STATISTIQUES

Grillon Pauline
Thépault Antoine
Sivayanama Perveena

TABLE DES MATIERES

RESUME

Ce travail étudie les déterminants du prix de logements résidentiels à partir d'un jeu de données comprenant 150 observations. La variable d'intérêt est le prix de vente, exprimé en milliers d'euros, et les variables explicatives décrivent les caractéristiques structurelles des logements, leur localisation, ainsi que certaines dimensions socio-économiques et temporelles.

Une analyse descriptive et préliminaire est menée afin de caractériser la distribution des variables et d'identifier les principales sources d'hétérogénéité. Les résultats mettent en évidence une asymétrie positive modérée des prix et une dispersion importante. Une analyse de corrélation montre que le prix est fortement associé à la surface habitable et négativement corrélé à la distance au centre-ville.

Des modèles de régression linéaire simple et multiple sont ensuite estimés par la méthode des moindres carrés ordinaires. Des transformations logarithmiques, des tests de diagnostic, une analyse de l'endogénéité par variables instrumentales et des méthodes de régularisation sont mobilisés afin d'évaluer la robustesse des résultats et les performances prédictives des modèles.

INTRODUCTION

L’analyse des prix immobiliers constitue un thème central en économie appliquée, dans la mesure où les prix des logements résultent d’une interaction complexe entre caractéristiques propres aux biens, facteurs de localisation et conditions socio-économiques. Comprendre les déterminants du prix immobilier permet d’éclairer les mécanismes de formation des prix et d’améliorer l’évaluation des biens sur le marché.

L’objectif de ce travail est d’identifier les principaux facteurs expliquant le prix de vente de logements résidentiels à partir d’un jeu de données observationnelles. La variable d’intérêt est le prix de vente, exprimé en milliers d’euros, et les variables explicatives décrivent les caractéristiques structurelles des logements, leur localisation, ainsi que certaines dimensions socio-économiques et temporelles.

L’étude débute par une analyse descriptive et une analyse de corrélation, permettant de caractériser la distribution des variables et d’identifier les relations linéaires entre le prix et ses principaux déterminants. Des modèles de régression linéaire simple et multiple sont ensuite estimés par la méthode des moindres carrés ordinaires afin d’interpréter les effets marginaux toutes choses égales par ailleurs. Des transformations logarithmiques et des diagnostics économétriques sont mobilisés pour améliorer les propriétés statistiques des modèles et tester leur robustesse.

Le rapport est organisé comme suit. La première partie présente l’analyse descriptive et préliminaire des données. La deuxième partie est consacrée à l’estimation et à l’interprétation des modèles linéaires. Les parties suivantes examinent les limites du cadre MCO à travers l’étude de l’endogénéité, l’utilisation de méthodes de régularisation et l’analyse des performances prédictives. Une conclusion générale synthétise les principaux résultats et en discute les limites.

PARTIE 1 : Analyse descriptive et modèle de base

1. Analyse descriptive des données

1.1. Présentation générale du jeu de données

Le jeu de données utilisé dans cette étude contient 150 observations, chacune correspondant à un bien immobilier résidentiel.
La variable d'intérêt principale est le prix de vente, exprimé en milliers d'euros (*Prix_milliers_euros*).

Les variables explicatives couvrent plusieurs dimensions du bien :

- Caractéristiques structurelles : surface habitable, nombre de chambres, étage, présence d'un ascenseur ;
- Caractéristiques de localisation : distance au centre-ville, distance à l'université ;
- Variables socio-économiques : revenu médian du quartier, qualité des écoles ;
- Variables temporelles : année de construction et année de vente.

Cette diversité de variables permet d'analyser les déterminants du prix immobilier selon une approche multidimensionnelle.

	ID	Surface_m2	Chambres	Annee_construction	Distance_centre_km	Etage	Ascenseur	Annee_vente	Qualite_ecole	Revenu_median_quartier
0	1	139.87	3	1982	21.33	2	1	2023	1.1	47.5
1	2	114.47	4	1991	2.90	5	1	2022	4.7	44.2
2	3	145.91	2	2005	3.00	3	1	2017	3.2	53.8
3	4	180.92	4	1995	29.61	3	1	2022	6.1	68.3
4	5	110.63	3	2016	11.54	0	0	2021	9.0	81.9

	ID	Surface_m2	Chambres	Annee_construction	Distance_centre_km	Etage	Ascenseur	Annee_vente	Qualite_ecole	Revenu_median_quart
147	148	67.18	1	2021	7.58	4	0	2019	5.4	7
148	149	140.88	3	1988	2.74	3	0	2020	5.0	5
149	150	131.88	2	2006	4.30	4	0	2020	4.3	6

Tableau 1: Description des variables)

1.2. Statistiques descriptives

L'analyse de ces statistiques met en évidence plusieurs éléments importants. Le prix moyen des biens (environ 2 108 milliers d'euros) est légèrement supérieur à la médiane, suggérant une asymétrie positive modérée de la distribution des prix. Cette configuration est classique sur les marchés immobiliers, où quelques biens de très grande valeur tendent à tirer la moyenne vers le haut.

La dispersion du prix est relativement élevée, comme en témoigne un écart-type important, indiquant une forte hétérogénéité des biens étudiés. Des variables telles que la surface habitable, le revenu médian du quartier et les distances géographiques présentent également une variabilité marquée, traduisant des différences significatives entre quartiers et types de logements.

Ces premiers résultats soulignent la nécessité d'une analyse plus approfondie des distributions, notamment à travers des analyses graphiques et des transformations appropriées

	n	mean	median	std	min	Q1	Q3	max
Ascenseur	150.0	0.460	0.000	0.500	0.00	0.000	1.000	1.00
Chambres	150.0	2.887	3.000	1.078	1.00	2.000	4.000	5.00
Etage	150.0	2.580	2.500	1.762	0.00	1.000	4.000	5.00
Qualite_ecole	150.0	5.469	5.600	1.868	1.00	4.125	7.000	10.00
Surface_m2	150.0	116.707	117.845	37.694	15.21	93.240	139.638	218.53
Distance_centre_km	150.0	16.500	16.865	9.017	0.83	9.105	24.698	29.99
Revenu_median_quartier	150.0	63.668	63.450	9.295	42.90	57.500	70.475	83.90
Distance_universite	150.0	8.064	8.300	3.747	1.00	5.300	10.875	17.10
Annee_construction	150.0	2001.827	2002.500	11.705	1980.00	1991.000	2012.000	2022.00
Annee_vente	150.0	2019.840	2020.000	2.288	2015.00	2018.000	2022.000	2023.00
Prix_milliers_euros	150.0	2107.905	2105.050	229.921	1500.77	1934.285	2272.780	2743.04

Tableau 2 : Statistiques descriptives

1.3 Analyse de la distribution du prix

La distribution du prix présente une asymétrie positive modérée, confirmée par une skewness positive. L'histogramme met en évidence une distribution globalement unimodale, avec une queue légèrement plus étendue vers les valeurs élevées.

Le boxplot montre une étendue relativement large des prix, traduisant une hétérogénéité notable des biens, sans toutefois faire apparaître de valeurs aberrantes extrêmement isolées.

Le calcul de la kurtosis excédentaire, positive, suggère des queues légèrement plus épaisses que celles d'une distribution normale, ce qui est cohérent avec les observations graphiques.

Ces caractéristiques justifient l'utilisation éventuelle d'une transformation logarithmique du prix afin de stabiliser la variance et de faciliter l'estimation des modèles linéaires. Les histogrammes et boîtes à moustaches (Figure 1) confirment visuellement ces propriétés de la distribution.

