Projet : Méthodes de traitement de données manquantes

Olga Silva / Marlène Chevalier

30/11/2019

Sujet : Valeur du logement en banlieue de Boston

Il s'agit de traiter les données manquantes du fichier Boston Housing. Ce fichier décrit la situation des logements dans les villes de la banlieue de Boston. Il est constitué de 506 enregistrements et de 14 variables quantitatives (soit 7084 données) :

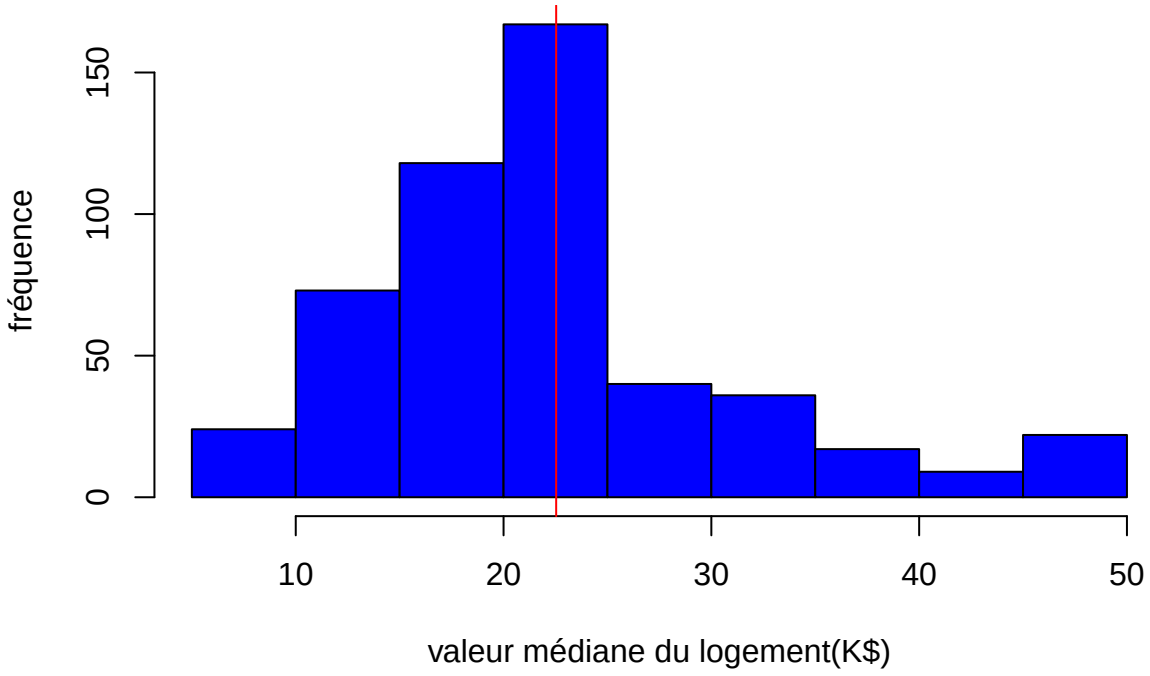
- **CRIM** : Taux de criminalité par habitant.
- **ZN** : Proportion de terrains résidentiels pour des lots de plus de 25000 pieds carré (environ 2300m^2).
- **INDUS** : Proportion d'espace, en acres, consacré aux affaires non commerciales (1 acre environ 4000m^2).
- **CHAS** : Proximité avec la rivière Charles (=1 si en bord de rivière / =0 si éloigné de la rivière)
- **NOX** : Concentration en oxyde d'azote (1 pour 10 millions)
- **RM** : Nombre moyen de chambres par logement
- **AGE** : Proportion des propriétés construites avant 1940
- **DIS** : Moyenne des distances aux 5 centres d'emploi de Boston
- **RAD** : Indice d'accessibilité aux autoroutes (de 1 à 8 et 24)
- **TAX** : Taux d'imposition foncier (1 pour 10 000$)
- **PTRATIO** : Ratio d'élèves-enseignants
- **B** : Proportion de population afro-américaine
- **LSTAT** : Proportion de population précaire
- **MDEV** : Valeur médiane des habitations privées (en K$)

Nous utiliserons ces données pour tenter d'expliquer la valeur médiane des habitations privées (MDEV) en fonction des autres variables du fichier.

1.Exploration des données incomplètes

Graphiques sur les données incomplètes

La valeur médiane du logement à Boston est comprise entre 5 K$ et 50 K$, en moyenne de 22.5 K$.



Graphiquement (*cf.annexe 1.1*), on observe que :

- Le **taux de criminalité** faible (inféreur à 10%) est le plus fréquent. La valeur du logement a tendance a diminué lorsque le taux de criminalité augmente. Mais la corrélation entre les 2 reste faible (0.4).
- La **proportion de terrains résidentiels** est majoritairement faible (inféreur à 10%), mais lorsqu'elle augmente, la valeur du logement a tendance a augmenté.
- La **proportion de surface d'activité industrielle** la plus fréquente est entre 18 et 20 acres (entre 72000m^2 et 80000m^2). A ce niveau, la valeur moyenne est bien souvent inférieure à la valeur médiane moyenne des logements (22.5K$).
- La **concentration d'oxyde d'*azote** est le plus souvent entre 0.4 et 0.6. Plus la concentration augmente, plus la valeur des logements diminue.
- Le **nombre moyen de chambre** est le plus souvent entre 5 et 7. Plus le nombre de chambre augmente, plus les logements ont de la valeur.
- La **proportion de propriétés construites avant 1940** est très importante (majoritairement autour de 90%). Plus cette proportion augmente, plus les logements ont de la valeur.
- La **moyenne des distances aux centres d'emploi** est fréquemment faible (<4). Cette variable influence peu la valeur des logements (corrélation=0.28).
- L'**imposition foncière** prend la plus importante partie de ses valeurs entre 200 et 500. Puis une autre serie importante de ses valeurs est autour de 666 ; à ce niveau d'impôt, la valeur du logement est plus faible (<moyenne 22.5).

- Le **ratio élèves-enseignants** est réparti quasi-équitablement autour de la valeur moyenne du logement.Une hausse de ce ratio a tendance à faire baisser le prix du logenment (corrélation =-0.54)
- La **proportion de population afro-américaine** est importante; mais son influence n'est pas significative sur la valeur du logement (cor =0.35)
- La **proportion de population précaire** influence négativement la valeur des logements (cor =-0.74).
- L'**indice d'accessibilité aux autoroutes** à 24 donne les valeurs des logements les plus basses (15K$ en moyenne) . Les indices 3 et 8 donnent les valeurs de logement les plus élévées.

- La **proximité avec la rivière Charles** augmente légèrement la valeur du logement.

Corrélation des variables

Les corrélations les plus significatives de la valeur du logement sont avec :

- RM (0.72) : plus le nombre de chambre est important, plus la valeur du biens est forte.

- INDUS et RAD (-0.51) : plus l'espace d'affaires non commerciales ou plus l'indice d'accessibilité aux autoroutes seront importants, moins la valeur du bien sera élevée.

- PTRATIO (-0.54) : plus le ratio elèves-enseignants est fort, moins la valeur des biens est élevée.

- LSTAT (-0.74) : une forte proportion de population précaire réduira très fortement la valeur des biens. (*cf. annexe 1.2*)

Modélisation sur les données incomplètes

Nous commençons par examiner les résultats d'une regression linéaire de MEDV sur les autres variables. Le modèle est correctement ajusté (R^2=0.75) mais il semble que les variables explicatives INDUS et AGE ne soient pas pertinentes pour ce modèle. (*cf. annexe 1.3*)

Nous voudrions maintenant utiliser la méthode **stepwise** pour choisir le meilleur modèle. Cependant, cette méthode ne fonctionne pas s'il y a des données manquantes. Afin de pouvoir l'utiliser, nous allons appliquer une méthode "listwise deletion" avec le risque de perte d'information et d'introduction de biais au dataset. Ce modèle écarte les variables INDUS et AGE, comme l'original, avec un R2 ajusté très proche de l'original. (*cf. annexe 1.4*)

En conclusion sur l'exploration de données

La valeur médiane du logement à Boston et ses environs est comprise entre 5K$ et 50K.$La distribution est croissante jusqu'a sava$
puis décroit fortement à partir de 25 K$.

La valeur du logement est influencée (p_value <5%):

- positivement principalement (coefficient estimé = 4.2) **par le nombre de chambres** et plus faiblement (coefficients estimés de 0.28 et 0.05) **par l'accessibilité aux autoroutes et la proportion de terrain résidentiel**

- négativement principalement (coefficient estimé = -18.5) **par la concentration en oxyde d'azote** et plus faiblement (coefficients estimés entre de -1.42 et -0.01) **par la distance aux centres d'emplois, le ration élèves-enseignants, la proportion de population précaire et le taux foncier**
Le R^2 du modèle de regression sur jeu de données incomplet est de **0.76 (R^2 de référence)**.
(**cf. annexe 1.4** : résultat de la regression linéaire après selection de variable)

Les variables INDUS et AGE ne sont pas significatives dans l'explication de la valeur du logement (p-value>5% **cf. annexe 1.3**).

Nous allons comparer maintenant ce modèle obtenu avec des données manquantes et les nouveaux modèles que nous allons obtenir avec des datasets complets à partir de differentes méthodes.

2.Inventaire des données manquantes

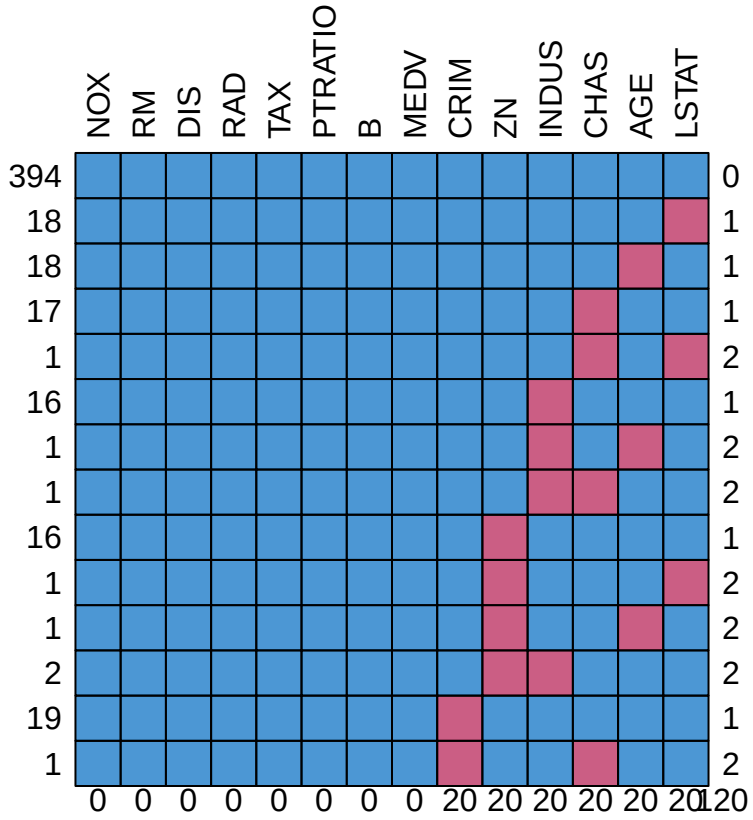
Il s'agit ici d'identifier les données manquantes par variable, représenter leur structure dans le jeu de données.

Structure des données manquantes

La fonction **md.pattern** du package MICE a pour résultat une matrice, dans laquelle chaque ligne correspond à des structures de données manquantes et chaque colonne à une variable du fichier. Les lignes et les colonnes sont triées selon le niveau de complétude des données.
A chaque ligne de la matrice (qui définit une structure de données manquantes du jeu de données) :

- la première colonne indique le nombre d'observations correspondant à la structure de données manquantes décrite ;

- la derniere colonne donne le nombre de variables incomplètes.

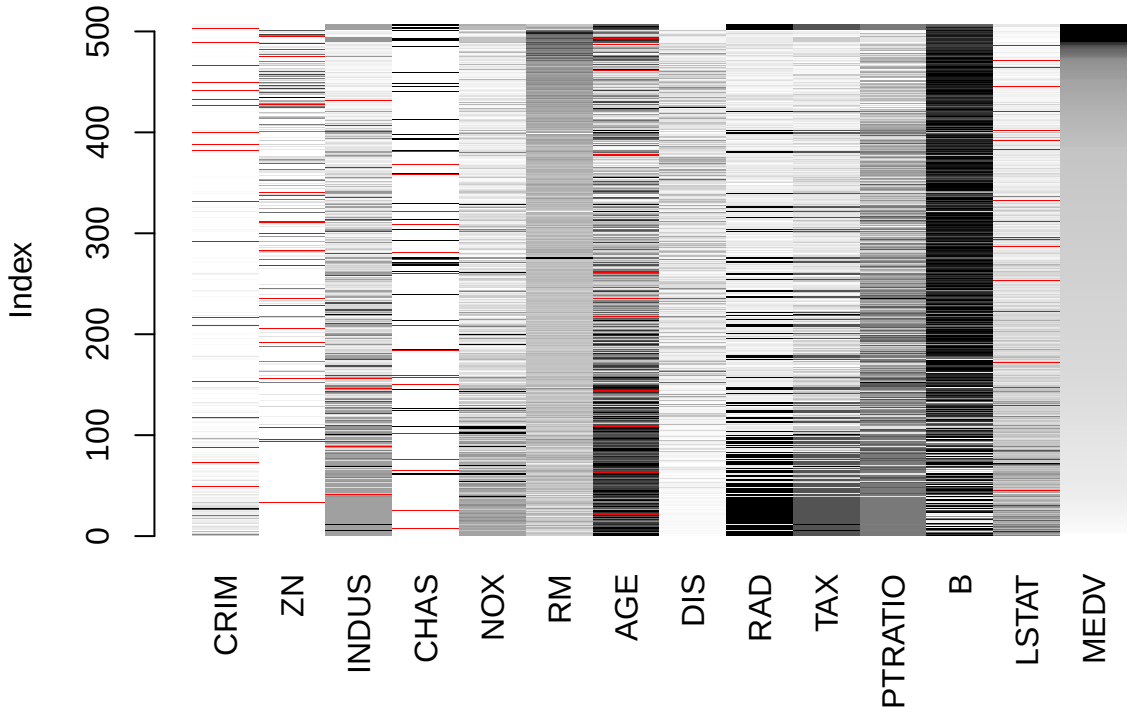


Au total, 120 observations sont manquantes : 20 pour chacune des variables CRIM, ZN, INDUS, CHAS, AGE, LSTAT.

- structure 1 : 394 observations pour lesquelles aucune donnée n'est manquante,

- structure 2 : 18 observations pour lesquelles seule la donnée de la variable LSTAT est absente,
- structure 3 : 18 observations pour lesquelles seule la donnée de la variable AGE est absente,
  ...
- structure 14 (dernière) : 1 observation pour laquelle les données des variables CRIM et CHAS sont absentes.

