TP STATISTIQUE EN GRANDE DIMENSION

AGBLODOE Komi/ M2 SSD

7 janvier 2020

TP2

2 - Sélection de modèle

2-1 Les données

Ces données proviennent d'une étude qui a examiné la corrélation entre le niveau d'antigène spécifique de la prostate et un certain nombre de mesures cliniques chez les hommes sur le point de subir une prostatectomie radicale.

Téléchargement des données

```
## R Package to solve regression problems while imposing
##   an L1 constraint on the parameters. Based on S-plus Release 2.1
## Copyright (C) 1998, 1999
## Justin Lokhorst   <jlokhors@stats.adelaide.edu.au>
## Berwin A. Turlach <bturlach@stats.adelaide.edu.au>
## Bill Venables     <wvenable@stats.adelaide.edu.au>
##
## Copyright (C) 2002
## Martin Maechler <maechler@stat.math.ethz.ch>
```

Nous vérifions que les données correspondent bien au tableau de description et faisons quelques statistiques descriptives : dim, names, summary, cor, hist

```
##      lcavol           lweight           age             lbph
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##       svi              lcp             gleason          pgg45
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
##  3rd Qu.:0.0000   3rd Qu.: 1.1787   3rd Qu.:7.000   3rd Qu.: 40.00
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
##       lpsa
```

```
##  Min.   :-0.4308
##  1st Qu.: 1.7317
##  Median : 2.5915
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

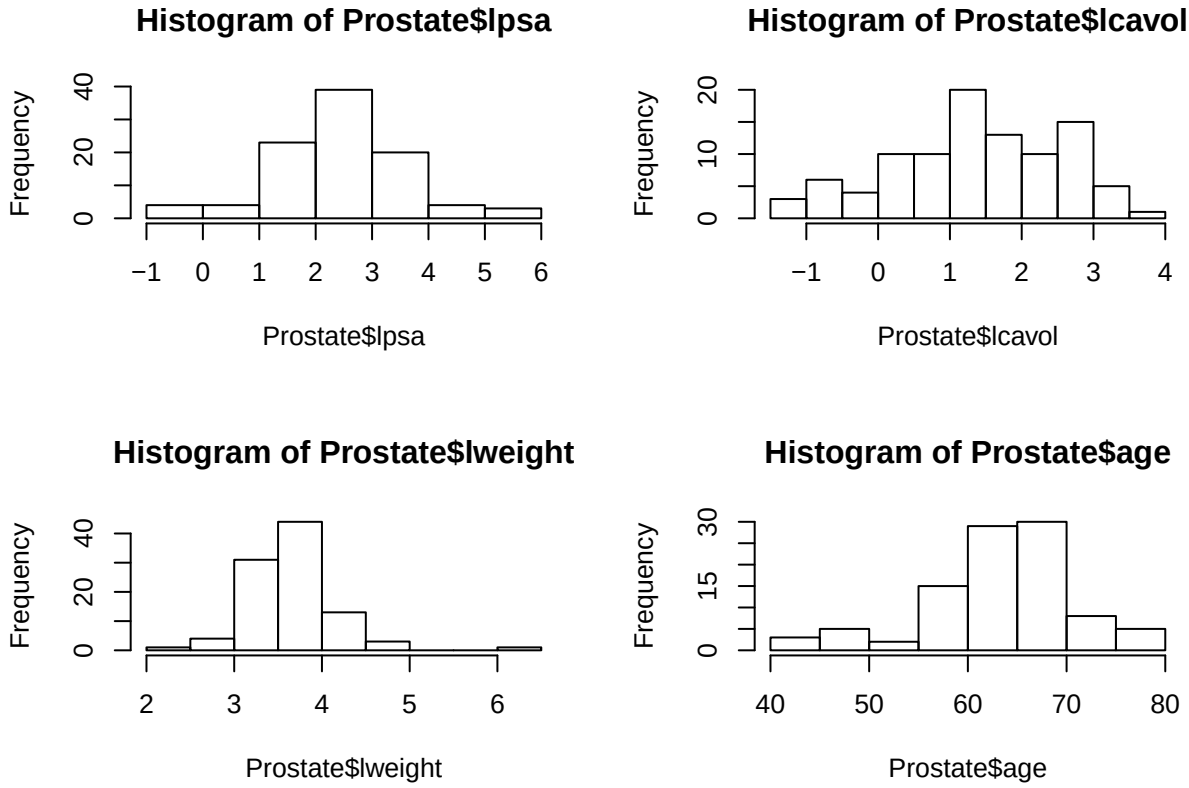
Les variables svi et gleason sont qualitatives.

```
## [1] 97  9
```

```
## [1] "lcavol"  "lweight" "age"     "lbph"    "svi"     "lcp"     "gleason"
## [8] "pgg45"   "lpsa"
```

```
##      lcavol           lweight           age              lbph         svi
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863   0:76
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863   1:21
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##       lcp          gleason     pgg45            lpsa
##  Min.   :-1.3863   6:35    Min.   :  0.00   Min.   :-0.4308
##  1st Qu.:-1.3863   7:56    1st Qu.:  0.00   1st Qu.: 1.7317
##  Median :-0.7985   8: 1    Median : 15.00   Median : 2.5915
##  Mean   :-0.1794   9: 5    Mean   : 24.38   Mean   : 2.4784
##  3rd Qu.: 1.1787           3rd Qu.: 40.00   3rd Qu.: 3.0564
##  Max.   : 2.9042           Max.   :100.00   Max.   : 5.5829
```

```
##            lcavol    lweight       age         lbph          lcp      pgg45
## lcavol  1.0000000 0.19412829 0.2249999  0.027349703  0.675310484 0.43365225
## lweight 0.1941283 1.00000000 0.3075286  0.434934636  0.100237795 0.05084682
## age     0.2249999 0.30752861 1.0000000  0.350185896  0.127667752 0.27611245
## lbph    0.0273497 0.43493464 0.3501859  1.000000000 -0.006999431 0.07846002
## lcp     0.6753105 0.10023780 0.1276678 -0.006999431  1.000000000 0.63152825
## pgg45   0.4336522 0.05084682 0.2761124  0.078460018  0.631528245 1.00000000
## lpsa    0.7344603 0.35412039 0.1695928  0.179809410  0.548813169 0.42231586
##              lpsa
## lcavol  0.7344603
## lweight 0.3541204
## age     0.1695928
## lbph    0.1798094
## lcp     0.5488132
## pgg45   0.4223159
## lpsa    1.0000000
```



Il s'agit d'une matrice de données comportant 97 lignes et 9 colonnes. On remarque aussi que la distribution des variables lpsa, lcavol, weight et âge semblent suivre une loi normale.

Données train et test

Les valeurs de lpsa sont rangées par ordre croissant. Nous divisons ici le jeu de données en un échantillon d'apprentissage pour estimer les modèles et en un échantillon de test pour comparer les erreurs de prédiction. On conserve 1/4 des données pour l'échantillon de test.

On donne quelques statistiques descriptives des données : Prostate.app et Prostate.test

