M2 TIDE

8 mars 2023

Projet d'économétrie des modèles linéaires : Boston House Price

Etude réalisée par :

Baptiste GOUMAIN

Adrien GARDONI

1 Évaluation des connaissances théoriques

1.1 Dans le cadre du modèle de régression multiple, que signifie "tester la significativité globale du modèle de la régression" ?

Tester la significativité globale du modèle de régression signifie évaluer si la relation entre les variables explicatives et la variable dépendante est statistiquement significative. Cela peut être fait en utilisant des tests statistiques tels que le test de Fisher permet de déterminer si l'ensemble des variables explicatives prises simultanément (à l'exception de la constante) permet d'expliquer les variations de la variable à expliquer.

Rappellons qu'être statistiquement significatif signifie que la relation observée entre deux variables ou des groupes de variables n'est pas simplement due au hasard, mais qu'elle est suffisamment importante pour être considérée significative d'un point de vue statistique.Pour cela on calcul une probabilité souvent appelée "p-value" ou "niveau de significativité", que l'on compare généralement à un seuil de 5%. Si la p-value est inférieure à ce seuil, cela indique qu'il est peu probable que le résultat observé soit dû au hasard, et que l'on peut donc conclure que la différence ou la relation entre les variables est statistiquement significative. En revanche, si la p-value est supérieure au seuil de significativité, on ne peut pas conclure que la différence ou la relation est statistiquement significative, et l'on considère alors que le résultat observé pourrait être dû au hasard.

1.2 Expliquez en mots et en détail ce que veut dire $\hat{\beta} \xrightarrow{d} N(\beta, \sigma^2\beta)$ pour l'estimateur $\hat{\beta}$ d'un paramètre β.

Cela signifie que l'estimateur $\hat{\beta}$ suit une distribution normale centrée sur la vraie valeur du paramètre β avec une variance $\sigma 2\beta$. En d'autres termes, cela signifie que si l'on répète l'échantillonnage et qu'on calcule l'estimateur $\hat{\beta}$ pour chaque échantillon, les valeurs de $\hat{\beta}$ suivront une distribution normale autour de la vraie valeur du paramètre β, avec une certaine variabilité donnée par la variance $\sigma 2\beta$. $\hat{\beta}$ est l'estimateur du paramètre β pour une régression multiple. Il mesure l'effet d'une variation d'une unité de la variable explicative sur la variable dépendante, en prenant en compte les effets des autres variables explicatives. $\hat{\sigma}2\ \beta$ mesure la variabilité de l'estimateur $\hat{\beta}$.

1.3 Définissez en mots le problème de multicollinéarité lors de la régression. Quelle est la conséquence de la multicollinéarité ? Quelle(s) solution(s) peut(peuvent) être envisagée(s) ?

La multicollinéarité se produit lorsque les variables explicatives sont fortement corrélées entre elles. Cela conduit à des estimations biaiseés des paramètres avec des variances importantes. Les conséquences de la multicollinéarité comprennent des coefficients de régression instables et des erreurs standard élevées. Les solutions pour la multicollinéarité comprennent la suppression de variables, la combinaison de variables, la réduction de la dimension, etc.

1.4 Lors de l'ajout d'une variable explicative supplémentaire à un modèle de régression, la mesure de l'ajustement R2 peut augmenter ou diminuer. Est-ce que cet énoncé est vrai, faux ou incertain ? Expliquez clairement.

Cet énoncé est incertain. L'ajout d'une variable explicative peut augmenter, diminuer ou ne pas affecter le mesure de l'ajustement R2. Cela dépend de la corrélation entre la variable explicative ajoutée et la variable dépendante ainsi que des autres variables explicatives déjà présentes dans le modèle. Lorsque la variable ajoutée est corrélée avec la variable à expliquer et/ou avec les variables explicatives existantes, l'ajout de cette variable peut améliorer la qualité de l'ajustement du modèle et ainsi augmenter la valeur de R2. Cependant, si la variable ajoutée n'est pas corrélée avec la variable à expliquer et/ou avec les variables explicatives existantes, elle n'apportera aucune information supplémentaire au modèle et peut même le rendre moins précis. Dans ce cas, l'ajout de la variable peut réduire la qualité de l'ajustement du modèle et diminuer la valeur de R2.

1.5 Dans quel cas pouvons-nous affirmer qu'un coefficient de régression est significatif à un niveau de 1% ?

Un coefficient de régression est considéré comme significatif à un niveau de 1% s'il est inférieur à 1% sur le test de significativité associé (par exemple, t-test ou test F). Par exemple, le test de student est basé sur le rapport entre l'estimateur du coefficient β et son erreur standard (une mesure de la précision de l'estimation). Si la valeur absolue de la statistique de test t est supérieure à la valeur critique correspondante du niveau de confiance choisi (ici 1% pour un test à 1%), on peut rejeter l'hypothèse nulle selon laquelle le coefficient est égal à zéro et conclure que le coefficient est significativement différent de zéro à ce niveau de confiance.