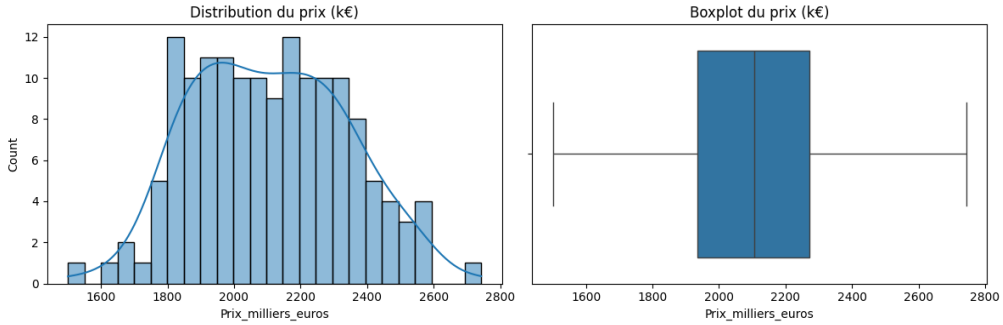


Figure 1 : Histogramme et boxplot du prix

1.4. Corrélations entre variables

La matrice de corrélation *(Figure 2)* présente les coefficients de corrélation linéaire entre les principales variables quantitatives. Les résultats montrent que le prix est fortement et positivement corrélé à la surface habitable (corrélation d'environ 0,83), confirmant l'importance de la taille du logement dans la formation des prix.

Le prix est également négativement corrélé à la distance au centre-ville, avec une corrélation modérée (environ –0,31), suggérant que les biens plus éloignés du centre tendent à être moins chers. Par ailleurs, le revenu médian du quartier présente une corrélation positive mais plus modérée avec le prix.

Les corrélations observées entre les variables explicatives restent globalement faibles à modérées, ce qui limite, à ce stade, les risques de multicolinéarité dans les modèles de régression multiple. Il convient toutefois de rappeler qu'une corrélation élevée n'implique pas nécessairement une relation causale directe.

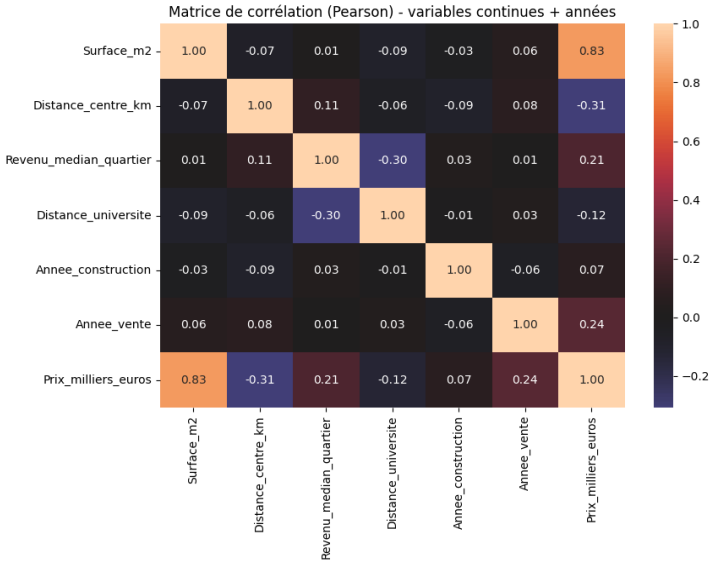


Figure 2 : Matrice de corrélation

2. Modèle linéaire simple & multiple

2.1. Modèle linéaire

Le coefficient associé à la variable explicative est positif et statistiquement significatif, indiquant qu'une augmentation de cette variable est associée à une hausse du prix.

Par exemple, dans le cas de la surface, le coefficient estimé suggère qu'une augmentation d'un mètre carré est associée à une augmentation moyenne du prix, toutes choses égales par ailleurs dans ce modèle simplifié.

	coef	std_err	t	p_value
const	1519.3743	34.5845	43.9323	0.0
Surface_m2	5.0428	0.2821	17.8771	0.0

Tableau 3 : Résultats du modèle linéaire simple

La significativité statistique du coefficient est évaluée à l'aide d'un test t de Student.

- L'hypothèse nulle $H_0 : \beta_1 = 0$est rejetée au seuil de 5 %.
- Le test global du modèle (test F) confirme que le modèle explique une part significative de la variabilité du prix.

Le coefficient de détermination (R^2) indique toutefois que ce modèle simple ne capture qu'une fraction limitée de la variance totale, ce qui motive l'introduction de variables supplémentaires.

2.2. Résultats du modèle multiple

Plusieurs variables apparaissent statistiquement significatives :

- la surface habitable ;
- la distance au centre-ville ;
- le revenu médian du quartier.

La présence d'un ascenseur a également un effet positif sur le prix, bien que son impact soit plus modéré.

Le coefficient de détermination du modèle multiple est sensiblement plus élevé que celui du modèle simple, indiquant une meilleure capacité explicative.

	coef	std_err	t	p_value
const	-46139.7002	5908.2589	-7.8094	0.0000
Surface_m2	4.2551	0.2134	19.9407	0.0000
Distance_centre_km	-6.9778	0.7297	-9.5621	0.0000
Revenu_median_quartier	2.9893	0.8728	3.4247	0.0008
Distance_universite	3.2584	1.9353	1.6836	0.0945
Chambres	36.4225	7.4741	4.8732	0.0000
Ascenseur	53.3182	13.0958	4.0714	0.0001
Qualite_ecole	21.8418	4.6021	4.7460	0.0000
Annee_vente	21.8370	2.8435	7.6797	0.0000
Annee_construction	1.6451	0.5577	2.9498	0.0037

Tableau 4 : Résultats du modèle linéaire multiple

2.3. Transformations logarithmiques

L'analyse descriptive a mis en évidence une asymétrie positive modérée de la distribution du prix ainsi qu'une dispersion importante, suggérant que l'hypothèse d'homoscédasticité pourrait être mise à mal dans un cadre linéaire en niveau. Dans ce contexte, l'introduction de transformations logarithmiques constitue une approche classique permettant de stabiliser la variance des résidus et de faciliter l'interprétation économique des coefficients.

Dans un premier temps, un modèle semi-logarithmique est estimé, dans lequel le logarithme du prix est expliqué par les variables explicatives conservées dans le modèle linéaire multiple. Cette spécification permet d'interpréter les coefficients associés aux variables quantitatives comme des variations relatives approximatives du prix. Les résultats montrent que les signes des coefficients restent globalement inchangés par rapport au modèle en niveau, tandis que certaines variables gagnent en significativité. Le pouvoir explicatif du modèle, mesuré par le R^2 ajusté, s'améliore légèrement, ce qui suggère une meilleure adéquation de cette spécification aux données.

Dans un second temps, un modèle log-log est estimé, dans lequel le logarithme du prix et le logarithme de certaines variables continues, notamment la surface habitable et la distance au centre-ville, sont introduits. Cette spécification permet une interprétation directe des coefficients en termes d'élasticités. Les résultats indiquent que l'élasticité du prix par rapport à la surface est positive et statistiquement significative, confirmant le rôle central de la surface dans la détermination des prix immobiliers. Toutefois, le gain en termes de qualité d'ajustement par rapport au modèle semi-logarithmique reste limité.

La comparaison des trois spécifications, modèle linéaire en niveau, modèle semi-logarithmique et modèle log-log, est réalisée à l'aide du R^2 ajusté ainsi que de critères d'information. Les résultats montrent que les modèles intégrant une transformation logarithmique présentent globalement de meilleures propriétés statistiques que le modèle

en niveau. Parmi eux, le modèle semi-logarithmique offre un compromis satisfaisant entre qualité d'ajustement et interprétabilité économique.

Au regard de ces éléments, le modèle semi-logarithmique est retenu comme spécification privilégiée pour la suite de l'analyse. Il permet de mieux prendre en compte l'hétérogénéité des prix, tout en offrant une interprétation claire des effets marginaux des variables explicatives en termes de variations relatives du prix.

2.4. Tests de significativité et qualité du modèle

La significativité des coefficients est évaluée à l'aide de tests t individuels, tandis que le test F global permet de tester la significativité conjointe des variables explicatives.

