UNIVERSITÉ PARIS 1

PANTHÉON SORBONNE

# Statistics Project

Analysis of Housing Prices

Lina Ibouchichene
Alexandre Mégard

December 2025

# Contents

# 1 Résumé exécutif

## Principaux résultats

Ce projet d'économétrie appliquée analyse les déterminants du prix de vente de 150 maisons vendues entre 2015 et 2023, à partir de modèles de régression linéaire et de méthodes économétriques avancées, incluant les diagnostics, les variables instrumentales et les techniques de régularisation.

L'analyse descriptive met en évidence une forte hétérogénéité des prix immobiliers, avec une distribution globalement proche de la symétrie. La surface habitable apparaît dès l'analyse préliminaire comme la variable la plus fortement corrélée au prix. Les corrélations observées entre les variables explicatives restent modérées, suggérant l'absence de multicolinéarité sévère.

Les estimations confirment le rôle central des caractéristiques physiques : la surface, le nombre de chambres, l'étage et la présence d'un ascenseur ont un effet positif et significatif sur le prix, tandis que l'éloignement du centre-ville exerce un effet négatif marqué. Le modèle semi-logarithmique a été retenu pour sa qualité d'ajustement ($R^2$ ajusté $\approx 0,78$) et sa capacité à stabiliser la variance des erreurs.

Le test de stabilité structurelle révèle une rupture significative liée au COVID-19, indiquant que la dynamique de formation des prix a changé après 2020. Par ailleurs, l'approche par variables instrumentales montre que l'effet de la  qualité des écoles , initialement jugé fort, devient non significatif une fois les biais d'endogénéité corrigés. Enfin, les méthodes de régularisation (Ridge/Lasso) valident la robustesse du modèle de base sans apporter de gain prédictif majeur sur cet échantillon.

## Recommandations

Au regard de ces conclusions, plusieurs recommandations stratégiques et méthodologiques sont formulées :

**Réévaluer l'impact de la carte scolaire :** L'analyse démontre que la qualité des écoles n'influence pas directement le prix immobilier lorsqu'on isole la richesse du quartier. Les acheteurs et investisseurs ne devraient pas payer de surprime pour ce seul critère, qui semble agir comme un proxy du niveau socio-économique plutôt que comme un déterminant structurel.

**Capitaliser sur les équipements structurels :** La présence d'un ascenseur génère une plus-value significative, confirmée par tous les modèles. C'est un levier de valorisation sûr, contrairement à des critères plus volatils.

**Adapter les modèles à la rupture post-COVID :** La relation entre les caractéristiques des biens et leur prix s'est modifiée depuis 2020. Il est impératif de privilégier les données récentes pour les estimations actuelles, les modèles basés sur l'historique pré-pandémie risquant de biaiser les évaluations.

**Prudence sur l'interprétation individuelle :** Bien que le modèle prédise le prix moyen avec une grande précision ($\pm 1,6\%$), l'intervalle de prédiction pour une transaction unique reste large. Une marge de négociation substantielle doit toujours être intégrée pour tenir compte des spécificités non observables de chaque bien.

# 2 Introduction

L'immobilier représente un secteur clé de l'économie, influençant à la fois les ménages, les investisseurs, et impacte grandement la dynamique économique locale et nationale. L'un des principaux défis du marché immobilier réside justement dans la détermination des facteurs influençant les prix des logements. Plusieurs éléments rentrent en jeu, tels que la taille des propriétés, leur localisation géographique, leur qualité, leur environnement, etc. Des facteurs qui peuvent avoir un impact direct sur la valeur des biens immobiliers.

Dans ce cadre, ce projet d'économétrie appliquée vise à analyser les facteurs qui déterminent le prix de vente des maisons dans une région donnée, en utilisant un ensemble de données réelles de 150 maisons vendues entre 2015 et 2023. En utilisant des outils économétriques, ce rapport propose d'identifier les relations entre les caractéristiques des maisons et leur prix afin de mieux comprendre ce qui influence la valeur des biens immobiliers et de proposer des recommandations pour les acheteurs, les vendeurs, et les investisseurs.

L'objectif principal de cette étude est de comprendre comment des variables telles que la surface habitable, l'année de construction, la distance au centre-ville, et d'autres variables socio-économiques influencent les prix des maisons. Pour ce faire, plusieurs modèles économétriques seront utilisés, allant des modèles de régression linéaire simple et multiple, à l'application de méthodes de régularisa- tion comme le Ridge et le Lasso, en passant par l'analyse de la multicolinéarité et la gestion des biais d'endogénéité.

## Structure du Rapport

Le rapport est structuré de la manière suivante :

- **Résumé exécutif** : Présentation des prinicpaux résultats et des recommandations
- **Introduction** : Présentation du contexte et de la problématique, suivi de la structure du rapport.
- **Partie 1 : Analyse descriptive et modèle de base** : Cette section inclut les statistiques descriptives des variables étudiées, ainsi que les résultats des régressions linéaires simples et multiples. Nous analyserons également les tests de significativité des variables dans le modèle.
- **Partie 2 : Diagnostics et corrections** : Nous y aborderons les tests de multicolinéarité, les tests d'hétéroscédasticité et les corrections nécessaires, ainsi que les tests de Chow pour la stabilité structurelle.
- **Partie 3 : Endogénéité** : Nous discuterons des sources possibles d'endogénéité dans notre modèle, en particulier pour la variable Qualite_ecole, et nous effectuerons une estimation par variables instrumentales (IV).
- **Partie 4 : Méthodes de régularisation** : Cette section sera dédiée à l'utilisation des modèles Ridge et Lasso, incluant la sélection du paramètre optimal de régularisation et la comparaison des performances prédictives.
- **Conclusion et recommandations** : Nous récapitulerons les principaux résultats obtenus et proposerons des recommandations pratiques en fonction des résultats obtenus.
- **Annexes** : Incluront les tableaux complets des résultats, le code Python utilisé, ainsi que les graphiques supplémentaires.

Cette structure a pour but d'offrir une analyse complète et rigoureuse des facteurs influençant les prix immobiliers tout en appliquant des outils économétriques appropriés.

# 3 Partie 1 : Analyse descriptive et modèle de base

## 3.1 Statistiques descriptives

| Variables | Moyenne | Médiane | Écart-type | Minimum | Q1 | Q3 | Maximum |
|---|---|---|---|---|---|---|---|
| Surface ($m^2$) | 116,71 | 117,85 | 37,69 | 15,21 | 93,24 | 139,64 | 218,53 |
| Chambres | 2,89 | 3,00 | 1,08 | 1 | 2 | 4 | 5 |
| Année de construction | 2001,83 | 2002,50 | 11,70 | 1980 | 1991 | 2012 | 2022 |
| Distance centre (km) | 16,50 | 16,87 | 9,02 | 0,83 | 9,11 | 24,70 | 29,99 |
| Étage | 2,58 | 2,50 | 1,76 | 0 | 1 | 4 | 5 |
| Ascenseur (0/1) | 0,46 | 0 | 0,50 | 0 | 0 | 1 | 1 |
| Année de vente | 2019,84 | 2020 | 2,29 | 2015 | 2018 | 2022 | 2023 |
| Qualité école | 5,47 | 5,60 | 1,87 | 1 | 4,13 | 7,00 | 10 |
| Revenu médian quartier (k€) | 63,67 | 63,45 | 9,30 | 42,90 | 57,50 | 70,48 | 83,90 |
| Prix (k€) | 2107,90 | 2105,05 | 229,92 | 1500,77 | 1934,29 | 2272,78 | 2743,04 |

Table 1: Statistiques descriptives des variables

| Indicateur | Valeur |
|---|---|
| Asymétrie (Skewness) | 0,15 |
| Aplatissement (Kurtosis) | -0,49 |

Table 2: Asymétrie et Applatissement pour le prix

Après avoir calculé les statistiques descriptives de chaque variable (cf Table 1 ci-dessus), nous allons analyser maintenant les variables afin de voir si la potentielle utilisation d'une transformation logarithmique permettrait de stabiliser la variance de ces variables, de rendre leur distribution plus régulière, et d'améliorer l'interprétation économique de leurs résultats.

- Le prix

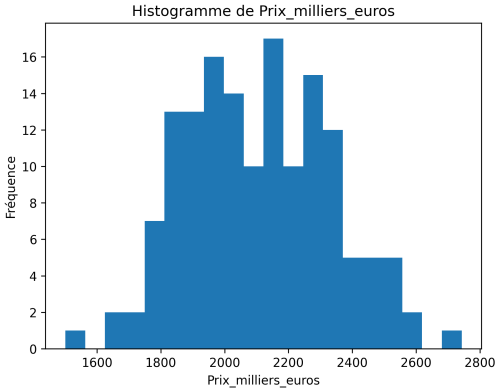


Figure 1: Histogramme du prix

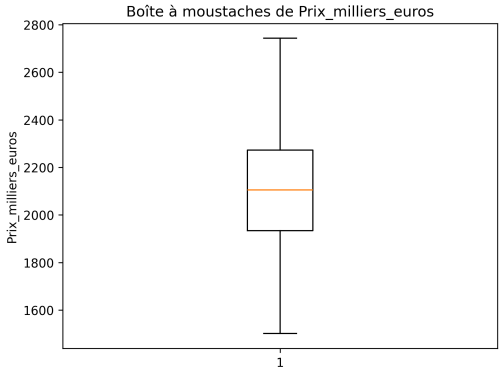


Figure 2: Boxplot du prix

Les histogrammes permettent de visualiser la distribution des principales variables continues. Pour le prix, la distribution apparaît relativement étalée et présente une légère asymétrie. Le coefficient d'asymétrie du prix (0,15) indique une asymétrie à droite de faible ampleur, tandis que le coefficient d'aplatissement (0,49) suggère une distribution légèrement plus plate que la loi normale. Ces éléments montrent que la distribution du prix reste globalement proche de la symétrie, tout en s'écartant modérément de la normalité. Cette analyse graphique et

descriptive suggère qu'une transformation logarithmique du prix peut être envisagée afin de stabiliser la variance et d'améliorer les propriétés statistiques du modèle, sans que celle-ci soit strictement indispensable.

- La surface

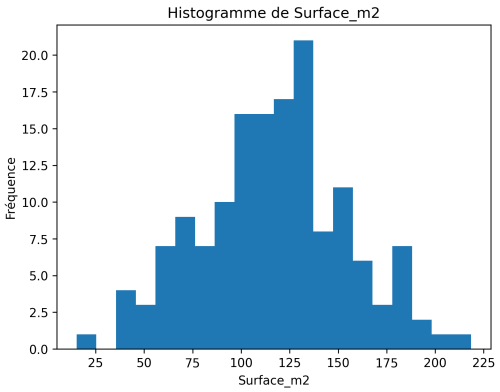


Figure 3: Histogramme de la surface

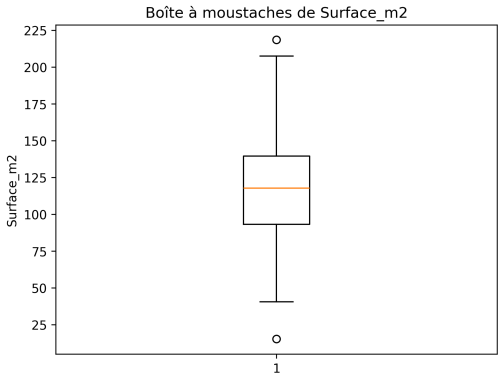


Figure 4: Boxplot de la surface

L'histogramme et la boîte à moustaches de la surface habitable mettent en évidence une dispersion modérée ainsi qu'une asymétrie à droite, avec la présence de quelques valeurs extrêmes, ce qui suggère qu'une transformation logarithmique de la surface pourrait être envisagée.

- La distance au centre ville

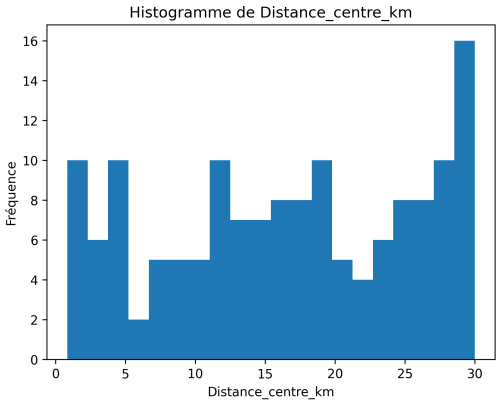


Figure 5: Histogramme de la distance au centre-ville

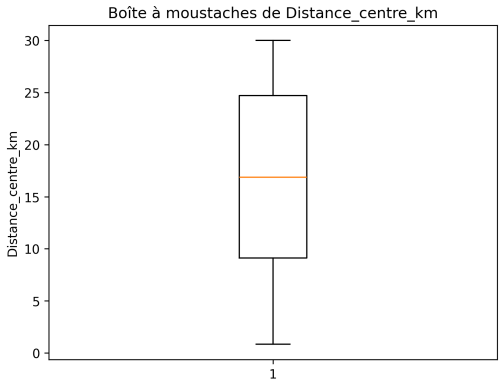


Figure 6: Boxplot de la distance au centre-ville

L'histogramme et la boîte à moustaches de la distance au centre-ville mettent en évidence une distribution relativement étalée, avec la présence de valeurs élevées. Cette structure suggère qu'une transformation logarithmique de cette variable pourrait être envisagée afin de mieux prendre en compte l'étendue des distances.

- Le Revenu médian du quartier

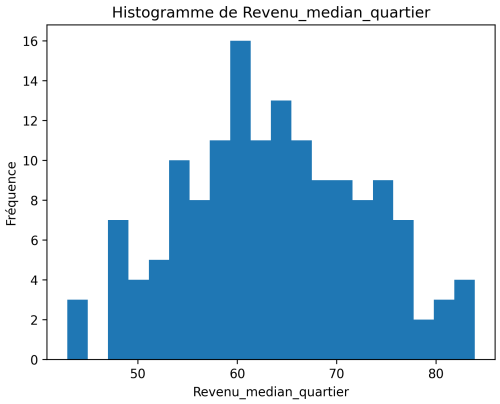


Figure 7: Histogramme du revenu médian du quartier

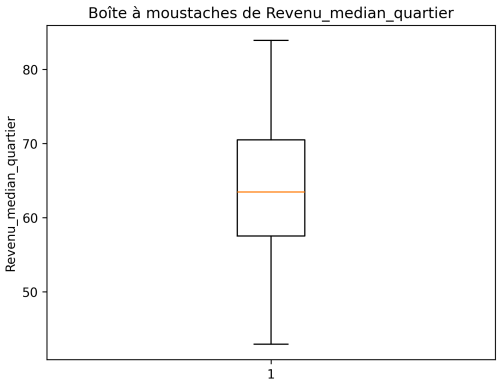


Figure 8: Boxplot du revenu médian du quartier

L'histogramme et la boîte à moustaches du revenu médian du quartier montrent une distribution relativement symétrique et modérément dispersée, ce qui ne rend pas indispensable l'utilisation d'une transformation logarithmique.

Afin de ne pas alourdir le corps du rapport, l'analyse graphique détaillée des distributions des autres variables explicatives (histogrammes et boîtes à moustaches) est présentée en annexe. Le texte principal se concentre sur les variables pour lesquelles ces graphiques apportent des enseignements particulièrement pertinents en vue de la suite du projet.

- L'année de construction

L'histogramme de la variable *Annee_construction* met en évidence une répartition relativement étalée des logements construits entre le début des années 1980 et le début des années 2020. La distribution ne présente pas d'asymétrie marquée, avec une concentration légèrement plus importante autour du début des années 2000, suggérant que la majorité des biens de l'échantillon a été construite au cours des deux dernières décennies.