1.6 Ci-après un extrait des résultats d'une régression multiple : R2 = 0.660, SSR = 8633.165, F(3, 1656) = 1.044e + 3, P rob > F = 0.000. Combien de variables explicatives ont été considérées dans ce modèle ? Quel est le nombre d'individus utilisés dans cet analyse ?

Il y a 3 variables explicatives considérées dans ce modèle (F (3, 1656)). Le nombre d'individus utilisés dans l'analyse est de 1656 + 3 + 1 = 1657.

2 Le projet

2.1 Plan :

A. Modèle explicatif

1) Statistiques descriptives
2) Définition d'un modèle de régression linéaire multiple avec stat model, test de student et fisher
3) Verification des hypothèses :
   - La normalité des résidus
   - L'Homoscédasticité des résidus
   - Autocorrélation des résidus
   - Multicollinéarité
4) Solution pour palier à la violation des hypothèse

B. Modèle prédictif

1) On divise X et y en train et test
2) Utiliser le modele sélectionné en I. sur les train test etc

2.2 Introduction

Comprendre la finalité d'un résultat par l'étude de facteurs pouvant influer dessus est la base d'une étude statistique linéaire. Il existe beaucoup d'analyses possibles en fonction des données disponibles. Le cadre de notre étude portera ici sur l'immobilier, en s'attardant sur l'explication des prix medians de logements.

Chaque enregistrement dans la base de données décrit une banlieue ou une ville de Boston. Les données proviennent de la zone statistique métropolitaine standard de Boston (SMSA) en 1970. Les variables sont expliquées ci dessous (tirés du référentiel de Machine Learning UCI1) :

CRIM : taux de criminalité par habitant dans la ville
ZN : proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés.
INDUS : proportion d'acres d'activités non commerciales par ville
CHAS : variable fictive de la rivière Charles (= 1 si la zone borde la rivière, 0 sinon)
NOX : concentration d'oxydes d'azote (en parties par 10 millions)
RM : nombre moyen de chambres par logement
AGE : proportion d'unités occupées par leur propriétaire construites avant 1940
DIS : distances pondérées vers les cinq centres d'emploi de Boston
RAD : indice d'accessibilité aux autoroutes radiales
TAX : taux de taxe foncière à la valeur totale par tranche de 10 000 $
PTRATIO : ratio élève-enseignant par ville
B : 1000 (Bk-0,63)2 où Bk est la proportion de Noirs par ville
LSTAT : % de la population ayant un statut inférieur
MEDV : Valeur médiane des maisons occupées par leur propriétaire en milliers de dollars.

A travers ces données, nous essaierons de répondre à la problématique suivante :

Pour cela nous présenterons, et décrierrons nos données, puis nous initialiseront notre modèle et les hypothèses qui le composent. Nous effectuerons les modifications nécessaires des données pour avoir des résuktats interpretables et non biaisés, puis nous expliquerons nos résultats et effectueront un modèle prédictif adapté.

2.3 A- Modele explicatif

Un modèle explicatif en économétrie est un modèle statistique qui vise à expliquer la relation entre une variable d'intérêt (la variable dépendante) et un ensemble de variables explicatives (les variables indépendantes). Le but est de déterminer si les variables indépendantes ont une influence significative sur la variable dépendante, et dans quelle mesure.

En général, un modèle explicatif cherche à expliquer une variable d'intérêt en fonction de variables explicatives qui sont supposées causer ou être liées à la variable d'intérêt. Les modèles explicatifs en économétrie sont souvent utilisés pour étudier des relations causales entre les variables, pour évaluer l'impact de politiques publiques ou de changements économiques, ou pour prédire des résultats futurs.