Les résultats montrent que :

- le modèle est globalement significatif ;
- la majorité des coefficients ont le signe attendu ;
- certaines variables perdent leur significativité lorsqu'elles sont introduites conjointement, ce qui peut être lié à des corrélations entre variables explicatives.

Ces éléments confirment l'importance de considérer la multicolinéarité et la spécification du modèle.

Cette première partie a permis de décrire le jeu de données, d'analyser les distributions et d'identifier les principaux déterminants du prix immobilier à l'aide de modèles linéaires de base.
Les résultats montrent que les caractéristiques structurelles et de localisation jouent un rôle central dans la formation des prix, tout en soulignant les limites d'une approche strictement linéaire.

PARTIE 2 : DIAGNOSTICS ET CORRECTIONS DU MODÈLE

Après l'estimation du modèle de régression visant à expliquer le prix des logements, il est indispensable d'en vérifier la validité économétrique. Les résultats obtenus par les moindres carrés ordinaires ne sont interprétables que si un certain nombre d'hypothèses classiques sont raisonnablement satisfaites. Cette partie est consacrée à l'analyse des diagnostics du modèle et aux corrections nécessaires en cas de violation de ces hypothèses.

Nous examinons successivement la présence éventuelle de multicolinéarité entre les variables explicatives, les problèmes liés à l'hétéroscédasticité et leurs corrections, puis la stabilité structurelle du modèle à travers un test de Chow. Cette démarche progressive permet de s'assurer de la robustesse des estimations et de la fiabilité de l'inférence statistique.

1. Multicolinéarité et observations influentes

1.1 Diagnostic de la multicolinéarité

La multicolinéarité correspond à une situation dans laquelle certaines variables explicatives du modèle sont fortement corrélées entre elles. Bien que la multicolinéarité n'entraîne pas de biais des estimateurs MCO, elle peut provoquer une forte inflation de leurs écarts-types, rendant les coefficients mal identifiés et les tests de significativité peu fiables.

Afin de diagnostiquer la présence de multicolinéarité, nous utilisons les facteurs d'inflation de la variance (Variance Inflation Factors, VIF). Le VIF mesure dans quelle mesure la variance estimée d'un coefficient est augmentée du fait de la corrélation avec les autres régresseurs. En pratique, des valeurs de VIF supérieures à 5, voire 10 selon les conventions, sont généralement considérées comme problématiques.

Les résultats obtenus montrent que l'ensemble des variables explicatives présentent des VIF compris entre 1 et 2. Ces valeurs sont très proches de l'unité et largement inférieures aux seuils critiques usuels. Elles indiquent l'absence de multicolinéarité significative entre les régresseurs. Les coefficients du modèle sont donc bien identifiés et peuvent être interprétés de manière fiable.

	variable	VIF
0	const	840442.999003
6	Qualite_ecole	1.788673
3	Revenu_median_quartier	1.588558
5	Chambres	1.567681
1	Surface_m2	1.557348
4	Distance_universite	1.270722
2	Distance_centre_km	1.042918
8	Annee_construction	1.019259
7	Annee_vente	1.016188

Diagnostic VIF

1.2 Conséquences et biais de variable omise

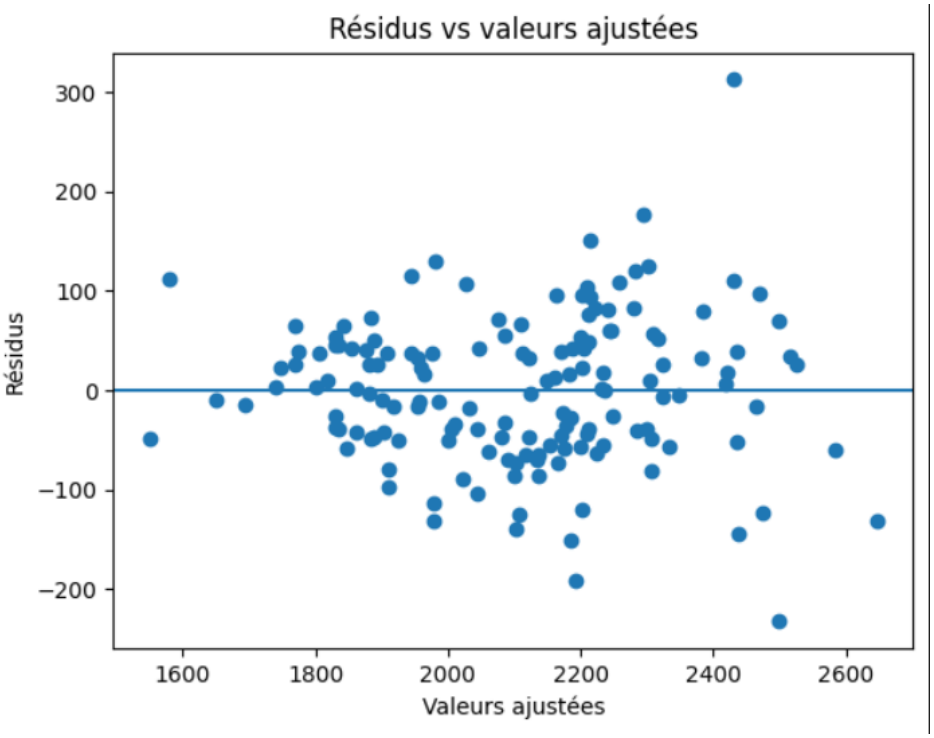
En l'absence de multicolinéarité problématique, il n'est pas nécessaire de supprimer certaines variables explicatives du modèle. Une telle suppression pourrait au contraire introduire un biais de variable omise.

Le biais de variable omise survient lorsqu'une variable explicative pertinente est exclue du modèle alors qu'elle est corrélée à la variable dépendante et à au moins une variable explicative incluse. Dans ce cas, le terme d'erreur devient corrélé avec les régresseurs, ce qui entraîne des estimateurs MCO biaisés et inconsistants. Afin d'éviter ce biais, il est préférable de conserver l'ensemble des variables explicatives pertinentes lorsque la multicolinéarité n'est pas un problème.

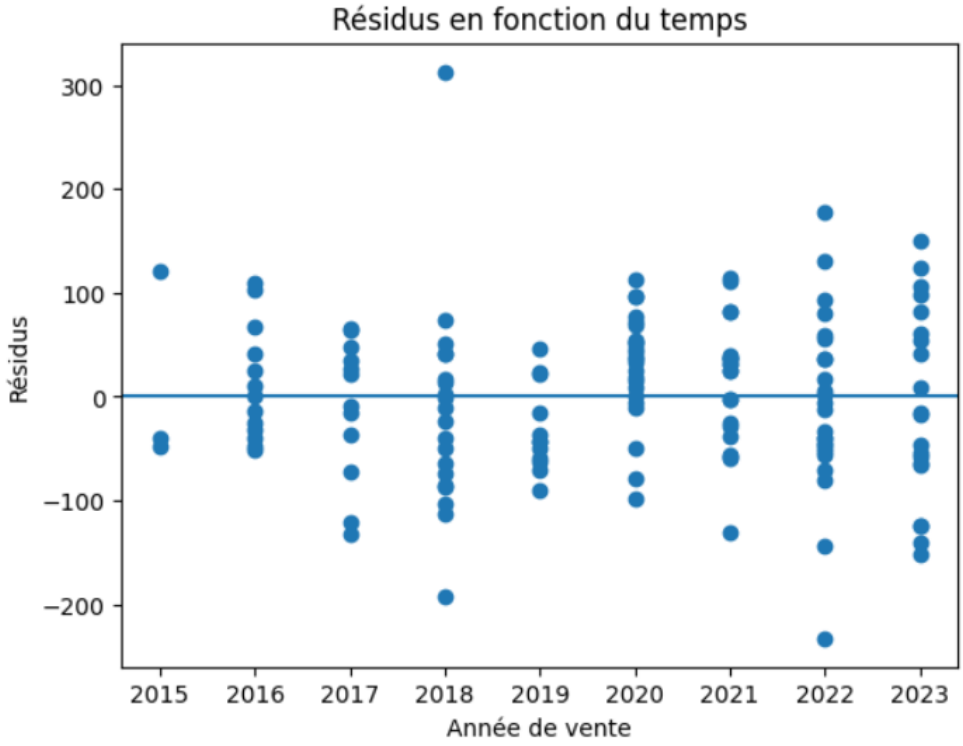
2. Tests d'hétéroscédasticité et corrections

