



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**



**UNIVERSITÉ
PARIS-EST CRÉTEIL
VAL DE MARNE**

MASTER :

METHODES APPLIQUEES DE LA STATISTIQUE ET DE L'ECONOMETRIE POUR LA RECHERCHE, L'ANALYSE ET LE TRAITEMENT DE L'INFORMATION

PARCOURS :

DATA SCIENCE/ DATA ANALYST

MODULE :

RAPPORT D'ETUDE

ANALYSE DES FACTEURS INFLUENCANT LE PRIX DE VENTE D'UNE MAISON

Etude de cas du marché immobilier de Windsor et Essex au canada

AUTEUR :

LAJOIE BENGONE AKOU
bengonelajoie@gmail.com

29 Novembre 2025

Tables des matières

1	DESCRIPTION DE LA BASE DE DONNEES :	3
2	STATISTIQUES DESCRIPTIVES :	3
2.1	TABEAU DE STATISTIQUES DESCRIPTIVES	3
2.2	DISTRIBUTION DE LA VARIABLE CIBLE	4
2.3	ANALYSE DE LA MATRICE DE CORRELATION	4
3	ESTIMATION DES MODELES DE REGRESSION :	5
3.1	MODELE LINEAIRE SIMPLE	5
3.2	MODELE LINEAIRE MULTIPLE	6
3.3	MODELE DE REGRESSION LINEAIRE MULTIPLE SANS CONSTANTE	7
3.4	MODELE DE REGRESSION LINEAIRE MULTIPLE AVEC SUPPRESSION DES VALEURS EXTREMES	7
4	VERIFICATION DES HYPOTHESES DE NORMALITE :	9
5	MODELE DE SELECTION ET TRANSFORMATION LOGARITHMIQUE :	10
6	CONCLUSION :	12
7	REFERENCE :	12

Tableau

Tableau 1 :	Description des variables	3
Tableau 2 :	Statistiques descriptives	3
Tableau 3 :	Matrice de corrélation	4
Tableau 4 :	Modèle (1) de régression linéaire simple	5
Tableau 5 :	Modèle (2) de régression linéaire multiple	6
Tableau 6 :	Modèle (3) de régression linéaire (2) sans constante	7
Tableau 7 :	Modèle (3) de régression avec la gestion des valeurs extrêmes	8
Tableau 8 :	Tableau récapitulatif des modèles de régression	8
Tableau 9 :	Test d'homoscédasticité de Breusch-Pagan	9
Tableau 10 :	Test de corrélation : vérification de l'exogénéité des variables	10
Tableau 11 :	Synthèse des modèles de sélection	11
Tableau 12 :	Synthèse des résultats des modèles de sélection	11

Graphique

Graphique 1 :	Histogramme du prix de vente d'une maison	4
Graphique 2 :	Nuage de points entre le prix et la surface d'une maison	5
Graphique 3 :	Visualisation 3D du modèle (2) de régression	6
Graphique 4 :	Distribution des résidus	9

1 DESCRIPTION DE LA BASE DE DONNEES :

La base de données provient des travaux de Anglin and Gencay (1996). Elle contient 546 observations et 12 variables qui représentent le prix de vente d'une maison ainsi que ses principales caractéristiques telles que : la surface du logement, le nombre de chambres etc.

Une description de ces variables est présentée dans le tableau suivant :

Nom d'origine	Description	Nature (unité)
price	sale price of a house	Continue (\$)
lotsize	lot size of a property	Continue (feet ²)
Bedrooms	number of bedrooms	Entier (#)
bathrms	number of full bathrooms	Entier (#)
stories	number of stories excluding basement	Entier (#)
driveways	dummy, 1 if the house has a driveway	Binaire (0,1)
recroom	dummy, 1 if the house has a recreational room	Binaire (0,1)
fullbase	dummy, 1 if the house has a full finished basement	Binaire (0,1)
gashw	dummy, 1 if the house uses gas for hot water heating	Binaire (0,1)
airco	dummy, 1 if there is central air conditioning	Binaire (0,1)
garagepl	number of garage places	Entier (#)
prefarea	dummy, 1 if located in the preferred neighbourhood of the city	Binaire (0,1)

