AIX-MARSEILLE UNIVERSITÉ
FACULTÉ D'ECONOMIE ET DE GESTION

LA MÉTHODE DES PRIX HÉDONIQUES APPLIQUÉE AU MARCHÉ DE L'IMMOBILIER À MELBOURNE

PIERRE DUMONTEL - VICTOR PION

MASTER 1 - AUTOMNE 2020

Table des matières

I. Introduction

Dans son article "*Hedonic Prices and Implicit Markets : Product Differentation in Pure Competition*", Sherwin Rosen (1974) définit la notion de prix "hédonique". Il pose l'hypothèse fondatrice qu'un agent considère un bien en fonction de l'utilité que peut lui apporter chacune de ses caractéristiques. Dans le cadre de biens hétérogènes et différentiables (logements, lacs, emplois…), chaque bien peut être alors décrit par un vecteur de ses caractéristiques observables. Le prix d'un tel bien dépend alors de la quantité de chaque caractéristique, on parle de prix hédonique.

Alors que le prix observé d'un produit dépend des fonctions d'offre et de demande du marché, la notion de prix hédonique de Rosen (1974) permet de détailler individuellement chaque caractéristique du bien en question ainsi que leur prix implicite. Il est alors possible de donner une valeur marchande à des attributs qui ne possèdent pas de marché explicite comme la qualité de l'air, un boîtier de vitesse automatique ou encore le nombre de places de parking d'une résidence.

La définition de Rosen (1974) étant très générale, son application a été faite dans des domaines variés tels que la valorisation des biens communs, l'économie du travail ou encore le marché immobilier. Cette étude se concentre sur ce dernier domaine. En effet, le marché immobilier concentre une multitude d'agents (locataires, propriétaires, investisseurs, promoteurs…) et une estimation précise du prix d'un logement en fonction de ses caractéristiques (nombre de pièces, proximité au centre ville, superficie du jardin…) constitue une donnée précieuse pour chacun d'entre eux. De nombreux articles académiques ont déjà utilisé la méthode des prix hédoniques de Rosen sur un marché immobilier pour déterminer le prix d'un logement et notre étude utilise leurs conclusions et résultats pour appuyer les nôtres.

Dans son étude, Rosen (1974) explique que les prix hédoniques sont les prix implicites des caractéristiques des biens. Il utilise ces prix implicites comme base empirique pour son modèle. Pour les obtenir, il effectue une régression dite de "première étape" du prix du produit en fonction de ses caractéristiques. Notre étude sera centrée sur cette étape. Dans le cadre du marché immobilier de Melbourne, nous allons régresser le prix d'un logement en fonction de ses caractéristiques. On considère comme logement toute habitation qui répond à la définition de l'INSEE : "un local utilisé pour l'habitation, séparé et indépendant". Notre étude prend en compte toutes les catégories de logements (résidences principales, secondaires, logements vacants…) quelle que soit l'utilisation qui en est faite.

Dans leur article "*House Price Prediction : Hedonic Price Model vs. Artificial Network*", Limsombunchain, Gan et Lee (2004) reconnaissent que l'utilisation d'un modèle de prix hédoniques sur un marché immobilier entraîne souvent des problèmes tels que la multicolinéarité des régresseurs, l'hétéroscédasticité des résidus ou encore la mauvaise spécification du modèle. Notre étude risque de se confronter à ces problèmes, nous essaierons d'en comprendre les causes et si possible de les solutionner pour rendre notre modèle le plus robuste possible.

II. Le modèle économique

Des différents articles académiques que nous avons consultés pour cette étude, on retient 3 catégories de variables déterminantes dans le prix du logement :

- Les caractéristiques privées du logement décrivent le logement en tant que tel
- Les variables environnementales décrivent l'environnement direct autour du logement
- Les variables de localisation décrivant la position du logement

Catégories	Exemples de variables
Caractéristiques privées du logement	Superficie de la surface habitable, du terrain… Nombre de pièces, salles de bains, chambres, places de parking, garages, balcon… Présence d'une piscine, sous-sol... Âge du logement Type de logement
Variables environnementales	Nombre de parcs à proximité, arrêts de bus, monuments historiques, musées… Taux de criminalité du quartier, taux d'échec scolaire...
Variables de localisation du logement	Distance au centre ville, autoroute, mer,… Nom du quartier

On cherche à estimer le modèle suivant :

$$Y = Xb + u$$

Où Y le prix observé du logement, X la matrice des variables explicatives, b les vecteurs des estimateurs de X et u le terme d'erreur.

Le tableau ci-dessous résume le signe attendu des variables en tant que déterminants du prix du logement. L'annexe n°1 présente la table de corrélation des variables, les corrélations des variables avec le prix du logement induisent le signe attendu par les variables explicatives.

Variables	**Signe attendu**	**Justifications**
Nombre de pièces, chambres, places de parking, salles de bains, chambres, balcons...	+	Un logement disposant de plus de pièces et de commodités peut accueillir plus d'individus justifiant un prix plus élevé
Superficie de la surface habitable, du terrain	+	De la même façon, la superficie d'un logement est valorisée car elle permet d'accueillir plus d'individus, de s'éloigner de ses voisins dans le cadre d'une maison...
Age du logement	-	Un logement plus âgé peut nécessiter des travaux, représentant un coût supplémentaire à l'achat
Type de logement	+	Une maison est généralement plus spacieuse qu'un appartement
Distance à un lieu d'intérêt	-/+	En fonction du lieu en question, sa proximité peut être recherchée ou non (proximité à une usine vs. proximité d'un supermarché)
Localisation	-/+	La segmentation géographique du marché immobilier induit des différences de prix. Notamment selon la valorisation de la localisation d'un logement

III. Les données

Les données utilisées pour notre étude proviennent du site Kaggle. La base de données a été construite à partir des informations de ventes de logements disponibles sur le site domain.com.au entre mars 2016 et décembre 2018 à Melbourne en Australie. A l'exception de la variable **Bedroom2**, provenant de sources différentes et inconnues.

