Compulsory exercise 2, version 07.03.2018

TMA4268 Statistical Learning V2018

Contact person: Thea Roksvåg, thea.roksvag@ntnu.no

To be handed in on Blackboard: deadline March 16 at 16.00

Version 07.03: corrected Q18 Europe coding

Maximal score is 10 points. You need a score of 4/10 for the exercise to be approved. Your score will make up 10% points of you final grade.

Supervision:

- Fridays 12-14 in Smia, in addition
- Monday March 12, 12-14 in R9,
- Wednesday March 14, 10-12 in Smia.

Practical issues:

- Maximal group size is 3 - join a group (self enroll) before handing in on Blackboard.
- Remember to write your names and group number on top of your submission.
- The exercise should be handed in as one R Markdown file and a pdf-compiled version of the R Markdown file (if you are not able to produce a pdf-file directly please make an html-file, open it in your browser and save as pdf - no, not landscape - but portrait please). We will print the pdf-files (and you get written comments) and use the Rmd file in case we need to check details in your submission.
- In the R-chunks please use both `echo=TRUE` and `eval=TRUE` to make it simpler for us to read and grade.
- Please do not include all the text from this file (that you are reading now) - we want your R code, plots and written solutions - see the template below.
- Please not more than 10 pages in your pdf-file! (This is a request, not a requirement.)
- Please save us time and NOT submit word or zip - or only Rmd - that only results in extra work for us!

Template file for submission of Compulsory exercise 2 is here: https://www.math.ntnu.no/emner/TMA4268/2018v/CompEx2mal.Rmd

You need to install the following packages in R:

```
install.packages("ISLR")
install.packages("ggplot2")
install.packages("GGally")
install.packages("leaps")
install.packages("glmnet")
install.packages("gam")
```

Problem 1 - Model selection and cross-validation

a) Theory [1 point]

- Q1. We will study a linear regression model with intercept β_0 present. We consider d possible predictors. How many different linear regression models can be fitted?
- Q2. Explain in a few words (presented as an algorithm) how we can use the best subset method with the BIC criterion to choose the best model out of all the possible models. What would happen if we instead used R^2 to choose the best model?

Hint: pages 244-247 in our textbook "Introduction to Statistical Learning".

b) Interpreting output [1 points]

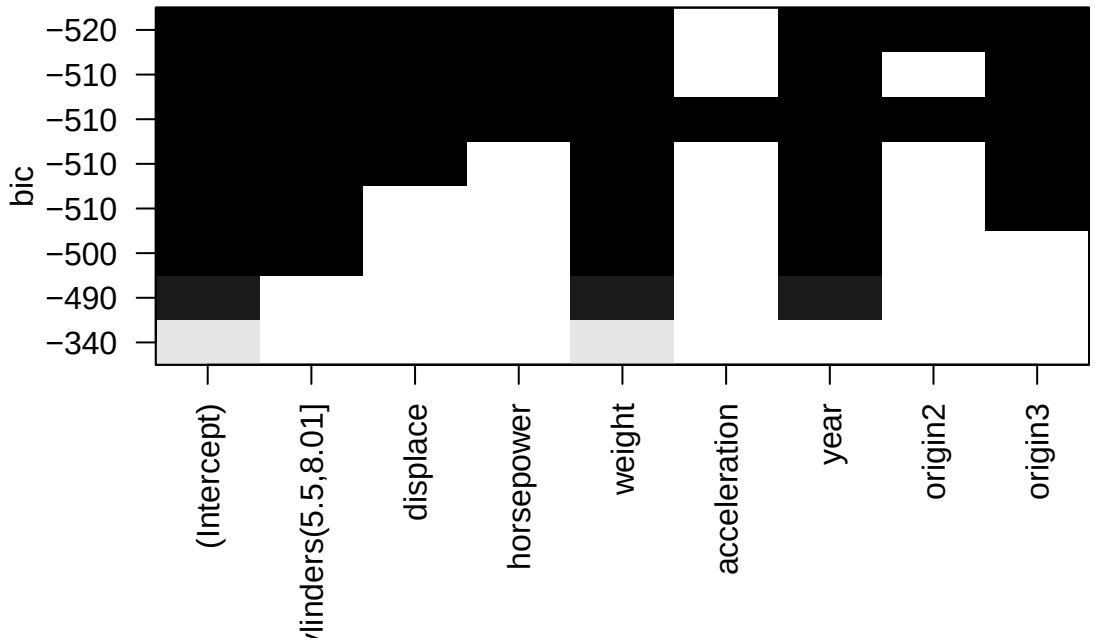
We will now study the `Auto` data set in the `ISLR` package. We will use `mpg` (miles pr gallon) as response, and selected covariates (see below, observe that we have recoded the `cylinders` into two groups, and that `origin` is a categorical covariate). We set aside 20% of the data (`ourAutoTest`) to be used to assess the selected model(s), and use the rest for model selection (`ourAutoTrain`).

```
library(ISLR)
ourAuto=data.frame("mpg"=Auto$mpg,"cylinders"=factor(cut(Auto$cylinders,2)),
                   "displace"=Auto$displacement,"horsepower"=Auto$horsepower,
                   "weight"=Auto$weight,"acceleration"=Auto$acceleration,
                   "year"=Auto$year,"origin"=as.factor(Auto$origin))
colnames(ourAuto)
ntot=dim(ourAuto)[1]
ntot
set.seed(4268)
testids=sort(sample(1:ntot,ceiling(0.2*ntot),replace=FALSE))
ourAutoTrain=ourAuto[-testids,]
ourAutoTest=ourAuto[testids,]
```

```
## [1] "mpg"          "cylinders"    "displace"     "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [1] 392
```