2.1 Observation graphique des résidus

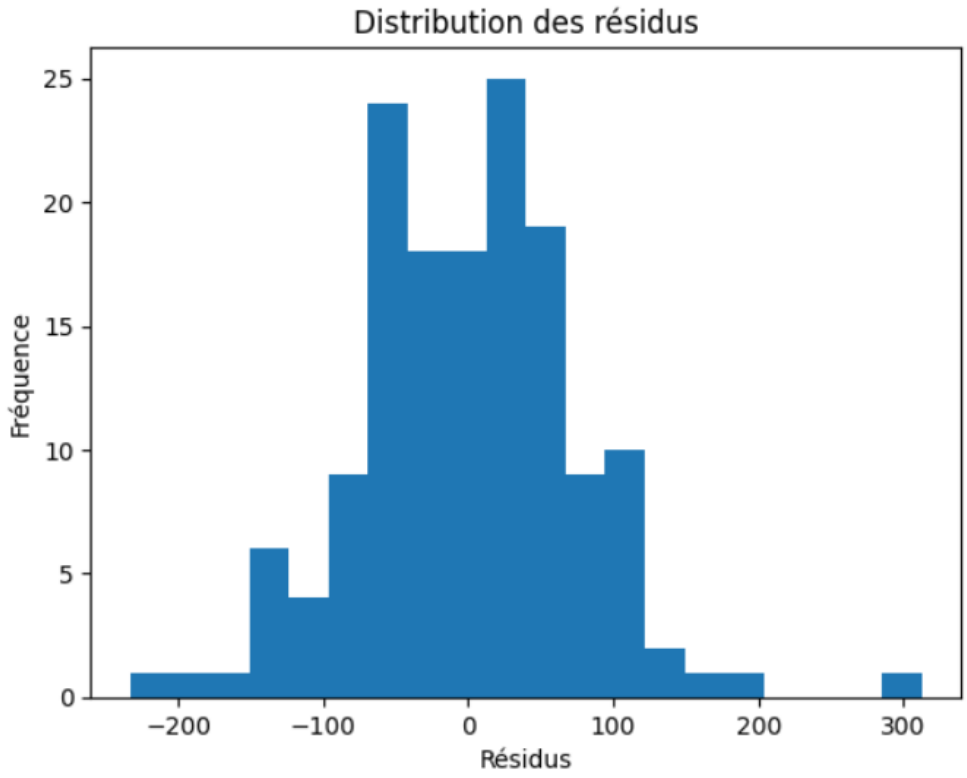
Avant de procéder à des tests formels, une analyse graphique des résidus permet d'évaluer visuellement la validité des hypothèses du modèle linéaire. Le graphique des résidus en fonction des valeurs ajustées montre que les résidus sont globalement centrés autour de zéro, ce qui est cohérent avec l'hypothèse d'espérance nulle des erreurs. Toutefois, leur dispersion tend à augmenter pour les valeurs ajustées élevées, suggérant une possible hétéroscédasticité.



Le graphique des résidus en fonction du temps ne met pas en évidence de tendance temporelle ni de structure systématique, ce qui ne suggère pas la présence d'une autocorrélation marquée des erreurs.



Enfin, l'histogramme des résidus indique une distribution approximativement symétrique et centrée autour de zéro, bien que quelques écarts à la normalité et des valeurs extrêmes soient observés. Ces caractéristiques sont fréquentes dans des données immobilières et motivent la réalisation de tests formels d'hétéroscédasticité.



2.2 Test formel de l'hétéroscédasticité

Afin de tester formellement la présence d'hétéroscédasticité, nous utilisons le test de Breusch–Pagan. L'hypothèse nulle de ce test est celle d'homoscédasticité des erreurs, tandis que l'hypothèse alternative correspond à une variance non constante.

Les résultats du test indiquent des p-values associées aux statistiques LM et F respectivement égales à 0.0048 et 0.0033. Ces valeurs étant inférieures au seuil de 5 %, l'hypothèse nulle d'homoscédasticité est rejetée. Il existe donc une hétéroscédasticité significative dans les résidus du modèle.

2.3 Correction de l'hétéroscédasticité

En présence d'hétéroscédasticité, les estimateurs MCO demeurent non biaisés mais les écarts-types classiques ne sont plus valides, ce qui rend les tests statistiques incorrects. Afin de corriger ce problème, nous utilisons des écarts-types robustes à l'hétéroscédasticité, selon la méthode de White (HC1).

Cette correction permet d'obtenir des écarts-types, des statistiques t et des p-values valides sans modifier les coefficients estimés. La comparaison entre les p-values

obtenues avec des écarts-types classiques et celles calculées à l'aide d'écarts-types robustes montre que, bien que les p-values robustes soient légèrement plus élevées, les conclusions principales demeurent inchangées. Toutes les variables explicatives clés restent statistiquement significatives après correction.

Pour rappel, les conclusions initiales du modèle MCO standard indiquaient que la surface, le nombre de chambres, la présence d'un ascenseur, la qualité des écoles et le revenu médian du quartier ont un effet positif et significatif sur le prix des logements, tandis que la distance au centre-ville a un effet négatif significatif. Ces conclusions sont confirmées après correction de l'hétéroscédasticité.

2.4 Comparaison des méthodes d'estimation

Afin d'approfondir l'analyse, nous comparons les résultats obtenus par trois méthodes d'estimation : les MCO standard, les MCO avec écarts-types robustes et les moindres carrés pondérés (WLS).

Les estimations obtenues par MCO standard et par MCO robustes présentent des coefficients strictement identiques, la correction affectant uniquement l'inférence statistique. Le modèle WLS, estimé à l'aide de poids dérivés des résidus, présente des instabilités numériques et ne permet pas d'obtenir l'ensemble des p-values. Cette sensibilité au choix des poids illustre les limites du WLS lorsque la structure de la variance des erreurs est mal spécifiée.

Dans ce contexte, les MCO avec écarts-types robustes apparaissent comme la méthode la plus appropriée : ils permettent de corriger efficacement l'hétéroscédasticité tout en conservant des résultats stables et une interprétation économique fiable.

	coef_MCO	p_MCO	p_MCO_robust	p_WLS
const	-45090.835923	2.378765e-12	1.367280e-10	1.096874e-55
Surface_m2	4.229384	5.519135e-43	2.042938e-37	NaN
Chambres	35.151060	4.551008e-06	4.618333e-07	NaN
Annee_construction	1.699645	2.492548e-03	4.217612e-03	NaN
Distance_centre_km	-7.056503	1.413744e-17	1.715150e-16	NaN
Etage	9.132240	1.390445e-02	1.051970e-02	NaN
Ascenseur	54.926324	4.123926e-05	3.338228e-05	NaN
Annee_vente	21.281043	5.599023e-12	1.267627e-10	NaN
Qualite_ecole	19.914911	7.359414e-06	1.915826e-06	NaN
Revenu_median_quartier	2.759262	1.746625e-03	2.307437e-04	NaN

Remarque : certaines p-values du modèle WLS ne sont pas disponibles en raison d'instabilités numériques liées au choix des poids. Cela illustre la sensibilité du WLS à la spécification de la variance et renforce l'intérêt des MCO avec écarts-types robustes.