La base de données originale est composée de 13 580 observations, qui correspondent à 13 580 logements situés à Melbourne et sa périphérie. Chaque logement est défini selon 21 caractéristiques distinctes, listées ci dessous :

Nom d'origine	Nom dans le rapport	Définition
Suburb	Quartier	Variable texte, quartier dans lequel le logement est situé
Address	Adresse	Variable texte, adresse du logement
Rooms	Pièces	Variable numérique discrète, nombre de pièces du logement
Type	Type	Variable texte, type de logement (maison, duplex…)
Price	Prix	Variable numérique continue, prix en dollars australiens ($) du logement
Method	Méthode	Variable texte, méthode de vente du logement (ventes aux enchères, vente classique…)
SellerG	Agent	Variable texte, nom de l'agent immobilier qui a vendu le logement
Date	Date	Variable horaire, date à laquelle le logement a été vendu
Distance	CBD	Variable numérique continue, distance du logement au CBD (Central Business District) en kilomètres
Postcode	Postcode	Variable numérique discrète, code postal du logement
Bedroom2	Chambres	Variable numérique discrète, nombre de chambres du logement
Bathroom	SDB	Variable numérique discrète, nombre de salles de bain du logement
Car	Parking	Variable numérique discrète, nombre de places de parking du logement
Landsize	Terrain	Variable numérique continue, superficie du terrain en mètre carré

BuildingArea	Surface	Variable numérique continue, superficie du logement en mètre carré
YearBuilt	Année	Variable horaire, date à laquelle le logement a été construit
CouncilArea	Syndicat	Variable texte, nom du syndicat de copropriété responsable du logement
Lattitude	Latitude	Variable numérique continue, latitude du logement
Longtitude	Longitude	Variable numérique continue, longitude du logement
Regionname	Région	Variable texte, position du logement à Melbourne (Nord, Sud, Est…)
Propertycount	Voisins	Variable numérique continue, nombre de propriétés dans le quartier du logement

Nous avons comparé les variables de notre base de données à celles utilisées dans les différents articles académiques pour que chaque variable soit pertinente.

Dans leur article, Limbsombunchai, Gan et Lee (2004) utilisent dans leur modèle les variables qui selon eux sont déterminantes dans le prix du logement. Les variables utilisées sont les suivantes : LAND, AGE, TYPE, BEDROOMS, BATHROOMS, GARAGES, AMENITIES ainsi que la position du logement par rapport à Christchurch (Nouvelle-Zélande), qu'on appellera CHRISTCHURCH. Les variables utilisées sont similaires aux variables **Terrain**, **Année, Type**, **Chambre**, **SDB**, **Parking**, **Region** et **Quartier** de notre étude. La variable AMENITIES indique pour chaque logement la proximité du logement a des lieux d'intérêt, comme la variable **CBD** qui indique la distance du logement au Central Business District.

Dans son article "*La Méthode Hédonique d'Évaluation des Biens Immobiliers : Intérêts et Limites pour les Parcs HLM*", Gravel (?) régresse le prix des logements dans le marché immobilier du Val d'Oise par de nombreuses variables parmi lesquelles on retiendra : "Nombre de pièces", "Parking", "Superficie du jardin", "Distance à Paris". Ces variables sont similaires aux variables **Pièces**, **Parking**, **Terrain** et **CBD** de notre étude. Gravel utilise également plusieurs variables d'environnement comme le "Nombre de musées pour mille habitants" et d'autres variables propres au logement comme la présence d'un "Balcon" ou d'une "Cave". Les variables comme "Nombre de terrains de petits jeux pour mille habitants" peuvent être assimilées à notre variable **Voisins** puisqu'elles comptent le nombre de bâtiments spécifiques dans la zone du logement.

Dans leur article "*L'Importance de la Localisation dans la Valorisation des Quartiers Marseillais*", Gravel, Bono et Trannoy (2007) utilisent 155 variables dans leur modèle pour déterminer l'importance de la localisation dans la valorisation des quartiers marseillais. Les variables communes à notre étude sont les suivantes : "Log de la surface habitable" soit **Surface** dans notre base de données, "Pas de salle de bain", "1 salle de bain" et "2 salles de bains ou plus" correspondent à **SDB**. "1 place de parking" et "2 places de parking ou plus" correspondent à **Parking**. Notre variable **Année** est similaire aux variables "Construit entre [...] et [...]" de leur étude. Finalement les nombreuses variables de localisation comme "Proximité immédiate de la mer" ou encore "Distance au vieux port" donnent une information comme **CBD** sur la proximité à un lieu d'intérêt.

Les combinaisons de variables explicatives utilisées dans ces études nous confortent sur la pertinence des variables disponibles dans notre étude. C'est pourquoi nous allons les utiliser et les combiner dans la construction de notre modèle. On retient alors les variables **Surface, Terrain**, **Pièces, Chambres, SDB, Parking, CBD, Type** pour expliquer le prix du logement. Par exemple, en comparant deux logements toutes choses égales par ailleurs dans la ville de Melbourne, un logement avec une surface habitable plus grande et un terrain plus vaste devrait avoir un prix plus élevé. Le prix d'un T2 est très souvent plus élevé que le prix d'un T1 dans une même ville. Un nombre plus élevé de places de parking permet d'accueillir plus de voitures, et donc plus de personnes dans le logement. En général, une maison est plus chère qu'un appartement car plus spacieuse.

Les tableaux suivants détaillent les statistiques descriptives des variables conservées dans la base de données originale :

Variable	Nombre	Moyenne	Ecart-type	Minimum	Maximum	Q1	Médiane	Q3
Prix	13 580	1 075 684.08	639 310.72	85 000	9 000 000	650 000	903 000	1 330 000
Surface	7 130	151.97	541.01	0	44 515	93	126	174
Terrain	13 580	558.42	3 990.67	0	433 014	177	440	651
CDB	13 580	10.14	5.87	0	48.1	6.1	9.2	13

N	Pièces	Chambres	SDB	Parking
0	.	0.12	0.25	7.59
1	5.01	5.09	55.3	40.7
2	26.9	27.5	36.6	41.4
3	43.3	43.4	6.75	5.53
4	19.8	19.1	0.78	3.74
5	4.39	4.09	0.21	0.47
6	0.49	0.46	0.04	0.40
7	0.07	0.07	0.01	0.06
8	0.06	0.04	0.01	0.07
9	0.01	0.02	.	0.01
10	.	0.01	.	0.02
20	.	0.01	.	.

Toutes les valeurs du tableau ci dessus sont en pourcentage.

Type	Pourcentage
h	69.58
t	8.20
u	22.22

h - house,cottage,villa, semi,terrace; **u** - unit, duplex; **t** – townhouse

Afin de servir au mieux notre étude, la base de données a été modifiée telle que :

- Les observations pour lesquelles une ou plusieurs données étaient manquantes ont été supprimées, soit 7 384 observations.
- Différentes variables ont été créées: L'**Age** du logement (à partir de la variable **Année**). **logterrain**, **logsurface**, **logprix**, **logage, logcbd** et **logvoisins** sont les logarithmes des variables **Terrain**, **Surface**, **Prix**, **Age, CBD** et **Voisins**. **house** est une variable binaire qui donne 1 si le logement est une maison (**Type** = "h") et 0 sinon.

