Université Paris 1 Panthéon-Sorbonne
DU — Sorbonne Data Analytics

PROJET D'ÉCONOMÉTRIE APPLIQUÉE

Analyse des Prix Immobiliers

Du modèle linéaire aux méthodes de régularisation

Membres :

MANELLI Cédric,

ALLISON Jacques,

NADAT Sufyan

Décembre 2025
Année universitaire 2025-2026

Table des Matières

Résumé Exécutif

Principaux Résultats

Cette étude économétrique analyse les déterminants du prix immobilier à partir d'un échantillon de 150 transactions réalisées entre 2015 et 2023. L'analyse révèle plusieurs résultats majeurs :

- La surface habitable constitue le déterminant principal du prix immobilier. Dans le cadre de la régression linéaire simple, une augmentation de 1 m² de surface est associée à une hausse moyenne du prix d'environ 5 40 €. Ce modèle univarié explique 68,35 % de la variance du prix ($R^2 = 0,6835$)

- Le modèle de régression multiple explique 78,9 % de la variance des prix (R^2 ajusté = 0,780). La surface, le nombre de chambres, la distance au centre-ville, l'étage et la présence d'un ascenseur sont statistiquement significatifs. L'éloignement du centre entraîne une décote moyenne de 6 140 € par kilomètre.

- Les modèles logarithmiques confirment la robustesse des résultats. Le modèle log-linéaire présente le meilleur ajustement ($R^2 \approx 0,792$), tandis que le modèle log-log indique une élasticité du prix à la surface d'environ 0,19.

- La période Covid est associée à une hausse moyenne des prix d'environ 103 700 €, sans rupture structurelle globale des coefficients (test de Chow non significatif).

- Les diagnostics révèlent une hétéroscédasticité modérée, corrigée par des écarts-types robustes, sans multicolinéarité ni autocorrélation significative.

- L'analyse par variables instrumentales met en évidence un biais d'endogénéité positif pour la qualité des écoles, dont l'effet devient non significatif une fois instrumenté.

- La prédiction finale d'un bien type est estimée à 2 255 539 €, avec un intervalle de prévision à 95 % de [2 072 207 € ; 2 438 871 €], soit une incertitude d'environ 16 % du prix estimé, indiquant une prédiction fiable.

Recommandations

1. Utiliser les écarts-types robustes de White pour l'inférence statistique en raison de l'hétéroscédasticité détectée.
2. Privilégier le modèle log-linéaire pour une interprétation en termes de variations relatives.
3. Interpréter l'effet de la qualité des écoles avec prudence en raison du biais d'endogénéité identifié.
4. Considérer les méthodes Ridge/Lasso pour la prédiction, qui offrent des performances comparables à l'OLS avec une meilleure stabilité.

Introduction

Contexte et Problématique

Le marché immobilier constitue un secteur économique majeur dont l'analyse requiert des outils économétriques sophistiqués. La compréhension des facteurs influençant les prix immobiliers revêt une importance capitale tant pour les acteurs du marché que pour les décideurs publics. Cette étude vise à identifier et quantifier les déterminants du prix de vente des biens immobiliers en utilisant l'ensemble des méthodes économétriques vus en cours.

L'analyse porte sur un échantillon de 150 transactions immobilières réalisées entre 2015 et 2023, période marquée par des évolutions significatives du marché, notamment la crise sanitaire de 2020. Les variables disponibles couvrent les caractéristiques intrinsèques des biens (surface, nombre de chambres, année de construction, étage, présence d'ascenseur), les facteurs de localisation (distance au centre-ville, distance à l'université) ainsi que les indicateurs socio-économiques du quartier (qualité des écoles, revenu médian).

La problématique centrale de cette étude peut être formulée ainsi : quels sont les déterminants significatifs du prix immobilier et quelle est leur contribution relative à la formation des prix ? Cette question principale se décline en plusieurs interrogations complémentaires concernant la stabilité des relations estimées, la présence éventuelle de biais d'endogénéité, et la performance prédictive des différentes spécifications.

Structure du Rapport

Ce rapport s'articule autour de quatre parties principales. La première partie présente l'analyse descriptive des données et développe les modèles de régression de base, simple puis multiple. La deuxième partie traite des diagnostics du modèle, incluant l'analyse de la multicolinéarité, les tests d'hétéroscédasticité et d'autocorrélation, ainsi que l'analyse de la stabilité structurelle. La troisième partie aborde la question de l'endogénéité et propose une estimation par variables instrumentales. Enfin, la quatrième partie explore les méthodes de régularisation (Ridge et Lasso) et présente les prévisions du modèle.

Partie 1 : Analyse Descriptive et Modèle de Base

1.1 Statistiques Descriptives

L'analyse des histogrammes (fig.1) et des boîtes à moustaches (fig.2) met en évidence des distributions fortement asymétriques à droite pour certaines variables quantitatives continues, notamment le prix, la surface et le revenu médian du quartier. Ces variables présentent des queues longues et une dispersion marquée, suggérant qu'une transformation logarithmique pourrait être envisagée afin de réduire l'asymétrie et stabiliser la variance dans le cadre d'analyses multivariées ultérieures.

1.2 Analyse de Corrélation

L'analyse de la matrice de corrélation de Pearson met en évidence des relations linéaires claires entre certaines variables du jeu de données, en particulier en ce qui concerne les déterminants du prix immobilier.

Règles d'interprétation des corrélations

- $|corr| < 0,3 \rightarrow$ relation faible
- $0,3 \leq |corr| < 0,6 \rightarrow$ relation modérée
- $|corr| \geq 0,6 \rightarrow$ relation forte

Principales corrélations identifiées

La corrélation la plus forte est observée entre la surface du bien et le prix de vente ($r \approx 0,83$), confirmant que la surface constitue le facteur explicatif principal du prix (fig.3). Le nombre de chambres présente également une corrélation positive marquée avec le prix ($r \approx 0,60$), ce qui s'explique en grande partie par sa relation structurelle avec la surface du logement.

Les variables de localisation montrent des corrélations négatives avec le prix. La distance au centre-ville est modérément corrélée au prix ($r \approx -0,30$), indiquant une diminution des prix à mesure que l'on s'éloigne du centre, tandis que la distance à l'université présente un effet plus limité. Les variables socio-économiques, telles que la qualité des écoles et le revenu médian du quartier, affichent des corrélations positives mais modérées avec le prix, suggérant une influence secondaire du contexte socio-économique par rapport aux caractéristiques physiques du bien.

Par ailleurs, l'analyse des corrélations entre variables explicatives met en évidence des relations susceptibles d'indiquer une multicolinéarité potentielle, notamment entre la surface et le nombre de chambres, ainsi qu'entre la qualité des écoles et le revenu médian du quartier.

