Analyse Économétrique des Déterminants du Marché Immobilier

Du modèle linéaire aux méthodes de régularisation et d'inférence causale

Rapport du Projet d'Économétrie Appliquée / Statistiques

Réalisé par : Driss Mikou et Omar Zeroual

En Décembre 2025

Université Paris 1 Panthéon-Sorbonne UFR02 Économie I Sorbonne Data Analytics

Ce projet analyse les déterminants des prix immobiliers à partir de 150 maisons vendues entre 2015 et 2023. L'objectif est d'identifier les facteurs expliquant le prix des logements et de comparer différentes approches économétriques selon qu'elles visent l'interprétation économique ou la prévision.

Principaux résultats.

Les estimations par régression linéaire montrent que la surface habitable est le principal déterminant du prix des logements. La distance au centre-ville exerce un effet négatif significatif, tandis que certaines caractéristiques du bien et du quartier contribuent positivement au prix.

Les diagnostics économétriques ne permettent pas de rejeter les hypothèses classiques du modèle linéaire, notamment l'homoscedasticité des résidus. En revanche, les tests de stabilité structurelle mettent en évidence une rupture significative en 2020, suggérant que la relation entre prix et caractéristiques des logements a été affectée par la période COVID.

Parmi les modèles linéaires classiques, la spécification log-log mixte apparaît comme la plus adaptée pour l'analyse économique. Dans une optique prédictive, les méthodes de régularisation dominent : le modèle Ridge, sélectionné par validation croisée, présente la meilleure performance de prévision sur l'échantillon test.

Recommandations.

Pour l'analyse économique des prix immobiliers, il est recommandé de tenir compte de la rupture structurelle identifiée en 2020 en estimant des modèles distincts selon les périodes. Les modèles linéaires restent adaptés pour l'interprétation des effets marginaux, en particulier sous des spécifications logarithmiques. Pour les objectifs de prévision, les méthodes de régularisation, et notamment le Ridge, doivent être privilégiées en raison de leur meilleure performance prédictive.

Introduction

Le marché immobilier occupe une place centrale dans l'analyse économique, tant pour les ménages que pour les décideurs publics. Les prix des logements reflètent à la fois des caractéristiques propres aux biens, des facteurs de localisation et des conditions économiques plus générales. Comprendre les déterminants des prix immobiliers constitue ainsi un enjeu important, aussi bien pour l'évaluation des biens que pour l'analyse des dynamiques urbaines et territoriales.

Dans ce contexte, l'économétrie fournit un cadre méthodologique particulièrement adapté pour analyser les relations entre le prix d'un logement et ses caractéristiques observables. Les modèles de régression linéaire permettent d'identifier et de quantifier l'effet marginal de variables telles que la surface habitable, la localisation ou la qualité du quartier. Toutefois, l'application de ces modèles soulève plusieurs questions méthodologiques, notamment en lien avec la forme fonctionnelle appropriée, la validité des hypothèses classiques, la stabilité des relations dans le temps ou encore la présence possible d'endogénéité.

L'objectif de ce projet est d'analyser les déterminants des prix de maisons à partir d'un ensemble de données portant sur 150 ventes réalisées entre 2015 et 2023. Plus précisément, il s'agit d'identifier les variables qui influencent significativement le prix de vente, d'évaluer la robustesse des résultats obtenus et de comparer différentes approches économétriques selon qu'elles visent l'interprétation économique ou la performance prédictive. Le projet mobilise ainsi une large gamme d'outils, allant du modèle linéaire simple aux méthodes de régularisation, en passant par les tests diagnostiques et les techniques d'estimation par variables instrumentales.

L'analyse est menée de manière progressive. Dans un premier temps, une exploration descriptive des données permet de caractériser les principales variables et d'identifier d'éventuelles relations simples entre elles. Dans un second temps, des modèles de régression linéaire, simples puis multiples, sont estimés afin d'analyser l'impact marginal des différentes caractéristiques des logements sur leur prix. Différentes spécifications fonctionnelles sont ensuite comparées afin d'évaluer leur pertinence empirique.

La validité des résultats est ensuite examinée à travers une série de diagnostics économétriques portant notamment sur l'hétéroscédasticité, la multicolinéarité et la stabilité structurelle des relations estimées, avec une attention particulière portée à la période de la crise sanitaire. Le projet aborde également la question de l'endogénéité, en discutant ses sources potentielles et en mettant en œuvre une estimation par variables instrumentales lorsque cela est pertinent.

Enfin, dans une perspective orientée vers la prévision, des méthodes de régularisation telles que Ridge et Lasso sont utilisées et comparées aux modèles linéaires classiques. Ces approches permettent d'évaluer les performances prédictives des différents modèles et de produire une prédiction du prix d'un logement type, accompagnée d'un intervalle de confiance.

Le rapport est structuré comme suit. La première partie est consacrée à l'analyse descriptive et à l'estimation des modèles linéaires de base. La deuxième partie traite des diagnostics économétriques et des corrections éventuelles. La troisième partie aborde la question de l'endogénéité et de l'estimation par variables instrumentales. La quatrième partie présente les méthodes de régularisation et compare leurs performances prédictives. Enfin, la conclusion synthétise les principaux résultats, discute les limites de l'analyse et propose des pistes de réflexion pour des travaux futurs.