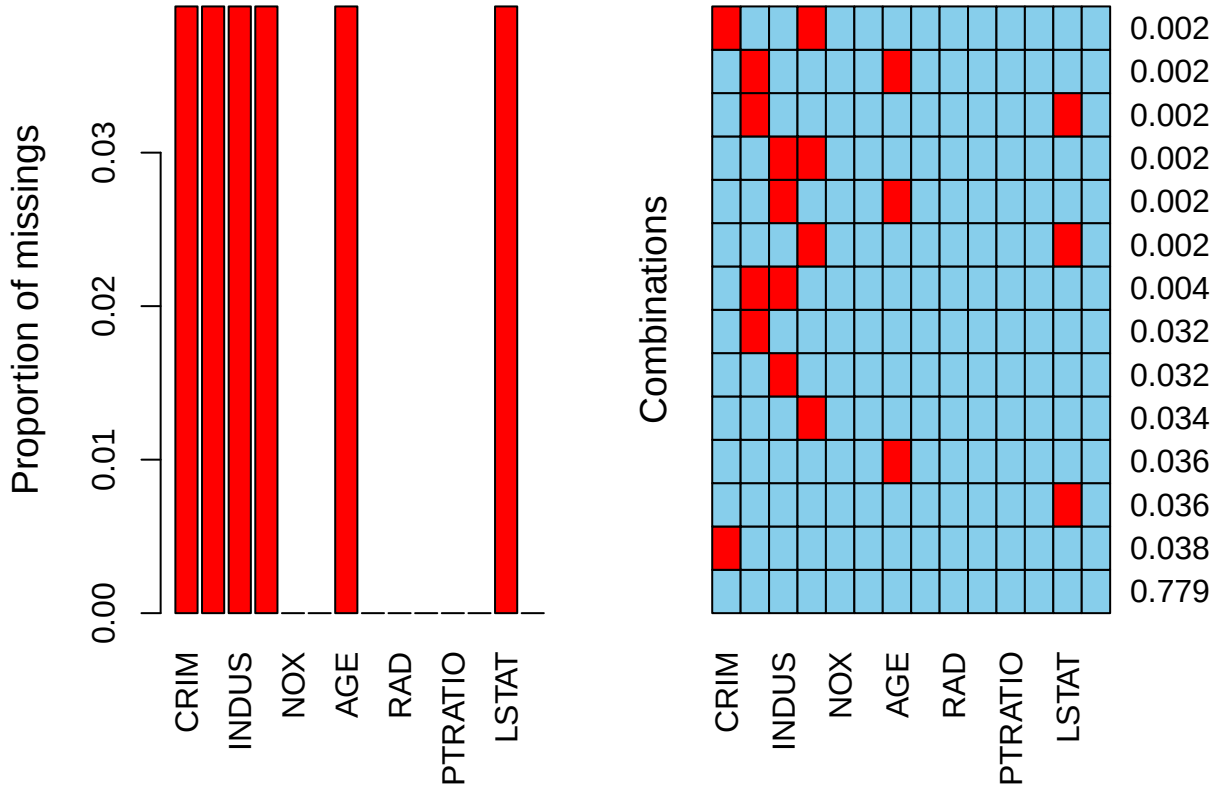
Voici une deuxième representation des données manquantes obtenue avec **matrixplot**. Elle permet d'identifier des dépendences des valeurs extrêmes et des données manquantes. Les données observées sont en gris et les manquantes en rouge.



Nous pouvons observer que les valeurs manquantes se repartissent bien dans l'ensemble du jeu de données.

Proportion de données manquantes

La fonction **aggr** permet d'appréhender les données par leur proportion dans le jeu complet.



En sortie, 2 graphiques :

- Le graphique de gauche donne la proportion de données maquantes de chaque variable : ici on retrouve des proportions égales pour les variables CRIM, ZN, INDUS, CHAS, LSTAT (autour de 4%), les autres variables sont complètes.

- Le graphique de droite donne la proportion de chaque structure de données.

Ici 78% du jeux de données est complet pour toutes les variables, 3.8% des individus ont uniquement la variable CRIM qui n'est pas renseignée, ...

Catégories de données manquantes

Rubin (1976) a classé les problèmes des données manquantes en trois catégories :

- **Données manquantes de façon complètement aléatoire : MCAR** (missing completely at random). L'absence de données est dûe au hasard, à la malchance. Cette hypothèse est peu réaliste.

- **Données manquantes de façon aléatoire : MAR** (missing at random). La probabilité d'absence de la valeur d'une variable dépend des valeurs prises par d'autres variables qui ont été observées. MAR est plus générale et plus réaliste que MCAR.
- **Données manquantes de façon non aléatoire : MNAR** (missing not at random). La cause d'absence de la valeur d'une variable est de raision inconnue. MNAR est le cas le plus complexe.

La plupart des méthodes modernes de traitement des données manquantes partent de la supposition MAR. Dans le cas du jeu de données Boston Housing, on peut aussi partir de cette hypothèse.

En conclusion sur l'inventaire des données manquantes

- 6 variables sont incomplètes, avec chacune 20 données manquantes (soit 120 au total) :
  - **CRIM** : Taux de criminalité par habitant
  - **ZN** : Proportion de terrains résidentiels
  - **INDUS** : Proportion d'espace consacré aux affaires non commerciales
  - **CHAS** : Proximité avec la rivière Charles
  - **AGE** : Proportion des propriétés construites avant 1940
  - **LSTAT** : Proportion de population précaire

Sur le jeu de données, 22% des individus sont incomplets

- Nous supposons qu'on est en situation **MAR (Données manquantes de façon aléatoire)**

3.Traitement des données manquantes

Imputation multiple

L'imputation multiple va créer m datasets complets, au lieu d'imputer une seule fois comme l'imputation simple.

Les méthodes d'imputation simples consistent à remplacer chaque valeur maquante par une valeur unique prédite ou simulée. Plusieurs solutions sont possibles (remplacer les données manquantes par la moyenne de la variable, faire une régression avec les données observées...). Ces solutions rapides sont à éviter car cela modifie la corrélation entre les variables, la distribution des variables, la variable peut-être sous-estimée, entre autres problèmes. Nous allons utiliser uniquement l'imputation multiple pour notre projet.

Tous les méthodes d'imputation multiple suivent les mêmes trois étapes :

Etape 1 : Imputation des données manquantes m fois
Etape 2 : Analyse de m datasets imputés
Etape 3 : Mise en commun des paramètres à travers m analyses

Plus de détail sur la méthodologie se trouve sur l'annexe 3.1

Nous allons ici en tester plusieurs méthodes d'imputation multiple : l'imputation par regression stochastique, les forêts aléatoires, predictive mean matching et ACP. Nous allons appliquer ces méthodes à des modèles sans INDUS et AGE, car elles ne semblent pas pertinentes, selon les résultats des analyses précèdentes.

Imputation par régression stochastique

Cette méthode consiste à imputer les données manquantes en utilisant la regression à laquelle on a ajouté du bruit. Cela permet de corriger le biais de corrélation qui existe par les méthodes plus rapides d'imputation simple. Nous fixons m=6, en suivant les conseils de Stef Van Buuren dans son livre "flexible imputation data"

Pour faire cette imputation, nous utilisons la fonction **mice (avec method = "norm.nob")**

```
##
## Call:
## lm(formula = MEDV ~ . - INDUS - AGE, data = dHBcompl.regsto)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5488  -2.7274  -0.4758   1.6642  26.2735
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.478048   5.063326   7.204 2.19e-12 ***
## CRIM         -0.114737   0.032401  -3.541 0.000436 ***
## ZN            0.049578   0.013452   3.686 0.000253 ***
## CHAS          2.879064   0.847209   3.398 0.000733 ***
## NOX         -17.060759   3.519179  -4.848 1.67e-06 ***
## RM            3.747276   0.406548   9.217  < 2e-16 ***
## DIS          -1.518824   0.184837  -8.217 1.84e-15 ***
## RAD           0.297555   0.063465   4.689 3.57e-06 ***
## TAX          -0.012166   0.003376  -3.604 0.000345 ***
## PTRATIO      -0.928884   0.129353  -7.181 2.56e-12 ***
## B             0.009037   0.002672   3.383 0.000775 ***
## LSTAT        -0.519437   0.046889 -11.078  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.726 on 494 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.736
## F-statistic:   129 on 11 and 494 DF,  p-value: < 2.2e-16
```

Avec cette méthode, on observe un R2 ajusté de 0.736 pour la régression linéaire faite avec le dataset completé avec la régression stochastique. C'est une R2 ajusté proche de celui obtenu avec la régression faite avec les données manquantes.

La description de la nouvelle structure des données se trouve sur l'annexe 3.2

Imputation par forêts aléatoires

Avec cette méthode, pour les variables continues les valeurs sont imputées en faisant des tirages aléatoires à partir des distributions gaussiennes independantes, centrées en les moyennes prédites par les forêts aléatoires. Pour faire cette imputation, nous utilisons la fonction **mice (avec method = "rf")**, avec m=6

```
##
## Call:
## lm(formula = MEDV ~ . - INDUS - AGE, data = dHBcompl.rf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6843  -2.8094  -0.5474   1.6836  26.6077
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.838360   5.066699   7.073 5.20e-12 ***
## CRIM         -0.116906   0.032291  -3.620 0.000325 ***
## ZN            0.044706   0.013423   3.331 0.000931 ***
## CHAS          2.857365   0.864289   3.306 0.001015 **
## NOX         -17.991771   3.522349  -5.108 4.66e-07 ***
## RM            3.878241   0.402788   9.629  < 2e-16 ***
## DIS          -1.462440   0.183844  -7.955 1.23e-14 ***
```

```
## RAD           0.298160   0.063479   4.697 3.43e-06 ***
## TAX          -0.011473   0.003374  -3.401 0.000727 ***
## PTRATIO      -0.952459   0.129410  -7.360 7.73e-13 ***
## B             0.009322   0.002680   3.479 0.000548 ***
## LSTAT        -0.497699   0.045467 -10.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.741 on 494 degrees of freedom
## Multiple R-squared:   0.74,  Adjusted R-squared:  0.7342
## F-statistic: 127.8 on 11 and 494 DF,  p-value: < 2.2e-16
```

Avec cette méthode on observe un R2 ajusté de 0.734 avec le dataset completé, presque le même obtenu avec la méthode d'imputation stochastique. Les valeurs estimées et les résidus restent aussi très proches.

La description de la nouvelle structure des données se trouve sur l'annexe 3.3

Imputation par predictive mean matching

La méthode pmm est très répandue car elle est facile à utiliser, elle sélectionne un sous ensemble à partir des cas complets, avec des valeurs prédites proches des valeurs prédites par les données manquantes. Un élement du sous ensemble est pris aléatoirement et la valeur observée remplace la donnée manquante.

Pour faire cette imputation, nous utilisons la fonction **mice (avec method = "pmm")**, m=6

```
##
## Call:
## lm(formula = MEDV ~ . - INDUS - AGE, data = dHBcompl.pmm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7781  -2.7927  -0.5505   1.6506  26.5719
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.632329   5.087189   7.004 8.16e-12 ***
## CRIM         -0.121857   0.032712  -3.725 0.000218 ***
## ZN            0.044661   0.013483   3.312 0.000993 ***
## CHAS          2.759776   0.859184   3.212 0.001404 **
## NOX         -17.752525   3.547748  -5.004 7.83e-07 ***
## RM            3.917625   0.406579   9.636  < 2e-16 ***
## DIS          -1.491754   0.188168  -7.928 1.50e-14 ***
## RAD           0.306176   0.063926   4.790 2.21e-06 ***
## TAX          -0.011778   0.003406  -3.458 0.000590 ***
## PTRATIO      -0.957221   0.129931  -7.367 7.37e-13 ***
## B             0.009503   0.002689   3.534 0.000447 ***
## LSTAT        -0.498575   0.047173 -10.569  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.77 on 494 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7311
## F-statistic: 125.8 on 11 and 494 DF,  p-value: < 2.2e-16
```

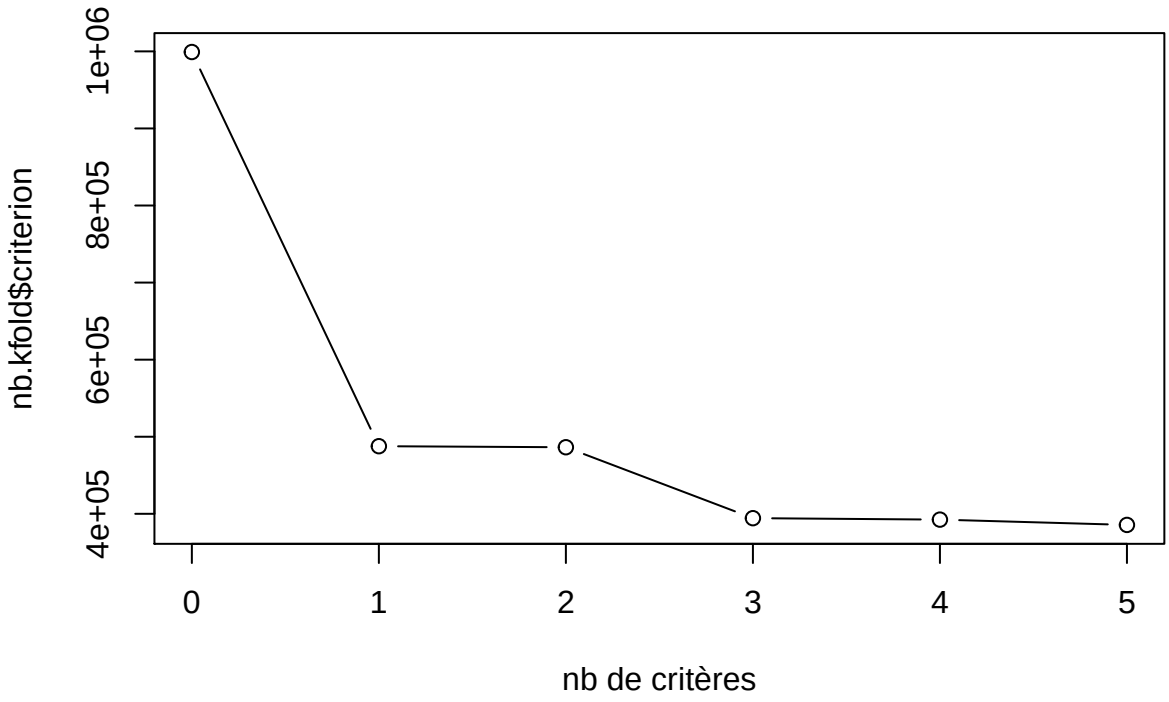
Pareil que pour les deux méthodes précédentes, le R2 ajusté est très proche du modèle avec des données manquantes. Les valeurs estimées et les résidus restent aussi très proches.

La description de la nouvelle structure des données se trouve sur l'annexe 3.4 et on observe que les valeurs min, max et médianne sont presque idéntiques à ceux obtenus avec les forêts aléatoires.

Imputation par ACP

Les données manquantes sont imputées en utilisant la ACP (Analyse des composants principales). Il s'agit :

- d'estimer le nombre de dimensions utilisées dans la formule de reconstruction avec la fonction **estim_ncpPCA(dHB,method.cv = "Kfold")**



Selon le graphique, le nombre de dimensions sera 3.