```
## [1] 22  9
```

```
## [1] 75  9
```

```
##      lcavol           lweight           age              lbph         svi
##  Min.   :-1.0498   Min.   :2.375   Min.   :41.00   Min.   :-1.3863   0:57
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863   1:18
##  Median : 1.4586   Median :3.593   Median :65.00   Median :-1.3863
##  Mean   : 1.3969   Mean   :3.607   Mean   :63.63   Mean   :-0.1226
##  3rd Qu.: 2.3491   3rd Qu.:3.858   3rd Qu.:68.50   3rd Qu.: 1.3737
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   : 2.3263
##       lcp          gleason     pgg45            lpsa
##  Min.   :-1.3863   6:28    Min.   :  0.00   Min.   :-0.4308
##  1st Qu.:-1.3863   7:42    1st Qu.:  0.00   1st Qu.: 1.7490
##  Median :-0.7985   8: 1    Median : 15.00   Median : 2.5915
```

```
##  Mean   :-0.1157   9: 4    Mean   : 25.05   Mean   : 2.5244
##  3rd Qu.: 1.3218           3rd Qu.: 40.00   3rd Qu.: 3.1751
##  Max.   : 2.9042           Max.   :100.00   Max.   : 5.5829
```

```
##      lcavol           lweight           age             lbph          svi
##  Min.   :-1.3471   Min.   :3.013   Min.   :56.00   Min.   :-1.38629   0:19
##  1st Qu.: 0.7192   1st Qu.:3.390   1st Qu.:62.25   1st Qu.: 0.07109   1: 3
##  Median : 1.2876   Median :3.832   Median :64.50   Median : 1.41580
##  Mean   : 1.1900   Mean   :3.808   Mean   :64.68   Mean   : 0.86047
##  3rd Qu.: 1.7792   3rd Qu.:4.006   3rd Qu.:66.75   3rd Qu.: 1.79964
##  Max.   : 3.1411   Max.   :6.108   Max.   :77.00   Max.   : 2.30757
##       lcp          gleason     pgg45            lpsa
##  Min.   :-1.3863   6: 7    Min.   : 0.00   Min.   :-0.1625
##  1st Qu.:-1.3863   7:14    1st Qu.: 0.00   1st Qu.: 1.7395
##  Median :-1.3863   8: 0    Median :12.50   Median : 2.4718
##  Mean   :-0.3963   9: 1    Mean   :22.09   Mean   : 2.3216
##  3rd Qu.: 0.3537           3rd Qu.:40.00   3rd Qu.: 2.9425
##  Max.   : 2.4204           Max.   :70.00   Max.   : 3.7124
```

On obteint donc un jeu d'apprentissage comportant 75 lignes et 9 colonnes et un jeu de test de 22 lignes et 9 colonnes.

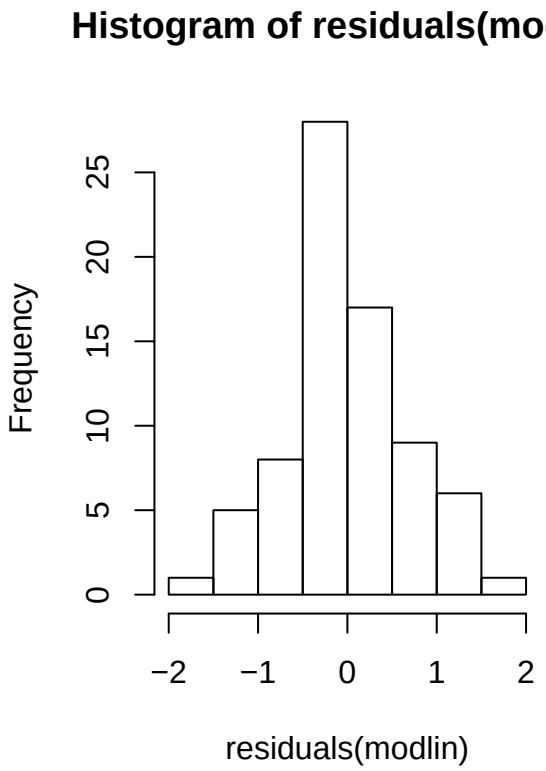
2.2 Modèle linéaire complet

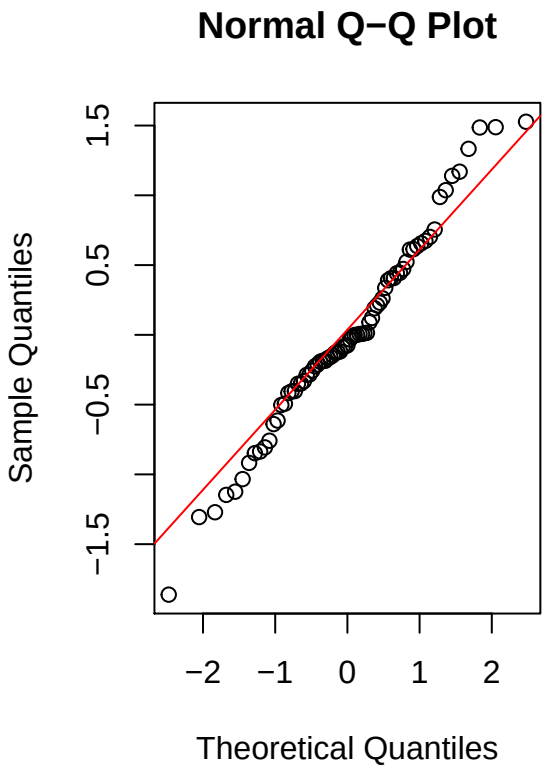
Une fonction utile de graphe des résidus

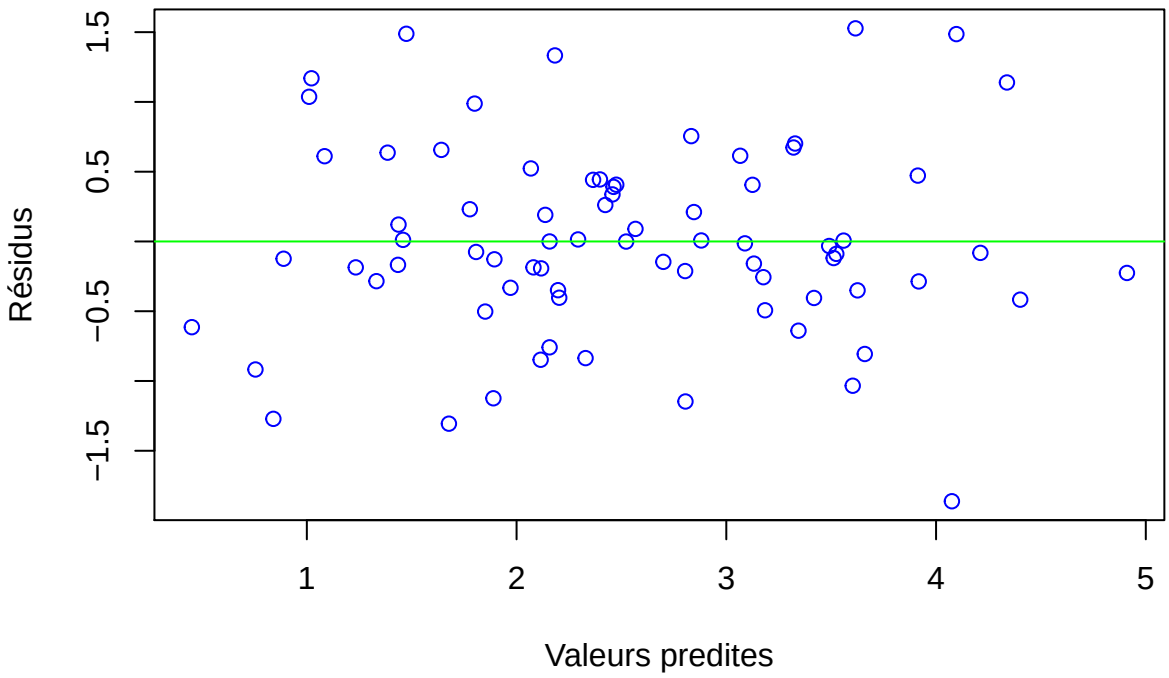
2.2.1 Estimation du modèle et graphes des résidus