Partie 1 – Analyse descriptive et modèle de base

1.1 Statistiques descriptives

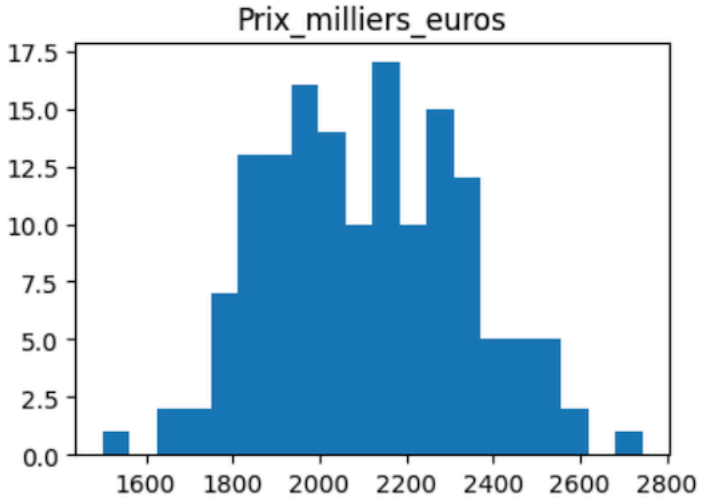
Cette section vise à décrire les principales caractéristiques du jeu de données et à fournir une première lecture des variables utilisées dans l'analyse. Le jeu de données comprend 150 maisons vendues entre 2015 et 2023. Les variables considérées décrivent à la fois les caractéristiques physiques des logements, leur localisation et certains attributs du quartier.

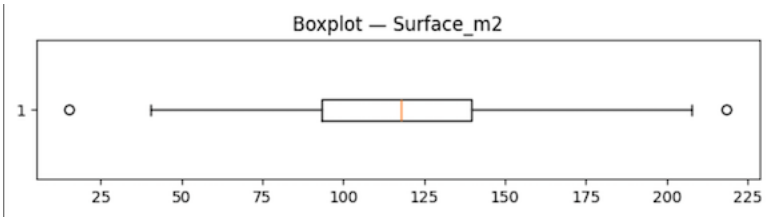
Les statistiques descriptives (moyenne, médiane, écart-type, minimum, maximum et quartiles) mettent en évidence une hétérogénéité raisonnable des biens observés, reflétant la diversité du marché immobilier étudié. La surface habitable varie sensiblement d'un logement à l'autre, ce qui en fait un candidat naturel pour expliquer les différences de prix. Le nombre de chambres et l'année de construction présentent également une dispersion modérée, traduisant des profils de logements relativement variés.

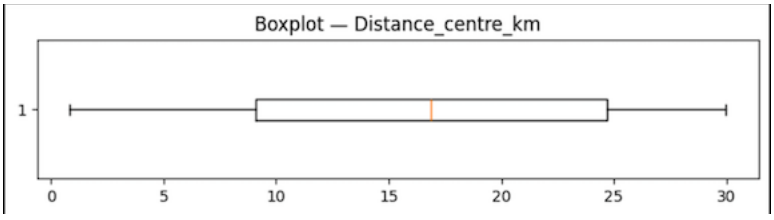
	mean	median	std	min	Q1	Q3	max
Surface_m2	116.706800	117.845	37.693819	15.21	93.240	139.6375	218.53
Chambres	2.886667	3.000	1.077760	1.00	2.000	4.0000	5.00
Annee_construction	2001.826667	2002.500	11.704841	1980.00	1991.000	2012.0000	2022.00
Distance_centre_km	16.500267	16.865	9.017430	0.83	9.105	24.6975	29.99
Etage	2.580000	2.500	1.761901	0.00	1.000	4.0000	5.00
Ascenseur	0.460000	0.000	0.500067	0.00	0.000	1.0000	1.00
Annee_vente	2019.840000	2020.000	2.288225	2015.00	2018.000	2022.0000	2023.00
Qualite_ecole	5.468667	5.600	1.868249	1.00	4.125	7.0000	10.00
Revenu_median_quartier	63.668000	63.450	9.295458	42.90	57.500	70.4750	83.90
Distance_universite	8.064000	8.300	3.746502	1.00	5.300	10.8750	17.10
Prix_milliers_euros	2107.904800	2105.050	229.921013	1500.77	1934.285	2272.7800	2743.04

Les variables de localisation jouent un rôle important dans la caractérisation des biens. La distance au centre-ville varie de manière significative, suggérant des situations géographiques contrastées. Les variables qualitatives, telles que la présence d'un ascenseur ou l'étage, permettent de distinguer différents niveaux de confort, tandis que les indicateurs de quartier, comme la qualité des écoles et le revenu médian, apportent une information complémentaire sur l'environnement socio-économique.

L'analyse de la distribution du prix de vente, à l'aide d'histogrammes et de boîtes à moustaches, montre une distribution globalement peu asymétrique. Les coefficients d'asymétrie et d'aplatissement confirment l'absence de déviation marquée par rapport à une distribution équilibrée. Sur cette base, aucune transformation logarithmique ne s'impose a priori pour corriger la forme de la distribution du prix. Néanmoins, ces transformations seront ultérieurement explorées dans une logique de comparaison de modèles.


Prix_milliers_euros
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1600
1800
2000
2200
2400
2600
2800


Boxplot — Surface_m2
1
25
50
75
100
125
150
175
200
225


Boxplot — Distance_centre_km
1
0
5
10
15
20
25
30

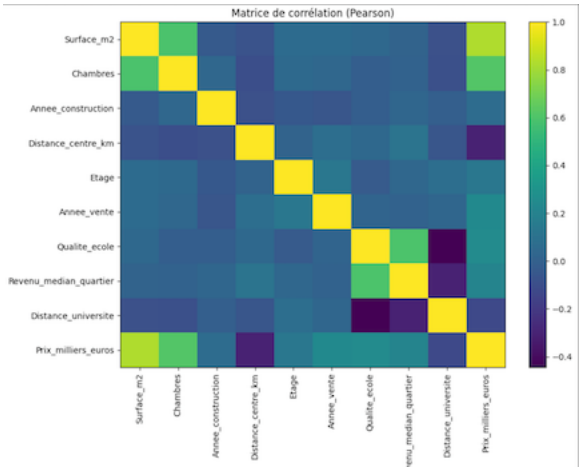
1.2 Analyse de corrélation

Afin de compléter l'analyse descriptive, une matrice de corrélation est calculée entre les principales variables continues du jeu de données. Les coefficients de corrélation permettent d'identifier les relations linéaires les plus marquées entre les variables, sans toutefois préjuger de relations causales.