- Plusieurs valeurs aberrantes ont été retirées. Les logements construits avant 1880 ont été retirés. Les logements ayant une superficie de moins de 10 mètres carrés et de plus de 300 mètres carrés ont été retirés. Les logements ayant plus de 7 chambres, plus de 6 salles de bains et plus de 7 places de parking ont été retirés. Pour tenir compte des problèmes de segmentation de marché induit dans "Standardised Price Indices for the Regional Housing Market" de Francke, Vos et Janssen (2000) et dans "A Critical Review of Literature on the Hedonic Price Model" de Chin et Chau (2003), les observations conservées sont celles présentes dans la métropole de Melbourne, les observations situées en dehors de la métropole ont été retirées.

Notre base de données nettoyée de ses valeurs manquantes et aberrantes contient 5 814 observations. Les statistiques descriptives sont résumées dans les tableaux ci dessous :

Variable	Nombre	Moyenne	Ecart-type	Min	Max	Q1	Médiane	Q3
Prix	5 814	1 013 394	572 110	131 000	9 000 000	611 000	861 250	1 280 000
Surface	5 814	129.7	55.22	10	298	90	120	160
Terrain	5 814	455	917	0	37000	144	343.5	611
CBD	5 814	9.4	5.1	0	38	5.9	8.8	12.3

Après nettoyage, la moyenne et l'écart type de toutes les variables ont diminué. Avec une forte baisse pour **BuildingArea** et **Landsize**, que l'on peut expliquer par le retrait des valeurs aberrantes de l'échantillon.

En moyenne, un logement dans notre échantillon coûte 1 013 394 dollars australiens. Le logement le moins cher coûte 131 000 dollars australiens. 75% des logements de l'échantillon coûtent plus de 622 500 dollars australiens. Le prix médian de l'échantillon est de 861 250 dollars australiens. D'après le rapport « plan compare property reports 25 years of housing trends property market report », le prix médian d'une maison (64% de l'échantillon) à Melbourne en 2018 est de 824 955 dollars australiens. Le prix médian d'un duplex (26% de l'échantillon) en 2018 à Melbourne est de 574 000 dollars australien. Outre le prix médian des logements, nous n'avons pas trouvé d'autres statistiques immobilières sur Melbourne pour comparer nos données, ce type d'information appartiennent aux professionnels du secteur de l'immobilier, il ne s'agit pas de données publiques.

La surface moyenne des logements est de 129,7 mètres carrés. 75% des logements de l'échantillon ont une superficie de plus de 90 mètres carrés.
Les logements ont une superficie de terrain moyenne de 455 mètres carrés. Le logement avec le plus grand terrain a une superficie de 37 000 mètres carrés. 50% des logements de l'échantillon ont un terrain avec une superficie de moins de 343,5 mètres carrés.
Enfin, les logements de l'échantillon sont en moyenne à 9.4 kilomètres du quartier des affaires. Le logement le plus éloigné est à 38 kilomètres. Seulement 25% des logements sont à moins de 5.9 kilomètres.

N	Pièces	Chambres	SDB	Parking
0	.	0.09	.	7.22
1	5.62	5.78	55.40	45.94
2	29.31	30.25	38.30	38.42
3	42.31	42.41	5.80	4.73
4	19.38	18.35	0.40	2.94
5	3.04	2.82	0.09	0.38
6	0.28	0.26	0.02	0.31
7	0.03	0.03	.	0.05
8	0.02	.	.	.

Toutes les valeurs du tableau ci dessus sont en pourcentage.

Après nettoyage, notre échantillon ne comprend plus de logements sans salle de bain. La majorité des logements de l'échantillon possèdent 3 pièces (42.3%), 3 chambres (42.4%), une salle de bain (53.3%) et 2 places de parking (44.3%).

Type	Pourcentage
h	64.38
t	10.01
u	25.61

Enfin, après nettoyage, 64.38% des logements de notre échantillon sont des logements, contre 10% pour les townhouses et 25.61% pour les duplex.

IV. Le modèle économétrique

Le point de départ de notre étude économétrique est l'estimation par les moindres carrés ordinaires (MCO) du modèle suivant :

$$logprice = b_0 + b_1 logsurface + b_2 logterrain + b_3 logcdb + b_4 logage + b_5 Pieces + b_6 Chambres + b_7 SDB + b_8 Parking + b_9 house + u$$

En conformité avec les précédentes études empiriques sur les prix hédoniques, la forme fonctionnelle retenue est le logarithme du prix de transaction.

Ce modèle est estimé sous les hypothèses suivantes :

H_1 : Le modèle est correctement spécifié.

H_2 : les variables explicatives ne sont pas aléatoires.

H_3 : $E(U \mid X) = 0$. Il y a exogénéité des variables explicatives.

H_4 : $V(U) = \sigma^2 I \ \forall i$. Il y a homoscédasticité.

H_5 : La Matrice X'X est inversible et de plein rang.

H_6 : $\underset{N->\infty}{plim} \frac{X'X}{N} = Q_{XX}$ est une matrice définie positive, il y a corrélation entre les X.

H_7 : $\underset{N->\infty}{plim} \frac{X'U}{N} = 0$, les X et les u ne sont pas corrélés entre eux.

V. Résultats

Les résultats de l'estimation par MCO de notre modèle initial sont présentés dans le tableau ci dessous :

Variable	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Constante	10.58	0.064	163.20	<.0001
Pièces	0.02	0.015	1.46	0.1437
Chambres	0.027	0.014	1.84	0.0652
SDB	0.124	0.008	14.29	<.0001
Parking	0.016	0.005	3.09	0.0020
logsurface	0.55	0.015	35.61	<.0001
logcdb	-0.28	0.007	-36.21	<.0001
logterrain	0.02	0.002	9.92	<.0001
house	0.12	0.012	9.70	<.0001
logage	0.136	0.005	24.48	<.0001

Ces résultats seront discutés dans la sous-partie "Résultats économiques". Il nous faut avant cela tester les hypothèses du modèle et s'assurer de sa bonne spécification.

Tests de Fisher et de Student de significativité des paramètres

Le test de Student renseigne sur la significativité des paramètres du modèle.

$$\begin{cases} H_0 : b_k = 0 \\ H_1 : b_k \neq 0 \end{cases}$$ Pour tous paramètres k.

