Projet_MLG

Karim Assaad, Wassim Ben Youssef et Yoann Dacruz

26 novembre 2015

Contents

Première Partie:Description du jeu de donnée et analyse exploratoire

Introduction

Le projet traité ici propose d'appliquer des modèles de régression et de régularisation dans le but d'élaborer un modèle prédictif d'un jeu de données.

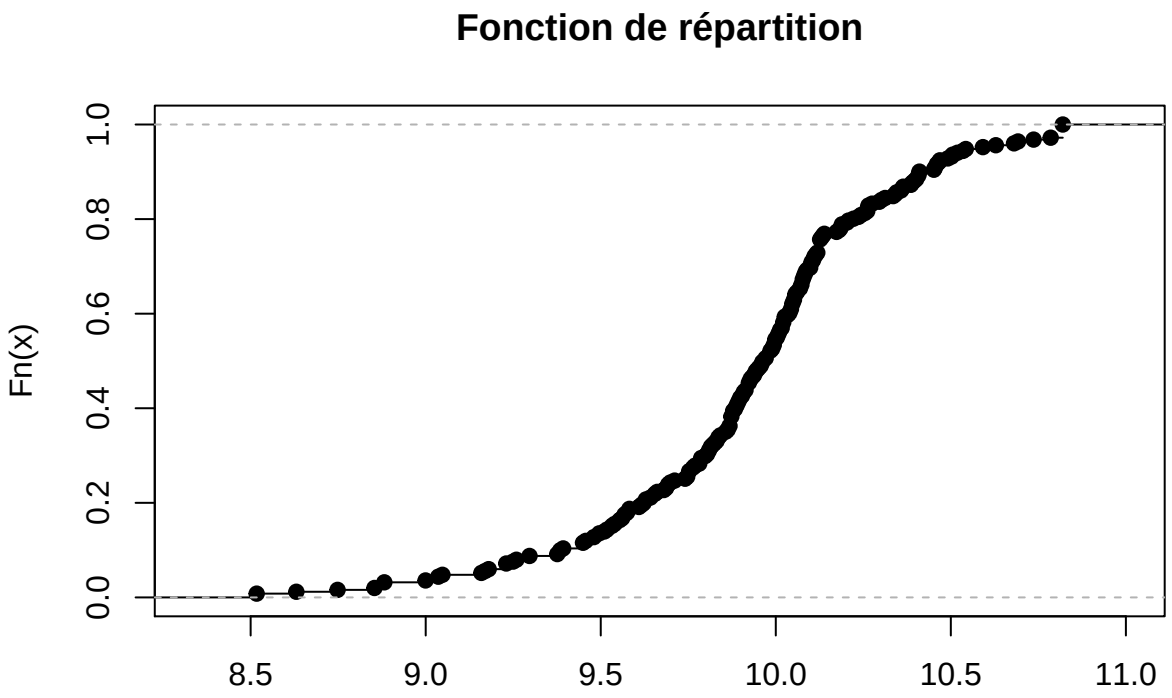
Le jeu de données "Hedonic" a ete choisi pour cette étude parmi les différents jeu de données proposés. Ce jeu de données recense le prix médian des maisons pour des secteurs de la ville de Boston évalué à partir de plusieurs facteurs économiques et sociaux en 1970.

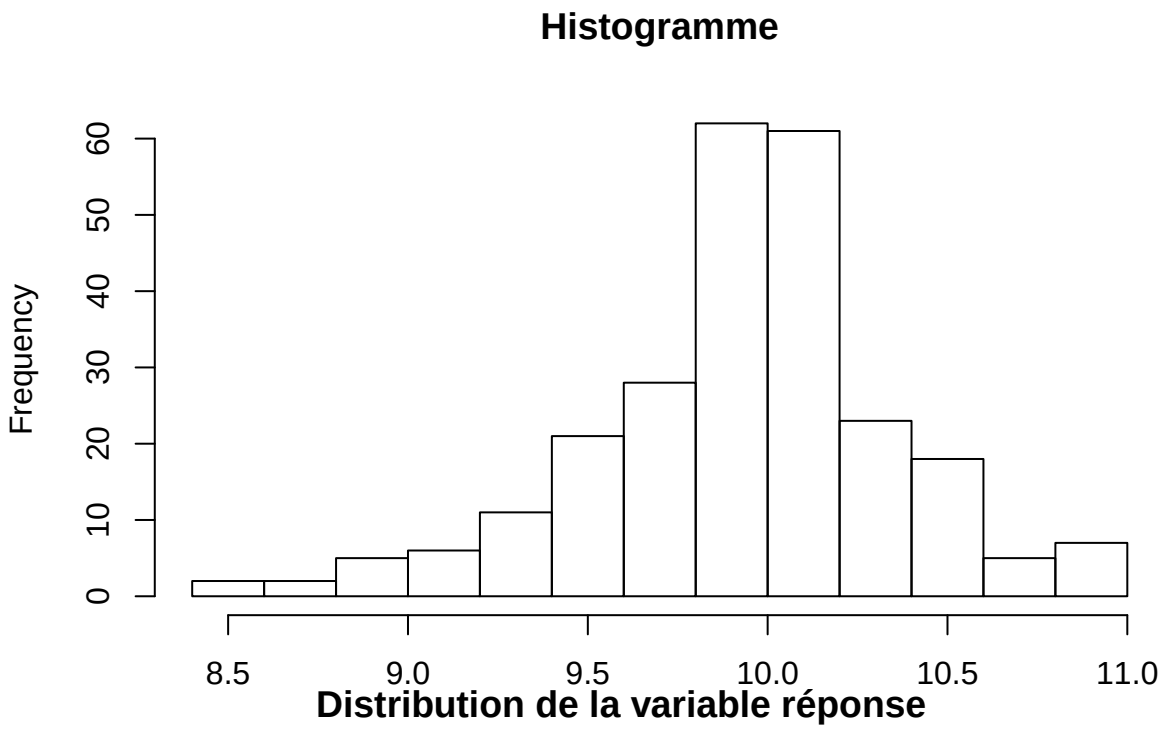
Le jeu de données Hedonic" possède ainsi 15 variables pour un total de 506 secteurs. La principale variable qu'il faut prédire en fonction des autres variables est la variable "mv" qui est le prix médian des maison d'un secteur. Parmi toutes ces variables, une est qualitative : la variable "chas" qui informe par oui ou par non si le secteur est situé dans les environs de la rivière Charles. Les autres variables, toutes quantitatives sont :

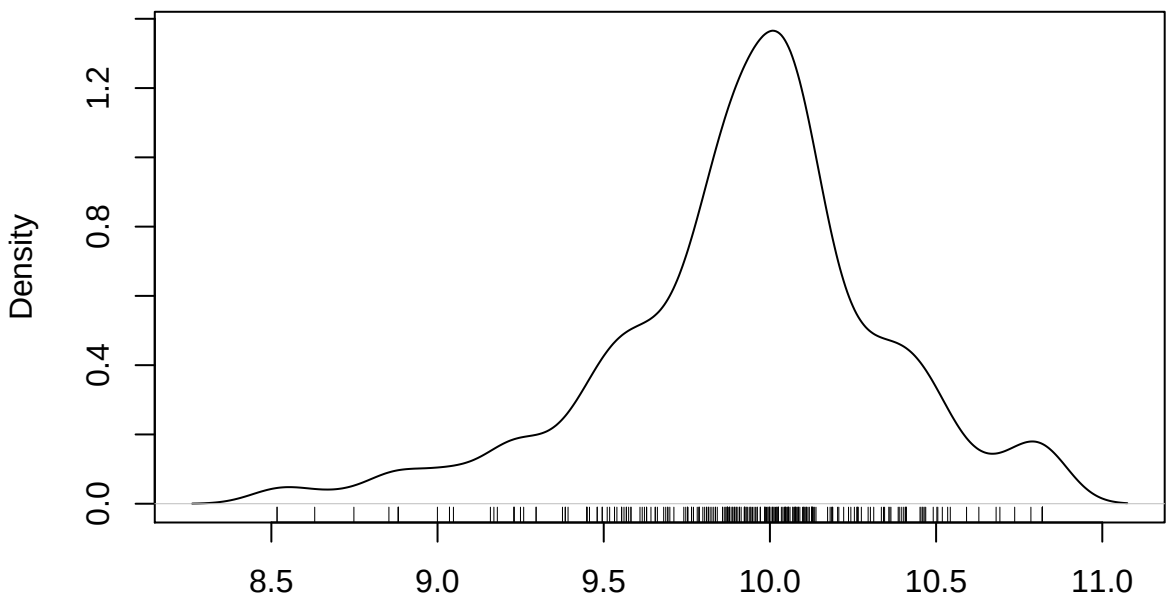
- "crim rate" qui informe du taux de criminalité du secteur

- “zn” qui est le nombre de lots résidentiels d’une superficie de 25000 pieds carrés
- “indus” qui est la proportion d’acres (une unité de mesure équivalente à l’hectare) d’industries et de commerces interentreprises (inversement au commerce de détail)
- “nox” la concentration d’oxyde d’azote annuelle moyenne en pour centaine de millions
- “rm” le nombre moyen de piéces par maison
- “age” la proportion de propriétés construites avant 1940
- “dis” les distance pondérées à cinq centre d’emplois dans la région de Boston
- “rad” l’index d’accessibilité à des voies de circulation importantes joignant le centre de l’agglomération à une voie périphérique ou à une ville de province
- “tax” le taux de la valeur de l’impôt foncier en /10000 (dollar par 10000 dollars)
- “pratio” le ratio d’éléves par professeurs
- “blacks” la proportion de personnes noirs dans la population
- “lstat” la proportion de la population ayant un statut social bas
- “townid” l’identifiant de la ville

