Master Traitement de l'Information
et Datascience en Entreprise (TIDE)

Econométrie des modèles linéaires

KORAIBI Kamar
YALAP Sophia

Mars 2022

Contents

Chapter 1

Introduction

Boston Housing est un jeu de données classique dans le monde de la science des données, utilisé pour comparer les algorithmes de régression de machine learning et appliquer des méthodes de sélection de variables. Il a été collecté par l'US Census Service Boston Mass area et a été initialement publié dans le document de recherche, Harrison, D. et Rubnfeld, D.L « Hedonic prices and the demand for clean air ». Chaque enregistrement de la base de données décrit une banlieue ou un quartier de Boston. Les données ont été tirées de la zone statistique métropolitaine standard (SMSA) de Boston en 1970.

Dans ce rapport nous allons réaliser une étude statistique sur ces données avec pour objectif de développer un modèle explicatif et prédictif. La variable à expliquer et prédire est la valeur médiane des maisons occupées par leur propriétaire en milliers de dollars.

Afin de répondre à cette problématique, nous allons dans un premier temps réaliser une analyse statistique descriptive des données à l'aide de graphiques et d'outils statistiques classiques. Nous allons ensuite modéliser les données à l'aide d'une méthode de régression linéaire simple puis multiple avant de réaliser une sélection de modèle qui nous permettra par la suite de mettre en place un modèle explicatif et un modèle prédictif. Enfin, nous finirons par conclure sur les résultats de cette étude.

Chapter 2

Analyse des données

La base de données Boston Housing contient 506 observations et 14 variables, telles que le taux de criminalité, la qualité de l'air, le nombre moyen de pièces par logement etc (voir la table 4.1 pour une description complète des variables). Après vérification, il n'y a aucune valeur manquante dans notre base de données.

La base contient exclusivement des variables numériques également appelées quantitatives. Deux types de donneés numériques sont présentes :

- **Variables continues** : ce sont des variables numériques décimales notées *num* sur l'image ci-dessous. Elles représentent 12 des 14 variables présentes dans le jeu de données.

- **Variables discrètes** : ce sont ici des variables entières notées *int* sur l'image ci-dessous. Nous avons effectué un traitement sur ces dernières afin de les transformer en facteurs. En effet, elles n'ont pas le même comportement que les variables continues et la valeur qui leur est attribuée n'est qu'une manière plus simple de les catégoriser. Dans la suite de notre analyse, nous les considèrerons donc comme des variables catégorielles.

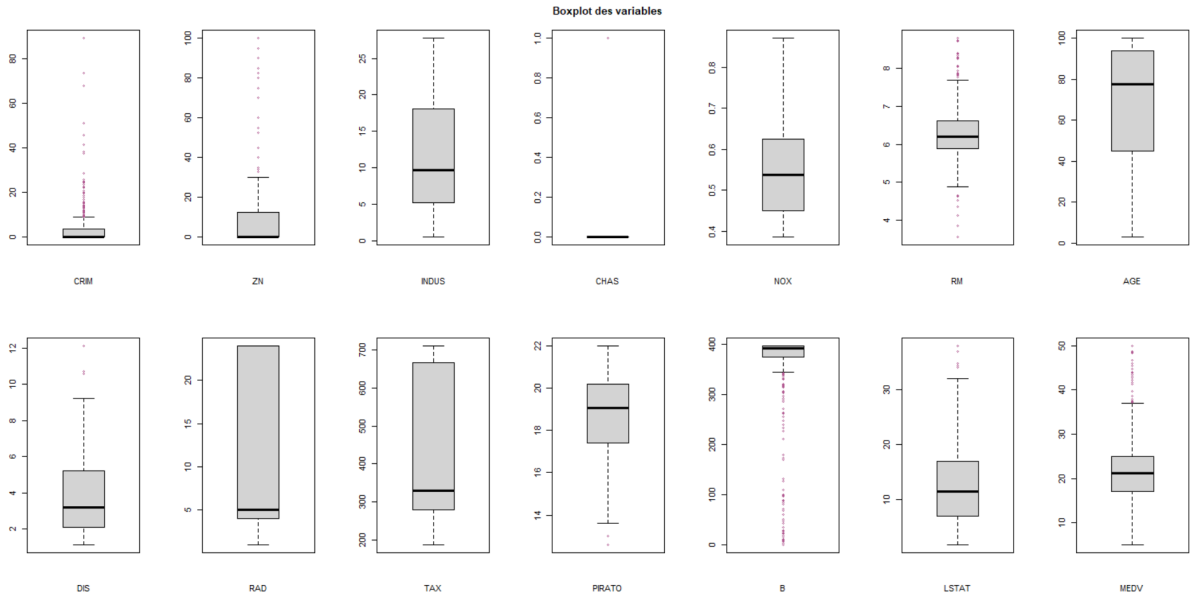
```
'data.frame':   506 obs. of  14 variables:
 $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ CHAS   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ RM     : num  6.58 6.42 7.18 7 7.15 ...
 $ AGE    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ RAD    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ TAX    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ B      : num  397 397 393 395 397 ...
 $ LSTAT  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

2.1 Visualisation

Avant toute chose, il est important de connaître la structure et le contenu de nos données. La commande *summary* nous permet d'avoir une description rapide de ces dernières.