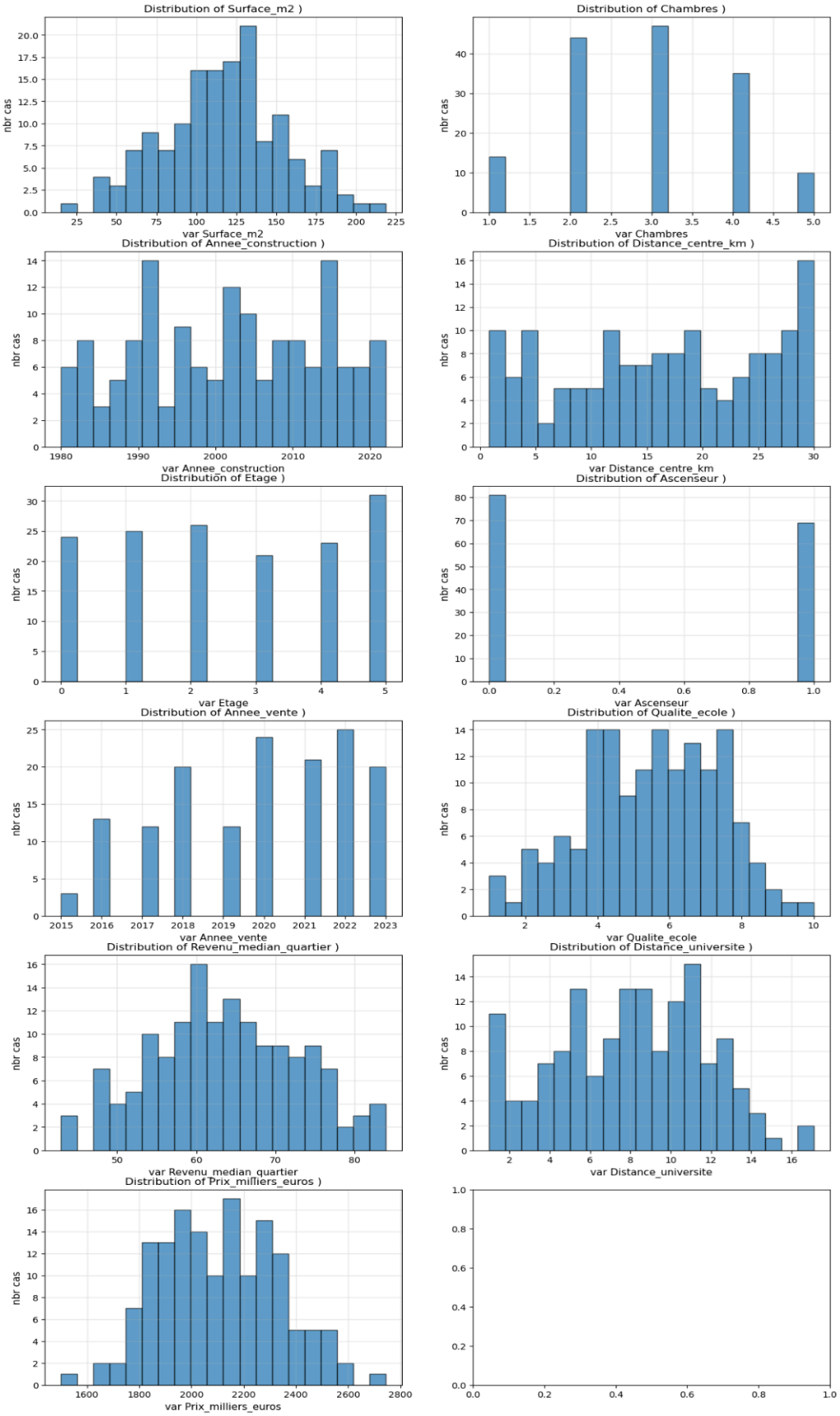

Figure 1 : Distribution des variables (Histogrammes)