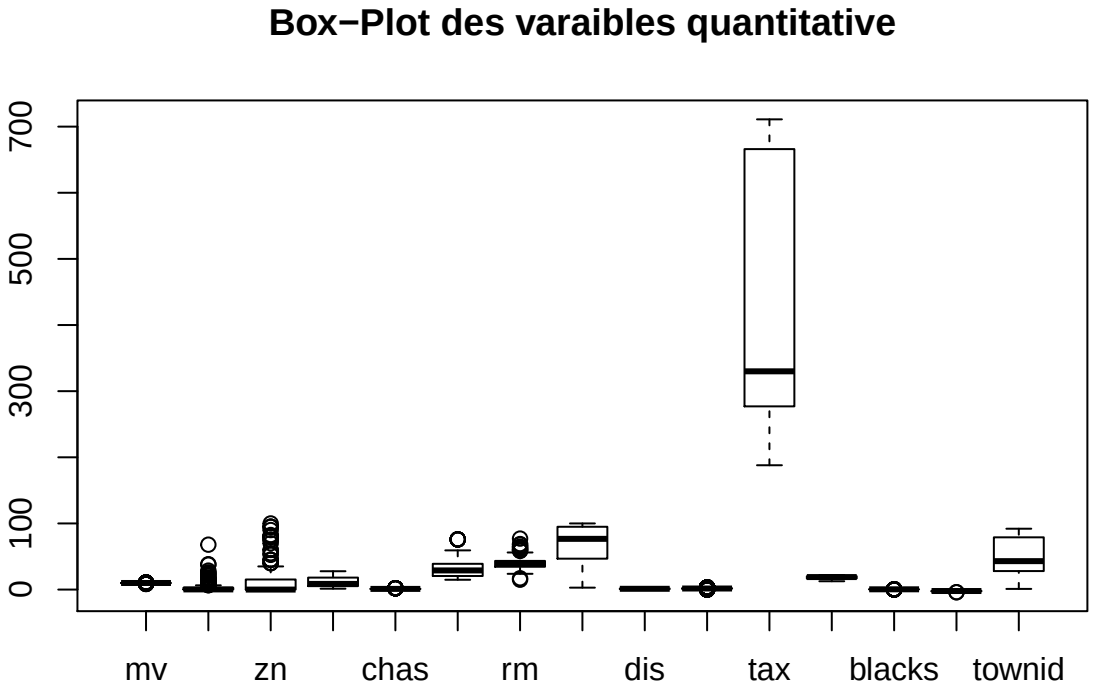
Description des données



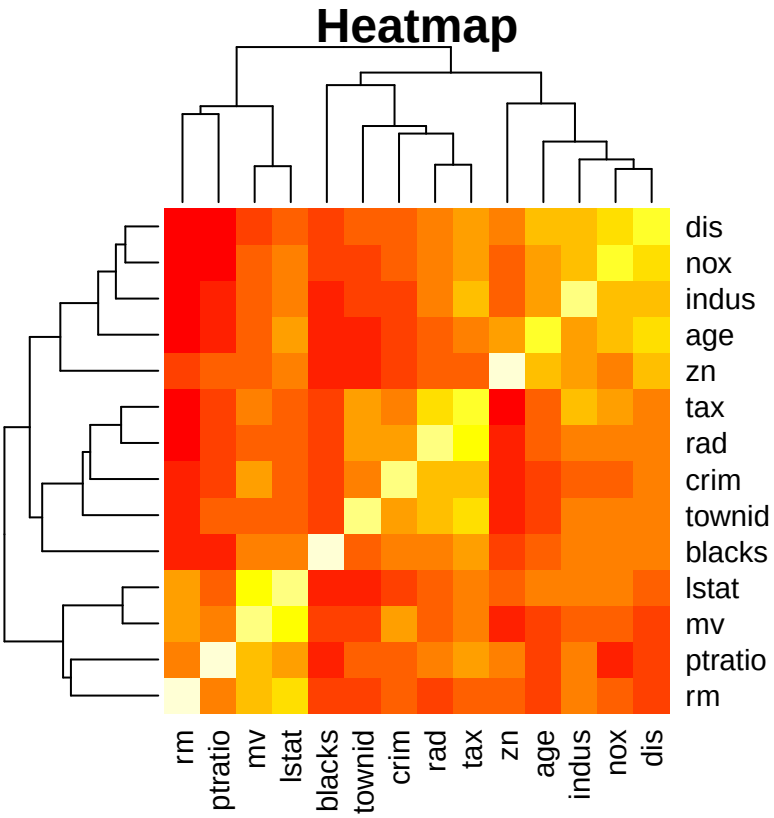




D'après l'histogramme de la variable mv, nous pouvons voir qu'on a une forte concentration des valeurs de mv aux alentours de 10. Nous observons aussi une fréquence maximale de 135 pour des valeurs de mv entre 9.75 et 10. Nous arrivons aux mêmes conclusion en effectuant l'analyse du tracé de l'estimation de densité de cette variable.



En traçant les box-plots pour chacune des variables, nous remarquons que pour les variables mv, crim, zn, rm et blacks ont des valeur aberantes



En observant la matrice des corrélations entre les variables, nous concluons que 2 paires de variables sont corrélées : mv avec lstat, tax avec rad et dis avec nox. Cela signifie que nous pouvons expliquer une variable par l'autre. Nous avons eglement les variables tax rad et townid qui sont correlee positivement entre elles de meme que les deux variables dis et nox. Nous avons egalement plusieurs variables correlees negativement

entre elles : tax et zn, crim et zn, rm et crim, rm et townid, rm et tax, ptratio et dis, ptratio et nox.

Valeur Manquante

```
##      mv    crim      zn   indus    chas     nox      rm     age     dis
##       0       0       0       0       0       0       0       0       0
##     rad     tax ptratio  blacks   lstat  townid
##       0       0       0       0       0       0
```

D'après ce tableau, le jeu de données Hedonic ne contient pas de valeurs manquantes.

Deuxieme Partie:Description des méthodes appliquées

Modèle prédictif et interprétable pour la variable mv

Modele qui depend de aucune variables (null)

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  9.92599 0.02671374 371.5687        0
```

```
## [1] 5.704322e-15
```

Modele qui depend de toute les variables (full)

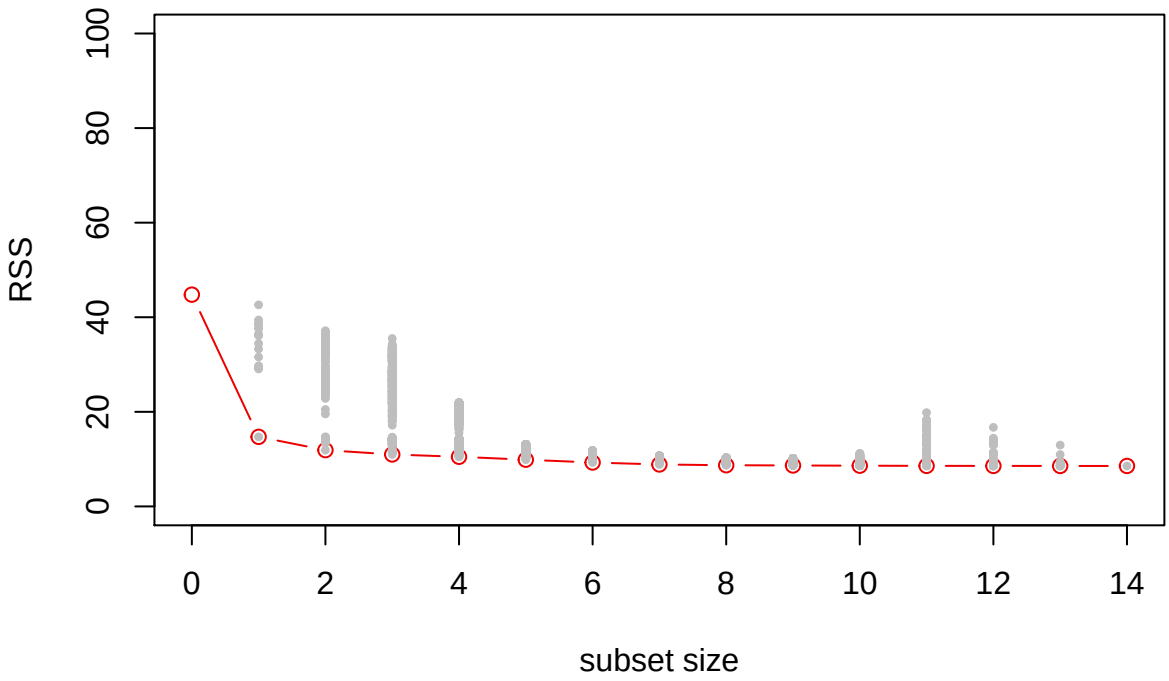
```
##                  Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept)  9.8431223301 0.2330058163  42.2441057 5.007140e-112
## crim        -0.0188267472 0.0023058196  -8.1648829  1.937945e-14
## zn           0.0002865865 0.0007650249   0.3746107  7.082864e-01
## indus        0.0018116093 0.0036299890   0.4990674  6.181968e-01
## chas         0.0564166970 0.0506526742   1.1137950  2.665002e-01
## nox         -0.0073794844 0.0017597467  -4.1934923  3.887898e-05
## rm           0.0036322024 0.0020302013   1.7890849  7.488353e-02
## age          0.0006592613 0.0007937258   0.8305907  4.070444e-01
## dis         -0.1983858066 0.0510843529  -3.8834946  1.337315e-04
## rad          0.1340437338 0.0280841013   4.7729401  3.184975e-06
## tax         -0.0004959807 0.0002040230  -2.4310034  1.580296e-02
## ptratio     -0.0352556353 0.0073935071  -4.7684590  3.250246e-06
## blacks       0.1796718461 0.1690400106   1.0628954  2.889158e-01
## lstat       -0.4002652113 0.0362813713 -11.0322515  4.290648e-23
## townid      -0.0003103795 0.0006563130  -0.4729138  6.367118e-01
```

```
## [1] 0.04546443
```

Dans le summary de full nous pouvons voir que les estimations de chacun des coefficients est assez bonne puisque l'erreur standard la plus importante est celle du coefficient "blacks" et est de 0.147. Cependant, l'intercept est assez mal estimé avec un modèle utilisant tout les coefficients, il est mieux estimé dans "null" lorsqu'on ne prend aucune variable et l'on voit que son erreur standard est de 0.025. Nous pouvons penser au vu des erreurs standards obtenues que notre modèle est assez bon.

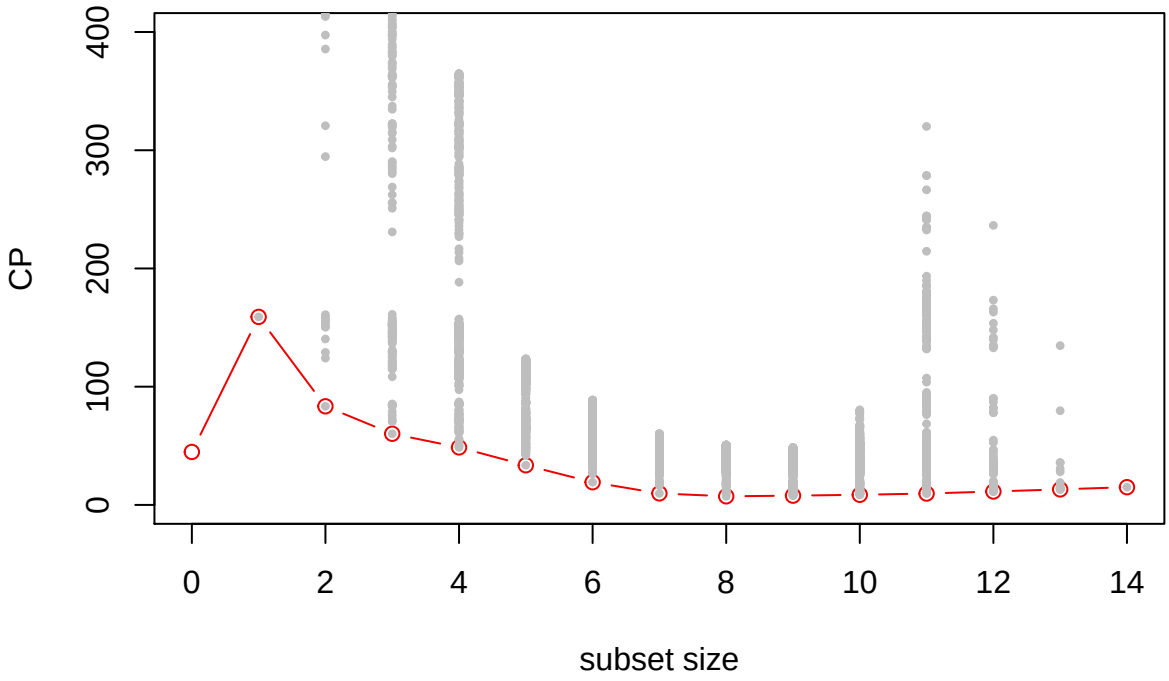
Pour faire une estimation de l'erreur, nous décidons de soustraire à la colonne des "mv" le modèle full obtenu, puis de faire la moyenne de cette colonne. Nous trouvons alors une erreur de 0.016, le modèle trouvé précedemment a donc une erreur assez faible pour ce modèle full.

Representation graphique



Le graphe montre bien que plus nous prenons un nombre important de variables dans le modèle, plus l'estimation des résidus est faible. Donc le meilleur modèle est celui qui prend en compte toutes les variables.