La boîte à moustaches confirme cette observation, la médiane se situe autour du début des années 2000 et l'intervalle interquartile est relativement large, traduisant une hétérogénéité modérée de l'âge des logements. Aucune valeur aberrante extrême n'est détectée, ce qui indique une distribution globalement régulière.

- L'année de vente

Compte tenu de la forme relativement symétrique de la distribution et de l'absence de valeurs extrêmes prononcées, la variable *Annee_construction* ne semble pas nécessiter de transformation logarithmique.

L'histogramme de la variable *Annee_vente* montre une concentration des observations sur la période récente, comprise entre 2018 et 2023. La distribution ne présente pas d'asymétrie marquée, les ventes étant relativement bien réparties sur les différentes années de la période considérée.

La boîte à moustaches confirme cela, la médiane se situe autour de l'année 2020 et l'intervalle interquartile est relativement resserré, indiquant une dispersion limitée des années de vente. Aucune valeur aberrante n'est observée, ce qui suggère une distribution homogène de cette

variable dans l'échantillon.

- Le nombre de chambres

L'histogramme de la variable *Chambres* met en évidence une concentration des observations autour de 2 à 4 chambres. La plus fréquente correspond à trois chambres, ce qui reflète une structure relativement homogène des biens immobiliers de l'échantillon.

La boîte à moustaches confirme cette concentration, la médiane est égale à trois chambres et l'intervalle interquartile est relativement resserré, indiquant une dispersion limitée de cette variable. Les valeurs extrêmes, comprises entre une et cinq chambres, restent peu nombreuses.

Compte tenu de la nature discrète de la variable *Chambres* et de sa faible dispersion relative, une transformation logarithmique n'est ni nécessaire ni pertinente.

- L'Etage

L'histogramme de la variable *Etage* met en évidence une distribution discrète des logements sur des niveaux compris entre le rez-de-chaussée (0) et le cinquième étage. Les observations sont relativement bien réparties entre les différentes catégories, sans concentration sur un étage particulier.

La boîte à moustaches indique une médiane située autour du deuxième étage et un intervalle interquartile modéré, traduisant une dispersion non négligeable mais limitée de cette variable. Les valeurs extrêmes observées correspondent aux étages minimum et maximum de l'échantillon et ne constituent pas des valeurs aberrantes au sens statistique.

Compte tenu de la nature discrète de la variable *Etage* et de l'absence d'asymétrie marquée ou de dispersion excessive, une transformation logarithmique n'est pas forcément pertinente.

- Qualite ecole

L'histogramme de la variable *Qualite_ecole* met en évidence une distribution relativement étalée des observations sur l'ensemble de l'échelle, comprise entre 1 et 10. La distribution apparaît globalement symétrique, avec une concentration plus importante des valeurs autour de la moyenne, située approximativement entre 5 et 6.

La boîte à moustaches confirme cette observation, la médiane se situe autour de 5,5 et l'intervalle interquartile est modéré. Les valeurs extrêmes observées correspondent aux bornes de l'échelle et ne constituent pas des valeurs aberrantes au sens statistique.

Compte tenu de la nature bornée et quasi continue de la variable *Qualite_ecole*, ainsi que de l'absence d'asymétrie marquée ou de valeurs extrêmes, une transformation logarithmique ne semble pas nécessaire.

## 3.2 Analyse de corrélation

La matrice de corrélation est calculée entre les principales variables numériques du jeu de données, à savoir le prix, la surface, le nombre de chambres, la distance au centre-ville, le revenu médian du quartier ainsi que l'année de vente et de construction. Ces variables,

mesurées sur des échelles quantitatives, se prêtent à l'analyse des relations linéaires à l'aide du coefficient de corrélation de Pearson. À l'inverse, les variables telles que l'étage ou la présence d'un ascenseur sont de nature discrète ou qualitative ; elles ne sont donc pas incluses dans la matrice de corrélation, mais seront analysées ultérieurement dans les modèles de régression.

| | Surf. | Cham. | An_c. | Dist. | Rev. | Etage | Prix | An_v. | Ecole |
|---|---|---|---|---|---|---|---|---|---|
| Surf. | 1 | 0,59 | −0,03 | −0,07 | 0,01 | 0,06 | 0,83 | 0,06 | 0,04 |
| Cham. | | 1 | 0,04 | −0,10 | 0,01 | 0,05 | 0,61 | 0,04 | −0,01 |
| An_c. | | | 1 | −0,09 | 0,03 | −0,04 | 0,07 | −0,06 | −0,01 |
| Dist. | | | | 1 | 0,11 | 0,02 | −0,31 | 0,08 | 0,04 |
| Rev. | | | | | 1 | 0,04 | 0,21 | 0,01 | 0,60 |
| Etage | | | | | | 1 | 0,13 | 0,13 | −0,03 |
| Prix | | | | | | | 1 | 0,24 | 0,25 |
| An_v. | | | | | | | | 1 | 0,02 |
| Ecole | | | | | | | | | 1 |

Figure 9: Matrice de corrélation des variables continues

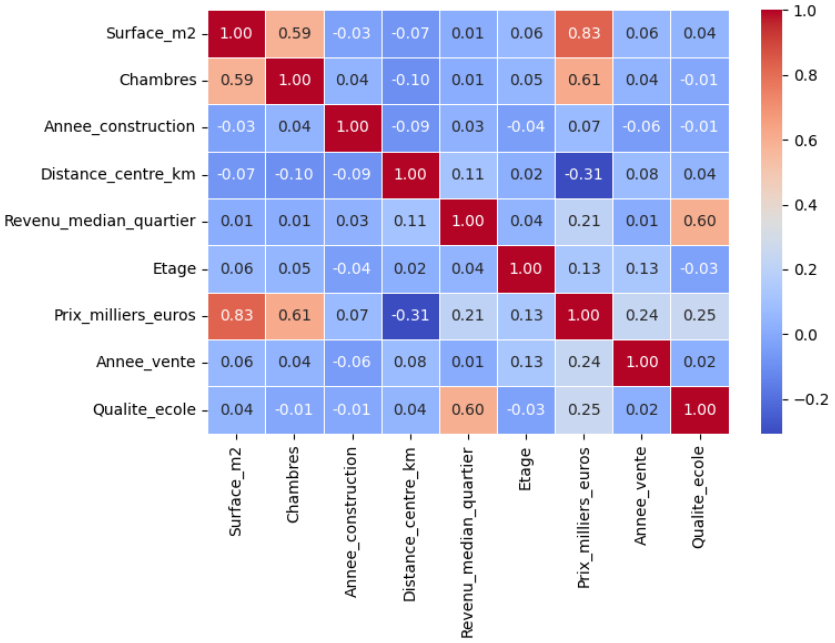


Figure 10: Graphique de Corrélation

L'analyse de la multicolinéarité repose sur l'étude des corrélations entre les variables ex-

plicatives, comme l'illustrent la matrice de corrélation et la heatmap associée. La multicolinéarité correspond à une situation dans laquelle certaines variables explicatives sont fortement corrélées entre elles, rendant l'estimation et l'interprétation des coefficients plus incertaines.

L'examen de la matrice montre que, si certaines variables présentent des corrélations élevées avec le prix du logement notamment la surface **(0,83)** et le nombre de chambres **(0,61)**,ce résultat s'inscrit dans une logique économique attendue. Toutefois, ces corrélations concernent la variable dépendante et ne relèvent donc pas directement du problème de multicolinéarité.

Concernant les corrélations entre variables explicatives, la relation la plus notable est observée entre la surface et le nombre de chambres **(0,59)**, ce qui indique une dépendance modérée entre ces deux variables sans atteindre un niveau problématique. On observe également une corrélation d'environ **(0,60)** entre le revenu médian du quartier et la qualité des écoles, cohérente avec la réalité socio-économique, mais restant inférieure au seuil critique généralement identifié autour de **(0,8)** dans la littérature.

Dans l'ensemble, les corrélations entre les autres variables explicatives demeurent faibles. Ces résultats suggèrent ainsi l'absence de multicolinéarité sévère dans le modèle, ce qui permet de considérer les estimations obtenues comme globalement fiables.

## 3.3 Modèle linéaire simple

Nous considérons dans un premier temps un modèle de régression linéaire simple reliant le prix de vente d'une maison à sa surface habitable :

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

où $Prix_i$ désigne le prix de la maison $i$, $Surface_i$ sa surface habitable, $\beta_0$ la constante du modèle, $\beta_1$ le coefficient mesurant l'effet marginal de la surface sur le prix, et $u_i$ le terme d'erreur regroupant l'ensemble des facteurs non observés influençant le prix du logement (localisation précise, état du bien, vue, environnement, etc.).

Les paramètres $\beta_0$ et $\beta_1$ sont estimés à l'aide de la méthode des moindres carrés ordinaires (MCO). Les résultats de l'estimation sont présentés dans le tableau ci-dessous et incluent les estimateurs des coefficients, leurs écarts-types, les statistiques de Student associées, les p-valeurs correspondantes, ainsi que le coefficient de détermination $R^2$ et le $R^2$ ajusté.

| **Variable** | $\hat{\beta}_j$ | Écart-type | Statistique $t$ | p-valeur (arrondie) |
|---|---|---|---|---|
| Constante ($\beta_0$) | 1519,37 | 34,58 | 43,93 | 0,000 |
| Surface ($\beta_1$) | 5,04 | 0,28 | 17,88 | 0,000 |

| | |
|---|---|
| $R^2$ | 0,683 |

Table 3: Modèle de régression linéaire simple

Le coefficient estimé associé à la surface ($\hat{\beta}_1$) est de 5,04 milliers d'euros. Il indique qu'une augmentation de 1 m$^2$ de la surface habitable entraîne, en moyenne, une hausse du prix de vente de la maison d'environ 5 040 euros. Ce résultat met en évidence l'impact positif et économiquement significatif de la surface sur le prix des logements.

Le coefficient associé à la constante ($\hat{\beta}_0$) est de 1 519,37 milliers d'euros. Bien que ce coefficient corresponde mathématiquement au prix théorique pour une surface nulle, cette situation n'ayant pas de sens économique, la constante doit être interprétée principalement comme un paramètre d'ajustement du modèle.

Les écarts-types relativement faibles des coefficients, en particulier pour la surface, indiquent que les estimations sont précises. Les p-valeurs associées aux coefficients sont toutes inférieures à 5%, ce qui conduit à rejeter l'hypothèse nulle de nullité des coefficients et à conclure à leur significativité statistique.

Le coefficient de détermination $R^2$ est de 0,683, ce qui signifie qu'environ 68,3% de la variation du prix des maisons est expliquée par la surface habitable. Ce résultat souligne le rôle central de la surface dans la détermination du prix, tout en indiquant qu'une part non négligeable de la variation du prix reste liée à d'autres caractéristiques non prises en compte dans ce modèle simple.

## 3.4 Modèle de régression linéaire multiple

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

| **Variable** | $\hat{\beta}_j$ | Écart-type | Statistique $t$ | p-valeur (arrondie) |
|---|---|---|---|---|
| Constante ($\beta_0$) | -1679,49 | 1535,67 | -1,09 | 0,276 |
| Surface ($\beta_1$) | 4,39 | 0,29 | 15,01 | 0,000 |
| Chambres ($\beta_2$) | 33,92 | 10,23 | 3,32 | 0,001 |
| Année de construction ($\beta_3$) | 1,61 | 0,77 | 2,10 | 0,037 |
| Distance au centre ($\beta_4$) | -6,14 | 0,99 | -6,19 | 0,000 |
| Étage ($\beta_5$) | 12,25 | 5,05 | 2,43 | 0,016 |
| Ascenseur ($\beta_6$) | 55,51 | 17,92 | 3,10 | 0,002 |

| | |
|---|---|
| $R^2$ | 0,789 |
| $R^2$ ajusté | 0,780 |

Table 4: Modèle de régression linéaire multiple

## 3.5 Tests de significativité

**1. Tous les coefficients sont-ils significatifs?**

L'examen des résultats de l'estimation montre que l'ensemble des coefficients associés aux variables explicatives est statistiquement significatif au seuil de 5 %, et pour la plupart au seuil de 1 %. La surface habitable, le nombre de chambres, l'année de construction, la distance au centre-ville, l'étage et la présence d'un ascenseur présentent tous des p-valeurs inférieures à 5 % et sont, pour la plupart, significatifs au seuil de 1 %. Cela indique qu'ils exercent tous un effet significatif sur le prix des logements, une fois les autres variables contrôlées.

La constante, en revanche, n'est pas statistiquement significative, comme l'indique la p-valeur associée à son coefficient estimé, égale à 0,276, largement supérieure aux seuils usuels de significativité de 5 % et de 10 %. La statistique de Student associée à la constante est de $-1{,}09$, ce qui ne permet pas de rejeter l'hypothèse nulle selon laquelle le coefficient de la

constante est égal à zéro. Cela s'explique par le fait que la constante représente la valeur du prix lorsque l'ensemble des variables explicatives est nul, une situation qui n'est pas économiquement pertinente et réaliste dans le contexte de l'immobilier.

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

Dans le cadre de notre modèle de régression linéaire multiple, l'impact marginal de chaque variable explicative correspond au coefficient estimé qui lui est associé. Il s'interprète comme la variation moyenne du prix du logement lorsque la variable considérée augmente d'une unité, toutes choses égales par ailleurs.

**Surface**
Le coefficient de la surface ($m^2$) est de $4,39$, avec une p-valeur arrondie à 0,000. Cela signifie que cette variable est très significative. En d'autres termes, chaque mètre carré supplémentaire de surface augmente le prix de la maison d'environ 4 388 €.

**Nombre de chambres**
Le coefficient des chambres est de $33,92$, et la p-valeur est de 0,001. Cela montre que le nombre de chambres est également très significatif. Une chambre supplémentaire augmente le prix de la maison de 33 920 €.

**Année de construction**
Le coefficient de l'année de construction est de $1,6$, et la p-valeur est de 0,037, ce qui montre que cette variable est significative au seuil de 5%. Cela signifie qu'un an de modernité supplémentaire dans la construction de la maison augmente son prix de 1 609 €.

**Distance au centre-ville**
Le coefficient de la distance au centre-ville est de $-6,14$, et la p-valeur est arrondie à 0,000. Cela indique que cette variable est très significative. Plus précisément, chaque kilomètre supplémentaire d'éloignement du centre-ville réduit le prix de la maison de 6 145 €.

**Étage**
Le coefficient de l'étage est de $12,25$, avec une p-valeur de 0,016, ce qui est significatif au seuil de 5%. Chaque étage supplémentaire dans un immeuble ajoute en moyenne 12 254 € à la valeur de la maison.

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

La variable Ascenseur étant une variable indicatrice, son coefficient estimé de 55,51 s'interprète comme une différence de prix. À caractéristiques comparables, un logement disposant d'un ascenseur se vend en moyenne 55,51 milliers d'euros plus cher qu'un logement sans ascenseur.

**4. Comment interprétez-vous la différence entre $R^2$ et $\bar{R}^2$ ?**

Dans le modèle de régression linéaire multiple, le coefficient de détermination $R^2$ est égal à 0,789, tandis que le $R^2$ ajusté, noté $\bar{R}^2$, est de 0,780. Le coefficient $R^2$ mesure la part de la variance du prix expliquée par l'ensemble des variables explicatives du modèle. Le $R^2$ ajusté corrige cette mesure en tenant compte du nombre de variables incluses et de la taille de l'échantillon.

L'écart relativement faible entre $R^2$ et $\bar{R}^2$, égal à 0,009, indique que l'ajout des variables explicatives améliore réellement le pouvoir explicatif du modèle et ne correspond pas à un simple effet mécanique lié à l'augmentation du nombre de variables. Ainsi, les variables

incluses dans le modèle apportent une information pertinente pour l'explication du prix des logements, sans générer de sur-ajustement trop important.