Tableau 1 : Description des variables

2 STATISTIQUES DESCRIPTIVES :

2.1 Tableau de Statistiques descriptives

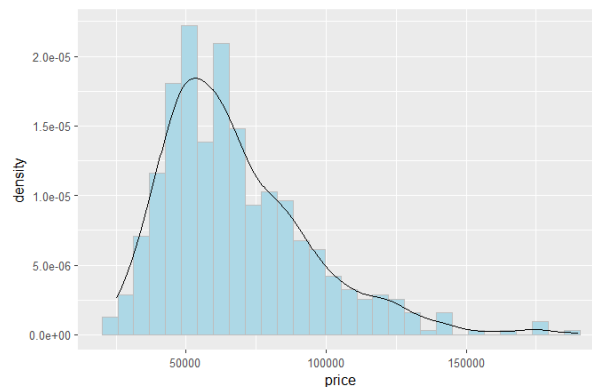
L'analyse descriptive qui comprend le calcul des paramètres de tendance centrale et de dispersion montre que la base de données ne contient aucune valeur manquante. Par ailleurs, elle permet de constater que la variable cible **price** est fortement dispersée au regard de son écart-type et de son intervalle interquartile.

Statistic	N	Mean	St. Dev.	Min	Max
price	546	68,121.600	26,702.670	25,000	190,000
lotsize	546	5,150.266	2,168.159	1,650	16,200
bedrooms	546	2.965	0.737	1	6
bathrms	546	1.286	0.502	1	4
stories	546	1.808	0.868	1	4
driveway	546	0.859	0.348	0	1
recroom	546	0.178	0.383	0	1
fullbase	546	0.350	0.477	0	1
gashw	546	0.046	0.209	0	1
airco	546	0.317	0.466	0	1
garagepl	546	0.692	0.861	0	3
prefarea	546	0.234	0.424	0	1

Tableau 2 : Statistiques descriptives

2.2 Distribution de la variable cible

Pour confirmer cette observation, un histogramme a été réalisé. L'analyse de ce graphique révèle une distribution asymétrique vers la droite et une courbe mésokurtique. On peut en déduire la présence des valeurs extrêmes.



Graphique 1 : Histogramme du prix de vente d'une maison

2.3 Analyse de la matrice de corrélation

La matrice de corrélation permet de mettre en évidence les relations linéaires entre les variables présentes dans la base de données. La force de ces relations affecte directement l'ajustement des paramètres du modèle. A titre d'exemple, la présence de variables fortement corrélées parmi les variables explicatives peut invalider l'hypothèse ($H3$) qui suppose que la matrice des variables explicatives ($X^T X$) soit de plein rang.

Dans cette étude, les résultats montrent que la surface de la maison **lotsize** est la variable la plus corrélée avec **price**, la variable cible. L'intensité de cette relation est égale à 0.536. On retrouve ensuite la variable associée au nombre de douches **bathrms** dont le coefficient de corrélation avec le prix immobilier est de 0.517.

	price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
price	1	0.536	0.366	0.517	0.421	0.297	0.255	0.186	0.093	0.453	0.383	0.329
lotsize	0.536	1	0.152	0.194	0.084	0.289	0.140	0.047	-0.009	0.222	0.353	0.235
bedrooms	0.366	0.152	1	0.374	0.408	-0.012	0.080	0.097	0.046	0.160	0.139	0.079
bathrms	0.517	0.194	0.374	1	0.324	0.042	0.127	0.103	0.067	0.185	0.178	0.064
stories	0.421	0.084	0.408	0.324	1	0.122	0.042	-0.174	0.018	0.296	0.043	0.043
driveway	0.297	0.289	-0.012	0.042	0.122	1	0.092	0.043	-0.012	0.106	0.204	0.199
recroom	0.255	0.140	0.080	0.127	0.042	0.092	1	0.372	-0.010	0.137	0.038	0.161
fullbase	0.186	0.047	0.097	0.103	-0.174	0.043	0.372	1	0.005	0.045	0.053	0.229
gashw	0.093	-0.009	0.046	0.067	0.018	-0.012	-0.010	0.005	1	-0.130	0.068	-0.059
airco	0.453	0.222	0.160	0.185	0.296	0.106	0.137	0.045	-0.130	1	0.157	0.116
garagepl	0.383	0.353	0.139	0.178	0.043	0.204	0.038	0.053	0.068	0.157	1	0.092
prefarea	0.329	0.235	0.079	0.064	0.043	0.199	0.161	0.229	-0.059	0.116	0.092	1