- générer les ensembles de données imputées avec la fonction MIPCA en utilisant le nombre de dimensions précédemment calculé
  fonction **MIPCA (avec method.mi = "Bayes")**

Regression linéaire sur le jeu de données complété par MIPCA

```
##                 estimate   std.error  statistic       df      p.value
## (Intercept)  35.978091089 5.165296996   6.965348 480.2792 1.083489e-11
## CRIM         -0.110818363 0.033373714  -3.320528 465.1297 9.691645e-04
## ZN            0.045105412 0.013851586   3.256335 460.7435 1.211848e-03
## CHAS          2.747853787 0.876961032   3.133382 461.9290 1.837967e-03
## NOX         -17.540270502 3.571025567  -4.911830 486.1882 1.234641e-06
## RM            3.842770936 0.414449969   9.271978 476.4948 0.000000e+00
## DIS          -1.470835496 0.186385108  -7.891379 481.3189 2.042810e-14
## RAD           0.293051782 0.064114273   4.570773 486.4481 6.169649e-06
## TAX          -0.011664569 0.003402662  -3.428072 488.2776 6.592967e-04
```

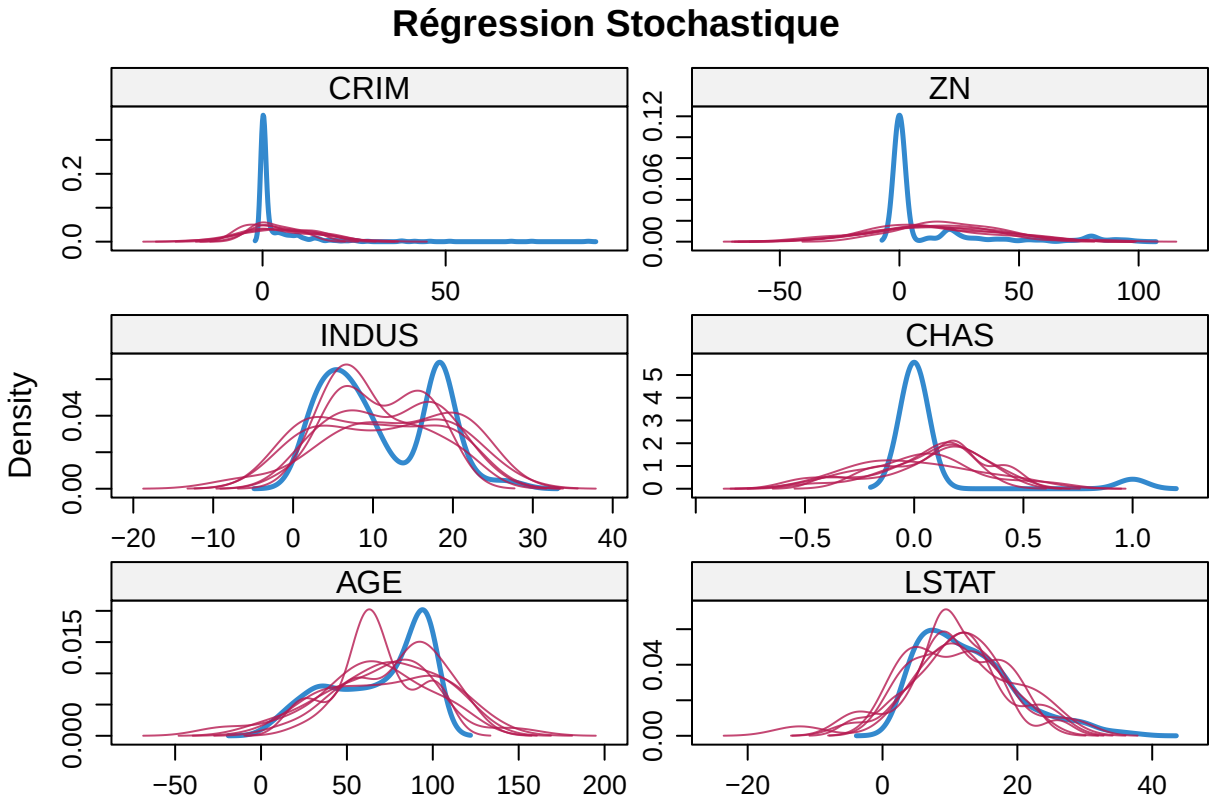
```
## PTRATIO      -0.944508946 0.131522172  -7.181367 483.8169 2.620126e-12
## B             0.009065249 0.002707681   3.347975 488.6123 8.768891e-04
## LSTAT        -0.507962286 0.048531983 -10.466547 456.8212 0.000000e+00
```

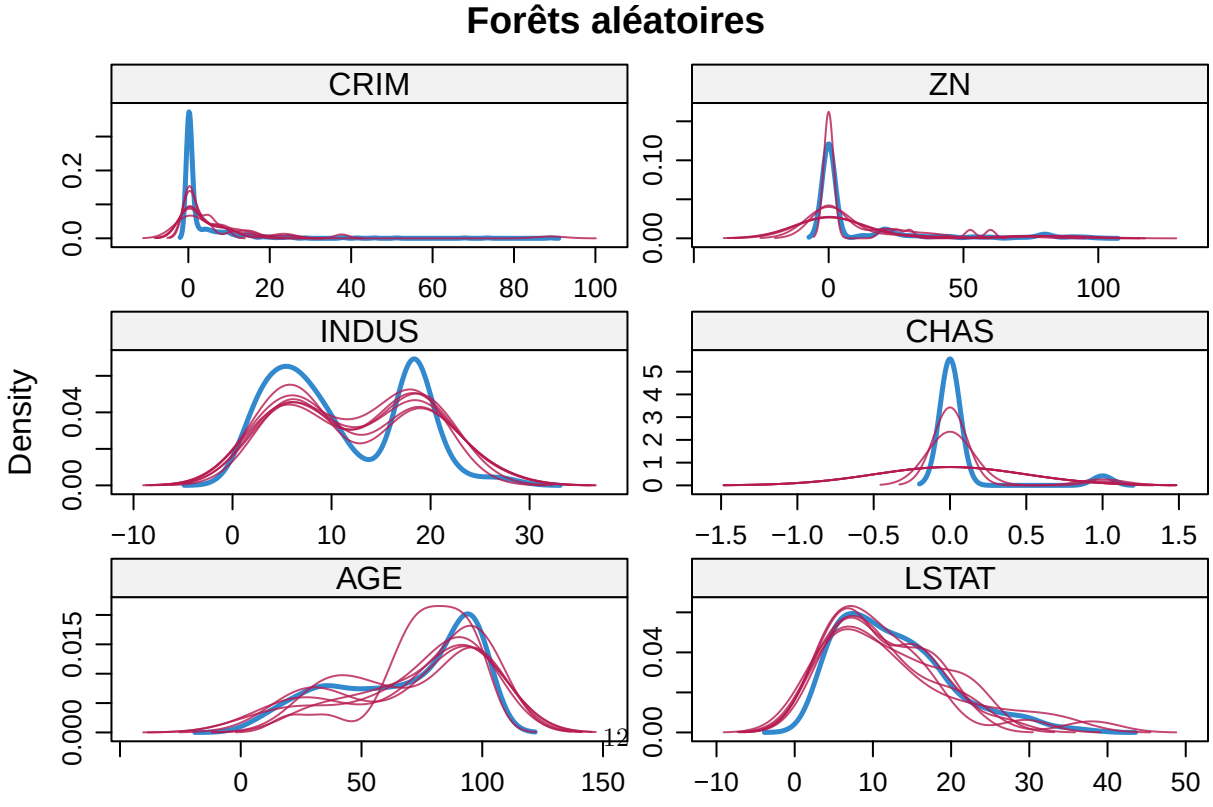
Les valeurs des estimateurs sont très proches aux celles trouvées par les autres méthodes d'imputation multiple.

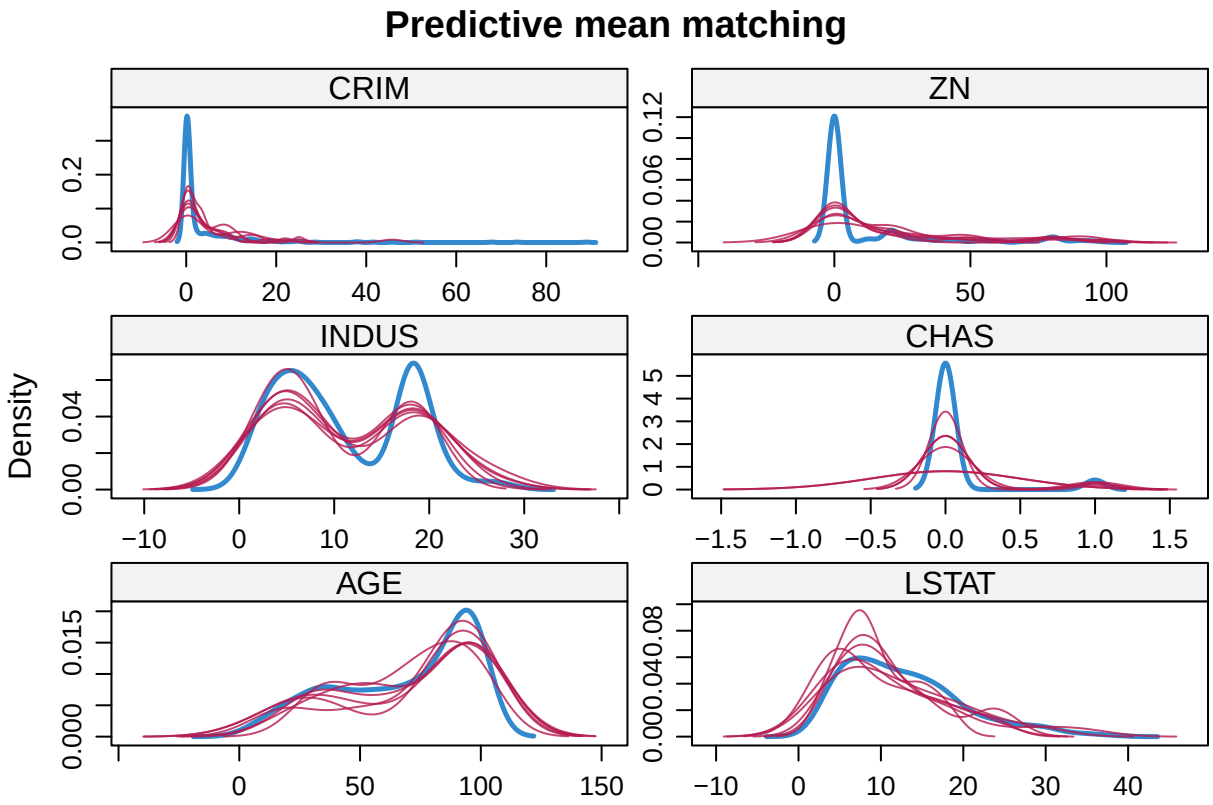
Nous avons testé 4 méthodes d'imputation, qui nous ont permis de faire des régressions avec des datasets complets. Ces régressions ont des résultats très proches en termes de R2 ajustée, des estimateurs et des résidus. Nous allons faire des diagnostics pour choisir le meilleur modèle d'imputation.

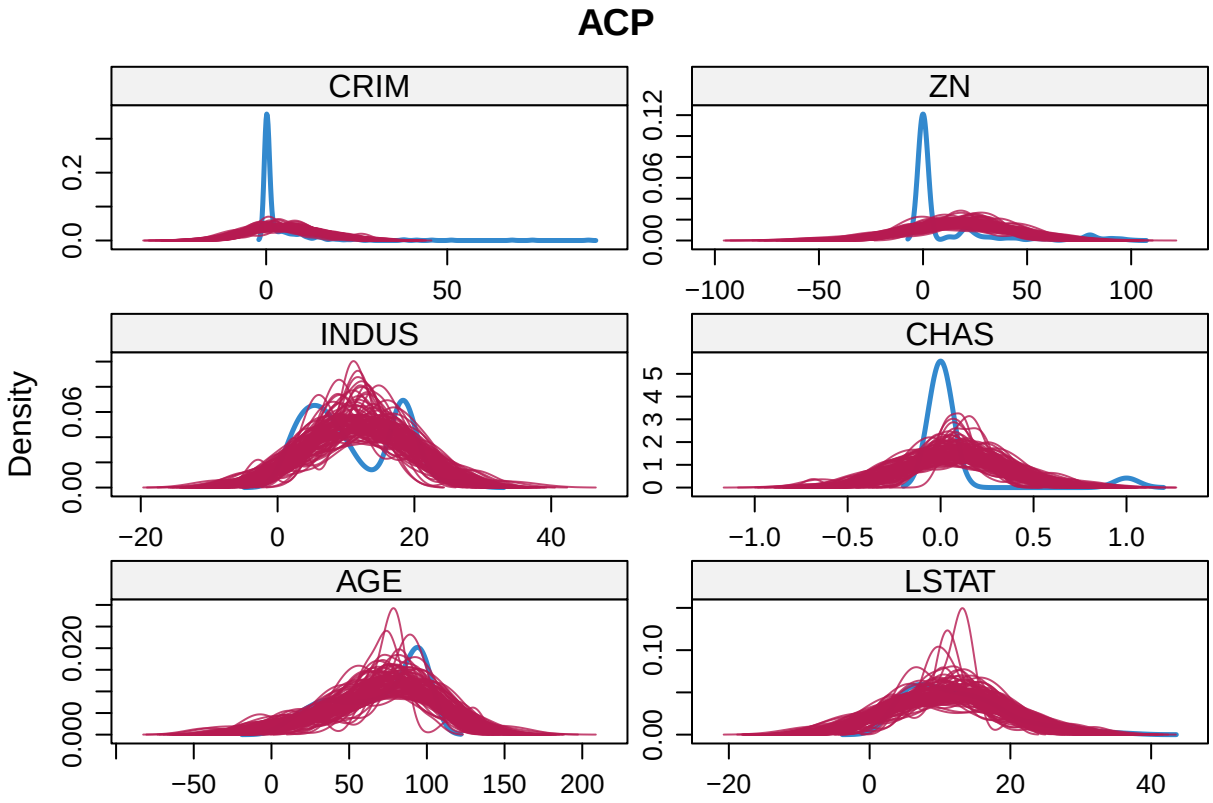
4. Diagnostics et conclusion

Diagnostic 1 : vérifier que la distribution des données imputées est similaire à celle des données d'origine.






Predictive mean matching
CRIM
ZN
INDUS
CHAS
AGE
LSTAT
Density



Pour la variable CRIM, aucune méthode ne semble parvenir à réproduire la même distribution. Par contre pour INDUS, AGE et LSTAT, tous les méthodes semblent réussir à le faire. Pour CHAS et ZN, les forêts aléatoires sont les plus proches de la vraie distribution.

Diagnostic 2 : vérifier la convergence des algorithmes.

