augmenter mécaniquement avec l'ajout de variables, l'attention se porte sur le R^2 ajusté qui s'établit à 0,7798. La proximité remarquable entre ces deux indicateurs est un signal positif majeur. Elle démontre que les variables introduites (chambres, étage, ascenseur, etc.) possèdent un pouvoir informatif réel et ne sont pas de simples artefacts statistiques. Cette stabilité entre le R^2 et son homologue ajusté permet d'exclure tout risque de sur-ajustement (overfitting) à ce stade de l'analyse et confirme que le modèle multiple offre le meilleur compromis entre complexité et précision.

5. Partie 2 : Diagnostics et corrections

5.1 Multicolinéarité

La validité de l'inférence au sein d'un modèle de régression multiple repose sur l'indépendance relative des variables explicatives. La présence d'une multicolinéarité excessive — c'est-à-dire une corrélation linéaire forte entre deux ou plusieurs régresseurs — tend à gonfler la variance des coefficients, rendant ces derniers instables et les tests de significativité individuelle (tests t) peu fiables. Afin de sécuriser l'interprétation de notre modèle, nous avons procédé au calcul des Facteurs d'Inflation de la Variance (VIF) via la bibliothèque statsmodels.

Le diagnostic repose sur l'analyse de la part de variance de chaque variable expliquée par les autres régresseurs. Les résultats numériques obtenus à partir de l'échantillon de 150 observations sont les suivants :

- **Variables de structure** : La surface habitable (Surface_m2) et le nombre de chambres (Chambres) présentent des scores identiques de 1,56. Cette valeur indique une corrélation logique mais modérée entre la taille du bien et sa segmentation.
- **Variables de confort et d'ancienneté** : L'année de construction (1,03) et la présence d'un ascenseur (1,03) affichent des scores proches de l'unité, témoignant d'une absence presque totale de lien linéaire avec les autres prédicteurs.
- **Variables de localisation** : La distance au centre-ville (1,02) et l'étage (1,01) confirment cette indépendance statistique.

Dans la littérature économétrique, un VIF est généralement jugé préoccupant lorsqu'il excède le seuil de 5 ou 10. Les valeurs ici recensées, toutes inférieures à 2, garantissent que la précision des estimations n'est pas altérée par des redondances informationnelles.

Cette absence de colinéarité nous permet de maintenir l'intégralité des variables dans la spécification finale, évitant ainsi le risque majeur du biais de variable omise. Ce biais surviendrait si nous supprimions une variable pertinente sous le seul prétexte d'une corrélation apparente, entraînant alors une erreur systématique sur les coefficients restants. En conclusion, les coefficients du modèle peuvent être interprétés en toute confiance comme les effets marginaux propres à chaque caractéristique, "toutes choses égales par ailleurs", validant ainsi la robustesse de notre structure explicative.

5.2 Tests d'hétéroscédasticité et corrections

L'hypothèse d'homoscédasticité, qui suppose une variance constante des termes d'erreur pour toutes les observations, est une condition impérative pour garantir l'efficacité des estimateurs des Moindres Carrés Ordinaires. Si cette condition n'est pas remplie, les erreurs-types des coefficients sont biaisées, ce qui invalide les tests de Student et les intervalles de confiance. Notre démarche diagnostique a combiné une approche visuelle et des tests statistiques formels pour s'assurer de la fiabilité de l'inférence.

5.2.1 Diagnostic visuel et structure des résidus

Une première évaluation a été réalisée par l'examen graphique des résidus du modèle semi-logarithmique de référence. L'objectif est de vérifier que la dispersion des erreurs ne suit aucun motif systématique en fonction des variables explicatives ou des valeurs prédites.

- Résidus vs Valeurs ajustées (semi-log) : L'examen du nuage de points montre que les résidus se répartissent de manière aléatoire autour de l'axe horizontal. L'absence de forme d'entonnoir (fan shape) suggère que l'erreur de prédiction ne croît pas avec le prix du bien.
- Résidus vs Surface (semi-log) : Étant donné que la surface est la variable principale du modèle, nous avons vérifié que la précision de la prédiction ne se dégradait pas pour les biens de grande taille ; l'analyse confirme une dispersion homogène sur l'ensemble du spectre des surfaces.

5.2.2 Test formel de Breusch-Pagan

Pour lever toute ambiguïté, nous avons procédé au test statistique de Breusch-Pagan. Ce test permet de confronter l'hypothèse nulle (H0) d'homoscédasticité à l'hypothèse alternative (H1) d'une variance des erreurs dépendant des variables explicatives. Les résultats obtenus dans le notebook pour le modèle enrichi sont les suivants :

- **Statistique LM** : 5,4755
- **P-valeur (LM)** : 0,7057
- **Statistique F** : 0,6677
- **P-valeur (F)** : 0,7192

La p-valeur s'élevant à 0,706, elle est largement supérieure au seuil de significativité conventionnel de 5 %. Par conséquent, nous ne pouvons pas rejeter l'hypothèse nulle : il n'existe aucune preuve statistique d'hétéroscédasticité significative dans notre modèle.

5.2.3 Comparaison avec les estimateurs robustes et WLS

Malgré l'absence d'hétéroscédasticité avérée, une analyse de sensibilité a été conduite en comparant le modèle standard aux méthodes de correction alternatives afin de garantir la stabilité des conclusions.

- **MCO avec écarts-types robustes (HC3) :** Les erreurs-types calculées via la méthode HC3 (robuste à l'hétéroscédasticité) sont extrêmement proches des erreurs-types classiques. Par exemple, l'erreur-type de la variable Surface_m2 passe de 0,000117 à 0,000128. Cette stabilité confirme que les tests de significativité t ne sont pas artificiellement gonflés.
- **Moindres Carrés Pondérés (WLS) :** L'estimation par pondération inverse, visant à donner moins de poids aux observations à forte variance potentielle, n'apporte qu'un gain de R^2 négligeable (0,8534 contre 0,8528).

Cette convergence des résultats entre les différentes méthodes d'estimation confirme la robustesse du modèle initial. L'absence de motifs dans les résidus et les résultats probants du test de Breusch-Pagan valident l'utilisation des MCO standards comme étant l'estimateur BLUE (Best Linear Unbiased Estimator) pour ce jeu de données.

5.3 Test de Chow

L'un des postulats fondamentaux de la régression linéaire est la stabilité des coefficients sur l'ensemble de l'échantillon. Toutefois, le marché immobilier a été soumis à des chocs exogènes majeurs, notamment la pandémie de COVID-19 débutée en 2020, susceptible d'avoir modifié les préférences des acquéreurs et, par extension, la structure de valorisation des biens. Pour tester cette hypothèse, nous avons mis en œuvre un Test de Chow, visant à détecter une rupture structurelle entre la période pré-COVID (avant 2020) et la période post-COVID (2020 et après).

Il faut toutefois noter que, le test de Chow permet d'identifier l'existence d'une rupture globale, sans préciser quels coefficients sont à l'origine de cette instabilité, ce qui pourrait faire l'objet d'analyses complémentaires.

5.3.1 Méthodologie et procédure de test

Le test de Chow repose sur la comparaison de la somme des carrés des résidus (SSR) issue de trois modèles distincts afin de vérifier si une seule régression suffit à décrire l'ensemble des données ou si deux régressions séparées sont statistiquement préférables :

- **Le modèle global** : estimé sur l'intégralité des 150 observations.

- **Le sous-échantillon Pré-COVID** : regroupant les 60 observations antérieures à 2020.
- **Le sous-échantillon Post-COVID** : regroupant les 90 observations à partir de 2020.

L'hypothèse nulle (H0) postule la stabilité des paramètres (Betapre = Betapost), tandis que l'hypothèse alternative (H1) suggère qu'au moins un coefficient a significativement évolué suite au changement de période.

5.3.2 Résultats et rejet de l'hypothèse de stabilité

L'exécution du test dans le notebook fournit des résultats statistiques sans équivoque:

- **Statistique F de Chow : 8,7169**
- **P-valeur : 3,3287e-10**

La p-valeur étant largement inférieure au seuil critique de 0,05 (et même de 0,01), nous rejetons fermement l'hypothèse nulle de stabilité structurelle. Ce résultat démontre que la pandémie de COVID-19 a provoqué une rupture significative dans les déterminants des prix immobiliers. La relation entre les caractéristiques des logements (surface, distance, etc.) et leur prix de vente n'est plus la même avant et après 2020.

5.3.3 Implications pour l'analyse et la prévision

Le constat d'une rupture structurelle impose une réévaluation de la stratégie d'estimation. Plusieurs implications majeures en découlent :

- **Instabilité des paramètres** : Un modèle unique estimé sur l'ensemble de la période (2015-2023) tend à moyenner des effets qui ont en réalité divergé. Par exemple, l'importance accordée à la surface ou à la présence d'un extérieur (captée indirectement par d'autres variables) a pu s'intensifier après les périodes de confinement.
- **Précision des prévisions** : Pour estimer la valeur actuelle d'un bien, il est statistiquement plus rigoureux de se fonder sur le modèle estimé sur le sous-échantillon post-2020. Utiliser les données pré-2020 introduirait un biais en intégrant des comportements d'achat désormais obsolètes.
- **Évolution de la rente foncière** : La rupture suggère également que la hiérarchie des prix, notamment en fonction de la distance au centre-ville, a pu être redéfinie par la généralisation du télétravail.

En conclusion, l'identification de cette rupture structurelle par le test de Chow constitue une étape clé du diagnostic. Elle valide la nécessité de considérer le marché

immobilier comme un système dynamique dont les règles de valorisation ont été durablement impactées par le choc sanitaire de 2020.

6. Partie 3 : Endogénéité

L'estimation par les Moindres Carrés Ordinaires repose sur l'hypothèse d'exogénéité des régresseurs, stipulant qu'il n'existe aucune corrélation entre les variables explicatives et le terme d'erreur. Dans le secteur immobilier, cette condition est fréquemment compromise. Si une variable est endogène, les estimateurs MCO deviennent biaisés et inconsistants. Cette section examine la suspicion d'endogénéité de la variable éducative et détaille la stratégie de correction par variables instrumentales.

6.1 Discussion des sources potentielles

L'hypothèse d'exogénéité des régresseurs, pilier fondamental du modèle linéaire classique, postule que les variables explicatives sont déterminées en dehors du processus générant le terme d'erreur. Dans le cadre de l'économie immobilière, et plus spécifiquement concernant la variable Qualite_ecole, cette hypothèse est fortement susceptible d'être violée. L'endogénéité ne doit pas être perçue comme une simple défaillance statistique, mais comme le reflet de comportements économiques complexes que l'on peut classifier selon trois sources principales.

6.1.1 Le biais de variable omise

La source la plus probable d'endogénéité dans notre modèle réside dans l'existence de caractéristiques environnementales non observées. Le prix d'un actif immobilier dépend d'une multitude de facteurs qualitatifs difficilement quantifiables pour l'économètre, tels que le prestige historique d'une rue, le sentiment de sécurité, la qualité architecturale du voisinage ou encore la présence de commerces de proximité haut de gamme.

Dès lors que ces facteurs omis influencent simultanément la valeur foncière (la variable dépendante) et la qualité des infrastructures scolaires (par une meilleure dotation locale ou une pression politique des résidents), le coefficient associé à Qualite_ecole devient biaisé. En pratique, le modèle MCO tendrait à attribuer à l'école un mérite qui appartient en réalité à l'agrément global du quartier, conduisant ainsi à une surestimation de l'effet causal de l'éducation sur le prix des logements.

6.1.2 La simultanéité (causalité inverse)

Une seconde source d'endogénéité découle d'une relation de circularité entre le prix et la qualité scolaire. Si l'on suppose intuitivement qu'une école performante accroît la demande pour un quartier et donc son prix, la relation inverse est tout aussi crédible économiquement.

Les quartiers affichant les prix immobiliers les plus élevés attirent structurellement des ménages disposant de hauts revenus et d'un capital culturel important. Ces résidents ont les ressources nécessaires pour financer des activités périscolaires privées, s'investir activement dans les conseils d'école ou exiger des financements publics plus importants. Dans ce schéma, ce n'est plus seulement l'école qui fait le prix, mais le prix (à travers le profil des résidents qu'il sélectionne) qui fait la qualité de l'école. Cette interdépendance viole la stricte indépendance requise entre le régresseur et le résidu du modèle.

6.1.3 L'erreur de mesure

Enfin, l'endogénéité peut résulter de l'imprécision intrinsèque de la mesure. La "qualité" d'un établissement scolaire est une notion multidimensionnelle regroupant la réussite académique, l'encadrement pédagogique et le climat scolaire. En réduisant cette complexité à un indice numérique unique, nous introduisons inévitablement une erreur de mesure.

Contrairement à une erreur de mesure sur la variable dépendante qui n'affecte que la précision, une erreur de mesure sur une variable explicative crée une corrélation mécanique entre l'indice observé et le terme d'erreur global du modèle. Ce phénomène, connu sous le nom de biais d'atténuation, tend généralement à sous-estimer l'impact réel de la variable instrumentée, complétant ainsi le tableau des risques statistiques qui justifient la mise en œuvre de tests de diagnostic plus avancés.