### 3.6 Transformation logarithmique

#### 1. Comparaison des trois modèles

Afin de capter d'éventuelles relations non linéaires et de corriger l'hétéroscédasticité suspectée dans le modèle linéaire initial, deux nouvelles spécifications du modèle ont été estimées par la méthode des moindres carrés ordinaires (MCO).

Le premier est le modèle *semi-logarithmique* (Semi-Log), dans lequel une transformation logarithmique est appliquée uniquement à la variable dépendante, le prix, tandis que les variables explicatives sont conservées à leur niveau initial. Ce modèle s'écrit sous la forme suivante :

$$\ln(Prix_i) = \beta_0 + \beta_1 Surface_i + \cdots + u_i.$$

Le second est le modèle *log-logarithmique* (Log-Log), dans lequel une transformation logarithmique est appliquée à la fois à la variable dépendante et aux variables explicatives continues. Il s'écrit alors :

$$\ln(Prix_i) = \beta_0 + \beta_1 \ln(Surface_i) + \cdots + u_i.$$

La comparaison des trois modèles repose sur le coefficient de détermination ajusté. Le modèle linéaire présente un $R^2$ ajusté de 0,780, indiquant qu'environ 78 % de la variance du prix est expliquée par les variables retenues. Le modèle semi-log améliore légèrement cette performance avec un $R^2$ ajusté de 0,783, ce qui en fait le modèle offrant la meilleure qualité d'ajustement parmi les trois. En revanche, le modèle log-log affiche un $R^2$ ajusté nettement plus faible, égal à 0,732, traduisant une perte de pouvoir explicatif.

| **Modèle** | $R^2$ | $R^2$ ajusté |
|---|---|---|
| Linéaire | 0,789 | 0,780 |
| Semi-log | 0,792 | 0,783 |
| Log-log | 0,743 | 0,732 |

Table 5: Comparaison des coefficients de détermination des modèles

## 4 Partie 2 : Diagnostics et corrections

Au regard des résultats obtenus dans la partie précédente, le modèle semi-log est retenu comme spécification de référence pour la suite de l'analyse.

### 4.1 Multicolinéarité et observations influentes

La multicolinéarité est analysée à l'aide des facteurs d'inflation de la variance (VIF), calculés à partir du modèle semi-log retenu, dans lequel le logarithme du prix est expliqué par la surface, le nombre de chambres, l'année de construction, la distance au centre-ville, l'étage et la présence d'un ascenseur.

Les VIF obtenus pour les variables explicatives sont globalement faibles et proches de 1. Les valeurs les plus élevées concernent la surface et le nombre de chambres, avec des VIF d'environ 1,56, tandis que les autres variables présentent des VIF compris entre 1 et 1,03. La variable constante présente un VIF élevé, ce qui est attendu car elle est mécaniquement

| Variable | VIF |
|---|---|
| Constante | 30387.38 |
| Surface_m2 | 1.56 |
| Chambres | 1.56 |
| Annee_construction | 1.03 |
| Distance_centre_km | 1.02 |
| Etage | 1.01 |
| Ascenseur | 1.03 |

Table 6: Facteurs d'inflation de la variance (VIF)

corrélée aux autres variables du modèle ; ce VIF n'est pas interprété et ne constitue pas un problème de multicolinéarité.
Ces résultats indiquent qu'aucune variable ne présente un VIF élevé et qu'il n'existe pas de multicolinéarité significative dans le modèle. La corrélation modérée observée entre certaines variables explicatives, en particulier entre la surface et le nombre de chambres, est intuitive et ne compromet pas la stabilité des coefficients estimés. Il n'est donc pas nécessaire de supprimer certaines variables sur la base du critère de multicolinéarité.

Supprimer une variable pertinente pourrait au contraire introduire un biais de variable omise. Ce biais apparaît lorsqu'une variable ayant un effet réel sur la variable expliquée est exclue du modèle alors qu'elle est corrélée avec une ou plusieurs variables explicatives conservées. Dans ce cas, l'effet de la variable omise est partiellement capté par les variables incluses, ce qui conduit à des coefficients biaisés et à une interprétation erronée des effets estimés. Au regard des VIF calculés dans le cadre du modèle semi-log, le maintien de l'ensemble des variables explicatives est donc justifié.

## 4.2 Tests et inférence

**Effet de la distance au centre-ville**
Afin d'évaluer l'impact de la distance au centre-ville sur le prix des biens immobiliers, un test de significativité est mené sur le coefficient associé à cette variable. L'estimation montre que le coefficient de la distance au centre est égal à 0,00301. Ce coefficient négatif indique qu'une augmentation de 1 km de la distance au centre-ville est associée, toutes choses égales par ailleurs, à une baisse d'environ 0,3 % du prix du bien.

| Statistique | Valeur |
|---|---|
| Coefficient (Distance_centre_km) | $-0.00301$ |
| Statistique $t$ | $-6.42$ |
| $p$-value bilatérale | $1.91 \times 10^{-9}$ |
| $p$-value unilatérale (effet négatif) | $9.53 \times 10^{-10}$ |

Table 7: Test de l'effet de la distance au centre sur le prix

La statistique de test associée à ce coefficient est égale à $-6.42$. La p-value ébilatérale est égale à $1,91 \times 10^{-9}$. Étant donné que l'hypothèse alternative est unilatérale et que la statistique de test est négative, la p-value unilatérale correspond à la moitié de la p-value bilatérale, soit $9,53 \times 10^{-10}$. Cette valeur étant extrêmement faible, l'hypothèse nulle est rejetée. On conclut donc à l'existence d'un effet négatif statistiquement très significatif de la distance au centre-ville sur le prix immobilier.
**Test de l'hypothèse des coefficients nuls et ajout des variables Qualité école et Revenu median quartier**
Dans le modèle semi-log de référence, incluant les variables *Surface*, *Chambres*, *Année de construction*, *Distance au centre*, *Étage* et *Ascenseur*, on teste l'hypothèse nulle selon laque-

lle l'ensemble des coefficients des variables explicatives, à l'exception de la constante, est simultanément nul.
Le test global de significativité (test de Fisher) fournit une statistique de test égale à

$$F = 90{,}56$$

associée à une p-value de

$$\text{p-value} = 3{,}309 \times 10^{-46}.$$

Cette p-value étant extrêmement faible, l'hypothèse nulle est rejetée aux seuils usuels de significativité. Le modèle est donc globalement significatif, ce qui indique qu'au moins une des variables explicatives exerce un effet statistiquement différent de zéro sur le logarithme du prix immobilier.

On teste ensuite si l'ajout des variables *Qualite_ecole* et *Revenu_median_quartier* améliore significativement le modèle, en comparant le modèle restreint (sans ces deux variables) au modèle non restreint (les incluant) à l'aide d'un test de Fisher de restrictions conjointes.
La statistique de test obtenue est égale à

$$F = 29{,}30$$

avec une p-value associée de

$$\text{p-value} = 2{,}278 \times 10^{-11}.$$

Cette p-value étant largement inférieure aux seuils usuels de significativité, l'hypothèse nulle selon laquelle les coefficients associés à *Qualite_ecole* et *Revenu_median_quartier* sont simultanément nuls est rejetée. L'ajout de ces deux variables améliore donc significativement le modèle.

Cette amélioration se reflète également dans le coefficient de détermination ajusté, qui passe de $R^2_{\text{ajusté}} = 0{,}7829$ pour le modèle de base à $R^2_{\text{ajusté}} = 0{,}8445$ pour le modèle étendu.
**Pourquoi un test t multiple est inadapté pour tester des restrictions conjointes**
L'utilisation successive de plusieurs tests de Student pour évaluer la pertinence d'un groupe de variables explicatives constitue une erreur méthodologique, car cette approche ne permet pas de tester rigoureusement une hypothèse conjointe. Le test de Student est conçu pour évaluer la significativité d'un paramètre pris seul, c'est-à-dire l'effet marginal d'une variable conditionnellement aux autres variables du modèle. Il ne permet donc pas de conclure sur la pertinence globale d'un ensemble de variables.

Le test de Fisher répond précisément à cette limitation, puisqu'il permet de tester simultanément plusieurs restrictions sur les paramètres du modèle. En évaluant l'hypothèse nulle selon laquelle l'ensemble des coefficients associés à un bloc de variables est nul, le test F fournit un cadre rigoureux pour juger de leur contribution globale à l'explication de la variable dépendante, tout en contrôlant le niveau de significativité du test.

Par ailleurs, l'utilisation exclusive de tests t individuels peut être trompeuse en présence de corrélations entre les variables explicatives. Lorsque certaines variables sont fortement corrélées, chacune peut apparaître individuellement non significative, car l'autre capte déjà une partie de l'information commune. Le test F, en considérant l'effet conjoint du groupe de variables, permet de mettre en évidence leur capacité explicative globale, même lorsque leurs contributions individuelles sont difficiles à identifier séparément.

## 4.3 Stabilité structurelle : test de Chow

Au regard des résultats obtenus dans la partie précédente, les variables *Qualite_ecole* et *Revenu_median_quartier*, dont l'ajout améliore significativement le modèle, sont intégrées à la spécification semi-log retenue pour la suite de l'analyse.

La stabilité structurelle du modèle est testée afin d'évaluer si la crise du COVID a modifié la relation entre le prix immobilier et ses déterminants. Le test est mené à partir du modèle semi-log étendu, incluant la surface, le nombre de chambres, l'année de construction, la distance au centre-ville, l'étage, la présence d'un ascenseur, la qualité des écoles et le revenu médian du quartier.

Une variable indicatrice *COVID*, égale à 1 pour les ventes réalisées à partir de 2020 et à 0 sinon, est introduite, ainsi que des termes d'interaction entre cette variable et l'ensemble des variables explicatives. Cette spécification permet de tester l'existence d'une rupture structurelle de type Chow, en autorisant les coefficients du modèle à différer avant et après le COVID.

L'hypothèse nulle testée est celle de la stabilité structurelle, selon laquelle les coefficients du modèle sont identiques sur l'ensemble de la période. Le test de Fisher associé fournit une statistique de test égale à

$$F = 9{,}79$$

pour $q = 8$ restrictions, avec une p-value de

$$\text{p-value} = 1{,}25 \times 10^{-10}.$$

Cette p-value étant extrêmement faible, l'hypothèse nulle est rejetée aux seuils usuels de significativité. On conclut donc à l'existence d'une rupture structurelle significative liée au COVID : la relation entre le prix immobilier et ses déterminants a été modifiée à partir de cette période.

Cette rupture structurelle a des implications importantes pour l'analyse. Elle suggère que les effets marginaux des variables explicatives ne sont pas stables dans le temps et qu'un modèle unique estimé sur l'ensemble de la période peut masquer des comportements différenciés avant et après le COVID. Dans ce contexte, il peut être pertinent d'estimer des modèles séparés pour les périodes pré- et post-COVID afin de mieux appréhender l'évolution des déterminants du prix immobilier. Toutefois, cette approche implique une réduction de la taille des échantillons et doit donc être envisagée en tenant compte du compromis entre précision des estimations et richesse de l'interprétation économique.

Finalement, à la suite du rejet de l'hypothèse de stabilité structurelle, on retient une spécification autorisant une rupture liée au COVID, dans laquelle les coefficients du modèle semi-log peuvent différer avant et après cette période :

$$\ln(Prix_i) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \gamma_0\, COVID_i + \sum_{k=1}^{K} \gamma_k\, (COVID_i \times X_{ki}) + u_i.$$

## 4.4 Hétéroscédasticité et autocorrélation

**Analyse graphique des résidus**
L'analyse graphique des résidus est menée à partir du modèle semi-log intégrant l'effet du COVID, afin d'évaluer visuellement le respect des hypothèses portant sur les erreurs.

Les graphiques mettent en évidence une dispersion non homogène des résidus : leur variance tend à augmenter avec le niveau des valeurs ajustées, ce qui se traduit par une forme d'éventail. Ce comportement suggère que les erreurs ne présentent pas une variance constante sur l'ensemble de l'échantillon.

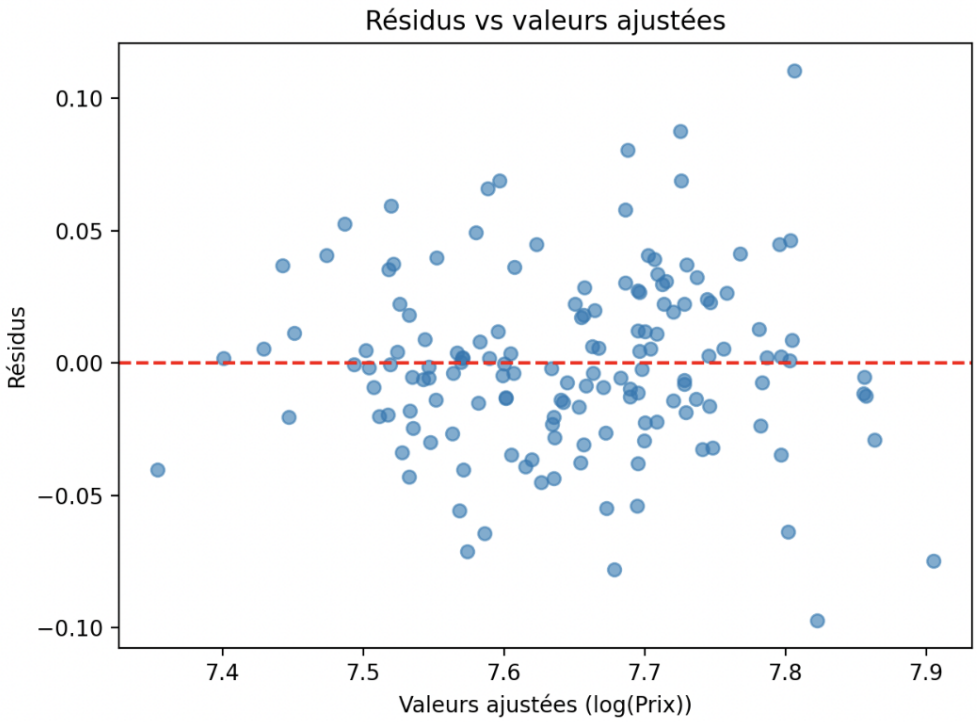


Figure 11: Graphique des résidus du modèle semi-log

Par ailleurs, aucun motif systématique de type courbe ou oscillation marquée n'est observé, ce qui indique que la spécification fonctionnelle du modèle est globalement correcte. En revanche, la variation de l'amplitude des résidus selon le niveau du prix constitue un indice visuel en faveur de la présence d'hétéroscédasticité. Cette observation justifie la mise en œuvre de tests formels d'hétéroscédasticité et, le cas échéant, l'adoption de méthodes d'inférence robustes dans la suite de l'analyse.

**Test formel d'hétéroscédasticité et comparaison des méthodes d'estimation**

Bien que la stabilité structurelle du modèle ait été analysée séparément, les tests d'hétéroscédasticité sont menés sur le modèle semi-log de référence afin d'évaluer exclusivement les propriétés de la variance des erreurs. Les variables de rupture liées au COVID, qui affectent la valeur moyenne du prix, ne sont pas incluses dans ces tests afin de ne pas confondre instabilité des coefficients et hétéroscédasticité.

À la suite de l'analyse graphique des résidus, l'hypothèse d'hétéroscédasticité est testée formellement à l'aide de tests statistiques standards. Deux tests complémentaires sont mobilisés : le test de Breusch–Pagan et le test de White.

Le test de Breusch–Pagan fournit une statistique égale à $BP = 5{,}48$, associée à une p-value de 0,706. Cette valeur étant largement supérieure aux seuils usuels de significativité, l'hypothèse nulle d'homoscedasticité ne peut pas être rejetée. De manière cohérente, le test de White conduit à une statistique de 40,68 avec une p-value de 0,573, confirmant l'absence d'hétéroscédasticité statistiquement significative dans les résidus du modèle semi-log.