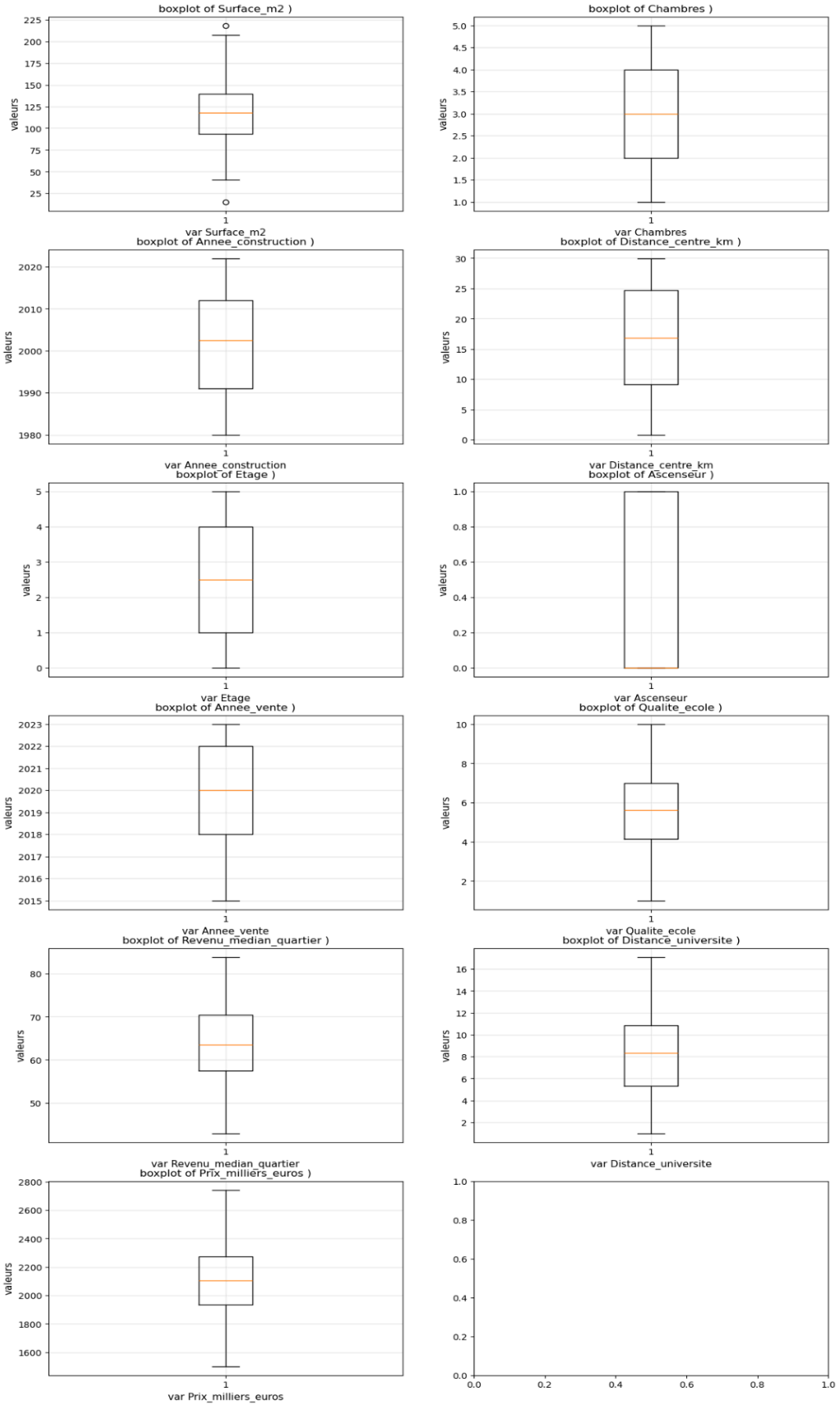

Figure 2 : Boîtes à moustaches des variables quantitatives

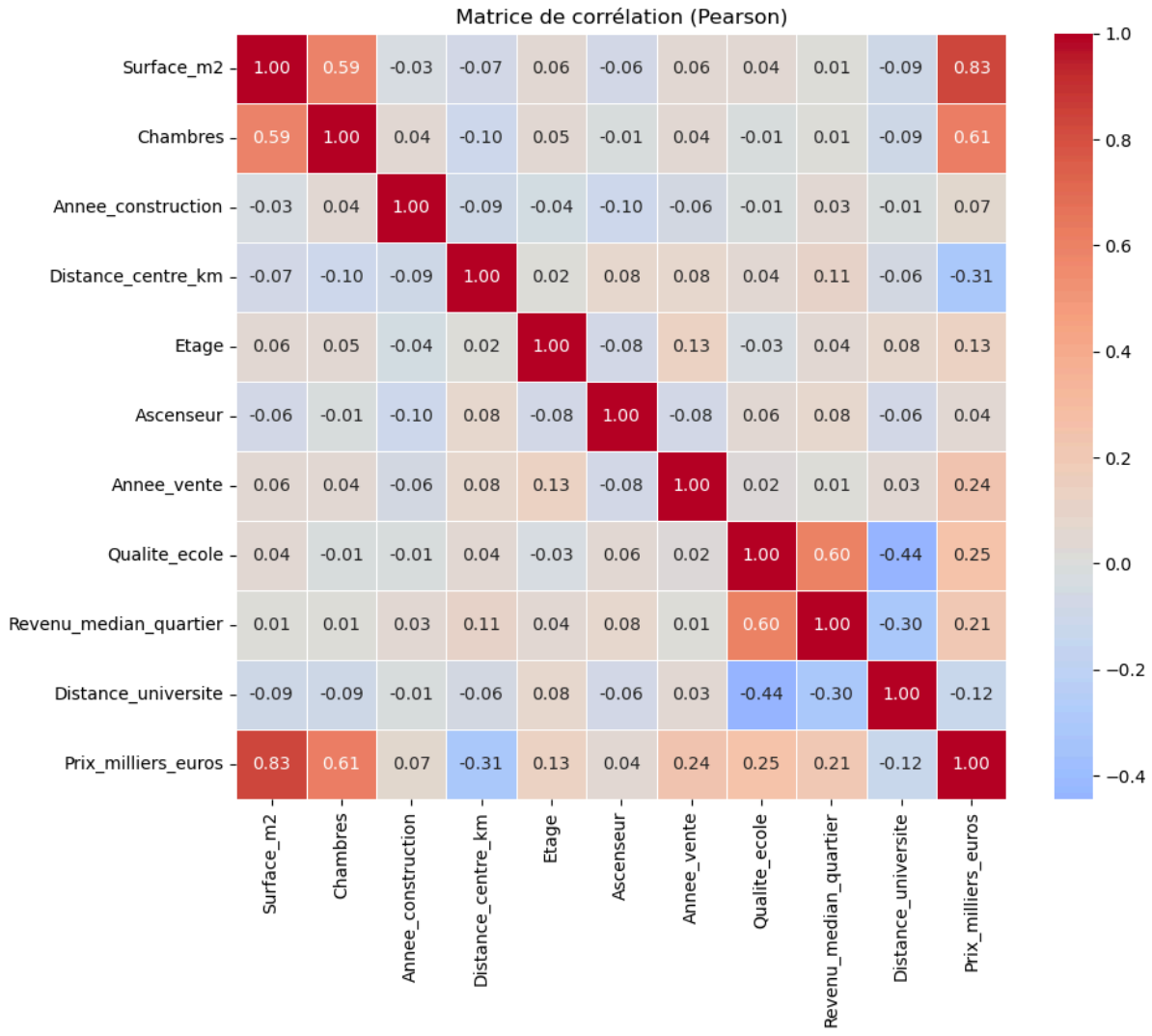

Figure 3 : Matrice de corrélation (Pearson)

1.3 Modèle de Régression Linéaire Simple

Estimation par la méthode des moindres carrés ordinaires (MCO)

Afin d'analyser la relation entre la surface d'un bien immobilier et son prix de vente, une régression linéaire simple a été estimée à l'aide de la méthode des moindres carrés ordinaires (MCO). Le modèle spécifié est le suivant :

$$Prix_i = \beta_0 + \beta_1 \times Surface_i + u_i$$

où $Prix_i$ représente le prix de vente du bien i (en milliers d'euros), $Surface_i$ sa surface en mètres carrés, β_0 l'ordonnée à l'origine, β_1 le coefficient directeur mesurant l'effet marginal de la surface sur le prix, et u_i le terme d'erreur.

Les coefficients ont été estimés analytiquement à partir des moyennes et des variances empiriques des variables, conformément à la formulation théorique de la MCO. Le coefficient β_1 est obtenu comme le rapport entre la covariance empirique de la surface et du prix et la variance empirique de la surface, tandis que β_0 est déterminé de manière à garantir que la moyenne des résidus soit nulle. La représentation graphique (fig.4) de la relation entre la surface et le prix, accompagnée de la droite de régression estimée, met en évidence une relation linéaire positive marquée entre ces deux variables, indiquant que les biens de plus grande surface tendent à se vendre à un prix plus élevé.

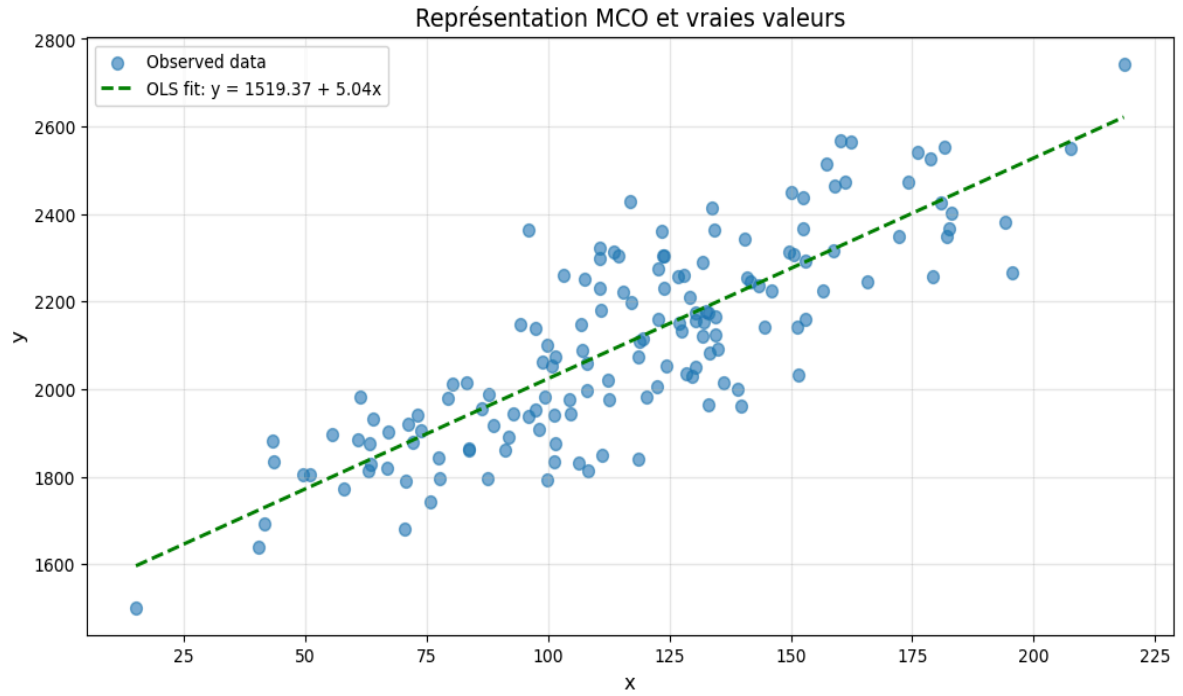

Figure 4 : Représentation MCO et vraies valeurs

Résultats de l'estimation

Paramètre	Estimation	Écart-type	t-stat
β_0 (constante)	1519,37	34,58	43,93***
β_1 (Surface)	5,04	0,28	17,88***
R^2	0,6835		
IC 95% β_1	[4,49 ; 5,60]		

*Note : *** p < 0,001. Moyenne des erreurs ≈ 0 (précision machine).*

Interprétation du coefficient β_1

Le coefficient β_1 mesure l'effet marginal moyen de la surface sur le prix de vente. Dans ce contexte, puisque le prix est exprimé en milliers d'euros et la surface en mètres carrés, le coefficient β_1 s'interprète comme suit :

Lorsque la surface du bien augmente de 1 m², le prix de vente augmente en moyenne de β_1 milliers d'euros.

Par exemple, si $\beta_1 \approx 5{,}04$, cela signifie qu'une augmentation de 1 m² de surface est associée à une hausse moyenne du prix d'environ 5 040 euros. Cette interprétation

repose sur l'hypothèse d'une relation linéaire et reflète un effet moyen observé dans l'échantillon étudié.