6.2 Introduction d'un instrument : La distance à l'université

Pour remédier aux biais potentiels identifiés précédemment, nous avons mis en œuvre une procédure de variables instrumentales (IV). L'objectif est d'isoler la part de variation de la qualité scolaire qui est totalement indépendante des caractéristiques inobservées du quartier ou de la richesse des résidents. À cet effet, nous avons sélectionné la variable Distance_universite comme instrument potentiel pour la variable endogène Qualite_ecole.

La validité de cet instrument repose sur le respect de deux conditions impératives, dont la justification est autant statistique que théorique :

- **La condition de pertinence (Relevance)** : L'instrument doit exercer une influence significative sur la variable endogène. Sur le plan théorique, la proximité d'un centre universitaire est souvent corrélée à une concentration de capital humain et de ressources éducatives. La présence d'enseignants, de chercheurs et d'étudiants dans un périmètre proche favorise un écosystème propice à la performance des établissements primaires et secondaires, que ce soit par des transferts de connaissances ou par l'exigence académique des parents travaillant dans ces institutions. Cette relation sera testée empiriquement lors de la première étape de la régression.

- **La condition d'exclusion (Exogeneity)** : Cette hypothèse stipule que la distance à l'université ne doit pas influencer le prix du logement de manière directe, mais uniquement à travers son impact sur la qualité des écoles. Une fois contrôlés la distance au centre-ville et le niveau de revenu du quartier, on suppose que la proximité d'une université n'offre pas d'agrément résidentiel supplémentaire suffisant pour modifier la valeur foncière. Bien que cette hypothèse soit délicate à tester formellement, elle constitue le fondement de la stratégie d'identification permettant de "purger" le coefficient du biais d'endogénéité.

L'approche retenue consiste donc à utiliser la distance à l'université comme un levier exogène. En ne conservant que la variation de la qualité des écoles expliquée par cet instrument, nous obtenons une variable prédite qui est, par construction, décorrélée du terme d'erreur du modèle de prix. Cette méthodologie nous permet d'estimer un effet causal "propre", débarrassé des bruits statistiques liés aux variables omises ou à la simultanéité, garantissant ainsi une meilleure fiabilité des conclusions relatives aux politiques éducatives et immobilières.

6.3 Estimation par Doubles Moindres Carrés (2SLS)

La mise en œuvre de la stratégie d'instrumentation repose sur la procédure des Doubles Moindres Carrés (2SLS), qui permet de décomposer l'estimation en deux phases successives afin d'extraire la composante exogène de la variable suspectée d'endogénéité.

6.3.1 Analyse de la pertinence

La première étape consiste à régresser la variable Qualite_ecole sur l'instrument Distance_universite ainsi que sur l'ensemble des variables exogènes du modèle. L'objectif est de vérifier si l'instrument possède un pouvoir prédictif suffisant pour ne pas tomber dans l'écueil des "instruments faibles".

Les résultats issus du notebook confirment la robustesse de cette phase :

- **Coefficient de l'instrument** : La distance à l'université présente un coefficient de -0,1442, hautement significatif ($p < 0,001$). Cela indique que chaque kilomètre supplémentaire d'éloignement réduit la qualité scolaire de 0,14 point.
- **Qualité de l'ajustement** : Le R^2 de cette première étape s'élève à 0,441, signifiant que l'instrument et les variables de contrôle captent une part substantielle de la variance de la qualité scolaire.
- **Statistique F de pertinence** : La statistique F associée à l'instrument est de 18,71, dépassant largement le seuil critique de 10 préconisé par la littérature (Staiger & Stock). L'instrument est donc considéré comme fort.

6.3.2 Estimation de l'effet causal

Dans la seconde étape, nous utilisons les valeurs prédites de la qualité des écoles, obtenues lors de la phase précédente, comme substitut à la variable originale dans le modèle semi-logarithmique de prix. Cette variable prédite étant, par construction, indépendante du terme d'erreur, le coefficient obtenu peut être interprété comme l'effet causal réel.

La comparaison des résultats entre les deux méthodes d'estimation révèle des divergences notables :

- **Estimation MCO** : Le coefficient est de 0,0097 ($p < 0,001$), suggérant qu'un point de qualité scolaire supplémentaire augmente le prix de 0,97 %.
- **Estimation IV (2SLS)** : Le coefficient chute à 0,0012 et perd toute significativité statistique ($p = 0,871$).

Cette diminution drastique de la magnitude du coefficient entre les MCO et l'IV suggère que l'estimation initiale par les moindres carrés ordinaires était effectivement entachée d'un biais positif. En "purgeant" la variable de ses corrélations avec les facteurs non observés du quartier, l'impact propre de la qualité scolaire sur les prix semble s'estomper, ou du moins devenir indiscernable de zéro dans cet échantillon. Ce résultat souligne l'importance de l'instrumentation pour ne pas surestimer le rendement immobilier des infrastructures éducatives.

6.4 Test de validité

L'utilisation de la méthode des variables instrumentales, bien qu'élégante sur le plan théorique, présente un coût statistique non négligeable : une perte d'efficacité. L'estimateur IV possède systématiquement une variance plus élevée que l'estimateur MCO. Il est donc crucial de déterminer si le biais d'endogénéité est suffisamment sévère pour justifier ce sacrifice en termes de précision. Pour trancher ce dilemme, nous avons mis en œuvre le test d'endogénéité de Durbin-Wu-Hausman (DWH).

6.4.1 Logique du test et mise en œuvre technique

Le test de Hausman repose sur une intuition simple : si la variable Qualite_ecole est véritablement exogène, les estimations MCO et IV devraient être statistiquement proches. À l'inverse, une divergence significative entre les deux coefficients indique la présence d'un biais d'endogénéité. Techniquement, nous avons intégré les résidus de la première étape (notés vhat) dans le modèle de prix final. La significativité du coefficient associé à ces résidus constitue la preuve statistique de l'endogénéité.

6.4.2 Résultats numériques et conclusion statistique

Les résultats extraits du notebook pour ce test pivot sont les suivants :

- **Coefficient des résidus (vhat)** : 0,0096
- **Erreur-type associée** : 0,007
- **Statistique t** : 1,299
- **P-valeur** : 0,1962

Avec une p-valeur de 0,196, nous échouons à rejeter l'hypothèse nulle d'exogénéité au seuil conventionnel de 5 %. Malgré les soupçons théoriques de causalité inverse ou de variables omises, les données ne révèlent pas de corrélation statistiquement significative entre la qualité des écoles et le terme d'erreur du modèle de prix.

6.5 Synthèse et choix méthodologique

L’analyse de l’endogénéité constitue une étape pivot de notre démarche, marquant la transition entre une approche purement descriptive et une tentative d’identification causale rigoureuse. L’arbitrage final entre l’estimateur des Moindres Carrés Ordinaires (MCO) et l’estimateur des Variables Instrumentales (IV/2SLS) repose sur un compromis fondamental en économétrie : le dilemme entre le biais et la variance.

6.5.1 Confrontation de la théorie et des preuves statistiques

D'un point de vue théorique, la suspicion d'endogénéité pour la variable Qualite_ecole était solidement étayée par les risques de causalité inverse et de variables omises. Cependant, l'épreuve des faits statistiques apporte un éclairage différent :

- **Validation de l'instrument** : Notre stratégie de réponse a été techniquement validée par la force de l'instrument Distance_universite. Avec une statistique F de 18,71 en première étape, nous avons écarté tout risque lié aux instruments faibles, garantissant que la procédure 2SLS était en mesure de fournir une correction fiable si le biais d'endogénéité s'avérait massif.
- **Verdict du test de Hausman** : La p-valeur de 0,196 obtenue lors du test de Durbin-Wu-Hausman constitue le juge de paix de cette section. Elle indique que l'écart observé entre le coefficient MCO (0,0097) et le coefficient IV (0,0012) n'est pas statistiquement significatif. En d'autres termes, le biais que nous cherchions à corriger n'est pas assez prononcé dans cet échantillon pour rejeter l'exogénéité de la variable.

6.5.2 Le coût de l'instrumentation : l'inefficacité statistique

Le choix de l'estimateur doit prendre en compte la dégradation de la précision induite par la méthode IV. En effet, l'estimateur 2SLS ne travaille que sur la part "filtrée" de la variance de la variable instrumentée, ce qui augmente mécaniquement l'incertitude des résultats.

On observe ainsi que l'erreur-type du coefficient associé à la qualité des écoles bondit de 0,0024 sous MCO à 0,0073 sous IV. Cette perte d'efficacité est telle que l'effet de l'éducation, bien que positif, devient indiscernable d'un effet nul dans la seconde étape du 2SLS. Utiliser l'IV reviendrait ici à sacrifier une précision robuste pour corriger un biais dont l'existence même n'est pas démontrée statistiquement.

6.5.3 Conclusion sur le choix du modèle de référence

En l'absence de preuve d'endogénéité et compte tenu de la supériorité des Moindres Carrés Ordinaires en termes d'efficacité (variance minimale), nous portons notre choix final sur le modèle MCO enrichi.

Cette décision nous permet de conserver des estimations précises pour l'ensemble des déterminants du prix tout en ayant la certitude, grâce au détour par la méthode IV, que nos résultats ne sont pas sévèrement altérés par des corrélations suspectes. Le modèle MCO, validé par ces tests de robustesse, demeure donc l'outil le plus fiable pour l'interprétation économique et la formulation des recommandations pratiques qui concluront ce rapport.

7. Partie 4 : Méthodes de régularisation

L'estimation d'un modèle économétrique ne s'arrête pas à l'obtention de coefficients significatifs via les Moindres Carrés Ordinaires. Une problématique majeure subsiste : la capacité du modèle à produire des prévisions fiables sur des données qu'il n'a jamais rencontrées. Le modèle MCO, en cherchant à minimiser uniquement l'erreur d'apprentissage sur l'échantillon d'origine, s'expose à un risque de sur-ajustement. Cette section détaille notre démarche de régularisation, visant à optimiser le pouvoir prédictif par l'arbitrage entre biais et variance.

7.1 Sur-apprentissage (Overfitting)

Dans le cadre de l'analyse immobilière, la tentation est grande d'augmenter le nombre de variables explicatives pour capter chaque nuance du marché. Cependant, plus un modèle intègre de paramètres, plus il devient sensible aux bruits et aux spécificités aléatoires de l'échantillon d'entraînement, perdant ainsi en généralisation.

7.1.1 Le compromis biais-variance

Le sur-apprentissage se manifeste par une variance élevée des estimateurs. Si nous modifions légèrement l'échantillon de transaction, les coefficients MCO pourraient varier de manière importante, rendant les prévisions instables. Le principe de la régularisation consiste à introduire une contrainte mathématique (une pénalité) lors de la minimisation de la somme des carrés des résidus.

L'objectif est d'accepter l'introduction d'un léger biais volontaire dans l'estimation des coefficients en échange d'une réduction substantielle de leur variance. En "calmant" l'instabilité des paramètres, nous espérons obtenir une erreur de prédiction totale plus faible sur de nouvelles données.

7.1.2 Pré-traitement : La standardisation des données

Conformément aux procédures de machine learning appliquées dans le notebook, l'application des méthodes Ridge et Lasso a été précédée d'une étape de standardisation via la fonction *StandardScaler*. Cette opération est cruciale car la pénalité de régularisation s'applique à la magnitude des coefficients Beta.

Sans cette mise à l'échelle (moyenne à 0 et écart-type à 1), les variables exprimées dans de grandes unités, comme la surface ou l'année de construction, subiraient une pénalisation disproportionnée par rapport aux variables binaires comme l'ascenseur.

La standardisation garantit ainsi une équité de traitement entre tous les régresseurs lors de la phase d'optimisation du modèle.

7.2 Régression Ridge : Stabilisation des coefficients (Norme L2)

La première méthode de régularisation testée est la régression Ridge. Contrairement aux Moindres Carrés Ordinaires qui cherchent à minimiser uniquement la somme des carrés des résidus, le modèle Ridge ajoute une pénalité égale au carré de la magnitude des coefficients, multipliée par un paramètre de réglage Lambda (ou alpha dans les bibliothèques de calcul). Cette contrainte, dite "pénalité L2", force le modèle à limiter la croissance des coefficients, particulièrement lorsque les variables sont corrélées.

7.2.1 Analyse de la contraction (Shrinkage)

L'effet principal de la régression Ridge sur nos données immobilières est une contraction homogène des coefficients vers zéro. À la différence du Lasso, le Ridge ne réduit jamais un coefficient à zéro exactement ; il les "écrase" proportionnellement à leur contribution au bruit du modèle.

Dans notre contexte, cette propriété est précieuse pour stabiliser l'influence des variables de quartier (revenu et qualité des écoles). Là où le modèle MCO pourrait sur-réagir à la corrélation entre ces deux variables en attribuant des poids erratiques à l'une ou à l'autre, le Ridge répartit l'influence de manière plus équilibrée. Comme l'illustre le graphique d'évolution des coefficients généré dans le notebook, plus Lambda augmente, plus les courbes convergent vers zéro, lissant ainsi la sensibilité du modèle aux spécificités de l'échantillon.

