Kryteria wyboru modelu - ćwiczenia

Piotr J. Sobczyk

8 grudnia 2016

Na dzisiejszych zajęciach:

1. Dowiemy się na dlaczego nie zawsze chcemy używać wszystkich dostępnych zmiennych w modelu
2. Poznamy dwie metody wyboru modelu AIC i BIC, które oparte są na penalizowanej funkcji wiarogodności
3. Poznamy jak wybierać model przy pomocy walidacji krzyżowej (CV)

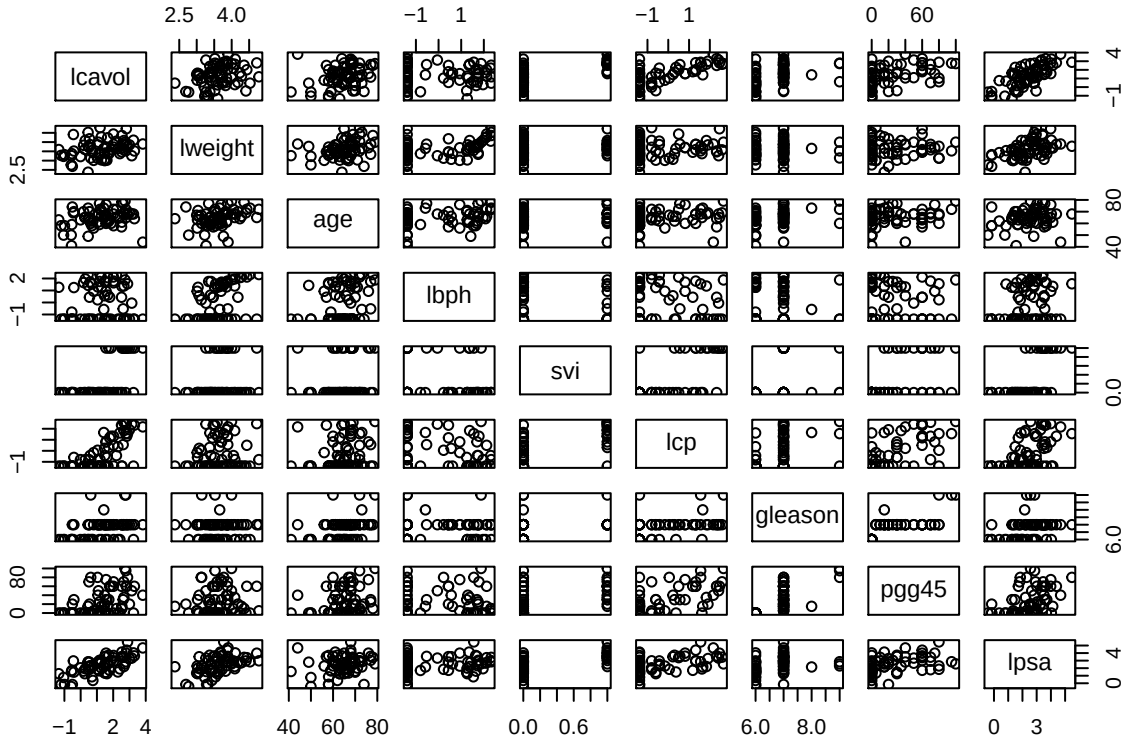
Będziemy pracować na danych z książki Elements of Statistical Learning.

```
url <- "http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data"
pcancer <- read.csv(url(url), header=TRUE, sep="\t", row.names = 1)
str(pcancer)
```

```
## 'data.frame':    97 obs. of  10 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
##  $ train  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```
train <- pcancer[which(pcancer$train),1:9]
validation <- pcancer[-which(pcancer$train),1:9]

plot(train)
```



Ze strony Trevora Hastiego możemy też ściągnąć opis danych:

```
url2=url("http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.info.txt")
cat(paste(readLines(url2), collapse="\n"))
```

```
## Prostate data info
##
## Predictors (columns 1--8)
##
## lcavol
## lweight
## age
## lbph
## svi
## lcp
## gleason
## pgg45
##
## outcome (column 9)
##
## lpsa
##
## train/test indicator (column 10)
##
## This last column indicates which 67 observations were used as the
## "training set" and which 30 as the test set, as described on page 48
## in the book.
##
## There was an error in these data in the first edition of this
## book. Subject 32 had a value of 6.1 for lweight, which translates to a
## 449 gm prostate! The correct value is 44.9 gm. We are grateful to
```

```
## Prof. Stephen W. Link for alerting us to this error.
##
## The features must first be scaled to have mean zero and  variance 96 (=n)
## before the analyses in Tables 3.1 and beyond.  That is, if x is the  96 by 8 matrix
## of features, we compute xp <- scale(x,TRUE,TRUE)
```

```
# * lcavol  : log-cancer volume
# * lweight : log-prostate weight
# * age     : age of patient
# * lbhp    : log-amount of benign hyperplasia
# * svi     : seminal vesicle invasion
# * lcp     : log-capsular penetration
# * gleason : Gleason Score, check http://en.wikipedia.org/wiki/Gleason_Grading_System
# * pgg45   : percent of Gleason scores 4 or 5
#
# And lpsa is the response variable, log-psa.
```

Polecenia:

1. Zbuduj model liniowy tłumaczący stężenie PSA (Prostate-specific antigen)
2. Jakie zmienne dostajemy dla kryterium AIC i BIC. Ile różnych modeli można zbudować w oparciu o dane zmienne? Policz wartość kryteriów informacyjnych dla wszsytkich modeli. Czy zachłanna strategia wyboru modelu okazała się skuteczna?
3. Wybierz zmienne za pomocą corss-validacji
4. Stwórz modele regresji grzbietowej i lasso. Jak wybierzesz parametr λ?
5. Porównaj wszystkie stworzone modele na zbiorze walidacyjnym. Jak duże są różnice? Który model wybierzesz?