UNIVERSITÉ D'ANGERS

FACULTÉ DE DROIT, D'ÉCONOMIE ET DE GESTION

MASTER 1 ÉCONOMIE APPLIQUÉE PARCOURS IEE

Cours de Data Mining2

Sujet : les facteurs de variation des prix de vente des maisons et leur impact sur ceux-ci

Préparé par : ADAYISSO Kokou

Professeur : Christophe DANIEL

Remis le 26 Mai 2020

Le Système SAS

1 INRODUCTION

Perçu comme l'un des secteurs les plus rentables au monde, bon nombre d'investisseurs rêvent investir leur argent dans l'immobilier. Ainsi pour tirer profit, il est important de tenir compte de certains aspects. Tout comme les autres domaines, l'immobilier est influencé par des facteurs. Ces facteurs sont souvent récurrents dans la fluctuation des prix des immobiliers ; les prix de vente de ces actifs varient en fonction de certains critères. Pour aider les agents du secteur à mieux adapter leur prix de vente des immobiliers, nous allons proposer une analyse statistique et économétrique. Cependant notre étude va principalement tourner autour des questions suivantes :

Quelles sont variables les plus importantes dans la variation des prix de vente des maisons ? quel impact ces variables exercent sur les prix de vente des logements ?

En effet, dans la perspective de fournir des réponses à ces questions, nous allons faire usage d'une base de données tirées sur le site Woodbridge[1]. Cette base de données est issue des données collectées par le bureau fiscal du Nord Andover, Massachusetts aux Etats-Unis dans les années 1981. Il faut noter que la base de données brutes est constituée des données de 1978 et des années 1981. Mais notre étude va s'intéresser uniquement des données de 1981 car la base de données adaptées cette étude est les données en coupe transversales et de plus les données de 1981 sont récentes par rapport aux données de 1978. La taille de l'échantillon de la population d'études est de 142 individus statistiques et nous avons au total 11 variables d'intérêt qui sont les facteurs de variations des prix de vente des immeubles. Ces variables sont l'âge des maisons en année(age), le nombre de connaissances dans le quartier(nbh), la distance de la maison par rapport à une station de bus mesuré en pieds(cbd), la distance par rapport à l'autoroute mesuré en pieds(intst), le nombre de chambres des maisons(rooms), la superficie des maisons en pieds carré(area), la superficie des terrains en pieds carré(land), le nombre de salles de bains(baths), la distance par rapport à l'incinérateur[2]

[1] https://faculty.utrgv.edu/diego.escobari/teaching/Datasets.html

[2] Un incinérateur, dans les années 1900, est un dispositif fonctionnant à base de vent, sert à détruire par combustion les ordures ménagères. Il permet également de transformer les déchets en électricité

mesuré en pieds(dist), le pourcentage de vent incinérateur qui rentre dans les maisons(wind) et enfin la variable binaire la maison proche de l'incinérateur(nearinc). Veuillez-vous référer à l'annexe dans la partie description des variables pour plus de détails sur les variables.

Par ailleurs nous sommes conscients qu'il peut y avoir d'autres bases de données plus récentes et d'autres facteurs qui décrivent aussi mieux la variation des prix des maisons mais nous allons nous contenter de cette base car c'est la seule qui nous est accessible pour le moment.
Pour mener à bien notre étude, nous allons dans un premier temps faire une statistique descriptive qui nous renouvelle les caractéristiques statistiques principales des différentes variables. Dans un deuxième temps effectuer la recherche du modèle économétrique optimal par la méthode de régression des moindre carrés ordinaires(MCO) non seulement permettra de déterminer le meilleur modèle économétrique mais va nous permettre également de connaitre les variables qui impactent de manière significative les prix de vente des maisons aux Etats-Unis et de mesurer des variations de ces variables sur les prix des maisons. Dans la section III, nous allons procéder à la régression PCR(Principal Composante Regression) ; son objectif est d'apporter un aperçu explicatif de cette méthode , de déterminer les variables les plus importantes dans la volatilité des prix des prix des immobiliers aux USA, de déterminer et d'interpréter les coefficients de régression dans le but de prédire les prix des maisons. La quatrième section va se consacrer à l'étude économétrique par PLS(Principale Least Square) ; cette section vise les mêmes objectifs que la méthode PCR(la section III) mais la seule différence est que nous réaliser une étude légèrement approfondie en comparant les résultats des deux dernières méthodes de régression(PCR et PLS).

2 STATISTIQUES DESCRIPTIVES

2.1 Visualisation et caractéristiques de la base de données

La base de données est constituée de 25 variables et de 142 observations. Et il n'y a pas de données manquantes.

Nom de la table	WORK.TRAV	**Observations**	142
Type de membre	DATA	**Variables**	25
Moteur	V9	**Index**	0
Créée	24/02/2020 07:08:30	**Longueur d'observation**	200
Dernière modification	24/02/2020 07:08:30	**Observations supprimées**	0
Protection		**Compressée**	NON
Type de table		**Triée**	NON

Tableau 1

Liste alphabétique des variables et des attributs					
#	**Variable**	**Type**	**Long.**	**Format**	**Libellé**
2	age	Num.	8	BEST.	age
3	agesq	Num.	8	BEST.	agesq
10	area	Num.	8	BEST.	area
12	baths	Num.	8	BEST.	baths
5	cbd	Num.	8	BEST.	cbd
13	dist	Num.	8	BEST.	dist
6	intst	Num.	8	BEST.	intst
11	land	Num.	8	BEST.	land
18	larea	Num.	8	BEST.	larea
14	ldist	Num.	8	BEST.	ldist
7	lintst	Num.	8	BEST.	lintst
21	lintstsq	Num.	8	BEST.	lintstsq
19	lland	Num.	8	BEST.	lland
16	lprice	Num.	8	BEST.	lprice
4	nbh	Num.	8	BEST.	nbh
22	nearinc	Num.	8	BEST.	nearinc
8	price	Num.	8	BEST.	price
9	rooms	Num.	8	BEST.	rooms
15	wind	Num.	8	BEST.	wind

Tableau 2

2.2 Les statistiques principales :

D'après la sortie SAS [voir tableau3], l'âge moyen des maisons mises en vente est environ de 14 ans. L'âge maximal des maisons mises en vente est de 131 ans cependant il existe des maisons mise en vente qui ont 0 durée de vie, n'ont pas encore au moins 1 an.

En moyenne, les maisons possèdent 2 salles de bain et possèdent au plus 4 salles de bain ; toutes les maisons possèdent au moins 1 salle de bain.

Par contre en moyenne, les maisons mise en vente contiennent environ 7 chambres. Cependant le nombre Minimal de chambre est 4 alors que le nombre maximal de chambres des maisons est de 9.

Le prix moyen de vente des maisons est de 120647 dollars. Cependant la maison la moins chère coute 41000 dollars tandis la maison la plus chère coute 270000 dollars.

Le Système SAS

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Maximum
age	age	142	13.9788732	23.9368164	0	131.0000000
agesq	agesq	142	764.3450704	2285.42	0	17161.00
area	area	142	2241.73	744.4012411	735.0000000	5136.00
baths	baths	142	2.3802817	0.8054107	1.0000000	4.0000000
cbd	cbd	142	15063.38	8344.71	1000.00	33000.00
dist	dist	142	20110.56	8047.15	5000.00	38900.00
intst	intst	142	15577.46	8466.31	1000.00	33000.00
land	land	142	35437.78	17976.25	3049.00	81273.00
larea	larea	142	7.6555919	0.3571932	6.5998710	8.5440300
ldist	ldist	142	9.8166745	0.4537116	8.5171930	10.5687500
lintst	lintst	142	9.4508204	0.7179780	6.9078000	10.4040000
lintstsq	lintstsq	142	89.8298756	13.0658329	47.7177000	108.2432000
lland	lland	142	10.2788935	0.7205404	8.0225690	11.3055700
lprice	lprice	142	11.6290187	0.3899205	10.6213273	12.5061772
lrprice	lrprice	142	11.3651173	0.3899207	10.3574300	12.2422800
nbh	nbh	142	1.9647887	2.1941637	0	6.0000000
nearinc	nearinc	142	0.2816901	0.4514154	0	1.0000000
price	price	142	120647.13	44359.89	41000.00	270000.00
rooms	rooms	142	6.5915493	0.8264676	4.0000000	9.0000000
rprice	rprice	142	92662.93	34070.58	31490.02	207373.30
wind	wind	142	6.7676056	2.4485210	3.0000000	11.0000000

Tableau 3

2.2.1 Fréquences

2.2.1.1 Pour la variable Price :

D'après la sortie SAS ci-dessous, on peut affirmer que 0.7% des maisons coute 41000 dollars alors que 2.82% des maisons coutent 70000 dollars.

Cependant 17.61% des maisons coutent moins de 73000 dollars et 18.31% des maisons valent au plus 73000 dollars.

price				
price	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
41000	1	0.70	1	0.70
47000	1	0.70	2	1.41
49000	1	0.70	3	2.11
50000	1	0.70	4	2.82
52000	1	0.70	5	3.52
52900	1	0.70	6	4.23
54000	1	0.70	7	4.93
60000	2	1.41	9	6.34
62000	1	0.70	10	7.04
65000	2	1.41	12	8.45
66000	1	0.70	13	9.15
66479	1	0.70	14	9.86
66500	2	1.41	16	11.27
67000	1	0.70	17	11.97
68000	3	2.11	20	14.08
70000	4	2.82	24	16.90
72500	1	0.70	25	17.61
73000	1	0.70	26	18.31
73900	1	0.70	27	19.01
74000	1	0.70	28	19.72
75000	1	0.70	29	20.42

Tableau 4

2.2.1.2 *Pour la variable nearinc :*

Parmi les maisons mises vente seulement 28.17% sont jugées proches de l'autoroute contre 71.83% qui sont jugées loin de l'autoroute.

nearinc				
nearinc	**Fréquence**	**Pourcentage**	**Fréquence cumulée**	**Pourcentage cumulé**
0	102	71.83	102	71.83
1	40	28.17	142	100.00

Tableau 5

2.2.1.3 Pour la variable baths :

La majorité (54.23%) des maisons dispose 3 salles de bain. Par contre 98.59% des maisons possède moins de 4 salles de bains.

baths				
baths	**Fréquence**	**Pourcentage**	**Fréquence cumulée**	**Pourcentage cumulé**
1	27	19.01	27	19.01
2	36	25.35	63	44.37
3	77	54.23	140	98.59
4	2	1.41	142	100.00

Tableau 6

3 REGRESSION MULTIPLE PAR METHODE DES MCO :

3.1 Modèle niveau : Modèle 1

3.1.1 Variable dépendante : price

D'après la sortie ci-dessous, le modèle est globalement significatif au seuil de 1% car la p-value(pr > F) est inférieure à 1%. Ce qui veut dire qu'au moins le coefficient d'une variable explicative est statistiquement différent de 0.

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	17	1.9863E11	11684118636	18.38	<.0001
Erreur	124	78829812517	635724294		
Total sommes corrigées	141	2.774598E11			

Tableau 7

La qualité d'ajustement du modèle est à 67.69%. C'est à dire les variables indépendantes expliquent 67.69% la variance de la variable prix.

Le Système SAS

Root MSE	25214	**R carré**	0.7159
Moyenne dépendante	120647	**R car. ajust.**	0.6769
Coeff Var	20.89861		

Tableau 8

Parmi toutes les variables explicatives du modele1, seulement quatre (nombre de salles de bain, distance de la maison à l'incinérateur, pourcentage de temps de vent incinérateur, log de la distance de la maison à l'incinérateur) sortent statistiquement significatives au seuil de 5% car leur p-value inférieure à 5%. On va donc essayer d'interpréter les coefficients des variables qui sont significatifs au seuil de 5% :

Toute chose égale par ailleurs, l'augmentation d'une unité le nombre de salles de bain entraine une hausse de prix de vente de maison de 15925 dollars us. Par contre Chaque 1 pied de plus de la distance de la maison à l'incinérateur entraine approximativement une baisse de prix de vente de la maison de 12 dollars us.

Si le pourcentage de temps incinérateur augmente de 1%, le prix de vente de la maison baisse de 79.15 de dollars us.

NB : On remarque que même après le test et la correction d'hétéroscédasticité, c'est toujours 4 coefficients de régression qui sont significatifs au seuil de 5% car p-value après la correction d'hétéroscédasticité sont inférieure à 0. De plus, d'après le test de vif (test de multi colinéarité), beaucoup de variables explicatives sont colinéaires entre elles car inflation de variance de vif est supérieure à 10. C'est-à-dire les variables qui ont leur inflation de variation supérieure à 10 cause le problème de multi colinéarité. Le modèle1 souffre donc de problème de biais d'estimation. Cependant il importe de changer de modèle. Ainsi nous allons passer au modèle log(prix).

NB : les variables âge et la distance par rapport à l'autoroute(linstst) sont en mode quadratique (le fait d'introduire une variable et son carré dans une régression) du fait qu'elles s'accommodent et ajustent mieux le modèle de régression car le coefficient de détermination ajusté devient plus important en mettant ces deux variables en mode quadratique.