2.5 Test d'autocorrélation

Le test de Durbin–Watson est utilisé afin de détecter une éventuelle autocorrélation des résidus. La statistique obtenue est égale à 2.28, valeur proche de 2, ce qui indique l'absence d'autocorrélation significative. Ainsi, bien que l'hétéroscédasticité ait été détectée, l'hypothèse d'indépendance des erreurs n'est pas rejetée. L'utilisation d'écarts-types de Newey–West n'est donc pas nécessaire.

3. Stabilité structurelle : Test de Chow

La stabilité structurelle du modèle est analysée à l'aide d'un test de Chow, avec une rupture supposée en 2020 afin d'évaluer l'impact du COVID sur le marché immobilier. Le test compare un modèle estimé sur l'ensemble de la période à deux modèles estimés séparément avant et après la rupture.

Le test fournit une statistique F égale à 1.31 et une p-value de 0.23. Cette valeur étant supérieure au seuil de 5 %, l'hypothèse nulle de stabilité des coefficients n'est pas rejetée. Il n'existe donc pas de preuve statistique d'une rupture structurelle liée au COVID. L'estimation d'un modèle unique sur l'ensemble de la période est donc appropriée.

L’analyse des diagnostics montre que le modèle est globalement bien spécifié. Aucune multicolinéarité problématique n’est détectée, l’hétéroscédasticité identifiée est correctement corrigée à l’aide d’écarts-types robustes, et aucune autocorrélation significative des résidus n’est mise en évidence. Enfin, le test de Chow ne révèle pas de rupture structurelle liée au COVID. Ces résultats garantissent la fiabilité de l’inférence statistique et la robustesse des conclusions du modèle.

PARTIE 3 : Endogénéité et Variables Instrumentales

Après avoir validé la spécification économétrique du modèle linéaire et corrigé les principales violations des hypothèses classiques (hétéroscédasticité, stabilité structurelle), il est nécessaire de s'interroger sur la possible endogénéité de certaines variables explicatives.
La présence d'endogénéité conduit à des estimateurs MCO biaisés et inconsistants, remettant en cause l'interprétation causale des coefficients estimés.

Dans cette partie, nous discutons les sources potentielles d'endogénéité dans notre modèle, puis nous mettons en œuvre une estimation par variables instrumentales (IV / 2SLS) afin d'évaluer la robustesse des résultats obtenus par MCO.

3.1. Endogénéité potentielle de la qualité des écoles

La qualité des écoles peut être corrélée à des caractéristiques non observées des quartiers, telles que leur attractivité globale ou leur composition socio-économique, qui influencent également les prix immobiliers. Dans ce cas, l'estimation MCO capterait non seulement l'effet propre de la qualité scolaire, mais aussi celui de ces facteurs omis. Par ailleurs, un mécanisme de causalité inverse peut exister : les quartiers les plus chers attirent des ménages favorisés, ce qui peut contribuer à améliorer la qualité des établissements scolaires. Ces éléments justifient l'examen d'une stratégie alternative afin de traiter un éventuel biais d'endogénéité.

D'autres variables, telles que le revenu médian du quartier, pourraient également être soupçonnées d'endogénéité. Toutefois, en l'absence d'instruments crédibles, l'analyse se concentre sur la variable Qualite_ecole, pour laquelle une stratégie IV peut être mise en œuvre.

3.2. Choix de la variable instrumentale

Pour corriger ce biais potentiel, nous utilisons la distance à l'université la plus proche (Distance_universite) comme instrument pour Qualite_ecole.

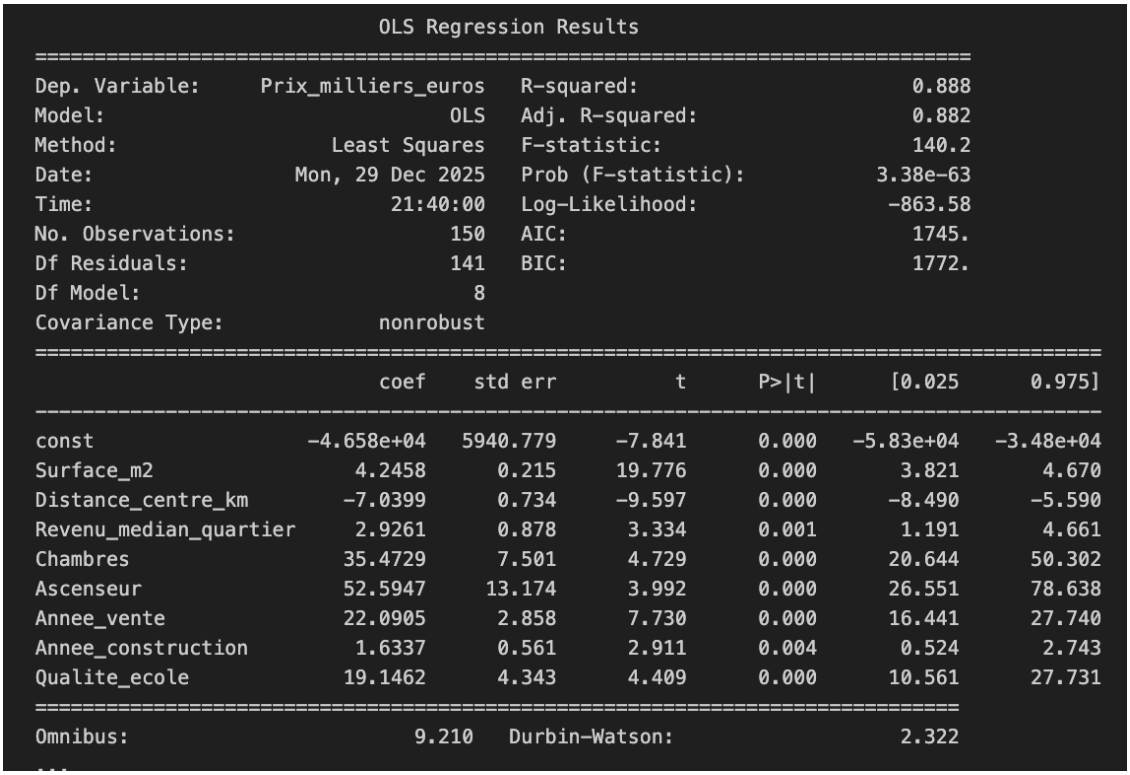
Un instrument valide doit satisfaire deux conditions :

- Pertinence : il doit être corrélé avec la variable endogène ;
- Exogénéité : il ne doit pas affecter directement la variable dépendante, conditionnellement aux contrôles.

Dans notre cas, la distance à l'université peut influencer la qualité des écoles via la composition socio-éducative du quartier, sans affecter directement le prix immobilier une fois les variables de localisation et de revenu contrôlées.

3.3. Estimation MCO de référence

L'estimation MCO de référence indique que Qualite_ecole a un effet positif et statistiquement significatif sur le prix des logements. Le coefficient estimé est d'environ 19,15, ce qui correspond à une augmentation moyenne d'environ 19 000 euros du prix pour une hausse d'un point de l'indice de qualité des écoles, toutes choses égales par ailleurs.



OLS Regression Results

Dep. Variable:	Prix_milliers_euros	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	140.2
Date:	Mon, 29 Dec 2025	Prob (F-statistic):	3.38e-63
Time:	21:40:00	Log-Likelihood:	-863.58
No. Observations:	150	AIC:	1745.
Df Residuals:	141	BIC:	1772.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.658e+04	5940.779	-7.841	0.000	-5.83e+04	-3.48e+04
Surface_m2	4.2458	0.215	19.776	0.000	3.821	4.670
Distance_centre_km	-7.0399	0.734	-9.597	0.000	-8.490	-5.590
Revenu_median_quartier	2.9261	0.878	3.334	0.001	1.191	4.661
Chambres	35.4729	7.501	4.729	0.000	20.644	50.302
Ascenseur	52.5947	13.174	3.992	0.000	26.551	78.638
Annee_vente	22.0905	2.858	7.730	0.000	16.441	27.740
Annee_construction	1.6337	0.561	2.911	0.004	0.524	2.743
Qualite_ecole	19.1462	4.343	4.409	0.000	10.561	27.731

Omnibus:	9.210	Durbin-Watson:	2.322

3.4. Test de pertinence de l'instrument (première étape)

La première étape de l'estimation IV consiste à régresser Qualite_ecole sur Distance_universite et l'ensemble des variables de contrôle.