```
      CRIM                ZN              INDUS            CHAS              NOX              RM            AGE       
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      DIS              RAD              TAX            PTRATIO             B              LSTAT            MEDV      
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

On y retouve pour chaque variable, sa valeur minimum, maximum, sa médiane, sa moyenne ainsi que son premier et troisième quartile. Nous pouvons avoir un premier apperçu de la structure de chacune de nos variables. Il est également possible de visualiser ces données à l'aide d'un *boxplot* autrement appelé *boîte à moustache*.



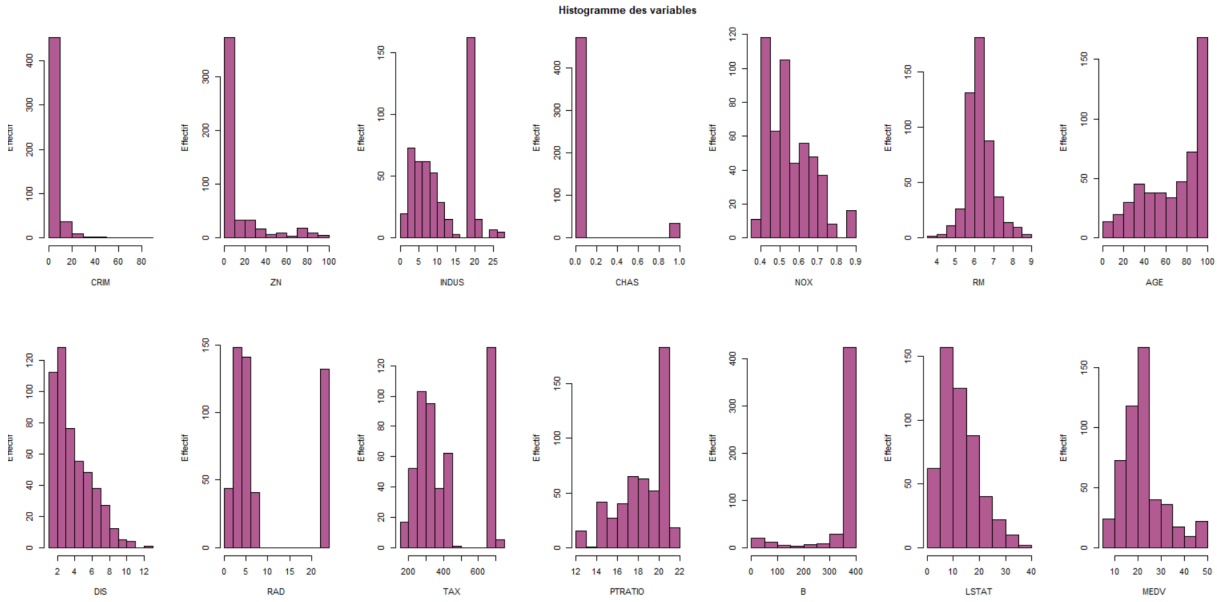
Dans la suite, nous allons réaliser nos analyses sur trois axes :

- **Analyse univariée**: permet d'étudier chacune des variables l'une après l'autre.
- **Analyse bivariée**: permet d'étudier d'éventuelles liaisons entre un couple de variables.
- **Analyse multivariée**: permet d'étudier d'éventuelles liasions avec toutes les variables entre elles.

2.1.1 Analyse univariée

Variables numériques

Il est possible de représenter les données sous forme d'histogrammes. On remarque que les distributions sont différentes, la variable RM semble suivre une loi normale tandis que d'autres semblent asymétriques. En effet, on peut voir que certaines variables telles que MEDV, LSTAT et DIS indiquent une distribution décalée à gauche contraiement à AGE et PTRATIO. Par ailleurs, les variables transformées en catégorielles lors de cette analyse ne prennent que certains valeurs (catégories) possibles. CHAS étant une variable binaire, elle prend une valeur de 0 ou 1. RAD, constituée de 9 catégories prendra donc 9 valeurs différentes (on pourrait changer le nombre de bacs pour une meilleure visibilité de cet histogramme). Enfin, on constate que la distribution de certaines variable peut être impactée par la présence de valeurs aberrantes.



Variables catégorielles

Nous pouvons représenter ces variables à l'aide d'un tableau d'effectif :