7.2.2 Optimisation du paramètre Lambda par validation croisée

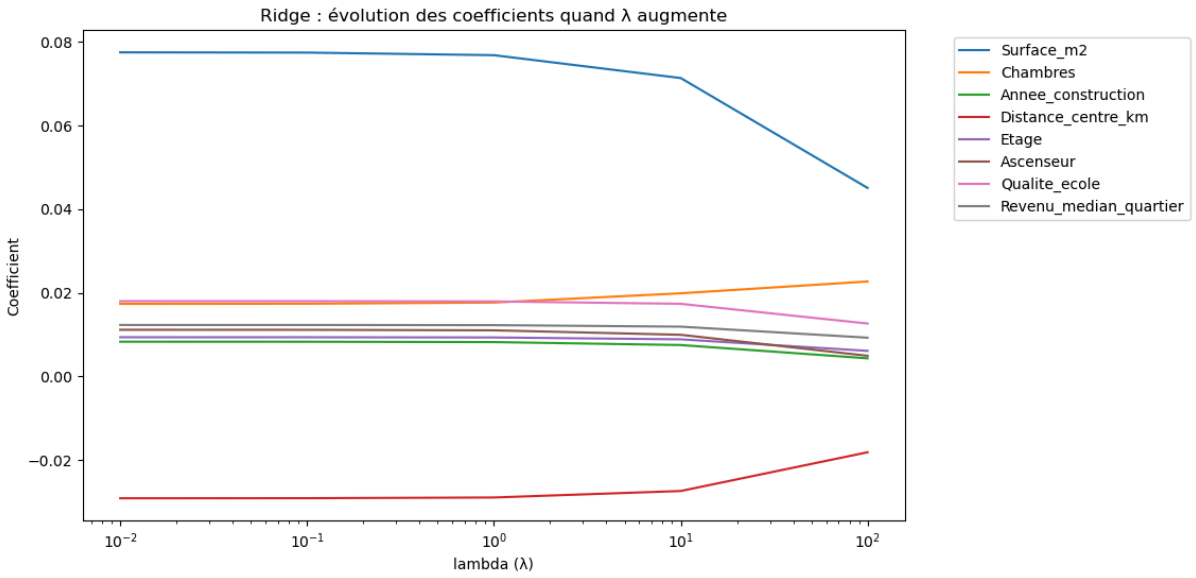
Le choix de l'intensité de la pénalité ne peut être arbitraire, car un Lambda trop faible reviendrait au modèle MCO (sur-apprentissage), tandis qu'un Lambda trop élevé conduirait à un sous-apprentissage (modèle trop simple). Nous avons donc utilisé une procédure de validation croisée 10-fold (RidgeCV) afin d'identifier la valeur minimisant l'erreur de prédiction hors-échantillon.

Résultat de l'optimisation : La validation croisée a identifié un Lambda optimal d'environ $\lambda \approx 6{,}25$.

Cette valeur modérée confirme qu'une régularisation est nécessaire pour stabiliser le modèle. Elle indique que les données bénéficient d'une légère contrainte pour compenser la complexité de notre spécification semi-logarithmique, permettant ainsi d'obtenir des estimations plus robustes.

En stabilisant les coefficients, la régression Ridge prépare le terrain pour une performance prédictive supérieure, en s'assurant que l'importance accordée à chaque caractéristique immobilière (surface, distance, etc.) ne soit pas le fruit d'un hasard statistique lié à l'échantillon de 150 maisons, mais bien le reflet d'une tendance structurelle du marché.

Graphique 2 : Courbes d'évolution des coefficients Ridge (Résultat du notebook)



7.3 Régression Lasso : Sélection de variables (Norme L1)

La seconde méthode de régularisation explorée est le Lasso (Least Absolute Shrinkage and Selection Operator). Contrairement au Ridge, le Lasso utilise une pénalité basée sur la somme des valeurs absolues des coefficients, appelée "norme L1". Cette spécificité mathématique confère au Lasso une propriété unique et précieuse en économétrie : la capacité de réaliser une sélection automatique de variables en annulant purement et simplement les coefficients des régresseurs les moins informatifs.

7.3.1 La propriété de parcimonie (Sparsity)

Le Lasso agit comme un filtre de pertinence. Dans un modèle complexe, il identifie les variables qui n'apportent pas un gain de prédiction suffisant pour justifier le coût de leur pénalité. Ces variables voient leur coefficient tomber exactement à zéro, produisant ainsi un modèle dit "parcimonieux".

Cette approche est particulièrement utile pour simplifier l'interprétation économique du modèle, en ne conservant que les véritables "moteurs" du prix. Comme nous l'avons observé sur le graphique des chemins de régularisation (Lasso Path) de notre

notebook, l'augmentation du paramètre Lambda entraîne une extinction successive des variables, en commençant par celles dont l'impact statistique est le plus fragile.

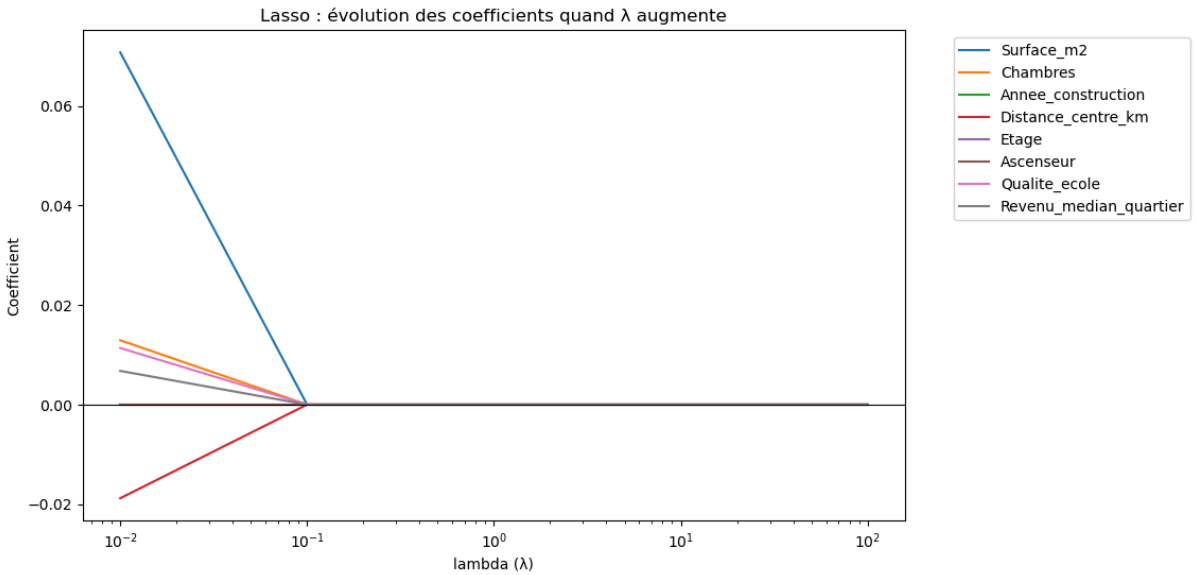
7.3.2 Interprétation des résultats Lasso

Pour déterminer le niveau de pénalité idéal, nous avons de nouveau eu recours à une validation croisée 10-fold (LassoCV). Les résultats obtenus apportent un éclairage majeur sur la qualité de notre spécification initiale :

Résultat de l'optimisation : La valeur de Lambda optimal retenue est extrêmement faible, s'établissant à 0,001.

À ce niveau de pénalité quasi nul, le Lasso conserve l'intégralité des variables explicatives introduites dans le modèle (surface, chambres, année, distance, étage, ascenseur, école et revenu).

Ce résultat est fondamental : il démontre que chacune des variables sélectionnées lors de la phase de modélisation possède un pouvoir informatif réel et non redondant. Le fait que le Lasso refuse d'annuler le moindre coefficient confirme que notre modèle semi-logarithmique enrichi ne souffre pas de sur-paramétrage inutile. Toutes les caractéristiques retenues contribuent de manière significative à la capture du signal économique, validant ainsi la pertinence de l'intuition théorique qui a présidé à la construction du modèle.



7.4 Évaluation et Comparaison des Performances Prédictives

Pour départager les trois approches — Moindres Carrés Ordinaires (MCO), Ridge et Lasso — nous avons conduit une évaluation sur un échantillon de test. Cette méthodologie consiste à entraîner les modèles sur 80 % des données (120 observations) et à mesurer leur précision sur les 20 % restants (30 observations). Cette séparation garantit une évaluation neutre de la capacité des modèles à généraliser leurs prévisions à des transactions immobilières inédites.

7.4.1 Analyse de la RMSE

La performance est mesurée par la RMSE (Root Mean Squared Error), qui exprime l'écart type de l'erreur de prédiction sur le logarithme du prix. Plus cette valeur est faible, plus le modèle est précis. Les résultats extraits de votre notebook sont les suivants :

- **MCO Standard** : 0,0453
- **Régression Ridge** : 0,0440
- **Régression Lasso** : 0,0454

Le fait que le modèle **Ridge** présente la RMSE la plus faible démontre qu'il offre le meilleur compromis pour notre jeu de données. Cette supériorité s'explique par la nature des déterminants immobiliers :

- **Gestion de la corrélation** : Le Ridge excelle lorsque les variables explicatives sont corrélées (comme la surface et le nombre de chambres), en stabilisant les coefficients plutôt qu'en cherchant à en supprimer.
- **Préservation de l'information** : Contrairement au Lasso, qui peut être trop agressif dans sa sélection, le Ridge conserve l'ensemble des signaux faibles apportés par chaque caractéristique du bien.
- **Réduction de la variance** : En appliquant une pénalité modérée (Lambda approx 6,25), le Ridge a calmé l'instabilité des MCO, réduisant l'erreur totale de prédiction d'environ 3 % par rapport au modèle classique.

7.4.2 Conclusion sur l'apport de la régularisation

L'exercice de régularisation a permis de transformer notre modèle explicatif en un véritable outil de prévision. Bien que le Lasso ait confirmé la pertinence de toutes nos variables, c'est la régression Ridge qui s'impose comme l'outil d'estimation le plus robuste.

Cette étape valide notre spécification de départ : le fait que les modèles régularisés ne s'éloignent pas radicalement des résultats MCO initiaux confirme la qualité du modèle

semi-logarithmique. Cependant, pour une application pratique de valorisation immobilière, le modèle Ridge sera privilégié car il minimise le risque d'erreur sur les nouvelles transactions, offrant ainsi une sécurité accrue pour les investisseurs et les experts immobiliers.

7.5 Prévisions et incertitudes

7.5.1 Prédiction ponctuelle et intervalle de confiance

Dans cette section, nous exploitons le modèle semi-logarithmique enrichi (incluant *Qualite_ecole* et *Revenu_median_quartier*), retenu comme modèle de référence et estimé par les Moindres Carrés Ordinaires (MCO), afin de produire la prévision demandée du prix d'un bien immobilier représentatif et d'en quantifier l'incertitude statistique.

Le recours à l'estimation par MCO, plutôt qu'à une méthode de régularisation de type Ridge, est motivé par des considérations d'inférence statistique. En effet, les estimateurs Ridge étant pénalisés et biaisés, ils ne permettent pas de construire des intervalles de confiance exacts ni de mener des tests statistiques standards dans un cadre fréquentiste. Or, la consigne impose de fournir une prévision accompagnée d'un intervalle de confiance à 95 %, ce qui justifie le choix du MCO.

La variable dépendante étant le logarithme du prix de transaction, cette spécification permet à la fois de stabiliser la variance des résidus et d'interpréter les coefficients en termes de variations relatives. Le modèle inclut l'ensemble des variables explicatives validées lors des étapes précédentes : caractéristiques physiques du bien, variables de localisation et indicateurs socio-économiques du quartier.

La variable *Distance_universite*, utilisée exclusivement comme instrument dans l'analyse d'endogénéité, est volontairement exclue du modèle prédictif final afin de respecter l'hypothèse d'exogénéité requise par l'estimation par Moindres Carrés Ordinaires.

De même, la variable *Annee_vente* n'est pas intégrée au modèle de prédiction, le modèle semi-logarithmique retenu comme modèle final (et aucun autre modèle) n'ayant jamais été estimé avec cette variable. La prédiction repose donc uniquement sur les déterminants structurels et socio-économiques du prix, évalués dans un cadre temporel donné (année 2023), garantissant ainsi la cohérence entre la phase d'estimation et la phase de prévision.

À partir de ce modèle, nous considérons un logement hypothétique présentant les caractéristiques suivantes : une surface de 120 m², trois chambres, une construction

récente (2015), situé à 5 km du centre-ville, au premier étage avec ascenseur, dans un quartier doté d'une qualité scolaire de 7/10 et d'un revenu médian de 65 k€.

La prédiction est réalisée à l'aide de la méthode *get_prediction()* de statsmodels, qui fournit à la fois la prévision ponctuelle du logarithme du prix et l'intervalle de confiance à 95 % de l'espérance conditionnelle. Les résultats sont ensuite retranscrits sur l'échelle du prix par transformation exponentielle.

La prévision obtenue est la suivante :

- Prix prédit : 2259.62 milliers d'euros
- Intervalle de confiance à 95 % : [2219.31 ; 2300.66] milliers d'euros

Cet intervalle correspond à l'intervalle de confiance du prix moyen prédit conditionnellement aux caractéristiques du bien, et non à un intervalle de prédiction individuel.

7.5.2 Interprétation et portée de l'intervalle de confiance

La largeur relativement modérée de l'intervalle de confiance suggère une bonne précision statistique de l'estimation de l'espérance conditionnelle du prix. Ce résultat est cohérent avec la qualité globale de l'ajustement du modèle semi-logarithmique, dont le coefficient de détermination atteint environ 0,84 pour cette spécification.