	Surface_m2	Chambres	Annee_construction	Distance_centre_km	Etage	Annee_vente	Qualite_ecole	Revenu_median_quartier
Surface_m2	1.000000	0.590959	-0.034333	-0.074765	0.061126	0.060417	0.043832	0.013908
Chambres	0.590959	1.000000	0.039398	-0.102319	0.048986	0.036140	-0.012775	0.011761
Annee_construction	-0.034333	0.039398	1.000000	-0.088117	-0.044559	-0.055920	-0.009887	0.030304
Distance_centre_km	-0.074765	-0.102319	-0.088117	1.000000	0.018235	0.076058	0.042029	0.111579
Etage	0.061126	0.048986	-0.044559	0.018235	1.000000	0.126383	-0.031142	0.038186
Annee_vente	0.060417	0.036140	-0.055920	0.076058	0.126383	1.000000	0.020798	0.007267
Qualite_ecole	0.043832	-0.012775	-0.009887	0.042029	-0.031142	0.020798	1.000000	0.598651
Revenu_median_quartier	0.013908	0.011761	0.030304	0.111579	0.038186	0.007267	0.598651	1.000000
Distance_universite	-0.088362	-0.089609	-0.006250	-0.055751	0.077101	0.033378	-0.444083	-0.304002
Prix_milliers_euros	0.826731	0.614825	0.068358	-0.306648	0.128403	0.236992	0.253632	0.205779

Revenu_median_quartier	Distance_universite	Prix_milliers_euros
0.013908	-0.088362	0.826731
0.011761	-0.089609	0.614825
0.030304	-0.006250	0.068358
0.111579	-0.055751	-0.306648
0.038186	0.077101	0.128403
0.007267	0.033378	0.236992
0.598651	-0.444083	0.253632
1.000000	-0.304002	0.205779
-0.304002	1.000000	-0.124521
0.205779	-0.124521	1.000000

La matrice de corrélation met en évidence une corrélation positive forte entre le prix de vente et la surface habitable. Cette relation apparaît comme la plus marquée parmi l'ensemble des variables considérées, ce qui suggère que la surface est le facteur le plus fortement associé au prix au niveau descriptif. Cette observation est cohérente avec l'intuition économique et sera confirmée ultérieurement par les estimations de régression.



D’autres variables présentent des corrélations plus modérées avec le prix. La distance au centre-ville est négativement corrélée au prix, indiquant que les logements situés plus loin du centre tendent à être moins chers. Les indicateurs de quartier, tels que le revenu médian et la qualité des écoles, affichent une corrélation positive avec le prix, bien que d’ampleur inférieure à celle observée pour la surface. Les caractéristiques du logement comme le nombre de chambres ou l’année de construction présentent des corrélations positives mais relativement limitées.

La matrice de corrélation met également en évidence certaines corrélations entre variables explicatives, en particulier entre la surface et le nombre de chambres, ainsi qu’entre les variables de quartier. Ces relations suggèrent un risque potentiel de multicolinéarité, qui sera examiné plus en détail dans la section consacrée aux diagnostics du modèle.

En réponse à la question posée, la variable qui semble avoir l’impact le plus fort sur le prix selon l’analyse de corrélation est **la surface habitable**. Il convient toutefois de rappeler que la corrélation ne permet pas d’établir un lien de causalité. Les relations observées peuvent refléter des effets indirects ou être influencées par des variables omises. Une analyse économétrique multivariée est donc nécessaire pour isoler l’impact propre de chaque variable, ce qui constitue l’objet des sections suivantes.

1.3 Modèle de régression linéaire simple

Dans un premier temps, un modèle de régression linéaire simple est estimé afin d’analyser la relation entre le prix de vente d’un logement et sa surface habitable. Ce modèle constitue une base de référence permettant d’évaluer l’importance de la surface dans la détermination du prix.

Le modèle estimé s’écrit comme suit :

$$Prix_i = \beta 0 + \beta 1 \times Surface_i + u_i$$

L’estimation par moindres carrés ordinaires met en évidence une relation positive et statistiquement significative entre la surface et le prix de vente. Le coefficient associé à la surface est estimé à **5,04**, avec une p-value inférieure à 1 %, ce qui permet de rejeter l’hypothèse nulle selon laquelle la surface n’aurait aucun effet sur le prix.

Sur le plan économique, ce résultat signifie qu’une augmentation de 1 m² de surface habitable est associée, en moyenne, à une hausse du prix d’environ **5 000 euros**. Le coefficient de détermination du modèle est égal à **R2 = 0,68**, indiquant que la surface seule explique environ 68 % de la variance du prix de vente. Ce niveau d’ajustement, relativement élevé pour un modèle univarié, confirme le rôle central de la surface, tout en suggérant que d’autres facteurs influencent également le prix des logements.

Ce modèle simple constitue une première étape utile, mais reste incomplet puisqu'il ne tient pas compte des autres caractéristiques du logement et de son environnement. La section suivante introduit donc un modèle de régression linéaire multiple afin d'analyser l'impact propre de plusieurs variables simultanément et d'en tester la significativité.