- Tableau d'effectif de la variable CHAS

0	1
471	35

- Tableau d'effectif de la variable RAD

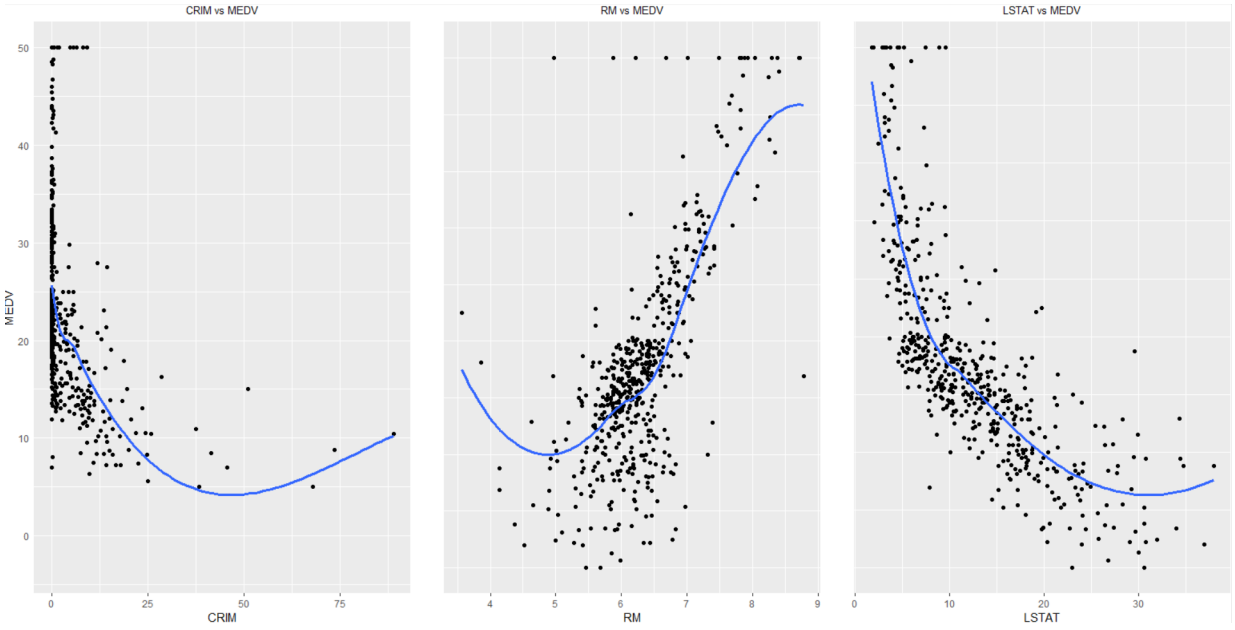
1	2	3	4	5	6	7	8	24
20	24	38	110	115	26	17	24	132

Nous pouvons observer que la variable CHAS qui vaut 1 si la parcelle borde la rivière et 0 sinon n'est pas équilibrée car la plupart des logements ne bordent pas la rivière.

2.1.2 Analyse bivariée

Variables numériques

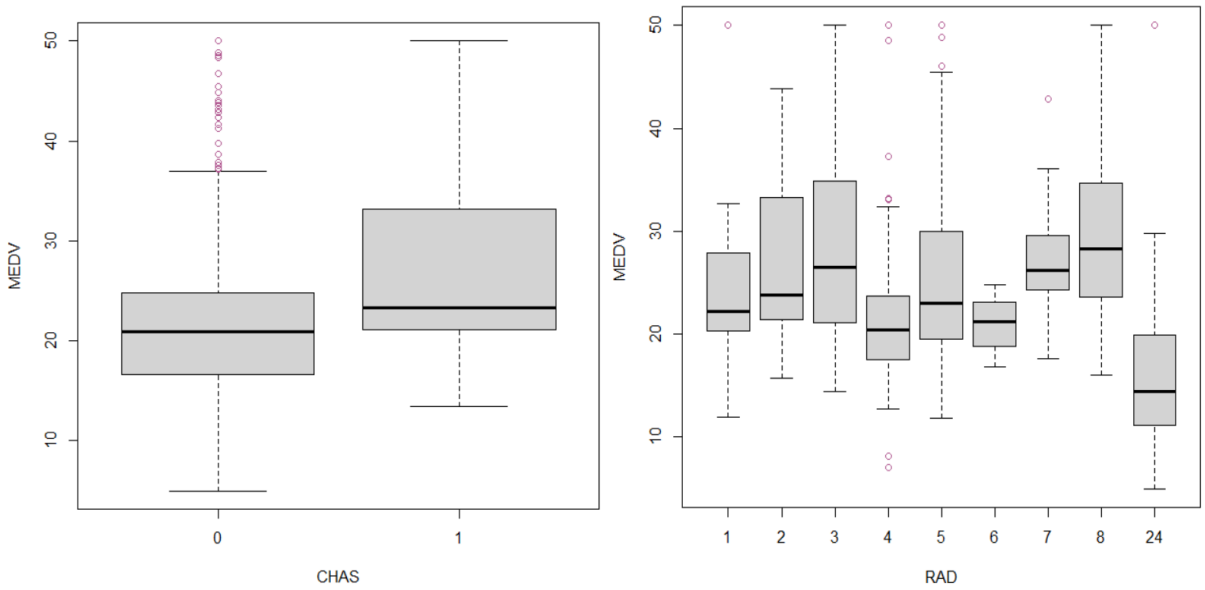
Puisque notre variable à expliquer est MEDV, nous allons coupler cette variable avec d'autres afin d'observer l'évolution de certaines variables en fonction de celle-ci.



Variables catégorielles

Le tableau d'effectif suivant permet de coupler les variables catégorielles et distinguer ainsi plus précisemment les relations entre ces deux variables.

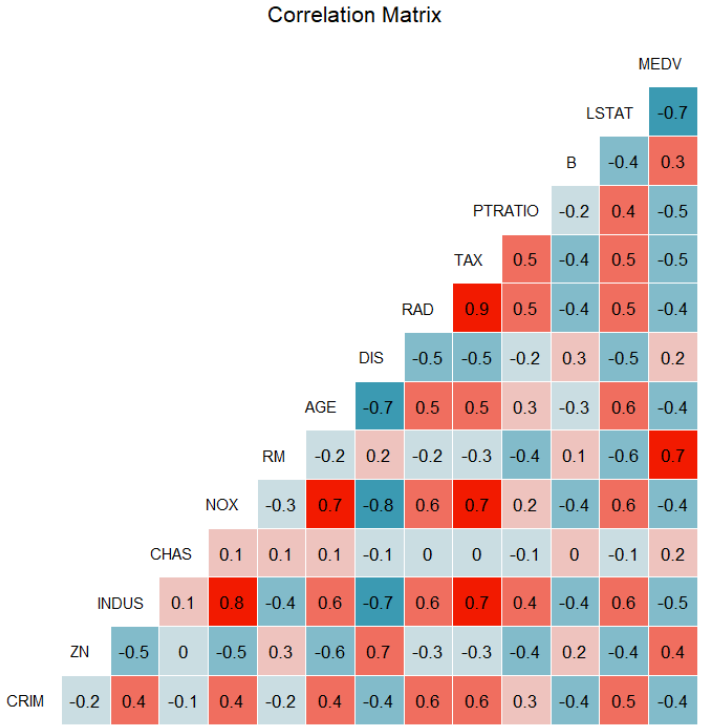
	1	2	3	4	5	6	7	8	24
0	19	24	36	102	104	26	17	19	124
1	1	0	2	8	11	0	0	5	8