En conséquence, les estimations obtenues par moindres carrés ordinaires (MCO) peuvent être considérées comme valides du point de vue de l'inférence, et les écarts-types usuels sont a priori appropriés.

Afin de vérifier la robustesse des résultats, le modèle est néanmoins réestimé en utilisant des écarts-types robustes à l'hétéroscédasticité au sens de White (HC1). Les coefficients estimés restent strictement identiques à ceux du modèle MCO standard, et les écarts-types robustes sont très proches des écarts-types usuels. Les niveaux de significativité des variables explicatives ne sont pas modifiés, ce qui indique que l'inférence n'est pas sensible à une éventuelle hétéroscédasticité résiduelle.

Enfin, une estimation par moindres carrés pondérés (WLS) est également conduite à titre de comparaison. Cette méthode conduit à un coefficient de détermination extrêmement élevé, ainsi qu'à des écarts-types artificiellement faibles. Toutefois, en l'absence d'hétéroscédasticité détectée par les tests formels et compte tenu de l'absence de modification substantielle des coefficients estimés, cette amélioration apparente ne reflète pas un gain économétrique réel. La méthode WLS n'est donc pas retenue pour la suite de l'analyse.

L'ensemble de ces résultats conduit à conclure que le modèle semi-log peut être estimé de manière fiable par MCO standard, sans recourir à des corrections spécifiques de l'hétéroscédasticité.

**Test d'autocorrélation**
L'hypothèse d'absence d'autocorrélation des erreurs est testée afin de vérifier une autre condition essentielle à la validité de l'inférence par les moindres carrés ordinaires (MCO). L'autocorrélation correspond à une dépendance entre les erreurs associées à différentes observations, susceptible d'apparaître lorsque les données sont ordonnées dans le temps ou lorsqu'un choc commun affecte plusieurs observations successives.
Le test retenu est le test de Durbin–Watson, dont l'hypothèse nulle est l'absence d'autocorrélation de premier ordre des erreurs. La statistique de Durbin–Watson obtenue est égale à

$$DW = 2{,}25.$$

Cette valeur étant proche de 2, elle ne met pas en évidence d'autocorrélation significative des résidus. L'hypothèse nulle d'absence d'autocorrélation n'est donc pas rejetée.
Dans ce contexte, les hypothèses usuelles des MCO concernant la structure des erreurs apparaissent satisfaites. Étant donné qu'aucune hétéroscédasticité statistiquement significative n'a été détectée précédemment et qu'aucune autocorrélation n'est mise en évidence ici, il n'est pas nécessaire de recourir à des écarts-types corrigés de type Newey–West. Les estimations MCO standard peuvent ainsi être conservées pour l'inférence.

# 5 Partie 3 : Endogénéité

## 5.1 Sources d'endogénéité

Plusieurs sources d'endogénéité peuvent être présentes dans notre modèle, notamment les variables omises, la causalité inverse et les erreurs de mesure.

Tout d'abord, un problème de **variables omises** peut survenir si certaines caractéristiques non observées des quartiers, telles que leur attractivité résidentielle, leur niveau de sécurité, leur image ou encore leur environnement socio-économique, influencent le prix des logements et la qualité des écoles. Ces facteurs, non pris en compte dans le modèle, sont intégrés au terme d'erreur et peuvent être corrélés à la variable *Qualite_ecole*, ce qui génère un biais

d'endogénéité.

Ensuite, une **causalité inverse** est également plausible. En effet, des prix immobiliers élevés attirent généralement des ménages plus aisés, disposant de davantage de ressources financières et d'un capital social plus important. Cette concentration de ménages peut contribuer à améliorer la qualité des établissements scolaires locaux, rendant difficile l'identification d'un effet allant uniquement de la qualité des écoles vers le prix des logements.

Enfin, la variable *Qualite_ecole* peut être affectée par une **erreur de mesure**, dans la mesure où un indicateur de qualité scolaire ne reflète pas parfaitement la réalité des choses et la qualité réelle des établissements. Cette approximation peut également entraîner une corrélation entre la variable explicative et le terme d'erreur.

Ainsi, la variable *Qualite_ecole* est potentiellement endogène, car elle est susceptible d'être corrélée au terme d'erreur du modèle. Cette endogénéité peut biaiser les estimations obtenues par la MCO et empêcher d'interpréter le coefficient associé comme un effet causal.

## 5.2 Estimation par Variables Instrumentales

L'intégration et le choix de la distance à l'université la plus proche comme instrument pour la qualité des écoles repose sur deux piliers. Premièrement il existe souvent une corrélation géographique et institutionnelle entre la présence d'universités et la qualité des établissements scolaires. La proximité d'un centre universitaire favorise un environnement éducatif dynamique, attire des familles au capital culturel élevé et peut faciliter des partenariats ou d'autres ressources, ce qui tend à améliorer le niveau des écoles primaires et secondaires du secteur.

Deuxièmement, la variable satisfait la condition d'exogénéité par opposition à l'endogénité. La distance géographique à une université est généralement une donnée historique et structurelle de l'aménagement du territoire qui n'influence pas directement le prix d'un logement, autrement que par l'intermédiaire de la qualité des infrastructures éducatives qu'elle suppose. Contrairement à la qualité de l'école, la distance à l'université est moins susceptible d'être corrélée aux caractéristiques non observées du quartier ou d'être affectée par une causalité inverse, car l'emplacement d'une université ne change pas en fonction de l'augmentation récente des prix immobiliers d'un quartier spécifique.

Nous allons donc maintenant vérifier cette hypothèse et soumettre une estimation par variables instrumentales réalisée à l'aide de la méthode des moindres carrés en deux étapes (2SLS). La procédure effectue automatiquement une première étape expliquant la qualité des écoles par l'instrument et les variables exogènes, puis une seconde étape estimant l'équation de prix à partir de la valeur prédite de la qualité des écoles.

## 5.3 Tests de validité

Dans la partie consacrée à l'endogénéité, l'analyse est menée à partir du modèle linéaire afin d'isoler l'effet de la correction par variables instrumentales. L'introduction de transformations logarithmiques ou de variables de rupture structurelle comme effectué précedemment est volontairement écartée afin de garantir la comparabilité entre les estimations MCO et IV.

L'estimation par variables instrumentales est réalisée à l'aide de la méthode des moindres carrés en deux étapes (Le tableau de résultat est référencé en Annexes : Table 15 ), en expliquant la qualité des écoles par l'instrument et les variables exogènes, puis en estimant l'équation de prix à partir de la valeur prédite de la qualité des écoles.Le tableau est référencé

en annexe

Le modèle est estimé sur 150 observations et présente un $R^2$ de **0,880** (avec un $R^2$ ajusté de **0,872**), ce qui indique un pouvoir explicatif très élevé. Par ailleurs, les coefficient des variables explicatives vu précedemment sont toujours aussi statistiquement significatifs

En revanche, le coefficient associé à la variable Qualite_ecole, instrumentée dans le cadre de l'estimation 2SLS, est faible et non significatif. Il est estimé à **2,09** avec une p-valeur de **0,874**, ce qui ne permet pas de mettre en évidence un effet significatif de la qualité des écoles sur le prix des logements, une fois l'endogénéité corrigée.

Cette différence par rapport à l'estimation MCO fait précédemment suggère que l'effet de la qualité des écoles était probablement surestimé , en raison d'un biais d'endogénéité. L'approche par variables instrumentales permet ainsi d'identifier une estimation plus fidèle de la relation réelle entre les variables, mais avec moins de précision dans les résultats.

Nous allons maintenant tester la validité de l'instrument en utilisant la méthode de la statistique F.

Cette statistique est calculée dans la première étape de l'estimation 2SLS, où nous régressons la variable endogène Qualite_ecole sur l'instrument Distance_universite et les autres variables exogènes. Le test d'hypothèse associé à la statistique F permet de vérifier si l'instrument (ici Distance_universite) explique suffisamment bien la variable endogène.

Selon la règle rappelée dans le cours de de Staiger & Stock, une valeur de F supérieure à 10 indique que l'instrument est fortement corrélé avec la variable endogène et permet d'écarter le problème d'un instrument faible. Si la statistique F est faible (inférieure à 10), cela suggère que l'instrument n'est pas pertinent, ce qui nuirait à la validité de notre estimation 2SLS.

| **First Stage Estimation Results** | |
|---|---|
| **R-squared** | 0.4419 |
| **Partial R-squared** | 0.1180 |
| **Shea's R-squared** | 0.1180 |
| **Partial F-statistic** | 20.911 |
| **P-value (Partial F-stat)** | 4.812e-06 |
| **Partial F-stat Distn** | chi2(1) |

Table 8: First Stage Estimation Results

Les résultats indiquent que l'instrument *Distance_universite* est significativement corrélé à la variable endogène *Qualite_ecole*. La statistique F partielle est égale à **20,91**, ce qui est supérieur au seuil de 10, permettant ainsi d'écarter le problème des instruments faibles. La p-value associée, égale à $\mathbf{4{,}81 \times 10^{-6}}$, confirme que cette relation est statistiquement significative.

Par ailleurs, le $\mathbf{R^2}$ partiel de la première étape est d'environ **0,118,** ce qui signifie que l'instrument explique près de 12 % de la variation de la qualité des écoles. L'instrument

*Distance_universite* peut ainsi être considéré comme valide pour l'estimation par variables instrumentales.

- Comparaisons des coefficients MCO et IV (Tableau en Annexes)

| **Comparaison des Coefficients** | | |
|---|---|---|
| **Variable** | **MCO** | **IV** |
| *Année de construction* | 1.699645 | 1.625046 |
| *Année de vente* | 21.281043 | 21.620536 |
| *Ascenseur* | 54.926324 | 55.583787 |
| *Chambres* | 35.151060 | 33.229231 |
| *Distance au centre* | -7.056503 | -7.167897 |
| *Étage* | 9.132240 | 8.029486 |
| *Qualité des écoles* | 19.914911 | 2.093095 |
| *Revenu médian quartier* | 2.759262 | 4.921996 |
| *Surface (m²)* | 4.229384 | 4.292819 |
| *Constante* | -45090.835923 | -45664.933000 |

Table 9: Comparaison des Coefficients MCO et IV

Le modèle MCO utilisé comme référence dans la comparaison avec l'estimation par variables instrumentales correspond au modèle linéaire enrichi incluant l'ensemble des variables explicatives du modèle 2SLS. Le modèle linéaire multiple présenté précédemment constitue une spécification de base, utilisée à des fins d'interprétation initiale, tandis que le modèle MCO retenu ici est celui pertinent pour l'analyse de l'endogénéité.

La comparaison des coefficients estimés par la méthode des moindres carrés ordinaires (MCO) et par variables instrumentales (IV) montre que, pour la majorité des variables explicatives, les résultats sont très proches en termes de signe et d'ordre de grandeur. Par exemple, le coefficient de la surface passe de **4,23** en MCO à **4,29** en IV, et celui de la distance au centre-ville reste proche de $-$**7,1** dans les deux modèles. Cette stabilité suggère que ces variables peuvent être considérées comme exogènes et que leurs estimations MCO ne sont pas fortement biaisées.

En revanche, une différence importante apparaît pour la variable *Qualite_ecole*. En MCO, son coefficient est relativement élevé **19,91**, suggérant un impact important de la qualité des écoles sur le prix des logements. En IV, ce coefficient chute fortement à **2,09** et devient non significatif. Cette différence marquée indique que l'estimation MCO était affectée par un biais d'endogénéité, son impact était artificiellement gonflé par des facteurs socio-économiques environnants conduisant à une surestimation de l'effet de la qualité des écoles sur le prix, comme on l'a démontré précédemment.

On observe également une augmentation du coefficient associé au revenu médian du quartier, qui passe de **2,76** en MCO à **4,92** en IV, suggérant que le niveau socio-économique du quartier capte une partie de l'effet auparavant attribué à la qualité des écoles. Ce résultat

est cohérent avec la forte corrélation observée entre ces deux variables (environ **0,60**), mise en évidence dans la matrice de corrélation. Lorsque l'endogénéité de la qualité des écoles est corrigée par la méthode des variables instrumentales, l'effet du revenu médian apparaît plus clairement, tandis que l'effet propre de la qualité des écoles est réévalué à la baisse.

La méthode par variable instrument fournit une estimation convergente et plus fidèle à la réalité économique, tandis que les caractéristiques structurelles du logement (surface, étage) confirment leur robustesse.

# 6 Partie 4 : Méthodes de régularisation

## 6.1 Régression Ridge

L'estimation Ridge, également appelée régression Ridge, permet de régulariser les coefficients du modèle afin de réduire la variance des estimateurs et de limiter le risque de surapprentissage, en introduisant une pénalité quadratique sur les coefficients :

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Dans cette partie, la régularisation est appliquée au **modèle semi-log** retenu précédemment, dans lequel le logarithme du prix est expliqué par les caractéristiques du logement ainsi que par leurs interactions avec une variable indicatrice COVID, afin de tenir compte de la rupture structurelle identifiée à partir de 2020. Les variables explicatives, y compris les termes d'interaction avec la variable COVID, sont préalablement standardisées (moyenne nulle et écart-type égal à 1), conformément aux recommandations usuelles, car la pénalisation dépend directement de l'échelle des coefficients.

La régression Ridge est estimée pour plusieurs valeurs du paramètre de régularisation $\lambda$, à savoir

$$\lambda = \mathbf{0.001,\ \ 0.01,\ \ 0.1,\ \ 1,\ \ 10,\ \ 50\ \ et\ \ 100,}$$

afin d'observer l'impact progressif de la pénalisation sur les coefficients des variables explicatives. **Le tableau complet des estimations est présenté en annexes (Table 18)** .

La variable *Distance_universite* n'est pas incluse dans cette analyse, car elle a été introduite exclusivement comme **instrument de la variable Qualite_ecole** dans la partie consacrée à l'endogénéité, et non comme déterminant direct du prix des logements, seules les variables explicatives structurelles du modèle de prix sont retenues.

Lorsque $\lambda$ est faible ($\lambda = 0.001$, $\lambda = 0.01$ ou $\lambda = 0.1$), les coefficients estimés sont très proches de ceux obtenus sans régularisation, ce qui traduit une pénalisation limitée. À mesure que $\lambda$ augmente, on observe une contraction progressive de l'ensemble des coefficients, illustrant clairement l'effet de la pénalité Ridge.

Cette contraction est particulièrement visible pour certaines variables. Par exemple, le coefficient associé à la surface diminue sensiblement lorsque $\lambda$ augmente, passant d'environ 0,082 pour $\lambda = 0.001$ à environ 0,042 pour $\lambda = 100$. Dans un modèle semi-log, cela signifie que l'effet marginal de la surface sur le prix, exprimé en pourcentage, est progressivement atténué par la régularisation. De même, le coefficient de la distance au centre-ville voit sa valeur absolue diminuer, passant approximativement de 0,031 à 0,018, traduisant une réduction de l'intensité de son effet négatif sur le prix.

Cette dynamique concerne également d'autres variables du modèle. Le coefficient associé au nombre de chambres reste relativement stable pour de faibles valeurs de $\lambda$, autour de 0,017, puis évolue modérément lorsque la pénalisation devient plus forte, ce qui suggère que l'effet du nombre de chambres est robuste mais partiellement redondant avec d'autres caractéristiques du logement, notamment la surface.

Les coefficients associés à l'année de construction et à l'année de vente diminuent également de manière notable lorsque $\lambda$ augmente, ce qui indique que ces variables partagent une partie de leur pouvoir explicatif avec d'autres dimensions temporelles ou structurelles du modèle.