Résultats estimés des paramètres									
							Cohérent avec l'hétéroscédasticité		
Variable	**Libellé**	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	Intercept	1	-1448661	1216815	-1.19	0.2361	906077	-1.60	0.1124
age	age	1	-522.09172	394.97512	-1.32	0.1887	387.12247	-1.35	0.1799
agesq	agesq	1	2.37244	3.42197	0.69	0.4894	2.91136	0.81	0.4167
area	area	1	-5.67575	13.59712	-0.42	0.6771	20.64010	-0.27	0.7838
baths	baths	1	15925	5807.98250	2.74	0.0070	5350.94662	2.98	0.0035
cbd	cbd	1	1.18163	3.39557	0.35	0.7284	2.27018	0.52	0.6036
dist	dist	1	-11.91354	4.13689	-2.88	0.0047	3.98682	-2.99	0.0034
intst	intst	1	4.27378	5.63383	0.76	0.4495	4.13621	1.03	0.3035
land	land	1	-0.30951	0.43351	-0.71	0.4766	0.37092	-0.83	0.4056
larea	larea	1	49112	30484	1.61	0.1097	38833	1.26	0.2084
ldist	ldist	1	139788	46012	3.04	0.0029	54751	2.55	0.0119
lintst	lintst	1	-36717	273177	-0.13	0.8933	184849	-0.20	0.8429

Le Système SAS

Résultats estimés des paramètres									
							Cohérent avec l'hétéroscédasticité		
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Erreur type	Valeur du test t	Pr > |t|
lintstsq	lintstsq	1	2147.38436	17758	0.12	0.9039	12371	0.17	0.8625
lland	lland	1	15509	11010	1.41	0.1615	8795.28473	1.76	0.0803
nbh	nbh	1	-2612.78575	1561.70776	-1.67	0.0968	1348.19144	-1.94	0.0549
nearinc	nearinc	1	13578	13485	1.01	0.3159	21805	0.62	0.5346
rooms	rooms	1	2599.19090	3997.40263	0.65	0.5168	3510.84103	0.74	0.4605
wind	wind	1	-7914.53712	2658.70153	-2.98	0.0035	2536.21055	-3.12	0.0022

Tableau 9

Résultats estimés des paramètres			
Variable	Libellé	DDL	Inflation de variance
Intercept	Intercept	1	0
age	age	1	19.82543
agesq	agesq	1	13.56550
area	area	1	22.72261
baths	baths	1	4.85328
cbd	cbd	1	178.07313
dist	dist	1	245.80111
intst	intst	1	504.59786
land	land	1	13.46948
larea	larea	1	26.29630
ldist	ldist	1	96.66134
lintst	lintst	1	8532.21369
lintstsq	lintstsq	1	11940
lland	lland	1	13.95952
nbh	nbh	1	2.60428
nearinc	nearinc	1	8.21915
rooms	rooms	1	2.42079
wind	wind	1	9.39934

Suite de tableau 9

3.2 Modèle log-Prix : Modele2

3.2.1 Variable dépendante :log(prix)

Le model est globalement significatif au seuil de 1% c'est-à-dire il existe au moins un coefficient Statistiquement différent de zéro car P-value est inferieure 1%.

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	17	17.38785	1.02281	31.32	<.0001
Erreur	124	4.04947	0.03266		
Total sommes corrigées	141	21.43732			

Tableau 10

Le modèle explique à 78.52% le log-prix. Autrement dire, les variables explicatives du modele2 explique 78.52% la variation de log-prix.

Root MSE	0.18071	**R carré**	0.8111
Moyenne dépendante	11.62902	**R car. ajust.**	0.7852
Coeff Var	1.55398		

Tableau 11

On note une augmentation de nombre de coefficients de régression significatifs par rapport au modèle précédent qui contient seulement 4 coefficients significatifs. On trouve ici 7 variables qui influencent significativement notre variable expliquée : le log-prix. On va essayer d'interpréter les variables significatives de ce modèle : L'âge n'a qu'un impact marginal sur le log-prix. On peut dire qu'un an de plus de vie de maison implique une baisse de prix de vente de maison de 0.64%.

Par ailleurs, le nombre de salles a un effet non négligeable sur le prix. Chaque salle de bain supplémentaire entraine 13.21% de prix de vente de maison de plus.

NB : il est clair que le modele2(modèle log-prix) est meilleur par rapport au modele1 car en plus son coefficient de détermination ajustée (R2 ajustée=78,52%) est le plus élevé, il regorge également un nombre de coefficients de régression significatifs le plus élevé. Ce qui nous permet de laisser le modele1 au profit de modele2 dans la suite de l'étude.

Cependant, malgré quelques des améliorations des indices statistiques notés au niveau du modele2 par rapport au modele1, le modele2 n'est le modèle le plus idéale qu'on puisse attendre en économétrie car certains coefficients de régressions restent non significatifs et certaines variables trainent avec les problèmes de multicolinéarité. Sur ce, nous allons passer un autre modèle qui va chercher à minimiser le plus possible le problème de multicolinéarité.

Le Système SAS

Résultats estimés des paramètres									
							Cohérent avec l'hétéroscédasticité		
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	-7.20454	8.72124	-0.83	0.4103	6.48844	-1.11	0.2690
age	age	1	-0.00644	0.00283	-2.28	0.0245	0.00275	-2.35	0.0206
agesq	agesq	1	0.00003582	0.00002453	1.46	0.1467	0.00002040	1.76	0.0815
area	area	1	-0.00014123	0.00009745	-1.45	0.1498	0.00013642	-1.04	0.3026
baths	baths	1	0.13207	0.04163	3.17	0.0019	0.04074	3.24	0.0015
cbd	cbd	1	0.00001319	0.00002434	0.54	0.5887	0.00001870	0.71	0.4819
dist	dist	1	-0.00007811	0.00002965	-2.63	0.0095	0.00002645	-2.95	0.0038
intst	intst	1	0.00003311	0.00004038	0.82	0.4138	0.00003075	1.08	0.2836
land	land	1	-0.00000453	0.00000311	-1.46	0.1476	0.00000283	-1.60	0.1118
larea	larea	1	0.62403	0.21849	2.86	0.0050	0.26467	2.36	0.0200
ldist	ldist	1	0.96373	0.32978	2.92	0.0041	0.32497	2.97	0.0036
lintst	lintst	1	0.99879	1.95793	0.51	0.6109	1.34931	0.74	0.4606
lintstsq	lintstsq	1	-0.06024	0.12727	-0.47	0.6368	0.08982	-0.67	0.5037
lland	lland	1	0.18309	0.07891	2.32	0.0220	0.06874	2.66	0.0088
nbh	nbh	1	-0.02089	0.01119	-1.87	0.0644	0.01048	-1.99	0.0485
nearinc	nearinc	1	0.10551	0.09665	1.09	0.2771	0.12095	0.87	0.3847
rooms	rooms	1	0.01585	0.02865	0.55	0.5811	0.02474	0.64	0.5230
wind	wind	1	-0.04822	0.01906	-2.53	0.0126	0.01838	-2.62	0.0098

Tableau12

3.3 Modèle 3

Ce modèle est un modèle issu de modele2, c'est le modèle résultant du retrait des certaines variables (area, dist, land, instst) qui peuvent accentuer le problème de multicolinearité. Ces variables peuvent être les variables qui sont à la fois en niveau et en log. En effet, nous savons qu'une variable et son log sont souvent fortement corrélées ; mettre une variable et son log en même temps dans un modèle de régression ne peut qu'aggraver le problème de multicolinearité. Ainsi pour amoindrir ce problème, nous avons retiré du modele2, les variables indépendantes qui sont à la fois en niveau et en log pour ne garder que leur log pour avoir le modele3 car le log des variables indépendantes (77%) explique mieux la variable endogène(log-prix) que leur niveau (75%).

Ici on note une nette amélioration en termes de multicolinearité, nous avons 7/13 variables indépendantes qui ne sont pas colinéaires aux autres variables indépendantes comparativement au modele2 qui ne compte 5/17 variables indépendantes non colinéaires aux variables indépendantes d'après les tests de vif (toutes variables qui possèdent l'inflation de variance inférieure à 10% sont non colinéaires aux autres variables indépendantes) [Voir Suite du tableau15].

Le modele3 est globalement significatif car probabilité critique(Pr>F<0.0001) est inférieure à 5% [voir tableau13] et explique à plus de 77% la variabilité de log(prix) [voir tableau14]. Il contient 4/13 variables significatives au seuil de 5% même après la correction d'hétéroscédasticité.

Nous remarquons que malgré nos efforts pour minimiser le risque de biais d'estimations et le problème de multicolinearité, beaucoup de variables (importantes pour notre étude) restent encore non significatives (9/13 variables) et colinéaires entre elles (6/13) dans le modele3.

Cependant, pour palier à ce problème, nous allons faire recours aux modèles de régressions sur des variables latentes qui consistent à régresser les variables synthétiques issus de l'analyse de données des variables indépendantes sur la viables dépendantes.

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	13	16.99307	1.30716	37.65	<.0001
Erreur	128	4.44426	0.03472		
Total sommes corrigées	141	21.43732			

Tableau 13

Root MSE	0.18634	**R carré**	0.7927
Moyenne dépendante	11.62902	**R car. ajust.**	0.7716
Coeff Var	1.60233		

Tableau 14

Le Système SAS

Résultats estimés des paramètres									
							Cohérent avec l'hétéroscédasticité		
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	4.31199	5.38153	0.80	0.4245	3.51181	1.23	0.2218
age	age	1	-0.00692	0.00289	-2.40	0.0180	0.00258	-2.68	0.0083
agesq	agesq	1	0.00003468	0.00002499	1.39	0.1676	0.00002109	1.64	0.1025
baths	baths	1	0.11543	0.04167	2.77	0.0064	0.04149	2.78	0.0062
cbd	cbd	1	-0.00001320	0.00001819	-0.73	0.4695	0.00001543	-0.86	0.3939
larea	larea	1	0.37622	0.07288	5.16	<.0001	0.08543	4.40	<.0001
ldist	ldist	1	0.16857	0.13011	1.30	0.1975	0.17612	0.96	0.3403
lintst	lintst	1	0.30783	1.32009	0.23	0.8160	0.84092	0.37	0.7149
lintstsq	lintstsq	1	-0.01205	0.08172	-0.15	0.8830	0.05340	-0.23	0.8219
lland	lland	1	0.08642	0.04050	2.13	0.0348	0.03568	2.42	0.0168
nbh	nbh	1	-0.01801	0.01015	-1.77	0.0784	0.01033	-1.74	0.0836
nearinc	nearinc	1	0.01186	0.08448	0.14	0.8886	0.11132	0.11	0.9153
rooms	rooms	1	0.02971	0.02887	1.03	0.3053	0.02773	1.07	0.2860
wind	wind	1	-0.01518	0.01381	-1.10	0.2739	0.01276	-1.19	0.2362

Tableau 15

Le Système SAS

Résultats estimés des paramètres			
Variable	**Libellé**	**DDL**	**Inflation de variance**
Intercept	Intercept	1	0
age	age	1	19.42409
agesq	agesq	1	13.24581
baths	baths	1	4.57468
cbd	cbd	1	93.60738
larea	larea	1	2.75183
ldist	ldist	1	14.15239
lintst	lintst	1	3648.06453
lintstsq	lintstsq	1	4629.48239
lland	lland	1	3.45857
nbh	nbh	1	2.01416
nearinc	nearinc	1	5.90596
rooms	rooms	1	2.31184
wind	wind	1	4.64491

Suite du tableau 15

4 METHODE REGRESSSION PCR (Principal Composantes Regression)

Le PCR ou le RCP (Régression sur Composantes Principales en Français) est la première méthode de régression de variables latentes que nous allons appliquer. Cette méthode consiste à établir des relations de causalité en maximisant les variances des variables exogènes en passant par l'analyse factorielles des données (ACP,...).

4.1 Procédure PCR

Table	WORK.TRAV1
Méthode d'extraction de facteurs	Régression des composantes principales
Nombre de variables de réponse	1
Nombre de paramètres du prédicteur	13
Gestion des valeurs manquantes	Exclude
Nombre de facteurs	13

Tableau 16

4.1.1 La part des facteurs dans l'explication des variables :

Nous avons au total 13 facteurs. L'axe 1 est la composante principale qui détient le pouvoir explicatif le plus élevé. Il explique à lui seul plus de la moitié (51.22%) des informations contenues dans les variables indépendantes et explique près de la moitié (47.79%) des informations contenues (la variance) dans la variable dépendante log(prix) [voir tableau17]. Quant à l'axe 2 et l'axe 3, ils expliquent respectivement 12.79% la variance des variables explicatives et 22% la variance de la variable log(prix) et 11.36% de la variance des variables indépendantes et 3.73% de la variance de variable dépendante log(prix). Par contre les trois premières composantes principales combinées (axe1, axe2, ax3) expliquent plus de 3/4(75.36%) l'inertie(variance) des variables indépendantes et plus de 2/3(73.57%) l'inertie de log(prix).

Ce qui veut dire que si nous retenons ces trois axes, on va passer d'un espace de 13 dimensions à un espace de trois dimensions. Ces trois dimensions sont orthogonales entres elles ; donc le problème de multi colinéarité est résolu et de plus nous aurons à expliquer plus de 2/3 des informations relatives aux variables indépendantes et à la variable log(prix).