```
##
## Call:
## lm(formula = lpsa ~ ., data = Prostate.app)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86180 -0.34983 -0.07532  0.42418  1.52730
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.710840   0.991216   0.717  0.47590
## lcavol       0.549812   0.107883   5.096  3.3e-06 ***
## lweight      0.668793   0.235243   2.843  0.00599 **
## age         -0.028194   0.012643  -2.230  0.02927 *
## lbph         0.124275   0.070797   1.755  0.08398 .
## svi1         0.752437   0.291861   2.578  0.01225 *
## lcp         -0.128543   0.110444  -1.164  0.24880
## gleason7     0.173159   0.264526   0.655  0.51507
## gleason8     0.464916   0.815868   0.570  0.57078
## gleason9    -0.342973   0.604472  -0.567  0.57243
## pgg45        0.006458   0.005407   1.194  0.23681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7363 on 64 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6283
## F-statistic: 13.51 on 10 and 64 DF,  p-value: 2.231e-12
```







Les p-valeurs de certaines variables comme lcavol, lweight, age et svi1 sont rès significatives. Selon la répartition des points (uniformément répartis), on peut dire que les résidus sont sans biais. En regardant l'histogramme des résidus, nous pouvons dire que leur distribution suit une loi normale avec des quantiles compris entre -1.86 et +1.5.

2.2.2 Erreur d'apprentissage

Calculons l'erreur d'apprentissage

```
## [1] 0.4626117
```

L'erreur d'apprentissage obtenue est de 0.46.

2.2.3 Erreur sur l'échantillon test

```
## [1] 0.4500765
```

On obtient 0.45 comme valeur de l'erreur sur l'échantillon de test.

Les erreurs d'apprentissage et de test sont donc presqu'égales.

2.2.4 Nouvelle paramétrisation

Afin de faciliter l'interprétation des résultats concernant les variables qualitatives, on introduit une nouvelle paramétrisation à l'aide de contrastes. Par défaut, la référence est prise pour la valeur 0 de svi et 6 de

gleason, qui sont les plus petites valeurs. Les paramètres indiqués pour les variables svi1 gleason 7, 8 et 9 indiquent l'écart estimé par rapport à cette référence. Il est plus intéressant en pratique de se référer à la moyenne des observations sur toutes les modalités des variables qualitatives, et d'interpréter les coefficients comme des écarts à cette moyenne.

```
##
## Call:
## lm(formula = lpsa ~ ., data = Prostate.app)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86180 -0.34983 -0.07532  0.42418  1.52730
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.160834   1.057759   1.097  0.27656
## lcavol       0.549812   0.107883   5.096  3.3e-06 ***
## lweight      0.668793   0.235243   2.843  0.00599 **
## age         -0.028194   0.012643  -2.230  0.02927 *
## lbph         0.124275   0.070797   1.755  0.08398 .
## svi1        -0.376219   0.145931  -2.578  0.01225 *
## lcp         -0.128543   0.110444  -1.164  0.24880
## gleason1    -0.073775   0.302549  -0.244  0.80813
## gleason2     0.099384   0.237560   0.418  0.67709
## gleason3     0.391140   0.620394   0.630  0.53063
## pgg45        0.006458   0.005407   1.194  0.23681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7363 on 64 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6283
## F-statistic: 13.51 on 10 and 64 DF,  p-value: 2.231e-12
```

Nom des variables : gleason1 =6, gleason2 =7, gleason3 =8, gleason4 =9 , la somme des coefficients associés à ces variables est nulle. svi1=0, svi2=1. La somme des deux coefficients est nulle.

3 Sélection de modèle par sélection de variables

3.1 Sélection par AIC et backward

```
## Start:  AIC=-35.82
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  3    1.2305 35.926 -39.201
## - lcp      1    0.7344 35.430 -36.244
## - pgg45    1    0.7731 35.469 -36.162
## <none>                 34.696 -35.815
## - lbph     1    1.6705 36.366 -34.288
## - age      1    2.6957 37.392 -32.203
## - svi      1    3.6032 38.299 -30.405
```