Les résultats indiquent que Distance_universite est statistiquement significative dans cette équation.
Le test de Fisher de pertinence de l'instrument donne une statistique F ≈ 19.4, largement supérieure au seuil conventionnel de 10, ce qui suggère que l'instrument n'est pas faible.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Qualite_ecole   R-squared:                       0.441
Model:                            OLS   Adj. R-squared:                  0.409
Method:                 Least Squares   F-statistic:                     13.90
Date:                Mon, 29 Dec 2025   Prob (F-statistic):           9.02e-15
Time:                        21:40:00   Log-Likelihood:                -262.48
No. Observations:                 150   AIC:                             543.0
Df Residuals:                     141   BIC:                             570.0
Df Model:                           8
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    -39.2664    108.067     -0.363      0.717    -252.907     174.374
Distance_universite       -0.1463      0.033     -4.406      0.000      -0.212      -0.081
Surface_m2                 0.0026      0.004      0.675      0.501      -0.005       0.010
Distance_centre_km        -0.0084      0.013     -0.631      0.529      -0.035       0.018
Revenu_median_quartier     0.1034      0.013      7.721      0.000       0.077       0.130
Chambres                  -0.1398      0.136     -1.026      0.307      -0.409       0.130
Ascenseur                  0.0141      0.240      0.059      0.953      -0.460       0.488
Annee_vente                0.0234      0.052      0.449      0.654      -0.079       0.126
Annee_construction        -0.0038      0.010     -0.374      0.709      -0.024       0.016
==============================================================================
Omnibus:                        0.139   Durbin-Watson:                   2.124
...
strong multicollinearity or other numerical problems.

F-test (pertinence) : Distance_universite = 0
<F test: F=19.414266862031596, p=2.0683792907013486e-05, df_denom=141, df_num=1>
```

3.5 Estimation IV (2SLS) et interprétation

Dans la seconde étape de l'estimation par variables instrumentales, le prix des logements est régressé sur la valeur prédite de la variable Qualite_ecole issue de la première étape, ainsi que sur l'ensemble des variables de contrôle. Cette approche permet d'isoler la composante exogène de la qualité des écoles afin d'en évaluer l'effet causal sur les prix immobiliers.

Les résultats montrent que le coefficient associé à Qualite_ecole_hat est estimé à –0,43 et n'est pas statistiquement significatif, avec une p-value élevée. Cette perte de significativité s'accompagne d'un écart-type sensiblement plus élevé que dans l'estimation MCO, reflétant une baisse de précision inhérente à l'utilisation d'un instrument.

Ainsi, bien que l'estimation IV corrige un biais potentiel d'endogénéité, elle conduit ici à des résultats beaucoup moins précis que ceux obtenus par MCO. En l'absence de significativité statistique du coefficient IV et compte tenu de la forte augmentation de sa variance, les résultats ne fournissent pas de preuve empirique robuste d'un effet causal distinct de la qualité des écoles sur les prix immobiliers par rapport à l'estimation MCO.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Prix_milliers_euros   R-squared:                       0.873
Model:                             OLS   Adj. R-squared:                  0.866
Method:                  Least Squares   F-statistic:                     121.1
Date:                 Mon, 29 Dec 2025   Prob (F-statistic):           2.89e-59
Time:                         21:40:00   Log-Likelihood:                -873.27
No. Observations:                  150   AIC:                             1765.
Df Residuals:                      141   BIC:                             1792.
Df Model:                            8
Covariance Type:             nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                  -4.701e+04   6343.126     -7.412      0.000   -5.96e+04   -3.45e+04
Qualite_ecole_hat         -0.4306     13.315     -0.032      0.974     -26.754      25.893
Surface_m2                 4.3137      0.233     18.508      0.000       3.853       4.774
Distance_centre_km        -7.1652      0.787     -9.110      0.000      -8.720      -5.610
Revenu_median_quartier     5.2920      1.776      2.980      0.003       1.782       8.802
Chambres                  33.3079      8.120      4.102      0.000      17.256      49.360
Ascenseur                 53.6316     14.068      3.812      0.000      25.820      81.443
Annee_vente               22.3574      3.053      7.322      0.000      16.321      28.394
Annee_construction         1.5601      0.601      2.598      0.010       0.373       2.747
==============================================================================
Omnibus:                        3.494   Durbin-Watson:                   2.470
```

3.6 Test d'endogénéité

Afin de tester formellement l'endogénéité de Qualite_ecole, un test de type Durbin–Wu–Hausman est mis en œuvre à l'aide d'une approche par fonction de contrôle.

Le coefficient du terme de correction issu de la première étape n'est pas statistiquement significatif au seuil de 5 %, avec une p-value d'environ 0,094. Par conséquent,

l'hypothèse nulle d'exogénéité de Qualite_ecole ne peut pas être rejetée. Autrement dit, les données ne fournissent pas de preuve statistique forte en faveur d'un biais d'endogénéité dans l'estimation MCO.

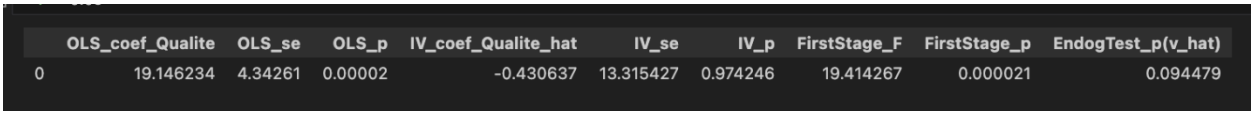
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Prix_milliers_euros   R-squared:                       0.891
Model:                             OLS   Adj. R-squared:                  0.884
Method:                  Least Squares   F-statistic:                     126.6
Date:                 Mon, 29 Dec 2025   Prob (F-statistic):           1.03e-62
Time:                         21:40:01   Log-Likelihood:                -862.08
No. Observations:                  150   AIC:                             1744.
Df Residuals:                      140   BIC:                             1774.
Df Model:                            9
Covariance Type:             nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                   -4.701e+04   5908.168     -7.958      0.000   -5.87e+04   -3.53e+04
Surface_m2                 4.3137      0.217     19.870      0.000       3.884       4.743
Distance_centre_km        -7.1652      0.733     -9.781      0.000      -8.614      -5.717
Revenu_median_quartier     5.2920      1.654      3.200      0.002       2.022       8.562
Chambres                  33.3079      7.563      4.404      0.000      18.356      48.260
Ascenseur                 53.6316     13.103      4.093      0.000      27.726      79.537
Annee_vente               22.3574      2.844      7.861      0.000      16.735      27.980
Annee_construction         1.5601      0.559      2.789      0.006       0.454       2.666
Qualite_ecole             -0.4306     12.402     -0.035      0.972     -24.951      24.090
v_hat                     22.2724     13.229      1.684      0.094      -3.881      48.426
==============================================================================
...
strong multicollinearity or other numerical problems.