1.4 Modèle de régression linéaire multiple

Afin d'analyser plus finement les déterminants du prix immobilier, un modèle de régression linéaire multiple est estimé en intégrant plusieurs caractéristiques du logement et de sa localisation.

Le modèle estimé s'écrit comme suit :

$$Prix_i = \beta0 + \beta1 \times Surface_i + \beta2 \times Chambres_i + \beta3 \times Annee\_construction_i + \beta4 \times Distance\_centre_i + \beta5 \times Etage_i + \beta6 \times Ascenseur_i + u_i$$

L'estimation par moindres carrés ordinaires montre une amélioration significative de la qualité d'ajustement par rapport au modèle simple. Le coefficient de détermination atteint **R2 = 0,79**, contre **0,68** précédemment, ce qui indique que l'ajout de variables explicatives permet de mieux rendre compte des variations du prix de vente.

La variable **Surface** demeure fortement significative, avec un coefficient estimé à **4,39** (p-value < 1 %), confirmant son rôle central dans la détermination du prix, même après contrôle des autres caractéristiques. La **distance au centre-ville** exerce un effet négatif et significatif : une augmentation d'un kilomètre est associée, en moyenne, à une baisse du prix d'environ **6 100 euros**, toutes choses égales par ailleurs.

Les variables **Chambres**, **Année de construction**, **Étage** et **Ascenseur** présentent également des coefficients positifs et statistiquement significatifs. En particulier, la présence d'un ascenseur est associée à une hausse moyenne du prix d'environ **55 500 euros**, ce qui suggère une valorisation importante de ce type d'équipement sur le marché immobilier. Le nombre de chambres et l'étage contribuent également positivement au prix, tandis que les logements plus récents tendent à être légèrement plus chers.

1.5 Transformation logarithmique et comparaison des spécifications

Bien que l'analyse descriptive n'ait pas mis en évidence d'asymétrie marquée dans la distribution du prix, différentes spécifications fonctionnelles sont néanmoins explorées afin de comparer leurs performances empiriques et d'évaluer leur pertinence en termes de prévision et d'interprétation économique.

Trois modèles sont estimés et comparés :

- un modèle en **niveau** (prix en niveau),
- un modèle **semi-logarithmique** (log(prix)),
- un modèle **log-log mixte**, dans lequel le prix et certaines variables continues sont exprimés en logarithme, tandis que les variables discrètes (étage, ascenseur) sont conservées en niveau.

Les modèles sont estimés sur un échantillon d'apprentissage et comparés sur un échantillon test représentant 20 % des observations. La performance prédictive est évaluée à l'aide de la racine de l'erreur quadratique moyenne (RMSE) calculée sur le prix en niveau.

Les résultats montrent que le modèle en niveau obtient un RMSE d'environ **105,05 k€**, tandis que le modèle semi-logarithmique présente une performance légèrement inférieure, avec un RMSE d'environ **107,39 k€**. Le modèle log-log mixte se distingue par la meilleure performance prédictive parmi les modèles linéaires classiques, avec un RMSE d'environ **103,55 k€**.

Ces résultats indiquent que la transformation logarithmique améliore la capacité prédictive du modèle, même en l'absence de problème majeur de distribution du prix. L'amélioration reste modérée mais systématique, ce qui suggère que la spécification log-log permet de mieux capturer certaines relations non linéaires entre le prix et ses déterminants.

Au-delà de la performance prédictive, le modèle log-log mixte présente également un avantage en termes d'interprétation économique. Les coefficients associés aux variables exprimées en logarithme peuvent être interprétés comme des élasticités, ce qui facilite l'analyse des effets proportionnels des caractéristiques du logement sur le prix.

Sur la base de ces éléments, le modèle log-log mixte est retenu comme **spécification de référence parmi les modèles MCO**, tant pour ses performances prédictives que pour la clarté de son interprétation économique. Les tests de significativité et les diagnostics économétriques sont ensuite appliqués à cette spécification.

1.6 Tests de significativité

Cette section évalue la significativité statistique des relations estimées dans le modèle de régression linéaire multiple, à la fois au niveau global et individuel.

1.6.1 Test de significativité globale du modèle (test F)

La significativité globale du modèle est évaluée à l'aide d'un test F, qui permet de tester l'hypothèse nulle selon laquelle l'ensemble des coefficients des variables explicatives (à l'exception de la constante) est égal à zéro.

Le test conduit à une statistique **F = 88,94**, associée à une p-value égale à $\mathbf{1{,}11 \times 10^{-16}}$. Cette p-value étant largement inférieure aux seuils usuels de significativité, l'hypothèse nulle est rejetée.

Ce résultat indique que le modèle, pris dans son ensemble, possède un pouvoir explicatif statistiquement significatif : les variables explicatives contribuent conjointement à expliquer le prix de vente des logements.

1.6.2 Test de significativité individuelle : effet de la distance au centre

On teste ensuite l'effet individuel de la distance au centre-ville sur le prix immobilier. L'hypothèse testée est formulée de manière unilatérale, conformément à l'intuition économique selon laquelle un éloignement du centre devrait réduire le prix :

H0 : $\beta\_distance \geq 0$
H1 : $\beta\_distance < 0$

La statistique de test associée à la variable *Distance_centre_km* est égale à **t = –6,19**. La p-value bilatérale est de $\mathbf{5{,}90 \times 10^{-9}}$, et la p-value unilatérale (à gauche) est de $\mathbf{2{,}95 \times 10^{-9}}$.

Ces valeurs conduisent à rejeter très nettement l'hypothèse nulle. On conclut ainsi que la distance au centre-ville a un **effet négatif et statistiquement significatif** sur le prix des logements : toutes choses égales par ailleurs, les biens situés plus loin du centre tendent à être moins chers.

