UNIVERSITÉ PARIS DESCARTES

FACULTÉ DE DROIT

Analyse de l'évolution du prix de l'Immobilier

BOQUANT Paul

HU Didier

Groupe de TD No 2

Chargé de TD M. Azzouz

LICENCE 3ÈME ANNÉE – Sciences Économiques et de Gestion

Semestre 1

Introduction à l'économétrie

Année universitaire 2018-2019

1. Introduction

Notre sujet porte sur l'explication de la valeur des biens immobiliers. Dans une société où l'accès à l'information économique est de plus en plus facilité, il parait évident que ce soit pour le vendeur ou l'acheteur d'un bien, d'estimer et de justifier le prix d'un bien à travers un modèle.

Ce sujet est important car il est simple et peut-être appliqué à la majorité des biens et services, il est applicable partout pourvu qu'on adapte le modèle au bien. Notre sujet est d'autant plus important qu'en immobilier la justification de la valeur d'un bien lors de la vente ou de l'achat est trop simple. Cette valeur est trop souvent basée uniquement sur un prix moyen du m^2, et de la présence ou non d'un ascenseur ou d'une terrasse et de quelques autres critères.

Nous pouvons voir depuis une dizaine d'année que l'économétrie s'empare de plus en plus de cette question. La volonté des investisseurs est souvent la même : comment prédire l'évolution du marché de l'immobilier dans tel ou tel secteur ?

Une autre tendance que l'on retrouve dans les thèses c'est la tentation des économètres de déceler des anticipations chez les agents économiques lors de tel ou tel impact sur le marché de l'immobilier.

Notre problématique est la suivante :« **Quels sont les facteurs qui justifient le prix d'un bien immobilier** ? »

Nous tenterons de répondre à cette problématique en posant plusieurs modèles économétriques à l'aide de notre base de données afin de garder celui qui explique au mieux le prix d'un bien immobilier. Nous procéderons ensuite à des tests pour déterminer si ce modèle est valable.

2. Données

2.1. Sources et échantillon

Dans un premier temps, avant de nous lancer dans un mémoire sur l'immobilier nous avions fait une demande au **Réseau Quetelet** pour avoir une base de données. Cependant cette base de données était trop compliquée à exploiter, elle contenait plus de 250 variables qui étaient presque toutes sous la forme booléenne. Nous avons donc recentré notre recherche vers des bases plus simples. Nous avons commencé à chercher des données sur le prix de l'immobilier en utilisant Google Dataset Search, divers sites comme Data.gouv.fr, l'Insee etc. Cependant les résultats de cette recherche ne nous ont pas permis de trouver une base de données. Nous avons alors dans un second temps, décidé de rechercher une base de données disponible sur R et trouvé la base de données intitulée ***hprice2*** via le package 'Wooldridge'. Elle résulte d'une étude menée par D.Harrisson et D.L. Rubinfeld en 1978 dont l'objectif était de rechercher une relation entre le prix d'un logement et la qualité de l'air, de la pollution environnante.

L'échantillon est constitué de 506 observations, chaque observation représente un lotissement (*communities* en anglais) dans les environs de Boston. L'échantillon ne présente aucune valeur manquante. Il s'agit d'un échantillon en coupe instantanée.

2.2. Variables

Tableau 1 : Statistiques descriptives des variables

Variable	Minimum	Médiane	Moyenne	Écart-type	Maximum
price	5000	21200	22512	9208.9	50001
crime	0.0060	0.2565	3.6115	8.5902	88.976
nox	3.85	5.38	5.55	1.16	8.71
rooms	3.560	6.210	6.284	0.703	8.780
dist	1.130	3.210	3.796	2.106	12.13
radial	1.000	5.000	9.549	8.707	24.00
proptax	18.70	33.00	40.82	16.85	71.10
stratio	12.60	19.10	18.46	2.166	22.00
lowstat	1.730	11.360	12.701	7.2381	39.070

Pour commencer nous pouvons dire que les écarts types (hormis stratio et rooms) sont assez élevés. On peut expliquer cette forte dispersion par le caractère très hétérogène des lotissements aux alentours de Boston. La volonté première de l'économètre dans sa recherche d'estimer la relation entre qualité de l'air et prix de l'immobilier, c'est de constituer un échantillon représentatif des alentours de Boston, donc de choisir des communautés aux profils variables. Stratio et rooms évoluent peu car il s'agit de variables qui ne peuvent évoluer significativement du fait de la proximité du choix des échantillons. La pollution ainsi que le droit foncier ne peuvent pas varier drastiquement à une telle proximité.