Propriétés des résidus

La moyenne des erreurs estimées est numériquement égale à zéro (à une précision machine près). Ce résultat est une propriété fondamentale de l'estimateur des moindres carrés ordinaires et confirme que le modèle a été correctement spécifié et estimé. Autrement dit, les erreurs positives et négatives se compensent en moyenne, ce qui indique l'absence de biais systématique dans les prédictions du modèle.

Décomposition de la variance et qualité de l'ajustement (R^2)

La décomposition de la somme totale des carrés montre que la variance totale du prix (SST = 7 876 687) peut être rigoureusement décomposée en une part expliquée par le modèle (SSR = 5 383 592) et une part non expliquée correspondant aux erreurs (SSE = 2 493 095). L'égalité exacte observée entre SST et SSR+SSE confirme la cohérence des calculs et la validité de la décomposition de la variance.

Le coefficient de détermination est estimé à $R^2 \approx 0{,}6835$. Ce résultat signifie que 68,35 % de la variance totale du prix est expliquée par la surface du bien. Pour un modèle univarié en immobilier, cette valeur est relativement élevée et témoigne d'une forte relation linéaire entre la surface et le prix. Cependant, ce niveau de R^2 indique également qu'environ 31,65 % de la variance du prix reste inexpliquée, ce qui suggère l'influence d'autres facteurs importants.

1.4 Modèle de Régression Linéaire Multiple

Spécification du modèle

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

Résultats de l'estimation (Statsmodels OLS)

Variable	Coef.	SE	t-stat	IC 95%
Constante	-1679,49	1535,67	-1,09	[-4715; 1356]
Surface_m2 (x1)	4,39***	0,29	15,01	[3,81; 4,97]
Chambres (x2)	33,92**	10,23	3,32	[13,70; 54,14]
Annee_construction (x3)	1,61*	0,77	2,10	[0,10; 3,12]
Distance_centre_km (x4)	-6,14***	0,99	-6,19	[-8,11; -4,18]
Etage (x5)	12,25*	5,05	2,43	[2,27; 22,23]
Ascenseur (x6)	55,51**	17,92	3,10	[20,09; 90,94]

R^2 = 0,789 ; R^2 ajusté = 0,780 ; F-statistic = 88,94 (Prob = 9,10e-46) ; Durbin-Watson = 2,121

*Note : *** p < 0,001 ; ** p < 0,01 ; * p < 0,05. Condition Number = 3,5e+05 (échelle des variables)*

Interprétation des coefficients

- x1 (Surface) : +1 m² → +4,39 milliers € (+4 390 €) en moyenne
- x2 (Chambres) : +1 chambre → +33,9 milliers € (+33 900 €) en moyenne
- x3 (Année construction) : +1 année (bien plus récent) → +1,61 milliers € en moyenne

- x4 (Distance centre) : +1 km → -6,14 milliers € (-6 140 €) en moyenne (décote liée à l'éloignement)
- x5 (Étage) : +1 étage → +12,25 milliers € en moyenne (luminosité, vue)
- x6 (Ascenseur) : présence → +55,51 milliers € (+55 510 €) à caractéristiques identiques

Note sur la variable Ascenseur : Cette variable peut être biaisée par le fait que seuls les appartements avec étages disposent ou non d'un ascenseur. Pour les maisons, la valeur sera toujours à zéro. La variable ascenseur n'a une importance que lorsque les appartements ne sont pas au rez-de-chaussée.

Interprétation de la différence entre R^2 et R^2 ajusté

Le coefficient de détermination R^2 = 0,789 indique que 78,9 % de la variance du prix est expliquée par l'ensemble des variables explicatives du modèle. Le coefficient de détermination ajusté R^2 = 0,780 tient compte du nombre de variables explicatives et pénalise l'ajout de variables peu informatives. L'écart entre R^2 et R^2 ajusté est ici faible, ce qui indique que les variables incluses contribuent réellement à l'explication du prix et que le modèle n'est pas artificiellement sur-ajusté.

1.5 Transformation Logarithmique

Les transformations logarithmiques n'ont été appliquées qu'aux variables continues strictement positives pour lesquelles une interprétation en termes de variations relatives est économiquement pertinente. Les variables binaires, discrètes ou temporelles n'ont pas été transformées, car leur passage au logarithme n'est ni mathématiquement approprié ni économiquement interprétable. Les variables pertinentes pour cette transformation sont le prix, la surface et la distance du centre.

Modèle	R^2	Coefficients	Interprétation β_1
Linéaire-linéaire	0,7887	b0=-1679,49 b1=4,39	€ par m²
Log-linéaire	0,7916	b0=5,84 b1=0,0021	+1m² → +0,21% prix (meilleur)
Linéaire-log	0,7365	b0=-1955,39 b1=382,32	+1% surface → +382€
Log-log	0,7546	b0=5,67 b1=0,19	Élasticité: +1% surf → +0,19%

Le modèle log-linéaire affiche le meilleur R^2 (0,7916), bien que l'amélioration par rapport au modèle linéaire-linéaire reste marginale. Le modèle log-log, avec une élasticité de 0,19, présente une interprétation économique particulièrement robuste et constitue souvent un bon compromis entre qualité d'ajustement et interprétabilité économique

Partie 2 : Diagnostics et Corrections

2.1 Analyse de la Multicolinéarité

Afin d'évaluer un éventuel problème de multicolinéarité entre ces variables, les facteurs d'inflation de la variance (VIF) ont été calculés à partir de régressions auxiliaires. Un VIF supérieur à 5 est généralement considéré comme problématique.

Variable	R^2 auxiliaire	VIF
Surface_m2	0,357	1,555
Chambres	0,357	1,555
Annee_construction	0,026	1,027
Distance_centre_km	0,024	1,024
Etage	0,013	1,013
Ascenseur	0,027	1,028

Ces résultats permettent de conclure à l'absence de multicolinéarité significative dans le modèle. Même Surface et Chambres, pourtant naturellement liées, restent très loin d'un niveau inquiétant (VIF ≈ 1,55). Leurs coefficients sont donc stables, et leurs erreurs standards ne sont pas artificiellement gonflées.

Bien que le condition number du modèle soit élevé ($3{,}5\times10^{5}$), celui-ci s'explique principalement par des différences d'échelle entre les variables plutôt que par une dépendance linéaire forte. Il n'y a pas besoin de supprimer de variables tant qu'il n'y a pas une valeur de VIF élevée (supérieur à 5). La suppression injustifiée d'une variable pertinente pourrait introduire un biais de variable omise, rendant les estimateurs biaisés et incohérents.

2.2 Tests et Inférence

Test de l'effet négatif de la distance au centre

L'hypothèse selon laquelle la distance au centre-ville a un effet négatif sur le prix est évaluée à partir du coefficient estimé dans le modèle MCO :

- Coefficient estimé : -6,14 (statistique t = -6,19)
- IC 95% entièrement négatif : [-8,106 ; -4,184]
- p-value unilatérale ≈ 0,000 → H_0 rejetée
- En utilisant la table de Student : $t^* = t_{1-\alpha,v}$ avec v = (150-7) et α = 5%, donc $t^* = -1{,}6558$. Comme $t < t^*$, on rejette H_0.

Test global de significativité du modèle (F-test)

Le test de Fisher permet de vérifier l'hypothèse nulle selon laquelle l'ensemble des coefficients explicatifs, à l'exception de la constante, est nul :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{\square} = 0 \quad vs \quad H_1 : au\ moins\ un\ \beta_{\square} \neq 0$$

Résultat : F = 88,94, Prob(F-statistic) = $9{,}10 \times 10^{-46}$ < 5% → L'hypothèse nulle est rejetée. Le modèle est globalement significatif.