1.6.3 Multicolinéarité : facteurs d'inflation de la variance (VIF)

Enfin, la présence éventuelle de multicolinéarité entre les variables explicatives est examinée à l'aide des facteurs d'inflation de la variance (VIF). Les valeurs obtenues sont toutes proches de 1, et inférieures à 2, avec un maximum d'environ **1,60** pour les variables *Revenu_median_quartier* et *Qualite_ecole*.

Ces résultats indiquent l’absence de multicolinéarité préoccupante dans le modèle. Les coefficients estimés peuvent donc être interprétés de manière fiable, sans risque significatif de variance excessive liée à des corrélations fortes entre variables explicatives.

Partie 2 – Diagnostics économétriques et corrections

L'estimation par moindres carrés ordinaires repose sur plusieurs hypothèses classiques concernant la structure des erreurs et la stabilité du modèle. Avant d'interpréter les coefficients estimés, il est donc nécessaire de procéder à une série de diagnostics afin d'évaluer la validité de ces hypothèses et, le cas échéant, d'apporter des corrections appropriées.

2.1 Hétéroscédasticité et robustesse des estimations

La validité des tests statistiques usuels repose sur l'hypothèse d'homoscedasticité des résidus, c'est-à-dire une variance constante du terme d'erreur conditionnellement aux variables explicatives.

Dans un premier temps, une analyse graphique des résidus est réalisée. Les nuages de points ne révèlent pas de structure particulière ni de motif systématique suggérant une variance non constante des erreurs. La dispersion des résidus apparaît globalement homogène sur l'ensemble des valeurs ajustées, bien qu'une légère augmentation de la variance puisse être observée pour les logements de grande surface.

Afin de compléter cette analyse visuelle, un test formel de Breusch–Pagan est mis en œuvre. Le test conduit à une statistique LM égale à **12,00** (degrés de liberté = 8), associée à une p-value de **0,151**. L'hypothèse nulle d'homoscedasticité ne peut donc pas être rejetée au seuil de 5 %. Aucune preuve statistique forte d'hétéroscédasticité n'est ainsi mise en évidence dans les données.

Par précaution, et afin de garantir la robustesse de l'inférence, les estimations sont complétées par le calcul d'écarts-types robustes à l'hétéroscédasticité (HC1). La comparaison entre les écarts-types classiques et robustes montre des valeurs proches, et les conclusions de significativité des variables principales (surface, distance au centre, nombre de chambres, etc.) demeurent inchangées. Les résultats apparaissent donc stables et robustes.

Enfin, une estimation par moindres carrés pondérés (WLS), fondée sur une pondération inversement proportionnelle au carré de la surface, est testée à titre exploratoire. Cette approche conduit toutefois à une dégradation de la performance prédictive (RMSE en hausse d'environ 4,5 % par rapport aux MCO). En l'absence d'hétéroscédasticité avérée et compte tenu de ces résultats, les estimations MCO standards sont conservées comme référence pour la suite de l'analyse.

Diagnostic / méthode	Résultat	Conclusion
Breusch–Pagan (LM, ddl = 8)	LM = **12,00**, p-value = **0,151**	**H0 (homoscedasticité) non rejetée**
Écarts-types robustes HC1	—	Conclusions **inchangées** (robustesse confirmée)
WLS (pondération liée à la surface)	RMSE **+4,5 %** vs MCO	WLS **non retenu** (pas d'hétéro avérée + prédiction dégradée)
Durbin–Watson	DW = **1,50**	**Autocorrélation positive légère** (modérée)

2.2 Autocorrélation des résidus

Bien que les données ne constituent pas une série temporelle au sens strict, les observations sont ordonnées dans le temps via l'année de vente. Il est donc pertinent de vérifier l'absence d'autocorrélation des résidus, qui pourrait biaiser les écarts-types estimés.

Un test de Durbin–Watson est réalisé après avoir ordonné les résidus par année de vente. La statistique obtenue est **DW = 1,50**, indiquant une légère autocorrélation positive des résidus.

Cette valeur reste toutefois dans une zone acceptable et ne suggère pas une dépendance sérielle forte. Ce résultat est cohérent avec la nature du marché immobilier, où certains facteurs macroéconomiques non observés (conditions de crédit, climat économique) peuvent introduire une inertie modérée des prix. L'utilisation d'écarts-types robustes, combinée aux tests de stabilité structurelle présentés ci-dessous, permet de garantir la fiabilité des conclusions statistiques.

2.3 Stabilité structurelle : test de rupture en 2020

La période étudiée (2015–2023) inclut la crise sanitaire du COVID-19, susceptible d'avoir modifié le fonctionnement du marché immobilier. Afin d'évaluer la stabilité des relations estimées, un test de Chow est réalisé en considérant **2020** comme date potentielle de rupture structurelle.

Le test compare un modèle estimé sur l'ensemble de l'échantillon à deux modèles estimés séparément sur les périodes **pré-2020** et **post-2020**, en utilisant les mêmes variables explicatives : Surface_m2, Chambres, Annee_construction, Distance_centre_km, Etage, Ascenseur, Qualite_ecole et Revenu_median_quartier. La variable Annee_vente n'est pas incluse comme régresseur, car elle sert uniquement à définir les sous-périodes.

Les résultats du test sont nets. Pour **n_pre = 60** et **n_post = 90**, la statistique de test est **F = 7,41**, associée à une p-value de **$1{,}04 \times 10^{-8}$**. L'hypothèse nulle de stabilité structurelle est donc rejetée à tous les seuils usuels.

