Tableau

Graphique

EXERCICE : RAPPELS D'ECONOMETRIE

Sujet : Analyse des facteurs influençant le prix de vente d'une maison : Etude de cas du marché immobilier de Windsor et Essex au Canada.

1. Description de la base de données

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 des variables (EN)	Description	Unité
price	sale price of a house	Dollar ($)
lotsize	lot size of a property	Feet² (m²)
Bedrooms	number of bedrooms	Integer
bathrms	number of full bathrooms	Integer
stories	number of stories excluding basement	Integer
driveways	dummy, 1 if the house has a driveway	Binary
recroom	dummy, 1 if the house has a recreational room	Binary
fullbase	dummy, 1 if the house has a full finished basement	Binary
gashw	dummy, 1 if the house uses gas for hot water heating	Binary
airco	dummy, 1 if there is central air conditioning	Binary
garagepl	number of garage places	Integer
prefarea	dummy, 1 if located in the preferred neighbourhood of the city	Binary

Tableau 1 : Description des variables

2. Statistiques descriptives

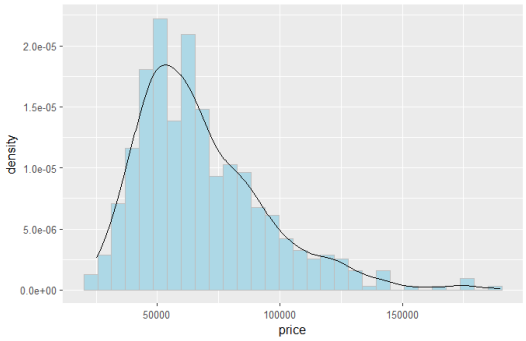
L'analyse descriptives qui comprend le calcul des paramètres de tendance centrale et de dispersion montre que la base de données ne contient aucune valeur manquante. Et que la variable cible à savoir ***price*** est fortement dispersée au regard de son écart-type et de son intervalle interquartile.

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

Pour vérifier ce constat autrement, un histogramme a été réalisé. L'analyse de ce graphique révèle une distribution asymétrique vers la droite et une courbe mésokurtique. Cela peut indiquer la présence des valeurs extrêmes.



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

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 $(X^T X)$ soit de plein rang.

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

Graphique 2 : Matrice de corrélation

3. Estimation des modèles de régression

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 :

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

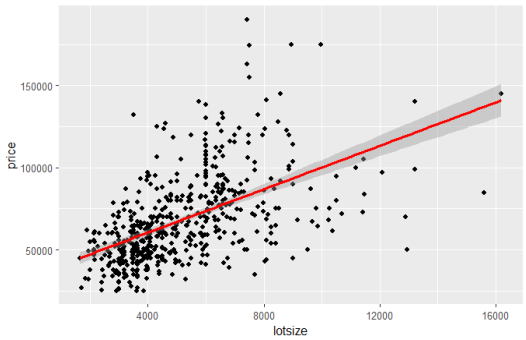
Les résultats de (1) se présentent comme suit :

	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 3 : Modèlé 1 de régression linéaire simple

La variable **lotsize** est très significative. Par conséquence, les résultats de ce modèle peuvent être expliqués selon le paradigme TCEPA. C'est-à-dire que l'augmentation d'une surface d'un m² entraine une hausse de prix du prix de vente de 6.499 $.

Toutefois, on observe que le R² 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 3 : 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 d'une maison et la surface de celle-ci n'est pas totalement linéaire mais plutôt quadratique.

Toutefois, tentons, 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

présentent 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.

A la suite de cela, l'équation du deuxième modèle de régression linéaire s'écrit donc comme suit :

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

Ce nouveau modèle a obtenu les résultats suivant :

	Dependent variable:
	price
lotsize	5.411***
	(0.388)
bedrooms	5,826.802***
	(1,206.571)
bathrms	19,750.210***
	(1,785.083)
Constant	-2,418.293
	(3,779.412)
Observations	546
R^2	0.486
Adjusted R^2	0.483
Residual Std. Error	19,191.610 (df = 542)
F Statistic	171.025*** (df = 3; 542)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tableau 4 : Modèlé 2 de régression linéaire multiple

Le R^2 est maintenant de 0.486 et β_1 est passé de de 6.59 à 5.411. 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.

NB : Les prédictions et les résidus de ce modèle ont été enregistrés dans un DataFrame nommé lm_1_results pour une analyse ultérieure.

Test de Fisher pour savoir si $\beta_1 = 0$

Le test de Fisher dans une régression multiple permet de tester la significativité des coefficients de régression pour savoir si une variable explicative influence réellement la variable à expliquer.

