Statistique bayésienne avec R - exercice régression données Prostate

Julien JACQUES

Chargeons les données

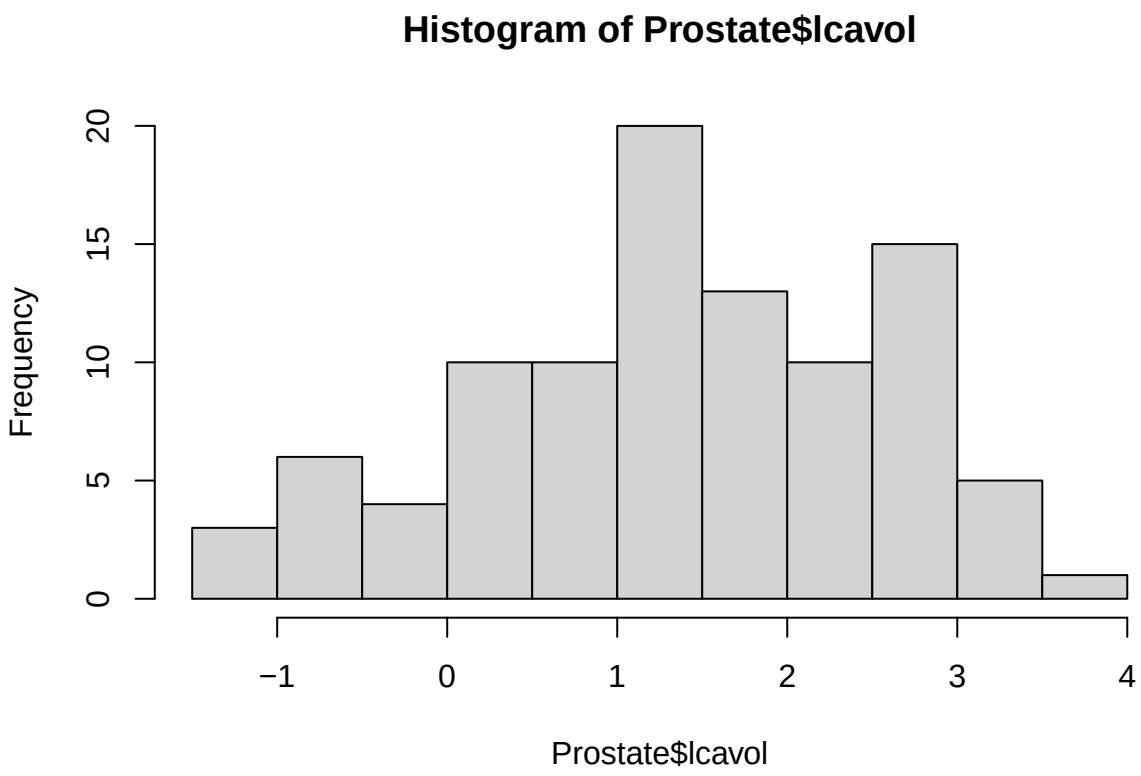
```
library(lasso2)
```

```
## 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>
```

```
data(Prostate)
```

Un premier examen rapide nous permet de juger que la variable a prédire est approximativement gaussienne, et les variables explicatives sont très corrélées entre elles

```
hist(Prostate$lcavol)
```



```
cor(Prostate)
```

```
##            lcavol      lweight       age         lbph         svi          lcp
## lcavol  1.0000000  0.194128286 0.2249999  0.027349703  0.53884500  0.675310484
## lweight 0.1941283  1.000000000 0.3075286  0.434934636  0.10877851  0.100237795
## age     0.2249999  0.307528614 1.0000000  0.350185896  0.11765804  0.127667752
## lbph    0.0273497  0.434934636 0.3501859  1.000000000 -0.08584324 -0.006999431
## svi     0.5388450  0.108778505 0.1176580 -0.085843238  1.00000000  0.673111185
## lcp     0.6753105  0.100237795 0.1276678 -0.006999431  0.67311118  1.000000000
## gleason 0.4324171 -0.001275658 0.2688916  0.077820447  0.32041222  0.514830063
## pgg45   0.4336522  0.050846821 0.2761124  0.078460018  0.45764762  0.631528245
## lpsa    0.7344603  0.354120390 0.1695928  0.179809410  0.56621822  0.548813169
##              gleason      pgg45      lpsa
## lcavol   0.432417056 0.43365225 0.7344603
## lweight -0.001275658 0.05084682 0.3541204
## age      0.268891599 0.27611245 0.1695928
## lbph     0.077820447 0.07846002 0.1798094
## svi      0.320412221 0.45764762 0.5662182
## lcp      0.514830063 0.63152825 0.5488132
## gleason  1.000000000 0.75190451 0.3689868
## pgg45    0.751904512 1.00000000 0.4223159
## lpsa     0.368986803 0.42231586 1.0000000
```