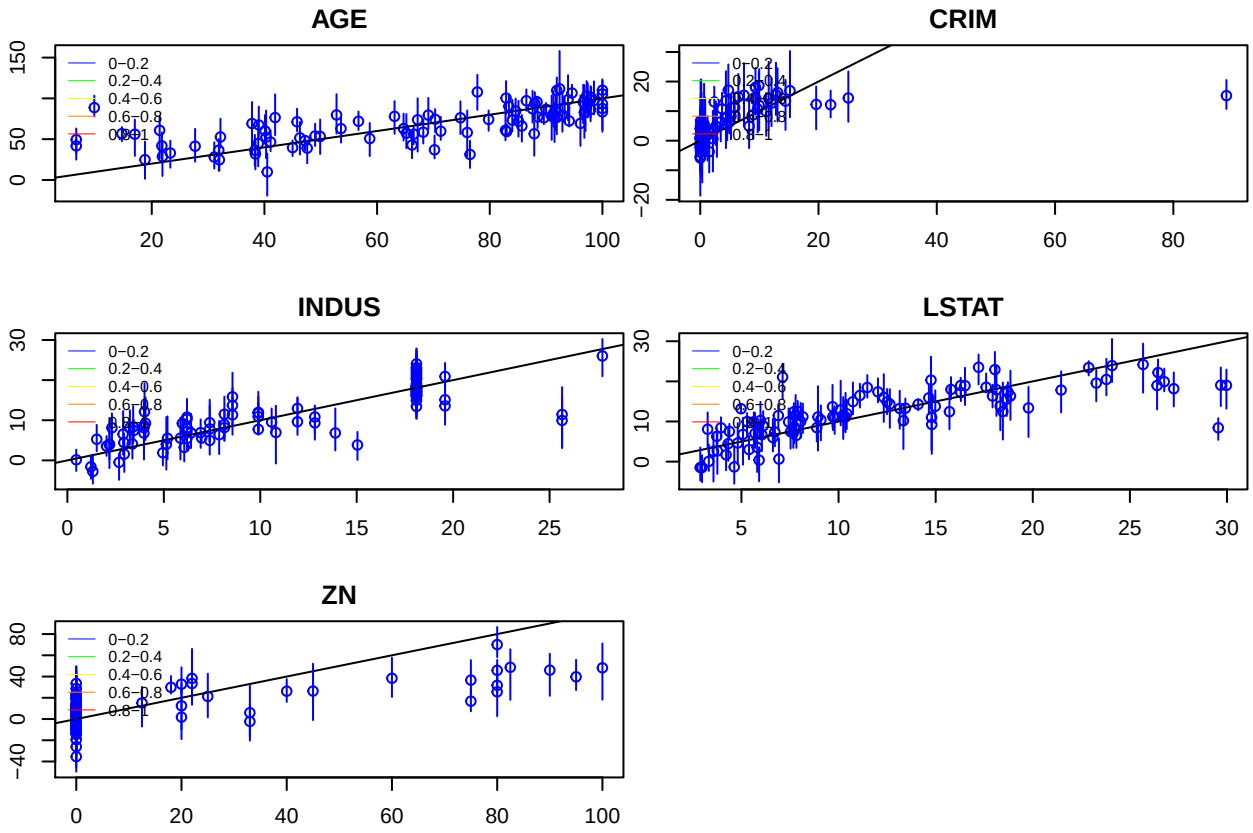
La verification de la convergence se fait à partir des graphiques de variations de la moyenne et de l'écart type, pour chaque méthode, pour chaque itération et chaque donnée imputée.Pour que la convergence soit vérifiée, il faut que les différentes courbes se mélangent librement, sans une tendance particulière. C'est bien le cas pour nos graphiques; donc, nous n'avons pas de problème de convergence. (**cf.annexe 4.1**)

Diagnostic 3 : vérifier l'ajustement du modèle d'imputation

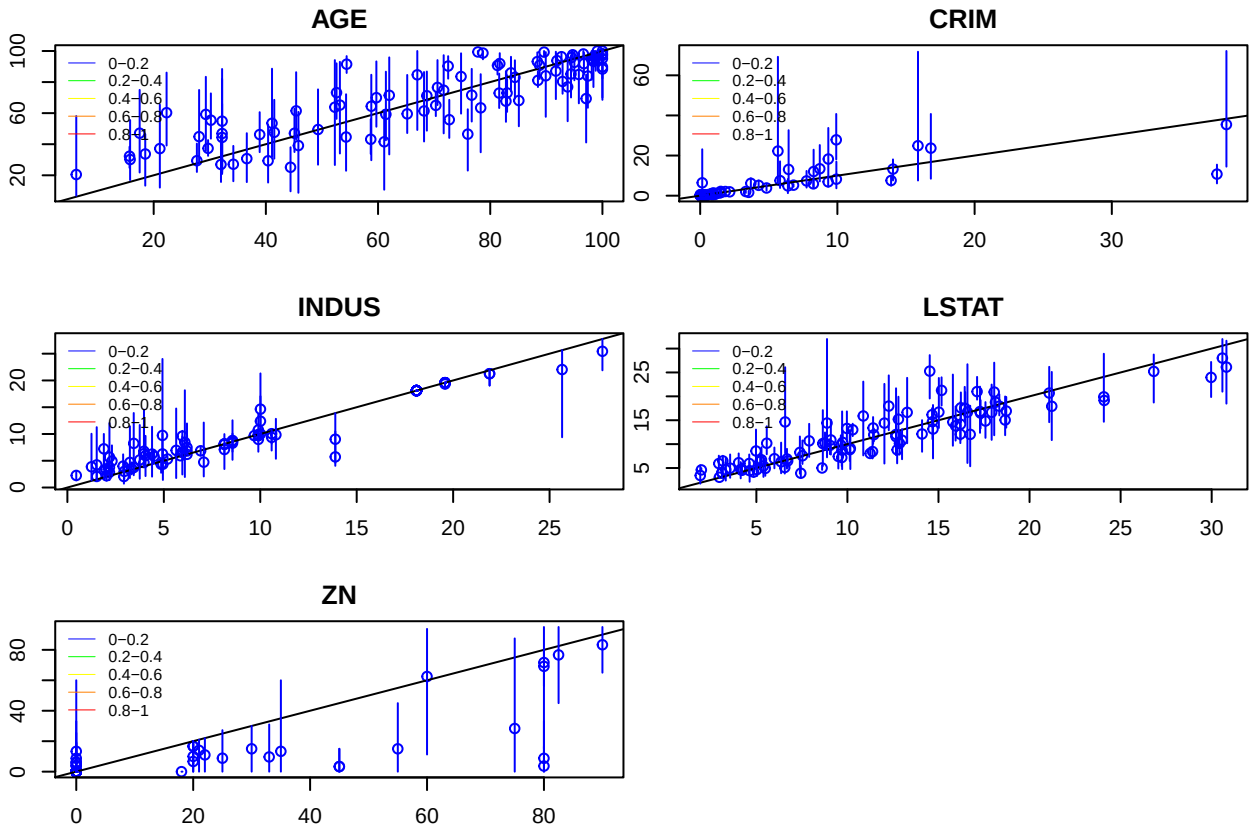
Pour cela, nous traçons le graphe d'overimputation. Dans ce cas, chaque donnée observée est supprimée et pour chacune d'entre elles, 100 valeurs sont predites (en utilisant la même méthode d'imputation choisie); la moyenne et des intervalles de confiance de 90% sont calculés pour ces valeurs.

Sur ces graphiques, la 1ere bissectrice (y=x) représente l'imputation parfaite. La qualité de l'imputation se mesure en observant la proximité des intervalles de confiance avec cette droite. On espère que 90% des intervalles traversent la 1ere bissectrice. La couleur des intervalles represente la fraction des données manquantes (entre 0 - 20% pour notre cas)

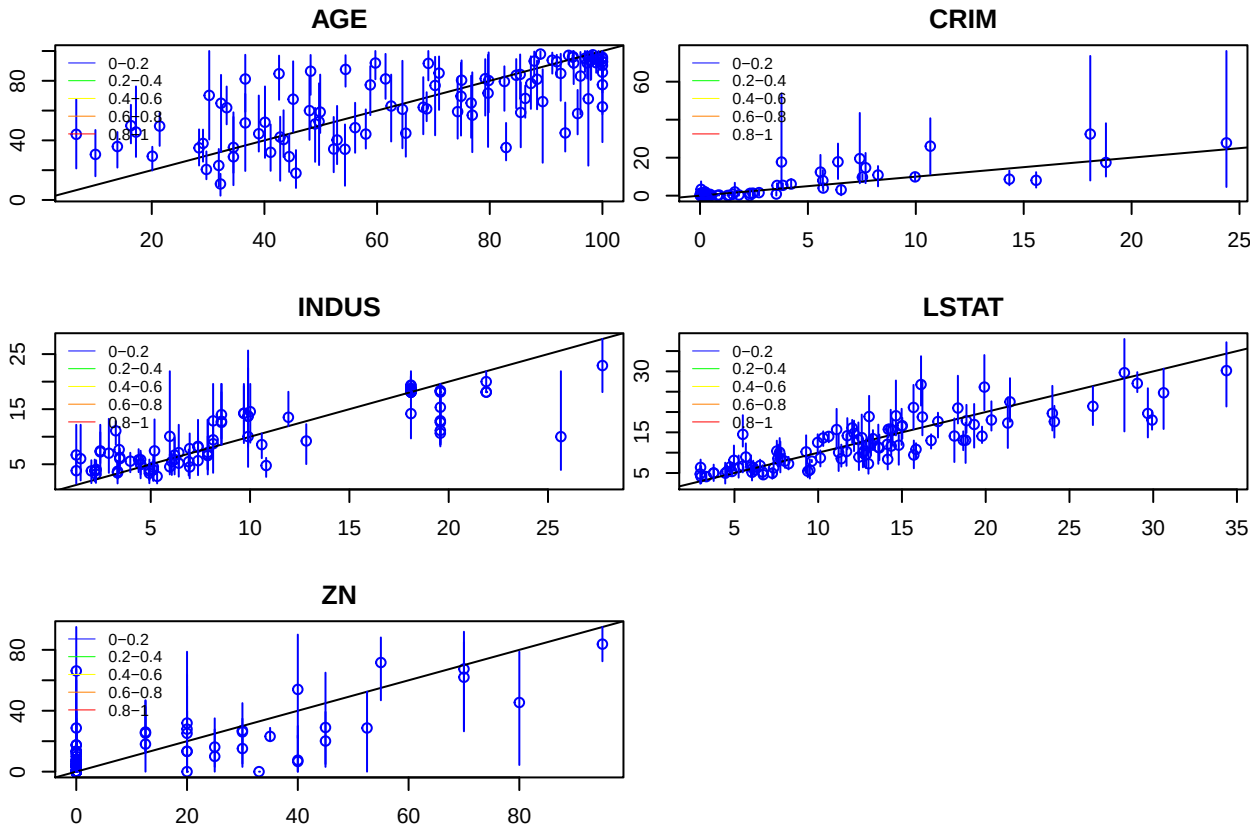
Ajustement de la regression stochastique :



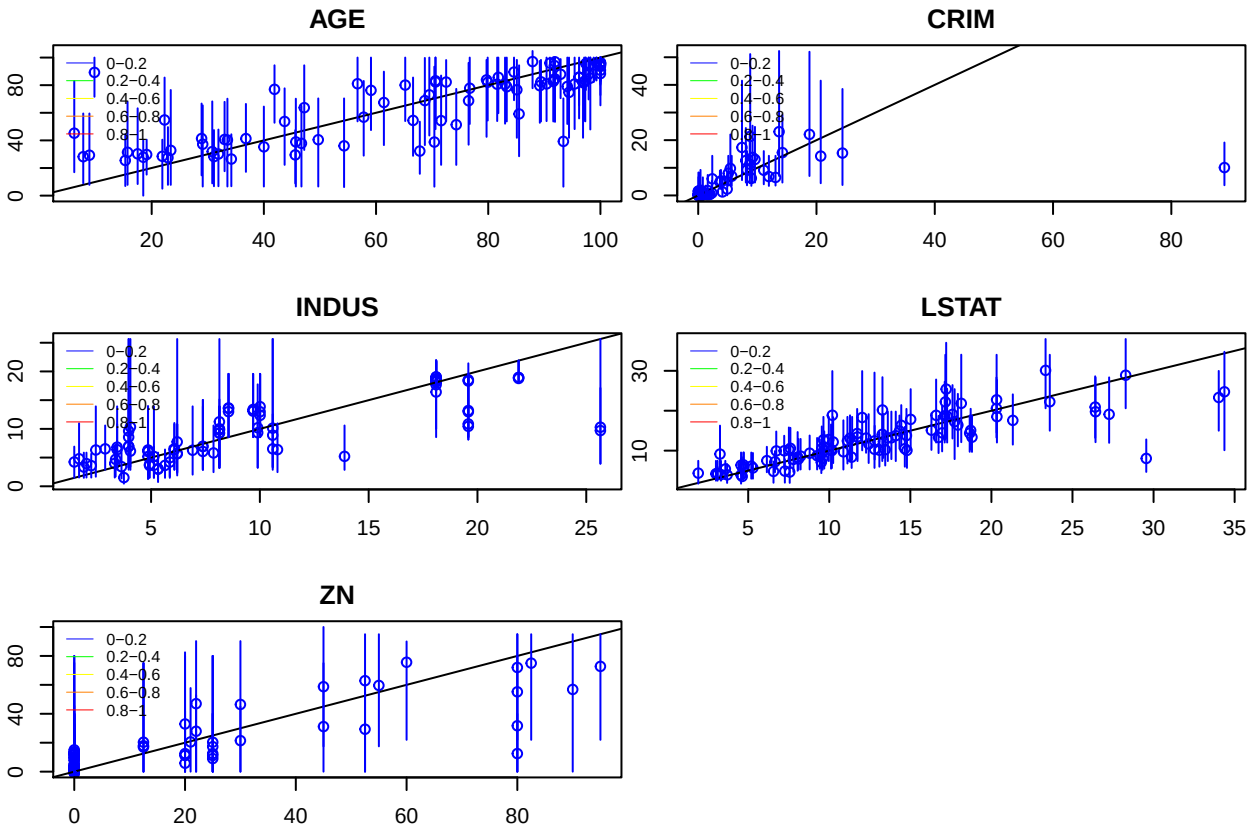
Ajustement des forêts aléatoires :



Ajustement de la predictive mean matching :



Ajustement de l'ACP :



Comme CHAS n'est pas une variable continue, elle prend uniquement les valeurs 0 et 1, elle n'apparaît pas dans les graphiques.

On observe que pour les variables INDUS, AGE et LSTAT la plupart des intervalles de confiance contiennent bien la ligne diagonale pour chacune des méthodes. Cependant, ce n'est pas le cas pour les variables CRIM et ZN.

Conclusions

Après nos trois diagnostics il semblerait que nos modèles avec les differents méthodes:

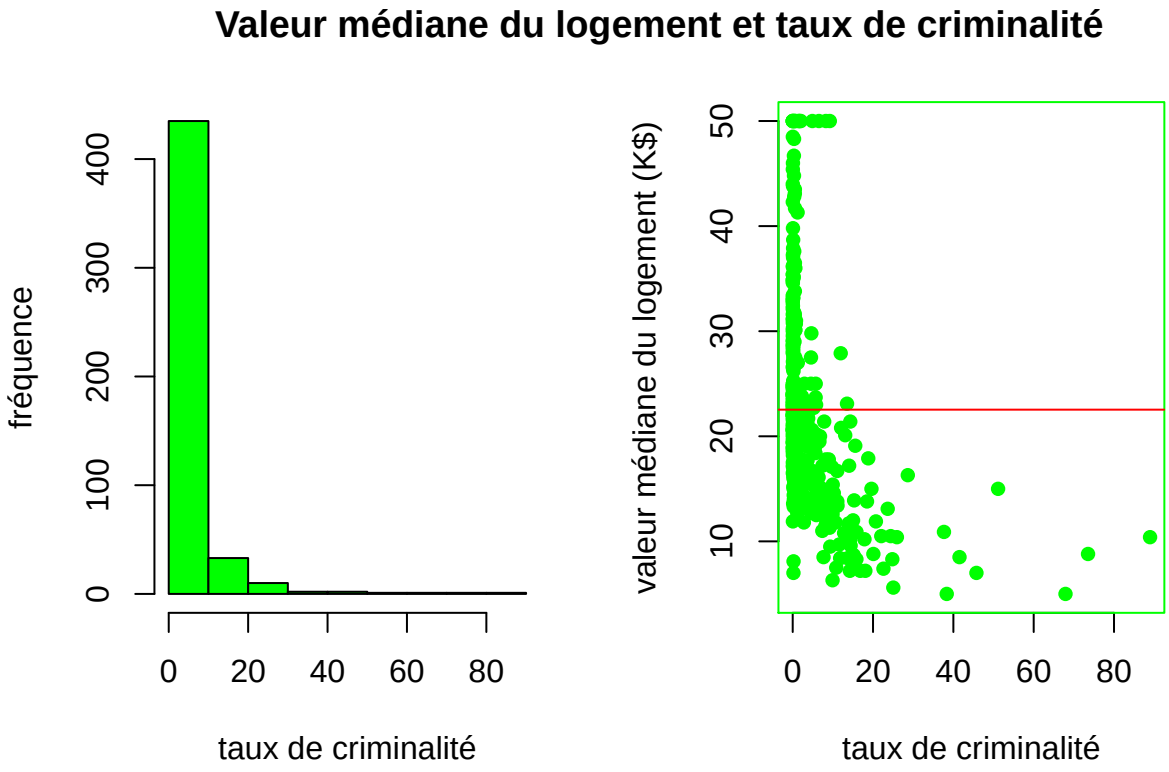
- Respectent bien la distribution des variables, même si la régression stochastique ne le fait pas si bien que les autres
- Nous n'avons pas des problèmes de convergence
- Avec l'overimputation nous voyons que les variables CRIm et ZN ne sont pas toujours bien representées. En CRIM c'est visible principalement pour les grands valeurs.