```
##                   1          2          3          4          5
##  44.779901 159.041226  83.414803  60.189787  48.653099  33.551831
##           6          7          8          9         10         11
##  19.177672   9.753495   7.281736   7.847199   8.560352   9.632154
##          12         13         14
##  11.342314  13.140333  15.000000
```



Le graphe montre que le coefficient Cp diminue plus le nombre de variables prises pour le modèle est important et ceci jusqu'à la dixième variable, ce qui signifie que plus l'on prend de variables, plus le modèle est bon jusqu'à la dixième variable. A partir de 10 variables, on observe cependant une légère augmentation du coefficient Cp, et ceci jusqu'à la quatorzième variable, mais pour le modèle à 14 variables, il n'y a qu'un seul modèle. Donc le modèle à 14 variables semble également bon.

Régression AIC et BIC

```
## Start:  AIC=-818.23
## mv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + blacks + lstat + townid
##
##           Df Sum of Sq     RSS     AIC
## - zn       1    0.0051  8.5560 -820.08
## - townid   1    0.0081  8.5590 -820.00
## - indus    1    0.0090  8.5599 -819.97
## - age      1    0.0250  8.5759 -819.50
## - blacks   1    0.0409  8.5919 -819.03
## - chas     1    0.0449  8.5959 -818.92
## <none>                  8.5509 -818.23
## - rm       1    0.1160  8.6669 -816.85
## - tax      1    0.2141  8.7650 -814.02
## - dis      1    0.5464  9.0974 -804.68
## - nox      1    0.6372  9.1881 -802.19
## - ptratio  1    0.8239  9.3748 -797.14
## - rad      1    0.8254  9.3763 -797.10
## - crim     1    2.4155 10.9664 -757.79
## - lstat    1    4.4099 12.9608 -715.84
##
## Step:  AIC=-820.08
## mv ~ crim + indus + chas + nox + rm + age + dis + rad + tax +
```

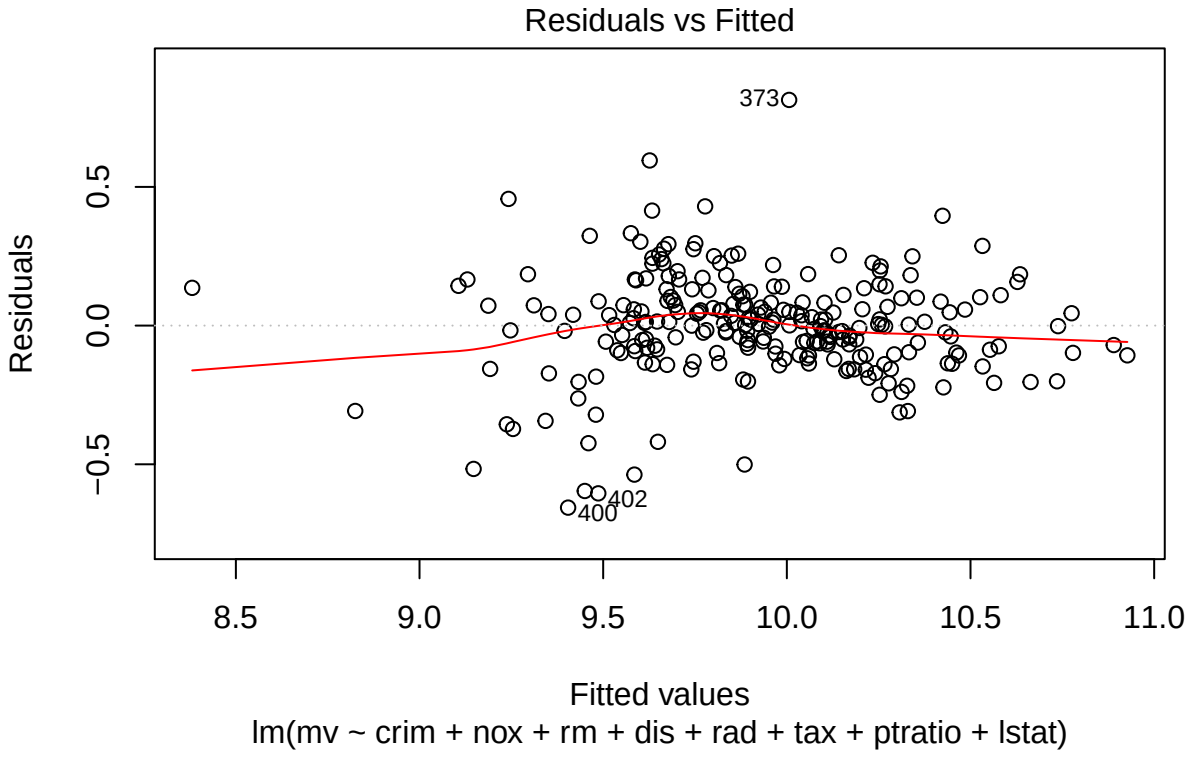
```
##     ptratio + blacks + lstat + townid
##
##           Df Sum of Sq     RSS     AIC
## - indus    1    0.0073  8.5633 -821.87
## - townid   1    0.0080  8.5640 -821.85
## - age      1    0.0224  8.5784 -821.43
## - blacks   1    0.0408  8.5968 -820.89
## - chas     1    0.0447  8.6007 -820.78
## <none>                  8.5560 -820.08
## - rm       1    0.1197  8.6757 -818.60
## - tax      1    0.2132  8.7692 -815.91
## - dis      1    0.5552  9.1112 -806.30
## - nox      1    0.6384  9.1945 -804.02
## - rad      1    0.8360  9.3920 -798.68
## - ptratio  1    0.9235  9.4795 -796.36
## - crim     1    2.4203 10.9763 -759.56
## - lstat    1    4.4547 13.0107 -716.88
##
## Step:  AIC=-821.87
## mv ~ crim + chas + nox + rm + age + dis + rad + tax + ptratio +
##     blacks + lstat + townid
##
##           Df Sum of Sq     RSS     AIC
## - townid   1    0.0105  8.5738 -823.56
## - age      1    0.0222  8.5855 -823.22
## - blacks   1    0.0400  8.6033 -822.70
## - chas     1    0.0524  8.6157 -822.34
## <none>                  8.5633 -821.87
## - rm       1    0.1155  8.6788 -820.51
## - tax      1    0.2230  8.7863 -817.42
## - nox      1    0.6436  9.2069 -805.68
## - dis      1    0.6683  9.2316 -805.01
## - rad      1    0.8391  9.4025 -800.40
## - ptratio  1    0.9269  9.4902 -798.07
## - crim     1    2.5013 11.0646 -759.55
## - lstat    1    4.4489 13.0122 -718.85
##
## Step:  AIC=-823.56
## mv ~ crim + chas + nox + rm + age + dis + rad + tax + ptratio +
##     blacks + lstat
##
##           Df Sum of Sq     RSS     AIC
## - age      1    0.0336  8.6075 -824.58
## - blacks   1    0.0383  8.6121 -824.44
## - chas     1    0.0502  8.6240 -824.10
## <none>                  8.5738 -823.56
## - rm       1    0.1105  8.6843 -822.35
## - tax      1    0.3168  8.8906 -816.45
## - nox      1    0.6414  9.2152 -807.45
## - dis      1    0.6583  9.2322 -806.99
## - rad      1    0.8310  9.4049 -802.34
## - ptratio  1    0.9400  9.5138 -799.45
## - crim     1    2.5212 11.0950 -760.86
## - lstat    1    4.4744 13.0483 -720.16
```

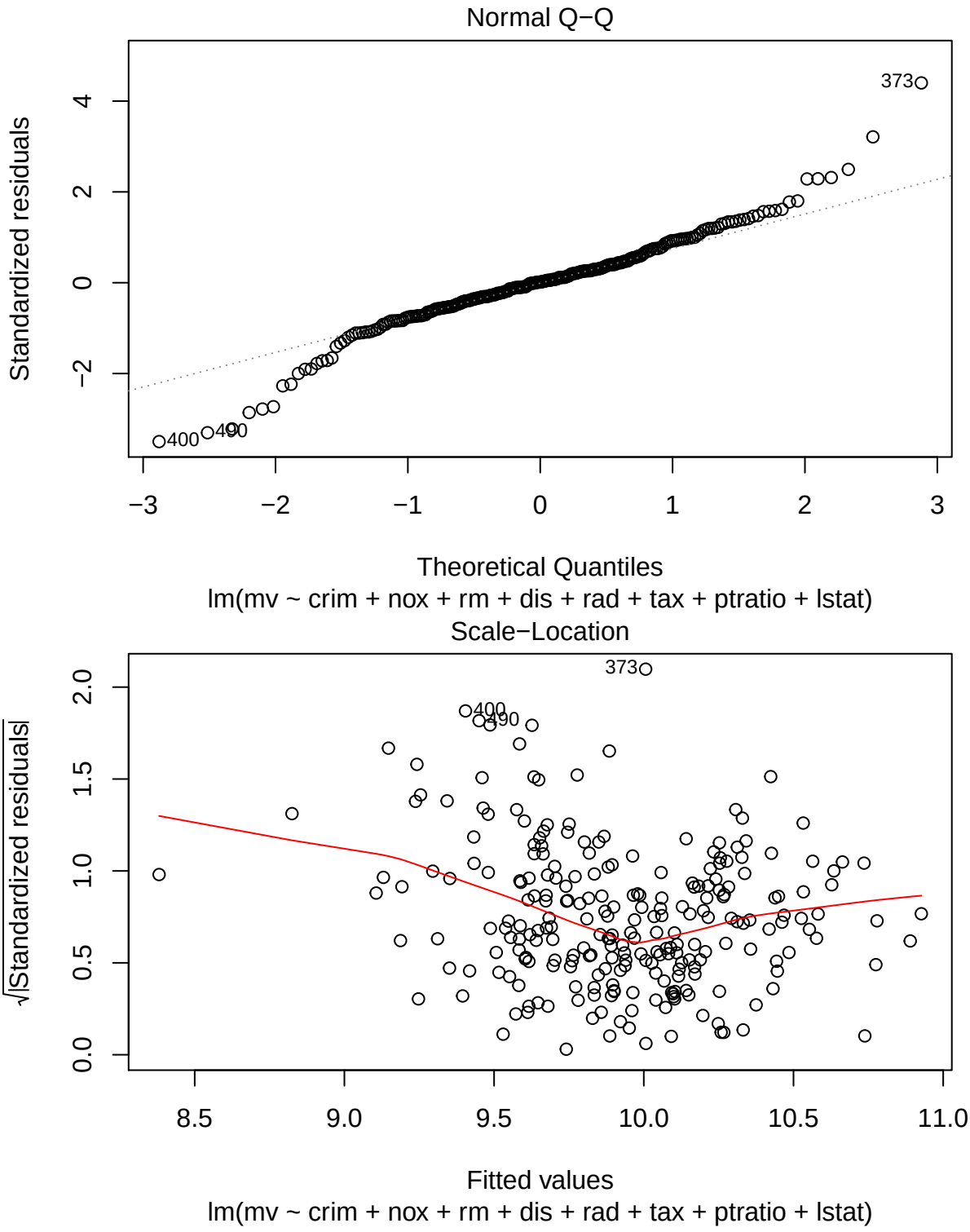
```
##
## Step:  AIC=-824.58
## mv ~ crim + chas + nox + rm + dis + rad + tax + ptratio + blacks +
##     lstat
##
##           Df Sum of Sq     RSS     AIC
## - blacks   1    0.0466  8.6541 -825.22
## - chas     1    0.0499  8.6574 -825.13
## <none>                  8.6075 -824.58
## - rm       1    0.1586  8.7660 -822.00
## - tax      1    0.3283  8.9357 -817.18
## - nox      1    0.6140  9.2215 -809.28
## - rad      1    0.8298  9.4372 -803.48
## - ptratio  1    0.9165  9.5240 -801.18
## - dis      1    1.0859  9.6934 -796.76
## - crim     1    2.5939 11.2013 -760.46
## - lstat    1    5.0448 13.6523 -710.80
##
## Step:  AIC=-825.22
## mv ~ crim + chas + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq     RSS     AIC
## - chas     1    0.0520  8.7061 -825.72
## <none>                  8.6541 -825.22
## - rm       1    0.1436  8.7977 -823.09
## - tax      1    0.3578  9.0119 -817.05
## - nox      1    0.6695  9.3236 -808.52
## - rad      1    0.8189  9.4730 -804.53
## - ptratio  1    0.9005  9.5546 -802.38
## - dis      1    1.1194  9.7735 -796.69
## - crim     1    2.6875 11.3415 -759.34
## - lstat    1    5.2868 13.9409 -707.55
##
## Step:  AIC=-825.72
## mv ~ crim + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq     RSS     AIC
## <none>                  8.7061 -825.72
## - rm       1    0.1620  8.8681 -823.09
## - tax      1    0.4016  9.1076 -816.40
## - nox      1    0.6764  9.3824 -808.94
## - rad      1    0.8893  9.5953 -803.31
## - ptratio  1    0.9242  9.6302 -802.40
## - dis      1    1.1976  9.9037 -795.37
## - crim     1    2.7357 11.4418 -759.13
## - lstat    1    5.3398 14.0459 -707.66