Tableau 3 : Matrice de corrélation

3 ESTIMATION DES MODELES DE REGRESSION :

3.1 Modèle linéaire simple

La première modélisation du prix de vente d'une maison est réalisée en utilisant un modèle de régression linéaire simple. La variable explicative choisie pour cela est **lotsize** qui présentait la plus forte corrélation avec **price**. L'équation du modèle s'écrit comme suit :

$$price_i = \beta_0 + \beta_1 \times lotsize_i$$

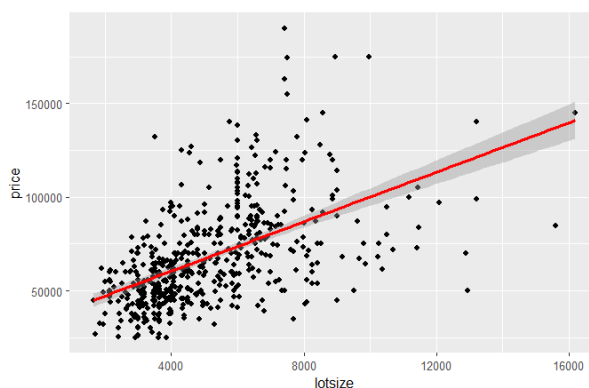
Les résultats de (1) se présentent comme suit :

Dependent variable:	
price	
lotsize	6.599*** (0.446)
Constant	34,136.190*** (2,491.064)
Observations	546
R ²	0.287
Adjusted R ²	0.286
Residual Std. Error	22,567.050 (df = 544)
F Statistic	219.056*** (df = 1; 544)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tableau 4 : Modèle (1) de régression linéaire simple

La variable **lotsize** est très significative. Par conséquent, les résultats de ce modèle peuvent être expliqués selon le paradigme TCEPA. Autrement dit, lorsque la surface d'une propriété augmente d'une unité (exprimée en feet²), son prix de vente augmente de 6.499 \$. De plus, on observe que la constante du modèle est très significative.

Cependant, le R² de ce modèle est égale à 0.287. Ce qui suggère un très mauvais ajustement du modèle que l'on peut analyser autrement grâce à un nuage de points entre la variable à expliquer et la variable explicative.



Graphique 2 : Nuage de points entre le prix et la surface d'une maison

Le graphique ci-dessus laisse penser que la relation entre le prix de vente et la surface d'une maison est quadratique au lieu de linéaire. Cela s'observe par la dispersion de la variance.

3.2 Modèle linéaire multiple

Toutefois, on va tenter d'améliorer ce premier modèle en nous appuyant sur l'idée des variables omises. Cette dernière suggère que, excepté la surface d'une maison, il existe d'autres variables présentes dans les résidus qui pourraient influencer les prix immobiliers. La méthode consiste donc à les extraire des résidus pour les inclure dans le modèle de régression.

C'est dans cette optique, que la variable **bathrms** a été employée pour élaborer un nouveau modèle de régression. Celle-ci a été choisie car elle est la variable la plus corrélée avec le prix de vente après la variable **lotsize**.