Below we have performed best subset selection with the BIC criterion using the function `regsubsets` from the `leaps` package.

```
library(leaps)
res=regsubsets(mpg~.,nbest=1,data=ourAutoTrain)
sumres=summary(res)
sumres
plot(res,scale="bic")
```




```
sumres$bic
which.min(sumres$bic)
```

```
coef(res,id=which.min(sumres$bic))
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., nbest = 1, data = ourAutoTrain)
## 8 Variables  (and intercept)
##                      Forced in Forced out
## cylinders(5.5,8.01]     FALSE      FALSE
## displace                FALSE      FALSE
## horsepower              FALSE      FALSE
## weight                  FALSE      FALSE
## acceleration            FALSE      FALSE
## year                    FALSE      FALSE
## origin2                 FALSE      FALSE
## origin3                 FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          cylinders(5.5,8.01] displace horsepower weight acceleration year
## 1  ( 1 ) " "                 " "       " "        "*"    " "          " "
## 2  ( 1 ) " "                 " "       " "        "*"    " "          "*"
## 3  ( 1 ) "*"                 " "       " "        "*"    " "          "*"
## 4  ( 1 ) "*"                 " "       " "        "*"    " "          "*"
## 5  ( 1 ) "*"                 "*"       " "        "*"    " "          "*"
## 6  ( 1 ) "*"                 "*"       "*"        "*"    " "          "*"
## 7  ( 1 ) "*"                 "*"       "*"        "*"    " "          "*"
## 8  ( 1 ) "*"                 "*"       "*"        "*"    "*"          "*"
##          origin2 origin3
## 1  ( 1 ) " "     " "
## 2  ( 1 ) " "     " "
## 3  ( 1 ) " "     " "
## 4  ( 1 ) " "     "*"
## 5  ( 1 ) " "     "*"
## 6  ( 1 ) " "     "*"
## 7  ( 1 ) "*"     "*"
## 8  ( 1 ) "*"     "*"
## [1] -343.1283 -494.5988 -501.7008 -505.3911 -505.5065 -514.2228 -516.9417
## [8] -512.8292
## [1] 7
##         (Intercept) cylinders(5.5,8.01]           displace
##       -19.410696099        -3.533041185         0.032788184
##          horsepower              weight               year
##        -0.048833206        -0.005823726         0.784897712
##             origin2             origin3
##         1.794007415         2.906285958
```

- Q3. How is the best model for each model complexity (number of parameters estimated) chosen? Write down the best model with 2 covariates (in addition to the intercept).
- Q4. How can you choose between models of different model complexity? According to the BIC criterion, which is the best model? Fit this best model on `ourAutoTrain` and comment on the model fit. Report the MSE on `ourAutoTrain`.
- Q5: Use this model fit to predict new values for `ourAutoTest` and report the MSE.

c) Cross-validation [1 point]

- Q6. Explain how k-fold cross-validation is performed (in a regression setting).
- Q7. Why may k-fold cross-validation be preferred to leave-one-out cross-validation?

d) Programming 10-fold cross-validation in R [2 points]

Refer to pages 248-249 in our textbook "Introduction to Statistical Learning", and also solutions to Recommended Problem 5 in Module 6 for hints on how to make the R code.