Les variables *Qualite_ecole* et *Revenu_median_quartier*, potentiellement corrélées entre elles, voient leurs coefficients diminuer progressivement lorsque $\lambda$ augmente. Par exemple, le coefficient de la qualité des écoles passe d'environ 0,017 pour de faibles valeurs de $\lambda$ à moins de 0,01 lorsque $\lambda = 100$. Ce comportement illustre le rôle spécifique de la régression Ridge dans la gestion de la multicolinéarité : plutôt que de sélectionner une seule variable, la pénalisation répartit l'effet explicatif entre variables corrélées.

Les termes d'interaction avec la variable COVID sont également affectés par la régularisation. L'interaction entre la surface et la période post-COVID voit notamment son coefficient se rapprocher progressivement de zéro lorsque $\lambda$ augmente, suggérant que l'effet différencié de la surface après 2020 est plus incertain et sensible à la pénalisation. D'autres interactions, comme celles associées à la qualité des écoles ou à la présence d'un ascenseur, restent positives mais faibles, indiquant que la rupture structurelle liée au COVID modifie certains effets marginaux sans remettre en cause l'ensemble des déterminants du prix.

Contrairement à la régression Lasso, aucun coefficient n'est annulé exactement à zéro, même pour des valeurs élevées de $\lambda$. La régression Ridge conserve ainsi l'ensemble des variables dans le modèle tout en réduisant leur influence relative. Ces résultats mettent en évidence le compromis biais–variance : une augmentation de $\lambda$ introduit davantage de biais, mais permet d'obtenir des estimations plus stables et potentiellement plus performantes en prédiction.

## 6.2 Régression Lasso

La régression Lasso est définie par le problème d'optimisation suivant :

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

La différence clé avec la régression Ridge réside dans la forme de la pénalité : le Lasso utilise la valeur absolue des coefficients plutôt que leur carré. Ce changement en apparence mineur a des implications importantes. En effet, la pénalité en norme $L_1$ conduit non seulement à une réduction de la magnitude des coefficients (*shrinkage*), mais également à une sélection automatique des variables. Contrairement à Ridge, certains coefficients peuvent être exactement égaux à zéro, ce qui permet d'éliminer les variables les moins pertinentes du modèle.

Comme pour la régression Ridge, le Lasso est estimé sur le modèle semi-log intégrant la rupture structurelle liée au COVID, après standardisation des variables explicatives (moyenne nulle et écart-type égal à 1). L'estimation est réalisée pour différentes valeurs du paramètre de régularisation, à savoir :

$$\lambda = 0.001,\ 0.01,\ 0.1,\ 1,\ 10,\ 50 \text{ et } 100.$$

Les résultats complets sont présentés en annexe.

Pour des valeurs très faibles de $\lambda$ ($\lambda = 0.001$ et $\lambda = 0.01$), plusieurs coefficients restent non nuls et proches de ceux obtenus dans les modèles non pénalisés. Les principales variables explicatives, telles que la surface, le nombre de chambres, la distance au centre-ville, la qualité des écoles et le revenu médian du quartier, conservent alors une contribution positive ou négative conforme aux résultats précédents.

En revanche, dès que la valeur de $\lambda$ atteint 0.1, la pénalisation devient suffisamment forte pour annuler l'ensemble des coefficients du modèle. À partir de ce seuil, et pour des valeurs plus élevées de $\lambda$ ($\lambda = 1, 10, 50, 100$), tous les coefficients sont égaux à zéro. Le Lasso conduit alors à un modèle extrêmement simplifié, dans lequel aucune variable n'est retenue comme suffisamment robuste face à la pénalisation.

Ce résultat met en évidence le caractère particulièrement agressif du Lasso dans un contexte où le nombre de variables explicatives et d'interactions est élevé par rapport à la taille de l'échantillon. Il suggère que, bien que plusieurs variables contribuent à l'explication du prix immobilier dans les modèles non pénalisés, leur pouvoir prédictif individuel devient insuffisant dès que la pénalisation augmente.

Ainsi, contrairement à la régression Ridge, qui conserve l'ensemble des variables tout en réduisant leur influence relative, le Lasso opère ici une sélection brutale des variables. Ces résultats illustrent clairement le compromis biais–variance : une pénalisation plus forte simplifie fortement le modèle, mais peut conduire à une perte d'information lorsque le signal est réparti entre plusieurs variables corrélées.

## 6.3 La valeur du paramètre $\lambda$

Maintenant nous allons choisir la valeur du paramètre $\lambda$ optimal à l'aide de la validation croisée 10-fold à la fois surle modèle ridge et sur le modèle lasso.

| **Lasso** — $\lambda = 0.001$ | |
|---|---|
| **Variable** | **Coef** |
| Surface_m2 | 0.075093 |
| Chambres | 0.017050 |
| Année construction | 0.006761 |
| Distance centre (km) | -0.031077 |
| Étage | 0.005407 |
| Ascenseur | 0.010204 |
| Qualité école | 0.015133 |
| Revenu médian quartier | 0.010842 |
| Surface_m2_COVID | 0.000000 |
| Chambres_COVID | 0.000000 |
| Année_construction_COVID | 0.017735 |
| Distance_centre_km_COVID | 0.000000 |
| Étage_COVID | 0.000000 |
| Ascenseur_COVID | 0.000790 |
| Qualité_ecole_COVID | 0.006888 |
| Revenu_median_quartier_COVID | 0.000000 |

Table 10: Coefficients Lasso

| **Ridge** — $\lambda = 2.9471$ | |
|---|---|
| **Variable** | **Coef** |
| Surface_m2 | 0.074965 |
| Chambres | 0.019424 |
| Annee_construction | 0.007982 |
| Distance_centre_km | -0.031198 |
| Etage | 0.008043 |
| Ascenseur | 0.010861 |
| Qualite_ecole | 0.016161 |
| Revenu_median_quartier | 0.010536 |
| Surface_m2_COVID | -0.005412 |
| Chambres_COVID | -0.002461 |
| Annee_construction_COVID | 0.022515 |
| Distance_centre_km_COVID | -0.000877 |
| Etage_COVID | -0.002876 |
| Ascenseur_COVID | 0.001488 |
| Qualite_ecole_COVID | 0.005907 |
| Revenu_median_quartier_COVID | 0.005891 |

Table 11: Coefficients Ridge

La valeur du paramètre de régularisation $\lambda$ est choisie séparément pour les modèles Ridge et Lasso à l'aide d'une validation croisée à 10 folds. Pour chaque méthode, une grille de valeurs de $\lambda$ est testée et la valeur retenue est celle qui minimise l'erreur de prédiction moyenne hors échantillon.

Les résultats indiquent que la valeur optimale de $\lambda$ est égale à 0,001 pour le modèle Lasso, ce qui correspond à une pénalisation très faible, le modèle conserve ainsi la quasi-totalité des variables, car une pénalisation plus forte dégraderait la capacité prédictive., tandis que pour le modèle Ridge, la valeur optimale est d'environ 2,95, traduisant une régularisation plus marquée.

Ces différences sont normales dans la mesure où Ridge et Lasso reposent sur des pénalités différentes (norme $L_2$ pour Ridge et norme $L_1$ pour Lasso) et poursuivent des objectifs distincts en termes de biais, de variance et de sélection des variables.

## 6.4 Comparaison des performances prédictives

Dans la continuité des résultats obtenus en partie 4, la comparaison des performances prédictives est réalisée à partir du modèle semi-log intégrant une rupture structurelle liée au

COVID. Les méthodes MCO, Ridge et Lasso sont estimées sur une spécification strictement identique, et comparées à l'aide de la RMSE calculée sur un échantillon test commun. Ce choix permet de conserver une cohérence globale dans l'analyse tout en évaluant l'apport des méthodes de régularisation dans un cadre économétrique réaliste.

| **Modèle** | **RMSE (échantillon test)** |
|---|---|
| MCO | 0.0413 |
| Ridge | 0.0408 |
| Lasso | 0.0411 |

Table 12: Comparaison des erreurs de prédiction (RMSE) des trois modèles

Les performances prédictives des trois modèles sont évaluées à l'aide de la RMSE sur un échantillon de test représentant 20 % des données. Les estimations portent sur le même modèle semi-log intégrant les interactions liées à la période COVID. Pour les modèles régularisés, les valeurs de $\lambda$ utilisées correspondent aux $\lambda$ optimaux déterminés précédemment par validation croisée 10-fold ($\lambda = 2{,}9471$ pour Ridge et $\lambda = 0{,}001$ pour Lasso).

Les résultats indiquent que les erreurs de prédiction des trois modèles sont très proches : la RMSE est égale à **0,0413** pour le modèle MCO, **0,0408** pour le modèle Ridge et **0,0411** pour le modèle Lasso. Étant donné que la variable dépendante est exprimée en logarithme, ces valeurs correspondent à une erreur moyenne d'environ **4 %** sur la prédiction du prix des logements. Ridge présente une performance légèrement supérieure, mais les écarts restent extrêmement faibles. Cela suggère que le modèle semi-log avec prise en compte de la rupture COVID est déjà bien spécifié et n'est pas affecté par un sur-ajustement important. Dans ce contexte, les méthodes de régularisation n'apportent donc qu'un gain marginal en termes de performance prédictive.

**Pourquoi les écarts-types et tests classiques ne sont-ils pas valides après Lasso?**

Les écarts-types et les tests statistiques classiques ne sont pas valides après l'estimation d'un modèle Lasso puisque le Lasso réalise simultanément l'estimation des coefficients et la sélection des variables à partir des données observées. Or, les tests d'inférence classiques reposent sur l'hypothèse que le modèle est spécifié indépendamment de l'échantillon utilisé pour l'estimation.

Par ailleurs, la pénalisation en norme $L_1$ introduite par le Lasso induit un biais dans l'estimation des coefficients, ceux-ci étant contraints vers zéro. Certains coefficients peuvent également être exactement nuls, ce qui entraîne des distributions non standards et discontinues des estimateurs. Dans ce contexte, les conditions asymptotiques justifiant l'utilisation des statistiques de Student et la construction d'intervalles de confiance classiques ne sont plus vérifiées.

Ainsi, le Lasso doit être principalement interprété comme un outil de sélection de variables et d'amélioration de la performance prédictive, plutôt que comme un cadre adapté à l'inférence statistique classique sur les coefficients estimés.

# 7 Conclusion et recommandations

## 7.1 Apports prédictifs du modèle et illustration empirique

**Prédiction ponctuelle**
Nous utilisons le modèle semi-logarithmique avec rupture structurelle (Post-COVID) pour prédire le prix d'un bien présentant les caractéristiques suivantes : Surface de 120 $m^2$, 3 chambres, construit en 2015, situé à 5 km du centre, au 1er étage avec ascenseur, vendu en 2023 (période Post-COVID).
En appliquant les coefficients du modèle aux caractéristiques du bien, nous obtenons la valeur prédite du logarithme du prix :

$$\ln(\widehat{\text{Prix}}) = 7{,}752$$

Pour revenir au prix en euros, une simple exponentielle (exp(7,752)) fournirait la médiane de la distribution des prix et non l'espérance mathématique, ce qui induirait un biais de sous-estimation systématique (inégalité de Jensen). Étant donné la nature log-normale du modèle, nous appliquons une correction de biais basée sur la variance des résidus ($\hat{\sigma}^2$) pour obtenir le prix moyen attendu :

$$\widehat{\text{Prix}} = \exp\left(7{,}752 + \frac{\hat{\sigma}^2}{2}\right)$$

Avec une erreur standard résiduelle (RMSE) d'environ 0,041 sur l'échantillon test, le facteur de correction reste marginal ($+0,08\%$), portant la prédiction finale à :

**Prix moyen estimé : 2 327 000 €**
**Intervalle de confiance à 95%**

L'intervalle de confiance à 95% encadre l'espérance conditionnelle du prix, c'est-à-dire le prix moyen théorique pour l'ensemble des maisons possédant exactement ces caractéristiques. Il est calculé comme suit :

$$IC_{95\%} = [\exp(7{,}736)\ ;\ \exp(7{,}768)]$$

**Résultat : [2 289 000 € ; 2 363 000 €]**

Cela signifie que si nous disposions d'un grand nombre de transactions pour des biens strictement identiques, leur prix moyen se situerait, avec 95% de certitude, dans cette fourchette.

**Cette prédiction est-elle fiable ?**
La fiabilité de cette prédiction doit être appréciée à deux niveaux distincts : la qualité statistique du modèle et la réalité opérationnelle d'une vente.
D'un point de vue statistique, la prédiction est très fiable. Le modèle présente un pouvoir explicatif élevé ($R^2$ ajusté proche de 0,90) et des erreurs de prédiction faibles (RMSE $\approx$ 4%). De plus, la prise en compte spécifique de la période Post-COVID permet d'intégrer les changements structurels récents du marché, évitant ainsi de baser l'estimation sur des dynamiques de prix obsolètes (2015-2019). L'étroitesse de l'intervalle de confiance (écart de 74 000 €) confirme que les paramètres du modèle sont estimés avec une grande précision.
Cependant, d'un point de vue opérationnel pour un vendeur unique, cette fiabilité doit être nuancée. L'intervalle de confiance présenté ci-dessus ne tient pas compte de l'aléa individuel ($\epsilon_i$), c'est-à-dire des caractéristiques non observées (charme, luminosité, état précis) qui font varier le prix d'une transaction à l'autre. Pour une vente isolée, l'incertitude réelle est mieux représentée par l'intervalle de prédiction, qui est nettement plus large (environ [2 166 000 € ; 2 497 000 €]).

En conclusion, si le modèle est un outil robuste pour déterminer la valeur de marché moyenne d'un bien (évaluation macro), il subsiste une marge de négociation et d'incertitude significative pour toute transaction individuelle.

## 7.2 Synthèse des résultats

L'analyse économétrique menée sur cet échantillon de 150 transactions immobilières a permis de mettre en évidence les déterminants fondamentaux du prix des logements tout en testant la robustesse de ces relations face aux chocs structurels et aux biais statistiques.

Premièrement, la spécification semi-logarithmique s'est révélée être la plus pertinente pour modéliser la formation des prix. Elle confirme que les caractéristiques structurelles jouent un rôle prépondérant : la surface habitable, le nombre de chambres et la présence d'un ascenseur exercent un effet positif significatif, tandis que l'éloignement du centre-ville déprécie la valeur du bien. À titre d'exemple, la présence d'un ascenseur est associée à une prime substantielle, toutes choses égales par ailleurs.

Deuxièmement, l'étude de la stabilité temporelle a identifié une rupture structurelle majeure liée à la période COVID-19. Les tests de Chow et l'analyse des interactions montrent que la dynamique de formation des prix a changé après 2020. Cela suggère que les préférences des acheteurs ou les conditions de marché ont évolué, rendant les modèles calibrés sur des données anciennes potentiellement obsolètes pour des prédictions actuelles.

Troisièmement, le traitement de l'endogénéité de la variable Qualité des écoles par la méthode des variables instrumentales (IV-2SLS) a apporté un éclairage crucial. Alors que l'estimation naïve (MCO) suggérait un fort impact de la carte scolaire sur le prix, l'approche IV révèle que cet effet est en réalité non significatif une fois corrigé des biais de variables omises (comme le niveau socio-économique du quartier). Ce résultat indique que la prime immobilière souvent attribuée aux bonnes écoles capture en réalité d'autres aménités locales ou la richesse du voisinage.

Enfin, l'application des méthodes de régularisation (Ridge et Lasso) a démontré que, sur un échantillon de cette taille, le modèle MCO standard reste performant et ne souffre pas d'un sur-ajustement critique. Ces techniques ont confirmé la robustesse des variables clés sélectionnées initialement.

## 7.3 Limites de l'analyse

Malgré la rigueur méthodologique appliquée, cette étude présente plusieurs limites intrinsèques qui invitent à la prudence quant à la généralisation des résultats.