Pour un niveau de confiance 5%, on rejette H0 si t > |1.96| ou si la p-value < 0,05. Les deux dernières colonnes du tableau ci-dessus nous donnent ces 2 valeurs.
L'estimateur associé à **Chambres** ne rejette pas l'hypothèse nulle, il n'est pas significativement différent de 0. Pour tous les autres estimateurs, on rejette l'hypothèse nulle.

Le test de Fisher détermine la nullité simultanée de tous les paramètres du modèle, à l'exception de la constante.

$$\begin{cases} H_0 : Tous\ les\ parametres\ sont\ nuls \\ H_1 : Il\ existe\ b_k \neq 0 \end{cases}$$

La règle de décision se fait en fonction de la valeur de statistique calculée et celle lue dans la table de Fisher ou identiquement si la p-value < 0,05.

Pour notre modèle on rejette H0, 0,05> 0,001 il n'y a pas nullité simultanée des paramètres.

Multicolinéarité

Plusieurs études décrivent des problèmes de multicolinéarité entre variables, dont les études de Marchand & Skhiri (1995) "Prix hedoniques et estimation d'un modele structurel d'offre et de demande de caracteristiques" et Limsombunchain, Gan et Lee (2004) qui rencontrent un problème de multicolinéarité entre des variables explicatives. Les méthodes utilisées par Anju Verma et Renu Bala (2013) dans "The Relationship between Life Insurance and Economic Growth Evidence from India" montrent comment étudier et corriger des problèmes de multicolinéarité.

La tolérance consiste à régresser chacune des variables explicatives sur les autres et de calculer le R_j^2 associé. L'option *tol* dans SAS calcule la tolérance, elle prend ses valeurs entre 0 et 1. Si la tolérance d'une variable est proche de 0 et inférieure ou égale à 0.10, alors elle est dite colinéaire. Si elle est proche 1, alors elle n'est pas colinéaire.

Le facteur d'inflation de la variance (VIF) nous montre dans quelle mesure la variance (le carré de l'écart-type de l'estimation) d'un coefficient de régression estimé est augmentée en raison de colinéarité.

$VIF_j = \frac{1}{1 - R_j^2}$ où R_j^2 est le coefficient de détermination de la régression de la variable explicative j sur l'ensemble des autres variables explicatives du modèle. Une VIF ≥ 10 indique une forte colinéarité entre variables.

Résultats des indicateurs Tolérance et VIF, avant et après le retrait de Chambres

Variable	Tolérance	VIF	Tolérance sans Chambres	VIF sans Chambres
Pièces	**0.08**	**12.38**	**0.32315**	**3.09455**
Chambres	**0.08**	**11.46**	.	.
SDB	0.52	1.88	0.53110	1.88289
Parking	0.77	1.29	0.77336	1.29306
logsurface	0.32	3.03	0.33049	3.02584
logcdb	0.76	1.3	0.76743	1.30305
logterrain	0.57	1.73	0.57737	1.73198
house	0.46	2.17	0.46123	2.16812
logage	0.67	1.47	0.67998	1.47063

La variable **Pièces** atteint les seuils critiques de VIF et tolérance lorsque **Chambres** est présente dans le modèle. On ajoute à ce point que selon « the Rule of Thumb », Anju Verma et Renu Bala (2013) puis Limsombunchain, Gan et Lee (2004) montrent que la forte corrélation entre ces 2 variables (supérieure à 0,8) suffit à expliquer un problème de multicolinéarité. Comme **Chambres** n'est pas significativement différente de 0, on décide de la retirer du modèle. Après le retrait de **Chambres, Pièces** sort des seuils critiques des indicateurs VIF et tolérance.
Une raison potentielle à cette multicolinéarité peut s'expliquer par une mauvaise construction de la variable **Chambres** dans notre base de données, puisque cette variable provenait d'une source différente du reste de la base de données.

La table de corrélation en annexe montre d'autres résultats cohérents avec les travaux de Limsombunchain, Gan et Lee. La faible corrélation entre la surface du terrain et le prix du logement (0.08) implique un faible degré d'association linéaire. Ceci ne signifie pas pour autant que les deux variables sont indépendantes. Comme le montre les indicateurs tolérance et VIF, les fortes corrélations entre les variables **logsurface**, **Pièces** et **SDB** ne suffisent pas à créer des problèmes de multicolinéarité entre ces variables.

Hétéroscédasticité

Cette partie cherche à vérifier que l'hypothèse H_4 de notre modèle est bien vérifiée, les autres hypothèses H_1 ,H_2, H_3, H_5 , H_6 et H_7 sont supposées vraies. On veut tester s'il y a bien homoscédasticité des résidus. Les estimateurs restent sans biais malgré la présence d'hétéroscédasticité des résidus. Mais la variance σ^2 est biaisée et non convergente en probabilité. La variance du modèle n'est alors plus la variance minimum. Et le théorème de Gauss Markov n'est plus vérifié pour notre modèle. On utilise le test de White :

$$\begin{cases} H_0 : V(u_i) = \sigma^2 \, \forall i \\ H_1 : V(u_i) = \sigma_i^2 \, \forall i \end{cases}$$

Avec l'instruction SPEC de PROC REG on obtient le résultat suivant :

Test de spécification du premier et du deuxième moment		
DDL	**khi-2**	**Pr > khi-2**
43	271.15	<.0001

Pour un risque α = 5%, (Pr> χ^2) < 0.001 < 0,05. On rejette H_0, Cela signifie que la variance de l'erreur est hétéroscédastique et que la variance de celle-ci n'est pas constante. L'hypothèse H_4 du modèle n'est pas respectée.

L'objectif est désormais de corriger l'hétéroscédasticité de notre modèle, 2 solutions s'offrent à nous :

La première méthode est la régression pondérée. L'instruction 'Weight' sous SAS permet de minimiser la somme des carrés résiduels pondérés :

$\sum_{i=1}^{5814} w_i(logprice_i - \hat{logprice_i})^2$ avec w_i la variable spécifiée dans l'instruction Weight.