## Start:  AIC=-765.35
## mv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + blacks + lstat + townid
##
##           Df Sum of Sq     RSS     AIC
## - zn       1    0.0051  8.5560 -770.73
## - townid   1    0.0081  8.5590 -770.64
```

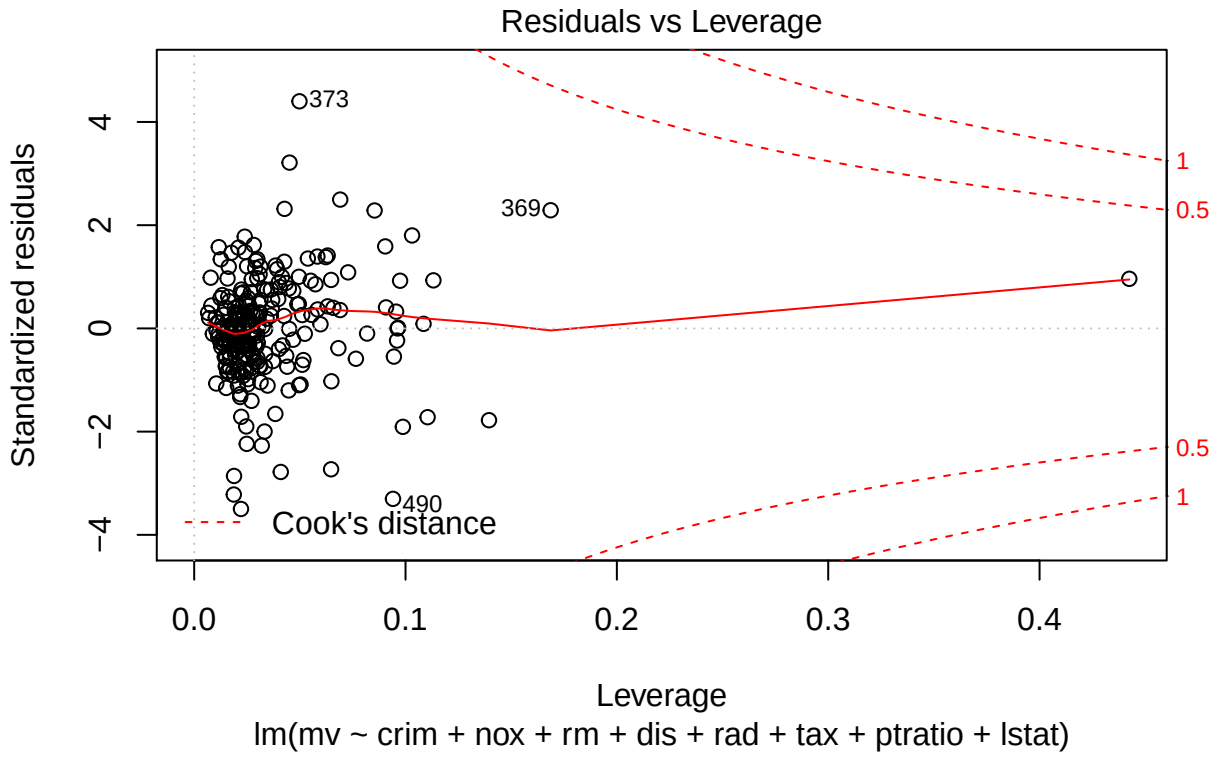
```
## - indus    1    0.0090  8.5599 -770.61
## - age      1    0.0250  8.5759 -770.14
## - blacks   1    0.0409  8.5919 -769.68
## - chas     1    0.0449  8.5959 -769.56
## - rm       1    0.1160  8.6669 -767.50
## <none>                  8.5509 -765.35
## - tax      1    0.2141  8.7650 -764.67
## - dis      1    0.5464  9.0974 -755.33
## - nox      1    0.6372  9.1881 -752.84
## - ptratio  1    0.8239  9.3748 -747.79
## - rad      1    0.8254  9.3763 -747.75
## - crim     1    2.4155 10.9664 -708.43
## - lstat    1    4.4099 12.9608 -666.49
##
## Step:  AIC=-770.73
## mv ~ crim + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + blacks + lstat + townid
##
##           Df Sum of Sq     RSS     AIC
## - indus    1    0.0073  8.5633 -776.04
## - townid   1    0.0080  8.5640 -776.02
## - age      1    0.0224  8.5784 -775.60
## - blacks   1    0.0408  8.5968 -775.06
## - chas     1    0.0447  8.6007 -774.95
## - rm       1    0.1197  8.6757 -772.76
## <none>                  8.5560 -770.73
## - tax      1    0.2132  8.7692 -770.08
## - dis      1    0.5552  9.1112 -760.47
## - nox      1    0.6384  9.1945 -758.19
## - rad      1    0.8360  9.3920 -752.85
## - ptratio  1    0.9235  9.4795 -750.53
## - crim     1    2.4203 10.9763 -713.73
## - lstat    1    4.4547 13.0107 -671.05
##
## Step:  AIC=-776.04
## mv ~ crim + chas + nox + rm + age + dis + rad + tax + ptratio +
##     blacks + lstat + townid
##
##           Df Sum of Sq     RSS     AIC
## - townid   1    0.0105  8.5738 -781.26
## - age      1    0.0222  8.5855 -780.91
## - blacks   1    0.0400  8.6033 -780.40
## - chas     1    0.0524  8.6157 -780.03
## - rm       1    0.1155  8.6788 -778.20
## <none>                  8.5633 -776.04
## - tax      1    0.2230  8.7863 -775.11
## - nox      1    0.6436  9.2069 -763.38
## - dis      1    0.6683  9.2316 -762.70
## - rad      1    0.8391  9.4025 -758.10
## - ptratio  1    0.9269  9.4902 -755.77
## - crim     1    2.5013 11.0646 -717.24
## - lstat    1    4.4489 13.0122 -676.54
##
## Step:  AIC=-781.26
```

```
## mv ~ crim + chas + nox + rm + age + dis + rad + tax + ptratio +
##     blacks + lstat
##
##           Df Sum of Sq     RSS     AIC
## - age      1    0.0336  8.6075 -785.80
## - blacks   1    0.0383  8.6121 -785.66
## - chas     1    0.0502  8.6240 -785.32
## - rm       1    0.1105  8.6843 -783.57
## <none>                  8.5738 -781.26
## - tax      1    0.3168  8.8906 -777.67
## - nox      1    0.6414  9.2152 -768.67
## - dis      1    0.6583  9.2322 -768.21
## - rad      1    0.8310  9.4049 -763.56
## - ptratio  1    0.9400  9.5138 -760.67
## - crim     1    2.5212 11.0950 -722.08
## - lstat    1    4.4744 13.0483 -681.38
##
## Step:  AIC=-785.8
## mv ~ crim + chas + nox + rm + dis + rad + tax + ptratio + blacks +
##     lstat
##
##           Df Sum of Sq     RSS     AIC
## - blacks   1    0.0466  8.6541 -789.97
## - chas     1    0.0499  8.6574 -789.87
## - rm       1    0.1586  8.7660 -786.74
## <none>                  8.6075 -785.80
## - tax      1    0.3283  8.9357 -781.93
## - nox      1    0.6140  9.2215 -774.03
## - rad      1    0.8298  9.4372 -768.22
## - ptratio  1    0.9165  9.5240 -765.93
## - dis      1    1.0859  9.6934 -761.50
## - crim     1    2.5939 11.2013 -725.21
## - lstat    1    5.0448 13.6523 -675.54
##
## Step:  AIC=-789.97
## mv ~ crim + chas + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq     RSS     AIC
## - chas     1    0.0520  8.7061 -793.99
## - rm       1    0.1436  8.7977 -791.36
## <none>                  8.6541 -789.97
## - tax      1    0.3578  9.0119 -785.32
## - nox      1    0.6695  9.3236 -776.79
## - rad      1    0.8189  9.4730 -772.80
## - ptratio  1    0.9005  9.5546 -770.65
## - dis      1    1.1194  9.7735 -764.96
## - crim     1    2.6875 11.3415 -727.61
## - lstat    1    5.2868 13.9409 -675.82
##
## Step:  AIC=-793.99
## mv ~ crim + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq     RSS     AIC
## - rm       1    0.1620  8.8681 -794.89
```

```
## <none>                  8.7061 -793.99
## - tax      1    0.4016  9.1076 -788.20
## - nox      1    0.6764  9.3824 -780.74
## - rad      1    0.8893  9.5953 -775.10
## - ptratio  1    0.9242  9.6302 -774.19
## - dis      1    1.1976  9.9037 -767.17
## - crim     1    2.7357 11.4418 -730.93
## - lstat    1    5.3398 14.0459 -679.46
##
## Step:  AIC=-794.89
## mv ~ crim + nox + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq     RSS     AIC
## <none>                  8.8681 -794.89
## - tax      1    0.4139  9.2820 -788.96
## - nox      1    0.6739  9.5420 -782.03
## - rad      1    1.0066  9.8746 -773.43
## - ptratio  1    1.0748  9.9429 -771.70
## - dis      1    1.2896 10.1577 -766.33
## - crim     1    2.8038 11.6719 -731.46
## - lstat    1    9.2467 18.1148 -621.13