3. Modèles

3.1. Choix et transformation des variables

Dans le cadre de ce mémoire, nous cherchons à déterminer les différents facteurs expliquant le prix d'un logement. Le prix est donc la variable qu'on cherche à expliquer et les autres variables, celles qui sont explicatives. Nous décidons d'utiliser toutes les variables que nous avons à notre disposition dans la base de données pour poser notre modèle. Ainsi nous cherchons à déterminer le prix médian d'un logement d'un lotissement en fonction du crime, de la pollution, du nombre de pièces, de la distance par rapport aux centres d'emplois, du ratio élève-professeur, de l'indice d'accessibilité aux autoroutes, de la taxe foncière et du ratio de personne sous le seuil de pauvreté.

Afin de déterminer s'il faut transformer les variables, on construit un histogramme (voir annexes) pour chaque variable, lorsque la distribution est dissymétrique à gauche, on décide de transformer la variable en logarithme, on passe alors les variables « price », « nox », « dist » et « lowstat » en logarithme. L'histogramme de la variable « crime » montre également une distribution dissymétrique à gauche mais comme la variable comprend des valeurs très proches de 0, on décide de ne pas la transformer.

3.2. Écriture du modèle et tests

Nous avons commencé par estimer un premier modèle de régression multiple, appelé Modèle 1. Pour un échantillon de 506 lotissements dans les environs de Boston, nous avons estimé un modèle déterminant le prix médian d'un logement au sein du lotissement en fonction de différentes caractéristiques du lotissement, à savoir, la pollution (la quantité de protoxyde d'azote dans l'air, en partie par millions), le nombre moyen de pièces par logement, la distance qui sépare le lotissement de 5 centres d'emplois et enfin le ratio élève-professeur.

Modèle 1 => $Log(price)_i = \beta_0 + \beta_1 Log(nox)_i + \beta_2 rooms_i + \beta_3 Log(dist)_i + \beta_4 stratio_i + \mu_i$; $\forall i \in \{1; N\}$ et μ_i = terme d'erreur.

En réalisant la régression linéaire sur Rstudio, on s'aperçoit que le modèle explique un peu plus de 58% de la variance de Y, soit Log(price) et que toutes les variables du modèle sont significatives au seuil de 0,1%. Cependant lorsqu'on effectue le test de Ramsey, la p-value est inférieure à 5%, on doit donc rejeter l'hypothèse H0 ce qui signifie que le modèle n'est pas correctement spécifié.

On décide alors d'estimer un second modèle pour tenter de corriger cela, appelé Modèle 2. On ajoute au Modèle 1, les variables explicatives suivantes : « crime » qui est le nombre de crimes signalés par habitant, « radial » qui est l'indice d'accessibilité aux autoroutes, « proptax » qui est le taux d'imposition de la taxe foncière pour 1000 dollars et « lowstat » qui est le pourcentage de personne sous le seuil de pauvreté.

Modèle 2 => $Log(price)_i = \beta_0 + \beta_1 crime_i + \beta_2 Log(nox)_i + \beta_3 rooms_i + \beta_4 Log(dist)_i + \beta_5 radial_i + \beta_6 proptax_i + \beta_7 stratio_i + \beta_8 Log(lowstat)_i + \mu_i$; $\forall i \in \{1; N\}$ et μ_i = terme d'erreur.

On s'aperçoit que le modèle explique plus de 78% de la variance de Y et que toutes les variables sont significatives au seuil de 0,1%. Ici lorsqu'on effectue le test de Ramsey, on obtient une p-value de 0.4408, soit une p-value supérieure à 5%, on ne peut pas rejeter l'hypothèse H0, le modèle est correctement spécifié.

Néanmoins lorsqu'on effectue le test de Rainbow, on obtient une p-value inférieure à 5%, on doit rejeter l'hypothèse H0, le modèle est non linéaire. Cela peut s'expliquer par la présence de valeurs extrêmes, de valeurs influentes. On décide alors de détecter les valeurs influentes, on s'aperçoit qu'elles représentent moins de 15% de la totalité des valeurs, on décide de laisser le modèle tel quel et de ne pas faire d'estimation robuste.