Variation de pourcentage expliquée par Composantes principales				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	51.2236	51.2236	47.7911	47.7911
2	12.7876	64.0111	22.0506	69.8417
3	11.3565	75.3676	3.7311	73.5728
4	9.3112	84.6788	3.0834	76.6561
5	6.0957	90.7745	0.0160	76.6721
6	2.9902	93.7647	0.9897	77.6619
7	2.0106	95.7753	0.2075	77.8694
8	1.5303	97.3056	0.2042	78.0736
9	1.4233	98.7289	0.0009	78.0745
10	0.7852	99.5141	0.0580	78.1324
11	0.2648	99.7789	0.9635	79.0960
12	0.2202	99.9991	0.1678	79.2637
13	0.0009	100.0000	0.0049	79.2686

Tableau 17

4.1.2 Les courbes de l'évolution de R2 en fonction de nombre de facteurs

La figure2 est un graphique de l'analyse de coefficients de détermination R2. Il nous montre les courbes des évolutions de R2 de la variable dépendante en bleu pointillé et des variables indépendantes en bleu gras en fonction de nombre de facteurs. Sur le graphique nous remarquons une augmentation accélérée de R2 de la variable à expliquer avec le nombre de facteurs et commence à se décélérer complètement à partir de quatrième composante alors que le R2 de variables explicatives connait une croissance accélérée avec le nombre de facteurs depuis le premier facteur jusqu'au cinquième et décélère par la suite. De deuxième au quatrième facteur, les deux courbes se sont coupées deux fois [Voir figure2 ci-dessous]. Ce qui signifie que le nombre de facteurs optimal à retenir pour estimer la PCR sera déterminé à partir de ces deux points d'intersection des courbes. Par conséquent, le nombre de composantes optimale se sera situé entre l'intervalle de [2-4[facteurs. Ce nombre exact de facteur sera déterminé par le test de validation croisée dans la partie B de la section IV.

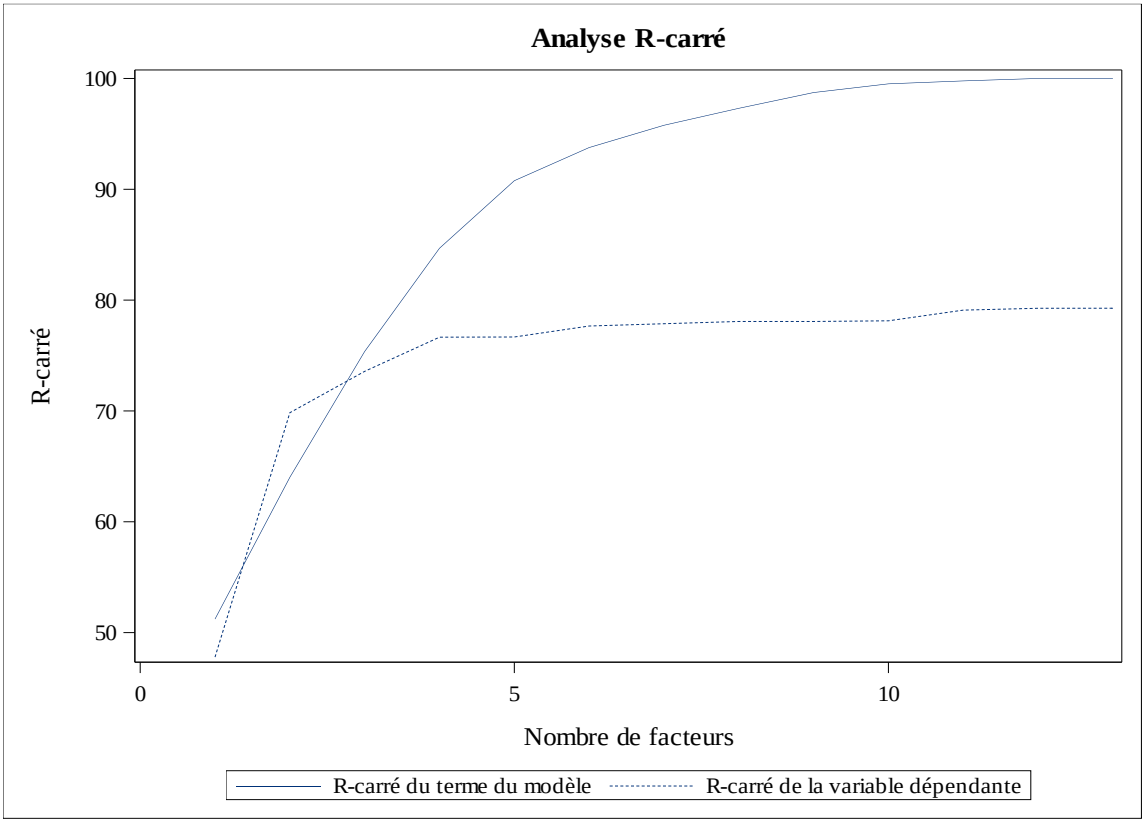


Figure 1

4.1.3 Cercle de corrélation : Mise en évidence des relations entre les variables

Ce cercle appelé cercle de corrélation en ACP, nous montre les degrés de relations entre les variables. Plus une variable est proche du plus grand cercle, mieux elle est représentée par les axes. Sur ce premier plan factoriel (axe1, axe2), nous avons les variables explicatives en bleu, les numéros des observations en vert et la variable expliquée en rouge.
Les deux premiers axes ensemble explique 64% la variance des variables indépendantes et 69.84% la variance de log(prix) dont le premier axe explique 51.2% la variance des x et 47.8% l'information relative à log(prix) tandis que l'axe2 explique 12.8% l'inertie des variables exogènes et 22.1% l'inertie de la variable endogène.
On note une corrélation positive d'un côté (à droite) et une corrélation négative de l'autre coté (à gauche) le long de l'axe1. La variable dépendante log(prix) est positivement corrélée avec l'axe1 et négativement corrélé à l'axe2. La variable dépendante est fortement corrélée à la variable nombre de bains. Les variables distance par rapport à l'autoroute, distance par rapport à l'incinérateur, distance par rapport à une station de bus, la superficie du terrain, le nombre de bains, le prix de vente sont fortes positivement corrélées entre elles le long de l'axe1 tandis que la variable âge de la maison et les maisons proches de l'incinérateur sont négativement corrélée le long de l'axe1. Nous pouvons donc déduire que l'axe1 est défini par une opposition entre les variables âge de la maison et les variables distance par rapport à l'incinérateur, par rapport à l'autoroute, par rapport à une station de bus, l'aire du terrain. Par contre nous pouvons observer que peu de variables sont fortement corrélées à laxe2 dans ce premier factoriel. Seule la variable nombre de bains et prix de vente sont fortes négativement corrélées avec l'axe2. Nous pouvons donc dire que l'axe2 est défini par un faible nombre de bains.

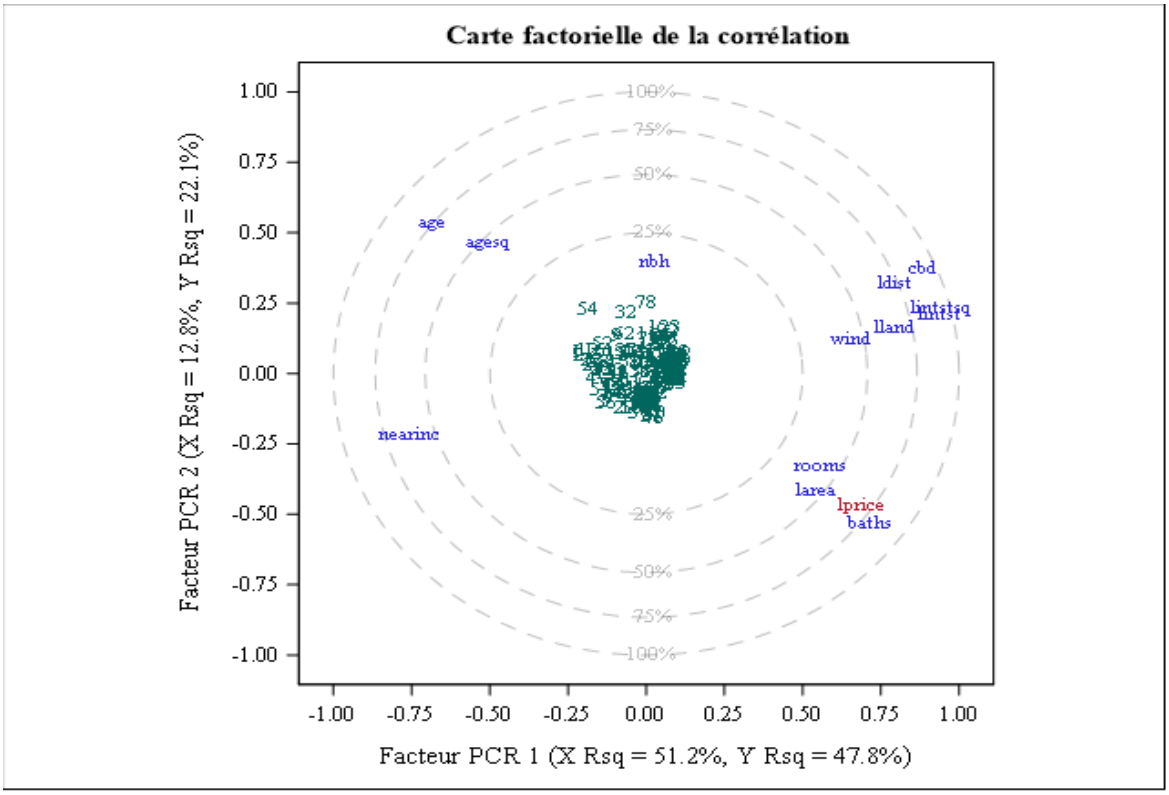


Figure 2

4.1.4 Les loadings ou les poids factoriels des effets du modèle :

Encore appelés les chargements des effets du modèle, ils sont en quelques sortes les corrélations entre les composantes principales et les variables. Ils permettent d'interpréter les variables par les facteurs et mesurent la qualité de représentation d'une variable par les composantes principales. Plus le loading est proche de 1, mieux cette variable est représentée par l'axe factoriel. On suppose souvent qu'un loading en valeur absolue supérieure à 4 indique une liaison significative. Il est déterminé par la lecture horizontale du tableau de poids factoriels des effets du modèle.
Si nous prenons comme exemple la variable âge, c'est laxe11 qui la représente le mieux car son loading est le plus élevé (0.705), suivi de l'axe2 qui représente à 0.419 la variable âge [voir tableau18]. Par contre la variable nombre de bains est le mieux représentée par le facteur9 qui la représente à 0.777 ;le second axe qui la représente le mieux est l'axe2(-0.413).

Poids factoriels des effets du modèle											
Nombre facteur extraits	**age**	**agesq**	**baths**	**cbd**	**larea**	**ldist**	**lintst**	**lintstsq**	**lland**	**nbh**	**nearinc**
1	-0.263030	-0.194334	0.278829	0.342661	0.211540	0.309261	0.363868	0.364756	0.307647	0.011220	-0.292830
2	0.419240	0.363499	-0.413420	0.288532	-0.320137	0.249751	0.164657	0.181429	0.126771	0.310074	-0.167485
3	0.335835	0.512751	0.175058	-0.039808	0.483188	-0.134519	-0.079166	-0.071333	0.165064	0.241825	0.129683
4	-0.200662	-0.290378	-0.062097	0.002432	-0.031127	-0.220242	0.001155	0.011467	-0.003914	0.646956	0.414367
5	0.080518	0.101620	-0.105295	-0.152070	-0.139850	-0.398130	0.146597	0.123609	0.437455	-0.536994	0.153998
6	0.122299	0.048981	0.258741	0.034424	0.539561	-0.006606	0.043534	0.047158	0.011655	-0.009789	0.021715
7	-0.058667	-0.055396	0.009217	0.218843	-0.079513	0.122420	-0.017170	0.000779	0.577054	0.017408	0.561876
8	0.052855	0.095434	-0.055424	0.315635	0.064372	-0.195936	0.392963	0.389237	-0.550443	-0.182838	0.416551
9	0.026736	0.269420	0.777086	0.037906	-0.542240	-0.042226	0.002399	0.004105	-0.071546	0.081557	0.048166
10	-0.018273	0.071091	-0.007930	0.330240	0.044040	0.525505	-0.350581	-0.271779	-0.164226	-0.309473	0.330506
11	0.704998	-0.600012	0.164069	0.247805	-0.031324	-0.145769	-0.116050	-0.044266	0.013986	-0.045491	-0.055740
12	0.275704	-0.142004	0.069284	-0.667158	-0.014918	0.513946	0.288896	0.172696	-0.047478	0.003363	0.263842
13	0.015572	-0.012151	0.001567	0.094403	0.002716	-0.011140	0.660513	-0.744408	-0.000353	-0.000588	-0.000884

Tableau 18

4.1.5 Les poids des effets du modèle ou les weights :

Ils sont très proches de loadings mais différents de ceux-ci. Ils sont en quelques sorte les contributions des variables à la définition des axes et indiquent la participation des variables à la construction des composantes principales. Ils sont également assimilables aux coordonnées car ils peuvent être négatifs. Plus le weight est élevé plus la variable contribue à la construction de l'axe. Pour le déterminer, on effectue la lecture horizontale du tableau [voir tableau19 ci-dessous]. Les variables distance par rapport à l'autoroute (0.36), par rapport à une station de bus (0.34), par rapport à l'incinérateur (0.31), la surface du terrain (0.31), la variable maison proche (-0.29) et l'âge (-0.26) sont les variables les plus contributrices à la définition de l'axe1 ; ce résultat vient de confirmer les ceux de cercle de corrélation étudiés précédemment. Par contre la deuxième composante est définie majoritairement par l'âge, le nombre de bains, la surface de la maison et le nombre connaissance dans le quartier. Les variables âge et le nombre de connaissance contribuant respectivement à 0.42 et à 0.31 s'oppose aux variables nombre de bains et la superficie de la maison contribuant respectivement à -0.41 et à -0.32 à la construction de la deuxième composante. Pour le facteur3, il est principalement construit par la coordination de trois variables : l'agesq (0.51), la superficie de la maison (0.48), nombre de chambres (0.46). Nous pouvons alors déduire que l'axe3 désigne les maisons vieilles, vaste et au nombre de chambre élevé.