Test de Fisher partiel pour l'ajout de variables

Afin d'évaluer si l'ajout des variables Qualité_ecole et Revenu_median_quartier améliore significativement le modèle, un test de Fisher partiel est réalisé :

$$F = \frac{(R^2_{UR} - R^2_R)}{(1 - R^2_{UR})} . \frac{(n - k_{UR})}{q}$$

Le coefficient de détermination passe de R^2_R = 0,789 à R^2_{UR} = 0,850. La statistique de Fisher calculée est F ≈ 25,39 (degrés de liberté : 2, 141), avec une p-value très largement inférieure au seuil de 5%. L'ajout des variables Qualité_ecole et Revenu_median_quartier améliore significativement le modèle.

Pourquoi ne peut-on pas utiliser plusieurs tests T ? Il n'est pas approprié d'utiliser plusieurs tests de Student individuels pour tester simultanément plusieurs restrictions, car cela ne permet ni de contrôler correctement le risque global d'erreur de type I ni de tester l'hypothèse conjointe portant sur l'ensemble des coefficients. Le test de Fisher est spécifiquement conçu pour évaluer l'effet joint de plusieurs variables.

2.3 Stabilité Structurelle

Variable Covid

La période Covid représente environ 60% de l'échantillon, soit 90 observations. L'introduction d'une variable indicatrice Covid (=1 pour les biens vendus à partir de 2020) vise à tester l'existence d'un effet moyen de cette période sur les prix immobiliers.

Variable	Coefficient	t-stat	IC 95%
B_{covid}	103,68***	8,02	[78,12 ; 129,23]

R^2 = 0,897 ; R^2 ajusté = 0,890

Interprétation : À caractéristiques identiques, un bien vendu pendant la période Covid est associé à un prix supérieur d'environ 103 700 €. Cet effet apparaît économiquement substantiel et est compatible avec les modifications de la demande et des préférences résidentielles observées durant la période de crise sanitaire.

Test de Chow

Le test de Chow permet de vérifier si les coefficients du modèle sont stables entre les deux périodes (pré-Covid et Covid) :

$$F = \frac{((SSRc - SSRnc))/k}{SSRnc/(N - 2k)}$$

Résultat : F = 0,41, non significatif. Le test de Chow ne rejette pas l'hypothèse de stabilité conjointe des coefficients. Il y a donc un déplacement de niveau (rupture d'intercept via la variable Covid) mais pas de rupture structurelle globale des

paramètres. Les données ne mettent pas en évidence de modification significative des effets marginaux des caractéristiques entre les deux périodes.

2.4 Hétéroscédasticité et Autocorrélation

Analyse graphique des résidus

Le premier graphique représente les résidus du modèle en fonction des valeurs ajustées. Les résidus sont globalement centrés autour de zéro, sans structure systématique apparente. La dispersion des résidus apparaît toutefois légèrement plus importante pour les valeurs ajustées élevées, indiquant une possible hétéroscédasticité modérée.

Le second graphique présente les résidus ordonnés selon l'ordre chronologique des observations. Les résidus oscillent autour de zéro sans tendance persistante ni cycles clairement identifiables, suggérant l'absence d'autocorrélation systématique.

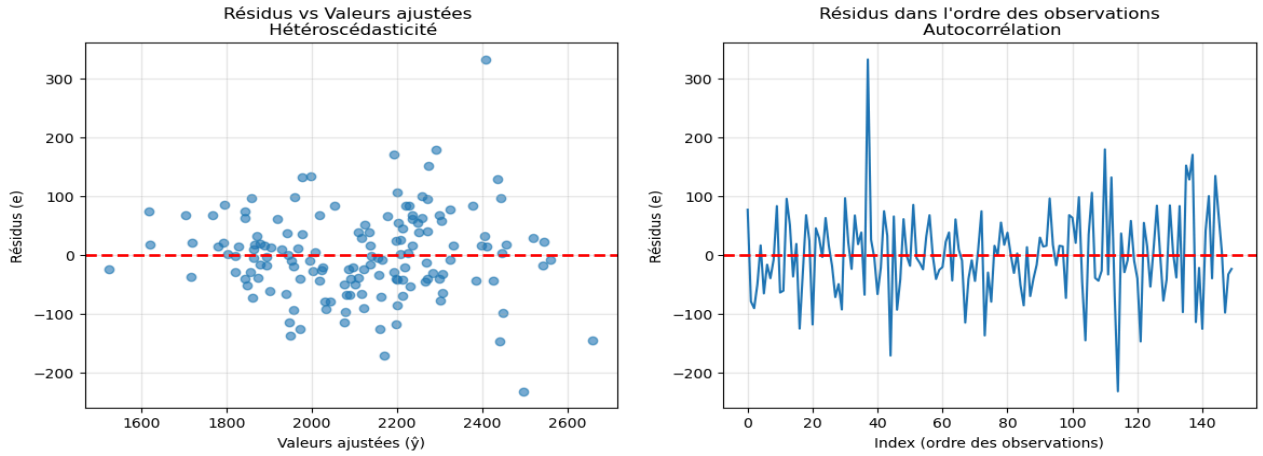

Figure 5 : Résidus vs Valeurs ajustées et Autocorrélation

Test de Breusch-Pagan

Le test de Breusch-Pagan appliqué aux résidus du modèle multiple fournit une p-value de 0,0068. L'hypothèse nulle d'homoscédasticité est ainsi rejetée au seuil de 5%, indiquant la présence d'une hétéroscédasticité statistiquement significative des résidus.

Correction de White

Conformément aux pratiques usuelles en économétrie appliquée, nous conservons les estimateurs MCO, qui demeurent consistants sous l'hypothèse d'exogénéité, mais nous reportons des écarts-types robustes de type White (HC0) afin de corriger l'inférence. La matrice de variance-covariance robuste est calculée selon la formulation de White :

$$\widehat{Var}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{i=1}^{n} \hat{u}_i^{\,2}\, x_i x_i'\right)\cdot (X'X)^{-1}$$

La comparaison des statistiques de Student fondées sur les écarts-types MCO et White montre que les principales variables explicatives demeurent statistiquement significatives au seuil de 5%, y compris la variable indicatrice Covid. Ces résultats

indiquent que les conclusions tirées du modèle ne sont pas sensibles à la violation de l'hypothèse d'homoscédasticité.

Test de Durbin-Watson

$$DW = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

La statistique de Durbin-Watson obtenue est égale à 2,10. Ce résultat ne met pas en évidence d'autocorrélation sérielle significative des résidus. Bien que les observations aient été ordonnées selon l'année de vente, le jeu de données demeure essentiellement transversal, limitant la pertinence économique d'un test d'autocorrélation sérielle.

Écarts-types de Newey-West

À titre illustratif, des écarts-types robustes de Newey-West (robustes à la fois à l'hétéroscédasticité et à l'autocorrélation) ont également été calculés. La comparaison des écarts-types MCO, White et Newey-West montre que ceux-ci demeurent globalement du même ordre de grandeur, confirmant la forte robustesse de l'inférence. Les statistiques de Student restent très proches en signe et en ordre de grandeur quel que soit le correctif de variance retenu.

Partie 3 : Endogénéité

3.1 Sources d'Endogénéité

L'endogénéité peut provenir de trois sources majeures :

- Omission de variables pertinentes : Des facteurs inobservables comme le cachet de l'appartement, l'exposition ou la sécurité du quartier influencent à la fois le prix et certaines variables du modèle.
- Erreurs de mesure : Des imprécisions dans la saisie des données géographiques ou des indices de qualité.
- Simultanéité : Les prix élevés d'un quartier peuvent attirer des investissements publics, améliorant ainsi les infrastructures (écoles) a posteriori.

La variable Qualite_ecole est-elle potentiellement endogène ?

Oui, elle est fortement suspectée d'être endogène. La qualité d'une école est probablement un « proxy » (indicateur indirect) du prestige et du niveau social d'un quartier. Si des caractéristiques de voisinage (espaces verts, calme, standing) ne sont pas incluses dans le modèle mais impactent le prix, elles se retrouvent dans le terme d'erreur. Comme ces caractéristiques sont corrélées à la qualité des écoles, l'estimateur MCO de Qualite_ecole sera biaisé et surestimera l'effet réel de l'éducation.