On observe à l'aide du graphique de gauche que les médianes ne sont pas éloignées. Cependant, les outliers ne sont observés que pour le cas 0 dans lequel le logement ne borde pas la rivière. On observe à l'aide du graphique de droite que la médiane de la valeur médiane des maisons est bien plus basse pour un indice d'accessibilité aux autoroutes radiales élevé.

2.1.3 Analyse multivariée

Correlation



La matrice de corrélation ci-dessus montre qu'il y a une forte corrélation entre les variables TAX et RAD, dont le coefficient s'élève à 0.9. Ensuite, on remarque une corrélation de 0.8 entre NOX et INDUS et de -0.8 entre NOX et DIS. Par ailleurs, les variables les plus corrélées à notre variable d'intérêt MEDV sont LSTAT et RM avec un coefficient de 0.7 en valeur absolue. Nous pouvons également afficher les variables qui sont le plus corrélées à la variable à expliquer MEDV.

2.1.4 Outliers

Un outlier est une donnée qui est éloignée des autres observations, c'est une donnée qui se trouve en dehors de la distribution globale de l'ensemble des données. A l'aide du graphique précédent, on remarque que certaines variables ont un grand nombre d'outliers. On a donc décidé d'afficher le pourcentage d'outliers par variable (table 2.1).
Même si le nombre d'outliers nous parraissait élevé pour certaines variables, nous observons que le pourcentage d'outliers ne dépasse pas 16%. Ainsi, nous considérons qu'il n'y a pas besoin de supprimer ces données car elles ne représentent pas une grande partie de ces dernières.

Table 2.1: Variables avec des outliers

Variable	B	ZN	CRIM	MEDV	CHAS	RM	PTRATIO	LSTAT	DIS
% d'outliers	15.02	13.44	13.04	7.31	6.92	5.93	2.96	1.19	0.99

2.2 Statistiques descriptives

2.2.1 Description intrinsèque

Dans cette section, on étudiera la variable d'intérêt MEDV via des critères de position et de dispersion.

Table 2.2: MEDV summary

Min	Quantile 1	Median	Mean	Quantile 3	Max
5.00	17.02	21.20	22.53	25.00	50.00

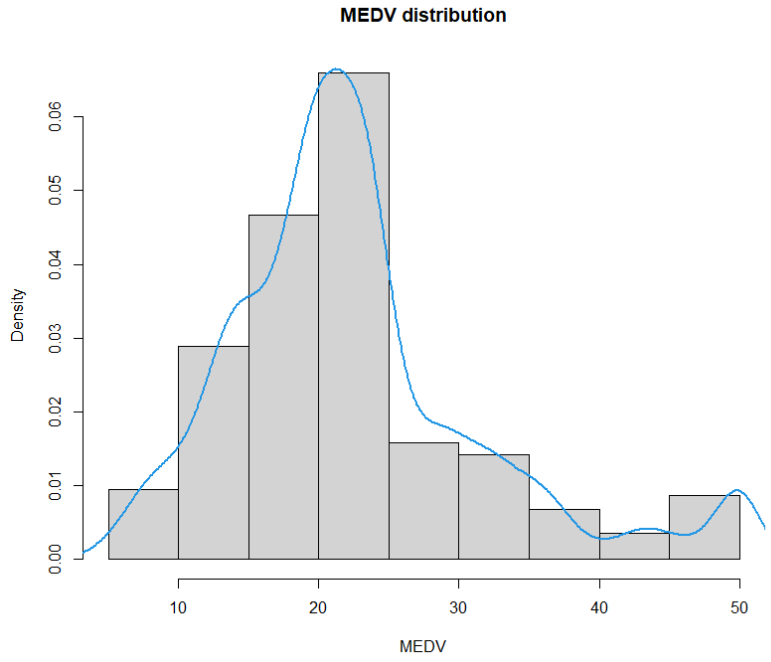
On remarque que la moyenne de la variable MEDV est de 22.53 (en $1000) tandis que la médiane est de 21.20 ce qui nous mène à penser que la queue de la distribution serait étalée vers la droite. Cela peut être du à la présence de valeurs aberrantes.

2.2.2 Skewness et Kurtosis

La skewness mesure l'asymétrie de la distribution de probabilité et par conséquent indique à quel point la distribution s'écarte de celle d'une loi normale tandis que le kurtosis est une mesure de l'aplatissement de la distribution.

$$skewness = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^3}{(N-1)s^3} \qquad kurtosis = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^4}{(N-1)s^4}$$

Dans notre base de données, les valeurs de la skewness et le kurtosis de la variable d'intérêt MEDV sont 1.104 et 4.468 respectivement. On peut conclure que la distribution de la variable MEDV est "right-skewed" et leptokurtique, comme le confirme l'histogramme ci-dessous.



