MS-C1620 Statistical Inference

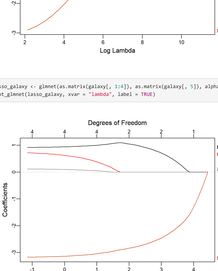
Exercise 11

Homework exercise

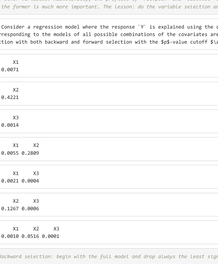
To be solved at home before the exercise session.

1. a. The data set `galaxy` from the package `ElemStatLearn` contains measurements on the position and radial velocity (the response) of the galaxy NGC7531. The following two plots show the ridge and LASSO coefficient profiles of the four explanatory variables. Compare the two plots and interpret them (for example, deduce which of the explanatory variables are the most ``important''?)

```
library(ElemStatLearn)
library(glmnet)
library(plotmo)

ridge_galaxy <- glmnet(as.matrix(galaxy[, 1:4]), as.matrix(galaxy[, 5]), alpha = 0)
plot_glmnet(ridge_galaxy, xvar = "lambda", label = TRUE)
```




```
lasso_galaxy <- glmnet(as.matrix(galaxy[, 1:4]), as.matrix(galaxy[, 5]), alpha = 1)
plot_glmnet(lasso_galaxy, xvar = "lambda", label = TRUE)
```




```
# We begin with LASSO as it is easier to interpret. Going from right to left, the variables enter the plot in the order of t
heir importance to the model fit. The most important is the north-south coordinate (`nrth.sth`) and second most important th
e radial position (`rdl.pstn`). The finnl two variables are almost equally important but the east-west coordinate (`east.wst
`) enters the model slightly before the angle.

# The interpretation of the ridge profiles is more difficult. For example, it looks like angle is the least important variab
le but as the coefficients never hit zero its small value might simply be an indication of it having a different scale than
the other variables. Additionally, the profiles of `rdl.pstn` and `east.wst` look mostly similar but the LASSO plot shows th
at the former is much more important. The lesson: do the variable selection and interpretation using LASSO.
```

b. Consider a regression model where the response `Y` is explained using the covariates `X1`, `X2` and `X3`. The p-values corresponding to the models of all possible combinations of the covariates are listed below. Use them to perform variable selection with both backward and forward selection with the p-value cutoff $\alpha_0 = 0.05$.

```
##    X1
## 0.0071
```

```
##    X2
## 0.4221
```

```
##    X3
## 0.0014
```

```
##    X1    X2
## 0.0055 0.2809
```

```
##    X1    X3
## 0.0021 0.0004
```

```
##    X2    X3
## 0.1267 0.0006
```

```
##    X1    X2    X3
## 0.0010 0.0516 0.0001
```

```
# Backward selection: begin with the full model and drop always the least significant variable until all have p-value < 0.0
5.

# X1 + X2 + X3 -> X1 + X3

# Forward selection: start with an empty model and add always the most significant new variable until no addition has p-valu
e < 0.05.

# X3 -> X1 + X3

# Both methods give the same result.
```

Class exercise

To be solved at the exercise session. Note: the R-script of lecture 10 might prove helpful in solving the below problems.

1. The data set `barro` in the package `quantreg` contains the annual GDP growth rates of several countries along with several explanatory variables. Our objective is to use variable selection to determine which factors are most helpful in predicting the growth rate of GDP.
   a. Visualize the data set.
   b. Fit a standard multiple regression to the data with `y.net` as the response.
   c. Use the function `step` to perform both backward and forward variable selection using the AIC as the criterion.
   d. Do the results of the backward and forward selections agree?
   e. Which variables would you conclude to be the most important in predicting the growth rate of GDP?