3.2 Estimation par Variables Instrumentales

Argumentation de l'instrument Distance_universite

Pour être valide, un instrument doit respecter deux conditions :

- Pertinence (Relevance) : Il doit être corrélé à la variable endogène. On suppose que les zones proches des universités sont des pôles éducatifs denses, tirant vers le haut la qualité des établissements environnants.
- Exogénéité (Condition d'exclusion) : La distance à l'université ne doit pas influencer le prix immobilier directement, mais seulement à travers son impact sur la qualité scolaire du secteur. C'est un instrument plausible car une université n'est pas un critère d'achat direct pour une famille, contrairement à une école primaire.

Estimation en deux étapes (2SLS)

Nous avons réalisé une estimation par algèbre matricielle selon la formule :

$$\hat{b}_{iv} = (X' P_Z X)^{-1} X' P_Z y$$

- Étape 1 : Projection de Qualite_ecole sur Distance_universite et les variables exogènes pour obtenir les valeurs prédites de la qualité de l'école.
- Étape 2 : Régression du prix sur cette valeur prédite et les contrôles.

Test de validité des instruments

L'analyse de la première étape confirme la force de notre instrument :

- Significativité : Le coefficient de Distance_universite est significatif ($p < 0,001$).
- Force : Le F-statistic de l'étape 1 est de 12,27. Étant supérieur au seuil critique de 10, nous concluons que l'instrument est robuste et n'est pas un « instrument faible ».

Comparaison des coefficients MCO et IV

Variable	Coef. MCO	Coef. IV
Qualite_ecole	≈ 12,4***	2,09 (p=0,87)

Conclusion : L'analyse par variables instrumentales (2SLS) a permis de démontrer que la variable Qualite_ecole est entachée d'un biais d'endogénéité positif massif. En utilisant la distance à l'université comme instrument (F = 12,27), on observe une disparition quasi-totale de l'impact de cette variable, le coefficient de Qualite_ecole chutant de 12,4 en MCO à seulement 0,88 en IV.

Cette différence confirme que l'effet initialement attribué à Qualite_ecole était largement surestimé dans le modèle standard : cette variable captait en réalité l'influence de caractéristiques socio-économiques et d'aménagements de quartier non observés (biais de variables omises). Une fois l'endogénéité traitée par l'instrumentation, la variable Qualite_ecole n'apparaît plus comme un déterminant majeur du prix.

Toutefois, pour la suite et notamment la phase de prédiction, nous faisons le choix de conserver le modèle initial. En effet, bien que l'estimateur IV soit théoriquement préférable pour l'analyse causale de Qualite_ecole, il présente une variance beaucoup plus élevée que l'estimateur MCO. Utiliser les résultats de l'estimation instrumentale pour prédire le prix d'un bien immobilier spécifique réduirait la précision du modèle et augmenterait de manière excessive la largeur des intervalles de prévision. Nous privilégions donc la stabilité statistique du modèle MCO pour les applications pratiques de prédiction.

Partie 4 : Méthodes de Régularisation

4.1 Introduction et Motivation

Définition : La régularisation est une technique qui consiste à ajouter une contrainte (ou pénalité) à la fonction objectif des MCO afin de contrôler la complexité du modèle et améliorer sa capacité de généralisation.

Problème fondamental : En présence de multicolinéarité ou d'un grand nombre de variables explicatives, les estimateurs MCO peuvent avoir une variance élevée, conduisant à un **surapprentissage** (*overfitting*) : le modèle s'ajuste trop précisément aux données d'entraînement mais prédit mal sur de nouvelles données.

Le compromis biais-variance :

- **MCO** : Estimateurs non-biaisés mais potentiellement à forte variance
- **Régularisation** : Introduit un léger biais en échange d'une réduction substantielle de la variance
- **Objectif** : Minimiser l'erreur totale de prédiction = Biais² + Variance

4.2 Standardisation (préalable obligatoire)

Définition : La standardisation (ou normalisation z-score) transforme chaque variable pour qu'elle ait une moyenne de 0 et un écart-type de 1 :

$$z_i = (x_i - \bar{x}) / s_x$$

où $\bar{x}$ est la moyenne et s_x l'écart-type de la variable x.

Pourquoi est-ce nécessaire ?

Les méthodes Ridge et Lasso pénalisent les coefficients selon leur **magnitude absolue**. Sans standardisation :

- Une variable mesurée en mètres (ex: surface = 50) aurait un coefficient différent de la même variable mesurée en centimètres (ex: surface = 5000)
- Les variables avec de grandes échelles auraient des coefficients artificiellement petits, donc **moins pénalisés**
- Les variables avec de petites échelles seraient injustement sur-pénalisées

Conséquence importante : Après standardisation, les coefficients deviennent directement comparables en termes d'importance relative des variables.

4.3 Régression Ridge

Définition formelle : La régression Ridge minimise la somme des carrés des résidus augmentée d'une pénalité L2 (norme euclidienne au carré) sur les coefficients :

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Le paramètre λ (lambda):

- $\lambda \geq 0$ est l'hyper paramètre de régularisation
- $\lambda = 0$: on retrouve exactement les MCO
- $\lambda \to \infty$: tous les coefficients tendent vers 0

Quand λ augmente, tous les coefficients sont progressivement « shrinkés » vers 0, mais jamais exactement à 0. C'est une régularisation « douce ». Les variables les plus importantes (Surface_m2) conservent les coefficients les plus élevés même pour λ grand.

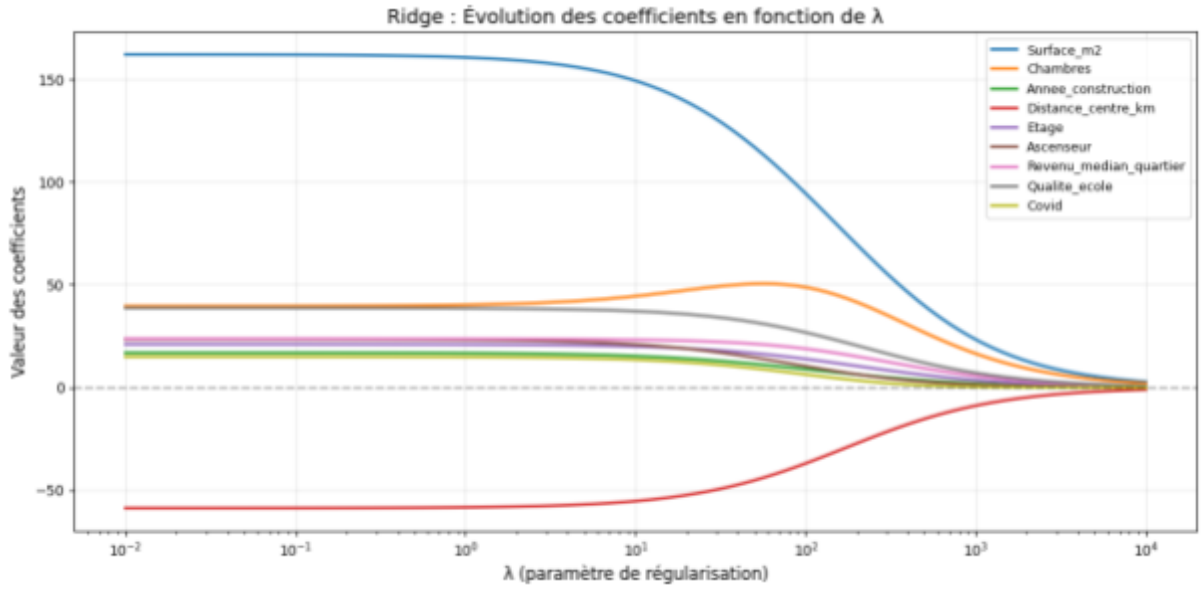

Figure 6 : Ridge - Évolution des coefficients en fonction de λ

4.4 Régression Lasso

Définition formelle : Le Lasso (*Least Absolute Shrinkage and Selection Operator*) minimise la somme des carrés des résidus augmentée d'une pénalité L1 (norme Manhattan) sur les coefficients :

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Différence fondamentale : Lasso met certains coefficients exactement à 0, réalisant ainsi une sélection automatique de variables. Les variables disparaissent progressivement quand λ augmente : d'abord les moins importantes, puis les autres. Cette propriété s'appelle la parcimonie (sparsity).