On procède ensuite à des tests sur les résidus afin de déterminer si le modèle suit les hypothèses du MRM. D'une part, lorsqu'on regarde le graphique des résidus, on peut apercevoir des points aberrants, on peut également penser à la présence d'hétéroscédasticité, d'autre part lorsqu'on regarde le QQplot (voir annexes) des résidus, on s'aperçoit que les résidus s'écartent de la droite de Henry aux extrémités, cela présage un problème de normalité.

Les tests de Jarque.Bera et de Shapiro-Wilk confirment cela, on obtient une p-value inférieure à 5% pour chacun des tests, on doit donc rejeter l'hypothèse H0, ce qui signifie que les résidus ne suivent pas une loi normale. Cependant, l'hypothèse de normalité des erreurs est admise lorsque N est grand, en effet selon le théorème central limite (TCL) et la loi des grands nombres (LDGN), lorsque N est grand, on dit que l'estimateur MCO est asymptotiquement convergent, la distribution des β tend vers une loi normale.

Quant à l'hétéroscédasticité, sa présence est confirmée par les tests de Breush Pagan et de White, lors desquels on obtient une p-value inférieure à 5%, on doit donc rejeter l'hypothèse H0, les résidus ne sont pas homoscédastiques mais par conséquent hétéroscédastiques.

Un problème d'hétéroscédasticité signifie qu'on ne peut plus interpréter les tests de Student et Fisher, ils sont faussés étant donné que les écarts-types sont biaisés, on doit alors procéder à une correction de la matrice variance-covariance. À la suite de la correction, on s'aperçoit que la variable « rooms » perd en significativité, elle n'est plus significative au seuil de 0,1% mais de 1%.

Pour terminer, on réalise le test de Wald sur la significativité globale du modèle, on obtient une p-value inférieure à 5%, on doit rejeter l'hypothèse H0, les coefficients ne sont pas nuls, le modèle est globalement significatif.

4. Résultats

4.1. Résultats des estimations des modèles économétriques

Tableau 2 : Résultats des estimations MCO

Variable expliquée : Log(price) Variables explicatives :	Modèle 1	Modèle 2
Crime		-0.013978 *** (0.003)
Log(nox)	-0.953539 *** (0.117)	-0.499088 *** (0.101)
Rooms	0.254527 *** (0.019)	0.070462 ** (0.025)
Log(dist)	-0.134339 *** (0.043)	-0.244996 *** (0.038)
Radial		0.014501 *** (0.003)
Proptax		-0.007343 *** (0.001)
Stratio	-0.052451 *** (0.005)	-0.031352 *** (0.004)
Log(lowstat)		-0.381653 *** (0.033)
Constante	11.083861 *** (0.318)	12.330492 *** (0.345)
N	506	506
R^2ajusté	0.5807	0.7869
Test de Fisher	Rejet H0	Rejet H0
Test d'hétéroscédasticité	Rejet H0	Rejet H0
Correction hétéroscédasticité	Non	Oui

Les écarts types sont entre parenthèses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2. Interprétation des résultats

Tout d'abord, nous notons une augmentation du R^2ajusté entre le Modèle 1 et le Modèle 2 (passant de 0.58 à 0.79). Nous allons donc nous intéresser, et interpréter les résultats des variables explicatives du Modèle 2.

Pour une unité de « crime » supplémentaire (soit un crime signalé par habitant supplémentaire), « price » (prix médian des logements) diminue de 1,40% toute chose égale par ailleurs.
Lorsque « nox » (concentration de protoxyde d'azote dans l'air) augmente de 1% alors « price » diminue de 0,50% toute chose égale par ailleurs.
Pour une unité de « rooms » supplémentaire (soit une pièce en plus), « price » augmente de 7,05% toute chose égale par ailleurs.
Lorsque « dist » (distance pondérée vers 5 centres d'emplois) augmente de 1% alors « price » diminue de 0,25% toute chose égale par ailleurs.
Pour une unité de « radial » supplémentaire (soit une unité d'indice d'accessibilité aux autoroutes en plus), « price » augmente de 1,45% toute chose égale par ailleurs.
Pour une unité de « proptax » supplémentaire (soit un pourcent en plus sur la taxe foncière sur une base de 1000$), « price » diminue de 0,73% toute chose égale par ailleurs.
Pour une unité de « stratio » supplémentaire (soit l'augmentation d'une unité du ratio élève/professeur), « price » diminue de 3,14% toute chose égale par ailleurs.
Lorsque « lowstat » (pourcentage de personne sous le seuil de pauvreté) augmente de 1% alors « price » diminue de 0,38% toute chose égale par ailleurs.
La constante est égale à 12,33, cela signifie que le Log(price) est égal à 12,33 lorsque les variables indépendantes sont nulles.