```
## - lweight  1    4.3817 39.078 -28.895
## - lcavol   1   14.0807 48.777 -12.268
##
## Step:  AIC=-39.2
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.4553 36.382 -40.257
## - pgg45    1    0.8821 36.809 -39.382
## <none>                  35.926 -39.201
## - lbph     1    1.7700 37.696 -37.594
## - age      1    2.4344 38.361 -36.284
## - svi      1    3.7232 39.650 -33.806
## - lweight  1    4.6200 40.546 -32.128
## - lcavol   1   15.8909 51.817 -13.732
##
## Step:  AIC=-40.26
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.5203 36.902 -41.192
## <none>                  36.382 -40.257
## - lbph     1    1.6557 38.037 -38.919
## - age      1    2.1996 38.581 -37.854
## - svi      1    3.2777 39.659 -35.787
## - lweight  1    4.5751 40.957 -33.373
## - lcavol   1   16.8805 53.262 -13.670
##
## Step:  AIC=-41.19
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                  36.902 -41.192
## - age      1    1.8283 38.730 -39.565
## - lbph     1    1.8859 38.788 -39.453
## - lweight  1    4.2368 41.139 -35.040
## - svi      1    4.8021 41.704 -34.017
## - lcavol   1   18.1823 55.084 -13.147


##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = Prostate.app)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93986 -0.37190  0.02246  0.47021  1.43568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.06729    0.99041   1.078  0.28495
## lcavol       0.52960    0.09083   5.831 1.61e-07 ***
## lweight      0.63898    0.22702   2.815  0.00636 **
## age         -0.02172    0.01175  -1.849  0.06875 .
## lbph         0.12957    0.06900   1.878  0.06463 .
```

```
## svi1        -0.36548    0.12197  -2.997  0.00379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7313 on 69 degrees of freedom
## Multiple R-squared:  0.6581, Adjusted R-squared:  0.6334
## F-statistic: 26.57 on 5 and 69 DF,  p-value: 7.267e-15
```

La sélection par crière AIC et backward nous fournit comme meilleur modèle celui en fonction des variables lcavol, lweight, age, lbph, svi. Ceci avec un AIC = 41.19. On note que certaines variables comme âge et lbph restent non significatifs.

3.2 Sélection par AIC et forward

```
## Start:  AIC=29.31
## lpsa ~ 1
##
##           Df Sum of Sq     RSS      AIC
## + lcavol   1    57.008  50.934 -25.0218
## + svi      1    37.507  70.435  -0.7097
## + lcp      1    33.011  74.931   3.9314
## + lweight  1    24.626  83.317  11.8870
## + gleason  3    22.657  85.285  17.6383
## + pgg45    1    16.316  91.626  19.0171
## + lbph     1     6.358 101.584  26.7552
## <none>                 107.942  29.3081
## + age      1     2.448 105.494  29.5878
##
## Step:  AIC=-25.02
## lpsa ~ lcavol
##
##           Df Sum of Sq    RSS     AIC
## + lweight  1    6.8773 44.057 -33.901
## + svi      1    4.8638 46.070 -30.549
## + lbph     1    3.1103 47.824 -27.747
## + pgg45    1    1.3620 49.572 -25.055
## <none>                 50.934 -25.022
## + lcp      1    0.8523 50.082 -24.287
## + age      1    0.0509 50.883 -23.097
## + gleason  3    1.9581 48.976 -21.962
##
## Step:  AIC=-33.9
## lpsa ~ lcavol + lweight
##
##           Df Sum of Sq    RSS     AIC
## + svi      1    4.1621 39.895 -39.343
## + pgg45    1    1.3831 42.674 -34.293
## + age      1    1.1984 42.858 -33.969
## <none>                 44.057 -33.901
## + lcp      1    0.7377 43.319 -33.167
## + lbph     1    0.6054 43.451 -32.938
## + gleason  3    2.0371 42.020 -31.451
##
```

```
## Step:  AIC=-39.34
## lpsa ~ lcavol + lweight + svi
##
##           Df Sum of Sq    RSS     AIC
## + lbph     1   1.16439 38.730 -39.565
## + age      1   1.10678 38.788 -39.453
## <none>                 39.895 -39.343
## + pgg45    1   0.31242 39.582 -37.933
## + lcp      1   0.02440 39.870 -37.389
## + gleason  3   1.11456 38.780 -35.469
##
## Step:  AIC=-39.56
## lpsa ~ lcavol + lweight + svi + lbph
##
##           Df Sum of Sq    RSS     AIC
## + age      1   1.82830 36.902 -41.192
## <none>                 38.730 -39.565
## + pgg45    1   0.14896 38.581 -37.854
## + lcp      1   0.07421 38.656 -37.709
## + gleason  3   0.85165 37.879 -35.233
##
## Step:  AIC=-41.19
## lpsa ~ lcavol + lweight + svi + lbph + age
##
##           Df Sum of Sq    RSS     AIC
## <none>                 36.902 -41.192
## + pgg45    1   0.52027 36.382 -40.257
## + lcp      1   0.09347 36.809 -39.382
## + gleason  3   1.13158 35.770 -37.528

##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = Prostate.app)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93986 -0.37190  0.02246  0.47021  1.43568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.06729    0.99041   1.078  0.28495
## lcavol       0.52960    0.09083   5.831 1.61e-07 ***
## lweight      0.63898    0.22702   2.815  0.00636 **
## svi1        -0.36548    0.12197  -2.997  0.00379 **
## lbph         0.12957    0.06900   1.878  0.06463 .
## age         -0.02172    0.01175  -1.849  0.06875 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7313 on 69 degrees of freedom
## Multiple R-squared:  0.6581, Adjusted R-squared:  0.6334
## F-statistic: 26.57 on 5 and 69 DF,  p-value: 7.267e-15
```

Avec la sélection par AIC et forward, on obtient presque les mêmes résultats qu'avec la sélection par AIC

backward.

3.3 Sélection par AIC et stepwise

```
## Start:  AIC=-35.82
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  3    1.2305 35.926 -39.201
## - lcp      1    0.7344 35.430 -36.244
## - pgg45    1    0.7731 35.469 -36.162
## <none>                 34.696 -35.815
## - lbph     1    1.6705 36.366 -34.288
## - age      1    2.6957 37.392 -32.203
## - svi      1    3.6032 38.299 -30.405
## - lweight  1    4.3817 39.078 -28.895
## - lcavol   1   14.0807 48.777 -12.268
##
## Step:  AIC=-39.2
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.4553 36.382 -40.257
## - pgg45    1    0.8821 36.809 -39.382
## <none>                 35.926 -39.201
## - lbph     1    1.7700 37.696 -37.594
## - age      1    2.4344 38.361 -36.284
## + gleason  3    1.2305 34.696 -35.815
## - svi      1    3.7232 39.650 -33.806
## - lweight  1    4.6200 40.546 -32.128
## - lcavol   1   15.8909 51.817 -13.732
##
## Step:  AIC=-40.26
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.5203 36.902 -41.192
## <none>                 36.382 -40.257
## + lcp      1    0.4553 35.926 -39.201
## - lbph     1    1.6557 38.037 -38.919
## - age      1    2.1996 38.581 -37.854
## + gleason  3    0.9515 35.430 -36.244
## - svi      1    3.2777 39.659 -35.787
## - lweight  1    4.5751 40.957 -33.373
## - lcavol   1   16.8805 53.262 -13.670
##
## Step:  AIC=-41.19
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                 36.902 -41.192
## + pgg45    1    0.5203 36.382 -40.257
```