Test : v_hat = 0 (pas d'endogénéité)
<F test: F=2.834661967226041, p=0.09447868902861963, df_denom=140, df_num=1>
```

3.7 Comparaison des estimations MCO et IV

La comparaison entre les estimations MCO et IV met en évidence plusieurs éléments. D'une part, l'estimation IV conduit à un coefficient plus élevé pour la qualité des écoles, suggérant un effet potentiellement sous-estimé par MCO. D'autre part, cette estimation est beaucoup moins précise, ce qui limite son intérêt pratique dans ce contexte.

Compte tenu de l'absence de preuve statistique forte d'endogénéité et de la forte augmentation de la variance associée à l'estimation IV, l'estimateur MCO apparaît comme une approximation satisfaisante de l'effet de Qualite_ecole sur le prix des logements.



	OLS_coef_Qualite	OLS_se	OLS_p	IV_coef_Qualite_hat	IV_se	IV_p	FirstStage_F	FirstStage_p	EndogTest_p(v_hat)
0	19.146234	4.34261	0.00002	-0.430637	13.315427	0.974246	19.414267	0.000021	0.094479

Cette partie a examiné la question de l'endogénéité de la variable Qualite_ecole dans un modèle de prix immobiliers. Bien que des arguments théoriques plausibles justifient une telle suspicion, les tests empiriques ne mettent pas en évidence de biais d'endogénéité statistiquement significatif. L'instrument proposé est pertinent mais conduit à une perte importante de précision.

Dans ce contexte, l'estimation MCO demeure appropriée pour l'analyse et l'interprétation économique. Ces résultats motivent, dans la partie suivante, un changement de perspective vers des méthodes de régularisation, davantage orientées vers la stabilité des estimations et la performance prédictive.

PARTIE 4 : Méthodes de régularisation

Les modèles estimés par moindres carrés ordinaires incluent plusieurs variables explicatives décrivant les caractéristiques des logements et de leur environnement. Même si les diagnostics précédents n'ont pas mis en évidence de multicolinéarité sévère, certaines variables restent naturellement corrélées entre elles, ce qui peut rendre les coefficients instables et sensibles aux variations de l'échantillon.

Dans une perspective plus orientée vers la prédiction, il est donc pertinent d'envisager des méthodes permettant de stabiliser les estimations et de limiter le risque de sur-ajustement. Les méthodes de régularisation répondent à cet objectif en ajoutant une pénalité sur la taille des coefficients. Dans cette partie, nous mettons en œuvre les régressions Ridge et Lasso, puis nous comparons leurs performances prédictives à celles du modèle MCO.

4.1 Régression Ridge : évolution des coefficients

La régression Ridge repose sur l'ajout d'une pénalisation quadratique des coefficients dans la fonction de coût. Avant l'estimation, toutes les variables explicatives sont standardisées afin que la pénalisation s'applique de manière homogène. Le paramètre de régularisation λ contrôle l'intensité de cette pénalité : lorsque λ augmente, les coefficients sont progressivement réduits en valeur absolue.

Les résultats montrent que pour des valeurs faibles de λ, les coefficients Ridge sont très proches de ceux obtenus par MCO. À mesure que λ augmente, l'ensemble des coefficients diminue de façon continue, sans qu'aucun ne devienne exactement nul. Par exemple, lorsque λ passe de 0,001 à environ 0,46, les coefficients restent du même signe mais leur amplitude est légèrement réduite, ce qui traduit un effet de stabilisation du modèle.

	lambda	Surface_m2	Distance_centre_km	Revenu_median_quartier	Distance_universite	Chambres	Ascenseur	Qualite_ecole	Annee_vente	Annee_construction
0	0.001000	165.885059	-64.531669	30.151485	12.379265	37.357394	25.349102	38.562551	49.507293	14.720076
1	0.004642	165.877783	-64.529593	30.151096	12.378183	37.360944	25.347190	38.561440	49.505955	14.719804
2	0.021544	165.844021	-64.519957	30.149293	12.373160	37.377409	25.338316	38.556286	49.499747	14.718541
3	0.100000	165.687593	-64.475282	30.140920	12.349863	37.453598	25.297197	38.532377	49.470953	14.712677
4	0.464159	164.967424	-64.269019	30.101964	12.242118	37.802188	25.107781	38.421730	49.337770	14.685448

Cette propriété fait de la régression Ridge une méthode particulièrement adaptée lorsque l'on souhaite limiter la variance des coefficients et améliorer la robustesse des prédictions, tout en conservant l'ensemble des variables explicatives dans le modèle.

4.2 Régression Lasso : évolution des coefficients et sélection

La régression Lasso repose sur une pénalisation de type L1, qui peut conduire certains coefficients à devenir exactement nuls lorsque la pénalisation augmente. Cette propriété permet, en théorie, d'effectuer une sélection automatique des variables les plus importantes.

Dans notre cas, les résultats montrent que, même pour des valeurs relativement élevées de λ, aucun coefficient ne devient nul. Tous les coefficients diminuent progressivement en valeur absolue, mais restent différents de zéro.

	lambda	Surface_m2	Distance_centre_km	Revenu_median_quartier	Distance_universite	Chambres	Ascenseur	Qualite_ecole	Annee_vente	Annee_construction	nb_coef_non_nuls
0	0.001000	165.886531	-64.531360	30.150947	12.377670	37.355659	25.348264	38.561375	49.506559	14.718963	9
1	0.004642	165.884408	-64.528097	30.148559	12.370891	37.353041	25.343275	38.556066	49.502555	14.714654	9
2	0.021544	165.875075	-64.513133	30.137630	12.339068	37.340491	25.320180	38.531152	49.483959	14.694610	9
3	0.100000	165.828380	-64.442590	30.086057	12.193408	37.284771	25.212591	38.417061	49.397721	14.601837	9
4	0.464159	165.607187	-64.113682	29.845499	11.520129	37.029512	24.712685	37.889635	48.997537	14.171566	9

Ce résultat indique que l'ensemble des variables incluses dans le modèle contribue à l'explication du prix des logements. Le modèle initial ne semble donc pas surchargé en variables peu informatives. Ici, le Lasso joue principalement un rôle de régularisation plutôt que de sélection.

4.3 Choix du paramètre λ par validation croisée

Afin de sélectionner la valeur optimale du paramètre de régularisation λ, une validation croisée à 10 plis est utilisée. Cette méthode consiste à estimer les modèles pour différentes valeurs de λ et à retenir celle qui minimise l'erreur de prédiction moyenne sur les échantillons de validation.

Les résultats indiquent que la valeur optimale de λ pour la régression Ridge est d'environ $\lambda \approx 2{,}15$, ce qui correspond à une pénalisation modérée. Pour la régression Lasso, la valeur optimale de λ est très faible, proche de zéro ($\lambda \approx 0{,}001$), confirmant une pénalisation limitée.

On choisit le λ qui minimise l'erreur en validation croisée.

```
ridge_cv = Pipeline([("scaler", StandardScaler()),
                     ("ridgecv", RidgeCV(alphas=alphas, cv=10))])
ridge_cv.fit(X_train, y_train)
best_ridge = ridge_cv.named_steps["ridgecv"].alpha_

lasso_cv = Pipeline([("scaler", StandardScaler()),
                     ("lassocv", LassoCV(alphas=alphas, cv=10, max_iter=20000))])
lasso_cv.fit(X_train, y_train)
best_lasso = lasso_cv.named_steps["lassocv"].alpha_

best_ridge, best_lasso
```

```
(np.float64(2.154434690031882), np.float64(0.001))
```

Ces résultats suggèrent que seule une régularisation modérée est nécessaire dans ce contexte, en particulier pour le modèle Ridge, ce qui est cohérent avec l'absence de multicolinéarité forte mise en évidence dans les diagnostics précédents.

4.4 Comparaison des performances prédictives (RMSE)

Les performances prédictives des modèles MCO, Ridge et Lasso sont comparées à l'aide de l'erreur quadratique moyenne (RMSE) calculée sur un échantillon de test représentant 20 % des données.

Les résultats montrent que :

- le modèle MCO présente une RMSE d'environ 78,83 k€ ;
- le modèle Ridge (avec λ optimal) obtient une RMSE légèrement plus faible, d'environ 77,91 k€ ;
- le modèle Lasso affiche une RMSE très proche de celle du modèle MCO, également autour de 78,83 k€.

	Modele	RMSE_test	lambda_choisi
0	OLS	78.828824	NaN
1	Ridge (CV)	77.910423	2.154435
2	Lasso (CV)	78.828882	0.001000

La régression Ridge présente la RMSE la plus faible (≈ 77,9 k€), mais l'amélioration par rapport au modèle MCO reste limitée. Les performances prédictives des trois modèles sont donc très proches.

Les méthodes de régularisation améliorent légèrement la performance prédictive du modèle. La régression Ridge obtient la RMSE la plus faible, bien que l'écart avec le modèle MCO reste modeste. La régression Lasso n'élimine aucune variable et fournit des performances proches de celles du MCO.

Dans ce contexte, la régression Ridge apparaît comme le meilleur compromis entre stabilité des coefficients et qualité de prédiction, et elle est retenue pour la prévision finale.