Poids des effets du modèle											
Nbrefacteurs extraits	**age**	**agesq**	**baths**	**cbd**	**larea**	**ldist**	**lintst**	**lintstsq**	**lland**	**nbh**	**nearinc**
1	-0.263030	-0.194334	0.278829	0.342661	0.211540	0.309261	0.363868	0.364756	0.307647	0.011220	-0.292830
2	0.419240	0.363499	-0.413420	0.288532	-0.320137	0.249751	0.164657	0.181429	0.126771	0.310074	-0.167485
3	0.335835	0.512751	0.175058	-0.039808	0.483188	-0.134519	-0.079166	-0.071333	0.165064	0.241825	0.129683
4	-0.200662	-0.290378	-0.062097	0.002432	-0.031127	-0.220242	0.001155	0.011467	-0.003914	0.646956	0.414367
5	0.080518	0.101620	-0.105295	-0.152070	-0.139850	-0.398130	0.146597	0.123609	0.437455	-0.536994	0.153998
6	0.122299	0.048981	0.258741	0.034424	0.539561	-0.006606	0.043534	0.047158	0.011655	-0.009789	0.021715

Tableau 19

Poids des effets du modèle		
Nombre facteurs extraits	**Rooms**	**wind**
1	0.216288	0.254453
2	-0.251791	0.096857
3	0.458838	0.126787
4	-0.120017	0.466125
5	-0.117055	0.465829
6	-0.784285	-0.058462

Suite de Tableau 19

4.1.6 L'importance des variables dans l'explication de log(prix) :

Le graphique ci-dessous nous montre le poids des variables indépendante dans l'explication de la variable à expliquer. Plus la tige représentant une variable est longue, plus elle est importante dans l'explication de la variable dépendante. La barre horizontale détermine (et passe par) la valeur (le seuil) critique à partir de laquelle les variables indépendantes participent de manière significative à l'explication de la variable dépendante selon le critère de wold. Ici ce seuil est de 0.8 et la variable nombre de connaissance est la seule variable en dessous de ce seuil et donc ne participe pas de manière significative à l'explication des prix de vente des maisons. La variable la plus importante dans l'explication de prix de vente des maisons est l'âge de la maison car elle est la plus longue dans le graphique. Les variables nombre de bains, distance par rapport à une station de bus, la distance par rapport à un incinérateur, la distance par rapport à l'autoroute sont également très importantes dans l'explication du prix de vente des maisons [voir figure3 ci-dessous].

Le Système SAS

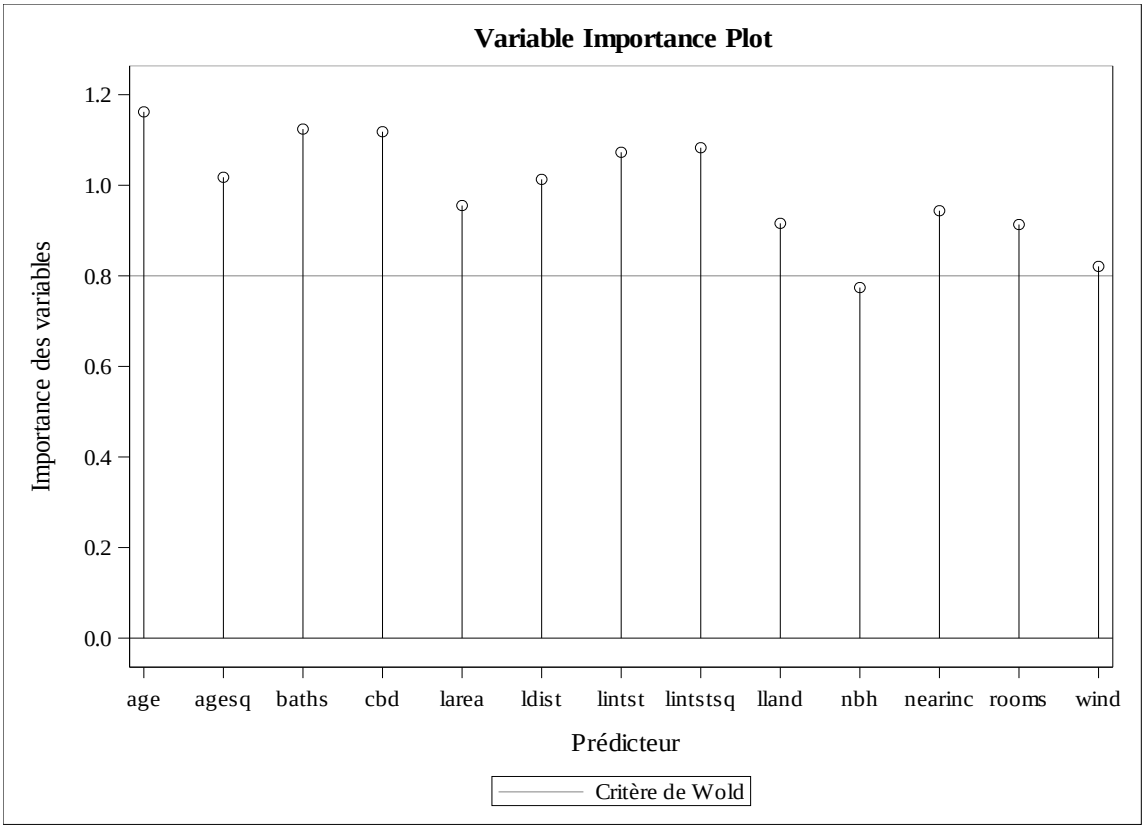


Figure 3

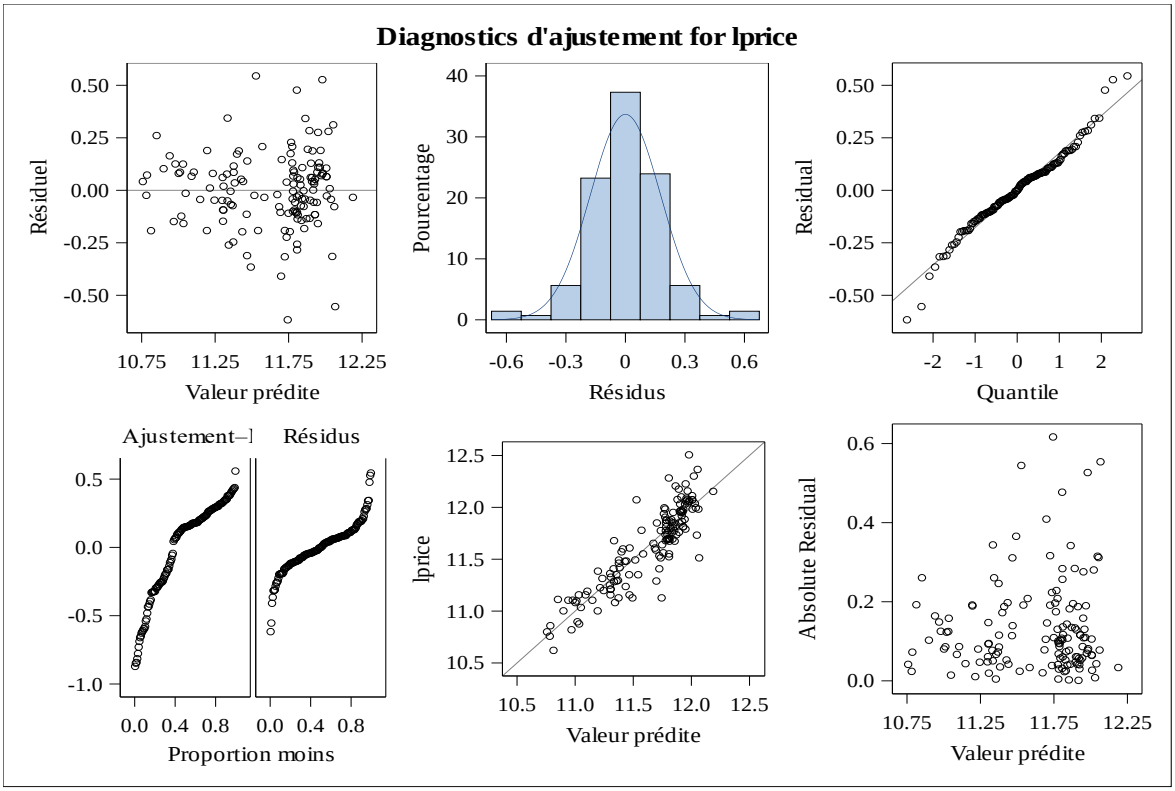


Figure 4

4.1.7 Test de validation croisée :

4.1.7.1 Méthode PRESS

Le test de validation croisée est un test statistique qui permet de déterminer le nombre de facteurs optimal à retenir dans l'application des méthodes de régression sur des variables latentes. Il se base sur la qualité d'approximation du modèle des valeurs de la variable expliquée pour des individus sur lequel est fondé le modèle. Il existe plusieurs tests ou critères de validation croisée dont les plus connues sont PRESS (Predicted Residuals Sum of Squares) et RSS(Residuals Sum of Squares). Le premier test sera celui adopté pour notre test de validation croisée ici. Le critère de validité de la méthode PRESS est de retenir le nombre le nombre de facteurs significatif prédictif au seuil de 10%. C'est-à-dire le nombre de facteurs optimal à retenir est le plus petit nombre de facteur à partir duquel la prob>T**2 est supérieure à 10%. Dans notre étude ici, ce nombre de facteur est 3[voir tableaux 20, 21 et 22]. Le nombre de facteur optimal approuvé par le test de validation pour notre étude est donc 3. Ce nombre est très important pour la suite de notre travail, il va nous permettre d'estimer les paramètres de régression par PCR.

Le Système SAS

Table	WORK.TRAV1
Méthode d'extraction de facteurs	Régression des composantes principales
Nombre de variables de réponse	1
Nombre de paramètres du prédicteur	13
Gestion des valeurs manquantes	Exclude
Nombre maximal de facteurs	13
Validation Method	7-fold Split-sample Validation
Validation Testing Criterion	Prob T**2 > 0.1
Number of Random Permutations	1000

Tableau 20

Split-sample Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	**Racine carrée du PRESS moyen**	**T**2**	**Prob > T**2**
0	1.077497	42.73388	<.0001
1	0.785972	20.3163	<.0001
2	0.660467	5.994842	0.0120
3	0.57902	2.685537	0.1060
4	0.54568	0.392632	0.5640

Tableau 21

Racine carrée du PRESS moyen min.	0.5386
Réduction du nombre de facteurs	6
Le plus petit nombre de facteurs avec p > 0.1	3

Tableau 22

4.1.8 La part des composantes dans l'explication des variables :

Variation de pourcentage expliquée par Composantes principales													
Nombre facteurs extraits	**Effets du modèle**												
	age	**agesq**	**baths**	**cbd**	**larea**	**ldist**	**lintst**	**lintstsq**	**lland**	**nbh**	**nearinc**	**rooms**	**wind**
1	46.0707	25.1484	51.7713	78.1883	29.7987	63.6888	88.1660	88.5966	63.0258	0.0838	57.1012	31.1516	43.1150
2	75.2891	47.1137	80.1841	92.0278	46.8360	74.0580	92.6730	94.0686	65.6974	16.0670	61.7644	41.6909	44.6746
3	91.9400	85.9286	84.7084	92.2618	81.3041	76.7295	93.5983	94.8198	69.7198	24.7005	64.2473	72.7725	47.0478

Tableau 23

Si nous retenons trois composantes principales comme nous l'a recommandé le test de validation croisée effectué précédemment, elles expliquent ensemble plus de 75% la variance de variables explicatives et plus de 73% la variance de log de prix de vente des maisons. En effet l'axe1 explique à lui seul plus de la moitié (51.22%) des informations contenues dans les variables indépendantes et explique à hauteur de 47.79% les informations relatives à la variable dépendante. Les deux derniers axes ensemble contribue moins à l'explication des variables par rapport à l'axe1 ; ils contribuent à 24.14% à expliquer la variabilité des variables indépendantes et à 25.78% expliquer la variance de log(prix) dont l'axe2 a une contribution de 12.79% dans l'explication des variables explicatives et 22% dans l'explication de la variable expliquée. Par contre l'axe3 se retrouve avec une contribution de 11.36% pour les variables explicatives et de 3.73% seulement pour la variable expliquée.