```
## - age      1    1.8283 38.730 -39.565
## - lbph     1    1.8859 38.788 -39.453
## + lcp      1    0.0935 36.809 -39.382
## + gleason  3    1.1316 35.770 -37.528
## - lweight  1    4.2368 41.139 -35.040
## - svi      1    4.8021 41.704 -34.017
## - lcavol   1   18.1823 55.084 -13.147

##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = Prostate.app)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93986 -0.37190  0.02246  0.47021  1.43568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.06729    0.99041   1.078  0.28495
## lcavol       0.52960    0.09083   5.831 1.61e-07 ***
## lweight      0.63898    0.22702   2.815  0.00636 **
## age         -0.02172    0.01175  -1.849  0.06875 .
## lbph         0.12957    0.06900   1.878  0.06463 .
## svi1        -0.36548    0.12197  -2.997  0.00379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7313 on 69 degrees of freedom
## Multiple R-squared:  0.6581, Adjusted R-squared:  0.6334
## F-statistic: 26.57 on 5 and 69 DF,  p-value: 7.267e-15
```

On retrouve ici le meme modèle qu'avec l'agorithme backward.

3.4 Sélection par BIC et stepwise

k=log(napp) pour BIC au lieu de AIC.

```
## Start:  AIC=-10.32
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS      AIC
## - gleason  3    1.2305 35.926 -20.6613
## - lcp      1    0.7344 35.430 -13.0693
## - pgg45    1    0.7731 35.469 -12.9873
## - lbph     1    1.6705 36.366 -11.1135
## <none>                 34.696 -10.3227
## - age      1    2.6957 37.392  -9.0283
## - svi      1    3.6032 38.299  -7.2298
## - lweight  1    4.3817 39.078  -5.7205
## - lcavol   1   14.0807 48.777  10.9070
##
```

```
## Step:  AIC=-20.66
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS      AIC
## - lcp      1    0.4553 36.382 -24.0342
## - pgg45    1    0.8821 36.809 -23.1595
## - lbph     1    1.7700 37.696 -21.3718
## <none>                 35.926 -20.6613
## - age      1    2.4344 38.361 -20.0614
## - svi      1    3.7232 39.650 -17.5831
## - lweight  1    4.6200 40.546 -15.9056
## + gleason  3    1.2305 34.696 -10.3227
## - lcavol   1   15.8909 51.817   2.4901
##
## Step:  AIC=-24.03
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS      AIC
## - pgg45    1    0.5203 36.902 -27.2868
## - lbph     1    1.6557 38.037 -25.0139
## <none>                 36.382 -24.0342
## - age      1    2.1996 38.581 -23.9490
## - svi      1    3.2777 39.659 -21.8820
## + lcp      1    0.4553 35.926 -20.6613
## - lweight  1    4.5751 40.957 -19.4678
## + gleason  3    0.9515 35.430 -13.0693
## - lcavol   1   16.8805 53.262   0.2354
##
## Step:  AIC=-27.29
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS      AIC
## - age      1    1.8283 38.730 -27.9775
## - lbph     1    1.8859 38.788 -27.8660
## <none>                 36.902 -27.2868
## + pgg45    1    0.5203 36.382 -24.0342
## - lweight  1    4.2368 41.139 -23.4528
## + lcp      1    0.0935 36.809 -23.1595
## - svi      1    4.8021 41.704 -22.4292
## + gleason  3    1.1316 35.770 -16.6701
## - lcavol   1   18.1823 55.084  -1.5593
##
## Step:  AIC=-27.98
## lpsa ~ lcavol + lweight + lbph + svi
##
##           Df Sum of Sq    RSS      AIC
## - lbph     1    1.1644 39.895 -30.0734
## <none>                 38.730 -27.9775
## + age      1    1.8283 36.902 -27.2868
## - lweight  1    3.3237 42.054 -26.1201
## + pgg45    1    0.1490 38.581 -23.9490
## + lcp      1    0.0742 38.656 -23.8039
## - svi      1    4.7211 43.451 -23.6684
## + gleason  3    0.8516 37.879 -16.6926
```

```
## - lcavol   1   17.0987 55.829  -4.8696
##
## Step:  AIC=-30.07
## lpsa ~ lcavol + lweight + svi
##
##           Df Sum of Sq    RSS      AIC
## <none>                 39.895 -30.0734
## + lbph     1    1.1644 38.730 -27.9775
## + age      1    1.1068 38.788 -27.8660
## - svi      1    4.1621 44.057 -26.9482
## + pgg45    1    0.3124 39.582 -26.3456
## + lcp      1    0.0244 39.870 -25.8018
## - lweight  1    6.1755 46.070 -23.5966
## + gleason  3    1.1146 38.780 -19.2461
## - lcavol   1   17.5291 57.424  -7.0747

##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = Prostate.app)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80400 -0.46271 -0.00071  0.44944  1.55985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.50668    0.74740  -0.678  0.50002
## lcavol       0.51598    0.09238   5.585 4.02e-07 ***
## lweight      0.68900    0.20783   3.315  0.00144 **
## svi1        -0.33677    0.12374  -2.722  0.00817 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7496 on 71 degrees of freedom
## Multiple R-squared:  0.6304, Adjusted R-squared:  0.6148
## F-statistic: 40.37 on 3 and 71 DF,  p-value: 2.452e-15
```

Avec la méthode BIC et stepwise, on obtient moins de variables dans le meilleur sélectionné comparé aux méthodes précédentes. Le modèle sélectionné est plus parcimonieux dans le cas de la méthode BIC et stepwise.

3.5 Erreur sur l'échantillon d'apprentissage

Modèle stepwise AIC

```
## [1] 0.4920266
```

Avec la méthode stepwise AIC, on trouve pour valeur 0.49 comme erreur sur l'échantillon d'apprentissage.

Modèle stepwise BIC