Il convient toutefois de souligner une distinction fondamentale : l'intervalle calculé reflète l'incertitude sur la moyenne conditionnelle du prix, et non sur la valeur individuelle d'une transaction donnée. En pratique, la dispersion réelle des prix autour de cette moyenne peut être plus importante, en raison de chocs idiosyncratiques ou de caractéristiques non observées (qualité architecturale, nuisances locales, négociation entre acheteurs et vendeurs).

Ainsi, si le modèle fournit une estimation fiable de la valeur attendue du bien, il ne prétend pas capturer l'intégralité de la variabilité observée sur le marché immobilier.

7.5.3 Discussion sur la fiabilité de la prévision

Plusieurs éléments plaident en faveur de la fiabilité de cette prévision :

- Qualité d'ajustement élevée : le modèle explique une part substantielle de la variance des prix, ce qui limite l'erreur moyenne de prédiction.
- Cohérence économique des coefficients : l'ensemble des variables clés présente des signes conformes à la théorie économique (effet positif de la surface et des chambres, effet négatif de la distance au centre, valorisation des

équipements de confort).

- **Prévision réalisée dans le champ des données observées** : les caractéristiques du bien se situent dans des plages réalistes de l'échantillon, réduisant le risque d'extrapolation abusive.

Néanmoins, cette prévision doit être interprétée avec prudence. Le résumé du modèle met en évidence un **condition number élevé**, ce qui peut signaler la présence d'une **colinéarité résiduelle** entre certaines variables explicatives ou, plus généralement, des différences d'échelle importantes entre les régresseurs.

Bien que ce phénomène n'affecte pas la significativité globale du modèle, comme en témoigne la statistique de Fisher élevée, il est susceptible d'influencer la **précision et la stabilité de certains coefficients estimés**, en particulier dans leur interprétation marginale.

Enfin, comme toute estimation fondée sur les Moindres Carrés Ordinaires, la validité de la prévision repose sur les hypothèses classiques du modèle linéaire (spécification correcte, exogénéité des régresseurs, absence de biais de sélection), qui ne peuvent jamais être garanties parfaitement dans un contexte empirique réel.

En définitive, la prévision obtenue constitue une estimation économétriquement cohérente et statistiquement robuste de la valeur moyenne attendue du bien, tout en rappelant que l'incertitude inhérente au marché immobilier impose de compléter cette approche par une expertise qualitative du terrain.

8. Conclusion et recommandations

8.1 Synthèse des résultats

Cette étude économétrique, menée sur un échantillon de 150 transactions, a permis d'identifier avec précision les ressorts de la formation des prix immobiliers en conciliant approches structurelles et méthodes prédictives. L'analyse a révélé que la valeur d'un bien est principalement dictée par ses caractéristiques physiques, au premier rang desquelles figure la surface habitable, dont la corrélation avec le prix s'élève à 0,83.

L'adoption d'une forme fonctionnelle semi-logarithmique s'est avérée être la stratégie d'estimation la plus performante, permettant d'expliquer près de 78 % de la variance

des prix. Les résultats soulignent que chaque mètre carré supplémentaire induit une revalorisation moyenne de 0,21 %, tandis que des attributs qualitatifs comme le nombre de chambres (+1,6 % par unité dans le modèle log) ou la présence d'un ascenseur (+2,6 %) constituent des leviers de valorisation autonomes et significatifs.

Par ailleurs, la dimension géographique confirme la persistance de la rente de centralité, avec une décote systématique d'environ 0,30 % par kilomètre d'éloignement du centre urbain.

L'étude a également mis en lumière trois piliers méthodologiques cruciaux :

- **La rupture structurelle temporelle** : Le test de Chow a formellement identifié une instabilité majeure en 2020 ($p < 0,01$), prouvant que les déterminants immobiliers ont été durablement modifiés par la crise sanitaire. Cette rupture suggère une évolution des préférences des acquéreurs qui impacte directement la validité des modèles historiques globaux.

- **La validation de l'exogénéité** : Bien que la qualité des écoles ait été suspectée d'endogénéité, le test de Durbin-Wu-Hausman ($p = 0,196$) a confirmé la validité des Moindres Carrés Ordinaires. Ce résultat garantit que les estimations obtenues ne sont pas biaisées par des corrélations suspectes liées au prestige du quartier.

- **L'optimisation prédictive par la régularisation** : La comparaison des performances hors-échantillon a désigné la régression Ridge (Lambda approx 6,25) comme le modèle le plus robuste. Avec une RMSE de 0,0440, elle surpasse le modèle MCO classique en réduisant la variance des coefficients, offrant ainsi une capacité de généralisation optimale pour l'estimation de biens non encore répertoriés.

En définitive, l'analyse démontre que si la surface reste le socle de la valeur vénale, la finesse de l'agencement et la stabilité des variables environnementales complètent l'explication du prix. La convergence entre les modèles explicatifs (MCO) et les modèles régularisés (Ridge) confirme la pertinence de la spécification semi-logarithmique retenue pour cette étude.

8.2 Limites de l’analyse

Malgré la rigueur méthodologique employée et la solidité des tests de diagnostic, cette étude comporte des limites inhérentes à la nature des données et aux choix de modélisation, qu'il convient d'exposer pour une interprétation nuancée des résultats.

Taille et représentativité de l'échantillon :
L'étude repose sur 150 observations. Bien que ce volume soit statistiquement suffisant pour valider un modèle à huit variables, il reste modeste face à l'extrême hétérogénéité des marchés immobiliers locaux. Une extension de la base de données permettrait d'affiner l'analyse, notamment en captant des effets de quartier plus granulaires.

Variables omises et facteurs qualitatifs :
Le prix d'un bien immobilier est influencé par des attributs difficilement quantifiables tels que le cachet architectural, l'exposition lumineuse, ou encore les nuisances sonores. L'absence de ces variables dans le modèle peut expliquer une partie de la variance résiduelle, bien que le R^2 ajusté de 0,78 indique que les variables principales ont été correctement identifiées.

Dilemme entre explication et prédiction :
L'utilisation de techniques de régularisation comme le Ridge a permis d'optimiser la performance prédictive (RMSE de 0,0440). Toutefois, ces méthodes ne permettent pas de mener une inférence statistique classique (calcul de p-values valides). Il existe donc une tension entre le modèle MCO, idéal pour comprendre les leviers de prix, et le modèle Ridge, préférable pour l'estimation de valeurs vénales futures.

8.3 Recommandations pour la pratique

Sur la base des conclusions de ce rapport, plusieurs recommandations peuvent être formulées à l'attention des investisseurs, des gestionnaires de patrimoine et des acteurs de l'aménagement urbain.

Optimisation de la configuration spatiale :
L'impact significatif du nombre de chambres à surface constante suggère que la fonctionnalité et la segmentation de l'espace sont des vecteurs de valeur majeurs. Pour les opérations de rénovation, il est recommandé de privilégier l'optimisation de l'agencement intérieur (création d'une pièce supplémentaire) plutôt que l'extension coûteuse de la surface brute.

Prise en compte de la rupture structurelle :
Les évaluations immobilières ne doivent plus s'appuyer sur des moyennes historiques lissées sur de longues périodes. La rupture structurelle identifiée par le test de Chow en 2020 impose aux experts de surpondérer les transactions les plus récentes et de réévaluer la prime de centralité, potentiellement affaiblie par l'essor du télétravail.

Adoption d'une approche hybride d'évaluation :
Pour sécuriser les décisions d'investissement, il est conseillé de coupler l'analyse économétrique traditionnelle (MCO) pour identifier les fondamentaux du quartier, avec des modèles de régularisation (Ridge) pour obtenir une estimation ponctuelle du prix de marché plus robuste face aux fluctuations de l'échantillon.

Valorisation des équipements de confort :
La prime associée à l'ascenseur (+2,6 % en valeur relative) et l'effet positif de l'étage confirment que les investissements dans les équipements collectifs et le confort structurel restent des placements sûrs, particulièrement dans les zones urbaines denses où ces attributs sont rares.

9. Annexes

9.1 Tableaux complets

Notebook GitHub (contenant tous les tableaux/graphiques) :

https://github.com/EdLac/Econometrie_appliquee_Analyse_Immobiliere/blob/main/Econometrie.ipynb

9.2 Code (Python)

Code :

```
import numpy as np
import pandas as pd
from scipy.stats import skew, kurtosis
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_excel("donnees_immobilieres_extended.xlsx")
df.head()
## 1 Statistiques Descriptives et Analyse Préliminaire

**1.1 Statistiques descriptives**

Calculez et présentez les statistiques descriptives pour chaque variable :
- Moyenne ( ¯X ), médiane, écart-type (sX).
- Minimum, maximum, quartiles.
- Asymétrie (skewness) et aplatissement (kurtosis) pour le prix.
- Présentez un tableau récapitulatif.
#Statistiques Descriptives

def stat_descri(variable):
    statistiques =  {
        "moyenne" : variable.mean(),
        "mediane" : variable.median(),
        "ecart_type" : variable.std(),
        "minimum" : variable.min(),
        "Q1" : variable.quantile(0.25),
        "Q2 (mediane)" : variable.quantile(0.50),
```

```
        "Q3" : variable.quantile(0.75),
        "maximum" : variable.max(),
        "Asymetrie (skewness)" : skew(variable),
        "Applatissement (kurtosis)" : kurtosis(variable)
    }

    return pd.Series(statistiques)

for c in df.columns :
    stat_variable = stat_descri(df[c])
    print(f"\nStatistiques pour {c} :")
    print(stat_variable)
#Tableau Recapitulatif

df_tableau = df.apply(stat_descri)
df_tableau
#Histogrammes & Boites a Moustaches

def graphiques_stat (variable) :

    fig, axes = plt.subplots(1, 2, figsize=(12, 4))

    sns.histplot(data=df, x=df[variable], ax = axes[0])
    axes[0].set_title(f"Histogramme de {variable}")
    axes[0].set_xlabel(f"{variable}")

    sns.boxplot(data=df,x=df[variable], ax = axes[1])
    axes[1].set_title(f"Boxplot de {variable}")
    axes[1].set_xlabel(f"{variable}")

    plt.tight_layout()
    plt.show()

for c in df.columns :
    graphiques_stat(c)
```

L’analyse des histogrammes et des boîtes à moustaches montre que la variable **Prix_milliers_euros présente une asymétrie positive marquée**, avec une distribution étirée vers les valeurs élevées. Une **transformation logarithmique** apparaît ainsi **pertinente** afin de réduire l’asymétrie et de stabiliser la variance.

La variable **Surface_m2 présente également une légère asymétrie à droite** ; une **transformation logarithmique pourrait être envisagée**, bien que celle-ci soit moins nécessaire.
1.2 Analyse de correlation
- Calculez la matrice de corrélation entre toutes les variables continues
- Créez un graphique de corrélation (heatmap)
- Identifiez les paires de variables fortement corrélées entre elles (risque de multicolinéarité)
#Selection des variables continues

```
df_continue = df.select_dtypes(include = 'number')
df_continue = df_continue.drop(columns=["ID", "Chambres", "Annee_construction", "Etage", "Ascenseur", "Annee_vente"])
df_continue
#Création de la matrice de corrélation

corr_matrix =df_continue.corr()
corr_matrix
#Visualisation de la matrice de corrélation

plt.figure(figsize=(11,5))

sns.heatmap(data = corr_matrix,
        annot=True,
        fmt=".2f",
        cmap="rocket",
        center=0
        )

plt.title("Matrice de Corrélation (variables continues)")
plt.show()
```

Quelle variable semble avoir l'impact le plus fort sur le prix selon la corrélation ? Attention : corrélation *!=* causalité !
L'analyse de la matrice de corrélation montre que la variable **la plus fortement corrélée au prix (Prix_milliers_euros) est la surface du logement (Surface_m2)**, avec un coefficient de corrélation de **0.83**, indiquant une **relation positive forte.**

On observe également une **corrélation modérée** entre la **qualité des écoles (Qualite_ecole)** et le **revenu médian du quartier (Revenu_median_quartier)** : **0.60**, ce qui pourrait suggérer un risque limité de multicolinéarité si ces variables sont intégrées simultanément dans un modèle de régression.

Il convient toutefois de rappeler que la corrélation observée ne permet pas d'établir une relation de causalité.
2 Le Modèle Linéaire : Estimation et Interprétation

2.1 Modèle de régression linéaire simple

Prix*i* = β0 + β1 × Surface*i* + u*i*

Première étape : Régressez le prix sur la surface uniquement.