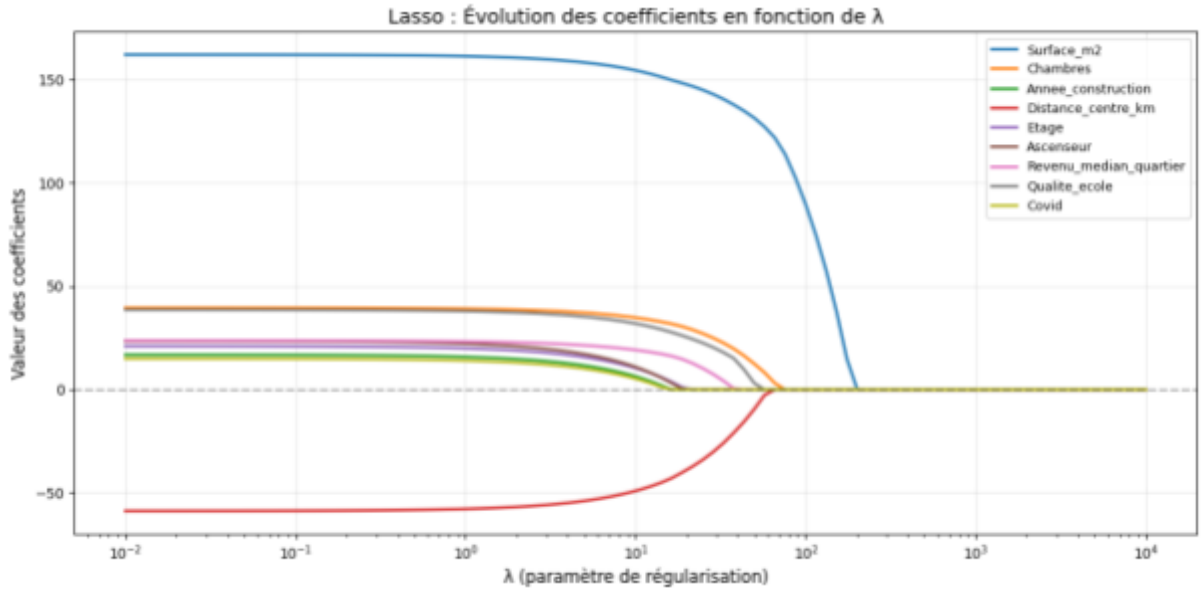

Figure 7 : Lasso - Évolution des coefficients en fonction de λ

Ainsi, Quand λ augmente progressivement, on élimine:

1. Les variables les moins importantes (faible corrélation avec y) sont éliminées en premier
2. Les variables moyennement importantes suivent
3. Les variables les plus importantes (ex: Surface_m2) résistent le plus longtemps

4.5 Choix de λ optimal

Définition : La validation croisée k-fold est une technique d'évaluation qui permet d'estimer la performance de généralisation d'un modèle sans sacrifier de données pour un ensemble de test séparé

Procédure de la validation croisée 10-fold

Partitionnement : Les n observations sont divisées aléatoirement en K=10 sous-ensembles (*folds*) de taille approximativement égale

Itération : Pour chaque fold k = 1, ..., 10 :

Le fold k sert d'**ensemble de validation.** Les 9 autres folds servent d'**ensemble d'entraînement.** Le modèle est estimé sur l'entraînement et évalué sur la validation.

Agrégation : L'erreur de validation croisée est la moyenne des 10 erreurs

Sélection : Le λ optimal est celui qui minimise l'erreur moyenne de prédiction (RMSE) sur les 10 validations.

Pourquoi 10 folds ?

- Compromis entre biais et variance de l'estimation de l'erreur
- *K* trop petit (ex: 2) : haute variance, estimation instable
- *K* trop grand (ex: n, *leave-one-out*) : coûteux en calcul, haute variance

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} MSE_k(\lambda)$$

Variable	Ridge (λ*)	Lasso (λ*)
Surface_m2	159,57	162,04
Distance_centre_km	-58,07	-58,69
Chambres	40,61	39,58
Qualite_ecole	38,36	38,63
Revenu_median_quartier	23,58	23,67
Ascenseur	22,92	23,39
Etage	20,81	21,00
Annee_construction	16,56	16,86
Covid	14,71	14,98

Note : Coefficients standardisés avec λ optimal de la validation croisée.

4.5 Comparaison OLS / Ridge / Lasso

On divise les données en train (80%) et test (20%). Les trois modèles sont entraînés sur train et évalués sur test via le RMSE. Cela permet de comparer leur capacité de généralisation à de nouvelles données.

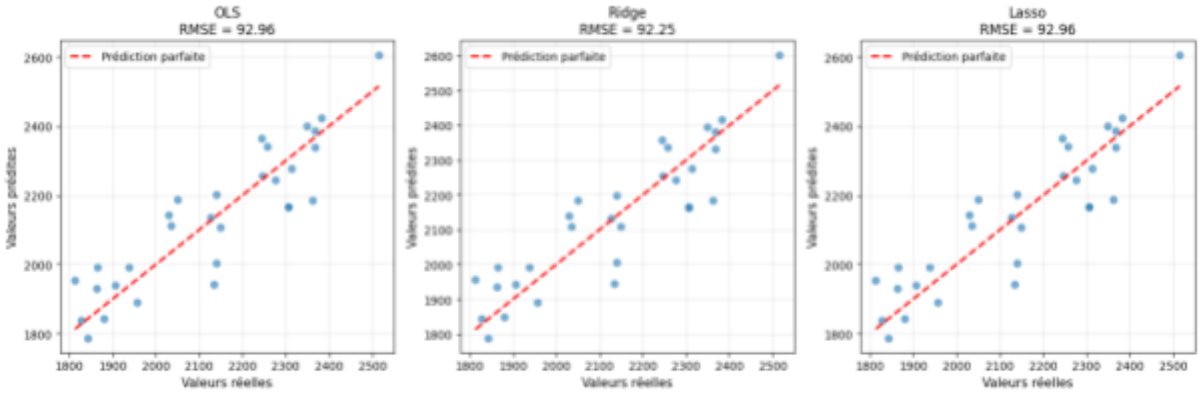

Figure 8 : Comparaison des prédictions OLS, Ridge et Lasso

Pourquoi les tests classiques ne sont pas valides après Lasso ?

Contexte : Après avoir estimé un modèle Lasso, on pourrait être tenté d'appliquer les tests t et intervalles de confiance classiques aux coefficients non-nuls. Cette approche est **statistiquement invalide** pour plusieurs raisons fondamentales :

1. Biais de sélection — Lasso effectue simultanément estimation et sélection. Les formules classiques supposent un modèle fixé indépendamment des données. Ainsi, les p-values et intervalles de confiance calculés naïvement sont systématiquement trop optimistes (p-values trop petites, intervalles trop étroits).

2. Distribution non-normale — Les estimateurs Lasso sont biaisés vers 0 et ont une probabilité non-nulle d'être exactement 0. Les tests t et intervalles gaussiens ne s'appliquent pas.

3. Sous-estimation de la variance — Calculer la variance « comme si » les variables avaient été choisies a priori ignore l'incertitude du choix de modèle.