```
## [1] 0.531929
```

Avec la méthode stepwise BIC, on trouve pour valeur 0.53 comme erreur sur l'échantillon d'apprentissage.

3.6 Calcul de l'erreur sur l'échantillon test

Modèle stepwise AIC

```
## [1] 0.4655829
```

Avec la méthode stepwise AIC, on trouve pour valeur 0.46 comme erreur sur l'échantillon test.

Modèle stepwise BIC

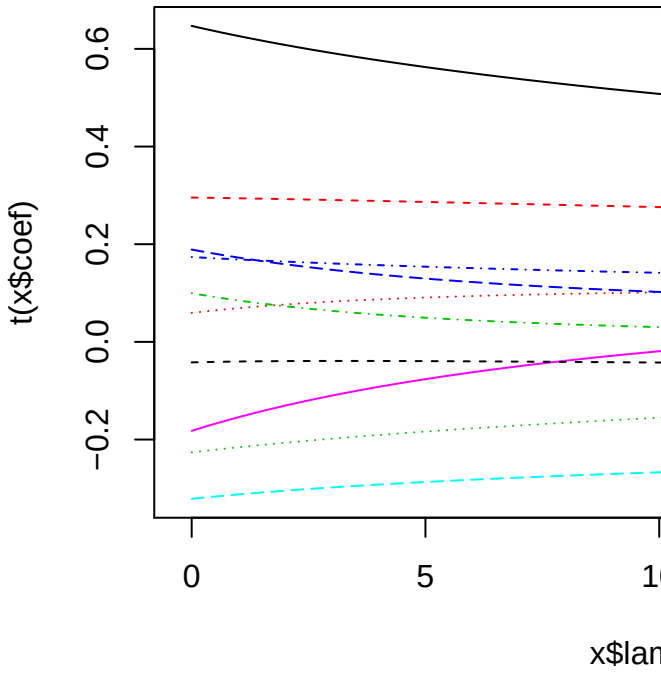
```
## [1] 0.4008493
```

Avec la méthode stepwise BIC, on trouve pour valeur 0.40 comme erreur sur l'échantillon test.

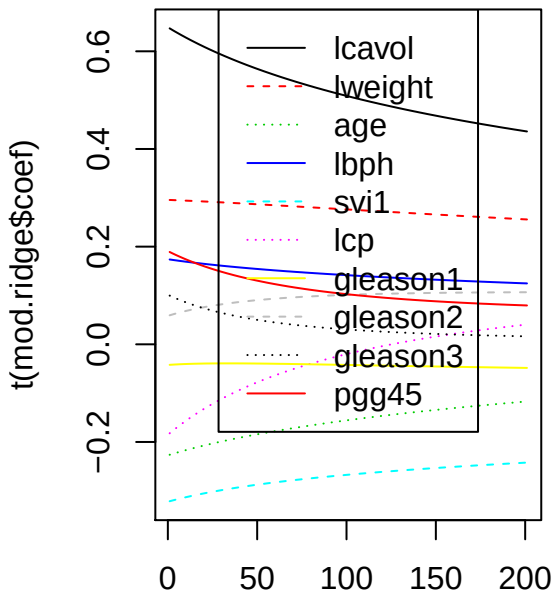
Les modèles sélectionnés ont une erreur plus grande que le modèle linéaire comprenant toutes les variables sur l'échantillon d'apprentissage. Sur l'échantillon test, le modèle qui minimise le critère BIC a de meilleures performances que le modèle initial. Les deux modèles sélectionnés sont beaucoup plus parcimonieux que le modèle initial.

4-Sélection de modèle par pénalisation Ridge

4.1 Comportement des coefficients



Calcul des coefficients pour différentes valeurs du paramètre lambda



Evolution des coefficients

4.2 Pénalisation optimale par validation croisée

```
## modified HKB estimator is 5.597491
## modified L-W estimator is 4.440824
## smallest value of GCV  at 10.4
```

4.3 Prévision et erreur d'apprentissage

Ici, on calcule les valeurs prédites à partir des coefficients.

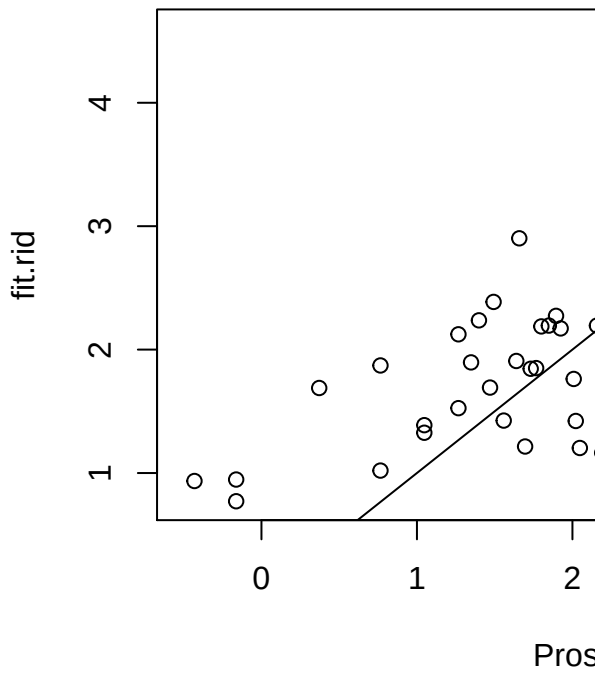
Coefficients du modèle sélectionné:

On crée des vecteurs pour :

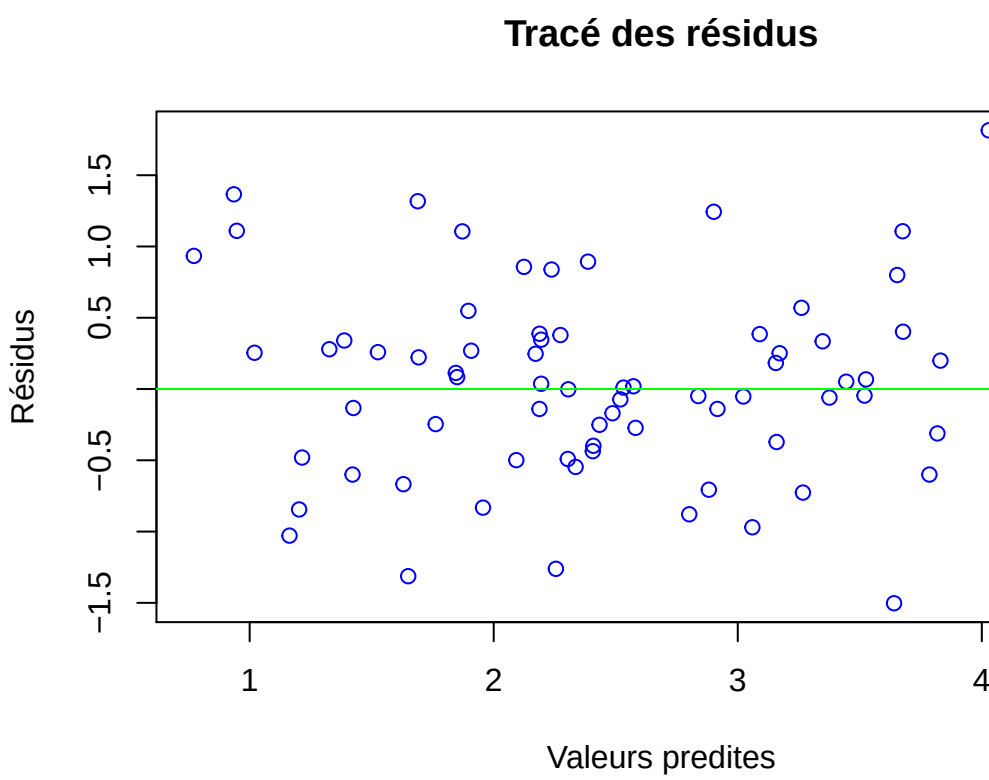
-les variables qualitatives

-les variables quantitatives

On calcule ici des valeurs prédites



Nous traçons ensuite des valeurs prédites en fonction des valeurs observées



On calcule et on fait le tracé des résidus
Selon la répartition des points (uniformément répartis), on peut dire que les résidus sont sans biais.

Erreurd'apprentissage