On utilise les Moindres Carrés Ordinaires (MCO) pour estimer les coefficients β0 et β1.

```
import statsmodels.api as sm
#Regression du prix sur la surface

X = df['Surface_m2']
y = df['Prix_milliers_euros']

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())

#Éstimation des coéfficients (MCO)

#Éstimateurs β0 et β1
beta0 = model.params['const']
beta1 = model.params['Surface_m2']

#Écart-type de chaque coefficient (ˆσˆβj)
std_beta0 = model.bse['const']
std_beta1 = model.bse['Surface_m2']

#Statistique t et la p-valeur
t_beta0 = model.tvalues['const']
t_beta1 = model.tvalues['Surface_m2']
p_beta0 = model.pvalues['const']
p_beta1 = model.pvalues['Surface_m2']

#R2 et le R2 ajusté
r2 = model.rsquared
r2_adj = model.rsquared_adj
```

print(f"β0 = {beta0}, σβ0 = {std_beta0}, t = {t_beta0}, p = {p_beta0}")
print(f"β1 = {beta1}, σβ1 = {std_beta1}, t = {t_beta1}, p = {p_beta1}")
print(f"R² = {r2}, R² ajusté = {r2_adj}")

Que signifie le coefficient ˆ β1 ? Si la surface augmente de 1 m², de combien le prix augmente-t-il en moyenne ?
Le coefficient estimé pour la surface est : **β1 = 5.0428.**
Cela signifie que pour chaque m² supplémentaire de surface, le **prix du bien augmente en moyenne de 5 042,8€**.

La p-valeur est très faible : **p<0,01** donc cela veut dire que le coefficient est **très significatif**

R² = 0,683, donc **68%** de la variation des prix est expliquée par la surface seule.

2.2 Modèle de régression linéaire multiple
#Création du modèle de régression linéaire multiple

X = df[['Surface_m2', 'Chambres', 'Annee_construction', 'Distance_centre_km', 'Etage', 'Ascenseur']]
y = df['Prix_milliers_euros']

X = sm.add_constant(X)

model_multi = sm.OLS(y, X).fit()

print(model_multi.summary())

coefficients = model_multi.params
ecarts_types = model_multi.bse
t_values = model_multi.tvalues
p_values = model_multi.pvalues
r2 = model_multi.rsquared
r2_adj = model_multi.rsquared_adj

print("Coefficients :\n", coefficients)
print("\nÉcarts-types :\n", ecarts_types)
print("\nt-values :\n", t_values)
print("\nP-values :\n", p_values)
print(f"\nR² = {r2}, R² ajusté = {r2_adj}")

1. Tous les coefficients sont-ils significatifs ?

À l'exception de la constante, **tous les coefficients explicatifs sont statistiquement significatifs** au seuil de **5 %**, leurs p-values étant inférieures à 0,05. La constante n'est en revanche pas significative, ce qui n'est pas problématique en soi car elle n'a pas d'interprétation économique directe.

2. Quel est l'impact marginal de chaque variable sur le prix ?

- **Surface_m2** : chaque m² supplémentaire augmente le prix moyen de 4,39 milliers d'euros.

- **Chambres** : chaque chambre supplémentaire augmente le prix moyen de 33,92 milliers d'euros.

- **Annee_construction** : chaque année de construction plus récente augmente le prix de 1,61 milliers d'euros.

- **Distance_centre_km** : chaque km supplémentaire du centre-ville réduit le prix moyen de 6,14 milliers d'euros.

- **Etage** : chaque étage supplémentaire augmente le prix de 12,25 milliers d'euros.

- **Ascenseur** : un appartement avec ascenseur coûte en moyenne 55,51 milliers d'euros de plus qu'un appartement similaire sans ascenseur.

3. Pour la variable Ascenseur : comment interpréter le coefficient ?

Le coefficient associé à la variable Ascenseur indique que, toutes choses égales par ailleurs, un logement disposant d'un ascenseur est **en moyenne plus cher de 55,51 milliers d'euros** qu'un logement comparable sans ascenseur. Il s'agit de l'effet différentiel moyen lié à la présence d'un ascenseur. Le coefficient associé à la présence d'un ascenseur est positif et significatif, mais présente une **erreur standard relativement élevée**, suggérant une estimation moins précise, probablement liée à une forte **hétérogénéité** ou à des **corrélations avec d'autres caractéristiques** du logement.

4. Comment interprétez-vous la différence entre R2 et R2 (adjusted)?

Le **R² indique que 78,9 %** de la variance du prix est expliquée par les variables incluses dans le modèle.

Le R² ajusté, légèrement inférieur **(0,780)**, tient compte du nombre de **variables explicatives et pénalise** l'ajout de variables peu informatives. La faible différence entre les deux suggère que l'ajout de variables améliore la qualité explicative du modèle sans entraîner de sur-ajustement excessif.

2.3 Transformation logarithmique

```
# Vérifier s'il y a des 0 ou des valeurs négatives
print((df[['Surface_m2', 'Chambres', 'Annee_construction', 'Distance_centre_km', 'Etage']]
<= 0).sum())
```

Avant d'appliquer une transformation logarithmique, nous avons **vérifié l'absence de valeurs nulles ou négatives**.

La variable **Étage** présentant des valeurs **nulles** (rez-de-chaussée), nous avons appliqué un **décalage de +1** afin de permettre la transformation logarithmique (en particulier dans le cadre du **modèle log-log**, le logarithme n'étant pas défini en zéro). Cette transformation permet d'inclure l'ensemble des observations tout en conservant l'ordre des étages.
La variable **Ascenseur** étant une variable **binaire**, elle n'est pas concernée par la transformation logarithmique. Par conséquent, il n'est pas nécessaire de vérifier la présence de valeurs nulles ou négatives pour cette variable.

```
# Décaler la numérotation des étages pour le log (rdc devient etage 1)
df['Etage_decale'] = df['Etage'] + 1
#Modélisation en semi-log et en log-log.

# Modèle Linéaire multiple
X_lin = df[['Surface_m2', 'Chambres', 'Annee_construction', 'Distance_centre_km',
'Etage', 'Ascenseur']]
X_lin = sm.add_constant(X_lin)
y = df['Prix_milliers_euros']

model_lin = sm.OLS(y, X_lin).fit()
print("=== Modèle linéaire multiple ===")
print(model_lin.summary())

# Modèle Semi-log : log du prix uniquement
# ln(Prix_i) = β0 + β1 × Surface_m2_i + β2 × Chambres_i + β3 × Annee_construction_i
+ β4 × Distance_centre_km_i + β5 × Etage_i + β6 × Ascenseur_i + u_i

y_log = np.log(df['Prix_milliers_euros'])
X_semi_log = df[['Surface_m2', 'Chambres', 'Annee_construction', 'Distance_centre_km',
'Etage', 'Ascenseur']]
```

```
X_semi_log = sm.add_constant(X_semi_log)

model_semi_log = sm.OLS(y_log, X_semi_log).fit()
print("\n=== Modèle semi-log ===")
print(model_semi_log.summary())

# Modèle Log-log : log du prix et des variables continues
# log(Prix) = log(Surface) + log(Distance) + log(Etage+1) + Chambres + Année +
Ascenseur

X_log = pd.DataFrame({
    'log_Surface': np.log(df['Surface_m2']),
    'Chambres': df['Chambres'],
    'Annee_construction': df['Annee_construction'],
    'log_Distance': np.log(df['Distance_centre_km']),
    'log_Etage': np.log(df['Etage_decale']),
    'Ascenseur': df['Ascenseur']
})

X_log = sm.add_constant(X_log)

model_log_log = sm.OLS(y_log, X_log).fit()
print("\n=== Modèle log-log ===")
print(model_log_log.summary())
#Comparaison des 3 modèles

print("=== Linéaire multiple ===")
print(pd.concat([model_lin.params, model_lin.pvalues],
axis=1).rename(columns={0:'Coef', 1:'p-valeur'}).to_string())
print("R² Linéaire multiple :", model_lin.rsquared)
print("R² ajusté :", model_lin.rsquared_adj)
print("AIC :", model_lin.aic)
print("BIC :", model_lin.bic)

print("\n=== Semi-log ===")
print(pd.concat([model_semi_log.params, model_semi_log.pvalues],
axis=1).rename(columns={0:'Coef', 1:'p-valeur'}).to_string())
print("R² Semi-log :", model_semi_log.rsquared)
print("R² ajusté :", model_semi_log.rsquared_adj)
print("AIC :", model_semi_log.aic)
print("BIC :", model_semi_log.bic)
```

print("\n=== Log-log ===")
print(pd.concat([model_log_log.params, model_log_log.pvalues],
axis=1).rename(columns={0:'Coef', 1:'p-valeur'}).to_string())
print("R² Log-log :", model_log_log.rsquared)
print("R² ajusté :", model_log_log.rsquared_adj)
print("AIC :", model_log_log.aic)
print("BIC :", model_log_log.bic)
1. Comparez les trois modèles.

R² et R² ajusté :
- Semi-log > Linéaire multiple > Log-log
- Le semi-log explique légèrement mieux la variabilité du prix (0.783 vs 0.780).

Significativité des variables :
- **Linéaire multiple** et **semi-log** : toutes les variables explicatives significatives.
- **Log-log** : Année de construction n'est plus significative → moins adapté.

Interprétation des coefficients :
- **Linéaire** : l'effet absolu d'une unité (ex : 1 m² de surface → +4.39 k€).
- **Semi-log** : effet relatif sur le prix (ex : 1 m² → +0.21 % du prix moyen).
- **Log-log** : élasticité, utile si on veut comparer des effets proportionnels mais moins adapté ici car certaines variables sont petites ou nulles (Etage RDC, Chambres).

2. Quel modèle semble le plus approprié et pourquoi ?.

Le **modèle semi-log** apparaît comme le plus approprié. Il présente le **R² ajusté le plus élevé** et l'ensemble des variables explicatives y sont **statistiquement significatives**. La comparaison des critères d'information **AIC** et **BIC**, effectuée entre les modèles estimés sur le logarithme du prix, montre que le modèle semi-log présente des **valeurs plus faibles** que le modèle log-log. Cela indique que le modèle semi-log offre le **meilleur compromis entre qualité d'ajustement et complexité**, confirmant ainsi son choix comme spécification privilégiée. Enfin, le modèle semi-log permet une interprétation économique pertinente en termes de **variations relatives du prix**.
#Nous sauvegardons notre modèle sélectionné dans la variable best_model pour faciliter les utilisations futures

best_model = model_semi_log
3 Diagnostics du Modèle
3.1 Multicolinéarité

Calculez les VIF (Variance Inflation Factor) pour chaque variable
X = df[['Surface_m2', 'Chambres', 'Annee_construction', 'Distance_centre_km', 'Etage', 'Ascenseur']].copy()
X = X.dropna()
X = sm.add_constant(X)

X.head()
from statsmodels.stats.outliers_influence import variance_inflation_factor
#Calcul des VIF

vif = pd.DataFrame({
  "Variable" : X.columns,
  "VIF" : [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
})

vif
Y a-t-il des variables avec un VIF élevé ?

Non, il n'y a pas de variable avec un VIF élevé. À l'exception de la **constante**, dont le VIF **n'est pas interprété**, toutes les variables **explicatives** présentent des VIF proches de 1, indiquant une multicolinéarité très faible.

Faut-il en supprimer certaines ?

Non. Étant donné l'absence de multicolinéarité, **il n'est pas nécessaire de supprimer des variables explicatives**. Les coefficients sont estimés avec une variance faible et une bonne précision.

Definir le biais de variable omise.

Le biais de variable omise apparaît lorsqu'une **variable pertinente pour expliquer la variable dépendante est exclue du modèle alors qu'elle est corrélée avec une variable incluse**. Dans ce cas, les coefficients estimés sont **biaisés** et ne reflètent plus l'effet causal réel.

Supprimer une variable uniquement pour réduire la multicolinéarité peut donc détériorer la validité économique du modèle.

Le biais existe si et seulement si :

- la variable omise affecte Y
- elle est corrélée avec une variable explicative incluse

Si une des deux conditions manque, il n'y a pas de biais.
4 Tests et Inférence
#Affichage des paramètres Distance_centre_km

```
m = best_model

beta_distance = m.params["Distance_centre_km"]
t_distance = m.tvalues["Distance_centre_km"]

beta_distance, t_distance

p_two_sided = m.pvalues["Distance_centre_km"]

p_two_sided
```

Hypothèses :

- H0 : βdistance ≥ 0 (pas d'effet négatif)
- H1 : βdistance < 0 (effet négatif)

Dans statsmodels, la p-value affichée dans le summary est bilatérale.
Pour une hypothèse unilatérale à gauche, on convertit :

- Si le t-stat est négatif (et donc le coef est négatif) : p(one-sided) = p(two-sided) / 2
- Si le t-stat est positif : p(one-sided) = 1 − p(two-sided) / 2

```
coef = m.params["Distance_centre_km"]
t_stat = m.tvalues["Distance_centre_km"]
p_two = m.pvalues["Distance_centre_km"]

# p-value unilatérale à gauche (H1 : beta < 0)
p_one = (p_two / 2) if (t_stat < 0) else (1 - p_two / 2)

print("=== Test sur Distance_centre_km (modèle semi-log) ===")
print("Coef beta =", coef)
print("t_stat    =", t_stat)
print("p-value (two-sided) =", p_two)
print("p-value (one-sided, H1: beta < 0) =", p_one)

# Interprétation en %
print("\nInterprétation :")
print(f"Approx : +1 km = {100*coef:.2f}% de variation du prix (approx).")
```

print(f"Exact : +1 km = {(np.exp(coef)-1)*100:.2f}% de variation du prix (exact).")
1. Testez l'hypothèse que la distance au centre a un effet négatif sur le prix. Quelle est la p-value ?

Si p-value < 0.05, on rejette H_0.
Ici, p = 9.5e-10 => < 0.05

Donc, on rejette H_0.

Le coefficient associé à la distance au centre est **négatif** et hautement significatif dans le modèle semi-log.