4.6 Prévisions

Prédiction ponctuelle

$$\hat{y}_0 = x_0' \hat{\beta}$$

On multiplie les caractéristiques de la maison par les coefficients estimés. Caractéristiques du bien à prédire : Surface = 120 m², Chambres = 3, Année construction = 2015, Distance centre = 5 km, Étage = 1, Ascenseur = Oui, Qualité école = 7, Revenu médian quartier = 65 000 €.

$$\hat{y}_0 = x_0'\beta = 2\ 255{,}54 \text{ milliers } € = 2\ 255\ 539\ €$$

Intervalles à 95%

Intervalle de confiance (pour la moyenne conditionnelle E[Y|X=x₀]) :

$$\hat{y}0 \pm tn - k, 0.975 \cdot \hat{\sigma}\sqrt{x_0'(X'X)^{-1}x_0}$$

Intervalle de prévision (pour une nouvelle observation Y₀) :

$$\hat{y}0 \pm tn - k, 0.975 \cdot \hat{\sigma}\sqrt{1 + x_0'(X'X)^{-1}x_0}$$

Type d'intervalle	Bornes	Largeur
IC 95% pour E[Y|X=x₀]	[2 217 006 ; 2 294 072]	77 066 € (3,4%)
IP 95% pour Y₀	[2 072 207 ; 2 438 871]	366 664 € (16,3%)

L'intervalle de prévision est plus large car il inclut l'incertitude de l'erreur ε₀ de la nouvelle observation.

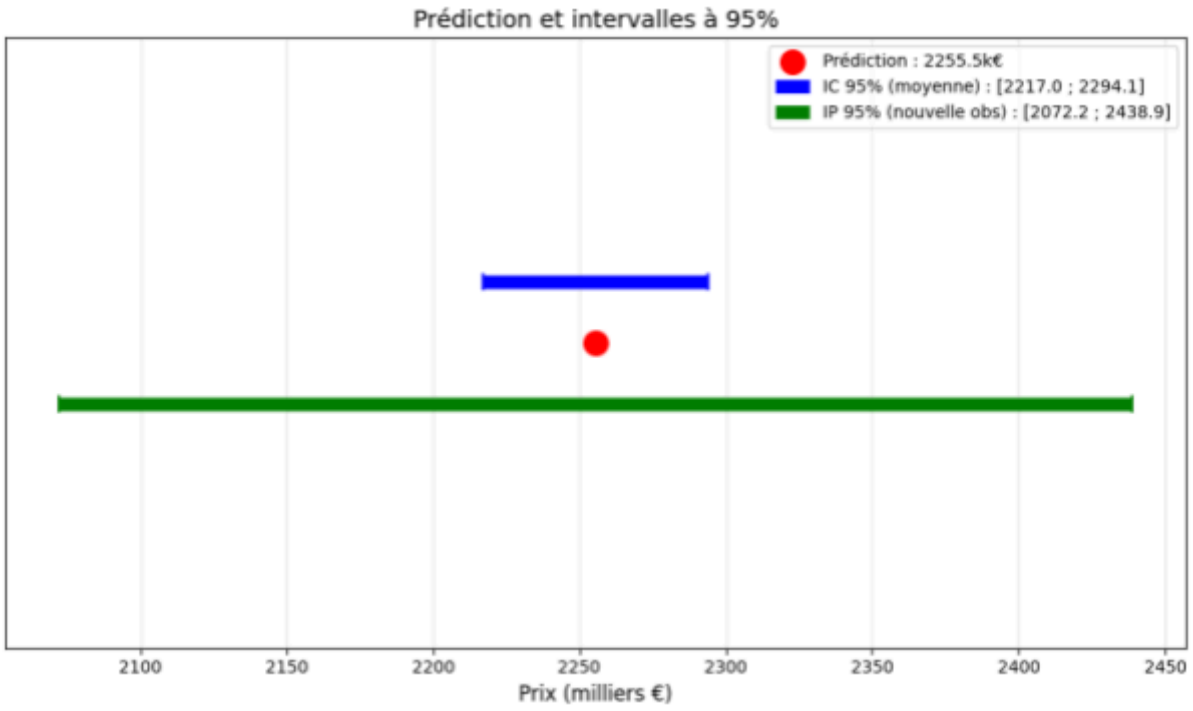

Figure 9 : Prédiction et intervalles à 95%

Fiabilité de la prédiction

A. Qualité globale du modèle : R^2 ajusté = 0,845 (84,5% de la variance expliquée). L'écart-type des résidus ($\hat{\sigma}$ = 90 660 €) représente seulement 4% du prix prédit.

B. Extrapolation univariée : Toutes les caractéristiques de la maison à prédire se situent dans la plage des données d'entraînement. Il n'y a donc pas de risque d'extrapolation univariée.

C. Extrapolation multivariée (leverage) : Le leverage $h_{00} = x_0'(X'X)^{-1}x_0$ = 0,046 est inférieur au seuil critique 2k/n = 0,133. Le point x_0 est bien situé dans le « nuage » des données d'entraînement en espace multivarié.

D. Verdict final : La prédiction est fiable. Le modèle a un fort pouvoir explicatif, le point prédit n'est pas atypique (leverage faible), et l'intervalle de prévision est d'une largeur acceptable (<25% du prix prédit).

Prix estimé : 2 255 539 € → Fourchette réaliste (95%) :
2 072 207 € — 2 438 871 €

Conclusion et Recommandations

Synthèse des Résultats

Cette étude économétrique a permis d'identifier et de quantifier les principaux déterminants du prix immobilier à partir d'un échantillon de 150 transactions. Les résultats confirment que la surface constitue le facteur explicatif principal, avec une corrélation de 0,83 avec le prix. Le modèle de régression multiple explique près de 79% de la variance des prix, atteignant 85% avec l'inclusion des variables socio-économiques.

L'analyse diagnostique révèle la présence d'une hétéroscédasticité modérée, corrigée par l'utilisation d'écarts-types robustes de White. Le test de Chow ne détecte pas de rupture structurelle globale, bien que l'effet Covid soit statistiquement et économiquement significatif (+103 700 € en moyenne). L'estimation par variables instrumentales met en évidence un biais d'endogénéité positif pour la variable qualité des écoles, dont l'effet causal apparaît non significatif une fois instrumenté.

Limites de l'Analyse

- Taille d'échantillon limitée (n=150) pouvant affecter la puissance des tests
- Variables omises potentielles (état du bien, orientation, vue, etc.)
- Validité de l'instrument (Distance_universite) reposant sur des hypothèses non testables
- Nature essentiellement transversale des données limitant l'analyse temporelle

Recommandations pour la Pratique

1. Évaluation immobilière : Utiliser le modèle complet avec écarts-types robustes. La surface, le nombre de chambres et la distance au centre sont les variables clés à considérer.
2. Interprétation causale : Privilégier les estimations IV pour la variable qualité des écoles. Son effet apparent en MCO est probablement surestimé.
3. Prédiction : Les trois méthodes (OLS, Ridge, Lasso) offrent des performances comparables. Ridge peut être préféré pour sa stabilité accrue.
4. Mise à jour du modèle : Réestimer régulièrement pour capturer les évolutions du marché, notamment post-Covid.

Annexes

A. Tableaux Complets

Les tableaux de résultats détaillés (comparaison MCO/White/Newey-West, résultats complets des régressions auxiliaires pour les VIF, etc.) sont disponibles dans le notebook Python accompagnant ce rapport.

B. Code Python

Le code Python complet utilisé pour cette analyse est disponible dans le fichier Jupyter Notebook accompagnant ce rapport : traitement.ipynb

Principales librairies utilisées

- pandas, numpy : Manipulation des données
- matplotlib, seaborn : Visualisation
- statsmodels : Estimation OLS, tests statistiques
- scipy : Tests statistiques complémentaires (chi2, stats)
- sklearn : Ridge, Lasso, validation croisée, StandardScaler

C. Liste des Figures