Regression linéaire avec sélection de variables

```
modele1=lm(lcavol~.,data=Prostate)
summary(modele1)
```

```
##
## Call:
## lm(formula = lcavol ~ ., data = Prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88603 -0.47346 -0.03987  0.55719  1.86870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.260101   1.259683  -1.794   0.0762 .
## lweight     -0.073166   0.174450  -0.419   0.6759
## age          0.022736   0.010964   2.074   0.0410 *
## lbph        -0.087449   0.058084  -1.506   0.1358
## svi         -0.153591   0.253932  -0.605   0.5468
## lcp          0.367300   0.081689   4.496 2.10e-05 ***
## gleason      0.190759   0.154283   1.236   0.2196
## pgg45       -0.007158   0.004326  -1.654   0.1016
## lpsa         0.572797   0.085790   6.677 2.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6998 on 88 degrees of freedom
## Multiple R-squared:  0.6769, Adjusted R-squared:  0.6475
## F-statistic: 23.04 on 8 and 88 DF,  p-value: < 2.2e-16
```

Nous pouvons réaliser une sélection de variables à l'aide d'une procédure step

```
modele2=step(modele1)
```

```
## Start:  AIC=-60.7
## lcavol ~ lweight + age + lbph + svi + lcp + gleason + pgg45 +
##     lpsa
##
##           Df Sum of Sq    RSS     AIC
## - lweight  1    0.0861 43.179 -62.507
## - svi      1    0.1792 43.272 -62.299
## - gleason  1    0.7486 43.842 -61.031
## <none>                 43.093 -60.701
## - lbph     1    1.1100 44.203 -60.234
## - pgg45    1    1.3403 44.433 -59.730
## - age      1    2.1058 45.199 -58.073
## - lcp      1    9.9002 52.993 -42.641
## - lpsa     1   21.8300 64.923 -22.946
##
## Step:  AIC=-62.51
## lcavol ~ age + lbph + svi + lcp + gleason + pgg45 + lpsa
##
##           Df Sum of Sq    RSS     AIC
## - svi      1    0.1752 43.354 -64.115
## - gleason  1    0.8357 44.015 -62.648
## <none>                 43.179 -62.507
## - pgg45    1    1.3195 44.499 -61.588
## - lbph     1    1.4818 44.661 -61.234
## - age      1    2.0198 45.199 -60.073
## - lcp      1    9.8752 53.054 -44.529
## - lpsa     1   23.1542 66.333 -22.862
##
## Step:  AIC=-64.11
## lcavol ~ age + lbph + lcp + gleason + pgg45 + lpsa
##
##           Df Sum of Sq    RSS     AIC
## <none>                 43.354 -64.115
## - gleason  1    0.9571 44.311 -63.997
## - lbph     1    1.3338 44.688 -63.175
## - pgg45    1    1.4298 44.784 -62.967
## - age      1    1.9355 45.290 -61.878
## - lcp      1   10.9352 54.289 -44.297
## - lpsa     1   24.9001 68.254 -22.093
```

```
summary(modele2)
```

```
##
## Call:
## lm(formula = lcavol ~ age + lbph + lcp + gleason + pgg45 + lpsa,
##     data = Prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90430 -0.51715 -0.02241  0.57388  1.87347
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.535920   1.111608  -2.281   0.0249 *
## age          0.021243   0.010598   2.004   0.0480 *
## lbph        -0.088505   0.053188  -1.664   0.0996 .
## lcp          0.344851   0.072379   4.765 7.22e-06 ***
## gleason      0.211762   0.150234   1.410   0.1621
## pgg45       -0.007353   0.004268  -1.723   0.0884 .
## lpsa         0.544422   0.075723   7.190 1.84e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6941 on 90 degrees of freedom
## Multiple R-squared:  0.6749, Adjusted R-squared:  0.6532
## F-statistic: 31.14 on 6 and 90 DF,  p-value: < 2.2e-16
```