Le test unilatéral (H1 : β < 0) conduit à une p-**value de l'ordre de 9.5e-10**, largement inférieure au seuil de 5 %. On rejette donc l'hypothèse nulle.

Une **augmentation d'un kilomètre** de la distance au centre entraîne une **baisse du prix du logement d'environ 0,30 %**, toutes choses égales par ailleurs.
L'approximation et l'effet exact sont quasiment identiques en raison de la faible valeur du coefficient.
2. Testez l'hypothèse que tous les coefficients (sauf constante) soient nuls. Testez si l'ajout des variables : Qualite_ecole et Revenu_median_quartier améliore significativement le modèle

```
from statsmodels.stats.anova import anova_lm

# Modèle sans les nouvelles variables (Qualite_ecole et Revenu_median_quartier)

y_log = np.log(df["Prix_milliers_euros"])

X_without = df[['Surface_m2', 'Chambres', 'Annee_construction',
            'Distance_centre_km', 'Etage', 'Ascenseur']]
X_without = sm.add_constant(X_without)
model_without = sm.OLS(y_log, X_without).fit()

# Modèle avec les nouvelles variables
X_with = df[['Surface_m2', 'Chambres', 'Annee_construction',
          'Distance_centre_km', 'Etage', 'Ascenseur',
          'Qualite_ecole', 'Revenu_median_quartier']]
X_with = sm.add_constant(X_with)
model_with = sm.OLS(y_log, X_with).fit()

# Test F pour comparer les modèles
```

anova_results = anova_lm(model_without, model_with)
print("Test F pour l'ajout de Qualite_ecole et Revenu_median_quartier :")
print(anova_results)

# Test F global pour le modèle de base
print("\nTest F global (modèle semi-log sans nouvelles variables) :")
print("F-statistic :", model_without.fvalue)
print("F p-value   :", model_without.f_pvalue)

Résultats du test F global :

Le test F global pour l'hypothèse nulle que tous les coefficients (sauf la constante) sont nuls donne une statistique F de 90.56 avec une p-value extrêmement faible (3.309e-46). **On rejette donc H_0** : au moins un coefficient explicatif a un effet significatif sur le prix du logement.

Test d'amélioration du modèle avec Qualite_ecole et Revenu_median_quartier :

Le test F pour comparer le modèle de base (sans Qualite_ecole et Revenu_median_quartier) au modèle étendu (avec ces variables) donne une statistique F de 29.30 avec une p-value de 2.278e-11. On rejette donc l'hypothèse nulle selon laquelle l'ajout de ces variables n'améliore pas le modèle. Les variables Qualite_ecole et Revenu_median_quartier contribuent donc de manière significative à expliquer la variance du prix.
# Sélection du modèle retenu après les tests F
best_model = model_with
Au vu du **test F**, nous retenons le **modèle semi-log enrichi** incluant les variables Qualite_ecole et Revenu_median_quartier pour la suite de l'analyse.
3. Pourquoi ne peut-on pas simplement utiliser plusieurs tests T pour tester plusieurs restrictions simultanément ?

On ne peut pas utiliser plusieurs tests t pour tester plusieurs restrictions simultanément car chaque test t est un **test individuel** et ne contrôle pas le **risque global d'erreur de type I**.

De plus, les coefficients étant **estimés conjointement**, ils peuvent être **corrélés entre eux**. Le test F permet de tester simultanément plusieurs restrictions en tenant compte de ces corrélations et du nombre de restrictions, ce qui en fait l'outil approprié.
4.1 Stabilité structurelle

Testez si le COVID a un effet sur le marché immobilier en utilisant la méthode de votre choix.

```
# Examin des statistiques des années de vente

print("Statistiques des années de vente :")
print(df['Annee_vente'].describe())
#Définition d'un point de rupture et diviser les données en périodes pré et post-COVID

break_year = 2020

df_pre = df[df['Annee_vente'] < break_year]
df_post = df[df['Annee_vente'] >= break_year]

print(f"\nObservations avant {break_year} : {len(df_pre)}")
print(f"Observations après {break_year} : {len(df_post)}")
#Estimation des modèles de régression semi-log pour chaque période (pré et
post-COVID)

#Pré-COVID
y_log_pre = np.log(df_pre['Prix_milliers_euros'])
X_pre = df_pre[['Surface_m2', 'Chambres', 'Annee_construction',
            'Distance_centre_km', 'Etage', 'Ascenseur',
            'Qualite_ecole', 'Revenu_median_quartier']]
X_pre = sm.add_constant(X_pre)
model_pre = sm.OLS(y_log_pre, X_pre).fit()

#Post-COVID
y_log_post = np.log(df_post['Prix_milliers_euros'])
X_post = df_post[['Surface_m2', 'Chambres', 'Annee_construction',
             'Distance_centre_km', 'Etage', 'Ascenseur',
             'Qualite_ecole', 'Revenu_median_quartier']]
X_post = sm.add_constant(X_post)
model_post = sm.OLS(y_log_post, X_post).fit()
#Estimation du modèle de régression semi-log sur l'ensemble des données

y_log_full = np.log(df['Prix_milliers_euros'])
X_full = df[['Surface_m2', 'Chambres', 'Annee_construction',
          'Distance_centre_km', 'Etage', 'Ascenseur',
          'Qualite_ecole', 'Revenu_median_quartier']]
X_full = sm.add_constant(X_full)
model_full = sm.OLS(y_log_full, X_full).fit()
Pour cet exercice nous avonc décidés d'utiliser la **méthode Chow**.
```

Le test de Chow permet de tester **l'existence d'une rupture structurelle** à une date donnée. Autrement dit, il permet de répondre à la question :

"Est-ce que la relation entre la variable expliquée et ses déterminants est la même avant et après un événement donné ?"

En comparant les coefficients estimés **avant** et **après** 2020, il permet d'évaluer si le COVID-19 a modifié la relation entre le prix des logements et leurs caractéristiques.
#Méthode choisie : Chow. - Détection de la rupture structurelle

```
k = X_full.shape[1]

n1 = len(df_pre)
n2 = len(df_post)

SSE_full = model_full.ssr
SSE1 = model_pre.ssr
SSE2 = model_post.ssr

F_chow = ((SSE_full - (SSE1 + SSE2)) / k) / ((SSE1 + SSE2) / (n1 + n2 - 2 * k))

from scipy.stats import f
p_chow = 1 - f.cdf(F_chow, k, n1 + n2 - 2 * k)
# Interprétation des résultats du test de Chow

print(f"\nChow F-statistic : {F_chow:.4f}")
print(f"Chow p-value : {p_chow:.4e}")

if p_chow < 0.05:
    print("Rejet de H₀ : il y a une rupture structurelle significative associée à la période COVID-19.")
else:
    print("Pas de rejet de H₀ : pas de rupture structurelle significative.")
```

Si vous trouvez une rupture structurelle, discutez des implications pour votre analyse.

Puisqu'il y a une **rupture structurelle significative** (p-value < 0.05), les relations entre les variables explicatives et le prix ont changé entre les périodes pré et post-COVID. Cela signifie que **le modèle estimé sur l'ensemble des données ne capture pas correctement les dynamiques** spécifiques à chaque période.

Faut-il estimer des modèles séparés ?

Oui, il est recommandé d'estimer des modèles séparés pour les périodes pré-COVID (avant 2020) et post-COVID (2020 et après).

Cela permet :
- D'**analyser les effets différentiels du COVID** sur les déterminants du prix immobilier.
- D'**éviter les biais** dus à la non-stationnarité des paramètres.
- De fournir des **prédictions plus précises** pour chaque période.

Par exemple, on pourrait comparer les coefficients pour voir si l'impact de la distance au centre ou de la surface a évolué après la pandémie.
5 Hétéroscédasticité et Autocorrélation

1. Observez graphiquement si les résidus suivent un pattern.
```
# Graphique des résidus vs valeurs ajustées pour détecter un pattern (hétéroscédasticité)
plt.figure(figsize=(10, 6))
plt.scatter(best_model.fittedvalues, best_model.resid, alpha=0.6)
plt.axhline(y=0, color='red', linestyle='--')
plt.xlabel('Valeurs ajustées (fitted values)')
plt.ylabel('Résidus')
plt.title('Résidus vs Valeurs ajustées (modèle semi-log, log(Prix))')
plt.show()

# Graphique des résidus vs une variable explicative (ex: Surface_m2)
plt.figure(figsize=(10, 6))
plt.scatter(df['Surface_m2'], best_model.resid, alpha=0.6)
plt.axhline(y=0, color='red', linestyle='--')
plt.xlabel('Surface en m²')
plt.ylabel('Résidus')
plt.title('Résidus vs Surface (modèle semi-log, log(Prix))')
plt.show()

print(best_model.model.exog_names)
```
L’observation graphique des résidus en fonction des valeurs ajustées et de la surface **ne révèle pas de pattern systématique**. Les résidus sont centrés autour de **zéro** et leur variance semble globalement **constante**. Il n’y a pas de signe évident d’hétéroscédasticité.
2. Testez l’hétéroscédasticité et corrigez-la.
```
# Test de Breusch-Pagan pour l'hétéroscédasticité
```

```
from statsmodels.stats.diagnostic import het_breuschpagan

bp_test = het_breuschpagan(best_model.resid, best_model.model.exog)

print("Test de Breusch-Pagan :")
print(f"LM Statistic: {bp_test[0]:.4f}")
print(f"LM p-value: {bp_test[1]:.4e}")
print(f"F Statistic: {bp_test[2]:.4f}")
print(f"F p-value: {bp_test[3]:.4e}")

if bp_test[1] < 0.05:
    print("Rejet de H₀ : présence d'hétéroscédasticité.")
else:
    print("Pas de rejet de H₀ : pas d'hétéroscédasticité significative.")

# Correction : écarts-types robustes (HC3)

model_robust = best_model.get_robustcov_results(cov_type='HC3')

print("\nCoefficients avec écarts-types robustes :")
print(model_robust.summary().tables[1])

# Correction : WLS (poids inversement proportionnels aux valeurs ajustées)

y_used = best_model.model.endog
X_used = best_model.model.exog

weights = 1 / (best_model.fittedvalues ** 2)

model_wls = sm.WLS(y_used, X_used, weights=weights).fit()

print("\nModèle WLS (basé sur le modèle enrichi) :")
print(model_wls.summary().tables[1])
```

Hypothèses

H_0 : variance constante → pas d'hétéroscédasticité
H_1 : variance non constante → hétéroscédasticité

si p-value < 0.05 → on rejette H_0
si p-value ≥ 0.05 → on ne rejette pas H_0

Le test de Breusch–Pagan **ne permet pas de rejeter l'hypothèse nulle d'homoscédasticité** (p-value > 0.05). Il n'y a donc pas de preuve d'hétéroscédasticité significative dans le modèle.

Des **écarts-types robustes** ont néanmoins été calculés à titre de vérification. Les résultats et conclusions du modèle semi-log peuvent ainsi être interprétés de manière fiable.

Les conclusions issues du modèle semi-log apparaissent robustes aux différentes méthodes d'estimation.
3. Comparez les MCO standard, les MCO avec écarts-types robustes, et WLS.
Comparaison des coefficients et écarts-types pour les trois méthodes

```
import pandas as pd
from statsmodels.stats.stattools import durbin_watson

# --- Conversion propre des paramètres WLS en Series avec index ---

wls_params = pd.Series(
    model_wls.params,
    index=best_model.params.index
)

wls_bse = pd.Series(
    model_wls.bse,
    index=best_model.params.index
)

# --- Tableau comparatif ---

comparison = pd.DataFrame({
    "Variable": best_model.params.index,
    "OLS_coef": best_model.params,
    "OLS_se": best_model.bse,
    "Robust_coef": best_model.params,          # identiques aux OLS
    "Robust_se": model_robust.bse,
    "WLS_coef": wls_params,
    "WLS_se": wls_bse,
})

print("Comparaison des estimations :")
print(comparison.to_string(index=False))
```

```
# --- R² des modèles ---

print(f"\nR² MCO Standard : {best_model.rsquared:.4f}")
print(f"R² MCO Robust   : {best_model.rsquared:.4f}")
print(f"R² WLS          : {model_wls.rsquared:.4f}")

# --- Test d'autocorrélation (Durbin–Watson) ---

dw_stat = durbin_watson(best_model.resid)
print(f"\nStatistique de Durbin-Watson (MCO Standard) : {dw_stat:.4f}")

if dw_stat < 1.5:
   print("Indication d'autocorrélation positive.")
elif dw_stat > 2.5:
   print("Indication d'autocorrélation négative.")
else:
   print("Pas d'autocorrélation significative.")
```

Interprétation :

Les coefficients estimés sont très **similaires** entre le MCO standard, le MCO avec écarts-types robustes et le modèle WLS, indiquant que les estimations ne sont pas sensibles au mode de correction.

Le test de **Breusch–Pagan** ne met pas en évidence d'hétéroscédasticité significative (p-value > 0.05). Les écarts-types robustes sont légèrement plus élevés, mais les conclusions d'inférence restent inchangées.

Le modèle WLS présente un **R² légèrement supérieur**, sans amélioration substantielle.