##       Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1          NA          NA       236   8.550922 -818.2329
## 2     - zn  1 0.005084652       237   8.556007 -820.0836
## 3  - indus  1 0.007318313       238   8.563325 -821.8690
## 4 - townid  1 0.010501694       239   8.573827 -823.5614
## 5    - age  1 0.033631152       240   8.607458 -824.5788
## 6 - blacks  1 0.046625953       241   8.654084 -825.2228
## 7   - chas  1 0.051977180       242   8.706061 -825.7198
```

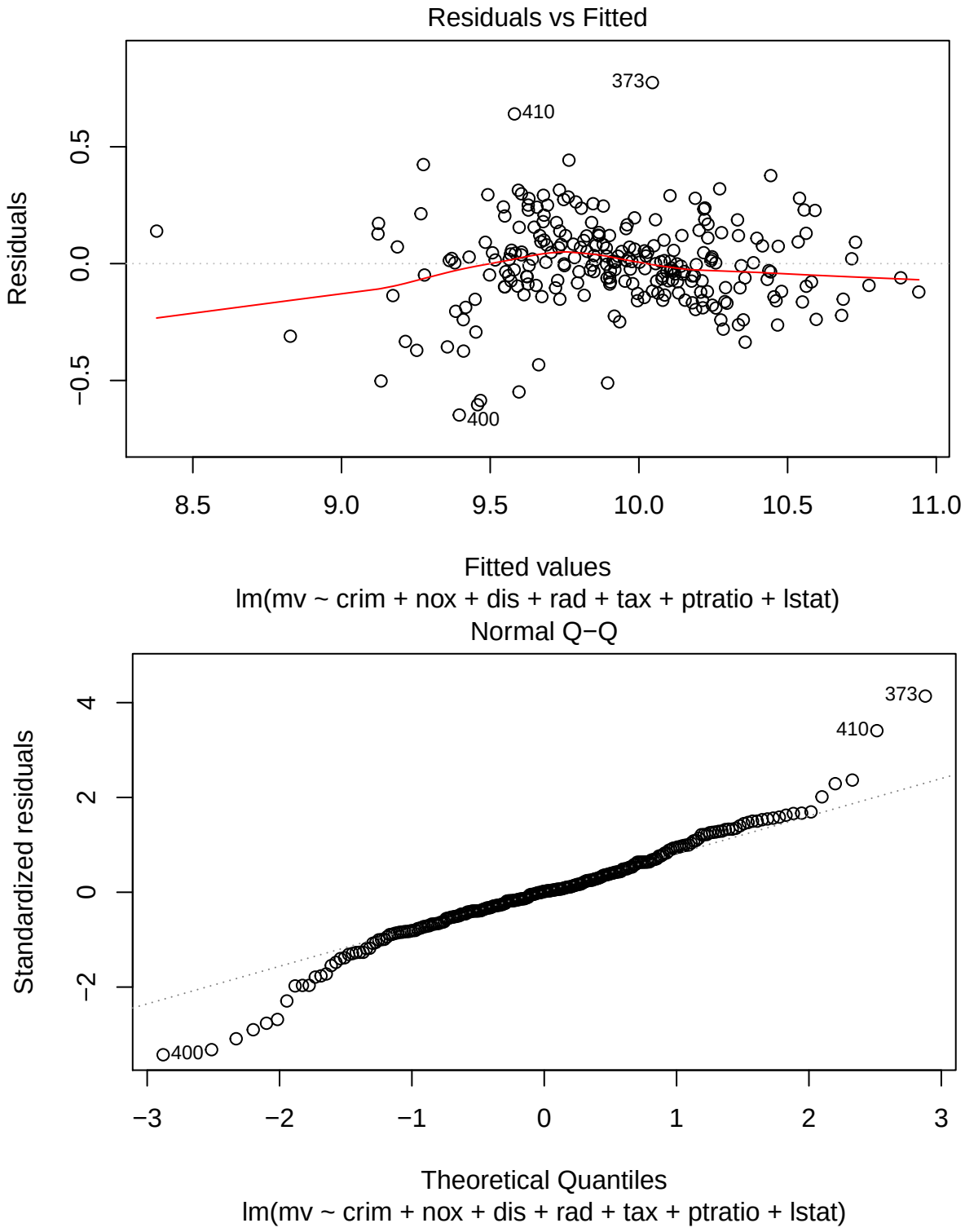


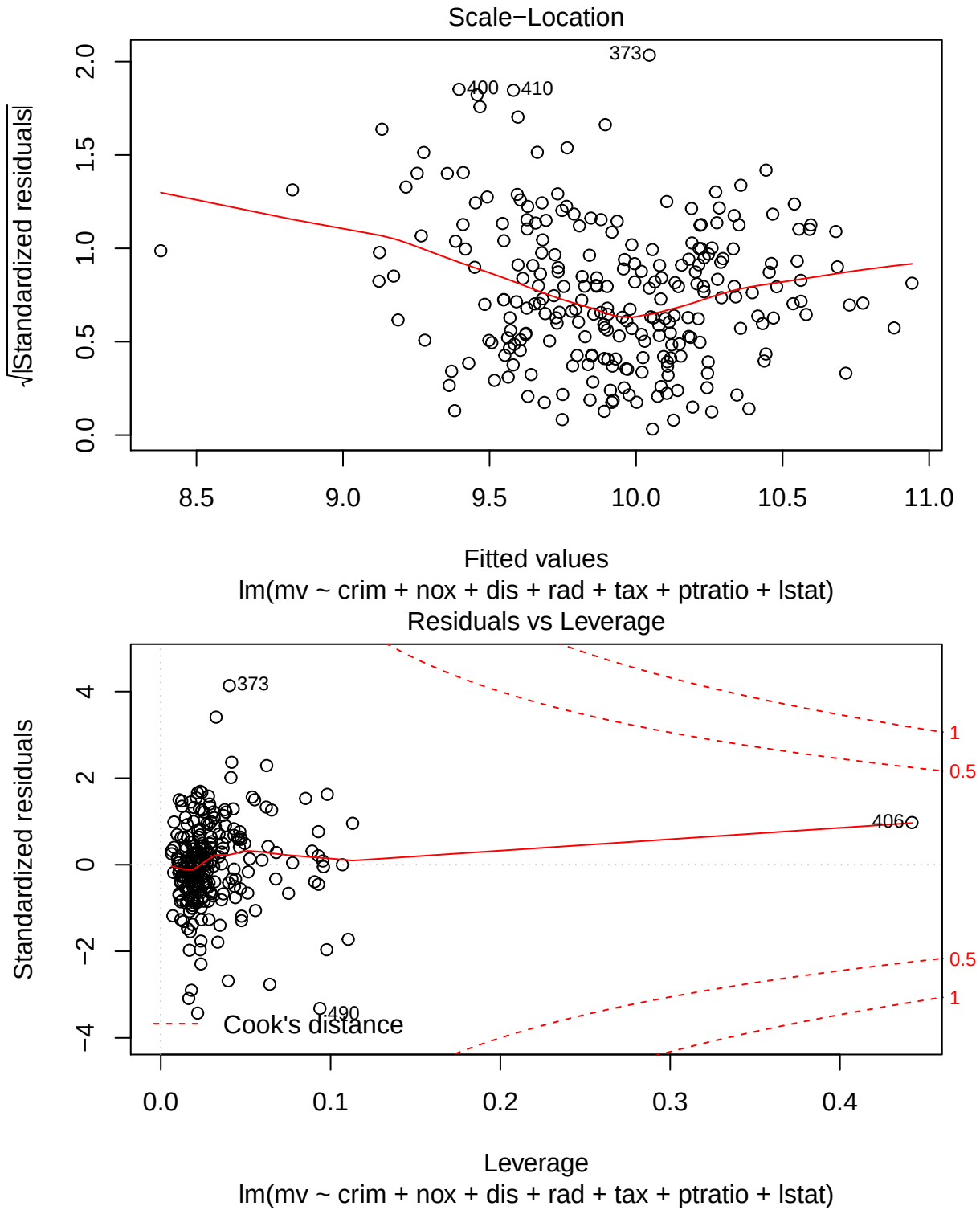

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mv ~ crim + nox + rm + dis + rad + tax + ptratio + lstat)
373
400
490
Scale−Location
√|Standardized residuals|
Fitted values
lm(mv ~ crim + nox + rm + dis + rad + tax + ptratio + lstat)



En utilisant le critère AIC, on trouve qu'il est plus précis d'utiliser 5 variables pour avoir le meilleur modèle, les variables "age", "indus", "zn", "townid" et "chas".