L'équation de ce modèle linéaire multiple s'écrit comme suit :

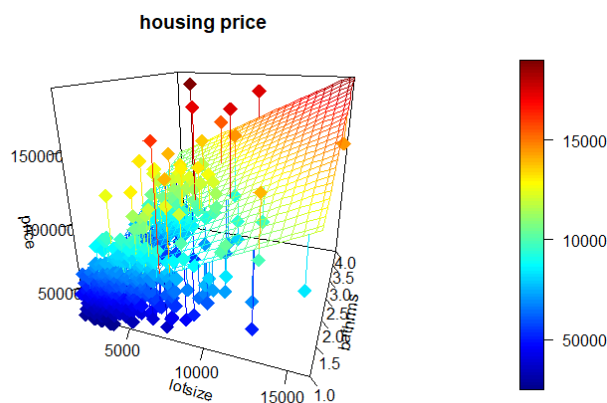
$$price_i = \beta_0 + \beta_1 \times lotsize_i + \beta_2 \times bedrooms_i + \beta_3 \times bathrms_i$$

Ce modèle (2) a obtenu les résultats suivant :

Dependent variable:	
price	
lotsize	5.575*** (0.394)
bathrms	22,811.460*** (1,702.693)
Constant	10,081.420*** (2,810.016)
Observations	546
R ²	0.464
Adjusted R ²	0.462
Residual Std. Error	19,582.100 (df = 543)
F Statistic	235.208*** (df = 2; 543)
Note: *p<0.1; **p<0.05; ***p<0.01	

Tableau 5 : Modèle (2) de régression linéaire multiple

Le R² est maintenant de 0.464 et β_1 est passé de 6.60 à 5.58. Ce qui signifie que l'effet de la surface du logement sur le prix de vente de la maison était surestimé dans le modèle (1).



Graphique 3 : Visualisation 3D du modèle (2) de régression

Cette visualisation confirme les observations réalisées à partir du Graphique 2. Il s'agit de l'existence de valeurs extrêmes et de la dispersion quadratique dans un espace à 3 dimensions. En effet, on peut aisément constater que les observations sont fortement concentrées à l'origine avant de se disperser brusquement à mesure que les valeurs des variables augmentent.

Cela peut indiquer que la relation entre la variable que l'on cherche à modéliser et les variables indépendantes est non linéaire. Cette hypothèse devrait être vérifiée en réalisant des analyses plus approfondies.

Avant cela, un test de significativité a été réalisé pour savoir si les variables **lotsize** et **bathrms** influencent significativement le prix de vente d'une maison. Pour se faire, on va tester si les coefficients de régression multiple sont nuls (H_0) avec une erreur de première espèce $\alpha = 5\%$.

Le test de significativité de Fisher donne une p-valeur de $2.2e-16$. Conclusion l'hypothèse nulle (H_0) peut être rejetée avec un seuil d'erreur de 5%. Au moins l'une des deux variables influence significativement le prix de vente.

3.3 Modèle de régression linéaire multiple sans constante

Le modèle devient :

$$price_i = \beta_1 \times lotsize_i + \beta_2 \times bedrooms_i + \beta_3 \times bathrms_i$$

Les résultats obtenus de ce nouveau modèle sont résumés dans le tableau suivant :

<i>Dependent variable:</i>	
	price
lotsize	6.384*** (0.327)
bathrms	26,714.700*** (1,323.990)
Observations	546
R ²	0.927
Adjusted R ²	0.927
Residual Std. Error	19,794.610 (df = 544)
F Statistic	3,457.130*** (df = 2; 544)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tableau 6 : Modèle (3) de régression linéaire (2) sans constante

Sans constante, l'ajustement du modèle de régression linéaire multiple passe de 0.464 à 0.927. On parvient à ce résultat parce que la suppression de la constante force le modèle à adopter davantage un ajustement linéaire au détriment de la fiabilité des résultats.

3.4 Modèle de régression linéaire multiple avec suppression des valeurs extrêmes

Pour préserver la qualité d'ajustement du modèle, il est préférable de conserver la constante dans l'équation de régression. Désormais, il sera question d'améliorer la qualité du modèle en explorant d'autres méthodes à l'instar de la méthode de gestion des valeurs extrêmes, qui ont été identifiées dans le Graphique 1.