La variable gleason n'étant pas significative dans le modèle obtenu par la sélection de variables (step), je la supprime.

```
modele3=lm(lcavol~age+lbph+lcp+pgg45+lpsa,data=Prostate)
summary(modele3)
```

```
##
## Call:
## lm(formula = lcavol ~ age + lbph + lcp + pgg45 + lpsa, data = Prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97761 -0.49431 -0.04114  0.55594  2.00901
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.299521   0.686526  -1.893   0.0616 .
## age          0.022662   0.010607   2.137   0.0353 *
## lbph        -0.089254   0.053473  -1.669   0.0985 .
## lcp          0.350498   0.072659   4.824 5.63e-06 ***
## pgg45       -0.003649   0.003381  -1.079   0.2834
## lpsa         0.549946   0.076031   7.233 1.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6978 on 91 degrees of freedom
## Multiple R-squared:  0.6677, Adjusted R-squared:  0.6495
## F-statistic: 36.57 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
modele3=lm(lcavol~age+lbph+lcp+lpsa,data=Prostate)
summary(modele3)
```

```
##
## Call:
## lm(formula = lcavol ~ age + lbph + lcp + lpsa, data = Prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92180 -0.49945 -0.09019  0.58214  2.01844
##
```

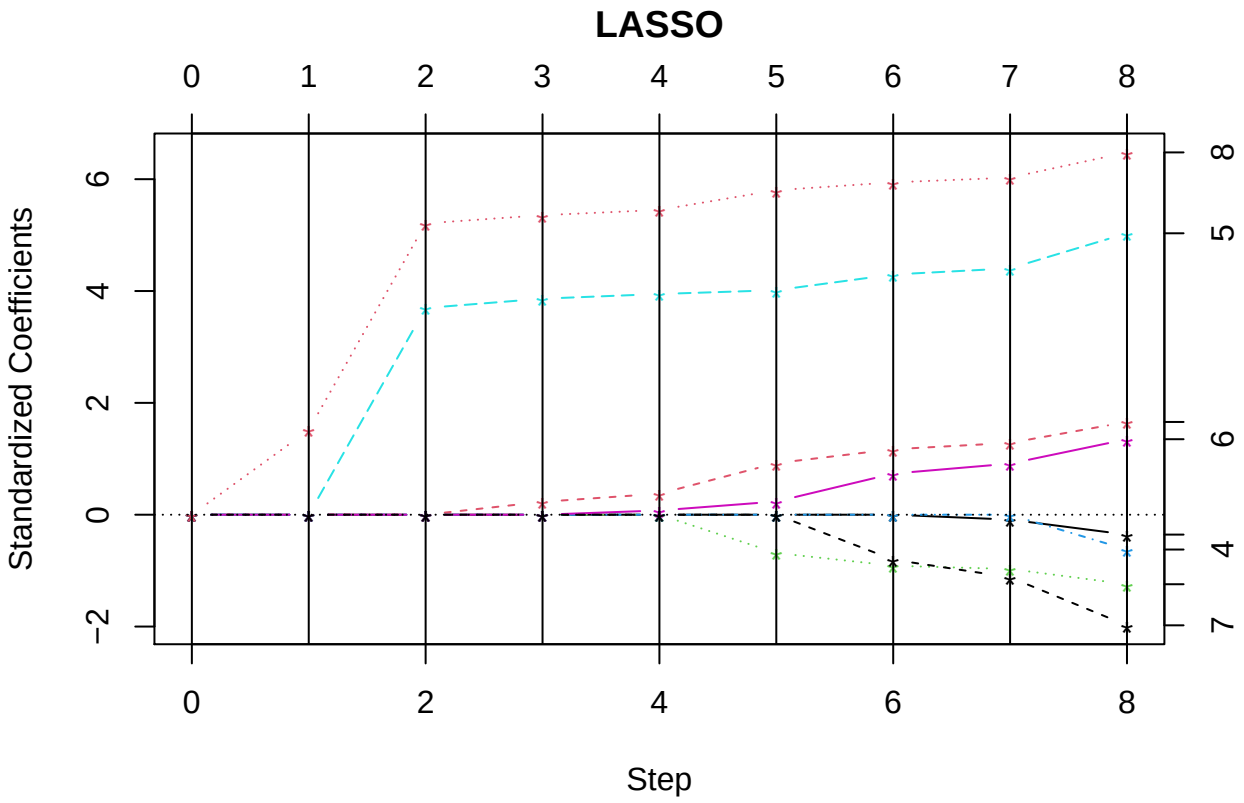
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.21022    0.68213  -1.774   0.0793 .
## age          0.02003    0.01033   1.939   0.0556 .
## lbph        -0.08935    0.05352  -1.669   0.0984 .
## lcp          0.30907    0.06175   5.006 2.67e-06 ***
## lpsa         0.54273    0.07580   7.160 1.93e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6984 on 92 degrees of freedom
## Multiple R-squared:  0.6635, Adjusted R-squared:  0.6488
## F-statistic: 45.35 on 4 and 92 DF,  p-value: < 2.2e-16
```

Regression LASSO