Toutes les variables du Modèle 2 sont significatives au seuil de 0.1% (hormis « rooms » qui est significative au seuil de 1%).

5. Conclusion

Nous avons donc tout au long de notre mémoire tenté d'expliquer l'évolution de la valeur des biens immobiliers avec des variables autres que celles utilisées traditionnellement. Nous avons établi deux modèles. Un premier dont le R^2ajusté est de 0.58, cependant ce modèle n'étant pas correctement spécifié, nous en avons estimé un second en ajoutant de nouvelles variables. Pour celui-ci, nous avons obtenu un R^2ajusté supérieur à 0.78. Nous nous retrouvions donc avec une meilleure qualité d'ajustement, de ce fait nous avons poursuivi les tests pour ce modèle. Nous avons corrigé le problème d'hétéroscédasticité en utilisant la méthode de White, ainsi nous avons pu interpréter les résultats obtenus, toutes les variables sont significatives au seuil de 0,1% sauf « rooms » qui l'est au seuil de 1%.

Notre modèle expliquant un peu plus de 78% de la variance, nous arrivons donc à expliquer une part importante de l'évolution de la valeur des biens immobiliers. Cependant il n'est pas parfait, 22% de l'évolution de la valeur reste inexpliquée. Ces 22% correspondent à des variables que nous n'avons pas, notre modèle théorique n'est pas complet. Des variables comme le montant des permis de construire de la zone, le salaire moyen pourraient expliquer encore un peu plus notre modèle. Nous avons donc pu répondre à notre problématique en exposant les variables qui expliquent le plus l'évolution de la valeur des biens.

Pour poursuivre notre projet nous pourrions nous-mêmes constituer une base de données auprès des agences immobilières et chercher à intégrer de nouvelles variables dans l'objectif d'estimer un modèle toujours plus complet.

Bibliographie

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 5th Edition, 2013

Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, *5*(1), 81-102.

Practicing Econometrics for Real estate Market Analysis
https://landeconomics.nccu.edu.tw/~LE/files/news/1082_df9cc446.pdf

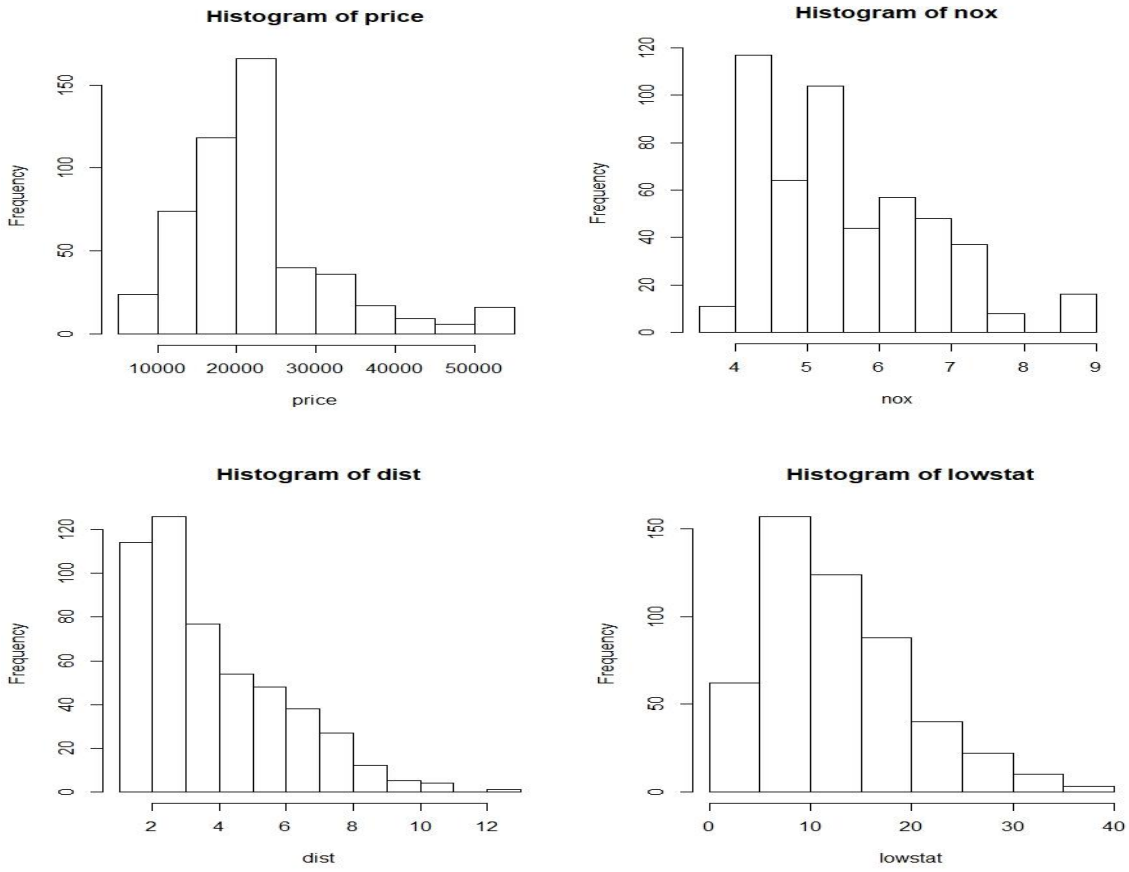
Mariusz Doszyńmariusz, & Sebastian Gnat. *Econometric Identification of the Impact of Real Estate Characteristics Based on Predictive and Studentized Residuals*
https://content.sciendo.com/view/journals/remav/25/1/article-p84.xml