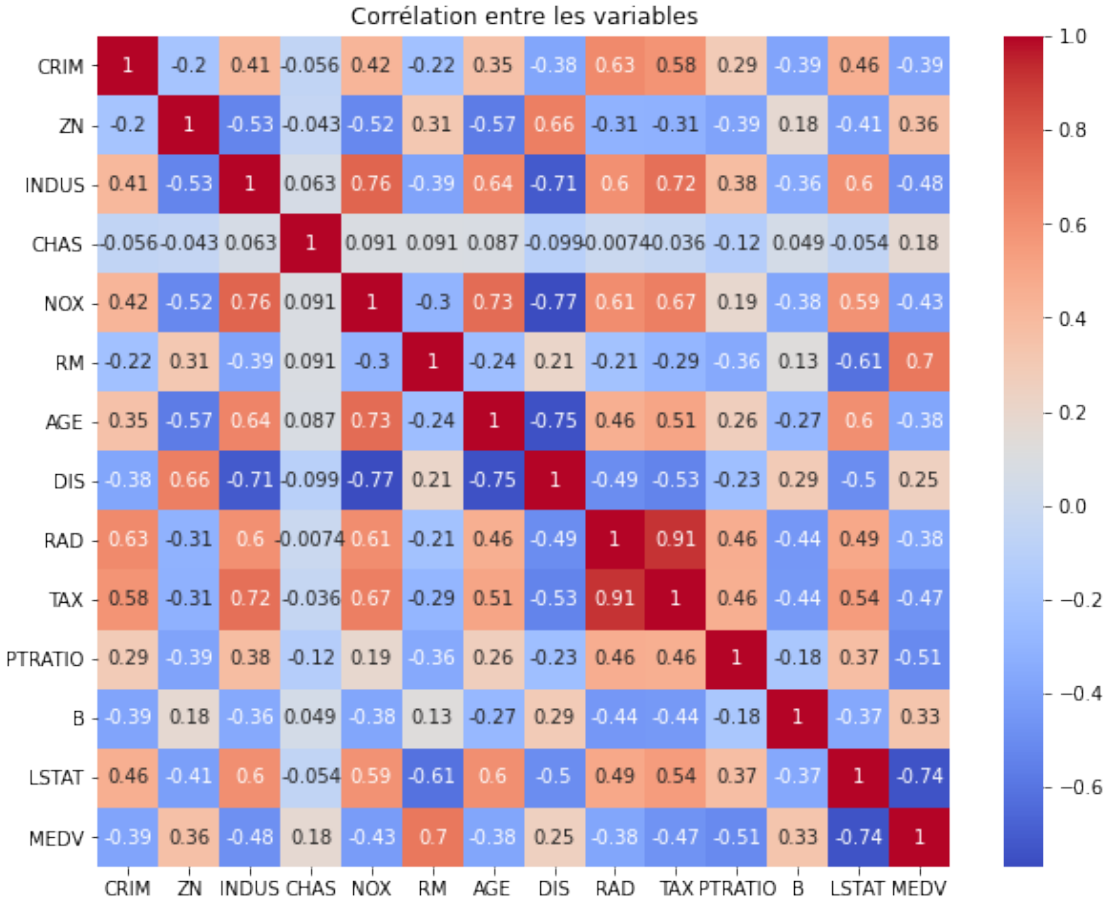
```
Nombre d'observations : 507
Nombre de variables : 14
Données manquantes : 0
```

Cette première approche est nécessaire pour constater la taille de notre jeu de données, et les éventuelles valeurs manquantes que nous aurions du traiter dans le cadre de notre étude. Nous pouvons voir que la base est relativement petite, et qu'aucune données manquante n'est à signaler.

2.3.1 Statistiques descriptives

```
            Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   CRIM     506 non-null    object
 1   ZN       506 non-null    object
 2   INDUS    506 non-null    object
 3   CHAS     506 non-null    object
 4   NOX      506 non-null    object
 5   RM       506 non-null    object
 6   AGE      506 non-null    object
 7   DIS      506 non-null    object
 8   RAD      506 non-null    object
 9   TAX      506 non-null    object
 10  PTRATIO  506 non-null    object
 11  B        506 non-null    object
 12  LSTAT    506 non-null    object
 13  MEDV     506 non-null    object
```

Nous commençons par regarder le type de nos variables, nous remarquons que toutes nos variables sont de types "object" et que nous allons donc devoir les convertir en variable numérique pour éviter tout problème par la suite.



L'objectif étant de modéliser la variable MEDV, la matrice de corrélation nous permet de faire une première analyse sur les relations entre les différentes variables de la base et MEDV. Cette étape indispensable nous permet de dire que visiblement toutes nos variables ont une corrélation avec notre variable cible et nous verrons par la suite si cette corrélation est significative quant à l'explication de la variable cible.

	CRIM	ZN	INDUS	CHAS	NOX	RM	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

	AGE	DIS	RAD	TAX	PTRATIO	B	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	
std	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	
min	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	

	LSTAT	MEDV
count	506.000000	506.000000
mean	12.653063	22.532806
std	7.141062	9.197104
min	1.730000	5.000000
25%	6.950000	17.025000
50%	11.360000	21.200000
75%	16.955000	25.000000
max	37.970000	50.000000

Dans ce tableau nous pouvons voir les statistiques descriptives de chacunes de nos variables et ce qui nous permet de mieux nous familiariser avec celles-ci.

Suppresion des valeurs aberrantes

Après plusieurs tentatives différentes pour tenter de supprimer intelligemment d'éventuelles valeurs aberrantes, nous décidons de les garder car au vu de la petite taille d'échantillon, nous ne voulons pas perdre plus d'informations. Nous fausserions probablement les résultats en retirant des observations.

2.3.2 Définition d'un modèle de régression linéaire multiple

Le modèle peut s'écrire par la formule suivante :

$$y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots + \beta k * Xk + \varepsilon$$

où y est la variable dépendante (la variable cible), β0 est l'interception, β1 à βk sont les coefficients de régression associés aux variables explicatives X1 à Xk, et ε est l'erreur aléatoire qui représente

la partie non expliquée de la variation de y.