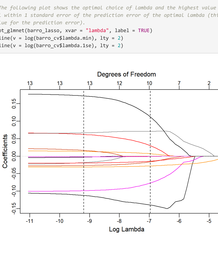
```
library(quantreg)
data(barro)

# a.
# The paired scatter plots reveal complicated relationships, not all of which are linear
# pairs(barro)

# b.
lm_barro <- lm(y.net ~ ., data = barro)
summary(lm_barro)
```

```
##
## Call:
## lm(formula = y.net ~ ., data = barro)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.037876 -0.009417  0.001339  0.009966  0.038882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.035312   0.052664  -0.671  0.50358
## lgdp2        0.027892   0.005480   5.091 1.15e-13 ***
## mse2         0.015117   0.005125   2.950  0.00370 **
## fse2        -0.004467   0.006238  -0.716  0.47503
## fhe2        -0.023602   0.027751  -0.850  0.39645
## mhe2         0.022279   0.023702   0.940  0.34871
## lexp2        0.067136   0.016310   4.116 6.39e-05 ***
## lintr2      -0.001695   0.005091  -0.333  0.73967
## gedy2       -0.104333   0.118510  -0.880  0.38009
## ly2          0.004770   0.022563   0.211  0.83270
## gcony2      -0.190927   0.038302  -4.985 1.79e-06 ***
## lblakp2     -0.031368   0.004933  -6.358 2.44e-09 ***
## pol2        -0.020233   0.006283  -3.220  0.00158 **
## ttrad2       0.178385   0.039877   4.473 1.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01651 on 147 degrees of freedom
## Multiple R-squared:  0.5924,	Adjusted R-squared:  0.5563
## F-statistic: 16.43 on 13 and 147 DF,  p-value: < 2.2e-16
```

```
# The model has good predictive power (R2 ~ 0.60) but, at least based on the p-values, seems to contain several variables wi
th no statistically significant relationship with the response. The variable selection in part c should help get rid of the u
nnecessary ones.

# c.
# The selection codes are not run here due to the enormous amount of produced output.

##### Backward selection
# We start with the full model and at each step drop the variable whose leaving out lowers the AIC most (lower AIC is bette
r).

# step(lm(y.net ~ ., data = barro), direction = "backward")


# The backward selection yields the model:
# y.net ~ lgdp2 + mse2 + lexp2 + lintr2 + ly2 + gcony2 + lblakp2 + pol2 + ttrad2


##### Forward selection
# We start with an empty model and at each step include the variable whose inclusion lowers the AIC most.

# step(lm(y.net ~ 1, data = barro),
#      scope = list(lower = lm(y.net ~ 1, data = barro),
#                   upper = lm(y.net ~ ., data = barro)),
#      direction = "forward")

# The forward selection yields the model:
# y.net ~ lblakp2 + ly2 + lgdp2 + gcony2 + lexp2 + ttrad2 + mse2 + pol2 + lintr2

# The forward selection output also provides us with an order of importance for the variables. The variable which enters the
model first is the most important (the variables are given in the output in the order of importance).

# Thus lblakp2 (Black Market Premium) is the most important variable



# d.
# The results of the backward and forward selection agree. This is a good sign and makes the result more reliable (two strat
egies arrived at the same result).

# e.
# The important variables are:
# Black Market Premium
# Investment/GDP
# Initial Per Capita GDP
# Public Consumption/GDP
# Life Expectancy
# Growth Rate Terms Trade
# Male Secondary Education
# Political Instability
# Human Capital
```

2. We continue the analysis of problem 1.
   a. Fit a LASSO model to the `barro` data set.
   b. Plot the LASSO coefficient profiles. Which variable does LASSO hold the most important?
   c. Use 10-fold cross validation to choose a suitable value for the parameter λ. Which variables are included in the corresponding model?

```
library(glmnet)
library(plotmo)

# a.
barro_lasso <- glmnet(as.matrix(barro[, 2:14]), as.matrix(barro[, 1]), alpha = 1)

# b.
plot_glmnet(barro_lasso, xvar = "lambda", label = TRUE)

# Reading from right to left, the variable (blakp2 (Black Market Premium) enters the model first and is thus the most import
ant. The second most important would be ly2 (Investment/GDP)

# c.
# There is an element of randomness in the results of the cross validation (the division of the data into the folds is done
randomly) and thus we fix the random seed below.
set.seed(08032019)
barro_cv <- cv.glmnet(as.matrix(barro[, 2:14]), as.matrix(barro[, 1]), alpha = 1, nfolds = 10)

# The following plot shows the optimal choice of lambda and the highest value of lambda for which the prediction error is st
ill within 1 standard error of the prediction error of the optimal lambda (this can still be considered a "reasonably good"
value for the prediction error).
plot_glmnet(barro_lasso, xvar = "lambda", label = TRUE)
abline(v = log(barro_cv$lambda.min), lty = 2)
abline(v = log(barro_cv$lambda.1se), lty = 2)
```




```
# To obtain a more narrow set of variables we take the 1 standard-error lambda and arrive at the model:
barro_lasso_1se <- glmnet(as.matrix(barro[, 2:14]), as.matrix(barro[, 1]), alpha = 1, lambda = barro_cv$lambda.1se)
coef(barro_lasso_1se)
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  0.023367695
## lgdp2       -0.016072043
## mse2         0.007882577
## fse2         .
## fhe2         .
## mhe2         .
## lexp2        0.030585681
## lintr2      -0.009922819
## gedy2       -0.119026661
## ly2          0.071141466
## gcony2      -0.076557654
## lblakp2     -0.027925369
## pol2        -0.014070818
## ttrad2       0.112016067
```

```
# Thus LASSO retains 10 variables in total. These include the same 9 as obtained with backward and forward selection and add
itionally also Education/GDP.
```

3. **(Optional)** Investigate how ridge regression and LASSO perform in the presence of ``noise'' variables. That is, simulate data where the response depends linearly on a few explanatory variables but include in the model also several explanatory variables (noise) which are independent of the response. Plot then the ridge and LASSO profiles of the variables. Do the profiles of the noise variables perform as expected?