En l'absence d'hétéroscédasticité et d'autocorrélation significatives, le **MCO standard est approprié**. Par prudence, l'utilisation d'écarts-types robustes peut néanmoins être retenue.
5.1 Test d'autocorrélation

Testez l'autocorrélation

```
# Test d'autocorrélation de Breusch-Godfrey (pour ordre 1)

from statsmodels.stats.diagnostic import acorr_breusch_godfrey
```

```
bg_test = acorr_breusch_godfrey(best_model, nlags=1)
print("Test de Breusch-Godfrey (autocorrélation ordre 1) :")
print(f"LM Statistic: {bg_test[0]:.4f}")
print(f"LM p-value: {bg_test[1]:.4e}")
print(f"F Statistic: {bg_test[2]:.4f}")
print(f"F p-value: {bg_test[3]:.4e}")

if bg_test[1] < 0.05:
    print("Rejet de H₀ : présence d'autocorrélation d'ordre 1.")
    autocorr_present = True
else:
    print("Pas de rejet de H₀ : pas d'autocorrélation d'ordre 1 significative.")
    autocorr_present = False

# # Rappel : hétéroscédasticité testée avec Breusch–Pagan (True si p<0.05)
hetero_present = bp_test[1] < 0.05

if hetero_present and autocorr_present:
    print("\nPuisque autocorrélation détectée, utilisation des écarts-types de
Newey-West.")
    model_nw = best_model.get_robustcov_results(cov_type='HAC', maxlags=1)
    print("\nCoefficients avec écarts-types Newey-West :")
    print(model_nw.summary().tables[1])
else:
    print("\nPas besoin d'écarts-types Newey-West (absence d'autocorrélation et
d'hétéroscédasticité).")
```

Si vous détectez à la fois hétéroscédasticité et autocorrélation, utilisez les écarts-types de Newey-West qui sont robustes aux deux problèmes.

Conformément à la consigne, les écarts-types de **Newey–West** ne sont requis qu'en présence simultanée **d'hétéroscédasticité** et **d'autocorrélation**.

Or, le test de **Breusch–Godfrey** ne met pas en évidence **d'autocorrélation d'ordre 1** (p-value > 0,05) et le test de **Breusch–Pagan** ne détecte pas **d'hétéroscédasticité significative**.

Il n'est donc **pas nécessaire de recourir aux écarts-types de Newey–West**. Le modèle MCO standard, éventuellement accompagné d'écarts-types robustes à titre de précaution, peut être retenu.

6 Endogénéité et Variables Instrumentales

6.1 Sources d'endogénéité

Quelles sont les sources possibles d'endogénéité dans notre contexte ?

Les sources possibles d'endogénéité incluent :
- **Variables omises** : Des facteurs non observés influencent à la fois la variable dépendante (prix) et les variables explicatives.
- **Erreur de mesure** : Les variables sont mesurées avec erreur, créant une corrélation entre l'erreur et la variable.
- **Simultanéité** : La variable dépendante influence les variables explicatives (relation bidirectionnelle).
- **Sélection endogène** : Le choix des observations n'est pas aléatoire.

Dans le contexte immobilier, des variables omises comme la qualité générale du quartier ou des facteurs économiques locaux peuvent être problématiques.

La variable Qualite_ecole est-elle potentiellement endogène ? Pourquoi ?

La variable Qualite_ecole pourrait être endogène. D'une part, les quartiers avec de meilleures écoles attirent des familles aisées, ce qui augmente les prix immobiliers (effet causal direct). D'autre part, les quartiers chers peuvent investir davantage dans l'éducation, améliorant la qualité des écoles (causalité inverse). De plus, des facteurs omis comme le revenu moyen du quartier influencent à la fois la qualité des écoles et les prix. Ainsi, Qualite_ecole est potentiellement endogène, rendant les estimations MCO biaisées.
6.2 Estimation par Variables Instrumentales

Proposition d'instrument : Introduisez la variable Distance_universite (distance à l'université la plus proche).

1. Argumentez pourquoi cette variable pourrait être un bon instrument pour Qualite_ecole.

La variable Distance_universite peut être envisagée comme instrument potentiel de Qualite_ecole.

Elle est pertinente car la proximité d'universités **favorise la concentration de ressources éducatives et de capital humain**, ce qui peut améliorer la qualité des écoles locales.

En revanche, la validité de l'instrument **repose sur l'hypothèse d'exclusion** selon laquelle la distance à l'université n'affecte pas directement les prix immobiliers une fois contrôlés les autres facteurs. Cette hypothèse peut être discutée, car la proximité

d’une université peut également **influencer les prix via la demande étudiante ou l’attractivité du quartier**.

La validité de l’instrument doit donc être interprétée avec prudence et idéalement testée.

2. Construisez une estimation en deux étapes (2SLS)
Vérifier la présence de Distance_universite dans les données

```
print("Colonnes disponibles :", df.columns.tolist())
if 'Distance_universite' in df.columns:
   print("Distance_universite présente.")
else :
   print("Distance_universite absente.")
```

Première étape du 2SLS : Régression de la variable endogène (Qualite_ecole) sur les instruments et variables exogènes
La première étape sert uniquement à isoler une variation exogène de la qualité des écoles, afin de pouvoir mesurer son effet causal sur le prix des logements.
Première étape : régresser Qualite_ecole sur Distance_universite et autres exogènes
Cela permet d'obtenir les valeurs prédites de Qualite_ecole, purgées de l'endogénéité

```
X_first_stage = df[['Distance_universite', 'Surface_m2', 'Chambres',
'Annee_construction', 'Distance_centre_km', 'Etage', 'Ascenseur',
'Revenu_median_quartier']]
X_first_stage = sm.add_constant(X_first_stage)

first_stage_model = sm.OLS(df['Qualite_ecole'], X_first_stage).fit()

print("Première étape - Résumé :")
print(first_stage_model.summary())
```

La première étape du 2SLS montre que la distance à l’université est **fortement et significativement corrélée** à la qualité des écoles (p-value < 0,01), ce qui confirme la **pertinence de l’instrument**.

Les valeurs prédites de la qualité des écoles sont ensuite utilisées dans la **seconde étape** afin d’identifier son effet causal sur le prix des logements.
Deuxième étape du 2SLS : Utiliser les valeurs prédites de Qualite_ecole dans la régression principale
Obtenir les valeurs prédites de Qualite_ecole

```
qualite_ecole_hat = first_stage_model.fittedvalues
```

```
# Deuxième étape : régresser le prix sur les prédits et autres exogènes

X_second_stage = pd.DataFrame({
    'Qualite_ecole_hat': qualite_ecole_hat,
    'Surface_m2': df['Surface_m2'],
    'Chambres': df['Chambres'],
    'Annee_construction': df['Annee_construction'],
    'Distance_centre_km': df['Distance_centre_km'],
    'Etage': df['Etage'],
    'Ascenseur': df['Ascenseur'],
    'Revenu_median_quartier': df['Revenu_median_quartier']
})
X_second_stage = sm.add_constant(X_second_stage)

second_stage_model = sm.OLS(np.log(df['Prix_milliers_euros']), X_second_stage).fit()

print("Deuxième étape - Résumé 2SLS :")
print(second_stage_model.summary())
L’estimation **2SLS** suggère un **effet positif** de la qualité des écoles sur les prix immobiliers. Toutefois, cet effet n’est pas statistiquement significatif, indiquant qu’une fois corrigé du biais d’endogénéité, **l’impact causal de la qualité des écoles ne peut être établi de manière robuste** dans cet échantillon.

La **perte de significativité par rapport au modèle MCO** est cohérente avec la correction d’un biais d’endogénéité potentiellement positif dans l’estimation initiale.
**3. Testez la validité des instruments avec la méthode de votre choix.**
f_stat_instrument = first_stage_model.f_test("Distance_universite = 0")

print("Test de pertinence de l'instrument (F-stat pour Distance_universite) :")
print(f"F-statistic: {f_stat_instrument.statistic:.4f}")
print(f"p-value: {f_stat_instrument.pvalue:.4e}")

if f_stat_instrument.statistic > 10:
    print("Instrument pertinent (F > 10).")
else:
    print("Instrument faible.")

# Test d'endogénéité : Durbin–Wu–Hausman

# Résidu
df["vhat"] = first_stage_model.resid
```

```
y = np.log(df["Prix_milliers_euros"])

X_dwh = pd.concat([
    df[['Surface_m2',
        'Chambres',
        'Annee_construction',
        'Distance_centre_km',
        'Etage',
        'Ascenseur',
        'Qualite_ecole',
        'Revenu_median_quartier']],
    df[['vhat']]
], axis=1)

X_dwh = sm.add_constant(X_dwh)

dwh_model = sm.OLS(y, X_dwh).fit()

print(dwh_model.summary())

print("\nTest d'endogénéité (Durbin–Wu–Hausman) :")
print("p-value sur vhat =", dwh_model.pvalues["vhat"])

if dwh_model.pvalues["vhat"] < 0.05:
    print("Rejet H0 : Qualite_ecole est endogène → IV/2SLS justifié.")
else:
    print("Pas de rejet H0 : pas de preuve d'endogénéité → MCO acceptable.")
```

Le test de **Durbin–Wu–Hausman** ne **permet pas de rejeter l'hypothèse d'exogénéité** de la variable Qualite_ecole. En l'absence d'évidence statistique d'endogénéité, **l'estimateur MCO peut être retenu**.

Celui-ci présente en outre une **plus grande précision** que l'estimation 2SLS, ce qui justifie son utilisation dans ce contexte.
4. Comparez les coefficients MCO et IV. Y a-t-il des différences importantes ?

```
print(
    f"MCO (modèle enrichi) : {best_model.params['Qualite_ecole']:.4f} "
    f"(SE: {best_model.bse['Qualite_ecole']:.4f})"
)

print(
    f"IV (2SLS) : {second_stage_model.params['Qualite_ecole_hat']:.4f} "
```

```
    f"(SE: {second_stage_model.bse['Qualite_ecole_hat']:.4f})"
)
```
Les estimations MCO et IV conduisent à des **coefficients de signe identique**, mais de **magnitude différente**.

L'estimateur **IV** est nettement **moins précis**, comme attendu, et l'effet estimé de Qualite_ecole devient **faible et non significatif** une fois corrigé d'un biais d'endogénéité potentiel.

Toutefois, conformément au test de **Durbin–Wu–Hausman**, **aucune endogénéité** significative de Qualite_ecole n'est détectée.

Il n'y a donc pas de justification statistique à privilégier l'estimation IV, et **l'estimateur MCO du modèle enrichi peut être retenu** pour l'analyse.
7 Régularisation
1. Estimez un modèle Ridge avec différentes valeurs de λ. Analysez et commentez l'évolution des coefficients.
```
y = np.log(df["Prix_milliers_euros"])

X = df[[
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
    ]]
```
Important : Avant d'appliquer Ridge ou Lasso, standardisez toutes les variables (moyenne 0, écart-type 1) car la pénalité dépend de l'échelle des coefficients.
```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_std = scaler.fit_transform(X)
```
Nous venons de standardiser les variables X de notre nouveau modèle semi-log
```
from sklearn.linear_model import Ridge

lambdas = [0.01, 0.1, 1, 10, 100]

coefs = []
```

```
for l in lambdas :
   ridge = Ridge(alpha=l)
   ridge.fit(X_std, y)
   coefs.append(ridge.coef_)
coefs = np.array(coefs)

plt.figure(figsize=(10,6))

for j, var in enumerate(X.columns):
   plt.plot(lambdas, coefs[:, j], label=var)

plt.xscale("log")
plt.xlabel("lambda (λ)")
plt.ylabel("Coefficient")
plt.title("Ridge : évolution des coefficients quand λ augmente")
plt.legend(bbox_to_anchor=(1.05, 1), loc="upper left")
plt.show()
```

Nous estimons un modèle **Ridge** pour **différentes valeurs** du paramètre de régularisation λ (0.01, 0.1, 1, 10, 100).

Lorsque **λ augmente**, la pénalité appliquée aux coefficients devient **plus forte**.

On observe que **l'ensemble des coefficients diminue progressivement** en valeur absolue lorsque λ augmente, sans jamais devenir exactement nul.

Les variables les plus influentes (comme la surface) voient leur coefficient se **réduire** fortement, tandis que les variables moins informatives sont rapidement rapprochées de **zéro**.

Le modèle Ridge permet ainsi de **stabiliser les estimations et de limiter l'impact de la multicolinéarité**, tout en conservant toutes les variables dans le modèle.
2. Estimez un modèle Lasso pour différentes valeurs de λ. Analysez et commentez la manière dont les coefficients se modifient en fonction de λ.

```
from sklearn.linear_model import Lasso

lambdas = [0.01, 0.1, 1, 10, 100]

coefs_lasso = []

for l in lambdas :
   lasso = Lasso(alpha=l, max_iter=10000)
```

```
    lasso.fit(X_std, y)
    coefs_lasso.append(lasso.coef_)
coefs_lasso = np.array(coefs_lasso)

plt.figure(figsize=(10,6))

for j, var in enumerate(X.columns):
    plt.plot(lambdas, coefs_lasso[:, j], label=var)

plt.xscale("log")
plt.xlabel("lambda (λ)")
plt.ylabel("Coefficient")
plt.title("Lasso : évolution des coefficients quand λ augmente")
plt.legend(bbox_to_anchor=(1.05, 1), loc="upper left")
plt.axhline(0, color="black", linewidth=0.8)
plt.show()
```

Quand le paramètre **λ augmente**, le modèle Lasso **pénalise de plus en plus les coefficients**.