Ce résultat indique que les coefficients du modèle (intercept et pentes) diffèrent significativement entre la période pré-COVID et la période post-COVID. Estimer un modèle unique sur l'ensemble de la période peut ainsi masquer des dynamiques distinctes dans la formation des prix.

Rupture testée	n_pre	n_post	Statistique F	p-value	Décision
2020	**60**	**90**	**7,41**	**$1{,}04 \times 10^{-8}$**	**Rejet de H0 : rupture structurelle**

2.4 Estimation de modèles séparés avant et après 2020

À la suite du rejet de l'hypothèse de stabilité structurelle, deux régressions linéaires multiples sont estimées séparément sur les sous-échantillons pré-2020 et post-2020, en conservant la même spécification.

Les résultats confirment des différences significatives dans l'impact de plusieurs variables. La **surface du logement** demeure le principal déterminant du prix dans les deux périodes, avec un effet positif et fortement significatif. Son impact marginal reste toutefois légèrement plus faible après 2020.

La **distance au centre-ville** exerce un effet négatif significatif dans les deux sous-périodes, mais son influence est plus marquée dans la période post-COVID, indiquant une pénalisation accrue de l'éloignement du centre.

Certaines variables gagnent en importance après 2020. L'**année de construction**, non significative avant la crise, devient significative après 2020, suggérant une valorisation accrue de la qualité du bâti. De même, la **qualité des écoles** et le **revenu médian du quartier** présentent des effets plus prononcés dans la période post-COVID, traduisant une attention renforcée portée à l'environnement socio-économique.

À l'inverse, des caractéristiques comme l'**étage** perdent en significativité après 2020, tandis que la présence d'un **ascenseur** conserve un effet positif relativement stable.

Ces résultats confirment que la crise sanitaire s'est accompagnée d'une modification de la structure des déterminants du prix immobilier, au-delà d'un simple choc sur le niveau général des prix.

2.5 Synthèse des diagnostics

L'ensemble des diagnostics économétriques indique que le modèle linéaire multiple satisfait globalement les hypothèses classiques relatives à l'homoscedasticité, à l'indépendance des erreurs et à la multicolinéarité. Les estimations par MCO apparaissent robustes et statistiquement fiables.

En revanche, une rupture structurelle significative est mise en évidence en 2020, justifiant l'estimation de modèles distincts avant et après la crise sanitaire. Cette rupture souligne l'importance de tenir compte des changements de régime dans l'analyse des prix immobiliers et motive l'examen de performances prédictives différenciées dans la section suivante.

Partie 3 – Endogénéité

3.1 Pourquoi l'endogénéité est à considérer dans ce cas?

Dans un modèle MCO, l'interprétation des coefficients suppose que les variables explicatives sont exogènes, c'est-à-dire non corrélées au terme d'erreur. Or, dans un contexte immobilier, certaines variables peuvent capter des facteurs non observés (attractivité générale du quartier, sécurité, accessibilité, prestige, projets urbains) qui influencent directement le prix.

La variable **Qualite_ecole** est un candidat naturel à l'endogénéité. Elle peut être corrélée à des caractéristiques non observées du quartier (tri résidentiel, composition sociale, investissements locaux) qui affectent aussi le prix. Dans ce cas, le coefficient MCO de Qualite_ecole peut être biaisé et ne pas refléter un effet causal.

3.2 Stratégie : variables instrumentales (2SLS)

Pour traiter cette endogénéité potentielle, une estimation par variables instrumentales (2SLS) est mise en œuvre en instrumentant **Qualite_ecole** par **Distance_universite**.

L'idée est d'utiliser une variable (l'instrument) qui fait varier Qualite_ecole, mais qui n'affecte pas directement le prix autrement que par Qualite_ecole. Deux conditions sont nécessaires :

1. **Pertinence** : l'instrument doit être suffisamment corrélé à Qualite_ecole.
2. **Exclusion (validité)** : l'instrument ne doit pas influencer directement le prix, une fois contrôlées les autres variables.

Dans la pratique, la pertinence est testable, tandis que l'exclusion repose sur une hypothèse économique, surtout lorsqu'on ne dispose que d'un seul instrument.

3.3 Première étape : pertinence de l'instrument

La première étape du 2SLS consiste à régresser **Qualite_ecole** sur **Distance_universite** et les autres contrôles exogènes. Les résultats montrent une relation statistiquement significative entre l'instrument et la variable instrumentée :

- **t = –4,33** pour Distance_universite dans la régression de première étape,
- statistique de "faiblesse de l'instrument" : **F = 18,71**, supérieure au seuil usuel de 10.

Ces résultats indiquent que l'instrument est **pertinent** et ne souffre pas d'un problème d'instrument faible au sens des règles empiriques courantes.

Variable instrumentale	Coefficient	t-stat	F (faiblesse)	Conclusion
Distance_universite → Qualite_ecole	–0.144	–4.33	**18.71**	Instrument pertinent

3.4 Deuxième étape : estimation IV et comparaison à MCO

La seconde étape estime l'équation du prix en remplaçant Qualite_ecole par sa composante prédite à partir de l'instrument.

La comparaison des coefficients met en évidence une différence très marquée :

- coefficient de **Qualite_ecole en MCO : +20,84**
- coefficient de **Qualite_ecole en IV : –0,94**

Le passage d'un effet fortement positif en MCO à un effet proche de zéro (et de signe opposé) en IV suggère que l'estimation MCO capte probablement un biais de quartier : Qualite_ecole peut refléter des caractéristiques socio-économiques et de désirabilité du quartier, elles-mêmes déterminantes du prix, plutôt qu'un effet causal pur de la qualité scolaire.