Le coefficient de détermination, noté ici R^2, est une mesure de la proportion de la variation totale de la variable dépendante qui est expliquée par le modèle de régression. Il est défini comme le rapport entre la variation expliquée par le modèle et la variation totale de la variable dépendante :

$$R = 1 - (SSE/SST)$$

avec SSE (Sum of Squared Errors) qui est la somme des carrés des résidus, c'est-à-dire la somme des différences entre les valeurs observées de la variable dépendante et les valeurs prédites par le modèle, et SST (Sum of Squared Total) qui est la somme des carrés des écarts par rapport à la moyenne de la variable dépendante.

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Tue, 07 Mar 2023	Prob (F-statistic):	6.72e-135
Time:	23:56:03	Log-Likelihood:	-376.55
No. Observations:	506	AIC:	781.1
Df Residuals:	492	BIC:	840.3
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.235e-16	0.023	-2.28e-14	1.000	-0.045	0.045
CRIM	-0.1010	0.031	-3.287	0.001	-0.161	-0.041
ZN	0.1177	0.035	3.382	0.001	0.049	0.186
INDUS	0.0153	0.046	0.334	0.738	-0.075	0.105
CHAS	0.0742	0.024	3.118	0.002	0.027	0.121
NOX	-0.2238	0.048	-4.651	0.000	-0.318	-0.129
RM	0.2911	0.032	9.116	0.000	0.228	0.354
AGE	0.0021	0.040	0.052	0.958	-0.077	0.082
DIS	-0.3378	0.046	-7.398	0.000	-0.428	-0.248
RAD	0.2897	0.063	4.613	0.000	0.166	0.413
TAX	-0.2260	0.069	-3.280	0.001	-0.361	-0.091
PTRATIO	-0.2243	0.031	-7.283	0.000	-0.285	-0.164
B	0.0924	0.027	3.467	0.001	0.040	0.145
LSTAT	-0.4074	0.039	-10.347	0.000	-0.485	-0.330

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	9.82

Pour notre première régression, nous ne pré-traitons pas nos données, et utilisons les résultats pour voir les corrections que nous devrons apporter. Nous constatons un R2 de 74%, ce qui est correct. Mais nous pouvons déjà voir avec les t-student que deux variables, INDUS et AGE, ne sont pas significatives. Nous les retirons donc pour la suite de l'étude. Les autres statistiques de test nous indiquent également que dans le cadre d'un modèle linéaire, plusieurs hypothèses sont à vérifier, c'est ce que nous verrons par la suite avant de revenir sur l'interprétation des différents coefficients.