```
## [1] 0.4789025
```

On trouve pour valeur 0.478 comme erreur d'apprentissage.

4.4 Prévision sur l'échantillon test

Les variables qualitatives

Les variables quantitatives

Erreur sur l'échantillon test

```
## [1] 0.424398
```

On trouve pour valeur 0.424 comme erreur sur l'échantillon test.

L'erreur d'apprentissage est légèrement plus élevée que pour le modèle linéaire sans pénalisation. L'erreur de test est plus faible.Les performances sont comparables sur l'échantillon test au modèle sélectionné par le critère BIC.En terme d'interprétation,les modèles sélectionnés par AIC et BIC sont préférables.

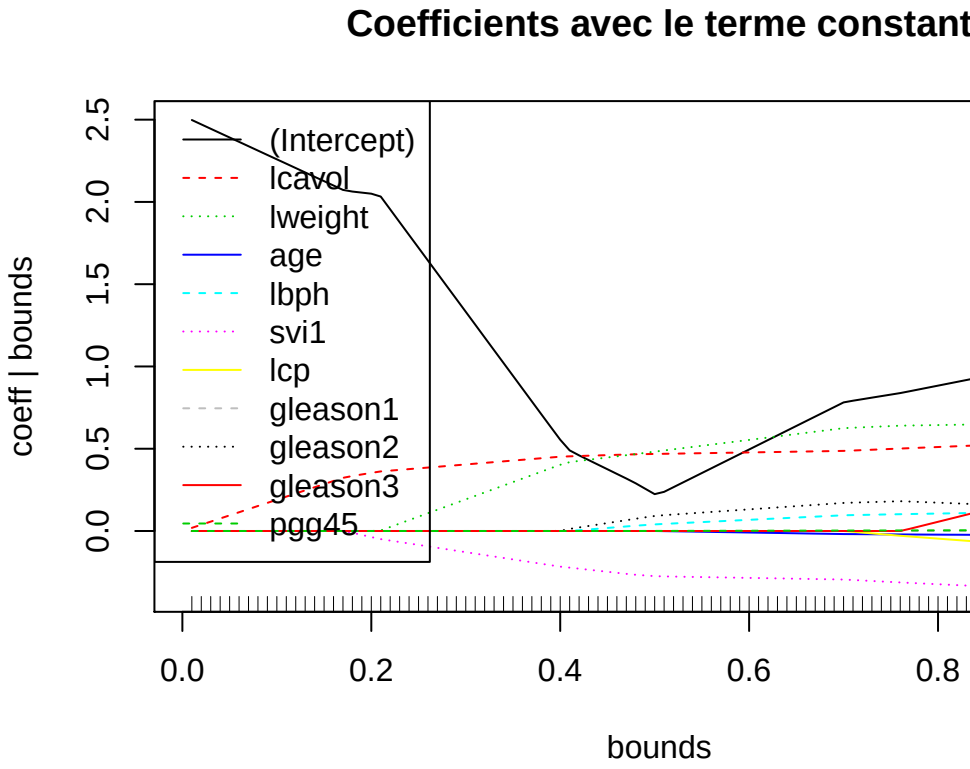
5 Sélection de modèle par pénalisation Lasso

5.1 Librairie Lasso2

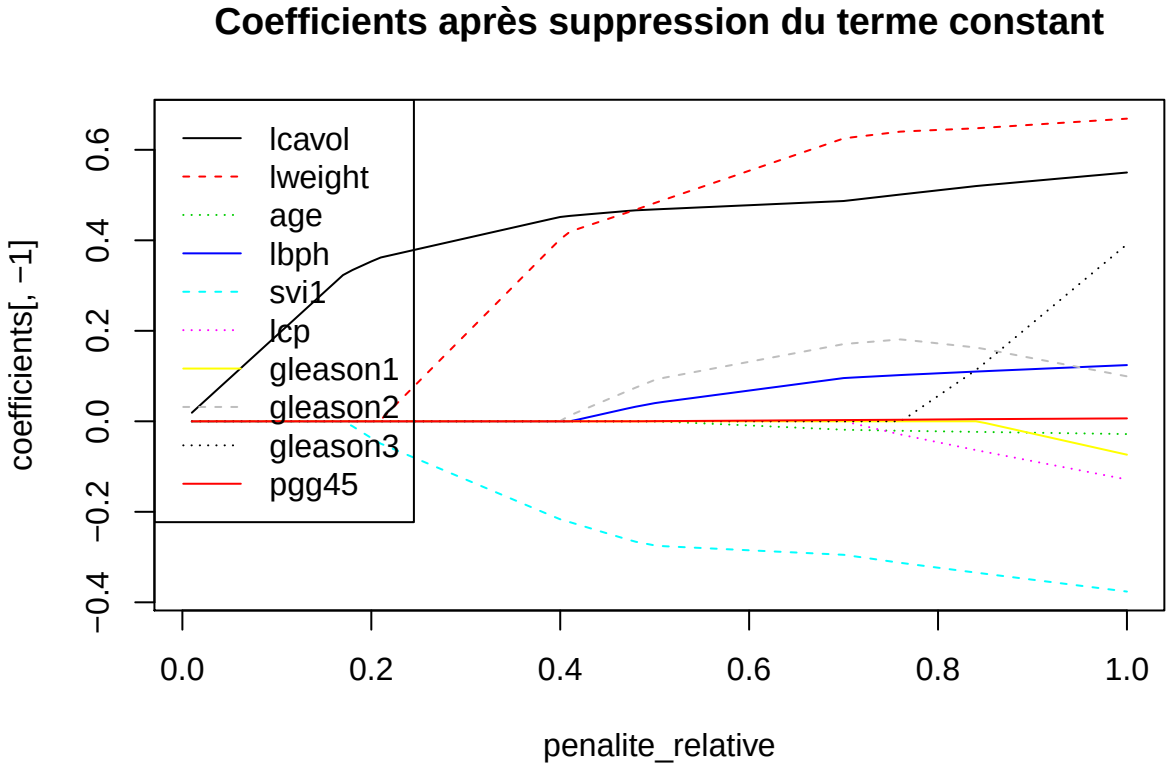
5.2 Construction du modèle

La borne est ici relative,elle correspond à une certaine proportion de la norme L1 du vecteur des coeficients des moindres carrés.Une borne égale à 1 correspond donc à l'absence de pénalité,on retrouve l'estimateur des moindres carrés.

5.3 Visualisation des coefficients



On visualise ici les coefficients du modèle



En fonction de différentes valeurs de la pénalité, on obtient différentes valeurs pour les coefficients. Afin de faire un bon choix de la pénalité, nous procédons par la méthode de validation croisée.

5.4 Sélection de la pénalité par validation croisée

On procède à la validation croisée pour sélectionner la pénalité.

Par la méthode de validation croisée, on obtient 0.95 comme valeur de la meilleure pénalité qui nous a servi à optimiser notre modèle précédent en utilisant que cette meilleure pénalité dans le modèle.

5.5 Erreur d'apprentissage