Le test de significativité de Fisher donne une pvaleur de 2.2e-16. Conclusion l'hypothèse nulle (H0) peut être rejetée avec un seuil d'erreur de 5%.

Estimons le modèle de régression linéaire multiple sans constante maintenant.

Le modèle devient :

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

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

	Dependent variable:
	price
lotsize	5.323***
	(0.362)
bedrooms	5,298.070***
	(878.726)
bathrms	19,530.840***
	(1,750.900)
Observations	546
R^2	0.932
Adjusted R^2	0.931
Residual Std. Error	19,181.170 (df = 543)
F Statistic	2,466.646*** (df = 3; 543)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tableau 5 : Modèle 3 de régression linéaire sans constante

Sans constante, l'ajustement du modèle de régression linéaire multiple passe de 0.486 à 0.932, soit une augmentation de près de 100%. On parvient à ce résultat parce que la suppression de la constante oblige le modèle à adopter davantage un ajustement linéaire au détriment de la fiabilité.

Pour préserver la qualité d'ajustement du modèle, il est préférable de conserver dans l'équation de régression la constante. Ainsi, on va plutôt chercher à améliorer la qualité du modèle grâce à une méthode de gestion des valeurs extrêmes qu'on a identifiées dans le Graphique 1.

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

	Dependent variable:
	price
lotsize	3.439***
	(0.302)
bedrooms	3,258.720***
	(935.095)
bathrms	11,810.330***
	(1,505.588)
Constant	23,362.370***
	(3,129.785)
Observations	441
R^2	0.367
Adjusted R^2	0.362

Residual Std. Error	13,273.930 (df = 437)
F Statistic	84.373*** (df = 3; 437)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tableau 6 : Modèle 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, une étude du poids des individus pourraient être effectuée. Celle-ci comprendrait l'identification des individus les plus << importants >> et l'impact de leur suppression sur l'ajustement 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 :

	Dependent variable:			
	price			
	(1)	(2)	(3)	(4)
lotsize	6.599***	5.411***	5.323***	3.439***
	(0.446)	(0.388)	(0.362)	(0.302)
bedrooms		5,826.802***	5,298.070***	3,258.720***
		(1,206.571)	(878.726)	(935.095)
bathrms		19,750.210***	19,530.840***	11,810.330***
		(1,785.083)	(1,750.900)	(1,505.588)
Constant	34,136.190***	-2,418.293		23,362.370***
	(2,491.064)	(3,779.412)		(3,129.785)
Observations	546	546	546	441
R^2	0.287	0.486	0.932	0.367
Adjusted R^2	0.286	0.483	0.931	0.362
Residual Std. Error	22,567.050 (df = 544)	19,191.610 (df = 542)	19,181.170 (df = 543)	13,273.930 (df = 437)
F Statistic	219.056*** (df = 1; 544)	171.025*** (df = 3; 542)	2,466.646*** (df = 3; 543)	84.373*** (df = 3; 437)
Note:				*p<0.1; **p<0.05; ***p<0.01

Tableau 7 : 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 R2 ajusté de 0.931 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 contenu des résultats des modèles (1) et (3), les estimations des paramètres de régression seraient biaisées.

Par conséquent, le modèle le plus fiable dans cette analyse est le modèle de régression (2) contenu de son coefficient de détermination ajusté et de la méthodologie employée pour sa construction. La suite de cette étude consistera à analyser 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, la non colinéarité entre les variables explicatives, et l'exogénéité.

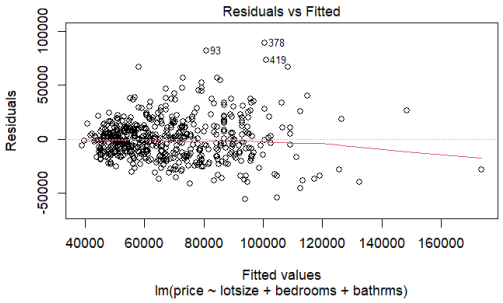
Etude des hypothèse de régression linéaire

- Espérance des résidus nulles : $E(u_i) = 0$

l'espérance des résidus est nulle : $-6.438936\text{e} - 13 \approx 0$

- Hétéroscédasticité

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



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 paramétrique de Sargan a été réalisé.

- Exogénéité des variables : $E\left(\frac{u_i}{xi}\right) = 0$

lotsize	bedrooms	bathrms
6.007705e-10	1.840208e-11	-1.093796e-12

Tableau 8 : Vérification de l'hypothèse d'exogénéité

https://stats.stackexchange.com/questions/26176/removal-of-statistically-significant-intercept-term-increases-r2-in-linear-mo/26205#26205