Variation de pourcentage expliquée par Composantes principales					
Nombre facteurs extraits	Effets du modèle		Variables dépendantes		
	En cours	Total	lprice	En cours	Total
1	51.2236	51.2236	47.7911	47.7911	47.7911
2	12.7876	64.0111	69.8417	22.0506	69.8417
3	11.3565	75.3676	73.5728	3.7311	73.5728

Suite du Tableau 23

4.1.9 Estimation de coefficients de régression par la méthode PCR

Variation de pourcentage expliquée par Composantes principales				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	51.2236	51.2236	47.7911	47.7911
2	12.7876	64.0111	22.0506	69.8417
3	11.3565	75.3676	3.7311	73.5728

Tableau 24

Cette section nous donne le tableau [Voir tableau26] de coefficients de régression issue de l'application de la méthode PCR sur nos variables. Nous pouvons donc écrire à partir dudit tableau l'équation de régression PCR suivante :

```
Log(prix)=-0.1698age-0.1029agesq     +0.2531baths-0.0196cbd+0.2501Larea     -
0.02950Ldist+0.0249Lintst+0.0203Llintstsq       +0.0625Lland       -0.0715nbh
+0.0032nearinc+0.2226rooms +0.0531wind.
```

D'après l'équation de régression, tous les coefficients possèdent le signe attendu sauf la distance par rapport à l'incinérateur et la variable binaire la maison proche. Selon la logique personne ne peut accepter payer davantage pour vivre tout près d'une machine(l'incinerateur) qui produit de fumée et du bruit. Cependant toutes les variables n'ont pas les mêmes degrés d'impact sur notre variable dépendante. La variable nombre de salles de bains est la variable qui influence le plus fortement et positivement le prix de vente des maisons. Toute chose égale par ailleurs, une unité supplémentaire de bain entraine une augmentation du prix de vente des maisons de 25.31%. La deuxième variable qui possède une plus forte influence est la variable la superficie de la maison. Toute comme la variable nombre de salles de bains, elle influe positivement le prix de vente des maisons ; toute augmentation de 1%(1% de pieds carré) de la superficie des maisons s'ensuit avec une hausse de 25.01% le prix de vente des maisons, ceteris paribus. Le nombre de chambre est également une variable très capitale dans la variation des prix des maisons, elle évolue dans le même sens que le prix des maisons. Le prix des maisons augmente de 22.26% suite à une augmentation d'une unité du nombre de chambre. L'âge de la maison constitue également un facteur non négligeable dans la fluctuation des prix des maisons. Contrairement aux variables précitées, l'âge de la maison exerce un impact négatif sur les prix de vente des maisons ; toute chose égale par ailleurs, si l'âge des maisons passe de 1 à 3 ans, le prix de la maison connait une diminution de prix des maisons de 1.16%.

Par ailleurs les autres variables comme la distance par rapport à une station de bus, la distance par rapport l'autoroute, la distance par rapport à un incinérateur, la superficie du terrain, le nombre de connaissances (qui ne contribue pas significativement à l'explication de prix des maisons selon le critères de Wold), la variable binaire nearinc(maison proche) impactent faiblement le prix et leur impact n'excède pas 10% sur le prix des maisons. D'un côté, nous avons des variables la distance par rapport à une station de bus, la distance par à l'autoroute le nombre de connaissances ont des impacts négatifs. Comme exemple d'interprétation, l'augmentation de distance par rapport à l'autoroute de 1000 pieds à 1300 pieds entraine l'augmentation de prix des immeubles de 0.08% toute chose égale par ailleurs. Et de l'autre côté, des variables la distance par rapport à l'incinérateur, la superficie du terrain, le pourcentage de vent incinérateur, la maison proche exercent une influence positive sur le prix des maisons. Parmi cette dernière catégorie variable, c'est la variable binaire maison proche qui a un impact le plus marginal (faible) sur le prix des maisons. En supposant que les autres variables ne varient pas, une maison proche de l'incinérateur est plus chère qu'une maison éloignée de l'incinérateur. La maison proche de l'incinérateur coute en moyenne plus de 0.32% de plus qu'une maison éloignée de l'incinérateur.

Le Système SAS

Paramètres estimés pour données centrées et mises à l'échelle	
	lprice
Intercept	0.0000000000
age	-.1697647620
agesq	-.1029354038
baths	0.2530961102
cbd	-.0196155507
larea	0.2500797538
ldist	-.0294953564
lintst	0.0249251105
lintstsq	0.0202996976
lland	0.0624876507
nbh	-.0714806883
nearinc	0.0031669190
rooms	0.2225891505
wind	0.0530471768

Tableau 25

Résultats estimés des paramètres	
	lprice
Intercept	8.307035130
age	-0.002765393
agesq	-0.000017562
baths	0.122530375
cbd	-0.000000917
larea	0.272992726
ldist	-0.025348339
lintst	0.013536352
lintstsq	0.000605798
lland	0.033815170
nbh	-0.012702682
nearinc	0.002735497
rooms	0.105015615
wind	0.008447616

Tableau 26

5 METHODE DE REGRESSION PLS (Partial Least Squares):

Cette partie va se focaliser sur la régression sur les variables par la méthode PLS qui signifie en Anglais Partial Least Square et qui veut dire en français Méthode des Moindres Carrées Partiels. Elle est une technique statistique qui permet de quantifier les relations plusieurs variables. Principalement conçue pour répondre aux problèmes d'insuffisance liées à l'usage des méthodes de régressions classiques. En plus la méthode des régressions multiples trouve ses limites dans le fait que les coefficients de régression sont instables avec les problèmes de multicolinéarités quand les variables explicatives sont fortement corrélées, l'impossibilité de l'appliquer si le nombre de variables explicatives est supérieur au nombre d'observations et s'il y a plusieurs variables à expliquer à mettre dans le modèle ; et elle se contente de maximiser la variance de variable dépendante. Par contre la méthode PCR se révèle peu robuste du fait qu'elle se fonde essentiellement sur la maximisation de variance des variables indépendantes comme critère de choix. Ainsi pour palier à toutes ces limites, PLS fut conçue et qui apparait plus robuste. Contrairement à PCR, elle base fondamentalement sur le critère de maximisation de covariance entre les variables explicatives et les variables expliquée en passant par l'analyse factorielle des données (ACP,..). Elle nous donne la possibilité d'appliquer et de continuer la régression multiple en cas de plusieurs variables à expliquer, en présence de problème de multicolinéarité et au cas de

faible degré de liberté. Il existe deux types de PLS : PLS1 encore appelé PLS simple, elle est appliquée lorsqu'il y a une seule variable expliquée et PLS2 qui est appliquée si on a plusieurs variables expliquées.

5.1 Procedure PLS

5.1.1 La part expliquée par les facteurs :

L'axe1 est celui qui contribue le plus à l'explication des informations contenues dans les variables et détient un pouvoir explicatif plus élevé comparativement à la méthode PCR. Il explique à lui seul près de la moitié (49.92%) la variance des variables indépendantes et plus de la moitié (60.64%) l'inertie de la variable dépendante. Et on remarque qu'il détient un pouvoir explicatif plus important pour la variable expliquée comparativement au cas de PCR qui se retrouve avec un pouvoir explicatif de l'axe1 pour la variable dépendante de 47.79%.
Par contre l'axe2 détient un pouvoir explicatif relativement faible par rapport à l'axe1. Il explique environ 13.64% de la variance des x et 16.56% de la variance de y [Voir tableau27].
Si nous retenons les deux premiers axes, nous allons expliquer 63.56% l'information contenue dans les variables explicatives et 77.20% l'information contenue dans le log(prix). Comparativement aux deux premiers facteurs de la PCR qui expliquent globalement 64.01% l'information contenue dans les variables explicatives et 69.79% l'information contenue dans le log(prix).
L'axe3 n'explique pas grande chose surtout pour la variable endogène et si on se limite aux trois premières composantes principales, on expliquerait plus de 2/3(plus précisément 71.76%) de la variance des variables exogènes et plus de 3/4(plus précisément 77.91%) de la variance de log(prix). Il apparait clair que la méthode PLS possède un pouvoir explicatif plus élevé pour la variable expliquée que la méthode PCR car dans tous les cas (1 ou 2 ou 3 axes retenus) elle explique une part plus importante de y mais en matière des variables explicatives, elle est légèrement dépassée par la méthode PCR car en terme de pouvoir explicatif, elle détient une part d'explication plus faible dans tous les cas (1 ou 2 ou axes retenus).

Le Système SAS

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	49.9173	49.9173	60.6434	60.6434
2	13.6410	63.5583	16.5593	77.2027
3	8.2043	71.7627	0.7147	77.9174
4	5.1500	76.9126	0.4754	78.3929
5	9.0371	85.9498	0.1421	78.5350
6	2.8151	88.7649	0.3966	78.9316
7	3.8086	92.5734	0.1883	79.1199
8	3.1995	95.7730	0.1196	79.2395
9	1.4699	97.2428	0.0195	79.2590
10	1.0886	98.3314	0.0040	79.2630
11	0.2625	98.5939	0.0009	79.2638
12	1.4048	99.9987	0.0000	79.2639
13	0.0013	100.0000	0.0048	79.2686

Tableau 27

5.1.2 Les courbes de l'évolution des R2 avec le nombre de facteurs retenus

Sur ce graphique[ci-dessous], on note une évolution croissante rapide de R2 du terme du modèle entre 0 et 2 facteurs et la courbe devient quasi constante entre 3 et 13 facteurs. Par contre la courbe de R2 de log(prix) connait une allure croissante avec le nombre de facteurs entre l'intervalle de facteurs [1 à 5] et cette évolution commence par faiblir après et tend à devenir constante avec le nombre de facteurs. Ce qui implique que le nombre de facteurs optimal à retenir pour effectuer la régression PLS se trouve entre 2 à 5 facteurs [voir sur la figure5]. Ce pendant le test de validation croisée effectué dans la partie B de cette section va nous permettre de déterminer le nombre de facteurs optimal.

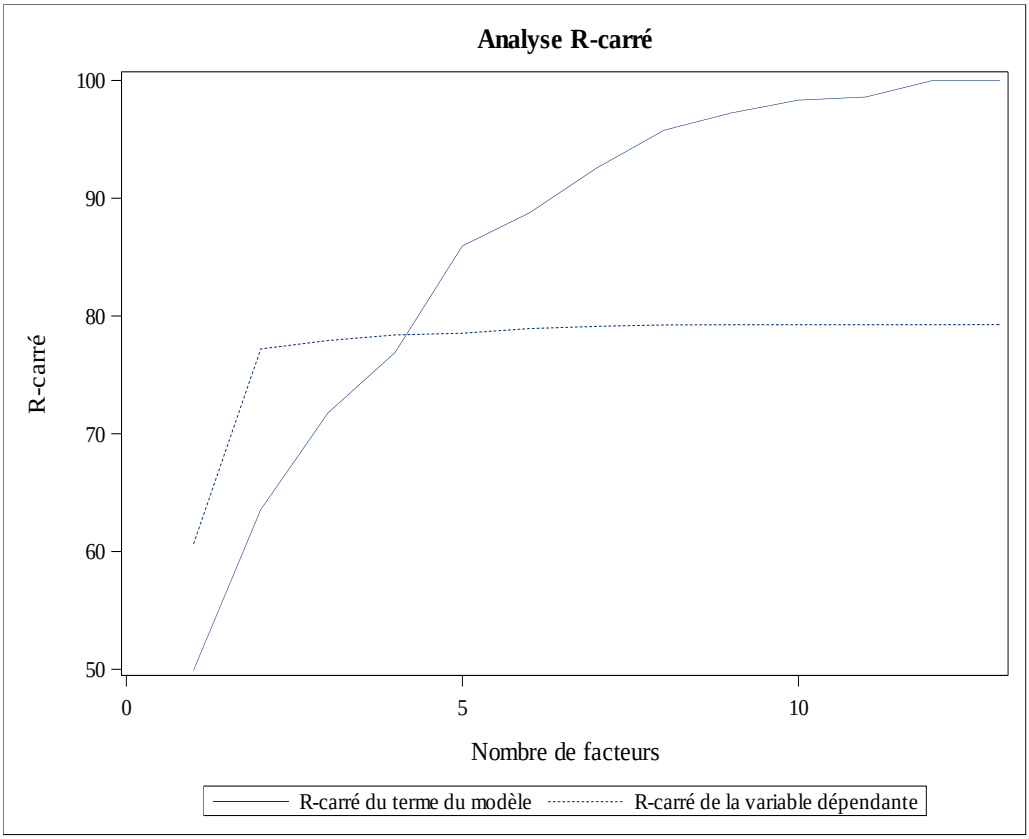


Figure 5

5.1.3 Cercle de corrélation : Mise en évidence des relations entre les variables

Ce cercle de corrélation nous présente le plan premier factoriel contenant l'axe1 et l'axe2. Il vient de confirmer ce que nous avons dire dans la partie part expliquée par les facteurs. Le premier facteur explique 49.9% des information relatives aux variables indépendantes et 60.6% des informations relatives à la variable dépendante. Cependant, l'axe2 se retrouve avec une part explicative de 13.6% des variables exogène contre 16.6% pour la variable endogène.

Par ailleurs, on remarque sur ce cercle, des variables explicatives en bleu et la variable expliquée en rouge et les observations en vert concentré au milieu du cercle. On note également un regroupe des variables dans le cercle. Le haut droit du cercle est marqué par une forte corrélation entre le log(prix), la superficie de la maison, le nombre de salles de bains, le nombre de chambre et ce regroupement de variable est corrélé positivement avec l'axe1 et l'axe2.