```
## [1] 0.4629727
```

On trouve pour valeur 0.46 comme erreur d'apprentissage.

5.6 Erreur sur l'échantillon test

```
## [1] 0.4438783
```

On trouve pour valeur 0.44 comme erreur sur l'échantillon test.

5.7 Librairie glmnet

L'utilisation de la librairie glmnet fournit des résultats plus rapides, ce qui peut s'avérer important pour des données de grande dimension. Par contre, on ne peut pas traiter à priori des variables qualitatives. Nous allons donc devoir créer des vecteurs avec des variables indicatrices des diverses modalités pour les variables qualitatives. Nous ne prendrons pas en compte les contrastes.

5.8 Mise en forme des variables

on construit une matrice xx.app d'apprentissage et xx.test de test

On construit ici des vecteurs indicatrices pour les variables qualitatives

On crée une matrice avec les vecteurs indicatrices

On nomme les colonnes avec les noms des variables

On fait de même pour l'echantillon test

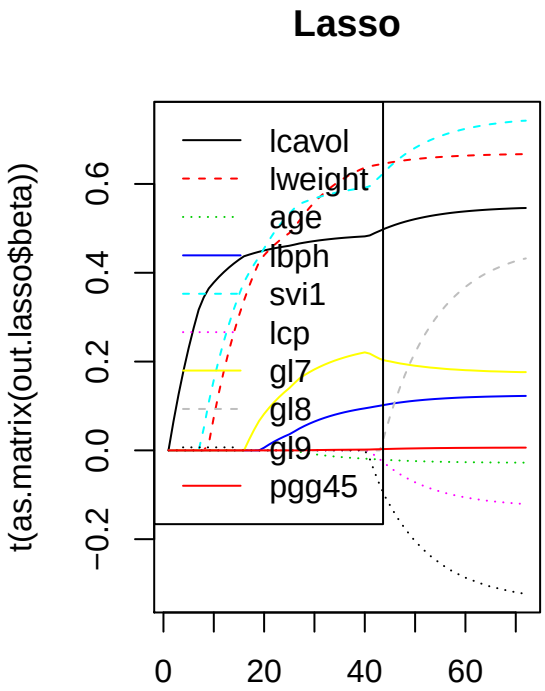
On construit une matrice avec les vecteurs indicatrices

5.9 Construction du modèle

```
## Loading required package: Matrix
```

```
## Loaded glmnet 3.0-2
```

5.10 Visualisation des coefficients



Chemin de régularisation du lasso
Ici, on a tracé les coefficients des différentes variables du modèle Lasso.

5.11 Sélection de la pénalité par validation croisée

Nous appliquons la méthode de validation croisée afin de sélectionner la meilleure pénalité.Ensuite, on optimise le modèle avec la valeur de la meilleure pénalité trouvée par la méthode de validation croisée.

Nous trouvons la valeur de 0.0444 comme meilleure pénalité.

5.12 Erreur d'apprentissage

```
## [1] 0.4802486
```

Avec le modèle optimisé, nous trouvons 0.48 comme valeur de l'erreur d'apprentissage.

5.13 Erreur sur l'échantillon test

```
## [1] 0.4032071
```

Avec le modèle optimisé, nous trouvons environ 0.39 comme erreur de prédiction commise sur le jeu de test.

Il est à noter que l'erreur de test est un peu plus petite que l'erreur d'apprentissage.

5.14 Elastic Net

La méthode Elastic Net est une méthode qui permet de résoudre les problèmes liés à la méthode Lasso (comme par exemples le problème de colinéarité entre les variables et le problème du nombre de variables à sélectionner lié au fléau de la dimension).

On peut jouer avec le paramètre alpha de glmnet

Ici, nous avons crée un modèle avec 0.5 comme paramètre alpha. On a procédé ensuite par la méthode de validation croisée pour choisir le meilleur paramètre lambda. Ceci nous a permis d'optimiser notre modèle. On utilise donc le modèle optmisé pour faire des prédictions et calculer les erreurs commises sur le jeu d'apprentissage et sur le jeu de test.

Erreur d'apprentissage

```
## [1] 0.5774533
```

Nous trouvons la valeur 0.6 comme erreur d'apprentissage.

Erreur de prédiction

```
## [1] 0.3670371
```

Nous trouvons la valeur 0.37 comme erreur de test.

L'erreur commise sur le jeu de test est plus petite que celle commise sur le jeu d'apprentissage.