Cette méthode consiste à supprimer toutes les observations qui sont supérieures (respectivement inférieures) au 1^{er} (respectivement 9^{ème}) décile. Une fois appliquées, on obtient une base de données qui contient 441 observations sur laquelle on va estimer de nouveau le modèle (2).

<i>Dependent variable:</i>	
price	
lotsize	3.478*** (0.306)
bathrms	13,614.450*** (1,431.657)
Constant	30,569.150*** (2,378.942)
Observations	441
R ²	0.349
Adjusted R ²	0.346
Residual Std. Error	13,441.740 (df = 438)
F Statistic	117.498*** (df = 2; 438)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Tableau 7 : Modèle (3) de régression avec la gestion des valeurs extrêmes

Le modèle (4) a obtenu des performances en dessous de ceux du modèle (2). L'explication probable associée à ce résultat est la suppression des individus importants dans l'ajustement du modèle de régression. Pour confirmer cette explication, l'analyse du poids des individus pourraient être effectuée. Celle-ci comprendrait l'identification des individus les plus << importants >> et l'impact de leur suppression sur la qualité du modèle de régression.

En résumé, 4 modèles de régression linéaire ont été réalisés pour modéliser le prix de vente. Les résultats de ces modèles ont été regroupés dans le tableau suivant :

<i>Dependent variable:</i>				
price				
	(1)	(2)	(3)	(4)
lotsize	6.599*** (0.446)	5.575*** (0.394)	6.384*** (0.327)	3.478*** (0.306)
bathrms		22,811.460*** (1,702.693)	26,714.700*** (1,323.990)	13,614.450*** (1,431.657)
Constant	34,136.190*** (2,491.064)	10,081.420*** (2,810.016)		30,569.150*** (2,378.942)
Observations	546	546	546	441
R ²	0.287	0.464	0.927	0.349
Adjusted R ²	0.286	0.462	0.927	0.346
Residual Std. Error	22,567.050 (df = 544)	19,582.100 (df = 543)	19,794.610 (df = 544)	13,441.740 (df = 438)
F Statistic	219.056*** (df = 1; 544)	235.208*** (df = 2; 543)	3,457.130*** (df = 2; 544)	117.498*** (df = 2; 438)
<i>Note:</i>				*p<0.1; **p<0.05; ***p<0.01

Tableau 8 : Tableau récapitulatif des modèles de régression

Au regard de ce tableau, on serait tenté de conclure que le modèle de régression multiple (3) est le modèle de régression le plus robuste puisqu'il a obtenu R^2 ajusté de 0.927 qui est largement supérieur aux autres modèles. Cependant, cette conclusion serait inexacte puisque ce dernier a été développé en supprimant la constante de l'équation.

La suppression de cette constante a pour effet de forcer une relation linéaire entre les variables explicatives et la variable à expliquer. En plus de cela, si la constante est significative, ce qui semble être le cas au regard des résultats des modèles (1) et (3), les estimations des paramètres de régression seraient biaisées dans le pire des cas.

Par conséquent, le modèle le plus fiable est le modèle de régression (2) qui présente le coefficient de détermination le plus élevé (sans suppression de la constante) et la méthodologie la plus exacte.

Ainsi, la suite de cette étude consistera à explorer de façon plus exhaustive ce modèle. Pour cela, il s'agira de vérifier dans un premier temps les hypothèses de régression linéaire que sont : l'espérance des résidus nulle, l'hétéroscédasticité, et l'exogénéité.