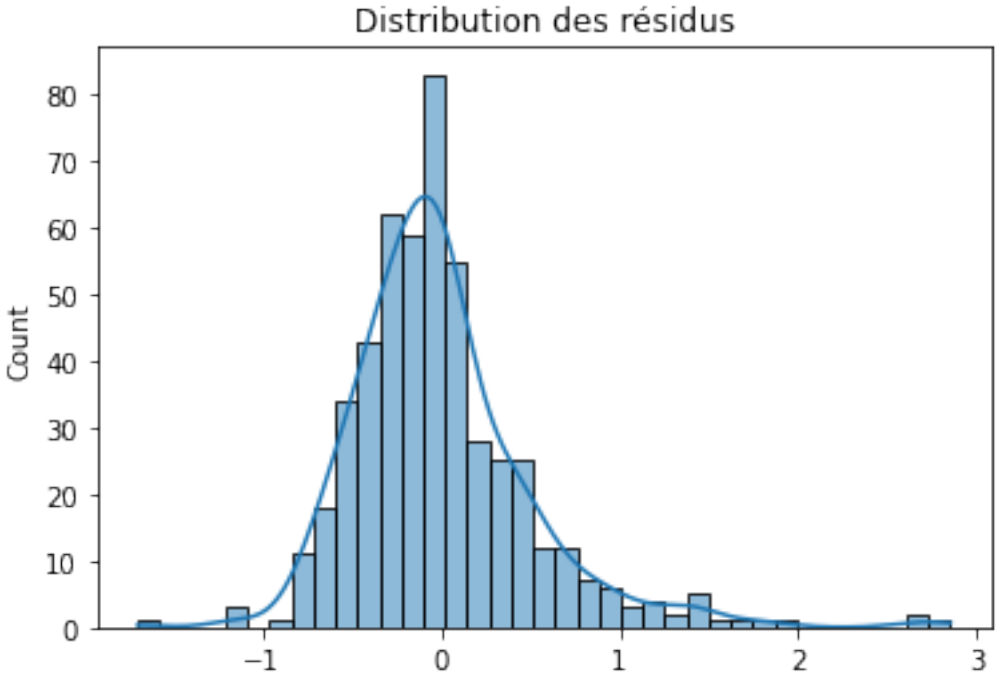
2.3.3 Vérification des hypothèses

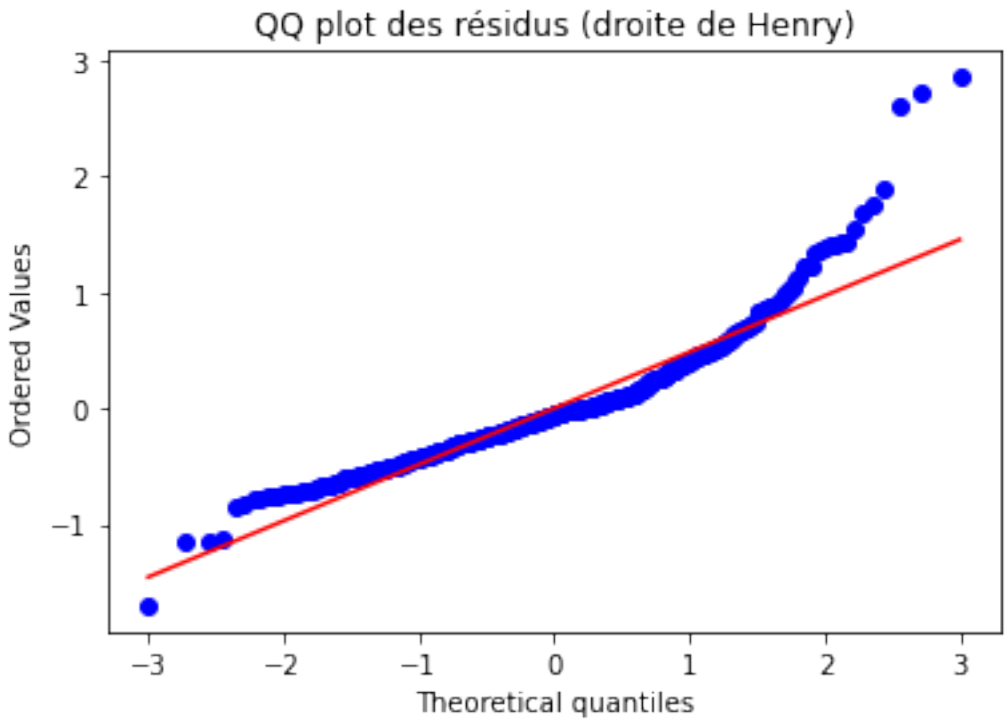
Nous allons vérifier certaines hypothèses :

H1 Normalité des résidus : Les résidus aléatoires doivent suivre une distribution normale.
H2 Homoscédasticité des résidus : L'écart-type de l'erreur aléatoire doit être constant pour toutes les valeurs de X.
H3 Autocorrélation des résidus : Les résidus aléatoires ε doivent être indépendantes les unes des autres.
H4 Multicollinéarité : Les variables explicatives doivent être linéairement indépendantes les unes des autres.

H1 : La normalité des résidus

Graphique de la distribution et droite de Henry





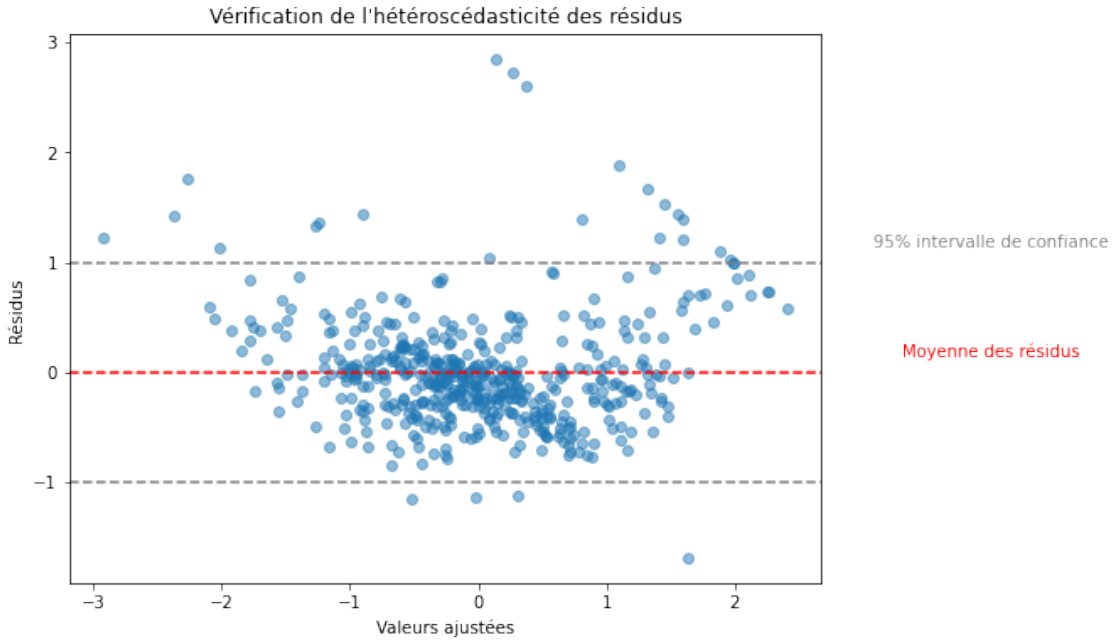
Le test de Shapiro-Wilk

```
Test de Shapiro-Wilk pour la normalité des résidus :
ShapiroResult(statistic=0.9013806581497192, pvalue=1.4801657562168808e-17)
```

Visuellement, nous aurions pu penser que les résidus étaient normaux, or, le test nous indique le contraire. Mais au vu de la taille restreinte de notre base de données, nous avons de fortes chances pour que la non normalité soit due à la taille du jeu de données. Nous considérerons pour la suite nos résidus comme normaux.