```
##       Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1           NA          NA       236   8.550922 -765.3511
## 2     - zn  1 0.005084652       237   8.556007 -770.7273
## 3  - indus  1 0.007318313       238   8.563325 -776.0382
## 4 - townid  1 0.010501694       239   8.573827 -781.2560
## 5    - age  1 0.033631152       240   8.607458 -785.7988
## 6 - blacks  1 0.046625953       241   8.654084 -789.9683
## 7   - chas  1 0.051977180       242   8.706061 -793.9907
## 8     - rm  1 0.162024000       243   8.868085 -794.8879
```


Residuals vs Fitted
373
410
400
Residuals
0.5
0.0
−0.5
8.5
9.0
9.5
10.0
10.5
11.0
Fitted values
lm(mv ~ crim + nox + dis + rad + tax + ptratio + lstat)
Normal Q−Q
373
410
400
Standardized residuals
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mv ~ crim + nox + dis + rad + tax + ptratio + lstat)

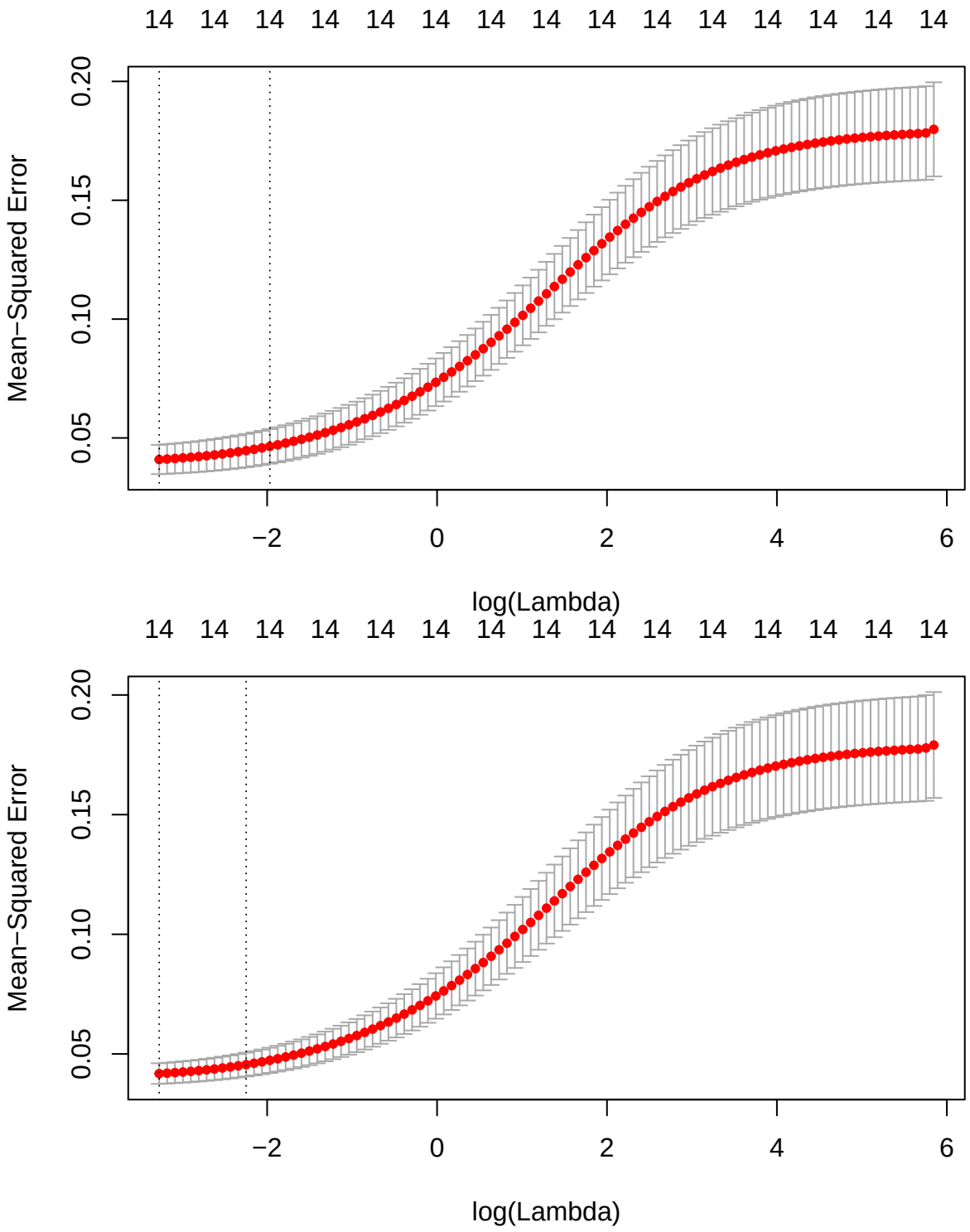


L'utilisation du critère BIC montre que le meilleur modèle est celui à 6 variables en ajoutant la variable "blacks" par rapport au modele AIC.

Régression ridge

A présent, utilisons la validation croisée pour trouver une valeur de lambda. Nous utiliserons un fold de 10 puis un fold qui sera le nombre d'individus contenus dans le tableau.

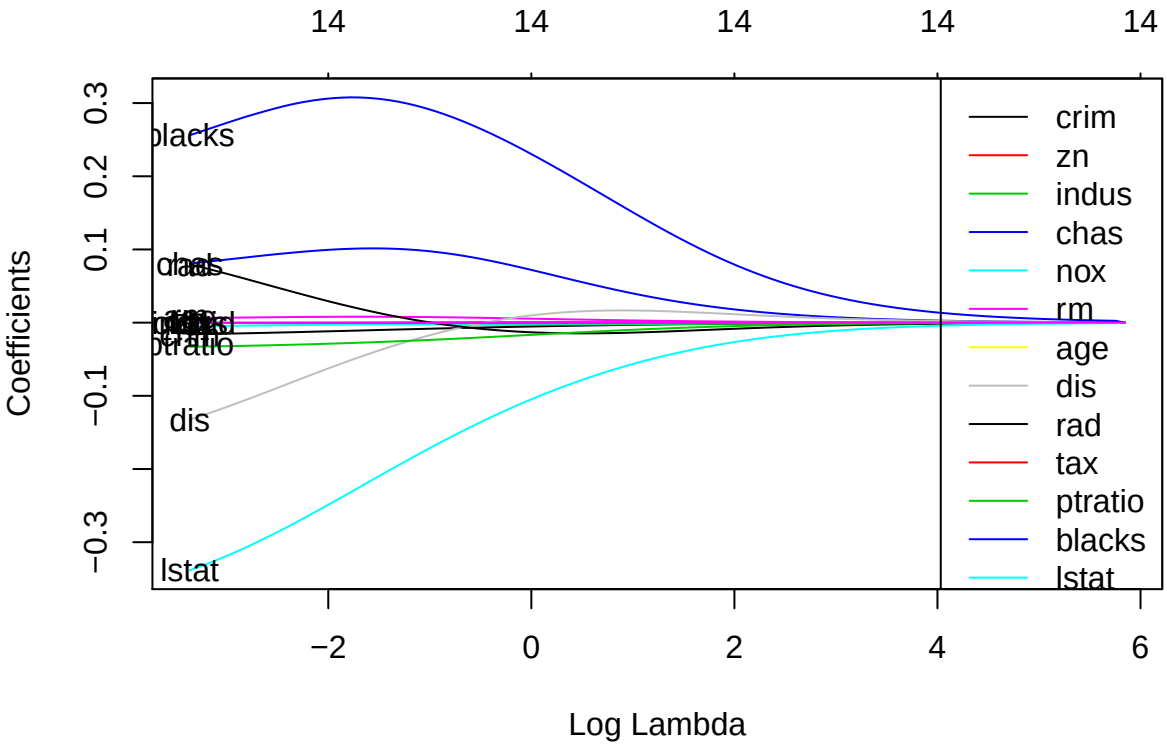
```
## Loading required package: Matrix
## Loading required package: foreach
## Loaded glmnet 2.0-2
```



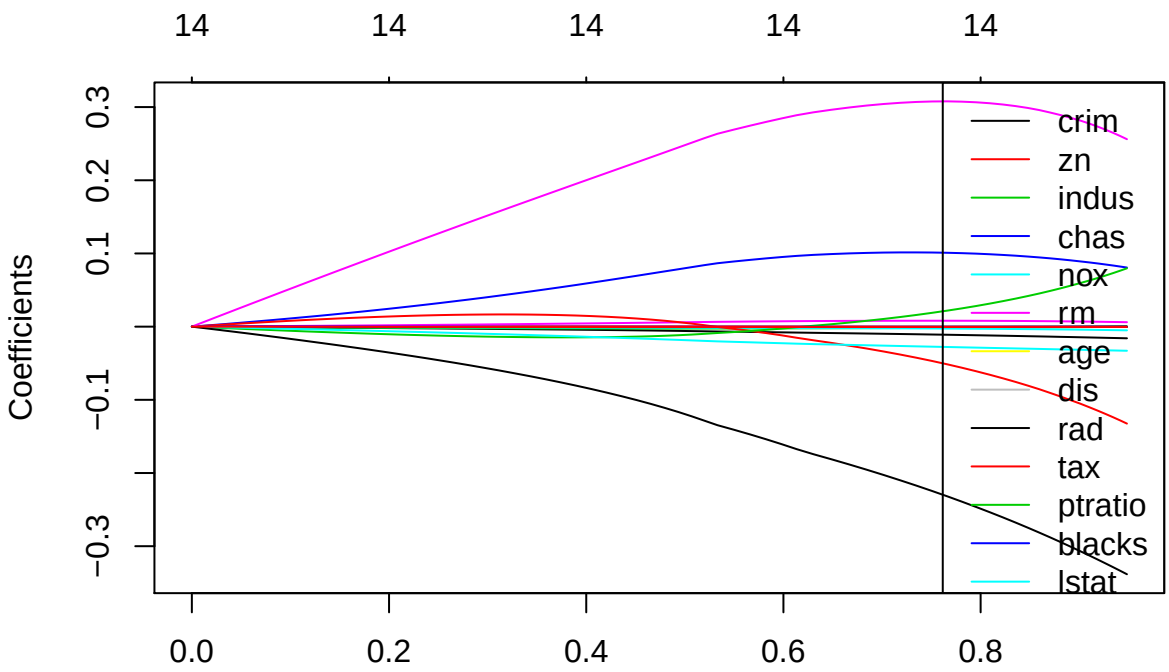
Le modèle du ridge calculé par la règle du minimum et celui calculé par la règle du "1 standard error" semblent quasiment identiques, la courbe d'évolution de l'erreur en fonction de lambda étant identiques pour les deux.

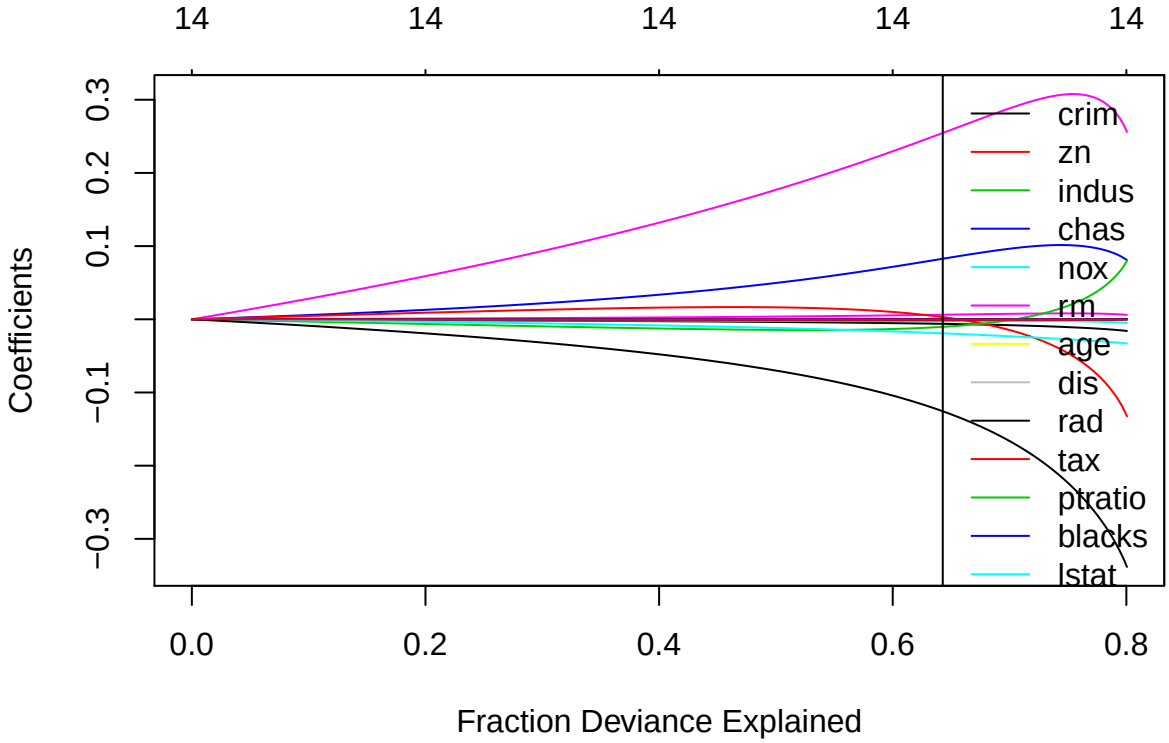
```
## [1] -3.270548
```

```
## [1] -1.968076
```



En augmentant la valeur de Lambda pour régulariser le modèle, nous pouvons voir que les paramètres qui convergent le moins rapidement vers 0 sont les variables “chas”, “lstat”, “blacks”, puis “dis”. Ces paramètres sont donc les meilleurs prédicteurs du modèle. Les paramètres “rad” et “indus” convergent également mois rapidement que tout les autres paramètres qui, eux, semblent convergent très rapidement vers 0 et sont donc les moins importants pour la prédiction et l’explication du paramètre “mv”.





Les graphes utilisant comme terme de régularisation la norme L1 et la fraction de déviance mettent en valeur la variable “rm” comme un paramètre également important pour la prédiction et l’explication de la variable “mv”.