**La taille de l'échantillon**
La limite principale réside dans la taille restreinte du jeu de données ($N = 150$). En économétrie, un échantillon aussi faible réduit la puissance des tests statistiques. Cela fragilise notamment la détection de l'hétéroscédasticité et rend les estimations par variables instrumentales (IV) particulièrement instables, ces méthodes étant connues pour être biaisées sur les petits échantillons (biais de petits échantillons des 2SLS). De même, la division de l'échantillon pour tester la rupture COVID réduit encore le nombre d'observations par période, augmentant la variance des estimateurs.

**La validité de l'instrument**
L'utilisation de la distance à l'université comme instrument pour la qualité des écoles est discutable sur le plan économique. Pour qu'un instrument soit valide, il doit respecter la condition d'exclusion (n'impacter le prix *que* via la qualité de l'école). Or, la proximité

d'une université influence probablement le prix de l'immobilier par d'autres canaux directs : demande locative étudiante, dynamisme culturel, commerces ou nuisances sonores. Si cette condition d'exclusion n'est pas strictement respectée, les estimateurs IV peuvent être incohérents.

**Variables omises et qualité de l'information**
Le modèle n'intègre pas certaines variables déterminantes sur le marché actuel, telles que le Diagnostic de Performance Énergétique (DPE), l'état intérieur du bien (rénové vs à rénover) ou l'exposition (balcon/terrasse). L'absence de ces facteurs dans le terme d'erreur peut introduire un biais si ces variables sont corrélées avec les variables explicatives retenues (par exemple, si les logements anciens sont moins bien isolés et situés plus près du centre).

**Validité externe** L'étude portant sur une région donnée non spécifiée, les résultats (notamment les élasticités-prix et la prime liée à l'ascenseur) ne sont pas nécessairement transposables à d'autres marchés immobiliers présentant des tensions ou des caractéristiques urbaines différentes.

## 7.4 Recommandations pour la pratique

Au regard des résultats empiriques et des limites identifiées, plusieurs recommandations peuvent être formulées à l'attention des acteurs du marché et pour de futurs travaux d'analyse.

**Pour les acteurs du marché (Acheteurs, Vendeurs, Investisseurs)**

- **Ne pas survaloriser le critère scolaire :** L'analyse suggère que la qualité des écoles, une fois isolée des autres facteurs de richesse du quartier, n'impacte pas significativement le prix. Les acheteurs devraient être vigilants à ne pas payer une surcote injustifiée pour ce seul critère.
- **Valoriser les équipements structurels :** La présence d'un ascenseur apparaît comme un levier de valeur puissant. Pour les investisseurs ou copropriétés, l'installation d'un ascenseur (lorsque cela est possible) semble être un investissement rentable au regard de la valorisation du bien.
- **Privilégier les données récentes :** La rupture structurelle post-COVID indique que le marché a changé. Pour estimer un bien aujourd'hui, il est recommandé d'accorder plus de poids aux transactions réalisées après 2020 et de se méfier des comparaisons avec des ventes plus anciennes.

**Recommandations méthodologiques pour de futures études**

- **Élargissement de la base de données :** Il est impératif d'augmenter la taille de l'échantillon pour valider la robustesse des méthodes avancées (Lasso, IV). Un échantillon de plusieurs milliers d'observations permettrait une segmentation plus fine (par type de bien ou micro-quartier).
- **Enrichissement des variables :** L'intégration de la performance énergétique (DPE) est devenue incontournable. Une future modélisation devrait impérativement inclure cette variable pour mesurer la valeur verte et son impact croissant sur les prix.
- **Prudence avec les modèles complexes :** Sur des échantillons de taille modeste, le principe de parcimonie doit prévaloir. Le modèle linéaire (OLS/MCO) bien spécifié reste souvent supérieur aux méthodes de Machine Learning ou aux estimateurs complexes qui risquent de capturer du bruit plutôt que du signal.

# 8 Annexes

## 8.1 Tableaux Complets

### 8.1.1 Modèle de régression linéaire simple

| **Variable** | **coef** | **std err** | **t** | **P-Value** | **[0.025, 0.975]** |
|---|---|---|---|---|---|
| *Constante* ($\beta_0$) | 1519,3743 | 34,584 | 43,932 | 0,000 | [1451,031, 1587,717] |
| *Surface* ($\beta_1$) | 5,0428 | 0,28 | 17,877 | 0,000 | [4,485, 5,600] |
| **R-squared:** | 0,683 | **Adj. R-squared:** | 0,680 | | |
| **F-statistic:** | 319,6 | **Prob (F-statistic):** | 8.45e-39 | | |
| **Log-Likelihood:** | -941,72 | **AIC:** | 1887,39 | | |
| **BIC:** | 1893,00 | **No. Observations:** | 150 | | |
| **Df Residuals:** | 148 | **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | | | |
| **Omnibus:** | 1,117 | **Durbin-Watson:** | 2,136 | | |
| **Prob(Omnibus):** | 0,572 | **Jarque-Bera (JB):** | 1,174 | | |
| **Skew:** | 0,131 | **Prob(JB):** | 0,555 | | |
| **Kurtosis:** | 2,655 | **Cond. No.:** | 400 | | |

Table 13: Modèle de régression linéaire simple

### 8.1.2 Modèle de régression linéaire multiple

| Variable | coef | std err | t | P-Value | [0.025, 0.975] |
|---|---|---|---|---|---|
| *Constante* ($\beta_0$) | -1679,49 | 1535,673 | -1,094 | 0,276 | [-4715,044 ; 1356,962] |
| *Surface* ($\beta_1$) | 4,39 | 0,28 | 15,01 | 0,000 | [3,846 ; 4,932] |
| *Chambres* ($\beta_2$) | 33,92 | 10,23 | 3,32 | 0,001 | [13,069 ; 54,770] |
| *Année de construction* ($\beta_3$) | 1,61 | 0,77 | 2,10 | 0,037 | [0,073 ; 3,148] |
| *Distance au centre* ($\beta_4$) | -6,14 | 0,99 | -6,19 | 0,000 | [-8,120 ; -4,162] |
| *Étage* ($\beta_5$) | 12,25 | 5,05 | 2,43 | 0,016 | [2,413 ; 22,087] |
| *Ascenseur* ($\beta_6$) | 55,51 | 17,92 | 3,10 | 0,002 | [20,676 ; 90,347] |
| **R-squared:** | 0,789 | **Adj. R-squared:** | 0,780 | | |
| **F-statistic:** | 84,3 | **Prob (F-statistic):** | 8.44e-38 | | |
| **Log-Likelihood:** | -911,43 | **AIC:** | 1887,39 | | |
| **BIC:** | 1858,00 | **No. Observations:** | 150 | | |
| **Df Residuals:** | 143 | **Df Model:** | 6 | | |
| **Omnibus:** | 0,955 | **Durbin-Watson:** | 2,136 | | |
| **Prob(Omnibus):** | 0,572 | **Jarque-Bera (JB):** | 1,174 | | |
| **Skew:** | 0,131 | **Prob(JB):** | 0,555 | | |
| **Kurtosis:** | 2,662 | **Cond. No.:** | 3.50e+05 | | |

Table 14: Résultats de la régression OLS

Modèle de régression linéaire multiple

### 8.1.3 Modèle Semi-Log Post-Covid

| Variable | coef | std err | t | P-Value | [0.025, 0.975] |
|---|---|---|---|---|---|
| Constante | 5.6515 | 0.532 | 10.633 | 0.000 | [4.600 ; 6.703] |
| Surface_m2 | 0.0022 | 0.000 | 14.616 | 0.000 | [0.002 ; 0.002] |
| Chambres | 0.0162 | 0.006 | 2.854 | 0.005 | [0.005 ; 0.027] |
| Année de construction | 0.0008 | 0.000 | 2.942 | 0.004 | [0.000 ; 0.001] |
| Distance au centre (km) | -0.0035 | 0.001 | -6.321 | 0.000 | [-0.005 ; -0.002] |
| Étage | 0.0050 | 0.003 | 1.811 | 0.072 | [-0.000 ; 0.010] |
| Ascenseur | 0.0243 | 0.010 | 2.536 | 0.012 | [0.005 ; 0.043] |
| Qualité des écoles | 0.0092 | 0.003 | 2.967 | 0.004 | [0.003 ; 0.015] |
| Revenu médian quartier | 0.0016 | 0.001 | 2.561 | 0.012 | [0.000 ; 0.003] |
| Surface_m2_COVID | -0.0003 | 0.000 | -1.490 | 0.139 | [-0.001 ; 0.0001] |
| Chambres_COVID | 0.0000 | 0.000 | -0.038 | 0.969 | [-0.014 ; 0.014] |
| Année_construction_COVID | 0.000058 | 0.000025 | 2.347 | 0.020 | [0.000009 ; 0.0001] |
| Distance_centre_km_COVID | -0.0001 | 0.001 | -0.181 | 0.856 | [-0.001 ; 0.001] |
| Étage_COVID | -0.0022 | 0.003 | -0.630 | 0.529 | [-0.009 ; 0.005] |
| Ascenseur_COVID | 0.0008 | 0.012 | 0.069 | 0.945 | [-0.024 ; 0.025] |
| Qualité_ecole_COVID | 0.0010 | 0.004 | 0.431 | 0.667 | [-0.006 ; 0.010] |
| Revenu_median_quartier_COVID | -0.0005 | 0.001 | -0.600 | 0.549 | [-0.002 ; 0.001] |
| **R-squared** | 0.907 | | | | |
| **Adj. R-squared** | 0.896 | | | | |
| **F-statistic** | 81.42 | | **Prob(F-stat)** | 1.71e-60 | |
| **Durbin-Watson** | 2.180 | | | | |

Table 15: Modèle Semi-Log Post-Covid

### 8.1.4 Effets des variables avant et après la période COVID

| Variable | Effet avant 2020 | Changement après 2020 | Effet après 2020 |
|---|---|---|---|
| Surface_m2 | 0.0022 | -0.0003 | 0.0019 |
| Chambres | 0.0162 | -0.0003 | 0.0159 |
| Année de construction | 0.0008 | 0.0001 | 0.0008 |
| Distance centre (km) | -0.0035 | -0.0001 | -0.0036 |
| Étage | 0.0050 | -0.0022 | 0.0028 |
| Ascenseur | 0.0243 | 0.0008 | 0.0252 |
| Qualité école | 0.0092 | 0.0018 | 0.0110 |
| Revenu médian quartier | 0.0016 | -0.0005 | 0.0011 |

Table 16: Effets des variables avant et après la période COVID

### 8.1.5 IV-2SLS

| IV-2SLS Estimation Summary | | | | | |
|---|---|---|---|---|---|
| **Dep. Variable:** | Prix_milliers_euros | **R-squared:** | 0.8798 | **Adj. R-squared:** | 0.8721 |
| **Estimator:** | IV-2SLS | **F-statistic:** | 1137.6 | **P-value (F-stat):** | 0.0000 |
| **No. Observations:** | 150 | **Distribution:** | chi2(9) | **Date:** | Sun, Dec 28 2025 |
| **Time:** | 18:16:57 | **Cov. Estimator:** | robust | | |
| **Parameter** | **Estimate** | **Std. Err.** | **T-stat** | **P-value** | **[Lower CI, Upper CI]** |
| *const* | -4.566e+04 | 6694.4 | -6.8214 | 0.0000 | [-5.879e+04, -3.254e+04] |
| *Surface_m2* | 4.2928 | 0.2587 | 16.5950 | 0.0000 | [3.7858, 4.7998] |
| *Chambres* | 33.229 | 7.5917 | 4.3770 | 0.0000 | [18.350, 48.109] |
| *Année_construction* | 1.6250 | 0.5828 | 2.7883 | 0.0053 | [0.4828, 2.7673] |
| *Distance_centre_km* | -7.1679 | 0.7671 | -9.3445 | 0.0000 | [-8.6713, -5.6645] |
| *Etage* | 8.0295 | 3.8934 | 2.0623 | 0.0392 | [0.3986, 15.660] |
| *Ascenseur* | 55.584 | 13.002 | 4.2750 | 0.0000 | [30.100, 81.067] |
| *Revenu_median_quartier* | 4.9220 | 1.7207 | 2.8604 | 0.0042 | [1.5494, 8.2946] |
| *Année_vente* | 21.621 | 3.1895 | 6.7787 | 0.0000 | [15.369, 27.872] |
| *Qualite_ecole* | 2.0931 | 13.183 | 0.1588 | 0.8738 | [-23.744, 27.930] |
| **Endogenous:** | Qualite_ecole | **Instruments:** | Distance_universite | | |
| **Robust Covariance:** | Heteroskedastic | **Debiased:** | False | | |

Table 17: IV-2SLS

### 8.1.6 F-Statistic

| First Stage Estimation Results | | |
|---|---|---|
| **Variable** | **Estimate** | **T-stat** |
| **R-squared** | 0.4419 | |
| **Partial R-squared** | 0.1180 | |
| **Shea's R-squared** | 0.1180 | |
| **Partial F-statistic** | 20.911 | |
| **P-value (Partial F-stat)** | 4.812e-06 | |
| **Partial F-stat Distn** | chi2(1) | |
| *const* | -44.594 | (-0.4411) |
| *Surface_m2* | 0.0027 | (0.6544) |
| *Chambres* | -0.1380 | (-1.0315) |
| *Année_construction* | -0.0041 | (-0.4623) |
| *Distance_centre_km* | -0.0083 | (-0.6734) |
| *Etage* | -0.0342 | (-0.4708) |
| *Ascenseur* | 0.0056 | (0.0247) |
| *Revenu_median_quartier* | 0.1039 | (7.5440) |
| *Année_vente* | 0.0263 | (0.5380) |
| *Distance_universite* | -0.1447 | (-4.5728) |

Table 18: First Stage Estimation Results

### 8.1.7 Modèle Ridge

| **Variable** | $\lambda = 0{,}001$ | $\lambda = 0{,}01$ | $\lambda = 0{,}1$ | $\lambda = 1$ | $\lambda = 10$ | $\lambda = 50$ | $\lambda = 100$ |
|---|---|---|---|---|---|---|---|
| Surface_m2 | 0.0819 | 0.0819 | 0.0815 | 0.0785 | 0.0684 | 0.0524 | 0.0423 |
| Chambres | 0.0174 | 0.0174 | 0.0175 | 0.0185 | 0.0205 | 0.0217 | 0.0208 |
| Année de construction | 0.0090 | 0.0090 | 0.0090 | 0.0085 | 0.0070 | 0.0050 | 0.0039 |
| Distance au centre (km) | -0.0310 | -0.0311 | -0.0311 | -0.0315 | -0.0292 | -0.0224 | -0.0183 |
| Étage | 0.0087 | 0.0087 | 0.0087 | 0.0085 | 0.0070 | 0.0050 | 0.0040 |
| Ascenseur | 0.0121 | 0.0121 | 0.0121 | 0.0118 | 0.0087 | 0.0050 | 0.0035 |
| Qualité des écoles | 0.0171 | 0.0171 | 0.0172 | 0.0171 | 0.0143 | 0.0119 | 0.0104 |
| Revenu médian quartier | 0.0147 | 0.0147 | 0.0142 | 0.0118 | 0.0100 | 0.0091 | 0.0080 |
| Surface_m2_COVID | -0.0191 | -0.0190 | -0.0181 | -0.0122 | 0.0037 | 0.0103 | 0.0102 |
| Chambres_COVID | -0.0004 | -0.0005 | -0.0007 | -0.0017 | -0.0018 | 0.0030 | 0.0050 |
| Année_construction_COVID | 0.0569 | 0.0565 | 0.0534 | 0.0359 | 0.0100 | 0.0024 | 0.0016 |
| Distance_centre_km_COVID | -0.0014 | -0.0014 | -0.0013 | -0.0008 | -0.0025 | -0.0058 | -0.0057 |
| Étage_COVID | -0.0042 | -0.0042 | -0.0042 | -0.0037 | -0.0014 | 0.0005 | 0.0011 |
| Ascenseur_COVID | 0.0004 | 0.0004 | 0.0004 | 0.0007 | 0.0031 | 0.0045 | 0.0043 |
| Qualité_ecole_COVID | 0.0054 | 0.0054 | 0.0051 | 0.0044 | 0.0086 | 0.0087 | 0.0077 |
| Revenu_median_quartier_COVID | -0.0161 | -0.0158 | -0.0130 | 0.0001 | 0.0065 | 0.0042 | 0.0035 |