Dans le cadre de l'hétéroscédasticité, on utilise comme pondération l'inverse des écarts types théoriques σ_i. Soit $w_i = \frac{1}{\sigma_i}$. L'annexe n°2 explique comment w_i a été estimé sous SAS.
L'application de cette méthode sur le modèle donne les résultats suivants :

Paramètres	Estimation	Err type approx.	Valeur du test t	Approx Pr > |t|
Constante	10.49	0.0643	163.28	<.0001
Pièces	0.046	0.00769	6.03	<.0001
SDB	0.12	0.00881	13.55	<.0001
Parking	0.016	0.00526	3.15	0.0016
logsurface	0.57	0.0157	36.57	<.0001
logcdb	-0.27	0.00742	-36.33	<.0001
logterrain	0.02	0.00213	9.88	<.0001
house	0.13	0.0122	10.78	<.0001
logage	0.13	0.00546	24.30	<.0001

Le test de White rejette une nouvelle fois H_0, indiquant toujours la présence d'hétéroscédasticité. Nous retrouvons ici un problème déjà rencontré par Limbsombunchai, Gan et Lee (2004) : malgré l'estimation d'une régression pondérée, le test de White indique toujours la présence d'hétéroscédasticité.
Comment l'expliquer ? Dans le cadre du cours d'Econométrie I, nous avons discuté les limites au test de White : si H_0 est rejetée c'est parce qu'il peut y avoir une erreur de spécification dans les variables explicatives autant qu'il y a hétéroscédasticité. On teste à la fois l'hétéroscédasticité et la spécification ensemble, pas seulement l'homoscédasticité. Cette hypothèse de mauvaise spécification est également reprise par Bono, Gravel et Trannoy (2007).

La seconde possibilité pour corriger le problème d'hétéroscédasticité est l'utilisation de l' estimation de la matrice de variance-covariance robuste à l'hétéroscédasticité (HCCME), proposé par White (1980), puis reprise par Davidson et MacKinnon (1993).

Cette matrice de variance-covariance est $(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$. L'apport de Davidson et MacKinnon est la spécification de $\hat{\Omega}$. Elle consiste à multiplier le carré des résidus u_i^2 $(u_i = Y - Xb_{MCO}^{\wedge})$ par $\frac{n}{n-df}$ où n et df sont respectivement le nombre d'observations et le nombre de variables explicatives. On obtient ainsi :

$$\hat{\Omega} = \begin{pmatrix} \frac{n}{n-df}u_i^2 & 0 & \cdots & 0 \\ 0 & \frac{n}{n-df}u_i^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{n}{n-df}u_i^2 \end{pmatrix}$$

Sous SAS, l'instruction HCCME = 1 applique cette méthode.

Paramètres	**Estimation**	**Err type approx.**	**Valeur du test t**	**Approx Pr > |t|**
Constante	10.58	0.116	91.23	<.0001
Pièces	0.047	0.011	4.30	<.0001
SDB	0.12	0.011	11.55	<.0001
Parking	0.016	0.0058	2.80	0.0052
logsurface	0.55	0.03	18.02	<.0001
logcdb	-0.28	0.0079	-35.60	<.0001
logterrain	0.022	0.002	9.57	<.0001
house	0.121	0.013	9.43	<.0001
logage	0.136	0.0057	23.78	<.0001

Endogénéité

La présence de variables endogènes parmi les variables explicatives impliquent 2 hypothèses alternatives :
$\bar{H}_2$: Les variables explicatives peuvent être aléatoires.
$\bar{H}_3$: $E(u/X) \neq 0$. Les variables explicatives et le terme d'erreur sont corrélés.

Sous $\bar{H}_2$ et $\bar{H}_3$, tous les estimateurs, pas seulement ceux qui sont endogènes, sont biaisés et non convergent.

Pour prendre en compte l'endogénéité, nous avons recours à la méthode des variables instrumentales. Cette méthode suppose 3 hypothèses supplémentaires :

$HZ_1 : \underset{N->\infty}{plim} \frac{Z'U}{N} = 0$ il n'y a pas corrélation entre Z et U
$HZ_2 : \underset{N->\infty}{plim} \frac{Z'X}{N} = Q_{ZX}$, il y a corrélation entre X et Z

$HZ_3 : \underset{N->\infty}{plim} \frac{Z'Z}{N} = Q_{ZZ}$, il n'y a pas de multicolinéarité stricte entre les variables instrumentales

Dans le cas de petits échantillons, comme pour une estimation par MCO, les estimateurs $\hat{b}_{IV}$ sont biaisés et $V(\hat{b}_{IV})$ est inconnue. Dans le cas asymptotique, $\hat{b}_{IV}$ converge, ce qui permet que si l'échantillon est grand, $\hat{b}_{IV}$ est un bon estimateur.

Dans le cas de notre modèle, les études empiriques (par exemple, Bono, Gravel et Trannoy (2007)) qui mentionnent des potentiels cas d'endogénéité sur l'étape de l'offre de logement ne les traitent pas (situation différente du côté de la demande), faute de variables instrumentales pertinentes pour appliquer la méthode homonyme. Marchand et Skhiri (1995) décrivent des problèmes d'endogénéité lié à l'utilisation du prix par mètre carré comme variable expliquée, ce qui n'est pas notre cas. L'endogénéité est ici induite par le fait qu'un individu rapporte implicitement ce qu'il paie à l'espace habitable qu'il désire.

Nos travaux nécessitent pour autant la démonstration économétrique du traitement de l'endogénéité. On a suspecté une seule variable d'être endogène : la surface du logement. Du fait du manque d'informations concernant cette variable outre son unité de mesure. On ne sait pas s'il s'agit de la surface au sol, ou s'il y a une réglementation dans la dimension retenue semblable à la loi Carrez en France (stipulant qu'on ne prend pas en compte dans le calcul les parties des locaux d'une hauteur inférieure à 1,80 mètre et des lots d'une superficie inférieure à 8 mètres carrés). Avec toutes les informations nécessaires sur la construction de la surface du logement, Francke, Vos et Janssen (2000) ne mentionnent pas de problème d'endogénéité de la surface du logement dans leurs travaux.

Comme Bono, Gravel et Trannoy (2007), sélectionner des instruments n'a pas été aisé parmi les variables présentes dans notre base de données. On traite l'endogénéité de la façon suivante : logprice et logsurface sont variables endogènes, toutes les autres sont exogènes, les instruments sont le logarithme de la variable **Voisins** (qu'on nomme **logvoisins**) et la variable **Postcode**. On estime ce modèle par MCO, puis par 2SLS pour lequel on applique également le test d'hétéroscédasticité de White.

Test d'hétéroscédasticité					
Equation	**Test**	**Statistique**	**DDL**	**Pr > khi-2**	**Variables**
logprice	Test de White	5793	43	<.0001	Croix de toutes les var

Le test de White rejette H_0, il y a hétéroscédasticité du modèle. On ajoute alors la matrice de variance-covariance robuste à l'hétéroscédasticité (HCCME) comme appliquée dans la section au dessus traitant l'hétéroscédasticité avec variables exogènes.