```
##           1
## 1 10.256858
## 3 10.372658
```

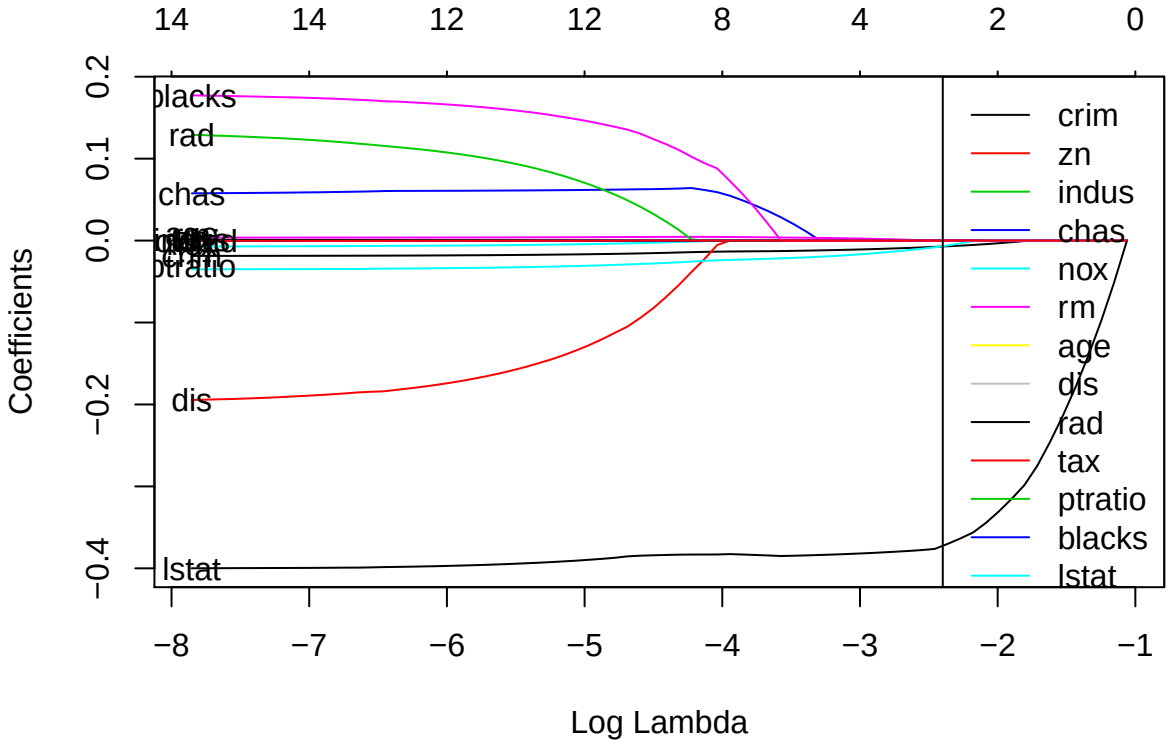
```
## 6 10.210320
## 7  9.993877
## 8  9.857658
```

```
##           1
## 1 10.273539
## 3 10.360692
## 6 10.199480
## 7 10.007150
## 8  9.896296
```

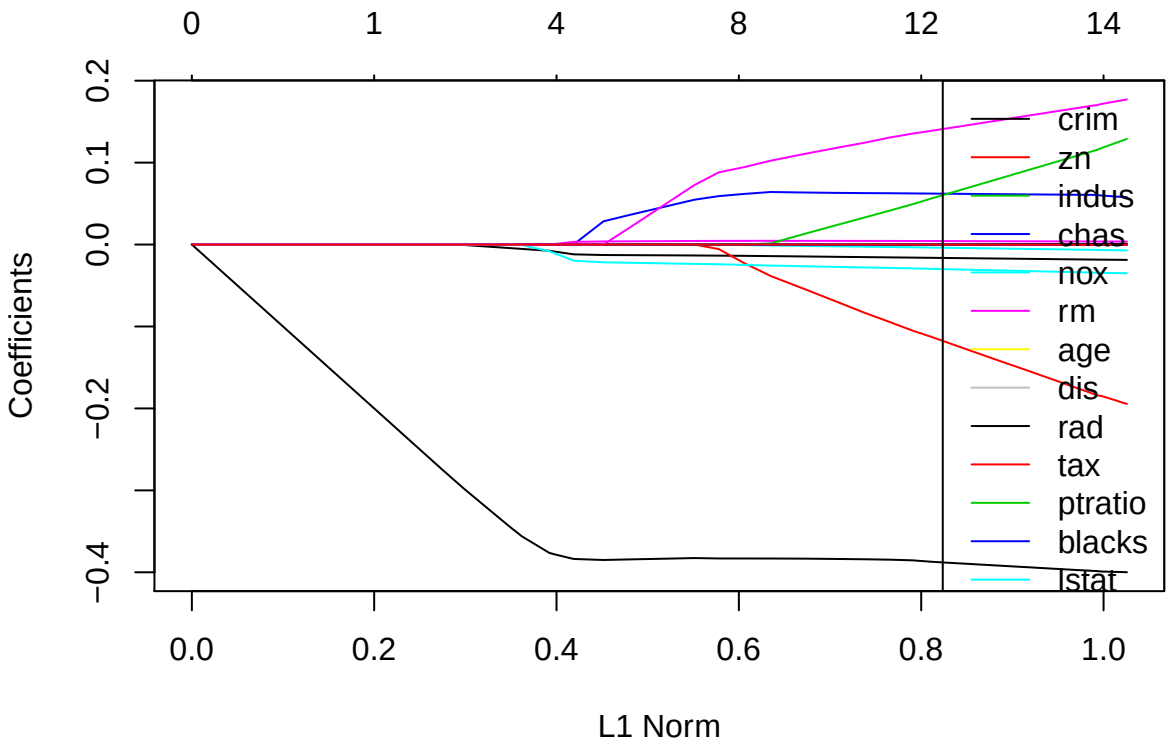
Nous avons ainsi un modele predictif trouvé en utilisant la regression de ridge.

Régression Lasso

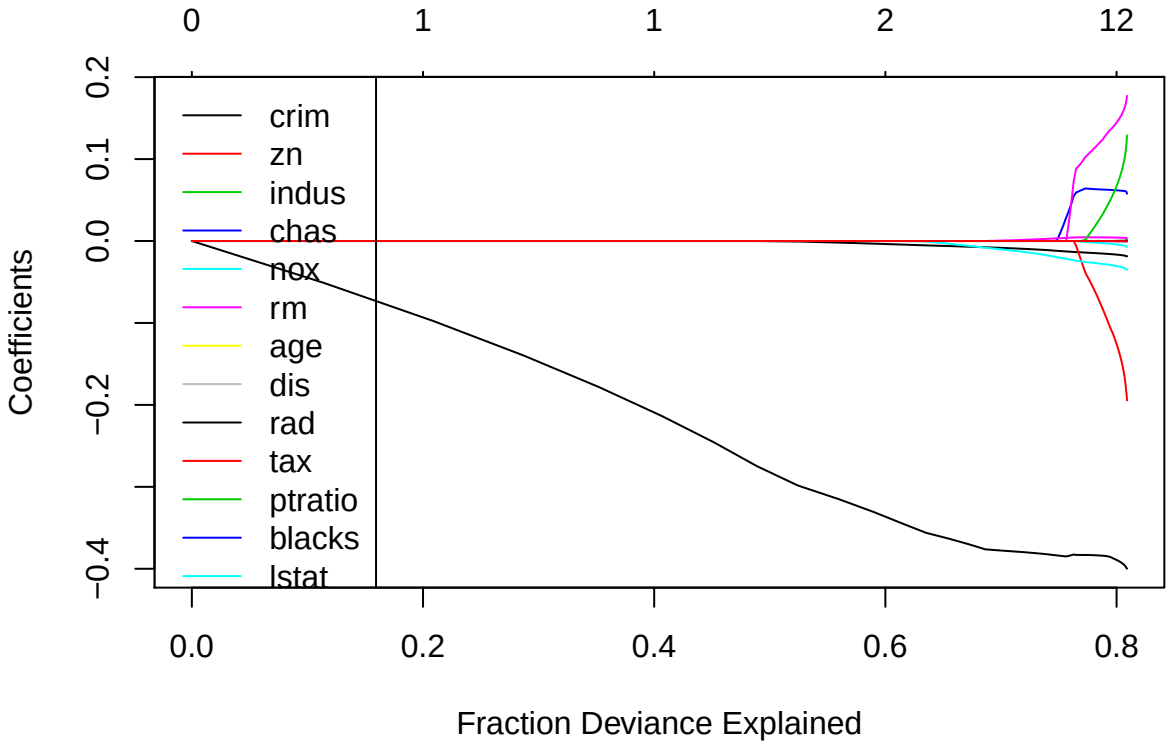
A présent, nous utilisons la régression Lasso pour avoir un choix de prédicteurs plus significatif. </>



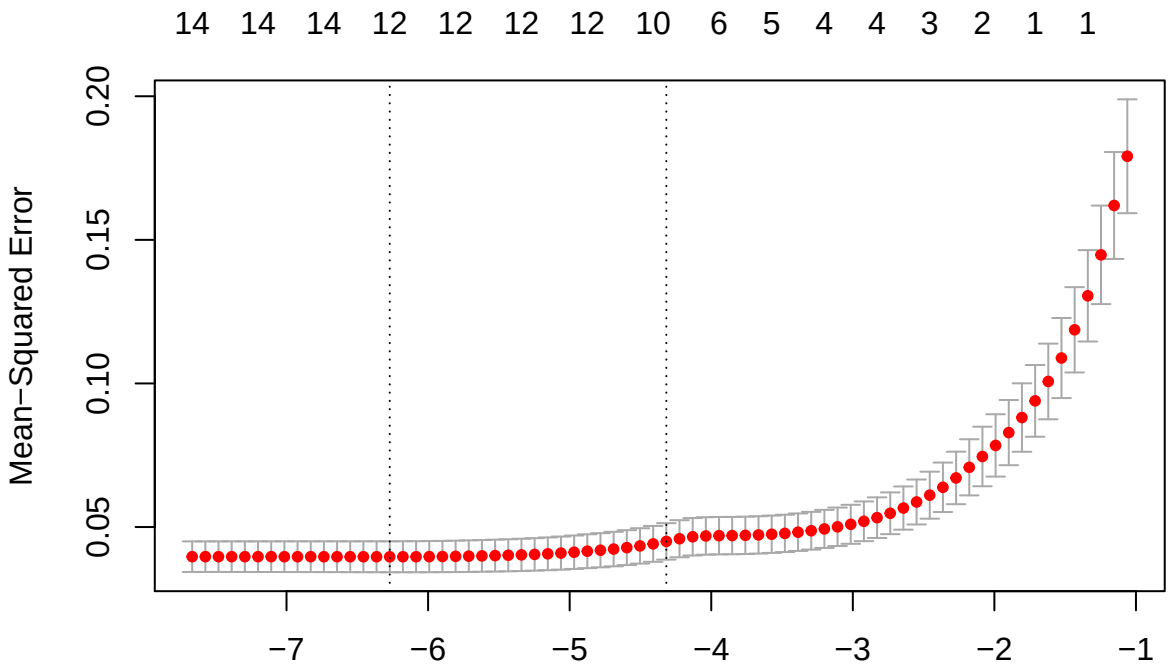
A présent, nous utilisons la méthode Lasso pour régulariser le modèle. Tout d'abord, ce premier graphe sur lequel on voit l'évolution des coefficients de chaque variable en fonction de Lambda nous permet de voir quels sont les variables qui semblent plus efficace pour servir de prédicteur : les variables dont les coefficients sont rapidement réduits à zéro sont les plus mauvais prédicteurs. Ainsi les variables dont les coefficients sont réduits à zéro avec une valeur de lambda plus importante que les autres sont les meilleurs prédicteurs. Grâce à ce graphe, nous pouvons donc observer que les variables "rad", "townid", "chas", "tax", "indus" et "nox" semblent être les plus à même de prédire le modèle.

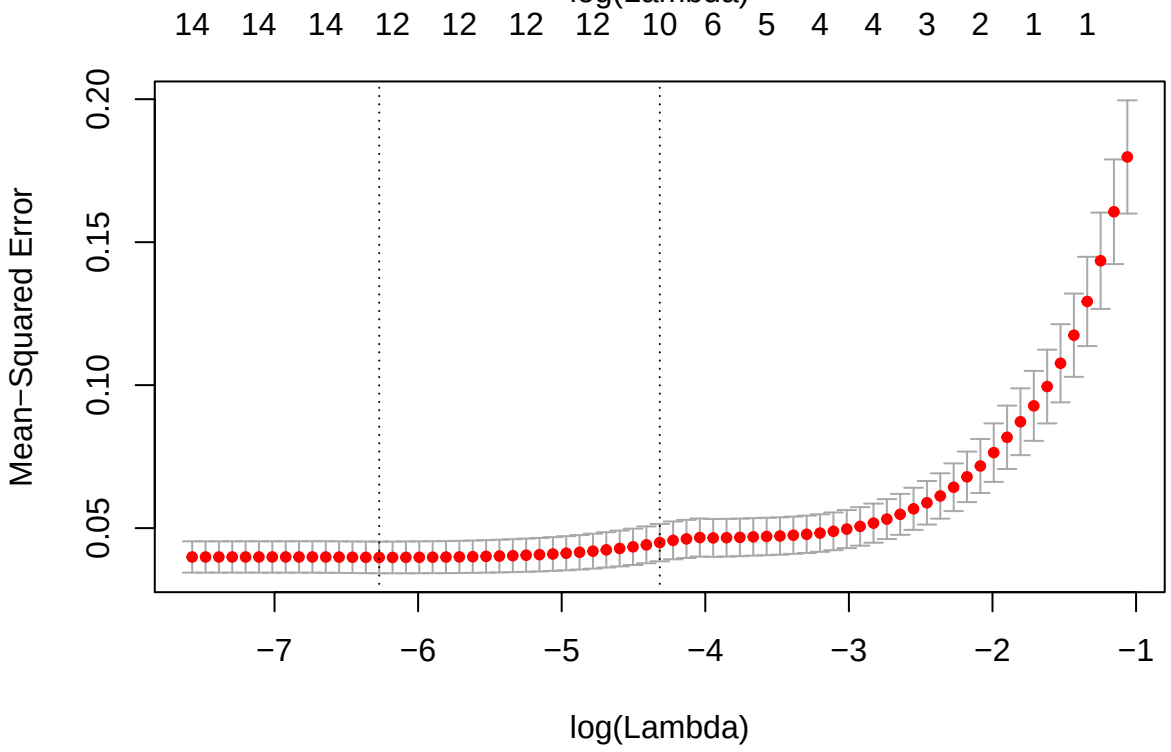


Ce second graphe nous amène aux mêmes conclusions que le dernier, à la différence qu'ici le terme de régularisation utilisé est la norme L1.



Le raisonnement et les conclusions sont identiques pour ce troisième graphe. Ici, le terme de régularisation utilisé est la fraction de déviance

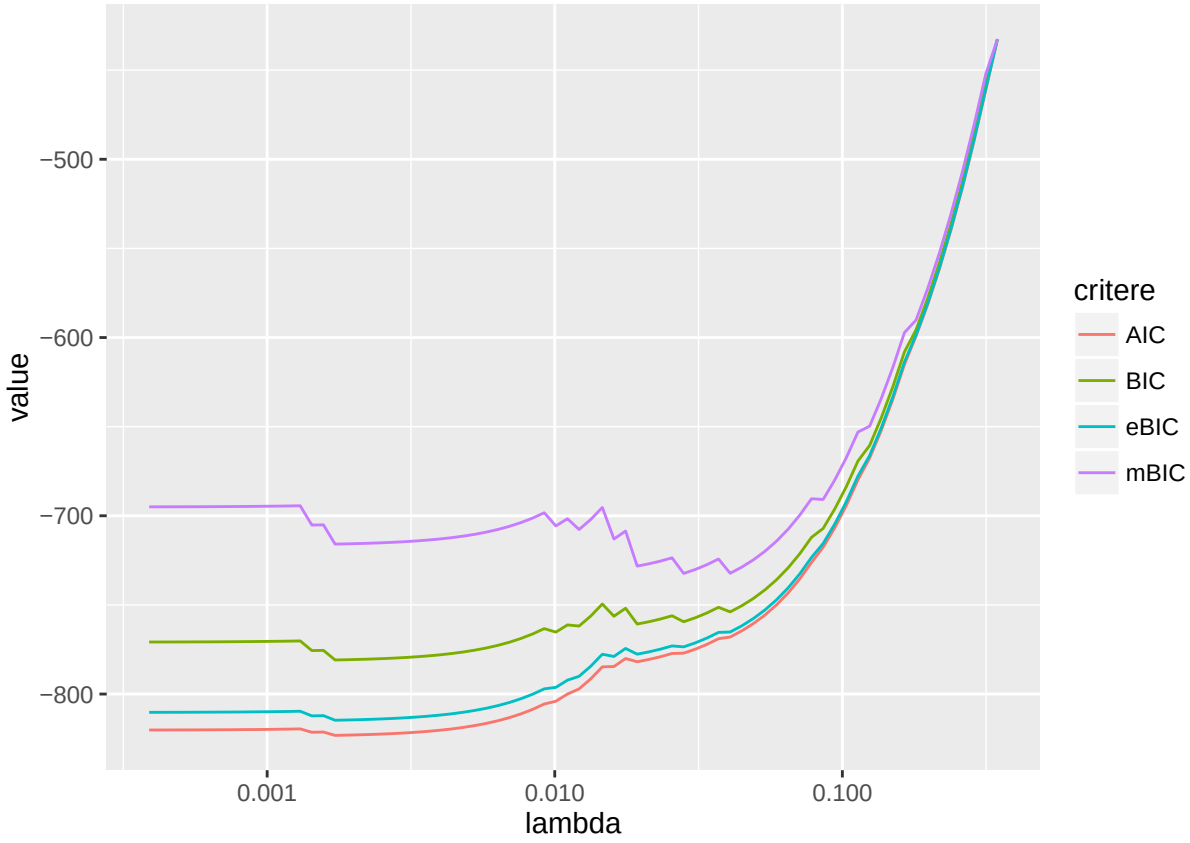