L'estimation IV indique au contraire que, une fois isolée la variation exogène de Qualite_ecole induite par l'instrument, l'effet estimé sur le prix devient **beaucoup plus faible**, voire nul dans cet échantillon.

Méthode	Coefficient Qualite_ecole	Interprétation
MCO	+20.84	Effet positif fort
IV (2SLS)	–0.94	Effet faible / nul

3.5 Limites et interprétation

Il est important de souligner que la validité de l'approche IV dépend de l'**hypothèse d'exclusion**, qui n'est **pas testable** ici avec un seul instrument. En effet, la proximité d'une université peut influencer directement le prix via d'autres canaux (demande étudiante, investissement locatif, dynamisme du quartier), indépendamment de la qualité des écoles.

Ainsi, les résultats IV doivent être interprétés comme une **analyse de robustesse** : ils montrent que l'effet positif de Qualite_ecole observé en MCO pourrait être surestimé en raison de facteurs non observés. Ils ne permettent pas, à eux seuls, d'affirmer définitivement un effet causal nul, mais ils renforcent l'idée que la relation MCO doit être lue avec prudence.

Synthèse

L'analyse d'endogénéité met en évidence que l'effet estimé de la qualité des écoles sur le prix est sensible à la méthode d'estimation. L'instrument retenu est statistiquement pertinent (F = 18,71), et l'estimation IV suggère que l'effet MCO positif pourrait être largement porté par des caractéristiques non observées du quartier. La validité économique de l'instrument restant discutible, les résultats IV sont présentés comme un complément informatif plutôt que comme une preuve définitive.

Partie 4

Méthodes de régularisation : Ridge et Lasso

Cette section vise à évaluer si des méthodes de régularisation permettent d'améliorer la performance prédictive du modèle linéaire multiple, en particulier en présence de variables potentiellement corrélées et d'un échantillon de taille modérée. Contrairement aux sections précédentes, l'objectif n'est plus l'interprétation causale des coefficients, mais la **qualité de la prédiction hors échantillon**.

Avant toute estimation, l'ensemble des variables explicatives est **standardisé** (moyenne nulle, variance unitaire). Cette étape est indispensable car les pénalisations Ridge et Lasso dépendent de l'échelle des variables : sans standardisation, une variable mesurée en kilomètres ou en années serait pénalisée différemment d'une variable binaire.

4.1 Choix du paramètre de pénalisation λ par validation croisée

Les paramètres de pénalisation λ (alpha) pour Ridge et Lasso sont sélectionnés par **validation croisée à 10 plis** sur l'échantillon d'apprentissage, à partir d'une grille de valeurs allant de 0.001 à 1000.

Les résultats montrent des comportements très différents entre les deux méthodes :

- Pour **Ridge**, la validation croisée sélectionne un paramètre optimal **α ≈ 6.25**, correspondant à une pénalisation modérée.
- Pour **Lasso**, le paramètre optimal est **α = 0.001**, soit la valeur la plus faible de la grille testée.

Ce résultat indique que, dans ce jeu de données, **une pénalisation forte dégrade la performance prédictive du Lasso**. Autrement dit, la sélection automatique de variables n'apporte pas de gain en termes de prédiction.

4.2 Évolution des coefficients avec la pénalisation

L'analyse de l'évolution des coefficients lorsque λ augmente permet de mieux comprendre ces résultats.

Pour **Lasso**, lorsque la pénalisation devient plus forte, certains coefficients diminuent rapidement et peuvent être annulés. Dans nos estimations, une pénalisation élevée (par exemple α = 10) conduit notamment à l'annulation du coefficient associé à l'année de construction. Cependant, cette simplification du modèle s'accompagne d'une **perte de performance prédictive**, ce qui explique pourquoi la validation croisée privilégie une pénalisation quasi nulle.

Pour **Ridge**, les coefficients diminuent progressivement en valeur absolue lorsque λ augmente, sans jamais être strictement annulés. La pénalisation agit principalement comme un mécanisme de **stabilisation** : elle réduit la variance des estimateurs en redistribuant l'effet entre variables corrélées (par exemple surface et nombre de chambres), tout en conservant l'ensemble de l'information explicative.

4.3 Interprétation des coefficients Lasso à l'alpha optimal

À l'alpha optimal sélectionné par validation croisée (α = 0.001), **aucune variable n'est supprimée** par le Lasso : tous les coefficients restent non nuls. Le modèle Lasso se comporte donc ici très proche d'une régression linéaire classique, avec un léger effet de shrinkage.

Ce résultat s'explique par deux caractéristiques des données :

- la taille relativement réduite de l'échantillon (environ 150 observations),
- l'absence de multicolinéarité forte entre les variables explicatives (VIF proches de 1 à 1.6).

Dans ce contexte, le Lasso n'a pas de rôle de "nettoyage" à jouer et la sélection de variables n'améliore pas la capacité de généralisation du modèle.

4.4 Comparaison des performances prédictives sur l'échantillon test

Les performances des différents modèles sont comparées sur un même échantillon de test, à l'aide de l'erreur quadratique moyenne (RMSE).

Les résultats sont les suivants :

- **OLS** : RMSE ≈ 93.10
- **Ridge (α optimal)** : RMSE ≈ **90.40**
- **Lasso (α optimal)** : RMSE ≈ 93.10

Le modèle **Ridge** obtient la meilleure performance prédictive, avec une réduction notable de l'erreur par rapport à la régression linéaire standard. Le **Lasso**, en revanche, n'apporte aucun gain par rapport à OLS.

4.5 Discussion et enseignements

Ces résultats mettent en évidence un point clé : dans ce jeu de données, **la régularisation par Ridge améliore la prédiction en stabilisant les coefficients**, tandis que la sélection de variables par Lasso n'est ni nécessaire ni bénéfique.