2.2.3 Analyse de variance et covariance

2.2.3.1. Analyse de variance

Nous allons tout d'abord réaliser un test de variance qui est l'ANOVA. Nous utilisons ce test quand on cherche un lien entre une variable quantitative qui est la variable à expliquer y (ici il s'agit donc de MEDV) et une variable qualitative à k modalités de facteurs. Lorsque notre variable qualitative a deux niveaux de facteurs, on peut utiliser le test de Student. En effet, l'ANOVA n'est qu'une généralisation du test de Student à k > 2 niveaux de facteurs. On utilisera donc un test de Student pour chercher le lien entre la variable quantitative y MEDV et la variable désormais qualitative CHAS car elle n'a que deux niveaux de facteurs.
Nous utiliseront l'ANOVA pour chercher le lien entre la variable quantitative y MEDV et la variable désormais qualitative RAD car elle a plus de 2 niveaux de facteurs.

L'utilisation de l'ANOVA implique de réaliser les conditions d'application qui sont les suivantes :
- **Les échantillons sont aléatoires et indépendants** : cette condition est réalisée car aucun individu ne peut appartenir simultanément au groupe 0 dans lequel la maison ne borde pas la rivière et au groupe 1 qui représente le cas inverse.
- **Homoscédasticité** : à première vue nous souhaiterions utiliser le test de Bartlett pour mesurer l'homoscédasticité. L'utilisation de ce test nécessite que les données soient normalisées. Or, le test

de Shapiro-Wilk nous amène à rejeter l'hypothèse de normalité. Puisque la variable à expliquer ne suit pas une loi normale, nous avons testé l'homogénéité des variables à l'aide du test de Levene.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1  4.5903 0.03263 *
      504
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On observe une p_value = 0.03 ce qui nous conduit à rejeter l'hypothèse nulle. Afin de contourner le problème d'hétéroscédasticité, nous transformons la variable à expliquer en prenons son log.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.0173 0.8953
      504
```

Cette manipulation nous permet d'obtenir une p_value = 0.89 et ainsi valider l'hypothèse d'homoscédasticité. Nous pouvons désormais appliquer les tests statiques.

Test de Student

```
        Welch Two Sample t-test

data:  log(MEDV) by as.factor(CHAS)
t = -3.744, df = 39.733, p-value = 0.0005735
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.3925812 -0.1172888
sample estimates:
mean in group 0 mean in group 1
       3.016879        3.271814
```

Avec une p_value < 0.05, nous pouvons affirmer que le test est significatif. Cela signifie que la valeur médiane des logements qui bordent la rivière est significativement différent du prix moyen des autres logements. Il s'agit donc d'une variable significative pour nos prédictions.

ANOVA

```
             Df Sum Sq Mean Sq F value Pr(>F)
RAD           8  25.61   3.201   27.07 <2e-16 ***
Residuals   497  58.77   0.118
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec une p_value < 0.05, nous pouvons là aussi affirmer que le test est significatif, la variable RAD est donc significative.

2.2.3.2. Analyse de covariance

Dans le cas de l'analyse de covariance, nous observons deux variables quantitatives x et y et une variable qualitative T. La variable quantitative y est la variable que l'on cherche à expliquer en fonction de la variable quantitative x et du facteur T à J niveaux. C'est ce que l'on appelle l'ANCOVA à un facteur. Le principe de l'ANCOVA à deux facteurs est le même excepté le fait qu'on aura deux variables de regroupement pour une ou plusieurs variables continues appelées covariables.

ANCOVA à un facteur

Pour réaliser l'ANCOVA, nous avons séléctionné la variable RM qui représente le nombre moyen de pièces par logement en tant que covariable x, la variable CHAS en tant que facteur et la variable MEDV comme variable à expliquer y. Les résultats de l'analyse de covariance à un facteur montrent qu'après l'ajustement pour tenir compte du nombre moyen de pièces par logement, il y avait une différence statistiquement significative du prix du logement entre les deux groupes 0 et 1.

```
ANOVA Table (type II tests)

  Effect DFn DFd       F        p p<.05   ges
1     RM   1 503 464.613 1.78e-73     * 0.480
2   CHAS   1 503  12.584 4.25e-04     * 0.024
```

ANCOVA à deux facteurs

Pour réaliser l'ANCOVA à deux facteurs, nous avons conservé la variable RM en tant que covariable x, la variable CHAS en tant que premier facteur, la variable RAD en tant que second facteur et la variable MEDV comme variable à expliquer y.

```
ANOVA Table (type II tests)

    Effect DFn DFd       F        p p<.05   ges
1       RM   1 490 353.233 9.66e-60     * 0.419
2     CHAS   1 490  14.342 1.71e-04     * 0.028
3      RAD   8 490  11.394 7.28e-15     * 0.157
4 CHAS:RAD   5 490   4.663 3.61e-04     * 0.045
```

Après ajustement pour le nombre moyen de pièces, il y avait une interaction statistiquement significative entre CHAS et RAD sur le prix de la maison, avec un p très bas. Cela indique que l'effet de CHAS sur le prix du logement dépend de RAD, et vice-versa.