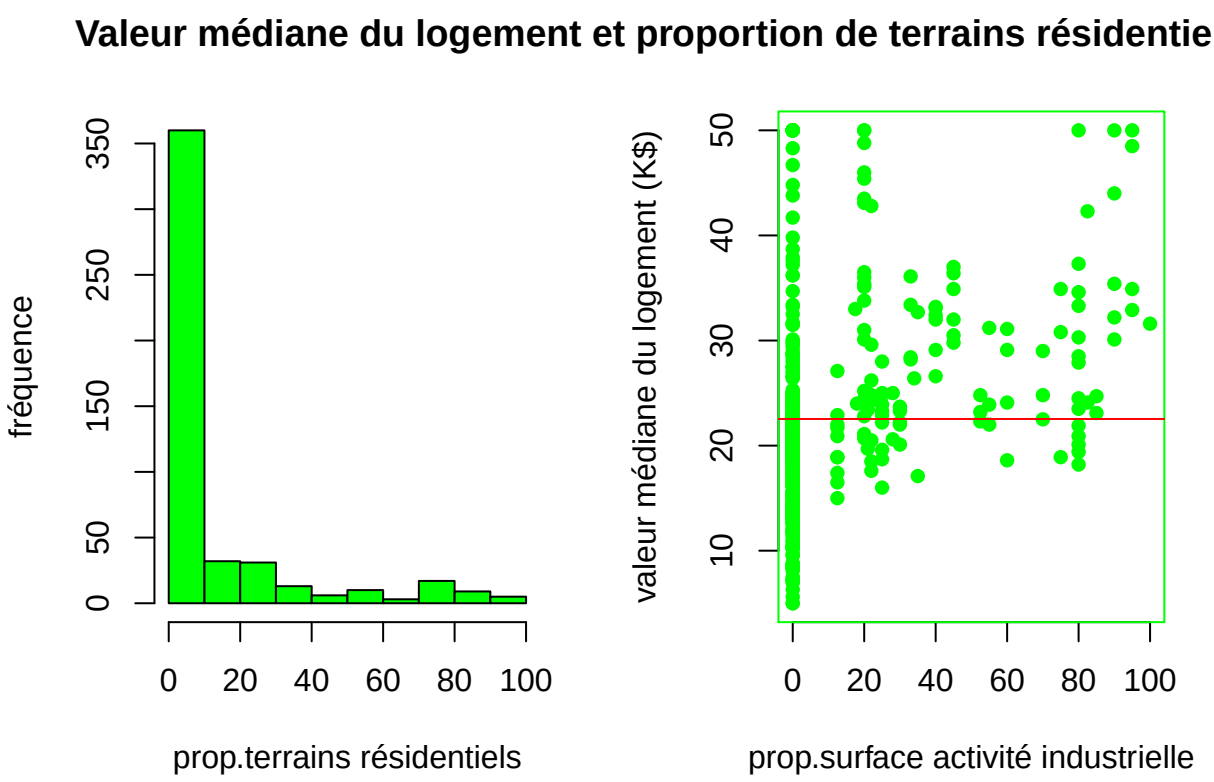
Comme le pourcentage des valeurs manquantes est bas

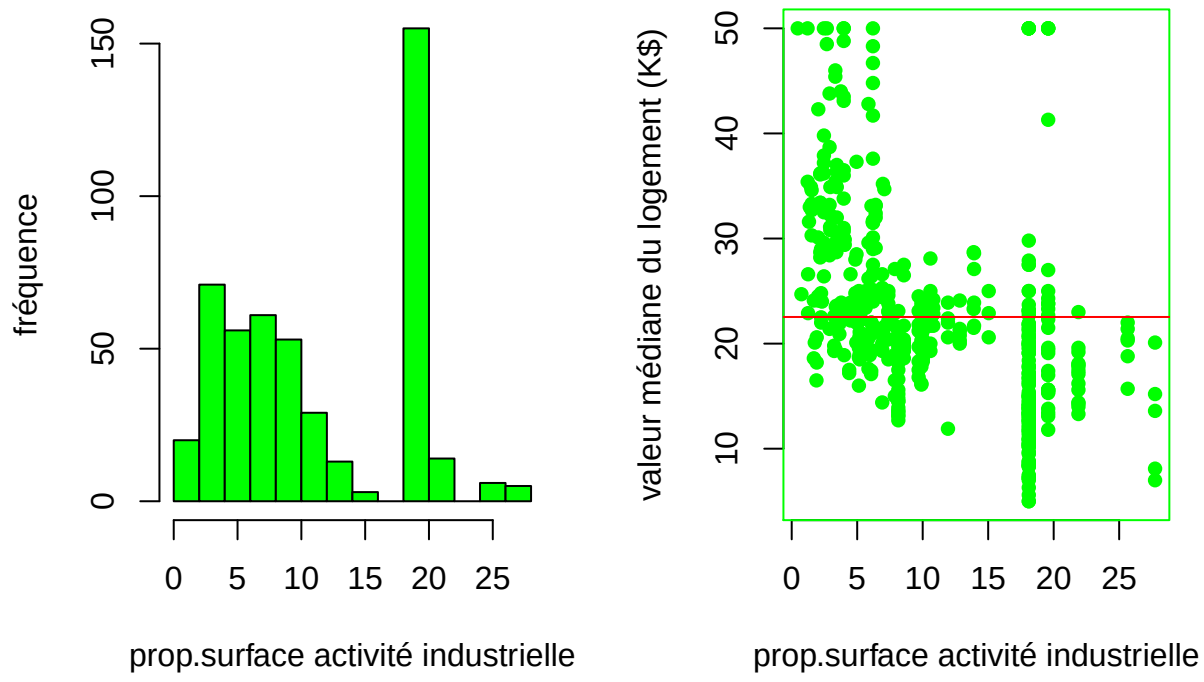
Annexes

Annexe 1.1 : graphiques exploratoires du jeu de données

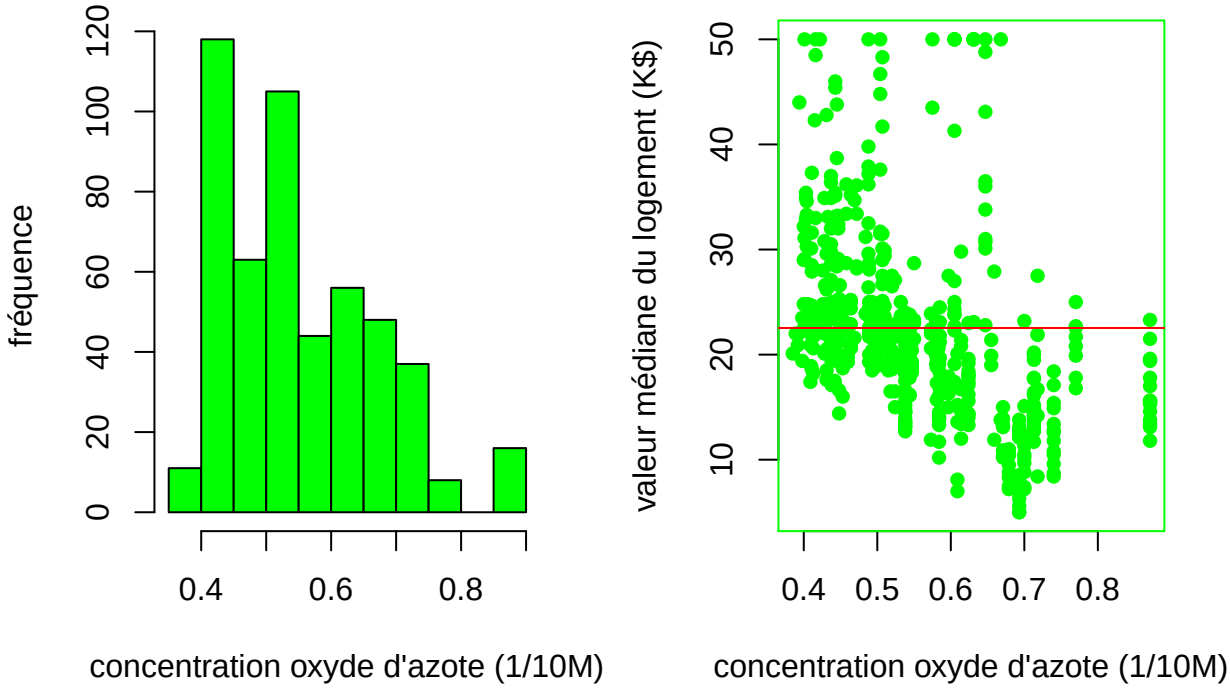
Les répresentations graphiques donnent, pour chaque variable explicative, la distribution sous forme d'histogramme et le nuage de point de la variable vs la valeur médiane du logement (en rouge la valeur moyenne = 22.5K$).



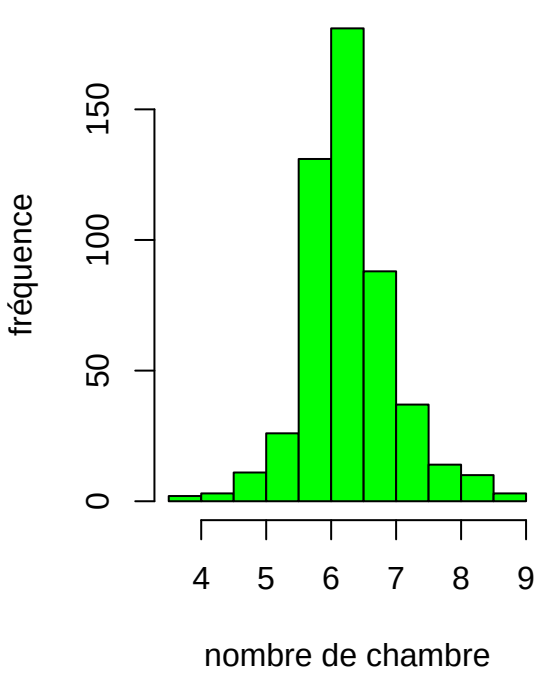

Valeur médiane du logement et proportion de terrains résidentiels
fréquence
0
50
150
250
350
0
20
40
60
80
100
prop.terrains résidentiels
valeur médiane du logement (K$)
10
20
30
40
50
0
20
40
60
80
100
prop.surface activité industrielle

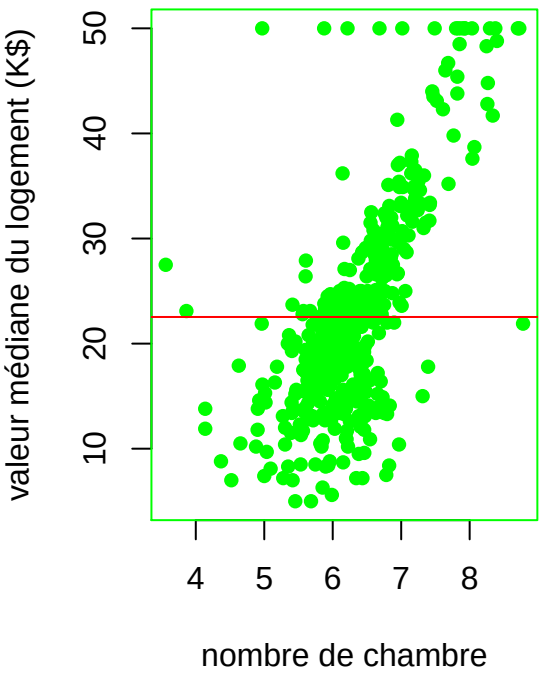

Valeur médiane du logement et proportion de surface d'activité industr
fréquence
0
50
100
150
0
5
10
15
20
25
prop.surface activité industrielle
valeur médiane du logement (K$)
10
20
30
40
50
0
5
10
15
20
25
prop.surface activité industrielle

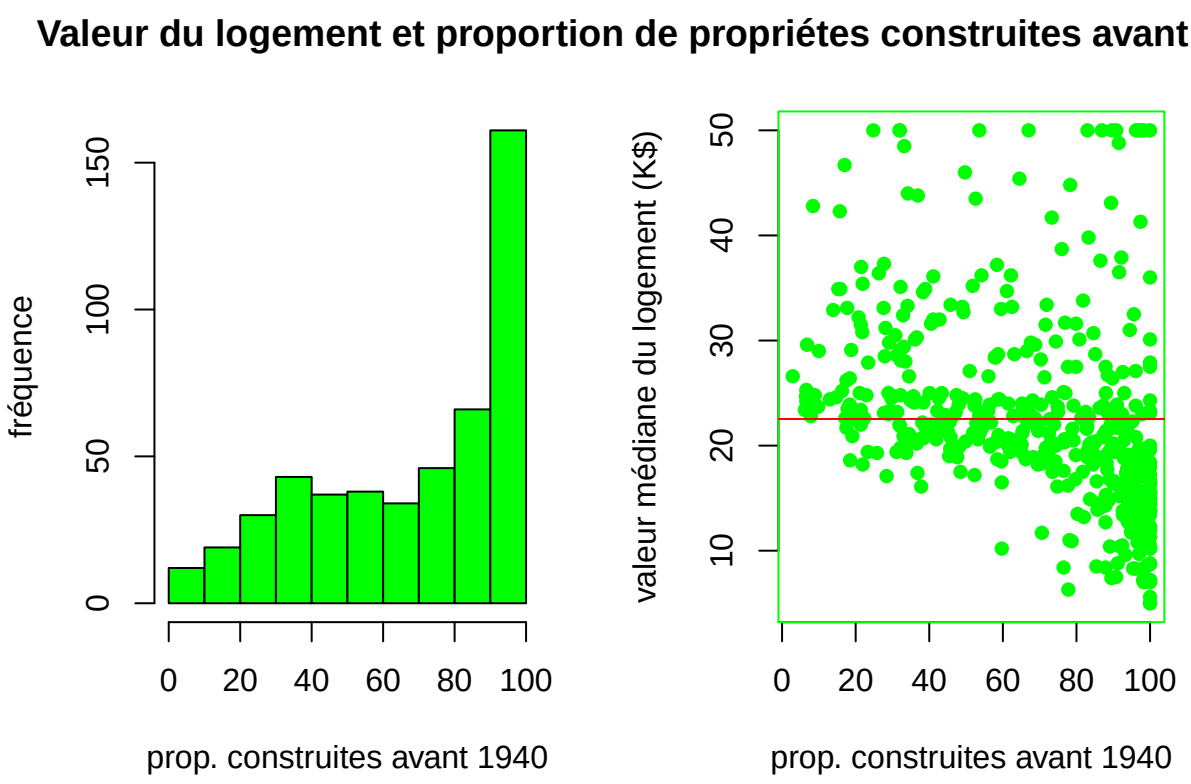
Valeur médiane du logement et concentration en oxyde d'azote