```
library('lars')
```

```
## Loaded lars 1.2
```

```
model_lasso=lars(as.matrix(Prostate[,-1]),Prostate$lcavol,type="lasso",trace=F,normalize=TRUE)
plot(model_lasso,xvar='step', plottype='coeff')
```



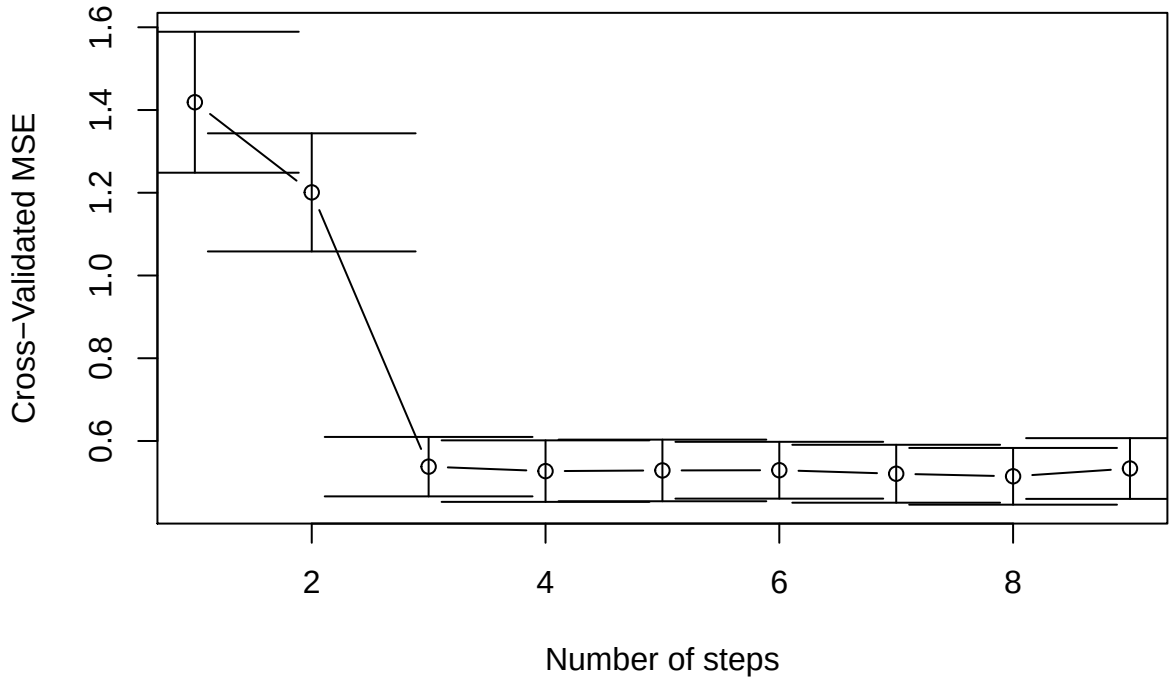

```
print(model_lasso)
```

```
##
## Call:
## lars(x = as.matrix(Prostate[, -1]), y = Prostate$lcavol, type = "lasso",
##     trace = F, normalize = TRUE)
```