Chapter 3

Modélisation

3.1 Régression linéaire simple

3.1.1 Hypothèses classiques

H1 : Le modèle s'écrit $y_i = \alpha + \beta xi + u_i$ (absence d'erreur de spécification)
H2 : Les perturbations sont d'espérance nulle $E(u_i) = 0$
H3 : Les x ne sont pas aléatoires $E(x_i) = x_i$
H4 : La variance empirique de x est non nulle
H5a : Homoscédasticité des perturbations $V(u_i) = \sigma^2$
H5b : Non autocorrélation (linéaire) des perturbations $Cov(u_i, u_j) = 0 \quad \forall i \neq j$
H6 : Les perturbations suivent une loi normale $u_i \sim N(0, \sigma^2)$

3.1.2 Modèle linéaire simple

Pour cette section, nous avons décidé de faire une régression simple sur la variable RM car elle est assez corrélée à MEDV.

$$MEDV = \alpha + \beta_1 RM$$

Table 3.1: Simple Regression Summary

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-34.671	2.650	-13.08	<2e-16	***
RM	9.102	0.419	21.72	<2e-16	***

On remarque que le coefficient de RM est positif ce qui est normal car plus le nombre de chambres augmente, plus cher est le logement. Cependant, l'intercept est négatif et logiquement parlant,

chaque maison a au moins une chambre donc l'ordonnée à l'origine n'a pas vraiement de signification intrinsèque dans cette régression.

3.1.3 Intervalle de confiance

L'intervalle de confiance encadre la valeur que l'on cherche à estimer en définissant une marge d'erreur, ici 5%. Ci-dessous on voit que l'estimation de RM se situe entre 8.27 et 9.92, ce qui est effectivement le cas d'après les résultats obtenus dans la régression. L'intervalle de confiance de l'intercept par ailleurs est assez large compte tenu de la valeur de $\hat{\alpha}$, l'estimation de α est assez imprécise mais pas totalement (que des valeurs négatives dans l'IDC).

Table 3.2: IC à 95%

	2.5%	97.5%
(Intercept)	-39.876641	-29.464601
RM	8.278855	9.925363

3.2 Régression linéaire multiple

3.2.1 Hypothèses classiques

H1 : Le modèle s'écrit $y = X\beta + u$ (absence d'erreur de spécification)
H2 : u est d'espérance nulle $E(u) = 0$
H3 : Les régresseurs ne sont pas aléatoires $E(X) = X$
H5 : $V(u) = \sigma^2 I_n$ avec I_n une matrice d'identité de dimension (N,N) et σ^2 un scalaire.
H5a : Homoscédasticité des perturbations $V(u_i) = \sigma^2 \quad \forall i \neq j$
H5b : Non autocorrélation (linéaire) des perturbations $Cov(u_i, u_j) = 0 \quad \forall i \neq j$
H6 : Les perturbations suivent une loi normale d'espérance nulle et de variance σ^2, $u \sim N(0, \sigma^2 I_n)$

3.2.2 Modèle linéaire multiple

Dans un premier temps, nous allons décrire le modèle explicatif en utilisant toutes les variables puis nous allons procéder par la suite à des méthodes de sélection de variables afin d'améliorer notre modèle. L'équation de la régression linéaire multiple en fonction de toutes les varibles est :

$$\begin{aligned} MEDV = \alpha + \beta_1 CRIM + \beta_2 ZN + \beta_3 INDUS + \beta_4 CHAS + \beta_5 NOX + \beta_6 RM + \beta_7 AGE \\ + \beta_8 DIS + \beta_9 RAD + \beta_{10} TAX + \beta_{11} PTRATIO + \beta_{12} B + \beta_{13} LSTAT \end{aligned} \tag{3.1}$$

Le taux de crimes évalue la menace pour le bien-être que les ménages perçoivent dans divers quartiers de la région métropolitaine de Boston (en supposant que le taux de criminalité est proportionnel à la perception du danger par les habitants), cela devrait avoir un effet négatif sur la valeur des logements et par conséquent on s'attend à ce que le coefficient soit négatif. La variable CHAS capture les commodités d'un emplacement riverain, le coefficient devrait être positif. La variable NOX, quant à elle, mesure la pollution de l'air et on s'attend donc à un coefficient négatif vis-à-vis du prix du logement. RM et RAD représentent le nombre de pièces et la proximité du lieu du travail respectivement et devraient donc être positifs. On s'attend à ce que le coefficient de la variable DIS soit négatif étant donné qu'il s'agit de la distance entre le logement et le lieu du travail, de même pour LSTAT qui représente la proportion de la population de statut inférieur. Un faible PTRATIO (pupil-teacher ratio) pourrait impliquer que chaque élève reçoit plus d'attention individuelle, le coefficient pourrait ainsi être négatif.

3.2.3 Test de significativité

Table 3.3: Regression 1 Summary

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.646e+01	5.103e+00	7.144	3.28e-12	***
CRIM	-1.080e-01	3.286e-02	-3.287	0.001087	**
ZN	4.642e-02	1.373e-02	3.382	0.000778	***
INDUS	2.056e-02	6.150e-02	0.334	0.738288	
CHAS	2.687e+00	8.616e-01	3.118	0.001925	**
NOX	-1.777e+01	3.820e+00	-4.651	4.25e-06	***
RM	3.810e+00	4.179e-01	9.116	< 2e-16	***
AGE	6.922e-04	1.321e-02	0.052	0.958229	
DIS	-1.476e+00	1.995e-01	-7.398	6.01e-13	***
RAD	3.060e-01	6.635e-02	4.613	5.07e-06	***
TAX	-1.233e-02	3.760e-03	-3.280	0.001112	**
PTRATIO	-9.527e-01	1.308e-01	-7.283	1.31e-12	***
B	9.312e-03	2.686e-03	3.467	0.000573	***
LSTAT	-5.248e-01	5.072e-02	-10.347	< 2e-16	***

On remarque que les signes des coefficients sont tels qu'attendus. Toutes les variables semblent être statistiquement très significatives sauf INDUS et AGE.