L'étape suivante consiste à s'assurer de l'exogénéité des instruments utilisés. On effectue le test de Hansen pour vérifier l'absence de corrélation entre instruments et résidus, avec hétéroscédasticité.

$$\begin{cases} H_0 : \underset{T->\infty}{plim} \frac{Z'U}{T} = 0 \\ H_1 : \underset{T->\infty}{plim} \frac{Z'U}{T} \neq 0 \end{cases}$$

L'utilisation de ce test nécessite l'utilisation d'une estimation par la méthode généralisée des moments (GMM). L'estimateur $\hat{b}_{GMM}$ est biaisé, mais converge en probabilité $\underset{T->\infty}{plim}\ \hat{b}_{GMM} = b$, sous l'hypothèse :

$HZ_4 : \underset{T->\infty}{plim} \frac{Z'\Omega Z}{T} = Q_{Z\Omega Z}$ est une matrice définie positive.

Statistique de test GMM			
Test	**DDL**	**Statistique**	**Prob.**
Restrictions de suridentification	1	0.64	0.4234

La statistique du test de Hansen doit être comparée à la table χ^2 avec p - (k+1) degrés de liberté.
k étant le nombre de variables exogènes (variable dans X). Et p le nombre de variables dans Z.
Dans notre cas, k = 8, p = 10.

Région critique du test : R= $\{H > F_{1-\alpha}[\chi^2_{(p-k+1)}]\}$. Avec $\alpha = 0{,}05$ et p-k+1 =1
R = $\{H > F_{0,95}[\chi^2_{(1)}]\}$ <=> R = $\{0{,}64 \not> 3.84\}$.
On ne rejette pas H_0, il y a exogénéité des instruments.

L'étape suivante consiste a tester la pertinence des instruments avec le test des instruments faibles. Il s'agit de s'assurer que X et Z sont suffisamment corrélés. Autrement : l'estimateur des variables instrumentales n'est pas sans biais, et les tests d'hypothèses n'ont pas la bonne taille.

On teste :

$$\begin{cases} H_0 : les\ instruments\ sont\ faibles \\ Contre\ H_O^c \end{cases}$$

On utilise la statistique de Fisher pour effectuer ce test, sur le modèle suivant :

$$logsurface = b_0 + b_1 logterrain + b_2 logcbd + b_3 logage + b_4 Pieces + b_5 logvoisins + b_6 SDB + b_7 Parking + b_9 house + b_{10} postcode + v$$

On réécrit le test tel que :

$$\begin{cases} H_0 : b_{logvoisins} = b_{postcode} = 0 \\ H_1 : b_{logvoisins} \neq 0, b_{postcode} \neq 0 \end{cases}$$

Résultats du test 1 pour la variable dépendante logsurface				
Source	**DDL**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Numérateur	2	0.22914	3.44	0.0320
Dénominateur	5804	0.06655		

Pour un niveau de confiance 5%, on rejette H_0 (Pr>F <=> 0,05>0,032) il n'y a pas nullité simultanée des instruments. Les instruments **logVoisins** et **Postcode** ne sont pas faibles, ce qui veut dire que X et Z sont suffisamment corrélés.

L'étape suivante est le test d'exogénéité des variables endogènes, que l'on effectue avec le test d'Hausman. Le test d'Hausman montre que :

$(b^* - \tilde{b})'[V(b^*) - V(\tilde{b})]^{-1}(b^* - \tilde{b}) \sim \chi^2_{[p-k+1]}$ où k est le nombre de variables exogènes (variable dans X). Et p le nombre de variables dans Z. Ici, k = 8, p = 10. Avec la présence d'hétéroscédasticité, $\tilde{b}$ est l'estimateur des moindres carrés généralisés (MCG) et b^* l'estimateur des variables instrumentales. Nous utilisons cependant $\tilde{b}$ estimé par les MCO corrigé par l'estimateur de la matrice de variance-covariance robuste à l'hétéroscédasticité de White (HCCME) à la place des MCG.

Dans notre cas, on veut tester :

$$\begin{cases} H_0 : logsurface\ est\ exogene \\ H_1 : logsurface\ est\ endogene \end{cases}$$

Résultats du test de la spécification de Hausman				
Efficace sous H0	**Cohérent sous H1**	**DDL**	**Statistique**	**Pr > khi-2**
MCO	2SLS	1~~(9)~~	9.25	0.4146

Région critique du test : R= $\{Ha > F_{1-\alpha}[\chi^2_{(p-k+1)}]\}$. Avec $\alpha = 0{,}05$ et p-k+1 =1
R = $\{Ha > F_{0,95}[\chi^2_{(1)}]\}$ <=> R = $\{9{,}25 > 3.84\}$.
On rejette H_0, logsurface est variable endogène.

Pour conclure cette partie concernant l'endogénéité : l'exogénéité de logsurface étant rejetée, l'estimation par MCO n'est pas convergente, contrairement aux estimations par variables instrumentales et GMM, et ce dernier doit être le modèle le plus efficace.

L'estimation par GMM donne les résultats suivants :

Valeurs estimées GMM Paramètre non linéaires				
Paramètre	**Estimation**	**Err type approx.**	**Valeur du test t**	**Approx Pr > |t|**
Constante	-28.98	13.1004	-2.21	0.0270
Logsurface	11.19	3.5230	3.18	0.0015
Logterrain	-0.208	0.0773	-2.70	0.0070
Pièces	-2.503	0.8485	-2.95	0.0032
SDB	-1.552	0.5628	-2.76	0.0058
Parking	-0.287	0.1104	-2.60	0.0093
CBD	-0.273	0.0668	-4.10	<.0001
Logage	0.465	0.1212	3.83	0.0001
Type	-1.408	0.5274	-2.67	0.0076

De tels résultats sont surprenants. L'estimation de 5 des 8 variables explicatives changent de signes par rapport aux estimations précédentes et à ce qui est économiquement attendu. L'unité de mesure la surface du logement est elle aussi surprenante (une augmentation de 10% de la surface du logement entraînerait une augmentation de 111,9% du prix du logement).
Une telle situation peut être causée par une mauvaise spécification du modèle ou par le choix des instruments. En effet, il semble peu intuitif d'utiliser le code postal et le nombre de logements dans le quartier comme instruments de la surface du logement. Il s'agit dans notre cas d'un choix contraint par le manque de possibilités proposées par notre base de données.
C'est pourquoi on suppose dans la suite par hypothèse l'exogénéité des variables explicatives et que l'on ne retient les résultats obtenus lors du traitement de l'endogénéité.