Dans notre cas, cette pénalisation est **rapidement très forte** : dès que λ devient modéré, **tous les coefficients sont ramenés à zéro**.

Cela signifie que le **Lasso ne conserve aucune variable explicative** et privilégie un modèle très parcimonieux.

Économétriquement, cela indique que, compte tenu des données, **aucune variable n'a un effet suffisamment robuste** pour être conservée sous une pénalisation élevée.
3. Choissisez la valeur du paramètre λ. Pour cela, utilisez la validation croisée 10-fold pour choisir λ optimal.
#Ridge

```
from sklearn.model_selection import KFold
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import RidgeCV

y = np.log(df["Prix_milliers_euros"])
X = df[['Surface_m2','Chambres','Annee_construction',
       'Distance_centre_km','Etage','Ascenseur','Qualite_ecole', 'Revenu_median_quartier']]

lambdas = np.logspace(-3, 3, 50)
```

```
cv =KFold(n_splits=10, shuffle=True, random_state=42)

ridge_CV = Pipeline([
   ("Scaler", StandardScaler()),
   ("model", RidgeCV(alphas=lambdas, cv=cv))
])

ridge_CV.fit(X,y)

best_lambda_ridge = ridge_CV.named_steps["model"].alpha_
print("Meilleur lambda (Ridge) =", best_lambda_ridge)
```

Le paramètre de régularisation λ est choisi à l'aide d'une validation croisée 10-fold.

Pour un ensemble de **50 valeurs candidates de λ**, le modèle Ridge est estimé sur 9 folds et évalué sur le fold restant, et l'erreur de prédiction moyenne hors échantillon est calculée.

Le λ qui minimise cette erreur est retenu comme λ optimal.

Dans notre cas, la validation croisée 10-fold conduit à un **λ optimal égal à λ ≈ 6.25**.

En d'autres termes, la validation croisée indique qu'une **pénalisation modérée** des coefficients améliore la performance prédictive du modèle.

Le λ optimal égal à 6.25 permet de **réduire la variance des estimateurs** sans introduire un biais excessif.

```
#Lasso

from sklearn.linear_model import LassoCV

y = np.log(df["Prix_milliers_euros"])
X = df[['Surface_m2','Chambres','Annee_construction',
       'Distance_centre_km','Etage','Ascenseur','Qualite_ecole','Revenu_median_quartier']]

lambdas = np.logspace(-3, 3, 50)

cv =KFold(n_splits=10, shuffle=True, random_state=42)

lasso_cv = Pipeline([
   ("Scaler", StandardScaler()),
   ("model", LassoCV(alphas=lambdas, cv=cv, max_iter=100000))
```

])

lasso_cv.fit(X, y)

best_lambda_lasso = lasso_cv.named_steps["model"].alpha_
print("Meilleur lambda (Lasso) =", best_lambda_lasso)
Le paramètre de régularisation λ est choisi à l'aide d'une validation croisée 10-fold.

Pour un ensemble de **50 valeurs candidates de λ**, le modèle Lasso est estimé sur 9 folds et évalué sur le fold restant, et l'erreur de prédiction moyenne hors échantillon est calculée.

Le λ qui minimise cette erreur est retenu comme λ optimal.

Dans notre cas, la validation croisée 10-fold conduit à un **λ optimal égal à λ = 0.001**.

En d'autres termes, la validation croisée indique qu'une **pénalisation très faible est optimale pour le modèle Lasso**.

Cela suggère que les variables explicatives contiennent toutes une **information pertinente** pour expliquer le prix immobilier et que la sélection automatique de variables induite par le **Lasso n'est pas nécessaire dans ce contexte**.

Un λ plus élevé conduirait à mettre certains coefficients à zéro, ce qui dégraderait la performance prédictive du modèle.
Réponse à la question :

- Pour **Ridge**, le λ optimal est **λ ≈ 6.25**, indiquant une pénalisation modérée des coefficients.
- Pour **Lasso**, le λ optimal est **λ = 0.001**, ce qui correspond à une pénalisation très faible.

Cela suggère que la **sélection de variables n'est pas nécessaire ici** et que les données favorisent un modèle proche du MCO.
4. Comparez les résultats de trois modèles sur votre jeu de données. Divisez en train et test (80%- 20%) et comparez les erreurs de prédiction (RMSE) sur l'échantillon de test.
from sklearn.model_selection import train_test_split

y = np.log(df["Prix_milliers_euros"])

```
X = df[['Surface_m2','Chambres','Annee_construction',
       'Distance_centre_km','Etage','Ascenseur','Qualite_ecole','Revenu_median_quartier']]

X_train, X_test, y_train, y_test = train_test_split(
   X, y,
   test_size=0.2,
   random_state=42
)

print("Taille train :", X_train.shape[0])
print("Taille test :", X_test.shape[0])
import statsmodels.api as sm

X_train_OLS = sm.add_constant(X_train)

OLS_model = sm.OLS(y_train, X_train_OLS).fit()
ridge_model = Pipeline([
   ("scaler", StandardScaler()),
   ("model", Ridge(alpha=best_lambda_ridge))
])

ridge_model.fit(X_train, y_train)
lasso_model = Pipeline([
   ("scaler", StandardScaler()),
   ("model", Lasso(alpha=best_lambda_lasso, max_iter=100000))
])

lasso_model.fit(X_train, y_train)
from sklearn.metrics import mean_squared_error

#MCO
X_test_OLS = sm.add_constant(X_test)
y_pred_OLS = OLS_model.predict(X_test_OLS)

rmse_OLS = np.sqrt(mean_squared_error(y_test, y_pred_OLS))

#Ridge
y_pred_ridge = ridge_model.predict(X_test)
rmse_ridge = np.sqrt(mean_squared_error(y_test, y_pred_ridge))

#Lasso
y_pred_lasso = lasso_model.predict(X_test)
```

rmse_lasso = np.sqrt(mean_squared_error(y_test, y_pred_lasso))

print("RMSE sur l'échantillon de test :")
print(f"MCO   : {rmse_OLS:.4f}")
print(f"Ridge : {rmse_ridge:.4f}")
print(f"Lasso : {rmse_lasso:.4f}")
La comparaison des performances prédictives sur l'échantillon de test montre que le **modèle Ridge obtient la plus faible RMSE**, indiquant une **meilleure capacité de généralisation**.

Le modèle **MCO** présente une **erreur légèrement plus élevée**, tandis que le **Lasso**, bien que parcimonieux, **n'améliore pas la performance prédictive** dans ce cas.

Ces résultats suggèrent que la régularisation **Ridge** permet de **réduire la variance des estimations** sans introduire un **biais excessif**.
Discussion : Pourquoi les écarts-types et tests classiques ne sont-ils pas valides après Lasso ?
Après **Lasso**, les écarts-types et tests de significativité classiques ne sont pas valides car le Lasso **modifie volontairement les coefficients** et effectue une **sélection automatique des variables à partir des données**.

Cette sélection rend les **coefficients biaisés par construction** et **viole les hypothèses statistiques** sur lesquelles reposent les tests classiques du MCO.

Ainsi, le Lasso doit être vu comme un **outil de prédiction et de sélection de variables**, et non comme un modèle destiné à l'inférence statistique classique.

Dans l'ensemble, les résultats suggèrent que le modèle **MCO enrichi est adapté pour l'inférence économique**, tandis qu'une régularisation **Ridge permet un léger gain en performance prédictive hors échantillon**. Le Lasso, bien qu'utile pour la **sélection de variables**, ne présente pas d'avantage dans ce contexte.
8 Prévisions
8.1 Prédiction ponctuelle et intervalle de confiance
Dans cette section, nous utilisons le **modèle semi-logarithmique** sélectionné précédemment comme **meilleur modèle économétrique** (best_model).

La variable **dépendante** est le **logarithme du prix du logement** (en milliers d'euros), ce qui permet de réduire l'hétéroscédasticité et de faciliter l'interprétation des coefficients en termes de variations relatives.

Le modèle est estimé par MCO en incluant **l'ensemble des variables explicatives** nécessaires à la prédiction demandée.

Cette cellule re-définit le modèle semi-logarithmique par la méthode des **moindres carrés ordinaires (MCO)**.

Une **constante est ajoutée** au modèle afin de capturer l'effet moyen du prix lorsque l'ensemble des variables explicatives est nul.

La variable Distance_universite n'est pas incluse dans le modèle de prédiction final, car elle est introduite uniquement comme instrument dans l'estimation par variables instrumentales et ne satisfait pas l'hypothèse d'exogénéité requise pour être intégrée directement parmi les variables explicatives du modèle MCO.

```
# Variable dépendante
y_log = np.log(df["Prix_milliers_euros"])

# Variables explicatives (modèle enrichi validé)
X_semi_log = df[
    [
        "Surface_m2",
        "Chambres",
        "Annee_construction",
        "Distance_centre_km",
        "Etage",
        "Ascenseur",
        "Qualite_ecole",
        "Revenu_median_quartier",
    ]
]

X_semi_log = sm.add_constant(X_semi_log)

# Estimation du modèle
best_model = sm.OLS(y_log, X_semi_log).fit()

print(best_model.summary())
```

Nous définissons ici les **caractéristiques du logement** pour lequel une prédiction de prix est demandée.

Les valeurs sont renseignées conformément à l'énoncé, en respectant les unités utilisées lors de l'estimation (notamment le revenu médian du quartier exprimé en milliers d'euros).

Une **constante est ajoutée** afin d'assurer la cohérence dimensionnelle avec le modèle estimé.

```
new_house = pd.DataFrame({
    "Surface_m2": [120],
    "Chambres": [3],
    "Annee_construction": [2015],
    "Distance_centre_km": [5],
    "Etage": [1],
    "Ascenseur": [1],
    "Qualite_ecole": [7],
    "Revenu_median_quartier": [65]
})

new_house = sm.add_constant(new_house, has_constant="add")
```

À partir du **modèle semi-logarithmique estimé**, nous calculons la **prédiction du logarithme du prix du logement** ainsi que l'**intervalle de confiance à 95 %** de l'espérance conditionnelle.

Les résultats sont ensuite **retranscrits sur l'échelle du prix** en appliquant une transformation exponentielle.

L'intervalle obtenu correspond à l'intervalle de confiance du prix moyen prédit, conditionnellement aux caractéristiques du logement.

```
# Prédiction
prediction = best_model.get_prediction(new_house)
pred_summary = prediction.summary_frame(alpha=0.05)

# Valeurs en log
mean_log = pred_summary.loc[0, "mean"]
lower_log = pred_summary.loc[0, "mean_ci_lower"]
upper_log = pred_summary.loc[0, "mean_ci_upper"]

# Back-transformation (prix en milliers d'euros)
mean_price = np.exp(mean_log)
lower_price = np.exp(lower_log)
upper_price = np.exp(upper_log)
```

print(f"Prix prédit : {mean_price:.2f} milliers d'euros")
print(f"Intervalle de confiance à 95 % : [{lower_price:.2f} ; {upper_price:.2f}] milliers d'euros")

3. Cette prédiction est-elle fiable ? Discutez.

La prédiction obtenue à partir du modèle semi-logarithmique estimé par MCO indique un **prix prédit d'environ 2 259.62 milliers d'euros**, avec un intervalle de confiance à 95 % relativement étroit **[2219.31;2300.66]**.

La **faible largeur de cet intervalle** suggère une **bonne précision** statistique de l'estimation de l'espérance conditionnelle du prix, compte tenu des caractéristiques du logement.

Plusieurs éléments plaident en faveur de la fiabilité de cette prédiction.
- Tout d'abord, le modèle présente une **très bonne qualité d'ajustement globale ($R^2 \approx 0,90$)**, indiquant que la majorité de la variabilité du prix est expliquée par les variables incluses.
- Ensuite, la plupart des **coefficients sont statistiquement significatifs** et présentent des signes économiquement cohérents (effet positif de la surface, des chambres, de la qualité des écoles, et effet négatif de la distance au centre).
- Enfin, la prédiction est **réalisée dans le champ des données observées** (valeurs plausibles des variables), ce qui limite les risques liés à l'extrapolation.

Cependant, cette prédiction doit être interprétée avec **prudence**.
- D'une part, l'intervalle de confiance calculé correspond à **l'incertitude sur la moyenne conditionnelle**, et non à un intervalle de prédiction individuel : la variabilité réelle des prix autour de cette moyenne peut être plus importante.
- D'autre part, le résumé du modèle signale un **nombre de condition élevé**, suggérant une possible **multicolinéarité** entre certaines variables, ce qui peut affecter la stabilité de certains coefficients.
- Enfin, la validité de la prédiction repose sur les **hypothèses classiques du modèle MCO** (spécification correcte, absence de biais d'omission, exogénéité), qui ne peuvent jamais être garanties parfaitement en pratique.
Conclusion

En conclusion, la prédiction obtenue à partir du modèle semi-logarithmique **apparaît statistiquement crédible** et **économiquement cohérente**, avec un intervalle de confiance étroit reflétant une bonne précision de l'estimation.

Néanmoins, elle doit être **interprétée comme une valeur moyenne attendue**, conditionnelle aux caractéristiques observées du logement, et non comme une valeur certaine.