3.2.4 Intervalle de confiance

A travers le tableau des intervalles de confiance ci-dessous, on remarque que les estimations obtenues précédemment tombent dans ces intervalles. Par contre, les intervalles sont plutôt larges et certaines estimations sont imprécises notamment celles de AGE et INDUS qui tombent dans des intervalles incluant 0 et donc avec des valeurs négatives et positives dans l'IDC.

Table 3.4: IC à 95%

	2.5%	97.5%
(Intercept)	26.432226009	46.486750761
CRIM	-0.172584412	-0.043438304
ZN	0.019448778	0.073392139
INDUS	-0.100267941	0.141385193
CHAS	0.993904193	4.379563446
NOX	-25.271633564	-10.261588893
RM	2.988726773	4.631003640
AGE	-0.025262320	0.026646769
DIS	-1.867454981	-1.083678710
RAD	0.175692169	0.436406789
TAX	-0.019723286	-0.004945902
PTRATIO	-1.209795296	-0.695699168
B	0.004034306	0.014589060
LSTAT	-0.624403622	-0.425113133

3.3 Sélection de modèle

3.3.1 Forward selection

On commence par chercher la variable qui explique le mieux MEDV, on se basera sur l'AIC mais il est également possible de comparer avec la p-value, le R^2 voire le C_p de Mallows. Puis on ajoute variable par variable jusqu'à ce que aucune variable restante n'améliore notre modèle, i.e. aucune variable n'a un AIC inférieur à celui obtenu dans l'étape d'avant. Suite à cette méthode de sélection ascendante, les variables INDUS et AGE n'ont pas été prises en compte par notre modèle. Précédemment, on a constaté que ces deux variables n'étaient pas significatives dans le modèle regroupant toutes les variables explicatives.

3.3.2 Backward selection

On part du modèle complet utilisant toutes les variables explicatives et on cherche celles qu'on pourrait supprimer afin d'obtenir un meilleur modèle. On se basera sur l'AIC des variables qu'on comparera à celui du modèle initial afin de décider de la suppression des variables. Suite à cette méthode de sélection descendante, on minimize l'AIC en supprimant les variables AGE et INDUS, on retrouve le même résultat qu'avec la selection Forward.

3.3.3 Stepwise selection

Une des méthode que nous pouvons utiliser pour choisir le meilleur modèle est la selection Subset, qui tente de choisir le meilleur modèle parmi tous les modèles possibles qui pourraient être construits avec l'ensemble de prédicteurs. Mais cette méthode peut être intense en termes de calcul. Pour un ensemble de p variables prédictives, il existe 2^p modèles possibles. Une alternative à la selection Subset est la sélection Stepwise, qui compare un ensemble beaucoup plus restreint de modèles. La table (3.5) représente les meilleurs modèles choisis selon le nombre de variables explicatives. Par exemple en prenant que deux variables explicatives, le meilleur modèle serait celui explqué par RM et LSTAT etc.

Table 3.5: Best Subsets Regression

Model Index	Predictors
1	LSTAT
2	RM LSTAT
3	RM PTRATIO LSTAT
4	RM DIS PTRATIO LSTAT
5	NOX RM DIS PTRATIO LSTAT
6	CHAS NOX RM DIS PTRATIO LSTAT
7	CHAS NOX RM DIS PTRATIO B LSTAT
8	ZN CHAS NOX RM DIS PTRATIO B LSTAT
9	CRIM CHAS NOX RM DIS RAD PTRATIO B LSTAT
10	CRIM ZN NOX RM DIS RAD TAX PTRATIO B LSTAT
11	CRIM ZN CHAS NOX RM DIS RAD TAX PTRATIO B LSTAT
12	CRIM ZN INDUS CHAS NOX RM DIS RAD TAX PTRATIO B LSTAT
13	CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT

La table (3.6) récapitule la comparaison des 13 modèles en utilisant trois critères différents. En retenant l'AIC on constate qu'effectivement le modèle 11, qui exclut les variables AGE et INDUS se caractérise par l'AIC le plus petit et donc comme vu avec les précédentes méthodes de sélection,

on retiendra le modèle suivant avec 11 variables explicatives :

$$\begin{aligned} MEDV = \alpha + \beta_1 CRIM + \beta_2 ZN + \beta_3 CHAS + \beta_4 NOX + \beta_5 RM + \beta_6 DIS \\ + \beta_7 RAD + \beta_8 TAX + \beta_9 PTRATIO + \beta_{10} B + \beta_{11} LSTAT \end{aligned} \tag{3.2}$$

Table 3.6: Subsets Regression Summary

Model	R^2	C(p)	AIC
1	0.5441	362.7530	3288.9750
2	0.6386	185.6474	3173.5423
3	0.6786	111.6489	3116.0973
4	0.6903	91.4853	3099.3590
5	0.7081	59.7536	3071.4386
6	0.7158	47.1754	3059.9390
7	0.7222	37.0589	3050.4384
8	0.7266	30.6240	3044.2750
9	0.7302	25.8659	3039.6381
10	0.7353	18.2049	3031.9965
11	0.7406	10.1145	3023.7264
12	0.7406	12.0027	3025.6114
13	0.7406	14.0000	3027.6086

3.4 Modèle explicatif

En effectuant un test de significativité sur le modèle retenu précédemment avec 11 variables explicatives, on obtient les résultats de la table (3.7).