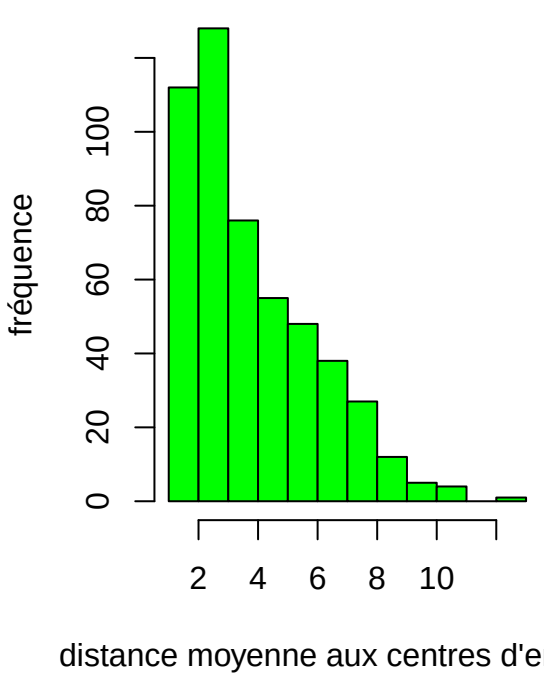
Valeur médiane du logement et nombre moyen de chambre

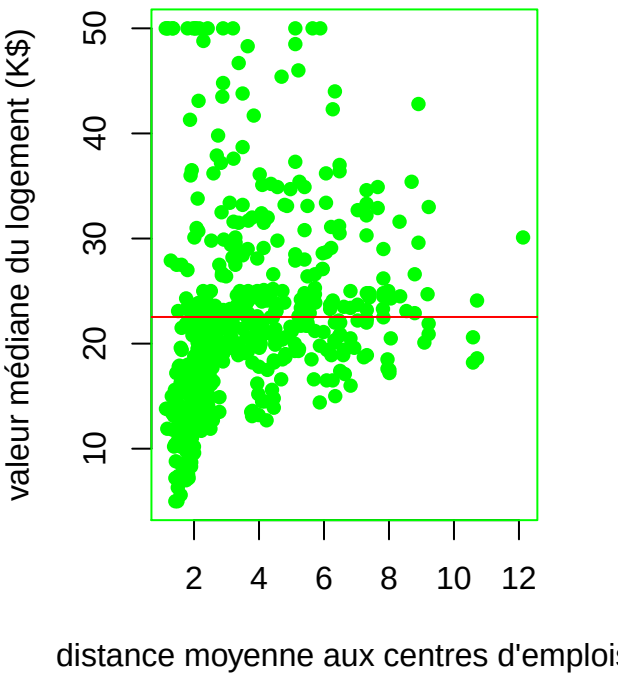


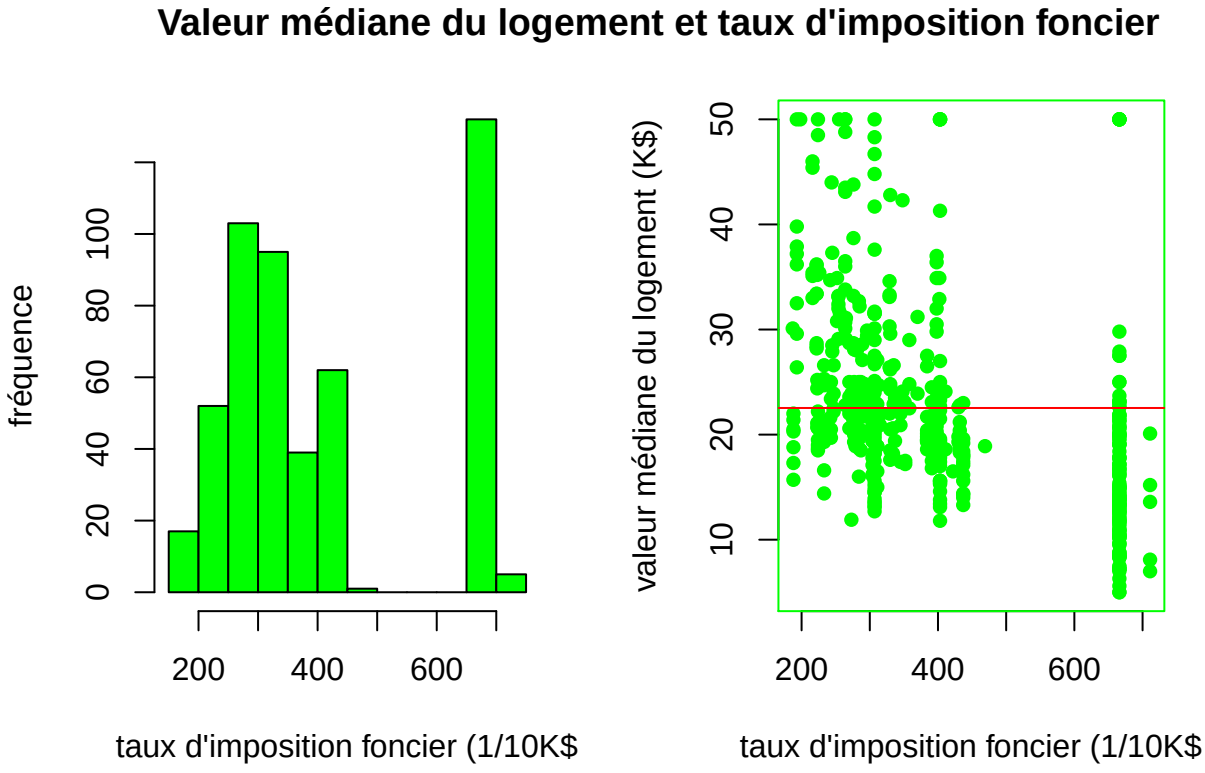


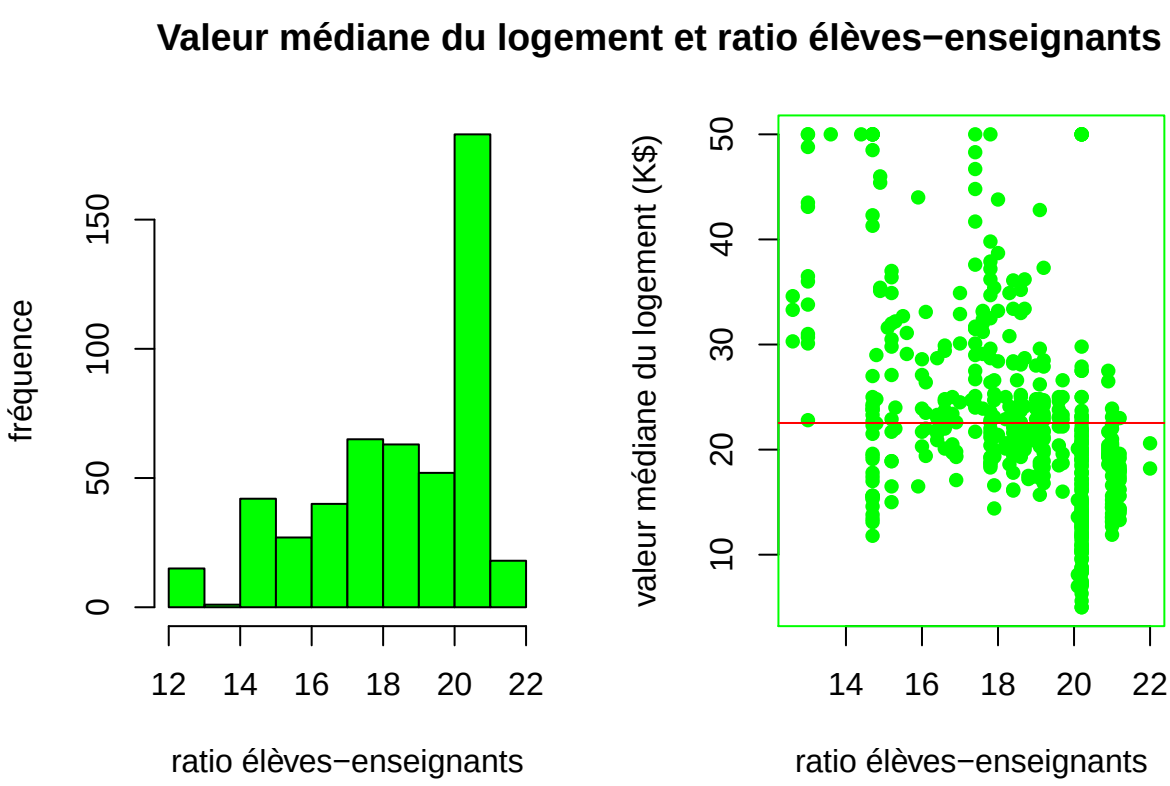

Valeur du logement et proportion de propriétes construites avant 19
fréquence
0
50
100
150
0
20
40
60
80
100
prop. construites avant 1940
valeur médiane du logement (K$)
10
20
30
40
50
0
20
40
60
80
100
prop. construites avant 1940