```python
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge

# Ici, best_ridge doit déjà exister (ex: best_ridge = ridge_cv.alpha_)
ridge_final = Pipeline([
    ("scaler", StandardScaler()),
    ("ridge", Ridge(alpha=best_ridge))
])

# X et y doivent être les mêmes que ceux utilisés pour Ridge/Lasso (features finales)
ridge_final.fit(X, y)

print("Ridge final entraîné. alpha =", best_ridge)
```

✓ 0.0s

```
Ridge final entraîné. alpha = 2.154434690031882
```

Contrairement aux MCO, les estimateurs issus du Lasso sont biaisés en raison de la pénalisation introduite dans la fonction de coût. De plus, le processus de sélection de variables dépend directement des données utilisées, ce qui rend la distribution asymptotique des coefficients non standard. Par conséquent, les écarts-types et les tests statistiques classiques ne sont pas valides après Lasso, et l'inférence doit être réalisée à l'aide de méthodes spécifiques telles que le bootstrap ou des approches post-sélection.

Synthèse des résultats de prédiction

Le modèle retenu pour la prédiction finale est la régression Ridge, avec un paramètre de régularisation optimal λ ≈ 2,15, choisi par validation croisée. Ce modèle est utilisé pour prédire le prix d'un logement présentant les caractéristiques spécifiées.

```python
# --- PARTIE 8 : Prediction avec le meilleur modèle (Ridge) ---

# 1) Ordre exact des features utilisé à l'entraînement Ridge
features = [
 'Surface_m2',
 'Distance_centre_km',
 'Revenu_median_quartier',
 'Distance_universite',
 'Chambres',
 'Ascenseur',
 'Qualite_ecole',
 'Annee_vente',
 'Annee_construction'
]

# 2) Nouveau logement à prédire (ATTENTION: pas de Etage ici)
X_new = pd.DataFrame([{
    "Surface_m2": 120,
    "Distance_centre_km": 5,
    "Revenu_median_quartier": 65,
    "Distance_universite": 4,
    "Chambres": 3,
    "Ascenseur": 1,
    "Qualite_ecole": 7,
    "Annee_vente": 2023,
    "Annee_construction": 2015
}])

# 3) On force le même ordre que l'entraînement
X_new = X_new[features]

X_new
```

✓ 0.0s

	Surface_m2	Distance_centre_km	Revenu_median_quartier	Distance_universite	Chambres	Ascenseur	Qualite_ecole	Annee_vente	Annee_construction
0	120	5	65	4	3	1	7	2023	2015

Pour ce logement, le prix prédit par le modèle Ridge est d'environ 2,34 millions d'euros.

```python
pred_k = ridge_final.predict(X_new)[0]   # résultat en k€
print(f"Prix prédit (Ridge) : {pred_k:.2f} k€")
print(f"Soit environ {pred_k*1000:,.0f} €")
```

✓ 0.0s

```
Prix prédit (Ridge) : 2339.73 k€
Soit environ 2,339,727 €
```

Cette valeur est cohérente avec les niveaux de prix observés dans les données, compte tenu de la surface du logement, de sa localisation et de la qualité du quartier. La prédiction doit toutefois être interprétée comme une valeur indicative, et non comme un prix exact, le modèle restant une approximation statistique.

CONCLUSION GÉNÉRALE

Ce travail avait pour objectif d'analyser les déterminants du prix de logements résidentiels à partir d'un jeu de données observationnelles, en mobilisant des outils de statistique descriptive et d'économétrie linéaire. L'approche adoptée s'est articulée autour d'une analyse progressive, allant de la caractérisation des données à l'estimation de modèles plus élaborés, afin d'assurer une interprétation rigoureuse et cohérente des résultats.

L'analyse descriptive a mis en évidence une hétérogénéité marquée des prix immobiliers ainsi qu'une asymétrie positive modérée de leur distribution, éléments caractéristiques des marchés immobiliers. Les statistiques de dispersion et les représentations graphiques ont montré des écarts importants entre les biens, reflétant la diversité des logements et des localisations étudiées. L'analyse de corrélation a permis d'identifier la surface habitable comme la variable la plus fortement associée au prix, tandis que la distance au centre-ville présente une corrélation négative modérée. Les corrélations observées entre les variables explicatives restent globalement limitées, ce qui réduit les risques de multicolinéarité au stade descriptif.
Les modèles de régression linéaire estimés par la méthode des moindres carrés ordinaires confirment ces constats. Le modèle linéaire simple met en évidence un effet positif et statistiquement significatif de la surface sur le prix. L'introduction de variables supplémentaires dans le modèle linéaire multiple améliore nettement la qualité de l'ajustement et permet d'interpréter les effets marginaux toutes choses égales par ailleurs. Les résultats montrent que les caractéristiques structurelles du logement et les variables de localisation jouent un rôle central dans la formation des prix, tandis que certaines variables perdent leur significativité une fois les autres contrôlées.

L'étude des transformations logarithmiques du prix et de certaines variables explicatives apporte un éclairage complémentaire. Les modèles semi-logarithmiques et logarithmiques facilitent l'interprétation économique des coefficients en termes de variations relatives et contribuent à une meilleure stabilité de la variance des résidus. Les diagnostics économétriques mettent en évidence la présence d'hétéroscédasticité, corrigée par l'utilisation d'écarts-types robustes, ainsi qu'une multicolinéarité globalement limitée. Par ailleurs, l'analyse de l'endogénéité potentielle de certaines variables, notamment la qualité des écoles, ne révèle pas de biais statistiquement significatif, ce qui conforte l'utilisation de l'estimateur MCO dans ce cadre.

Dans une perspective plus orientée vers la prévision, les méthodes de régularisation, telles que les régressions Ridge et Lasso, ont été mises en œuvre afin d'évaluer la robustesse et la performance prédictive des modèles. Les résultats montrent que la régularisation permet de stabiliser les estimations et d'améliorer légèrement les performances prédictives par rapport au modèle linéaire standard, bien que les gains restent modestes. Ces résultats soulignent le compromis classique entre interprétabilité et performance prédictive.

Malgré ces apports, plusieurs limites doivent être soulignées. L'analyse repose sur des données observationnelles et sur un ensemble restreint de variables, ce qui peut conduire à un biais lié à des facteurs non observés. De plus, le cadre linéaire impose des hypothèses restrictives sur la forme fonctionnelle des relations entre les variables et ne permet pas de capturer d'éventuelles non-linéarités. Enfin, les résultats obtenus ne doivent pas être interprétés comme des relations causales, mais comme des associations conditionnelles.

Des prolongements naturels de ce travail consisteraient à enrichir la base de données par des variables supplémentaires décrivant plus finement l'environnement des logements, à explorer des modèles non linéaires ou des approches semi-paramétriques, et à approfondir l'analyse prédictive à l'aide de méthodes d'apprentissage automatique.

En définitive, ce travail met en évidence l'importance d'une démarche statistique rigoureuse et progressive pour analyser les prix immobiliers et souligne le rôle central des caractéristiques structurelles et de localisation dans la formation des prix.

ANNEXES

- Notebook.ipynb
- Diagnostic_VIF
- donnees_immobilieres_extended
- Tableau2_stats_descriptives
- Tableau3_modele_simple
- Tableau4_modele_multiple