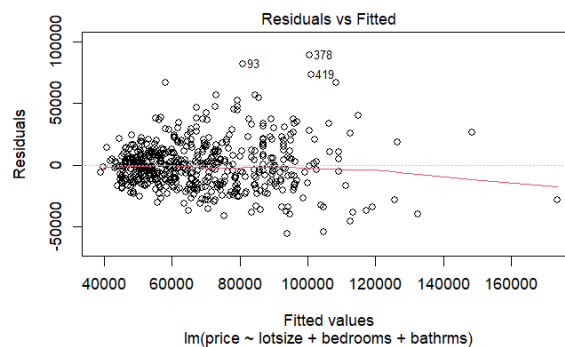
4 VERIFICATION DES HYPOTHESES DE NORMALITE :

a. *Espérance des résidus nulles* : $E(u_i) = 0$

l'espérance des résidus est nulle : $1.469997e - 12$

b. *Hétéroscédasticité*

Pour vérifier l'hétéroscédasticité, on va procéder à une visualisation de la distribution des résidus et réaliser un test paramétrique.



Graphique 4 : Distribution des résidus

La distribution des résidus montre la présence d'une structure polynomiale. Cela laisse induire que les résidus sont hétéroscédastiques. Pour confirmer ce constat, le test asymptotique d'homoscédasticité de Breusch-Pagan a été réalisé.

<i>studentized Breusch-Pagan test</i>	
Test statistic	38.975
DF	2
P value	3.44e-09

Tableau 9 : Test d'homoscédasticité de Breusch-Pagan

Comme les p-values de ces deux tests sont inférieures à 5%, par conséquent, on peut rejeter H_0 qui suggère que les résidus sont homoscedastiques. Cela signifie que l'estimateur des Moindres Carrés $\hat{\beta}$ n'est pas l'estimateur le plus précis parmi les estimateurs linéaires sans biais.

Pour résoudre ce problème, il est préférable d'employer la méthode des moindres carrés quasi-généralisée aussi appelée méthodes des doubles moindres carrés.

c. *Exogénéité des variables* : $E(u_i/x_i) = 0$

La vérification de l'hypothèse d'exogénéité des variables est une étape essentielle dans l'évaluation d'un modèle de régression. Elle contrôle l'existence d'une relation entre les variables explicatives et les résidus du modèle. Si cette liaison est avérée, les estimateurs deviennent biaisés et l'inférence des résultats se trouve compromise voire impossible.

C'est pour vérifier cette hypothèse que plusieurs méthodes ont été élaborées. Parmi elles, on peut retrouver les tests statistiques tels que le test de Hausman qui est particulièrement robuste.

Toutefois, dans cette étude, le choix s'est porté sur le test de corrélation en raison de la facilité de son interprétation et de son implémentation. Les résultats de ce dernier sont résumés dans le tableau suivant :

<i>Correlation test : price with independ variables</i>		
	lotsize	Bathrms
Test statistic	1.0849e-15	-3.8837e-15
DF	544	544
P value	1	1

Tableau 10 : Test de corrélation : vérification de l'exogénéité des variables

Les deux tests ont tous deux obtenu des p-values égales à 1. Par conséquent, on ne peut pas rejeter l'hypothèse d'un coefficient de corrélation nulle entre les résidus du modèle et les variables **lotsize** et **bathrms**. En d'autres termes, les variables ne sont pas exogènes.

5 MODELE DE SELECTION ET TRANSFORMATION LOGARITHMIQUE :

Ces analyses ont permis de mettre en évidence que la modélisation linéaire reste très limitée en termes d'explicabilité (R^2 relativement faible) et de validité (résidus hétéroscédastiques).

C'est pour améliorer les estimations des paramètres et la validité du modèle que les modèles dits de sélection ont été explorés. Les modèles de sélection sont des modèles obtenus à partir des transformations logarithmiques des variables indépendantes et/ou de la variable dépendante. Ces modèles sont résumés dans le tableau ci-dessous.