Et le bas droit du cercle par contre, est marqué par une forte liaison entre la superficie du terrain, la distance par rapport à l'autoroute, par rapport à l'incinérateur, par rapport à une station de bus et au pourcentage de vent. Ce groupe de variables sont d'un part corrélé positivement le long de l'axe1 et négativement corrélé le long de l'axe2 de l'autre part. Et à gauche du cercle, on a les variables maisons proches et l'âge des maisons sont négativement corrélé le long de l'axe1 mais qui sont non significativement bien représentées par l'axe2. Par conséquent, nous pouvons dire que l'axe1 désigne l'opposition entre les maisons proches, âge des maisons d'un côté et la superficie des terrains, la distance par rapport à l'incinérateurs, le nombre de salles de bains, la distance par rapport à une station de bus de l'autre côté tandis que l'axe2 peut être défini comme l'opposition entre la superficie des maisons, le nombre de bains d'une part et la distance par rapport à une station de bus, et la distance par rapport à l'autoroute d'autre part.
Cependant, comparativement à la méthode PCR, nous remarquons une tendance renversée des choses. Sur le cercle de corrélation de PCR, le groupe de variable{larea, baths, price, rooms} se trouvait en bas droit du cercle positivement corrélé avec l'axe1 mais négativement corrélé avec l'axe2 ; alors que le groupe de variables{lland, wind, ldist lintst, cbd} se situe en haut droit du cercle positivement corrélé le long de l'axe1 et l'axe2 ; pour le cas des variables{nearinc, age}, la variable âge se situait à gauche au-dessus de l'axe1 et la variable nearinc se trouvait à gauche en dessous de l'axe2[figure2]. Alors que nous observons l'inverse sur ce cercle PLS[figure6].

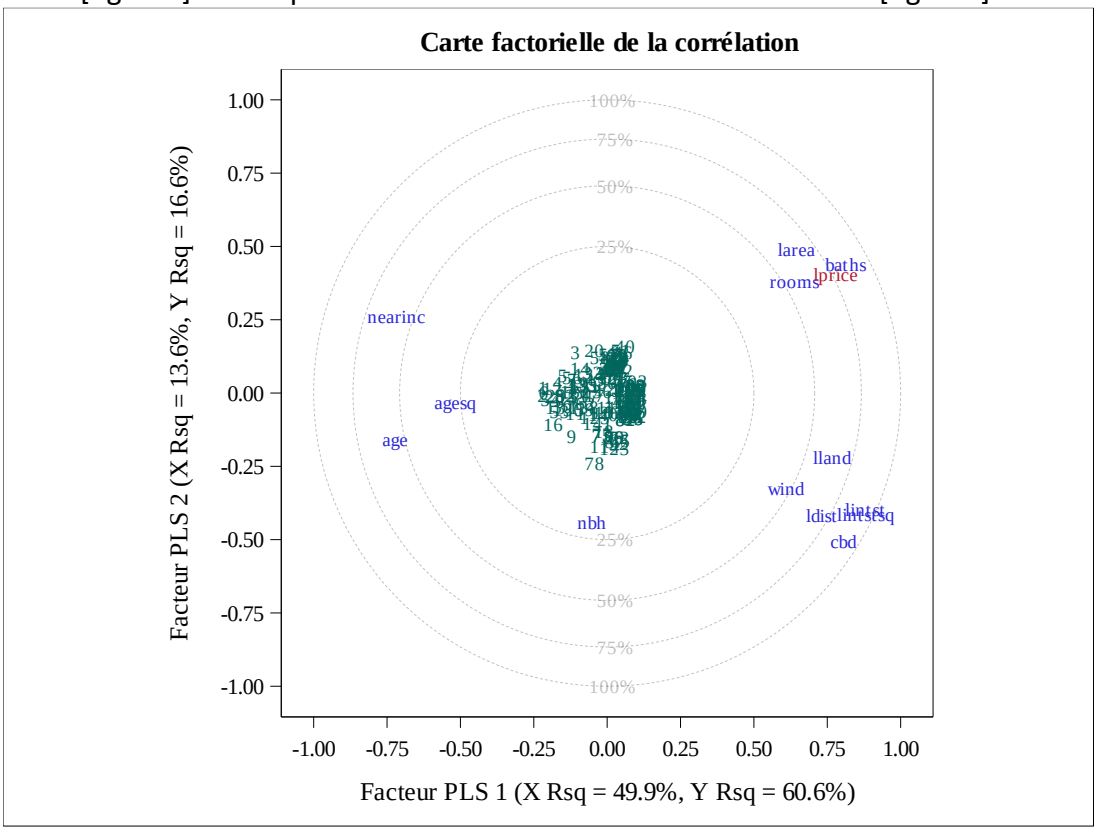


Figure 6

5.1.4 Les loadings ou les poids factoriels des effets du modèle :

Comme déjà mentionné ci-haut, les loadings sont en quelques sortes les coordonnées des variables sur les axes et représente la qualité de représentation des variables par les axes. Ils constituent les combinaisons entre les variables et les axes et sont déterminés par une lecture verticale de tableau de poids factoriels des effets du modèle [voir

tableau28]. Par exemple si nous considère les variables âge des maisons et le nombre de salles de bains, nous aurons ces types de liaisons entre les deux variables et les composantes principales ci-dessous.

Age=-0.28 facteur1-0.12facteur2+0.47facteur3-0.39facteur4+0.23facteur5+……-0.08facteur13
Baths=0.32facteur1+0.33facteur2+0.01facteur3+0.14facteur4+0.03facteur5+……..-0.40facteur13

Plus les loadings sont élevés mieux les variables sont bien représentées.
La variable âge est le mieux représenté par l'axe11, suivi de l'axe3, l'axe4, l'axe6 et l'axe1. L'axe1 représente l'âge à -028 alors que l'axe11 la représente à 0.49. Pour la variable nombre de salles de bains, c'est seulement quatre facteurs qui la représentent de manière significative au seuil de 0.25. Elle possède une qualité de représentation par rapport à l'axe12 de 0.78 alors qu'elle est représentée respectivement par les axe13, axe2, axe1 à -0.40, à 0.33 et à 0.32

Poids factoriels des effets du modèle										
Nombre de facteurs extraits	**Age**	**agesq**	**baths**	**cbd**	**larea**	**ldist**	**lintst**	**lintstsq**	**lland**	**nbh**
1	-0.283506	-0.203622	0.319448	0.315822	0.253452	0.287701	0.346003	0.345521	0.300721	-0.020173
2	-0.117510	-0.022159	0.329378	-0.379643	0.367878	-0.311990	-0.296220	-0.310443	-0.163601	-0.330204
3	0.471271	0.562995	0.004903	0.129219	0.202413	0.279148	0.035850	0.040393	0.174717	-0.222676
4	-0.390494	-0.534922	0.139274	-0.066427	0.077537	-0.037883	-0.026649	-0.030788	0.035461	0.110085
5	0.229009	0.361345	0.028561	-0.041402	0.306382	-0.198480	-0.040575	-0.033763	0.323482	0.542461
6	-0.308309	-0.160864	-0.191870	-0.233508	-0.535205	0.130762	-0.000038337	-0.042010	0.300403	-0.480337
7	-0.115151	0.022063	0.089965	0.052491	0.214545	0.303650	-0.044827	-0.044069	-0.594476	0.600138
8	0.054076	0.215646	-0.050388	-0.158877	0.038398	-0.380601	0.282807	0.242490	0.036665	-0.556610
9	-0.101563	0.068366	-0.017713	0.402426	-0.086043	0.300811	0.035812	0.068291	0.044248	-0.105127
10	-0.012119	0.019186	0.012127	-0.141516	0.058532	0.449689	-0.422892	-0.390356	-0.034668	-0.109538
11	0.493587	-0.431314	-0.108240	-0.505514	0.120797	0.224826	0.336968	0.246040	-0.003722	0.023844
12	0.051134	0.241971	0.784484	-0.016955	-0.546311	-0.025172	-0.002151	-0.006706	-0.037376	0.100412
13	-0.077845	-0.085376	-0.404108	0.161122	0.277321	-0.046286	0.531269	-0.655598	0.022218	-0.051747

Tableau 28

5.1.5 Les weights ou les poids des effets du modèle

Les poids sont les coefficients de combinaisons entre les variables et les facteurs. Ils correspondent aux contributions des variables à la définition des facteurs. Ils sont obtenus par la lecture horizontale du tableau ci-dessous. Comme exemple, nous avons ces types de relations entre l'axe1 et l'axe2 le cas suivant :

Facteur1=-033age-0.21agesq+0.44baths+0.23cbd+0.39larea+……+0.17Wind
Facteur2=-0.09age+0.04agesq+0.34baths-0.39cbd+0.41larea+……-0.30Wind

Les variables qui contribuent les plus à construire l'axes1 sont le nombre de salles de bains, la superficie des maisons, âge des maisons. Les variables nombre de salles, superficie des maisons, nombre de chambres et âge des maisons y contribuent respectivement à 0.44, à 0.39, à 0.34 et à -0.33 à la définition de l'axe1. L'axe1 peut designer alors l'opposition entre les maisons moins âgées et celles vastes au nombre de salles de bains et de chambres élevés.

Par contre, environ cinq variables participent de manières significatives à la construction de l'axe2. La variable qui contribue le plus est la superficie des maisons ; elle contribue à hauteur de 0.41 à la définition de l'axe2 ; la deuxième variable la plus contributrice est la distance par rapport à une station de bus qui y participe à -0.39. Les autres variables les plus importantes à la construction du deuxième facteur sont le nombre de connaissances (0.34), le nombre de salles de bains (0.34), la distance par rapport à l'autoroute au carré (-0.31) et enfin le pourcentage de vent (0.30). L'axe2 oppose donc les variables {la superficie des maisons, nombre de salles de bains} corrélées positivement avec l'axe2 d'un côté et les variables {la distance par rapport à une station de bus, nombre de connaissances, nombre de salles de bains, la distance par rapport à l'autoroute au carré et le pourcentage de vent} corrélé négativement avec l'axe2 de l'autre côté.

Poids des effets du modèle											
Nombre de facteurs extraits	**age**	**agesq**	**baths**	**cbd**	**larea**	**ldist**	**lintst**	**lintstsq**	**lland**	**nbh**	**Nearinc**
1	-0.332939	-0.208118	0.443845	0.227190	0.393475	0.230956	0.288424	0.282670	0.289758	-0.121463	-0.253527
2	-0.093230	0.040100	0.342234	-0.392283	0.414214	-0.274508	-0.293315	-0.311383	-0.119998	-0.344206	0.174483
3	0.239182	0.568332	0.063551	-0.053496	0.361050	0.390278	0.074084	0.041490	0.421884	-0.073808	-0.243454
4	-0.612168	-0.356723	0.074570	-0.325692	0.079481	-0.033076	0.027514	-0.024647	0.214728	0.341218	0.338164
5	-0.302938	0.858268	-0.271836	-0.636961	-0.051015	0.040138	0.162417	0.038521	0.448004	0.531011	0.209635
6	-0.558409	-0.014747	-0.221389	-0.347518	-0.517385	0.339767	0.169439	0.078345	-0.220884	-0.513217	-0.344843

Tableau 29

Poids des effets du modèle			
Nombre de facteurs extraits	**rooms**	**wind**	**Coefficients de régression interne**
1	0.336334	0.170233	0.305699
2	0.226910	-0.302854	0.305581
3	-0.575533	-0.486371	0.081860
4	-0.610186	0.049489	0.084269
5	0.699831	0.285791	0.034779
6	0.059826	-0.373382	0.104100

Suite de la tableau 29

5.1.6 L'importance des variables dans l'explication de log(prix) selon le critère de Wold:

Par ce graphique, nous pouvons déceler les variables qui sont importantes dans la construction de notre modèle. Plus une tige est longue, plus la variable qu'elle représente contribue l'explication de la variable dépendante. Les variables qui dépassent la barre horizontale sont variables qui participent de manière significative à la construction de notre modèle. Neuf variables indépendantes sortent significativement à l'explication de la variable dépendante. La première variable la plus importante pour notre modèle de régression PLS est la variable le nombre de salle de bains. Les autres variables dans l'ordre les plus importantes sont : l'étendue des maisons, le nombre de chambres, l'âge des maisons, la distance par rapport l'incinérateur, la distance par rapport à une station de bus, la surface de terrain, la distance par rapport à l'autoroute, la maison proche. Par ailleurs deux variables sortent de manière insignifiante à la construction du modèle. Ces variables sont : le pourcentage de vent incinérateur et le nombre de connaissances.