Résultats économiques

	OLS		WLS		HCCME	
Variable	**Valeur estimée**	**Erreur type**	**Valeur estimée**	**Erreur type**	**Valeur estimée**	**Erreur type**
Constante	10.58	0.06485	10.49	0.0643	10.58	0.116
Pieces	0.05	0.00767	0.05	0.00769	0.05	0.011
SDB	0.13	0.00868	0.12	0.00881	0.12	0.011
Parking	0.02	0.00515	0.02	0.00526	0.016	0.0058
logsurface	0.55	0.01557	0.57	0.0157	0.55	0.03
logcbd	-0.28	0.00774	-0.27	0.00742	-0.28	0.0079
logterrain	0.02	0.00224	0.02	0.00213	0.022	0.002
house	0.12	0.01235	0.13	0.0122	0.121	0.013
logage	0.14	0.00557	0.13	0.00546	0.136	0.0057

Il est nécessaire de rappeler que les estimations par OLS et par WLS ou HCCME diffèrent sur une hypothèse, les résidus de ces deux derniers modèles sont hétéroscédastiques.
En comparant les résultats des différentes estimations que nous avons réalisées, l'estimation par Weighted Least Squares a globalement moins d'erreur-type pour chaque variable que l'estimation utilisant HCCME. Par équivalence, les estimations du modèle WLS sont globalement les plus précises.
Nous pouvons tirer nos résultats économiques du modèle WLS. Les résultats de ce modèle nous donnent les prix implicites des caractéristiques des logements. Notre variable expliquée étant le logarithme du prix, l'interprétation des résultats diffère en fonction de la variable explicative : une variable explicative sous forme logarithmique doit être interprété comme un paramètre d'élasticité, tandis qu'une variable explicative dite de "niveau" doit être interprété comme un paramètre de semi-élasticité.
Les résultats nous apprennent qu'une pièce supplémentaire entraîne une augmentation du prix du logement de 5%. Une salle de bain supplémentaire augmente le prix de 12%, tandis qu'une place de parking de plus augmente de 2% le prix du logement. Si la surface du logement augmente de 10%, alors le prix devrait augmenter de 5.7%. De la même manière, une augmentation de la superficie du terrain de 10% devrait augmenter le prix du logement de 0.2%. Un logement plus éloigné de 10% du Central Business District par rapport à un autre affiche un prix 2.7% plus faible. Si le logement en question est une maison, alors son prix devrait être plus élevé de 13% que celui d'un appartement, toutes choses égales par ailleurs. Enfin, si le logement affiche 10% d'ancienneté supplémentaire par rapport à sa date de construction, son prix devrait augmenter de 1.3%.
Les signes des coefficients sont tous en accord avec la théorie économique et nos attentes. Sauf pour l'interprétation de la valeur estimée de **logage**, qui est moins évidente. Les résultats indiquent

que toute chose égale par ailleurs, plus un logement est ancien, plus son prix devrait être élevé. Nous pourrions imaginer que dans une certaine mesure, les anciens logements de Melbourne peuvent être recherchés pour leur architecture d'époque. La rareté de ces anciens bâtiments se refléterait alors dans le prix. Un autre argument est l'état général du logement en question, une information à laquelle nous n'avons pas accès.

Critiques et discussions

Les résultats énoncés ci-dessus sont pour la plupart en accord avec les conclusions de la littérature économique des prix hédoniques appliqués à l'immobilier. Toutefois, notre étude souffre de quelques limites, principalement dues à nos données.
Durant la phase de recherches, nous nous sommes heurtés à la difficulté d'obtenir des bases de données officielles concernant les marchés immobiliers. La plupart des informations sur les logements sont détenues par les professionnels du secteur, et leurs données ne sont pas forcément publiques. Ceci fait du marché immobilier un milieu relativement opaque. Nous devions pour autant nous tourner vers l'open-data.

Nos données proviennent du site Kaggle, spécialisé dans la data-science. Le site propose de nombreuses bases de données, créées par les utilisateurs dans le but de s'entraîner au traitement et à l'analyse de données. Ainsi, la plupart des bases de données sont construites dans le but d'obtenir des résultats et il n'est pas forcément possible de pouvoir se fier à la provenance des données. Cela peut conduire à des problèmes de mauvaise spécification des modèles économétriques. Par exemple, l'auteur de notre base de données a précisé que la variable **Chambres** a été créée indépendamment des autres variables.
Comme nous l'avons vu, **Chambres** souffre d'un problème de non-significativité et de multicolinéarité, et nous avons dû la retirer du modèle. Le nombre de chambres est considéré comme un déterminant fiable du prix du logement, l'absence de données sur cette variable est problématique.

De plus, en ne construisant pas nous même notre base de données, nous avons été contraints dans le choix de nos variables. Nos travaux manquent de variables environnementales, beaucoup utilisées dans la littérature économique sur le sujet. Il aurait été intéressant d'intégrer dans nos variables explicatives le taux de criminalité par quartier, la nuisance sonore, la réussite scolaire, le taux de pauvreté… Par exemple, la variable **logage** a un impact positif sur le prix du logement : plus le logement est ancien, plus son prix devrait augmenter. Une solution pour justifier ce résultat serait l'intégration dans notre modèle du "gros-oeuvre" à l'instar des travaux de Marchand et Skhiri (1995). Une telle variable tiendrait compte de la détérioration physique du bâtiment et de la nécessité d'effectuer des travaux justifierait une relation positive entre l'âge du bâtiment et son prix, en supposant que le coût des travaux soit intégré au prix de transaction.

Ensuite, de nombreuses variables de notre base de données originale n'ont pas été utilisées. Les variables **Quartier**, **Adresse**, **Méthode**, **Agent, Date**, **Syndicat**, **Latitude, Longitude** et **Région** ont été mises de côté. Les variables **Quartier**, **Adresse**, **Syndicat**, **Latitude**, **Longitude** et **Région** sont des variables environnementales ou de localisation. Les variables **Quartier**, **Syndicat**, **Région** et **Voisins** donnent différentes informations propres à chaque quartier de l'échantillon.

Notre base nettoyée contient 244 quartiers différents, nous n'avons pas trouvé le bon moyen d'intégrer ces variables à notre régression. Enfin, les variables **Latitude** et **Longitude** sont des variables qui donnent la latitude et la longitude de chaque logement de l'échantillon. Il est difficilement justifiable d'imaginer que le prix du logement varie en fonction des variations de ces données toute chose égale par ailleurs, elles auraient besoin d'être combinées, soit ensemble, soit avec une autre variable de localisation. La variable **Date** permettrait de comparer le prix de transaction avec le climat économique dans laquelle la transaction a eu lieu. C'est-à-dire de comparer le prix de transaction tout chose égale par ailleurs entre 2 périodes différentes, par exemple avant et après une crise économique. Notre base de données ne renseigne que sur 2 années de vente, 2016 et 2017, nous ne l'avons pas utilisé dans le modèle.