Table 19: Évolution des coefficients estimés par la régression Ridge (semi-log + COVID) selon le paramètre de régularisation $\lambda$

### 8.1.8 Modèle Lasso

| **Variable** | $\lambda = 0{,}001$ | $\lambda = 0{,}01$ | $\lambda = 0{,}1$ | $\lambda = 1$ | $\lambda = 10$ | $\lambda = 50$ | $\lambda = 100$ |
|---|---|---|---|---|---|---|---|
| Surface_m2 | 0.075093 | 0.068718 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Chambres | 0.017050 | 0.012851 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Année de construction | 0.006761 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Distance au centre (km) | -0.031077 | -0.020626 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Étage | 0.005407 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Ascenseur | 0.010204 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Qualité des écoles | 0.015133 | 0.004750 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Revenu médian quartier | 0.010842 | 0.005461 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Surface_m2_COVID | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Chambres_COVID | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Année_construction_COVID | 0.017735 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Distance_centre_km_COVID | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Étage_COVID | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Ascenseur_COVID | 0.000790 | 0.000124 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Qualité_ecole_COVID | 0.006888 | 0.019898 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Revenu_median_quartier_COVID | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Table 20: Coefficients estimés par la régression Lasso (semi-log + COVID) selon le paramètre de régularisation $\lambda$

## 8.2 Code Python

### Tableau des Variables

```
import pandas as pd
from spipy.stats import skew, kurtosis
df = pd.read_excel("donnes_immob.xlsx")
variables = df.select_dtypes(include="number")
stats = pd.DataFrame({
   "Moyenne␣(X)": variables.mean(),
   "Mediane": variables.median(),
   "Ecart-type␣(sX)": variables.std(),
   "Minimum": variables.min(),
   "Q1␣(25%)": variables.quantile(0.25),
   "Q2␣(50%)": variables.quantile(0.50),
   "Q3␣(75%)": variables.quantile(0.75),
   "Maximum": variables.max(),
   " Asymtrie ␣(Skewness)": variables.apply(skew),
   "Aplatissement␣(Kurtosis)": variables.apply(kurtosis)
})
print(stats)
```

### Histogrammes et Boxplot

```
import pandas as pd
import matplotlib.pyplot as plt
import os

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

variables = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Revenu_median_quartier",
    "Etage",
    "Prix_milliers_euros",
    "Annee_vente",
    "Qualite_ecole"
]

os.makedirs("figures", exist_ok=True)

# Histogrammes
for var in variables:
    plt.figure()
    plt.hist(df[var].dropna(), bins=20)
    plt.title(f"Histogramme␣de␣{var}")
    plt.xlabel(var)
    plt.ylabel("Frequence")
    plt.savefig(f"figures/histogramme_{var}.png", dpi=300, bbox_inches="tight")
    plt.close()

# Boxplots
for var in variables:
    plt.figure()
    plt.boxplot(df[var].dropna(), vert=True)
    plt.title(f"Boite␣a␣moustaches␣de␣{var}")
    plt.ylabel(var)
    plt.savefig(f"figures/boxplot_{var}.png", dpi=300, bbox_inches="tight")
    plt.close()
```

Listing 1: Génération des histogrammes et boxplots des variables

## Matrice de Corrélation

```
import pandas as pd

df = pd.read_excel("donnees_immobilieres_extended.xlsx")
variables_continues = df[
   [
   "Surface_m2",
   "Chambres",
   "Annee_construction",
   "Distance_centre_km",
   "Revenu_median_quartier",
   "Etage"
   "Prix_milliers_euros",
   "Annee_vente",
   "Qualite_ecole"
   ]
]
matrice_correlation = variables_continues.corr()
print(matrice_correlation)
```

## Heat Map

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

df = pd.read_excel("donnees_immobilieres_extended.xlsx")
variables_continues = df[
   [
"Surface_m2",
   "Chambres",
   "Annee_construction",
   "Distance_centre_km",
   "Revenu_median_quartier",
   "Etage"
   "Prix_milliers_euros",
   "Annee_vente",
   "Qualite_ecole"
   ]
]
corr = variables_continues.corr()

os.makedirs("figures", exist_ok=True)

plt.figure(figsize=(8, 6))
sns.heatmap(
   corr,
   annot=True,
   fmt=".2f",
   cmap="coolwarm",
   linewidths=0.5
)
plt.tight_layout()
plt.savefig("figures/heatmap_correlation.png", dpi=300, bbox_inches="tight")
plt.show()
```

## Régression Linéaire Simple

```
import pandas as pd
import statsmodels.api as sm
import os
df = pd.read_excel("donnees_immobilieres_extended.xlsx")
```

```
Y = df["Prix_milliers_euros"]
X = df["Surface_m2"]

X = sm.add_constant(X)
model = sm.OLS(Y, X)
results = model.fit()
print(results.summary())
with open("resultats_regression.txt", "w") as f:
   f.write(results.summary().as_text())
```

## Régression Linéaire Multiple

```
import pandas as pd
import statsmodels.api as sm

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

variables = [
   "Prix_milliers_euros",
   "Surface_m2",
   "Chambres",
   "Annee_construction",
   "Distance_centre_km",
   "Etage",
   "Ascenseur"
]

df = df[variables].dropna()

Y = df["Prix_milliers_euros"]
X = df[
   [
       "Surface_m2",
       "Chambres",
       "Annee_construction",
       "Distance_centre_km",
       "Etage",
       "Ascenseur"
   ]
]
X = sm.add_constant(X)

model_multi = sm.OLS(Y, X)
results_multi = model_multi.fit()
print(results_multi.summary())
```

## Modèles linéaire, log-log, semi-log

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = df["Prix_milliers_euros"]
X = df[[
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur"
]]
```

```
X = sm.add_constant(X)

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

model_semilog = sm.OLS(np.log(y), X).fit()

X_loglog = X.copy()
X_loglog["Surface_m2"] = np.log(df["Surface_m2"])
X_loglog["Chambres"] = np.log(df["Chambres"])
X_loglog["Distance_centre_km"] = np.log(df["Distance_centre_km"])

model_loglog = sm.OLS(np.log(y), X_loglog).fit()

results = pd.DataFrame({
    "R2": [
        model_linear.rsquared,
        model_semilog.rsquared,
        model_loglog.rsquared
    ],
    "R2_adjusted": [
        model_linear.rsquared_adj,
        model_semilog.rsquared_adj,
        model_loglog.rsquared_adj
    ]
}, index=["Linaire", "Semi-log", "Log-log"])

print(results)
```

## Variance Inflation Factors (VIF)

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Chargement du dataset
df = pd.read_excel("donnees_immobilieres_extended.xlsx")

# Variables explicatives du modle semi-log
X = df[[
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur"
]]

# Ajout de la constante
X = sm.add_constant(X)

# Calcul des VIF
vif_df = pd.DataFrame()
vif_df["Variable"] = X.columns
vif_df["VIF"] = [
    variance_inflation_factor(X.values, i)
    for i in range(X.shape[1])
]

print(vif_df)
```

## Hypothèse des coefficients nuls

```
import pandas as pd
import numpy as np
```

```
import statsmodels.api as sm

df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = np.log(df["Prix_milliers_euros"])
vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur"
]

X = sm.add_constant(df[vars_base])

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

print("Test global (H0 : tous les coefficients sauf constante = 0)")
print("F =", model_base.fvalue)
print("p-value =", model_base.f_pvalue)
print("R2 ajust =", model_base.rsquared_adj)
```

## Hypothèse effet négatif de la distance au centre

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy import stats


df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = np.log(df["Prix_milliers_euros"])
X = df[[
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur"
]]

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())

coef = model.params["Distance_centre_km"]
t_stat = model.tvalues["Distance_centre_km"]
p_value_two_sided = model.pvalues["Distance_centre_km"]

if coef < 0:
    p_value_one_sided = p_value_two_sided / 2
else:
    p_value_one_sided = 1 - (p_value_two_sided / 2)

print("Coefficient Distance_centre_km :", coef)
print("Statistique t :", t_stat)
print("p-value bilatrale :", p_value_two_sided)
print("p-value unilatrale (effet ngatif) :", p_value_one_sided)
```

## Ajout variables: Qualite-ecole et Revenu-Median-quartier

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
```

```
df = pd.read_excel("donnees_immobilieres_extended.xlsx")

# Variable dpendante : semi-log
y = np.log(df["Prix_milliers_euros"])

#  Modle  de base
X_base = df[["Surface_m2","Chambres","Annee_construction","Distance_centre_km","Etage","
    Ascenseur"]]
X_base = sm.add_constant(X_base)
m_base = sm.OLS(y, X_base).fit()

print("Test global (H0: tous les coefficients sauf constante = 0)")
print("F =", m_base.fvalue)
print("p-value =", m_base.f_pvalue)
print("R2 ajust =", m_base.rsquared_adj)

#  Modle  tendu
X_ext = df[["Surface_m2","Chambres","Annee_construction","Distance_centre_km","Etage","
    Ascenseur",
          "Qualite_ecole","Revenu_median_quartier"]]
X_ext = sm.add_constant(X_ext)
m_ext = sm.OLS(y, X_ext).fit()

print("\nModle  tendu  : R2 ajust =", m_ext.rsquared_adj)

# Test F d'ajout conjoint de Qualite_ecole et Revenu_median_quartier
F_stat, p_val, df_diff = m_ext.compare_f_test(m_base)
print("\nTest d'ajout conjoint (H0: coefficients Qualite_ecole = Revenu_median_quartier = 0)
    ")
print("F =", F_stat)
print("p-value =", p_val)
print("df diff =", df_diff)
```

## Stabilité structurelle

```
import pandas as pd
import numpy as np
import statsmodels.api as sm


df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = np.log(df["Prix_milliers_euros"])

df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)

vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X_base = sm.add_constant(df[vars_base])

model_base = sm.OLS(y, X_base).fit()

X_break = X_base.copy()
for var in vars_base:
    X_break[f"{var}_COVID"] = df["COVID"] * df[var]

model_break = sm.OLS(y, X_break).fit()
```

```
F_stat, p_value, df_diff = model_break.compare_f_test(model_base)

print("Statistique F :", F_stat)
print("p-value :", p_value)
print("Nombre de restrictions testes :", df_diff)
```

## Modèle Semi-Log Post Covid

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)

vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X = sm.add_constant(df[vars_base])

for var in vars_base:
    X[f"{var}_COVID"] = df["COVID"] * df[var]

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

print(model.summary())

results = []

for var in vars_base:
    beta_before = model.params[var]
    gamma_change = model.params[f"{var}_COVID"]
    beta_after = beta_before + gamma_change

    results.append({
        "Variable": var,
        "Effet_avant_2020": beta_before,
        "Changement_apres_2020": gamma_change,
        "Effet_apres_2020": beta_after
    })

effects_df = pd.DataFrame(results)

print("\nEffets␣avant␣/␣aprs␣COVID␣:")
print(effects_df.round(4))
```

## Observation des résidus

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

```
df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = np.log(df["Prix_milliers_euros"])
X = df[[
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]]

X = sm.add_constant(X)

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

residuals = model.resid
fitted = model.fittedvalues

plt.figure(figsize=(7, 5))
plt.scatter(fitted, residuals, alpha=0.6)
plt.axhline(0, color="black", linestyle="--")
plt.xlabel("Valeurs ajustes")
plt.ylabel("Rsidus")
plt.title("Rsidus en fonction des valeurs ajustes (modle sans COVID)")
plt.show()
```

## Hétéroscédasticité

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.stats.diagnostic import het_breuschpagan, het_white
from statsmodels.stats.stattools import durbin_watson
df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

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

X = sm.add_constant(X)

model_ols = sm.OLS(y, X).fit()
print("\n=== MCO standard ===")
print(model_ols.summary())

bp_stat, bp_pvalue, _, _ = het_breuschpagan(
    model_ols.resid,
    model_ols.model.exog
)

print("\nTest de BreuschPagan")
print("Statistique BP :", bp_stat)
```

```
print("p-value :", bp_pvalue)

white_stat, white_pvalue, _, _ = het_white(
    model_ols.resid,
    model_ols.model.exog
)

print("\nTest de White")
print("Statistique White :", white_stat)
print("p-value :", white_pvalue)

model_robust = model_ols.get_robustcov_results(cov_type="HC1")

print("\n=== MCO avec carts-types robustes (HC1) ===")
print(model_robust.summary())

weights = 1 / (model_ols.resid ** 2)

model_wls = sm.WLS(y, X, weights=weights).fit()

print("\n=== Moindres carrs  pondrs  (WLS) ===")
print(model_wls.summary())
```

## Comparaison MCO standards, MCO avec écart-types robustes, WLS

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = np.log(df["Prix_milliers_euros"])
df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)
vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X = sm.add_constant(df[vars_base])

for var in vars_base:
    X[f"{var}_COVID"] = df["COVID"] * df[var]

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

model_robust = model_ols.get_robustcov_results(cov_type="HC0")

weights = 1 / (model_ols.resid ** 2)
model_wls = sm.WLS(y, X, weights=weights).fit()

comparison = pd.DataFrame({
    "Coef_MCO": model_ols.params,
    "SE_MCO": model_ols.bse,
    "SE_Robustes": model_robust.bse,
    "Coef_WLS": model_wls.params
})

print("\n=== Comparaison MCO / Robust / WLS ===")
print(comparison.round(4))
```

## Autocorrélation

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.stats.stattools import durbin_watson

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

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

X = sm.add_constant(X)

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

dw_stat = durbin_watson(model.resid)

print("Test de DurbinWatson")
print("Statistique DW :", dw_stat)
```

## Estimation en deux étapes (2SLS)

```
import pandas as pd
import statsmodels.api as sm
from linearmodels.iv import IV2SLS

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

y = df["Prix_milliers_euros"]
exog = df[
   [
       "Surface_m2",
       "Chambres",
       "Annee_construction",
       "Distance_centre_km",
       "Etage",
       "Ascenseur",
       "Revenu_median_quartier",
       "Annee_vente"
   ]
]
exog = sm.add_constant(exog)

endog = df["Qualite_ecole"]
instrument = df["Distance_universite"]

iv_model = IV2SLS(
   dependent=y,
   exog=exog,
   endog=endog,
```

```
   instruments=instrument
)

iv_results = iv_model.fit(cov_type="robust")

print(iv_results.summary)
```

## F-Statistic

```
import pandas as pd
import statsmodels.api as sm
from linearmodels.iv import IV2SLS


df = pd.read_excel("donnees_immobilieres_extended.xlsx")
y = df["Prix_milliers_euros"]
exog = df[["Surface_m2", "Chambres", "Annee_construction", "Distance_centre_km", "Etage", "
     Ascenseur", "Revenu_median_quartier", "Annee_vente"]]

exog = sm.add_constant(exog)
endog = df["Qualite_ecole"]
instrument = df["Distance_universite"]

iv_model = IV2SLS(
   dependent=y,
   exog=exog,
   endog=endog,
   instruments=instrument
)
iv_results = iv_model.fit(cov_type="robust")

print(iv_results.first_stage)
```

## Estimation par Moindres Carrés Ordinaires (MCO)

```
import pandas as pd
import statsmodels.api as sm
from linearmodels.iv import IV2SLS

df = pd.read_excel("donnees_immobilieres_extended.xlsx")


y = df["Prix_milliers_euros"]