On constate que toutes les variables sont significatives. Les variables CRIM, NOX, DIS, TAX, PTRATIO et LSTAT dépendent négativement de MEDV tandis que les autres variables en dépendent positivement. L'équation devient la suivante :

$$\begin{aligned} MEDV = 36.3411 - 0.108 CRIM + 0.0458 ZN + 2.7187 CHAS - 17.376 NOX \\ + 3.8 RM - 1.4927 DIS + 0.299 RAD - 0.0117 TAX \\ - 0.9465 PTRATIO + 0.0092 B - 0.5225 LSTAT \end{aligned} \tag{3.3}$$

L'intercept (ordonnée à l'origine), ici 36.34 (en $1000) peut être interprétée comme la valeur qu'on obtient lorsque toutes les variables sont égales à 0. Lorsque CRIM augmente de 1 point de pourcentage, MEDV baisse de 0.108. Par ailleurs, une hausse de 1 de la moyenne des chambres

Table 3.7: Regression 2 Summary

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	36.341145	5.067492	7.171	2.73e-12	***
CRIM	-0.108413	0.032779	-3.307	0.001010	**
ZN	0.045845	0.013523	3.390	0.000754	***
CHAS	2.718716	0.854240	3.183	0.001551	**
NOX	-17.376023	3.535243	-4.915	1.21e-06	***
RM	3.801579	0.406316	9.356	< 2e-16	***
DIS	-1.492711	0.185731	-8.037	6.84e-15	***
RAD	0.299608	0.063402	4.726	3.00e-06	***
TAX	-0.011778	0.003372	-3.493	0.000521	***
PTRATIO	-0.946525	0.129066	-7.334	9.24e-13	***
B	0.009291	0.002674	3.475	0.000557	***
LSTAT	-0.522553	0.047424	-11.019	< 2e-16	***

disponibles conduit à une hausse de la valeur de MEDV de 3.8, la présence de Charles River augmente MEDV de 2.7 et ainsi de suite. Quant à la qualité de l'ajustement, obtient un $R^2 = 0.7406$, cela signifie que le modèle explique 74% de la variance de MEDV ce qui est plutôt pas mal mais cela n'indique rien sur la qualité du modèle.

3.5 Modèle prédictif

Dans cette section, on prédit les valeurs de MEDV par une régression linéaire multiple en fonction de quelques variables choisies selon certains critères tels que la corrélation et la significativité. En effet, les variables RAD et TAX étant très corrélées (à 0.9), on a décidé d'omettre la variable RAD durant cette procédure ainsi que INDUS et AGE jugées non significatives à posteriori. Après avoir divisé les données en training et test sets et effectué les prédictions sur le test, on remarque que les valeurs prédites sont proches des vraies valeurs, les prédictions sont bonnes mais pas parfaites. On obtient les résultats suivants sur les 5 premières observations du test :

Predicted	31.18	23	19.86	13.17	16.79	16.35
True	34.7	22.9	20.4	13.6	15.2	15.6

Chapter 4

Conclusion

La méthode de régression linéaire multiple effectuée lors de cette analyse nous a permi d'étudier et expliquer la relation entre la variable MEDV et plusieurs facteurs. La variable AGE s'est avérée positive et statistiquement non significative, ceci est probablement dû au fait que l'âge ne soit pas significativement corrélé à la médiane des prix des logements étant donné qu'on pourrait bien avoir des anciens biens à Boston de très bonne qualité. Nous avons également effectué des prédictions sur la valeur médiane des logements qui se sont avérées plutôt bonnes.

Néanmoins, il serait intéressant d'essayer d'autres modèles non linéaires tels qu'un modèle log-lin, log-log voire des fonctions puissance (X^p) et évaluer les effets marginaux. Cependant, il serait préférable d'estimer l'exposant en effectuant un grid search sur des valeurs alternatives de p.

Pour aller plus loin, on pourrait essayer d'utiliser des méthodes de machine learning plus sophistiquées par exemple en ajoutant des termes de régularisation tels que Ridge et Lasso tout en ajustant les hyperparamètres, ou encore des modèles de régression comme les arbres de décision ou les Support Vector Regressors. L'utilisation de ces différentes méthodes linéaires et polynomiales nous permettra de comparer différents modèles plus avancés via leur performance grâce à une ou plusieurs métriques comme la racine de l'erreur quadratique moyenne (RMSE) par exemple.

Table 4.1: Variables used in the Boston Housing analysis

Variable	Definition	Source
CRIM	Per capita crime rate by town	FBI (1970)
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft	Metropolitan Area Planning Commission (1972)
INDUS	Proportion of non-retail business acres per town	Vogt, Ivers and Associates
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)	1970 US Census Tract maps
NOX	Nitric oxides concentration (parts per 10 million)	TASSIM
RM	Average number of rooms per dwelling	1970 US Census
AGE	Proportion of owner-occupied units built prior to 1940	1970 US Census
DIS	Weighted distances to five Boston employment centers	Schnare
RAD	Index of accessibility to radial highways	MIT Boston Project
TAX	Full-value property-tax rate per $10,000	Massachusetts Tax-payers Foundation (1970)
PTRATIO	Pupil-teacher ratio by town 12	Massachusetts Dept. of Education (1971-1972)
B	B : $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town 13	1970 US Census
LSTAT	% lower status of the population	1970 US Census
MEDV	Median value of owner-occupied homes in $1000s	1970 US Census