Le Système SAS

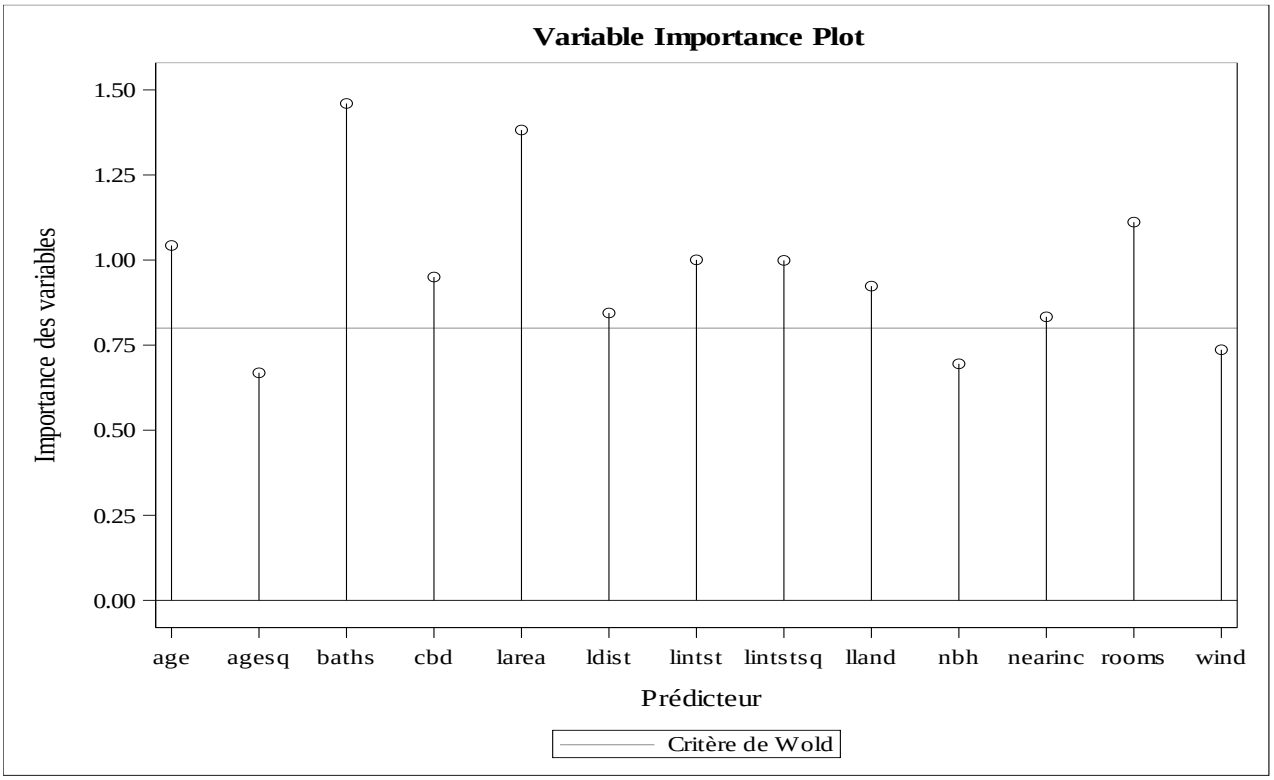
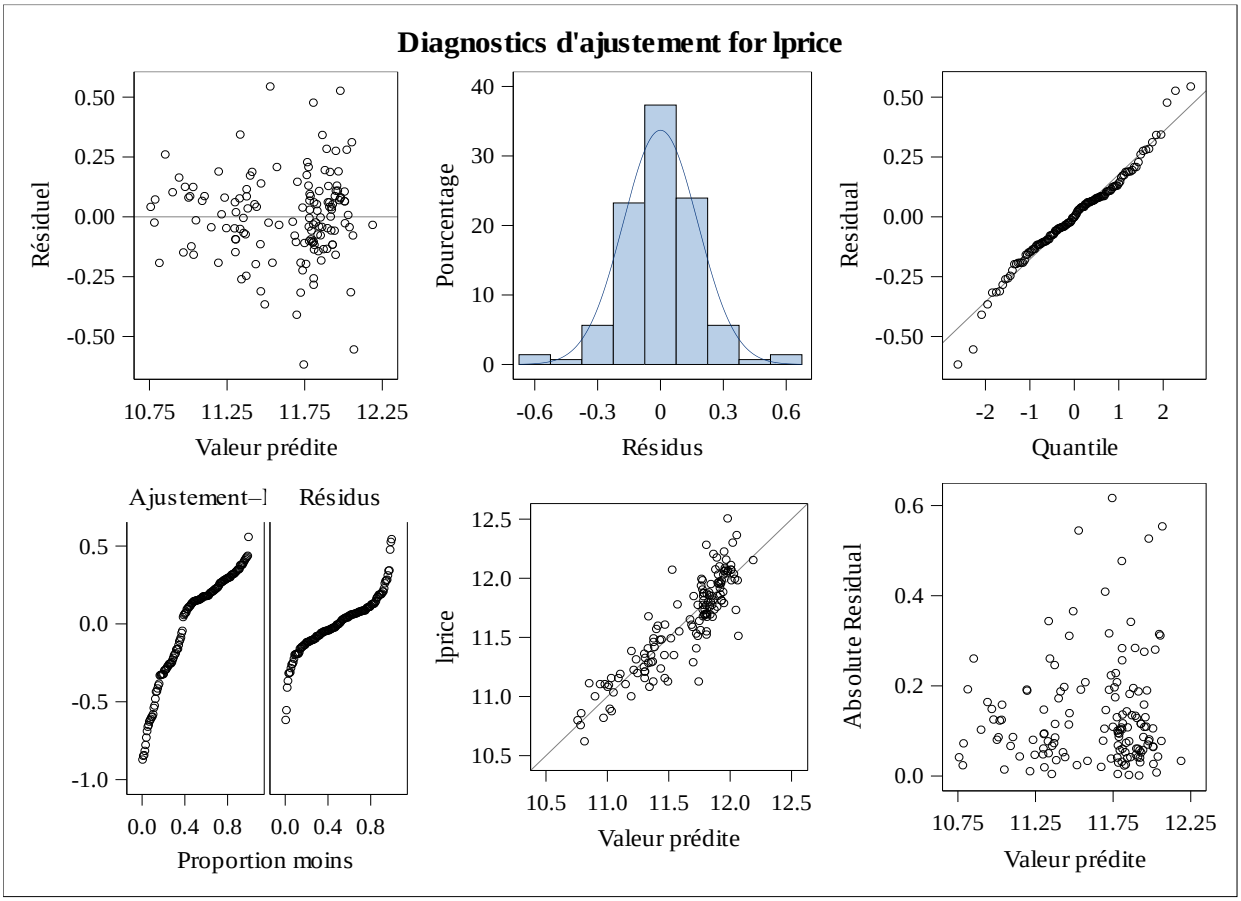


Figure 7



5.1.7 TEST DE VALIDATION CROISEE :

5.1.7.1 Critère de validation par la méthode PRESS

Comme mentionné précédemment, le test de validation permet déterminer le nombre de composantes optimale à retenir pour l'estimation de coefficient de régression en PCR ou en PLS. Contrairement au cas de PCR, le test nous propose que le nombre de composantes qu'il faut retenir est 2 dans le cas de PLS car 2 est le plus petit nombre de composante à partir duquel la p-value(Prob>T**2) est supérieure à 10% [Voir Tableau32].

Le Système SAS

Table	WORK.TRAV1
Méthode d'extraction de facteurs	Moindres carrés partiels
Algorithme PLS	NIPALS
Nombre de variables de réponse	1
Nombre de paramètres du prédicteur	13
Gestion des valeurs manquantes	Exclude
Nombre maximal de facteurs	13
Validation Method	7-fold Split-sample Validation
Validation Testing Criterion	Prob T**2 > 0.1
Number of Random Permutations	1000
Random Permutation Seed	637210001

Tableau 30

Racine carrée du PRESS moyen min.	0.5448
Réduction du nombre de facteurs	2
Le plus petit nombre de facteurs avec p > 0.1	2

Tableau 32

Split-sample Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	**Racine carrée du PRESS moyen**	**T**2**	**Prob > T**2**
0	1.077497	42.99469	<.0001
1	0.689504	12.97502	<.0001
2	0.544803	0	1.0000
3	0.545603	0.006175	0.9440
4	0.553264	0.35106	0.5710
5	0.556153	0.540359	0.4830
6	0.561553	1.115838	0.3140
7	0.563386	1.013728	0.3270
8	0.558806	0.529734	0.5180
9	0.555713	0.33412	0.6000
10	0.556664	0.372041	0.5900
11	0.55724	0.388804	0.5880
12	0.55756	0.40713	0.5780
13	0.558078	0.430633	0.5620

Tableau 31

5.1.8 La part des facteurs dans l'explication des variables

Si nous retenons deux composantes principales comme nous l'a recommandé le test de validation croisée effectué ci-haut, ensemble elles expliquent environ près de 64% la variance de variables explicatives et plus de 77% la variance de prix de vente des maisons. En effet l'axe1 explique à lui seul près de la moitié (49.92%) des informations contenues dans les variables indépendantes et explique à hauteur de 60.64% les informations relatives à la variable dépendante. Le deuxième axe contribue moins à l'explication des variables par rapport à l'axe1 ; l'axe2 contribuent donc à plus de 13.64% à expliquer la variabilité des variables indépendantes et à plus de 16.56% à expliquer la variance de log(prix) [voir suite de Tableau 33].
Par ailleurs, parmi toutes les variables exogènes, celle qui est la plus expliquée par l'axe1 est la distance par rapport à l'autoroute. L'axe1 explique donc à 77.69% la variance de la variable distance par rapport à l'autoroute. La deuxième variable la plus expliquée est le nombre de salles de bain et son inertie est élucidée à 66.22%. Et le nombre de connaissances dans le quartier est la variable la moins expliquée par le premier facteur ; le premier facteur explique seulement 0.26% de la variance de cette variable. Par ailleurs, la variable la plus expliquée par le deuxième facteur est la distance par rapport à l'autoroute comme cela en est pour le premier facteur. Mais par contre, l'axe2 détient une part d'explication plus importante de cette variable par rapport au premier facteur. La première composante explique 93.24% l'information contenue dans la variable distance à l'autoroute. La deuxième variable la plus expliquée est la distance par rapport à la station de bus. Et la variable la moins expliquée par l'axe2 est le nombre de connaissances dans le quartier [voir tableau 33].

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels													
Nombre de facteurs extraits	Effets du modèle												
	age	agesq	baths	cbd	larea	ldist	lintst	lintstsq	lland	nbh	nearinc	rooms	wind
1	52.1577	26.9057	66.2207	64.7259	41.6855	53.7126	77.6879	77.4716	58.6844	0.2641	51.5210	40.8088	37.0790
2	54.6064	26.9928	85.4596	90.2847	65.6848	70.9738	93.2483	94.5621	63.4308	19.5996	58.4597	55.2419	47.7137

Tableau 33

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels					
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes		
	En cours	Total	lprice	En cours	Total
1	49.9173	49.9173	60.6434	60.6434	60.6434
2	13.6410	63.5583	77.2027	16.5593	77.2027

Suite de Tableau 33

Le Système SAS

5.1.9 Estimation de coefficient de régression par PLS

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	49.9173	49.9173	60.6434	60.6434
2	13.6410	63.5583	16.5593	77.2027

Tableau 34

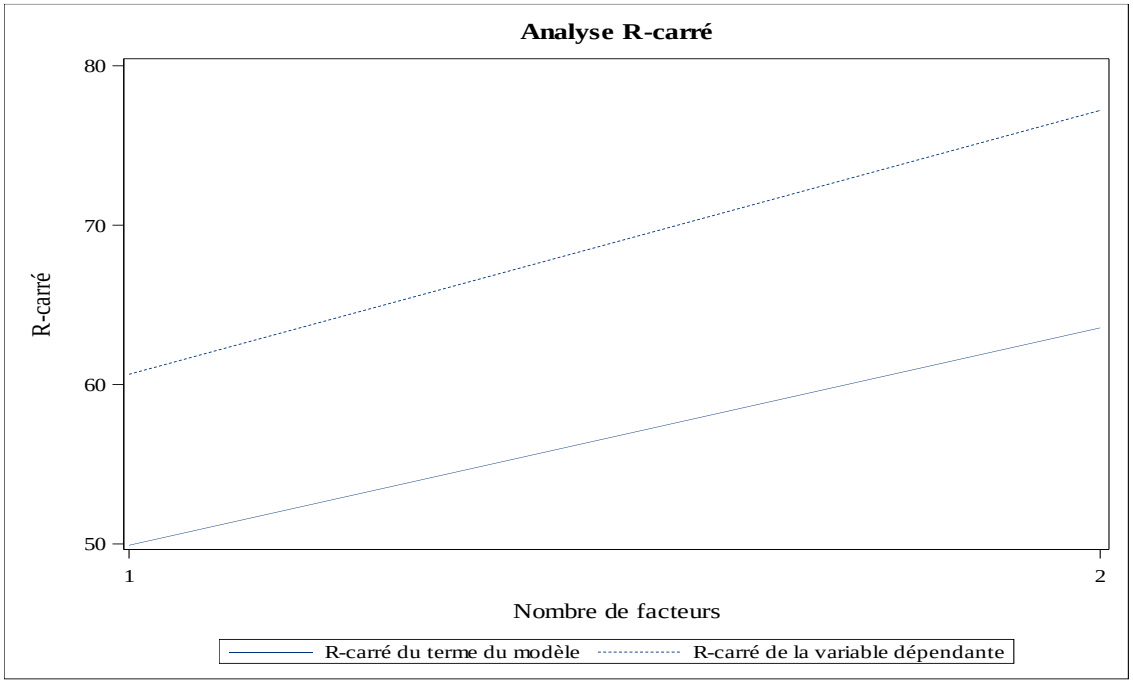


Tableau 35

Cette partie constitue l'une des parties la plus importante de notre étude car elle nous permet de mesurer l'impact des variables indépendantes sur le prix de vente des maisons en tenant compte de la maximisation de la covariance entre les variables dépendantes et les variables indépendantes contrairement à la méthode PCR qui se contente de maximiser la variance des variables indépendantes. Grace aux coefficients de régression PLS du tableau 36, nous pouvons déterminer l'équation de régression suivante.