Modèles de Sélection – Log Transformation				
Modèles	Equation	Estimateur	Transformation	Interprétation
(1) Lin-lin	$Y_i = \beta X_i + u_i$	$\frac{\Delta Y}{\Delta X}$	Effet marginal (unité)	Lorsque X varie d'une unité, Y varie de β unités
(2) Log-log	$\log(Y_i) = \beta \log(X_i) + u_i$	$\frac{\Delta \log(Y)}{\Delta \log(X)} = \frac{100 \Delta \log(Y)}{100 \Delta \log(X)}$	Elasticité (%)	Lorsque X varie d'un pourcent, Y varie de β pourcent.
(3) Lin-log	$Y_i = \beta \log(X_i) + u_i$	$\frac{\Delta Y}{\Delta \log(X)} = \frac{100 \Delta Y}{100 \Delta \log(X)}$	Semi-élasticité (unité)	Lorsque X varie d'un pourcent, Y varie de $(\beta / 100)$
(4) Log-lin	$\log(Y_i) = \beta X_i + u_i$	$\frac{\Delta \log(Y)}{\Delta X}$	Semi-élasticité (pourcentage)	Lorsque X varie d'une unité, Y varie de $(\beta \times 100\%)$

Tableau 11 : Synthèse des modèles de sélection

Ces transformations permettent d'obtenir un ajustement non linéaire des modèles de régression. L'explication à l'origine de ce phénomène est que la transformation logarithmique minimise le poids des valeurs les plus grandes et amplifie celui des plus petites (fichier/log_transformation_visualisation.xls).

Cela a pour l'effet de réduire l'asymétrie de la distribution et de linéariser des données de façon artificielle. Car les nouveaux moments d'ordre 1 et 2 obtenus sont plus proches de 0 et de 1 qu'auparavant qui représentent respectivement la moyenne et la variance d'une distribution centrée-réduire.

	Dependent variable:			
	price (1)	log(price) (2)	price (3)	log(price) (4)
lotsize	5.575*** (0.394)			0.0001*** (0.00001)
bathrms	22,811.460*** (1,702.693)			0.292*** (0.024)
log(lotsize)		0.469*** (0.030)	31,994.640*** (2,114.477)	
log(bathrms)		0.443*** (0.036)	34,280.880*** (2,582.158)	
Constant	10,081.420*** (2,810.016)	7.004*** (0.249)	-209,330.100*** (17,828.310)	10.271*** (0.040)
Observations	546	546	546	546
R ²	0.464	0.481	0.483	0.444
Adjusted R ²	0.462	0.479	0.481	0.442
Residual Std. Error (df = 543)	19,582.100	0.269	19,238.580	0.278
F Statistic (df = 2; 543)	235.208***	251.510***	253.465***	217.089***

Note:

*p<0.1; **p<0.05; ***p<0.01

Tableau 12 : Synthèse des résultats des modèles de sélection

Au regard de ces résultats, on peut conclure que la transformation lin-log est la modélisation la plus appropriée. En effet, son R² qui vaut .493 est supérieur aux coefficients de détermination des modèles (1), (2) et (4) qui sont respectivement de 0.464, 0.481 et 0.444.

En plus de cela, les résultats du test de Breusch-Pagan indiquent que les résidus du modèle de sélection retenu dans cette étude ne sont pas hétéroscédastiques. Par conséquent, l'hypothèse d'homoscédasticité qui n'était pas respectée dans le modèle (2), est ici vérifiée.

6 CONCLUSION :

En définitive, cette étude montre le modèle de régression caractérisé par la transformation lin-log est le modèle avec le meilleur ajustement des données pour analyser la variation des prix immobiliers. Car, ce dernier a obtenu le coefficient de détermination (R^2) le plus élevé qui vaut 0.483.

En plus de cela, il s'avère être le plus fiable car il respecte toutes les hypothèses de régression linéaire notamment une variance des résidus constante. Ce qui n'était pas le cas des premiers modèles de régression linéaire.

Ainsi, on peut conclure que lorsque les variables **lotsize** et **bathrms** augmentent d'un pourcent, la variable **price** augmente respectivement de 0.32 (32/100) et 0.342 (34,2/100).

7 REFERENCE :

- Anglin, P. M., & Gencay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of applied econometrics*, 11(6), 633-648.
- Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, London*, 22(1), 23-36.
- Hlavac, M. (2018). Stargazer: Well-formatted regression and summary statistics tables. *R package version*, 5(2), 2.