```
## R-squared: 0.677
## Sequence of LASSO moves:
##      lpsa lcp age gleason lbph pgg45 lweight svi
## Var     8   5   2       6    3     7       1   4
## Step    1   2   3       4    5     6       7   8
```

Cherchons le lambda optimal

```
cv=cv.lars(as.matrix(Prostate[,-1]),Prostate$lcavol,K=10,trace=F,plot.it=T,se=T,type=c("lasso"),mode='s
```



Le meilleur CVMSE semble être à la troisième étape.

On peut examiner les valeurs des variables sélectionnées à la troisième étape

```
print(model_lasso$lambda[3])
```

```
## [1] 1.242805
```

```
print(model_lasso$beta[3,])
```

```
##   lweight       age      lbph       svi       lcp   gleason     pgg45      lpsa
## 0.0000000 0.0000000 0.0000000 0.0000000 0.2698036 0.0000000 0.0000000 0.4606733
```

Ré-estimons le modèle avec les variables sélectionnées par le LASSO

```
modele4=lm(lcavol~lpsa+lcp,data=Prostate)
summary(modele4)
```

```
##
## Call:
## lm(formula = lcavol ~ lpsa + lcp, data = Prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65744 -0.54398 -0.05502  0.57163  2.07959
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09135    0.20527   0.445    0.657
## lpsa         0.53162    0.07501   7.087 2.49e-10 ***
## lcp          0.32837    0.06193   5.303 7.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7092 on 94 degrees of freedom
## Multiple R-squared:  0.6455, Adjusted R-squared:  0.6379
## F-statistic: 85.57 on 2 and 94 DF,  p-value: < 2.2e-16
```

On peut finalement comparer les modèles suivant le critère de notre choix (AdjR2,AIC,BIC).

```
library(broom)
rbind(glance(modele1),glance(modele2),glance(modele3),glance(modele4))
```

```
## # A tibble: 4 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.677         0.647 0.700      23.0 1.34e-18     8  -98.3  217.  242.
## 2     0.675         0.653 0.694      31.1 5.52e-20     6  -98.6  213.  234.
## 3     0.663         0.649 0.698      45.3 5.51e-21     4 -100.   213.  228.
## 4     0.645         0.638 0.709      85.6 6.81e-22     2 -103.   214.  224.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Le meilleur modèle suivant les critères AIC et R2ajusté est le modèle 5, avec les 2 variables lpsa et lcp

On pourrait aussi implétenter une validation croisée

```
library(lmvar)
modele1=lm(lcavol~.,data=Prostate,x=TRUE,y=TRUE)
modele2=step(modele1,trace = FALSE)
modele3=lm(lcavol~age+lbph+lcp+pgg45+lpsa,data=Prostate,x=TRUE,y=TRUE)
modele4=lm(lcavol~lpsa+lcp,data=Prostate,x=TRUE,y=TRUE)
cv.lm(modele1)
```

```
## Mean absolute error        :  0.5709444
## Sample standard deviation  :  0.1204477
##
## Mean squared error         :  0.4843254
## Sample standard deviation  :  0.2244983
##
## Root mean squared error    :  0.6828113
## Sample standard deviation  :  0.1417909
```

```
cv.lm(modele2)
```

```
## Mean absolute error        :  0.6021423
## Sample standard deviation  :  0.1833568
##
## Mean squared error         :  0.537672
## Sample standard deviation  :  0.2789312
##
## Root mean squared error    :  0.711776
## Sample standard deviation  :  0.1857324
```

```
cv.lm(modele3)
```

```
## Mean absolute error        :  0.6123086
```

```
## Sample standard deviation  :  0.1478697
##
## Mean squared error         :  0.5410866
## Sample standard deviation  :  0.2426156
##
## Root mean squared error    :  0.7198675
## Sample standard deviation  :  0.1594343
```

```
cv.lm(modele4)
```

```
## Mean absolute error        :  0.6147869
## Sample standard deviation  :  0.1509818
##
## Mean squared error         :  0.5530135
## Sample standard deviation  :  0.2807446
##
## Root mean squared error    :  0.7250806
## Sample standard deviation  :  0.1740742
```

Le meilleur modèle suivant le critère de validation croisée est le modèle 4, avec les 3 variables séléctionnées par LASSO