Annexes

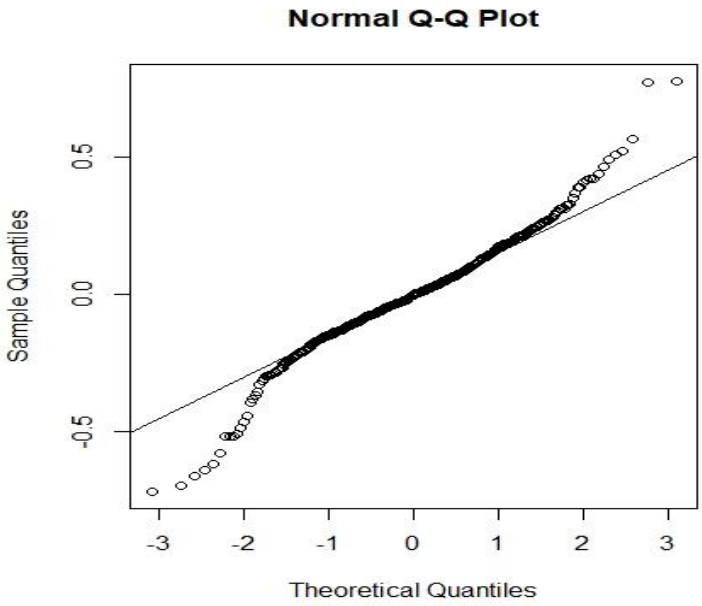
a. Définition des variables

Nom de la variable	Définition de la variable	Type de variable
price	Prix médian des logements en dollars	Variable à expliquer
crime	Nombre de crimes signalés par habitant	Variable explicative
nox	Concentration de protoxyde d'azote dans l'air, en ppm	Variable explicative
rooms	Nombre moyen de pièces par logement	Variable explicative
dist	Distance pondérée vers 5 centres d'emplois	Variable explicative
radial	Indice d'accessibilité aux autoroutes	Variable explicative
proptax	Taux d'imposition de la taxe foncière pour 1000 dollars	Variable explicative
stratio	Ratio élève-professeur	Variable explicative
lowstat	Pourcentage de personne sous le seuil de pauvreté	Variable explicative

b. Histogrammes des variables transformées



c. Diagramme Quantile-Quantile des résidus



d. Tableau des tests effectués pour le Modèle 2

Test	P-value	Acceptation ou rejet de H0
Ramsey	0.4408	Acceptation de H0
Rainbow	8.73e-11	Rejet de H0
Jarque.Bera	<2.2e-16	Rejet de H0
Shapiro-Wilk	1.731e-09	Rejet de H0
Breush Pagan	1.175e-15	Rejet de H0
White	5.007e-15	Rejet de H0
Wald	<2.2e-16	Rejet de H0

***