En pratique :

- Ridge constitue un bon compromis entre biais et variance, particulièrement adapté à la prédiction des prix immobiliers.
- Lasso est plus pertinent dans des contextes de très grande dimension ou de forte multicolinéarité, ce qui n'est pas le cas ici.

Enfin, il convient de souligner que les **tests de significativité classiques ne sont pas valides après Lasso**, car la sélection des variables est elle-même guidée par les données. Pour cette raison, l'évaluation des modèles pénalisés repose exclusivement sur des critères prédictifs (RMSE, performance hors échantillon), et non sur des tests statistiques traditionnels.

Conclusion et recommandations

Synthèse des principaux résultats

Ce projet a analysé les déterminants du prix des logements résidentiels à partir d'un ensemble de caractéristiques structurelles, de localisation et de qualité du quartier, en mobilisant des outils économétriques avancés.

L'analyse descriptive et les modèles de base montrent que **la surface du logement constitue le déterminant le plus important du prix**, suivie par la distance au centre-ville, le nombre de chambres et certains attributs de confort (étage, ascenseur). Ces effets sont statistiquement significatifs et économiquement cohérents.

Les diagnostics économétriques indiquent que le modèle linéaire multiple satisfait globalement les hypothèses classiques. Aucune hétéroscédasticité significative n'est détectée, et les résultats restent robustes à l'utilisation d'écarts-types robustes. Une **rupture structurelle majeure est toutefois mise en évidence en 2020**, confirmant que la relation entre prix et caractéristiques des logements a été profondément modifiée par la crise sanitaire. Les estimations pré- et post-COVID révèlent des changements notables dans la valorisation relative de certaines caractéristiques, en particulier celles liées à la qualité du quartier et à l'environnement.

La question de l'endogénéité est abordée via une estimation en variables instrumentales. En instrumentant la qualité des écoles par la distance à l'université la plus proche, les résultats suggèrent que l'effet estimé par les moindres carrés ordinaires est probablement biaisé à la hausse. L'estimation IV indique un effet causal beaucoup plus faible, voire nul, soulignant l'importance des mécanismes de sélection résidentielle et des variables omises dans l'analyse immobilière.

Enfin, l'objectif prédictif est exploré à l'aide de méthodes de régularisation. La comparaison hors échantillon montre que **Ridge améliore la performance prédictive par rapport au modèle OLS**, tandis que Lasso n'apporte aucun gain et ne sélectionne aucune variable pertinente dans ce contexte. Ces résultats confirment que, pour des données de dimension modérée et faiblement colinéaires, la stabilisation des coefficients est préférable à une sélection agressive des variables.

Afin d'illustrer concrètement l'usage prédictif des modèles estimés, nous appliquons les modèles retenus à un logement type correspondant à l'exemple fourni dans l'énoncé.

Les modèles **OLS** et **Ridge** produisent une prédiction ponctuelle très proche, de l'ordre de **2,33 M€**, ce qui témoigne d'une bonne robustesse des estimations. Le modèle Ridge, sélectionné comme meilleur modèle prédictif sur la base du RMSE hors échantillon, fournit ainsi une estimation stable et cohérente.

Le modèle linéaire log-log permet en outre de construire des **intervalles de confiance analytiques**. L'intervalle de confiance à 95 % pour la moyenne conditionnelle du prix est compris entre **2,26 M€ et 2,38 M€**, avec une largeur d'environ **120 k€**, ce qui reflète une incertitude raisonnable compte tenu de la variabilité des données. L'intervalle de prédiction pour une transaction individuelle est, sans surprise, plus large, soulignant que les modèles économétriques capturent une valeur moyenne attendue et non un prix exact.

Ceci soulève un point central : les modèles estimés constituent des **outils d'aide à la décision**, permettant de fournir des ordres de grandeur crédibles et des fourchettes de prix, mais ils ne sauraient se substituer à une expertise de terrain intégrant des éléments qualitatifs non observés (état précis du bien, négociation, contexte local immédiat).

Limites de l'analyse

Plusieurs limites doivent être soulignées.
Premièrement, la taille relativement réduite de l'échantillon limite la précision des estimations, en particulier pour les modèles plus complexes (IV, sous-périodes).
Deuxièmement, la validité de l'instrument utilisé repose sur une hypothèse d'exclusion forte, qui ne peut être testée formellement avec un seul instrument.
Enfin, l'analyse repose sur un découpage temporel simplifié (avant/après 2020), qui ne capture pas nécessairement l'ensemble des dynamiques progressives du marché immobilier.

Recommandations pour la pratique

D'un point de vue opérationnel, plusieurs enseignements peuvent être tirés :

- Pour **l'analyse explicative et l'interprétation économique**, un modèle linéaire multiple bien spécifié, accompagné de diagnostics rigoureux et d'une prise en compte explicite des ruptures structurelles, constitue une base solide.
- Pour **la prédiction des prix**, les méthodes pénalisées de type **Ridge** sont à privilégier, car elles offrent un meilleur compromis biais–variance sans sacrifier l'information contenue dans les variables explicatives.
- Toute analyse immobilière couvrant la période récente doit impérativement tenir compte de **la discontinuité introduite par la crise du COVID-19**, sous peine de masquer des changements profonds dans les préférences des acheteurs.
- Enfin, les résultats liés aux variables de quartier doivent être interprétés avec prudence, en raison des problèmes d'endogénéité et de sélection résidentielle.

En conclusion, ce travail montre que la combinaison d'outils économétriques classiques, de diagnostics approfondis et de méthodes modernes de régularisation permet d'obtenir une compréhension à la fois rigoureuse et opérationnelle des déterminants du prix immobilier.