Finalement, les variables **Méthode** et **Agent** renseignent sur la transaction même du logement en question sur le site domain.com.au. Nous considérons que ces données ne sont pas déterminantes du prix du logement.

VI. Conclusion

Cette étude a permis de construire une régression de “première étape” au sens de Rosen (1974) sur le prix d’un logement à Melbourne.

L’objectif principal était d’obtenir les prix implicites des caractéristiques de l’offre. Pour obtenir une telle régression, il a fallu tenir compte de problèmes économétriques classiques : hétéroscédasticité des résidus, endogénéité et multicolinéarité entre variables. Les problèmes d’hétéroscédasticité et de multicolinéarité ont été corrigés par des méthodes déjà utilisées dans la littérature économique, notamment Limbsombunchai, Gan et Lee (2004). Le traitement de l’endogénéité a été compliqué par le manque de pertinence des variables instrumentales. Un problème déjà rencontré par Bono, Gravel et Trannoy (2007). Si celles-ci ont permis de tester toutes les étapes du traitement de l’endogénéité d’une variable explicative, les estimations associées perdaient leur cohérence économique, avec notamment des signes de paramètres contraires aux corrélations entre variables et à la vraisemblance économique. C’est pourquoi les résultats d’une régression de “première étape” discutées ci-dessus sont ceux supposant par hypothèse l’exogénéité des variables explicatives.

On rappelle dans cette conclusion les autres limites économétriques de nos travaux déjà citées ci dessus, notamment sur la spécification du modèle.

Un tel modèle peut être complété par sa confrontation à un modèle représentant la demande de logement, soit une régression de « seconde étape » au sens de Rosen (1974).

Bibliographie & webographie

Source des données

DanB (2017) "*Melbourne Housing Snapshot*" Kaggle.com

Théorie des prix hédoniques

S. Rosen (1974) "*Hedonic Prices and Implicit Markets : Product Differentation in Pure Competition*", The Journal of Political Economy Vol. 82, No. 1, pp. 34-55

M. Greenstone (2017) "*The Continuing Impact of Sherwin Rosen's "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition"*" The Journal of Political Economy Vol. 125, No. 6, pp. 1891-1902

Application de la théorie des prix hédoniques

V. Limsombunchai, C. Gan, M. Lee (2004) "*House Price Prediction : Hedonic Price Model vs. Artificial Network*" American Journal of Applied Sciences Vol. 1, No. 3, pp. 193-201

N. Gravel (?) "*La Méthode Hédonique d'Évaluation des Biens Immobiliers : Intérêts et Limites pour les Parcs HLM*"

P-H. Bono, N. Gravel, A. Trannoy (2007) "*L'Importance de la Localisation dans la Valorisation des Quartiers Marseillais*" Economie Publique No. 20

K-W. Chau, T-L. Chin (2003) "*A Critical Review of Literature on the Hedonic Price Model*" International Journal for Housing Science and Its Applications Vol. 27, No. 2, pp. 145-165

G-A. Vos, M-C. Francke, J-E. Janssen (2000) "*Standardised Price Indices for the Regional Housing Market*" 7th European Real Estate Society Conference No. 40

E. Shkiri, O. Marchand (1995) "*Prix hédoniques et estimation d'un modèle structurel d'offre et de demande de caractéristiques*" Economie et Prévision Vol. 121, No. 5, pp. 127-140

R. Bala, A. Verma (2013) "*The Relationship Between Life Insurance and Economic Growth: Evidence from India*" Global Journal of Management and Business Studies Vol. 3, No. 4, pp. 413-422

P. Graves, J. Murdoch et M. Thayer (1988) « *The Robustness of Hedonic Price Estimation Urban Air Quality*"

M.Pitzer, S.Sebastian (2004) "*Hedonic price indices for the Paris housing market*"

Utilisation de SAS

J. Confais, M. Le Guen (2007) "Premier pas en régression linéaire avec SAS"

SAS User's Guide "The Model Procedure"

Regis de Bourbonnais 3th édition, Econométrie, Chapitre 4 : Multicolinéarité

"Le problème de la multicolinéarité", Sidi Mohamed Maouloud

Le marché immobilier à Melbourne

"plan compare property reports 25 years of housing trends property market report", Corelogic (2018)

Annexes

Annexe 1 : Table de corrélations

Table de corrélations									
	Price	**Rooms**	**Distance**	**Bathroom**	**Car**	**Landsize**	**BuildingArea**	**YearBuilt**	**Bedroom2**
Price	1								
Rooms	0.5	1							
Distance	-0.14	0.30	1						
Bathroom	0.42	0.57	0.13	1					
Car	0.21	0.4	0.29	0.31	1				
Landsize	0.05	0.07	0.04	0.05	0.09	1			
BuildingArea	0.61	0.76	0.23	0.63	0.39	0.07	1		
YearBuilt	-0.39	-0.1	0.25	0.15	0.12	0.02	-0.03	1	
Bedroom2	0.49	0.95	0.31	0.56	0.4	0.07	0.74	-0.09	1

Annexe 2 : régression pondérée

L'instruction 'Weight' sous SAS permet de minimiser la somme des carrés résiduels pondérés

$\sum_{i=1}^{6084} w_i(logprice_i - \hat{logprice_i})^2$ avec w_i la variable spécifiée dans l'instruction Weight.

Dans le cadre de l'hétéroscédasticité, on utilise comme pondération l'inverse des écarts types théoriques σ_i. Soit $w_i = \frac{1}{\sigma_i}$.

La démarche est la suivante : on récupère les résidus $\hat{u}_i$ du modèle initial, auxquels on applique 2 opérations. On les met au carré avant de passer ce dernier en logarithme : $log(\hat{u}_i^2)$

On régresse ensuite l'opérateur précédent sur toutes les variables explicatives du modèle.

$$log(\hat{u}_i^2) = b_0 + b_1 logsurface + b_2 logterrain + b_3 logdistance + b_4 logage + b_5 Rooms + b_6 SDB + b_7 Parking + b_8 house + v$$

On récupère ensuite les résidus de ce modèle, qu'on met sous forme exponentielle, avant de créer w_i tel que l'inverse de la racine carré de l'exponentiel des résidus estimés.

$$w_i = \frac{1}{\sqrt{\sigma_i^2}} = \frac{1}{\sigma_i}.$$