X_mco = df[
   [
       "Surface_m2",
       "Chambres",
       "Annee_construction",
       "Annee_vente",
       "Distance_centre_km",
       "Etage",
       "Ascenseur",
       "Revenu_median_quartier",
       "Qualite_ecole"
   ]
]

X_mco = sm.add_constant(X_mco)

mco_results = sm.OLS(y, X_mco).fit(cov_type="HC1")

print("=====␣RSULTATS␣MCO␣=====")
```

```
print(mco_results.summary())
```

## Comparaison des Coefficients MCO vs IV (2SLS)

```
comparison = pd.DataFrame({
   "MCO": mco_results.params,
   "IV␣(2SLS)": iv_results.params
})

print("\n=====␣COMPARAISON␣DES␣COEFFICIENTS␣=====")
print(comparison)
```

## Ridge

```
import pandas as pd
import numpy as np
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)

vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X_base = df[vars_base]

X_covid = X_base.mul(df["COVID"], axis=0)
X_covid.columns = [f"{col}_COVID" for col in X_covid.columns]

X = pd.concat([X_base, X_covid], axis=1)

scaler = StandardScaler()
X_std = scaler.fit_transform(X)

lambdas = [0.001, 0.01, 0.1, 1, 10, 50, 100]

coef_ridge = pd.DataFrame(index=X.columns)

for l in lambdas:
    ridge = Ridge(alpha=l)
    ridge.fit(X_std, y)
    coef_ridge[f"lambda={l}"] = ridge.coef_

print(coef_ridge.round(4))
```

## Lasso

```
import pandas as pd
import numpy as np
```

```
from sklearn.linear_model import Lasso
from sklearn.preprocessing import StandardScaler

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)

vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X = df[vars_base].copy()
for var in vars_base:
    X[f"{var}_COVID"] = df[var] * df["COVID"]

scaler = StandardScaler()
X_std = scaler.fit_transform(X)

lambdas = [0.001, 0.01, 0.1, 1, 10, 50, 100]

coef_lasso = pd.DataFrame(index=X.columns)

for l in lambdas:
    lasso = Lasso(alpha=l, max_iter=10000)
    lasso.fit(X_std, y)
    coef_lasso[f"lambda={l}"] = lasso.coef_

print(coef_lasso)
```

## La validation croisée 10-fold

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LassoCV
from sklearn.preprocessing import StandardScaler

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)

vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X = df[vars_base].copy()

for var in vars_base:
    X[f"{var}_COVID"] = df[var] * df["COVID"]
```

```
scaler = StandardScaler()
X_std = scaler.fit_transform(X)

alphas = np.logspace(-3, 2, 50)

lasso_cv = LassoCV(
    alphas=alphas,
    cv=10,
    max_iter=10000,
    random_state=0
)

lasso_cv.fit(X_std, y)

lambda_opt = lasso_cv.alpha_
print("Lambda optimal :", lambda_opt)

coef_lasso_cv = pd.Series(
    lasso_cv.coef_,
    index=X.columns
)

print("\nCoefficients pour lambda optimal :")
print(coef_lasso_cv)
```

## Comparaison des trois modèles

```
import pandas as pd
import numpy as np
from sklearn.linear_model import Ridge, Lasso
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm

df = pd.read_excel("donnees_immobilieres_extended.xlsx")

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

df["COVID"] = (df["Annee_vente"] >= 2020).astype(int)

vars_base = [
    "Surface_m2",
    "Chambres",
    "Annee_construction",
    "Distance_centre_km",
    "Etage",
    "Ascenseur",
    "Qualite_ecole",
    "Revenu_median_quartier"
]

X_base = df[vars_base]

X_covid = X_base.mul(df["COVID"], axis=0)
X_covid.columns = [f"{col}_COVID" for col in X_covid.columns]

X = pd.concat([X_base, X_covid], axis=1)

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

X_train_ols = sm.add_constant(X_train)
X_test_ols = sm.add_constant(X_test)
```

```
ols_model = sm.OLS(y_train, X_train_ols).fit()
y_pred_ols = ols_model.predict(X_test_ols)
rmse_ols = np.sqrt(mean_squared_error(y_test, y_pred_ols))

scaler = StandardScaler()
X_train_std = scaler.fit_transform(X_train)
X_test_std = scaler.transform(X_test)

ridge = Ridge(alpha=2.9471)
ridge.fit(X_train_std, y_train)
y_pred_ridge = ridge.predict(X_test_std)
rmse_ridge = np.sqrt(mean_squared_error(y_test, y_pred_ridge))

lasso = Lasso(alpha=0.001)
lasso.fit(X_train_std, y_train)
y_pred_lasso = lasso.predict(X_test_std)
rmse_lasso = np.sqrt(mean_squared_error(y_test, y_pred_lasso))

print("RMSE␣(semi-log␣+␣COVID)␣:")
print(f"MCO␣␣␣:␣{rmse_ols:.4f}")
print(f"Ridge␣:␣{rmse_ridge:.4f}")
print(f"Lasso␣:␣{rmse_lasso:.4f}")
```

Listing 2: Comparaison MCO, Ridge et Lasso avec division Train/Test et RMSE

## Prédiction ponctuelle et intervalle de confiance

```
import numpy as np
import pandas as pd

maison = {
    "Constante": 1,
    "Surface_m2": 120,
    "Chambres": 3,
    "Annee_construction": 2015,
    "Distance_centre_km": 5,
    "Etage": 1,
    "Ascenseur": 1,
    "Qualite_ecole": 7,
    "Revenu_median_quartier": 65,
    "COVID": 1 # Vente aprs 2020
}

coefs = {
    "Constante": 5.6515,
    "Surface_m2": 0.0022,
    "Chambres": 0.0162,
    "Annee_construction": 0.0008,
    "Distance_centre_km": -0.0035,
    "Etage": 0.0050,
    "Ascenseur": 0.0243,
    "Qualite_ecole": 0.0092,
    "Revenu_median_quartier": 0.0016,

    "Surface_m2_COVID": -0.0003,
    "Chambres_COVID": 0.0000,
    "Annee_construction_COVID": 0.000058,
    "Distance_centre_km_COVID": -0.0001,
    "Etage_COVID": -0.0022,
    "Ascenseur_COVID": 0.0008,
    "Qualite_ecole_COVID": 0.0010,
    "Revenu_median_quartier_COVID": -0.0005
}

RMSE = 0.0413
```

```
log_prix_reporte = 7.752
SE_mean = (7.768 - 7.752) / 1.96

print("--- 1. PRDICTION PONCTUELLE ---")

log_pred_calc = 0
for var, val in maison.items():
    if var == "COVID": continue
    log_pred_calc += val * coefs.get(var, 0)
    if maison["COVID"] == 1:
        log_pred_calc += val * coefs.get(f"{var}_COVID", 0)

print(f"Log(Prix) calcul avec les coefficients arrondis : {log_pred_calc:.4f}")
print(f"Log(Prix) retenu (valeur exacte du logiciel) : {log_prix_reporte:.4f}")

facteur_correction = np.exp(RMSE**2 / 2)
prix_median = np.exp(log_prix_reporte)
prix_moyen_corrige = prix_median * facteur_correction

print(f"\nFacteur de correction (Biais Jensen) : {facteur_correction:.6f} (+{(
     facteur_correction-1)*100:.2f}%)")
print(f"Prix Mdian (exp simple)          : {prix_median:.0f} k")
print(f"Prix Moyen Corrig (FINAL)        : {prix_moyen_corrige:.0f} k")


print("\n--- 2. INTERVALLES ---")

borne_inf_log_IC = log_prix_reporte - 1.96 * SE_mean
borne_sup_log_IC = log_prix_reporte + 1.96 * SE_mean

ic_inf = np.exp(borne_inf_log_IC)
ic_sup = np.exp(borne_sup_log_IC)

print(f"Intervalle de Confiance (95%) [Moyenne thorique] :")
print(f" Log : [{borne_inf_log_IC:.3f} ; {borne_sup_log_IC:.3f}]")
print(f" Euros : [{ic_inf:.0f} k ; {ic_sup:.0f}  k ]")

SE_prediction = np.sqrt(SE_mean**2 + RMSE**2)

borne_inf_log_IP = log_prix_reporte - 1.96 * SE_prediction
borne_sup_log_IP = log_prix_reporte + 1.96 * SE_prediction

ip_inf = np.exp(borne_inf_log_IP)
ip_sup = np.exp(borne_sup_log_IP)

print(f"\nIntervalle de Prdiction (95%) :")
print(f" Log : [{borne_inf_log_IP:.3f} ; {borne_sup_log_IP:.3f}]")
print(f" Euros : [{ip_inf:.0f} k ; {ip_sup:.0f}  k ]")
```

## Graphiques

- L'année de construction

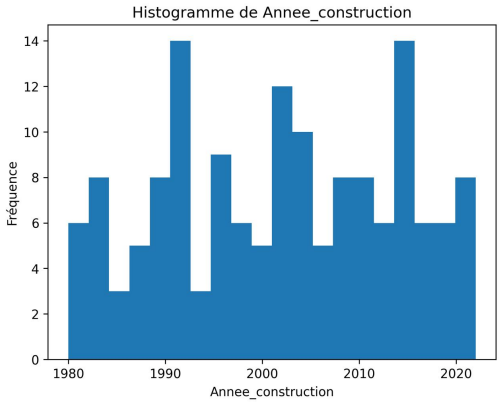


Figure 12: Histogramme de *Annee_construction*

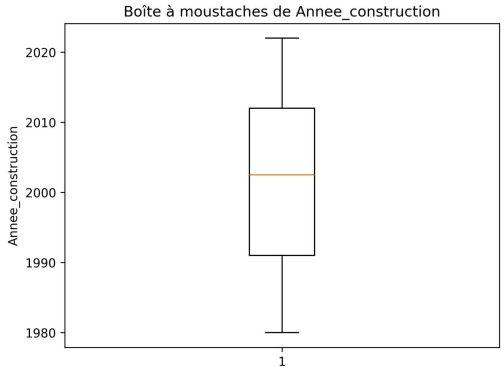


Figure 13: Boîte à moustaches de *Annee_construction*

- L'année de vente

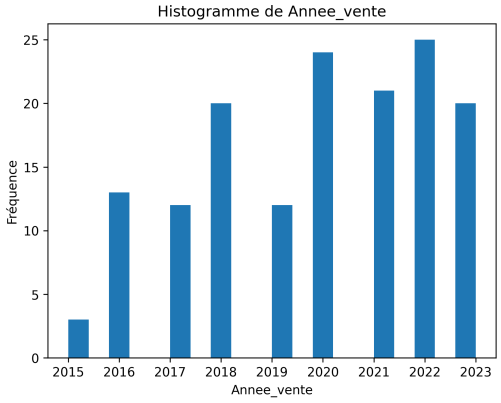


Figure 14: Histogramme de *Annee_vente*

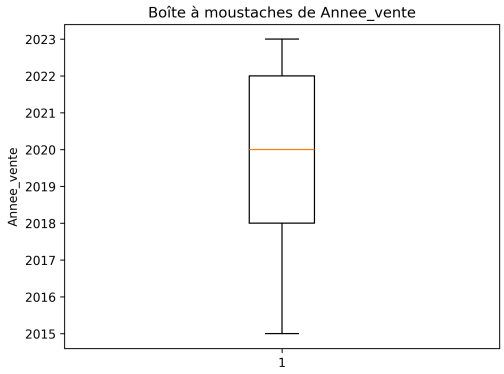


Figure 15: Boîte à moustaches de *Annee_vente*

- Le nombre de chambres

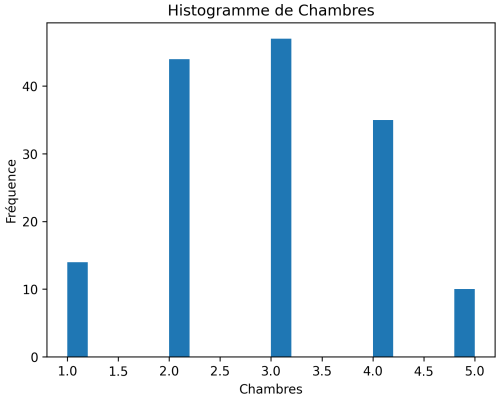


Figure 16: Histogramme de *Chambres*

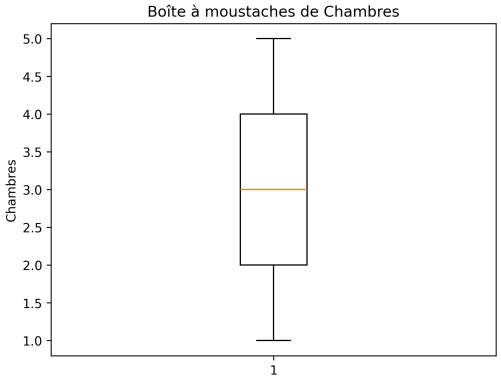


Figure 17: Boîte à moustaches de *Chambres*

- L'Etage

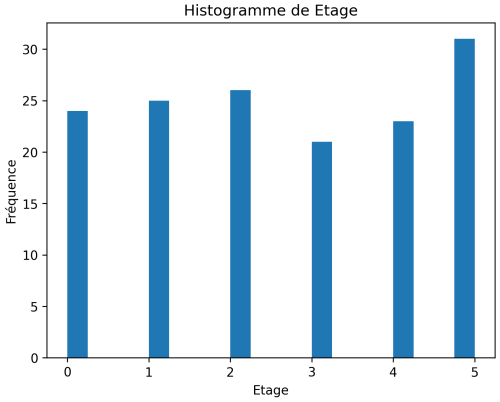


Figure 18: Histogramme de *Etage*

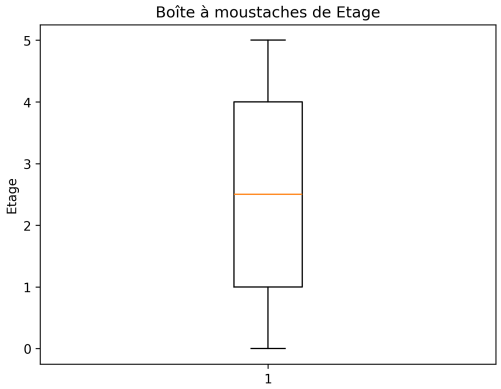


Figure 19: Boîte à moustaches de *Etage*

- Qualité Ecole

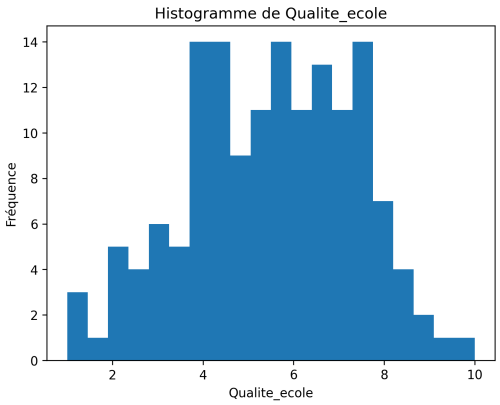


Figure 20: Histogramme de *Qualite_ecole*

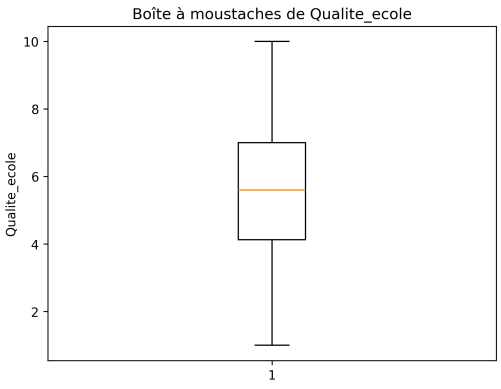


Figure 21: Boîte à moustaches de *Qualite_ecole*