Valeur du logement et distance moyenne aux 5 centres d'emplois

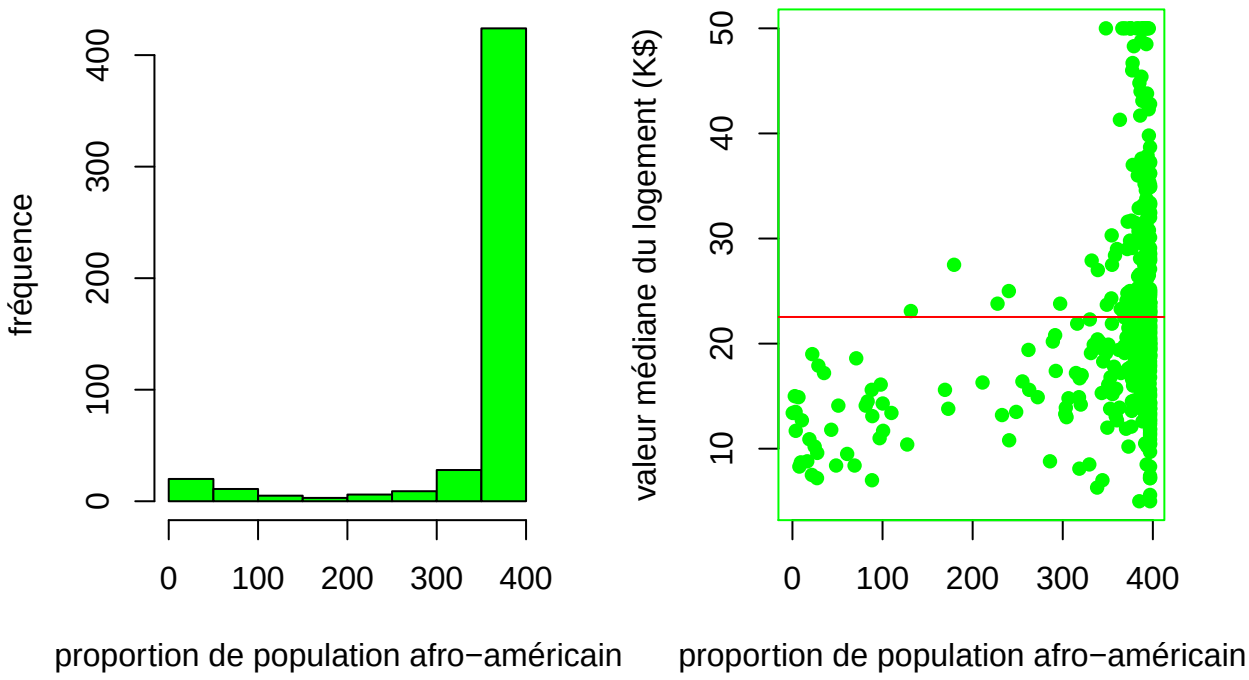


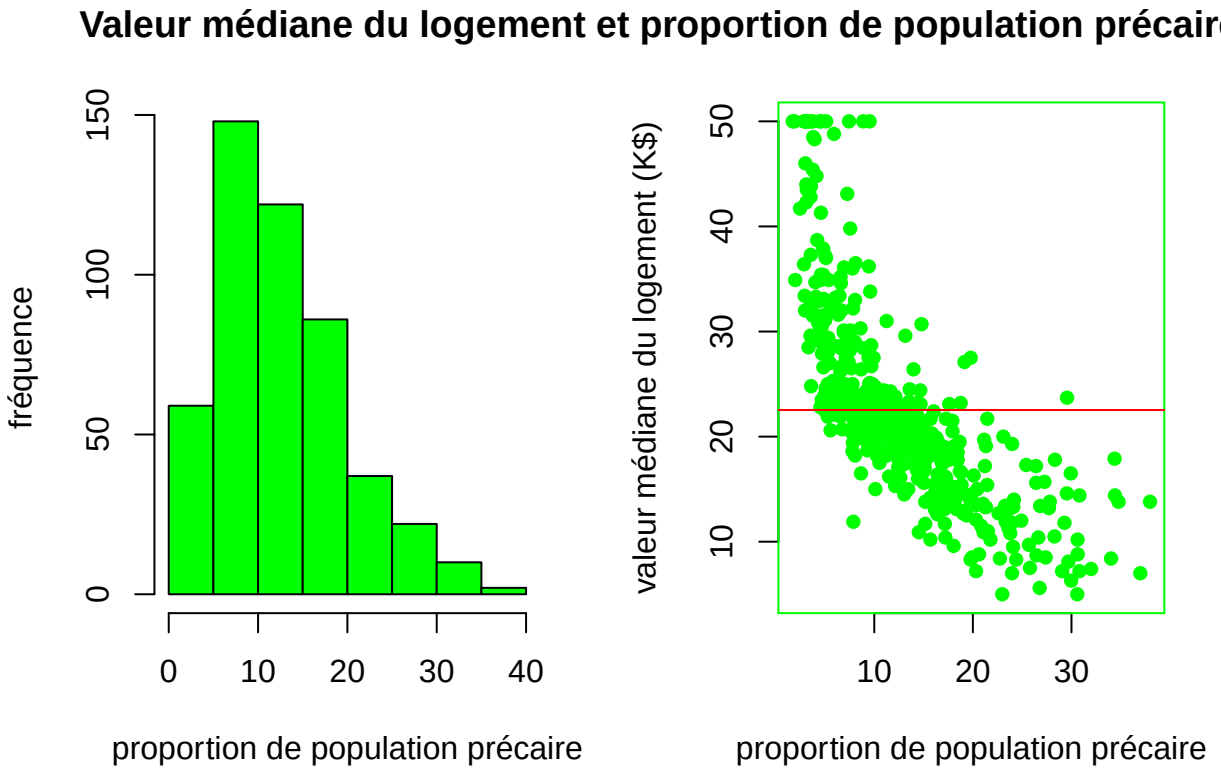


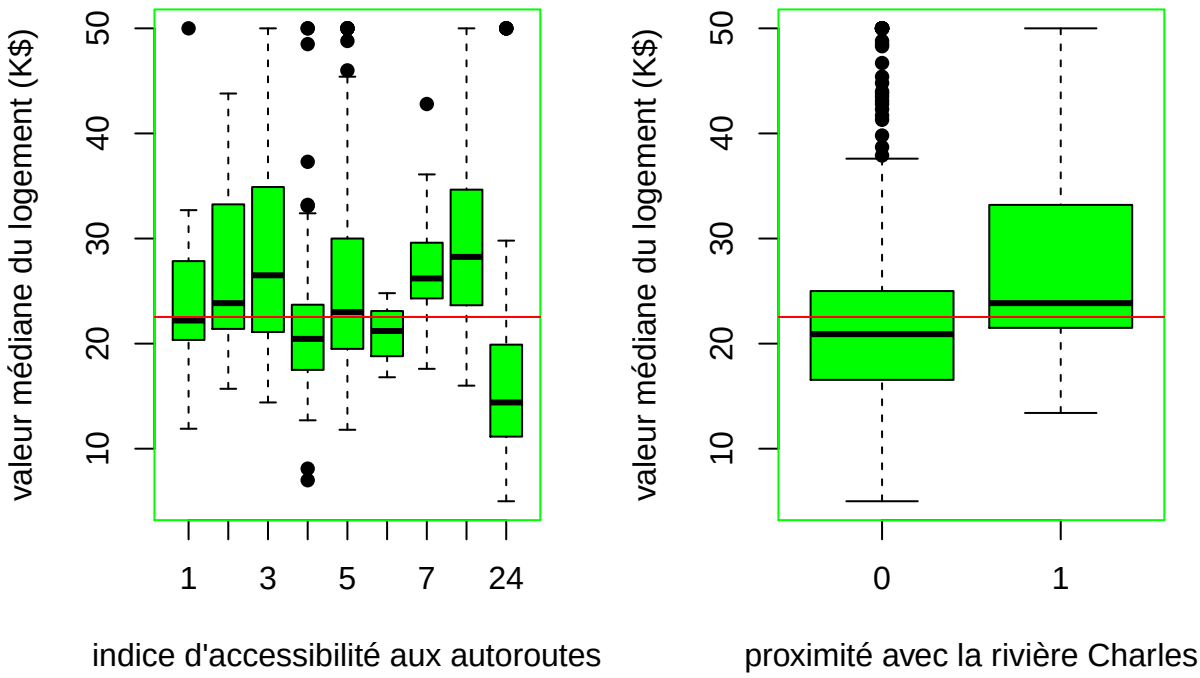

Valeur médiane du logement et taux d'imposition foncier
fréquence
0
20
40
60
80
100
200
400
600
taux d'imposition foncier (1/10K$
valeur médiane du logement (K$)
10
20
30
40
50
200
400
600
taux d'imposition foncier (1/10K$


Valeur médiane du logement et ratio élèves−enseignants
fréquence
0
50
100
150
12
14
16
18
20
22
ratio élèves−enseignants
valeur médiane du logement (K$)
10
20
30
40
50
14
16
18
20
22
ratio élèves−enseignants

Valeur du logement et proportion de population afro−américaine




Valeur médiane du logement et proportion de population précaire
fréquence
0
50
100
150
0
10
20
30
40
proportion de population précaire
valeur médiane du logement (K$)
10
20
30
40
50
10
20
30
proportion de population précaire



Annexe 1.2 : corrélation entre les variables du jeu de données

Les corrélations les plus significatives apparaissent en rouge.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
CRIM	1	-0.19	0.39	-0.05	0.42	-0.23	0.34	-0.37	0.61	0.56	0.27	-0.39	0.46
ZN	-0.19	1	-0.52	-0.03	-0.52	0.34	-0.57	0.65	-0.3	-0.31	-0.42	0.17	-0.42
INDUS	0.39	-0.52	1	0.05	0.76	-0.4	0.64	-0.7	0.59	0.73	0.4	-0.34	0.6
CHAS	-0.05	-0.03	0.05	1	0.08	0.1	0.07	-0.1	0.01	-0.03	-0.1	0.07	-0.04
NOX	0.42	-0.52	0.76	0.08	1	-0.32	0.73	-0.77	0.63	0.68	0.21	-0.38	0.59
RM	-0.23	0.34	-0.4	0.1	-0.32	1	-0.25	0.22	-0.24	-0.32	-0.39	0.12	-0.64
AGE	0.34	-0.57	0.64	0.07	0.73	-0.25	1	-0.75	0.44	0.5	0.26	-0.28	0.6
DIS	-0.37	0.65	-0.7	-0.1	-0.77	0.22	-0.75	1	-0.48	-0.53	-0.23	0.29	-0.51
RAD	0.61	-0.3	0.59	0.01	0.63	-0.24	0.44	-0.48	1	0.9	0.44	-0.44	0.51
TAX	0.56	-0.31	0.73	-0.03	0.68	-0.32	0.5	-0.53	0.9	1	0.45	-0.44	0.57
PTRATIO	0.27	-0.42	0.4	-0.1	0.21	-0.39	0.26	-0.23	0.44	0.45	1	-0.18	0.4
B	-0.39	0.17	-0.34	0.07	-0.38	0.12	-0.28	0.29	-0.44	-0.44	-0.18	1	-0.38
LSTAT	0.46	-0.42	0.6	-0.04	0.59	-0.64	0.6	-0.51	0.51	0.57	0.4	-0.38	1
MEDV	-0.4	0.41	-0.51	0.17	-0.46	0.72	-0.41	0.28	-0.42	-0.51	-0.54	0.35	-0.74

Annexe 1.3 : regression lineaire MEDV sur l'ensemble des variables

##

```
## Call:
## lm(formula = MEDV ~ ., data = dHB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4234  -2.5830  -0.5079   1.6681  26.2604
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.680059   5.681290   5.752 1.81e-08 ***
## CRIM         -0.097594   0.032457  -3.007 0.002815 **
## ZN            0.048905   0.014398   3.397 0.000754 ***
## INDUS         0.030379   0.065933   0.461 0.645237
## CHAS          2.769378   0.925171   2.993 0.002940 **
## NOX         -17.969028   4.242856  -4.235 2.87e-05 ***
## RM            4.283252   0.470710   9.100  < 2e-16 ***
## AGE          -0.012991   0.014459  -0.898 0.369504
## DIS          -1.458510   0.211007  -6.912 2.03e-11 ***
## RAD           0.285866   0.069298   4.125 4.55e-05 ***
## TAX          -0.013146   0.003955  -3.324 0.000975 ***
## PTRATIO      -0.914582   0.140581  -6.506 2.44e-10 ***
## B             0.009656   0.002970   3.251 0.001251 **
## LSTAT        -0.423661   0.055022  -7.700 1.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.487 on 380 degrees of freedom
##   (112 observations deleted due to missingness)
## Multiple R-squared:  0.7671, Adjusted R-squared:  0.7591
## F-statistic: 96.29 on 13 and 380 DF,  p-value: < 2.2e-16
```

Annexe 1.4 : regression linéaire MEDV sur variables choisies par stepwise

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = dHB2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.214  -2.552  -0.503   1.768  26.027
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.975051   5.630782   5.856 1.02e-08 ***
## CRIM         -0.098151   0.032405  -3.029 0.002621 **
## ZN            0.049962   0.014169   3.526 0.000473 ***
## CHAS          2.788061   0.919721   3.031 0.002600 **
## NOX         -18.467815   3.895303  -4.741 3.01e-06 ***
## RM            4.166982   0.455473   9.149  < 2e-16 ***
## DIS          -1.420599   0.197272  -7.201 3.20e-12 ***
## RAD           0.282322   0.065525   4.309 2.09e-05 ***
## TAX          -0.012400   0.003471  -3.573 0.000398 ***
```

```
## PTRATIO      -0.914756   0.138631  -6.599 1.39e-10 ***
## B             0.009477   0.002961   3.201 0.001483 **
## LSTAT        -0.439994   0.051567  -8.532 3.41e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.481 on 382 degrees of freedom
## Multiple R-squared:  0.7665, Adjusted R-squared:  0.7598
## F-statistic:   114 on 11 and 382 DF,  p-value: < 2.2e-16
```

Annexe 3.1 : Imputation - méthodologie

L'imputation correspond à l'action de convertir un échantillon incomplet en un échantillon complet. Le but de l'imputation multiple est d'affecter plusieurs fois des données manquantes, d'analyser les données complétées et ensuite d'intégrer les résultats des analyses.

Les 7 étapes de l'imputation:

Etape 1 - Décider si supposition de MAR est plausible.
(vu en partie 2)

Etape 2 - Identifier la forme du modèle d'imputation.
Le choix sera orienté par l'échelle de la variable à imputer, et intègre de préférence la connaissance de la relation entre les variables. L'algorithm MICE a besoin d'avoir une méthode univariée d'imputation pour chaque variable incomplète.

Etape 3 - Sélectionner le groupe de variables à inclure comme predicteurs dans le modèle d'imputation (fonction **mice**)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEI
CRIM	0	1	1	1	1	1	1	1	1	1	1	1	1	
ZN	1	0	1	1	1	1	1	1	1	1	1	1	1	
INDUS	1	1	0	1	1	1	1	1	1	1	1	1	1	
CHAS	1	1	1	0	1	1	1	1	1	1	1	1	1	
NOX	1	1	1	1	0	1	1	1	1	1	1	1	1	
RM	1	1	1	1	1	0	1	1	1	1	1	1	1	
AGE	1	1	1	1	1	1	0	1	1	1	1	1	1	
DIS	1	1	1	1	1	1	1	0	1	1	1	1	1	
RAD	1	1	1	1	1	1	1	1	0	1	1	1	1	
TAX	1	1	1	1	1	1	1	1	1	0	1	1	1	
PTRATIO	1	1	1	1	1	1	1	1	1	1	0	1	1	
B	1	1	1	1	1	1	1	1	1	1	1	0	1	
LSTAT	1	1	1	1	1	1	1	1	1	1	1	1	0	
MEDV	1	1	1	1	1	1	1	1	1	1	1	1	1	

Selon la matrice de résultat, CRIM sera prédit à partir de toutes les autres variables (indicateur = 1); idem pour ZN, INDUS, CHAS, AGE et LSTAT. Nous allons utiliser toutes les variables comme predicteurs. Cela est possible car le dataset est encore de taille raisonnable, (difficile sur les grands datasets , à cause de la multicolinearité ou de la capacité des machines)

Etape 4 - Imputer ou non des variables qui sont des fonctions d'autres variables incomplètes.
Dans le cas de notre dataset, les variables avec des données manquantes ne sont pas des fonctions d'autres variables du dataset. Chacune répresente une thématique différente, utile pour l'estimation de la valeur de la maison.

Etape 5 - Définir l'ordre d'imputation des variables (influe sur la convergence de l'algorithme). Par défaut, algorithme MICE impute les données incomplètes du dataset de gauche à droite. L'ordre est à changer si nous avons des soucis de convergence des algorithmes.

Etape 6 - Définir les imputations de départ et le nombre d'itérations

Etape 7 - Imputer et ajuster le modèle la taille de m, le nombre d'ensembles de données imputées. L'imputation du dataset demande de faire des "essais-erreur", pour adapter et améliorer le modèle. Pour démarrer il est conseillé de mettre m = 5 et l'augmenter lors de la dernière étape si on est déjà satisfait avec le modèle.

Annexe 3.2 : Structure des données imputées par régression stochastique

```
##       CRIM                ZN              INDUS              CHAS        
##  Min.   :-6.91103   Min.   :-32.60   Min.   :-0.5783   Min.   :-0.53767  
##  1st Qu.: 0.08057   1st Qu.:  0.00   1st Qu.: 5.1900   1st Qu.: 0.00000  
##  Median : 0.26600   Median :  0.00   Median : 9.6900   Median : 0.00000  
##  Mean   : 3.69822   Mean   : 11.40   Mean   :11.0728   Mean   : 0.06747  
##  3rd Qu.: 3.75547   3rd Qu.: 14.71   3rd Qu.:18.1000   3rd Qu.: 0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.7400   Max.   : 1.00000  
##       NOX               RM              AGE               DIS        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.80   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 69.03   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :119.89   Max.   :12.127  
##       RAD              TAX           PTRATIO            B        
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      LSTAT            MEDV      
##  Min.   :-4.002   Min.   : 5.00  
##  1st Qu.: 7.125   1st Qu.:17.02  
##  Median :11.395   Median :21.20  
##  Mean   :12.626   Mean   :22.53  
##  3rd Qu.:16.860   3rd Qu.:25.00  
##  Max.   :37.970   Max.   :50.00
```

Annexe 3.3 : Structure des données imputées par forêts aléatoires

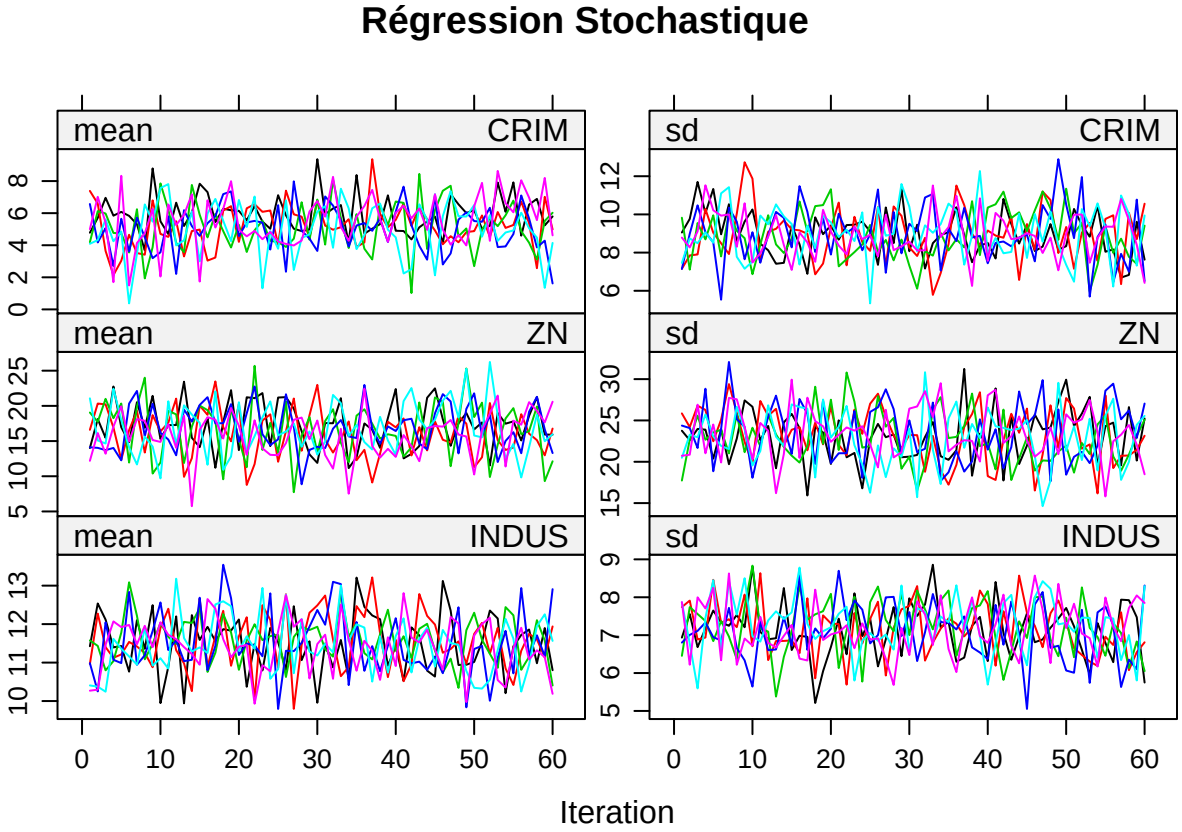
```
##       CRIM                ZN              INDUS             CHAS        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08190   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.26042   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.65851   Mean   : 11.14   Mean   :11.12   Mean   :0.06719  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       NOX               RM              AGE               DIS        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.17   1st Qu.: 2.100
```

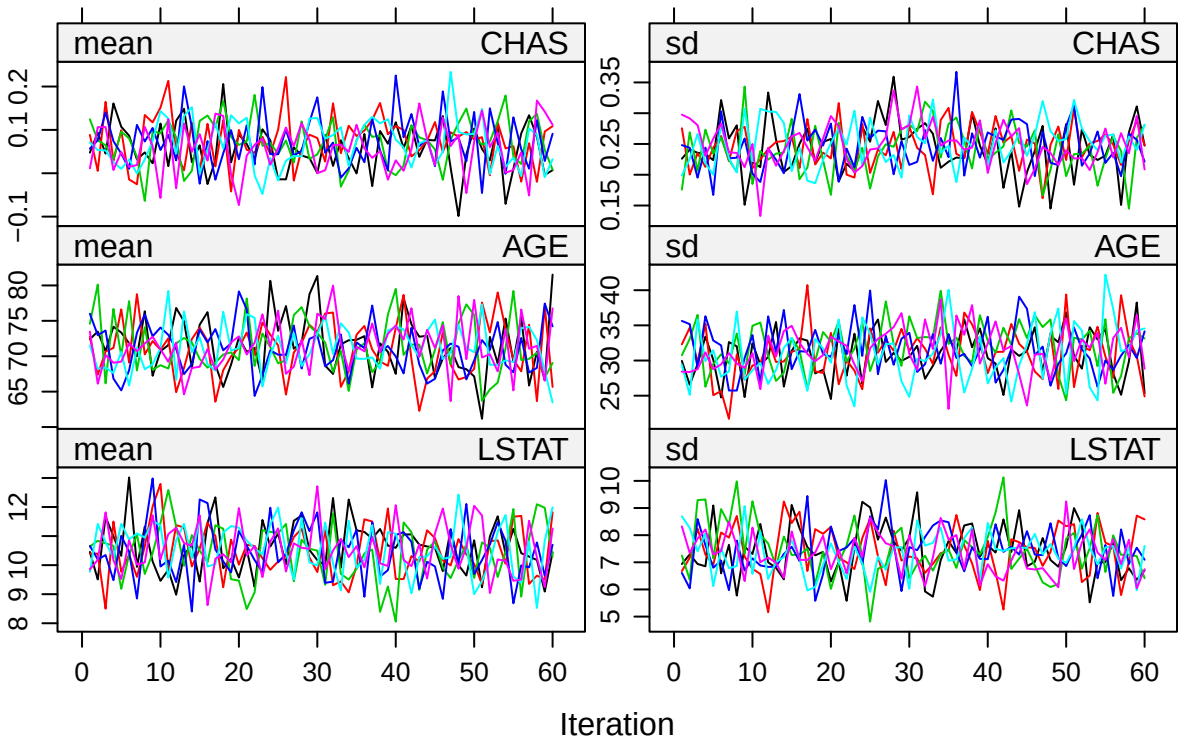
```
##  Median :0.5380   Median :6.208   Median : 76.95   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.61   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       RAD              TAX           PTRATIO            B
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      LSTAT            MEDV
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.70   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Annexe 3.4 : Structure des données imputées par pmm