H2 : L'Homoscédasticité des résidus

Graphique



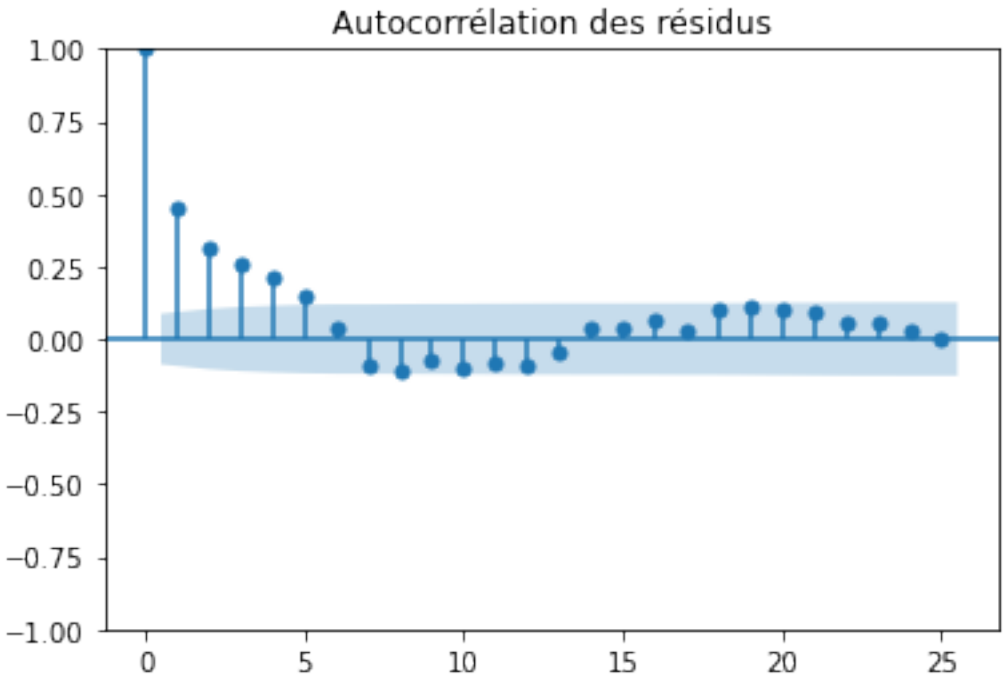
Test de Breusch-Pagan

```
Test de Breusch-Pagan:
['Lagrange multiplier statistic', 'p-value', 'f-value', 'f p-value']
(59.7718556889478, 1.0222228508295184e-08, 6.015532043780897,
2.9637191487426812e-09)
L'hypothèse d'hétéroscédasticité est acceptée.
```

Encore une fois ici, nous avons un graphique qui tend plus à ce que l'hypoyhèse d'hétéroscédasticité des résidus soit rejetée. Nous n'avons pas une forme de cone pour les résidus, et la variance semble stable et comprise dans l'intervale de confiance. Mais en effectuant le test de Breusch-Pagan, l'hypothèse d'homoscédasticité est rejetée. Nous ajusteront donc le modèle en conséquence.

H3 : Autocorrélation des résidus

Graphique ACF



Le test de Durbin-Watson

```
Test de Durbin-Watson : 1.0783751186797246
Présence d'une autocorrélation positive des résidus.
```

Nous constatons que plusieurs barres sortent de l'intervalle de confiance, ce qui suppose que les résidus sont autocorrelés. Nous validons cette hypothèse en réalisant le test de Durbin-Watson qui nous confirme bien cela. Nos résidus sont donc autocorrelés et hétéroscédastique.

H3 : Multicollinéarité

VIF

	VIF Factor	features
0	1.0	const
1	1.8	CRIM
2	2.2	ZN
3	1.1	CHAS
4	3.8	NOX
5	1.8	RM
6	3.4	DIS
7	6.9	RAD
8	7.3	TAX
9	1.8	PTRATIO
10	1.3	B
11	2.6	LSTAT

La dernière hypothèse à vérifier est la multicolinéarité. Les VIF (variance inflation factor) mesurent le degré de multicollinéarité entre les variables explicatives. Plus précisément, un VIF élevé pour une variable donnée indique que cette variable est corrélée avec une ou plusieurs autres variables explicatives du modèle. En général, le seuil acceptable est autour de 5. Le VIF de TAX étant trop élevé, nous supprimerons cette variable pour corriger le problème de multicolinéarité.