```
##           1
## 1 10.235644
## 3 10.370328
## 6 10.221729
## 7  9.975711
## 8  9.812819
```

```
##           1
## 1 10.337581
## 3 10.401957
## 6 10.230882
## 7  9.962082
## 8  9.800869
```

Voici donc le modele trouvé a partir de la regression Lasso.



Dans le modele lasso, on peut voir que AIC, BIC, eBIC et mBIC se superposent apres que lambda soit egale a 0.1.

Avant la superposition on peut voir que mBIC a les plus grandes valeurs donc il est meilleur que les autres methodes

évaluation de la qualité des modèles

```
##   (Intercept)          crim           nox            rm           dis
## 10.0651124789 -0.0195147970 -0.0072492777  0.0040496657 -0.2279671996
##           rad           tax       ptratio         lstat
##  0.1316715839 -0.0005254904 -0.0350612964 -0.3939180571
```

```
##   (Intercept)          crim           nox           dis           rad
## 10.1826554224 -0.0197348846 -0.0072358044 -0.2355837355  0.1389165933
##           tax       ptratio         lstat
## -0.0005333613 -0.0373487076 -0.4318681852
```

```
##   (Intercept)          crim           nox            rm           dis
## 10.0651124789 -0.0195147970 -0.0072492777  0.0040496657 -0.2279671996
##           rad           tax       ptratio         lstat
##  0.1316715839 -0.0005254904 -0.0350612964 -0.3939180571
```

```
##
## Call:
## lm(formula = mv ~ ., data = Hedonic.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63949 -0.08836  0.00114  0.08761  0.76451
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.8431223  0.2330058  42.244  < 2e-16 ***
## crim        -0.0188267  0.0023058  -8.165 1.94e-14 ***
## zn           0.0002866  0.0007650   0.375 0.708286
## indus        0.0018116  0.0036300   0.499 0.618197
## chas         0.0564167  0.0506527   1.114 0.266500
## nox         -0.0073795  0.0017597  -4.193 3.89e-05 ***
## rm           0.0036322  0.0020302   1.789 0.074884 .
## age          0.0006593  0.0007937   0.831 0.407044
## dis         -0.1983858  0.0510844  -3.883 0.000134 ***
## rad          0.1340437  0.0280841   4.773 3.18e-06 ***
## tax         -0.0004960  0.0002040  -2.431 0.015803 *
## ptratio     -0.0352556  0.0073935  -4.768 3.25e-06 ***
## blacks       0.1796718  0.1690400   1.063 0.288916
## lstat       -0.4002652  0.0362814 -11.032  < 2e-16 ***
## townid      -0.0003104  0.0006563  -0.473 0.636712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1903 on 236 degrees of freedom
## Multiple R-squared:  0.809,  Adjusted R-squared:  0.7977
## F-statistic: 71.42 on 14 and 236 DF,  p-value: < 2.2e-16
```

```
##         modele     erreur
## 1         null 0.15580831
## 2         full 0.03780815
## 3     step.AIC 0.03916586
## 4     step.BIC 0.04148585
## 5  ridge.CVmin 0.03395566
## 6  ridge.CV1se 0.03395566
## 7  lasso.CVmin 0.03676067
## 8  lasso.CV1se 0.03676067
## 9    lasso.BIC 0.03683542
## 10  lasso.mBIC 0.03889894
```

Dans aic et bic on a utiliser 9 vriables donc on peut negliger les autre et dans full on peut voir que ces variable sont negligeables

Le tableau ci dessus indique que le meilleur modele est celui du methode full avec une erreur egale a 0.03814 mais AIC et BIC sont legerement moin bon mais ils utilisent 5 variables en moin.

Maintenant nous allons chercher un modele plus pertinent.

Modele de base polynomiale et de splines

```
##         modele     erreur
## 1         null 0.15580831
## 2         full 0.03780815
## 3     step.AIC 0.03916586
## 4     step.BIC 0.04148585
## 5  ridge.CVmin 0.03395566
## 6  ridge.CV1se 0.03395566
## 7  lasso.CVmin 0.03676067
## 8  lasso.CV1se 0.03676067
## 9    lasso.BIC 0.03683542
## 10  lasso.mBIC 0.03889894
## 11        poly 0.05009442
```

```
step.AIC.poly2 <-step(lm(mv~.,Hedonic2),k=2, direction="both")
```

```
step.BIC.poly2 <-step(lm(mv~.,Hedonic2),k=log(nrow(Hedonic2)))
```

On eu dans ce modele une

Troisieme Partie:Analyse du jeu de données et comparaison de méthodes

```
##         modele     erreur
## 1         null 0.15580831
## 2         full 0.03780815
## 3     step.AIC 0.03916586
## 4     step.BIC 0.04148585
## 5  ridge.CVmin 0.03395566
## 6  ridge.CV1se 0.03395566
## 7  lasso.CVmin 0.03676067
## 8  lasso.CV1se 0.03676067
## 9    lasso.BIC 0.03683542
## 10  lasso.mBIC 0.03889894
## 11        poly 0.05009442
```

Si on regarde le tableau de tous les modeles on peut voir que le seul modele qui est meilleur de full c'est le modele polynomiale

Et si on compare le r^2 ajusté des modele on peut voir que celui du modele polynomiale a la valeur la plus grande. Donc qu'il est le meilleur modele malgré son grand nombre de valeur