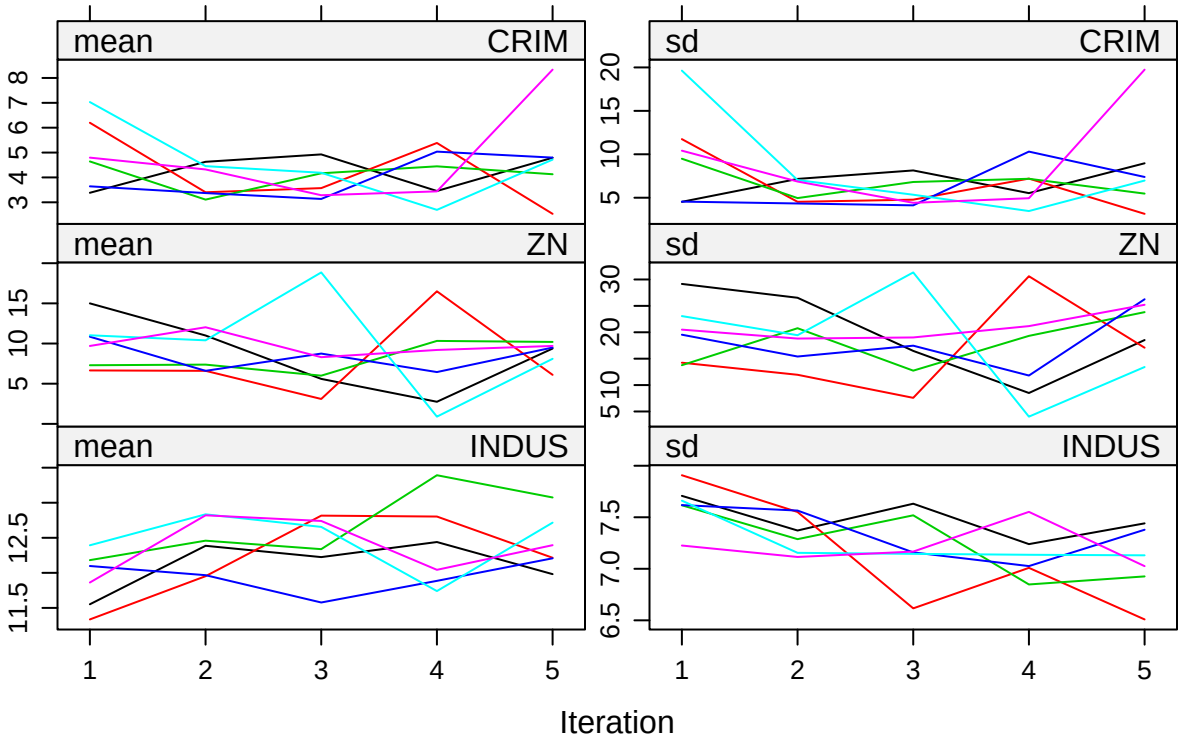
```
##       CRIM                ZN              INDUS            CHAS
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.07987   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.26266   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.59812   Mean   : 11.59   Mean   :11.11   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       NOX               RM             AGE              DIS
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.45   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 76.95   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.67   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.10   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       RAD              TAX           PTRATIO            B
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      LSTAT            MEDV
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.43   Median :21.20
##  Mean   :12.68   Mean   :22.53
##  3rd Qu.:17.06   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Annexe 4.1 Diagnostic 2 : convergence des algorithmes.

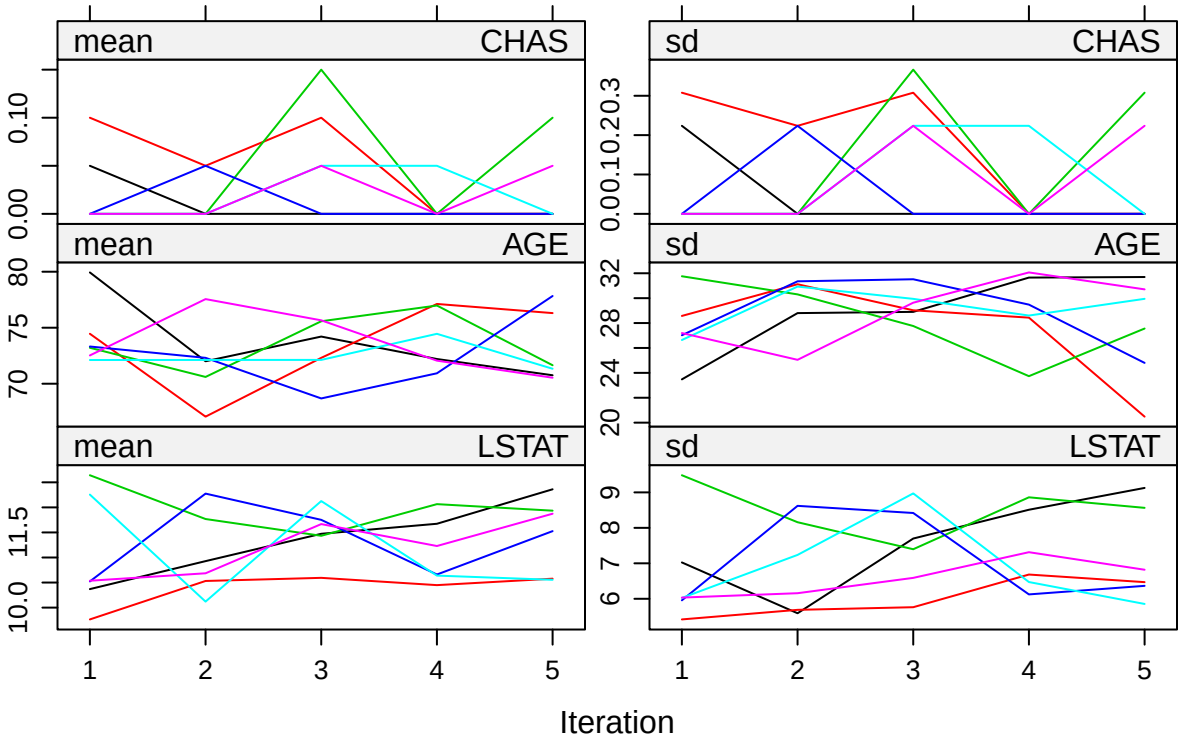


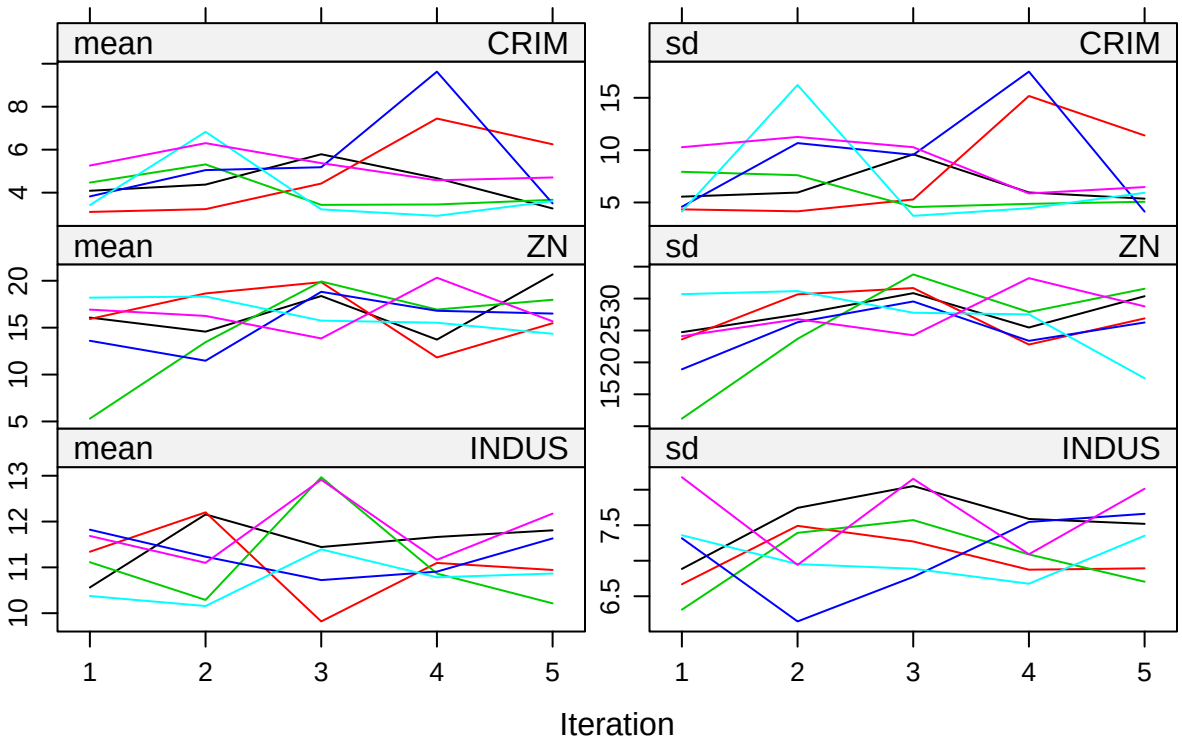


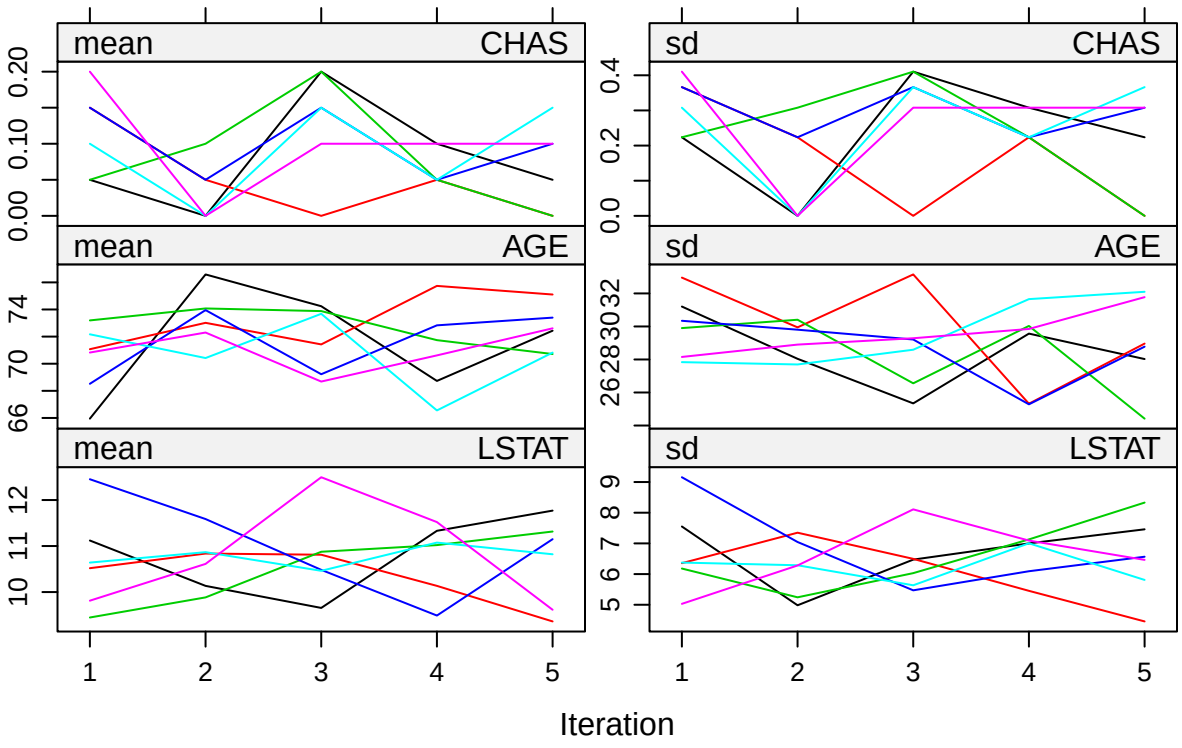
Forêts aléatoires



Forêts aléatoires




Predictive mean matching
mean
CRIM
sd
CRIM
mean
ZN
sd
ZN
mean
INDUS
sd
INDUS
Iteration


Predictive mean matching
mean
CHAS
sd
CHAS
mean
AGE
sd
AGE
mean
LSTAT
sd
LSTAT
0.00
0.10
0.20
0.0
0.2
0.4
66
70
74
26
28
30
32
10
11
12
5
6
7
8
9
1
2
3
4
5
Iteration