2.3.4 Solution pour palier à la violation des hypothèse

Nous allons donc procéder à une régression linéaire multiple avec MCOP (Méthode des Moindres Carrés Ordinaires Partiels) qui permet de pallier à la violation des hypothèses de la régression linéaire classique, notamment en ce qui concerne la normalité des résidus et l'homoscédasticité.

En effet, lorsque les hypothèses de la régression linéaire classique sont violées, les estimations des coefficients de régression peuvent être biaisées et les intervalles de confiance et tests d'hypothèses associés peuvent être incorrects. La méthode des moindres carrés ordinaires partiels permet de minimiser l'impact de ces violations d'hypothèses en ajustant le modèle linéaire de manière robuste.

Plus précisément, la méthode MCOP consiste à projeter les variables explicatives sur un sous-espace de plus faible dimension, tout en conservant la plus grande quantité d'information possible. Cela permet de réduire la dimension de l'espace des variables explicatives et de se concentrer sur les variables les plus significatives pour expliquer la variation de la variable cible.

En outre, la méthode MCOP permet également de traiter des données multicollinéaires. Cette technique permet de sélectionner les variables explicatives les plus pertinentes pour expliquer la variation de la variable cible, en évitant les biais dus à la redondance de l'information.

En résumé, la méthode des moindres carrés ordinaires partiels permet de modéliser la relation entre notre variable cible et nos variables explicatives de manière robuste, en minimisant l'impact des violations des hypothèses de la régression linéaire classique. Cette méthode est particulièrement utile lorsque les données présentent des caractéristiques qui ne sont pas conformes aux hypothèses classiques de la régression linéaire.

WLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.998
Model:	WLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.863e+04
Date:	Tue, 07 Mar 2023	Prob (F-statistic):	0.00
Time:	23:56:05	Log-Likelihood:	-394.84
No. Observations:	506	AIC:	811.7
Df Residuals:	495	BIC:	858.2
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0023	0.007	0.344	0.731	-0.011	0.015
CRIM	-0.0938	0.010	-9.872	0.000	-0.112	-0.075
ZN	0.1030	0.004	22.900	0.000	0.094	0.112
CHAS	0.0725	0.013	5.608	0.000	0.047	0.098
NOX	-0.1561	0.008	-18.389	0.000	-0.173	-0.139
RM	0.3266	0.005	68.355	0.000	0.317	0.336
DIS	-0.3025	0.007	-45.675	0.000	-0.316	-0.290
RAD	0.0283	0.010	2.777	0.006	0.008	0.048
PTRATIO	-0.2153	0.004	-51.594	0.000	-0.224	-0.207
B	0.0787	0.005	15.322	0.000	0.069	0.089
LSTAT	-0.4314	0.007	-57.716	0.000	-0.446	-0.417

Omnibus:	335.619	Durbin-Watson:	1.772
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47273.796
Skew:	-1.889	Prob(JB):	0.00
Kurtosis:	50.201	Cond. No.	117.

Dans ce cas, il semble que le modèle n'a pas de problème d'endogénéité car le test Durbin-Watson est proche de 2 (1.772) indiquant qu'il n'y a pas d'autocorrélation des résidus et le test d'Omnibus indique que les résidus ne sont pas distribués de manière non normale (Prob(Omnibus) < 0.05).

Après suppression de la variable TAX, nous pouvons voir que tous nos VIF sont bas. Ce qui confirme l'absence de multicolinéarité.

Voici les interprétations des coefficients de régression sur la variable cible MEDV :

CRIM : le coefficient négatif de -0.0938 suggère que le taux de criminalité a un impact négatif sur la valeur médiane des maisons. Ainsi, une hausse de 1 point dans le taux de criminalité par habitant dans la ville est associée à une baisse de 0,0938 milliers de dollars (soit environ 93 dollars) dans la valeur médiane des maisons.

ZN : le coefficient positif de 0.1030 indique que la proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés a un impact positif sur la valeur médiane des maisons. Ainsi, une augmentation de 1 point de la proportion de terrains zonés pour des lots de plus de 25 000 pieds

carrés est associée à une augmentation de 0,1030 milliers de dollars (soit environ 103 dollars) dans la valeur médiane des maisons.

CHAS : le coefficient positif de 0.0725 suggère que la variable fictive de la rivière Charles a un impact positif sur la valeur médiane des maisons. Ainsi, une maison qui borde la rivière Charles peut valoir en moyenne 0,0725 milliers de dollars (soit environ 72 dollars) de plus que celle qui ne la borde pas.

NOX : Le coefficient de -0.1561 indique une relation négative entre la concentration d'oxydes d'azote et la valeur médiane des maisons. Cela peut s'expliquer par le fait que les oxydes d'azote sont des polluants atmosphériques qui peuvent avoir des effets négatifs sur la santé humaine et la qualité de vie en général. Les acheteurs de maisons peuvent donc être prêts à payer moins cher pour des maisons situées dans des zones avec une concentration plus élevée de polluants atmosphériques.