Log(prix)= -0.1581age -0.0687agesq +0.2773baths -0.0315cbd + 0.2797Larea +0.0060Ldist +0.0226Lintst + 0.0149Lintstsq +0.0761Lland -0.1525nbh -0.0454nearinc +0.2002rooms -0.0263wind.

En effet, on remarque que les coefficients de régression possèdent tous le signe attendu mais en revanche les variables ne contribuent pas de la même manière à l'explication de la variabilité des prix des maisons. Il y a des variables qui impactent fortement et d'autres impactent faiblement les prix de vente des maisons. La variable qui influence plus le prix des maisons aux USA est la surface occupée par la maison (comparativement au cas de PCR qui nous montre que c'est le nombre de salles de douches qui influe plus le prix de vente de maisons) et cette influence est positive. Toute chose égale par ailleurs, pour 1% de pieds carré de plus, le prix de vente des maisons augmente de 0.28%. Le nombre de salles de douches est la deuxième variable qui se révèle la plus influente sur le prix de vente des maisons et elle exerce un impact positif sur le prix. Un américain est prêt à augmenter le prix d'achat de 27.73% d'une maison pour une salle de douche supplémentaire. La troisième variable la plus importante dans l'explication de la variation des prix des maisons est le nombre de chambres. Ceteris paribus, le prix de vente des maisons augmente de 20.02% quand le nombre de salle augmente d'une unité.
Cependant, les variables qui influencent faiblement le prix de vente des immeubles et dans l'ordre croissant selon le degré d'impact sont : la superficie du terrain, la variable binaire maison proche, la distance par rapport à une station de bus, la distance par rapport à l'autoroute, la distance par rapport à l'incinérateur. Toute chose égale par ailleurs, l'augmentation de de la superficie de terrain de 1% entraine une hausse de prix de maison de 0.08%. Contrairement à la méthode PLS, les maisons proches de l'incinérateur coutent moins chères que les maisons éloignées de l'incinérateur aux USA. Une maison proche de l'incinérateur coute 4.54% de moins qu'une maison non proche de l'incinérateur. Toute chose égale par ailleurs, si la distance par rapport l'autoroute passe de 1000 pieds à 1300 pieds, le prix de vente des maisons connait une augmentation de 0.06%. Comparativement à la méthode PCR, la distance par rapport à l'incinérateur impacte positivement le prix des maisons dans le cadre PLS. Pour 1% de distance d'éloignement par rapport à l'incinérateur, les Américains sont prêts à payer 0.006% supplémentaire.

Le Système SAS

Paramètres estimés pour données centrées et mises à l'échelle	
	lprice
Intercept	0.0000000000
age	-.1580602813
agesq	-.0687404924
baths	0.2773131995
cbd	-.0314579634
larea	0.2797062911
ldist	0.0059976765
lintst	0.0226155553
lintstsq	0.0148550853
lland	0.0760971481
nbh	-.1524529111
nearinc	-.0453476295
rooms	0.2002319792
wind	-.0262960675

Tableau 36

Résultats estimés des paramètres	
	lprice
Intercept	7.879175902
age	-0.002574732
agesq	-0.000011728
baths	0.134254495
cbd	-0.000001470
larea	0.305333726
ldist	0.005154409
lintst	0.012282077
lintstsq	0.000443316
lland	0.041179944
nbh	-0.027092084
nearinc	-0.039170033
rooms	0.094467697
wind	-0.004187576

Tableau 37

6 CONCLUSION

Les résultats de notre étude suggèrent que les maisons coutent en moyenne plus de 12.000 dollars aux USA dans les années 1980.

Cependant, nous pouvons repartir nos variables d'intérêts en trois catégories de variables : la catégorie des variables d'emplacement, la catégorie des caractéristiques internes des immeubles et autres caractéristiques. La case des facteurs d'emplacements est constituée de critères : distance par rapport à un incinérateur, la distance par rapport à l'autoroute, la distance par rapport à la station de bus, le pourcentage de vent incinérateur, la maison proche. Ensuite dans la catégorie caractéristique interne des immeubles se trouve l'âge des maisons, l'aire du terrain des maisons, la superficie des maisons, le nombre de salles de bains, le nombre de chambres. Et enfin dans la case des autres caractéristiques, on a la variable nombre de connaissances dans le quartier.

En effet, l'étude économétrique par la régression multiple (MCO) révèle que seulement quatre variables parmi les onze variables d'études impactent significativement le prix de vente des logements aux USA car leur p-value inférieure à 5%. Ces variables sont : l'âge des maisons, le nombre de salles de bains, la superficie des maisons, l'aire du terrain. Elles se trouvent toutes dans la première catégorie. L'âge des maisons affecte faiblement et négativement (-0.00692) le prix des logements. Le nombre de salles de bains a un effet fort et positif (0.11543) sur le prix de vente des maisons ; ce qui veut dire qu'en général, les Américains ont une préférence pour les maisons avec beaucoup de salles de bains et ils sont prêts à débourser 11.54% de plus pour avoir une salle de douche supplémentaire. La taille des maisons impacte fortement et positivement (0.37622) notre variable endogène. Elle est le critère plus influent à la variabilité des prix des immeubles. Autrement dire, une augmentation de l'aire de maisons de 1% entraine une hausse de prix de maisons de 0.38% Et enfin la superficie du terrain impacte (0.08642) positivement le prix de logement. Ces variables font partie toutes de la catégorie des caractéristiques internes. Ce qui laisse comprendre que les Américains ne tiennent pas trop compte de l'emplacement des logements avant de les acheter mais ils accordent plutôt plus d'importance aux critères internes des maisons plus précisément à la taille des maisons, au nombre de salles, à la superficie de terrain, à l'âge des maisons dans leur processus d'achat des immeubles.

Toutefois, grâce au test de vif, il a été mis en exergue que de nombreuses variables sont colinéaires avec d'autres variables. Ce qui expliquait l'instabilité des coefficients de régression par les MCO (Moindre Carré Ordinaire) et qui rendait des variables non significatives au seuil de 5%. Par conséquent, d'autres méthodes économétriques telles que PCR (Principale Composantes Regression), PLS (Partial Least Square) s'invitent pour contourner ce problème.

Tout d'abord, PCR est une méthode de régression basée sur la maximisation des variances des variables exogènes. Elle est très utilisée dans le domaine de la médecine. Ainsi l'application de cette méthode permet d'enrichir notre étude et de résoudre certains problèmes cités précédemment. Comparativement à la méthode des Régressions Multiples, la méthode PCR

nous laisse voir que toutes les variables contribuent de manière significative à l'explication des prix de vente des immeubles sauf la variable nombre de connaissances. Ce qui veut dire que les Américains tiennent compte non seulement des critères de l'emplacement mais aussi des caractéristiques internes des maisons dans le processus de définition des prix de vente de leur maison. On peut donc classer les variables selon le sens de relation qui les animent avec la variable expliquée. D'un côté, on a les variables qui influent positivement les prix de logement et de l'autre côté les variables qui influent négativement les prix de logement. Le premier groupe de variable est constitué dans l'ordre les plus impactant à la variabilité des prix de logement de : le nombre de salles de bains (0.2531), la surface de la maison (0.2501), le nombre de chambres, l'aire du terrain (0.0625), le pourcentage de vent incinérateur (0.0531), la distance par rapport à l'autoroute, la maison proche. Ainsi l'augmentation de ces variables vont entrainer l'augmentation des prix de logement. En gros, les Américains, généralement, aiment plus les maisons avec plusieurs salles de bains et plusieurs chambres, maisons larges sur de grandes superficies, maisons éloignées des autoroutes et des fumées des usines. C'est pourquoi ils sont prêts à payer plus pour les avoir. Par contre le second groupe regorge les variables : âge de maisons, la distance par rapport à l'incinérateur, distance par rapport à une station de bus (-0.0196). Comparativement au premier groupe, une variation positive de ces variables fait baisser les prix de vente des maisons. Ce qui implique qu'en générale les acheteurs américains ne valorisent pas trop les maisons vielles, maisons trop éloignées des stations de bus et des incinérateurs. C'est pourquoi ils dépensent moins pour acheter ces types de maisons.

Certes la méthode PCR nous a permis de résoudre des problèmes d'instabilité de coefficients de régression auxquels est confronté la régression multiple mais elle demeure moins robuste car elle se contente de maximiser la variance des variables explicatives.

C'est dans ce contexte que nous avons fait recours à une troisième méthode qui s'avère plus robuste car elle permet non seulement de résoudre les problèmes d'instabilité mais elle cherche également à la covariance entre les variables indépendantes et les variables dépendantes : PLS.

Par ailleurs, l'étude par PLS nous laisse croire que, d'après le critère de Wold, toutes les toutes variables sauf le nombre de connaissances dans le quartier et le pourcentage de vent incinérateur participent de manière significative à l'explication de prix de logements ; contrairement à la méthode PCR qui insinue que c'est seulement le nombre de connaissances dans le quartier est non significatif. Les coefficients de régression de ces variables ont tous le signe attendu. Ainsi Ces variables peuvent être classée selon leur relation avec les prix de de vente de logement. D'une part, on a les variables qui exerce un impact positif sur les prix de vente des maisons et de l'autre part celles qui influent négativement les prix. Dans la première famille de variable, nous avons les mêmes variables que dans le cas de PCR sauf qu'ici la variable distance par rapport à l'incinérateur qui possède un coefficient négatif et se retrouve dans la deuxième catégorie. Et une augmentation de ces variables entraine une hausse du prix des immeubles. En effet, parmi cette catégorie, la superficie des terrains est la première

variable la plus influente sur les prix (car son coefficient de régression est le plus important) comparativement à PCR qui suggère plutôt le nombre de salles de bains. Ce qui veut qu'en moyenne ; les acheteurs américains donnent une place très considérable au facteur superficie des maisons dans leur choix d'achat des immeubles et ils sont prêts à augmenter le prix de 27.80% pour 1% de terrain supplémentaire. La deuxième variable la plus influençant est le nombre de salles de bains. Bref les Américains sont aptes à dépenser plus pour s'acquérir des maisons larges sur des terrains vastes, des maisons éloignées des autoroutes et des incinérateurs, des maisons avec beaucoup de chambres et de salles de bains. Par contre, le second groupe est composé de l'âge des maisons, la distance par rapport à une station de bus, la distance par rapport à l'incinérateur, maison proche. Et le prix des maisons diminue suite à la variation positive de ces variables. Ce qui signifie que les Américains ont moins de préférence pour les maisons trop vielles, des maisons éloignées des stations de bus et des maisons proches des incinérateurs.

NB : le variables nombre de connaissances dans le quartier dans le cadre de PCR et les variables nombre de connaissances dans le quartier et le pourcentage de vent incinérateur dans le cadre de PLS ne font pas objet de notre conclusion car elles ne contribuent pas de manière significative à l'explication de la variation des prix des maisons selon le critère de Wold. Par ailleurs, les résultats de la méthode PLS sont les plus fiables pour notre étude car la méthode PLS est la plus robuste et la plus adaptée à notre base de données.

Néanmoins les résultats de notre étude peuvent ne plus être vrai aujourd'hui et peuvent susciter des critiques car les données datent depuis 1981 et que les envies peuvent changer selon les réalités de temps. Par exemple une famille avec beaucoup d'enfants va chercher à acheter une maison avec plusieurs chambres. De plus, de nos jours on peut penser à d'autres variables qui peuvent contribuer significativement à expliquer la variation des prix de vente des immeubles telles que la présence de piscine, de l'ascenseur, le système électrique, distance par rapport à des écoles, à des parcs, à des jardins etc.

7 BIBLIOGRATPHIE

Brigitte Escofier, Jérôme Pagés. *Analyse Factorielle Simples et Multiples*. 5e Edition. 2016

James Stock, Mark Watson. *Principes d'Econométrie*. 3e Edition.2012

Michel VOLLE. *Analyse des données*. 4e Edition. 1997

Pierre-André Cornillon et al. *Statistiques avec R.* 3e Edition. 2012

8 ANNEXE

- **année:** 1981
- **age:** âge de la maison
- **agesq:** âge ^ 2
- **nbh:** quartier,=neighborhood identication =nombre de connaissance du quartier

- **cbd:** distance to central busing distric=distance par rapport à une station de bus, feet =pied (1 feet= 0.3048 mètre).
- **intst:** distance à l'autoroute, feet(pieds)
- **lintst:** log (intst)
- **price:** prix de vente en dollar
- **rooms:** nombre de chambres dans la maison
- **area:** superficie de la maison, square footage=pieds carré (1 pieds carré=0.092903 mètre carré)
- **land:** superficie du terrain, square footage
- **baths:** nombre de salles de bain
- **dist:** distance de la maison à l'incinérateur, feet
- **ldist:** log (dist)
- **wind:** pourcentage temps vent incinérateur à la maison
- **lprice:** log (price)
- **y81:** = 1 si année == 1981
- **larea:** log (area)
- **lland:** log (land)
- **y81ldist:** y81 * ldist
- **lintstsq:** lintst ^ 2
- **nearinc:** = 1 si dist <= 15840
- **y81nrinc:** y81 * nearinc
- reprice**:** prix, 1978 dollars
- **lrprice:** log (rprice)

  NB : un incinérateur, dans les années 1900, est un dispositif fonctionnant à base de vent, sert à détruire par combustion les ordures ménagères.