RM : Le coefficient de 0.3266 indique une relation positive entre le nombre moyen de chambres par logement et la valeur médiane des maisons. Cela peut s'expliquer par le fait que les maisons avec plus de chambres sont généralement plus grandes et offrent plus d'espace pour les occupants, ce qui peut augmenter leur valeur.

DIS : Le coefficient de -0.3025 indique une relation négative entre la distance pondérée vers les cinq centres d'emploi de Boston et la valeur médiane des maisons. Cela peut s'expliquer par le fait que les maisons situées plus près des centres d'emploi sont plus pratiques pour les travailleurs et donc plus demandées, ce qui peut augmenter leur valeur.

PTRATIO : Le coefficient de -0.2153 indique une relation négative entre le ratio élève-enseignant par ville et la valeur médiane des maisons. Cela peut s'expliquer par le fait que les acheteurs de maisons peuvent être prêts à payer plus cher pour des maisons situées dans des zones avec un ratio élève-enseignant plus faible, car cela peut indiquer des écoles de meilleure qualité et donc une meilleure qualité de vie pour les enfants.

B : Le coefficient de 0.0787 indique une relation positive entre la proportion de Noirs par ville et la valeur médiane des maisons, une relation qui peut sembler contre-intuitive. Cependant, il convient de noter que la formule utilisée pour calculer B inclut une valeur de référence de 0,63, ce qui signifie que les coefficients reflètent la relation entre la proportion de Noirs par ville et la valeur médiane des maisons après avoir tenu compte de cette valeur de référence. De plus, il est possible que la proportion de Noirs soit liée à d'autres facteurs qui influencent la valeur des maisons, tels que la diversité culturelle, l'accès aux services et aux infrastructures, etc.

LSTAT : Le coefficient de -0.4314 indique une relation négative entre le pourcentage de la population ayant un statut inférieur et la valeur médiane des maisons. Cela peut s'expliquer par le fait que les acheteurs de maisons peuvent être prêts à payer moins cher pour des maisons situées dans des zones avec une proportion plus élevée de personnes ayant un statut inférieur, car cela peut indiquer des problèmes socio-économiques dans la région, tels que la pauvreté, le chômage, la criminalité, etc.

En fonction des coefficients de régression, on peut dire que les variables les plus importantes pour prédire la variable cible MEDV sont RM, LSTAT et DIS.

Le coefficient positif de RM indique une forte relation positive entre le nombre moyen de chambres par logement et la valeur médiane des maisons. Cela suggère que les maisons avec plus de chambres ont tendance à être plus grandes et à offrir plus d'espace, ce qui peut augmenter leur valeur.

Le coefficient négatif de LSTAT indique une forte relation négative entre le pourcentage de la population ayant un statut inférieur et la valeur médiane des maisons. Cela suggère que les maisons

situées dans des zones où la population a un statut socio-économique plus faible ont tendance à avoir une valeur médiane plus faible. Cela peut être dû à des facteurs tels que la pauvreté, le chômage, la criminalité, etc.

Enfin, le coefficient négatif de DIS indique une forte relation négative entre la distance pondérée vers les cinq centres d'emploi de Boston et la valeur médiane des maisons. Cela suggère que les maisons situées plus près des centres d'emploi ont tendance à avoir une valeur médiane plus élevée, car elles sont plus pratiques pour les travailleurs.

Néanmoins toutes les variables ont une influence sur la valeur médiane des maisons, même si certaines ont un impact plus fort que d'autres. Il est donc important de prendre en compte toutes les variables lors de la prédiction de la valeur des maisons.

2.4 Modele prédictif

Un modèle prédictif est conçu pour prédire la valeur de la variable dépendante en fonction des valeurs des variables indépendantes.Ils sont souvent utilisés dans des contextes où il est important de prédire l'avenir ou de prendre des décisions en temps réel. Contrairement aux modèles explicatifs, les coefficients des modèles prédictifs ne sont pas nécessairement interprétables en termes de causalité, mais ils sont choisis pour maximiser la capacité du modèle à prédire le résultat de la variable cible sur de nouvelles données.

2.4.1 Utilisons le modèle sélectionné en I. sur nos données d'entrainement et de test

Après avoir divisé notre jeu de données en données d'entraînement et de test nous avons entraîner le modèle sélectionné sur celles-ci. Nous obtenons les résultats suivant :

```
RMSE: 0.5360159851419014
R^2: 0.6692541090701765
```

Le modèle prédictif est nettement moins moins précis que le modèle explicatif, ce qui est normal au vu de la petite taille du jeu de données. Nous avons néanmoins trouvé qu'il était intéressant de l'intégrer à l'étude pour nuancer l'analyse, et montrer la necessité de certains éléments (comme la taille de l'échantillon) dans les études statistiques.