- Q8. Write R code to perform 10-fold cross-validation with the best subset method on `ourAutoTrain` (no, do not use `ourAutoTest` here). (Hint: smart to make a `predict.regsubsets` function as on page 249 in "Introduction to Statistical Learning".)
- Q9. What is the optimal model complexity (number of parameters) in your regression?
- Q10. Use the model complexity you found in Q9 to find the best model on the `ourAutoTrain`. Report interesting features of this final model. Also use this model fit to predict new values for `ourAutoTest` and report the MSE. (If you get the same best model as Q4, just refer back to Q4 and Q5.)

Problem 2 - Shrinkage methods

In this exercise we will study lasso and ridge regression. We continue using the `ourAutoTrain` dataset from Problem 1.

a) Lasso and ridge regression [2 points]

In a regression model with p predictors the ridge regression coefficients are the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

while the lasso regression coefficients are the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

In Figure 1 and Figure 2 you see the results from lasso and ridge regression applied to `ourAutoTrain`. Standardized coefficients $\hat{\beta}_1, ..., \hat{\beta}_8$ are plotted against the tuning parameter λ.

- Q11. Which figure (1 or 2) corresponds to ridge and which figure corresponds to lasso? Justify your answer.
- Q12. Use the two figures and the above formulas to explain the impact of the tuning parameter λ on the coefficients β_j, and on the bias and variance of the resulting predictions. In particular, what happens when $\lambda = 0$ and when $\lambda \to \infty$?
- Q13. Can you use lasso and/or ridge regression to perform model selection similar to what you did in Problem 1? Explain. Compare what you see in Figure 1 and Figure 2 to the results in Problem 1b.

b) Finding the optimal λ [1 point]

In the following, we will use functions in the `glmnet` package to perform *lasso* regression. The first step is to find the optimal tuning parameter λ. This is done by cross-validation using the `cv.glmnet()` function:

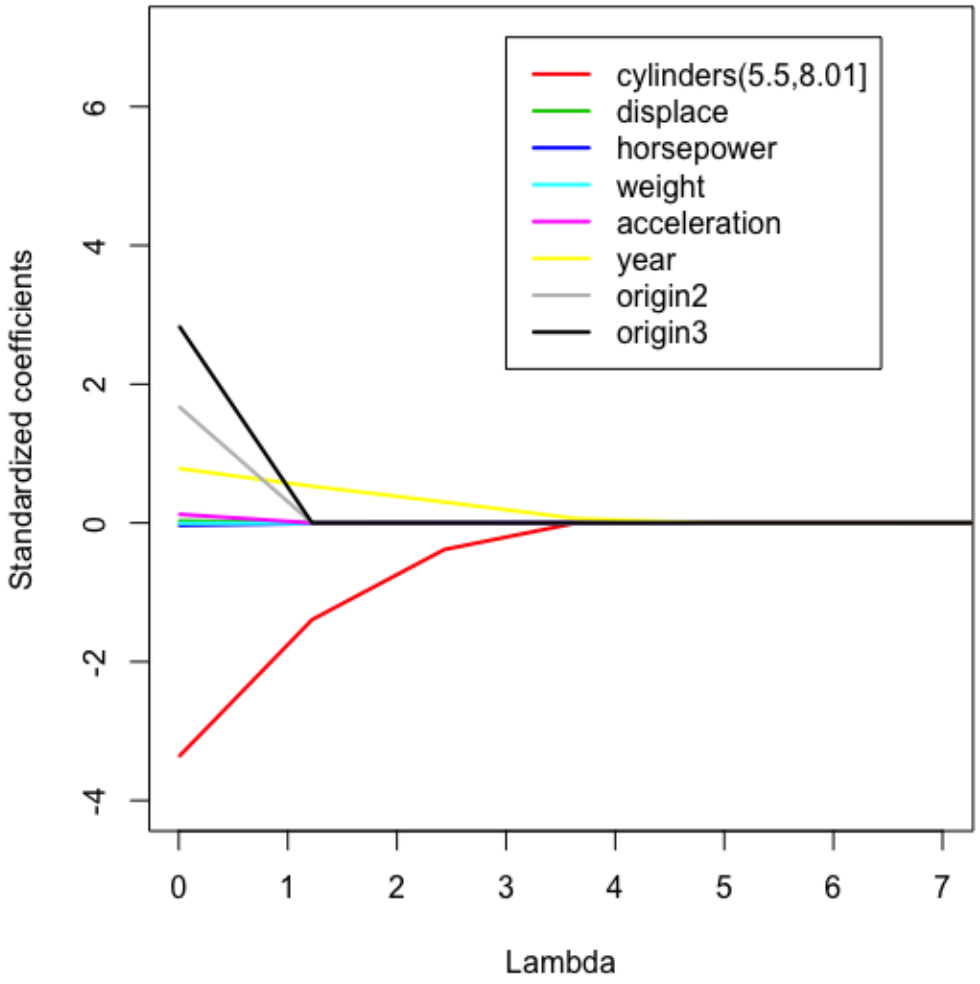


Figure 1: Figure 1.

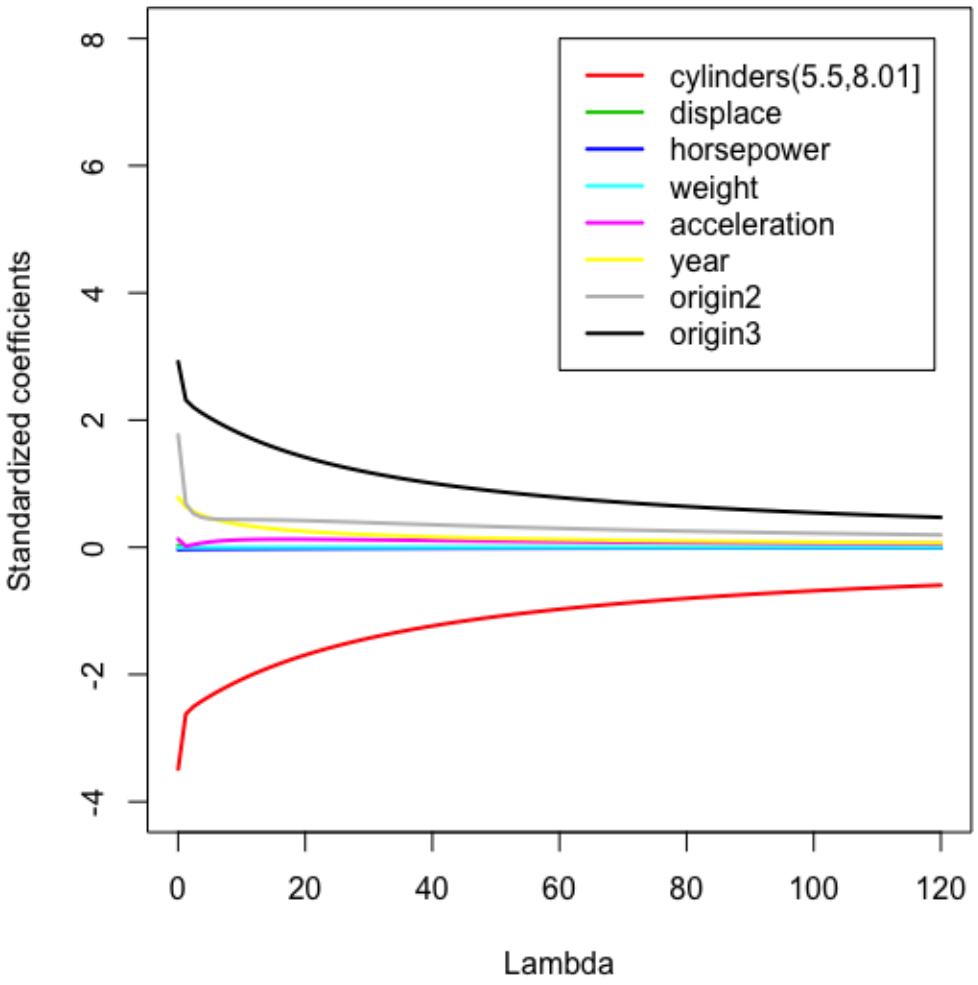


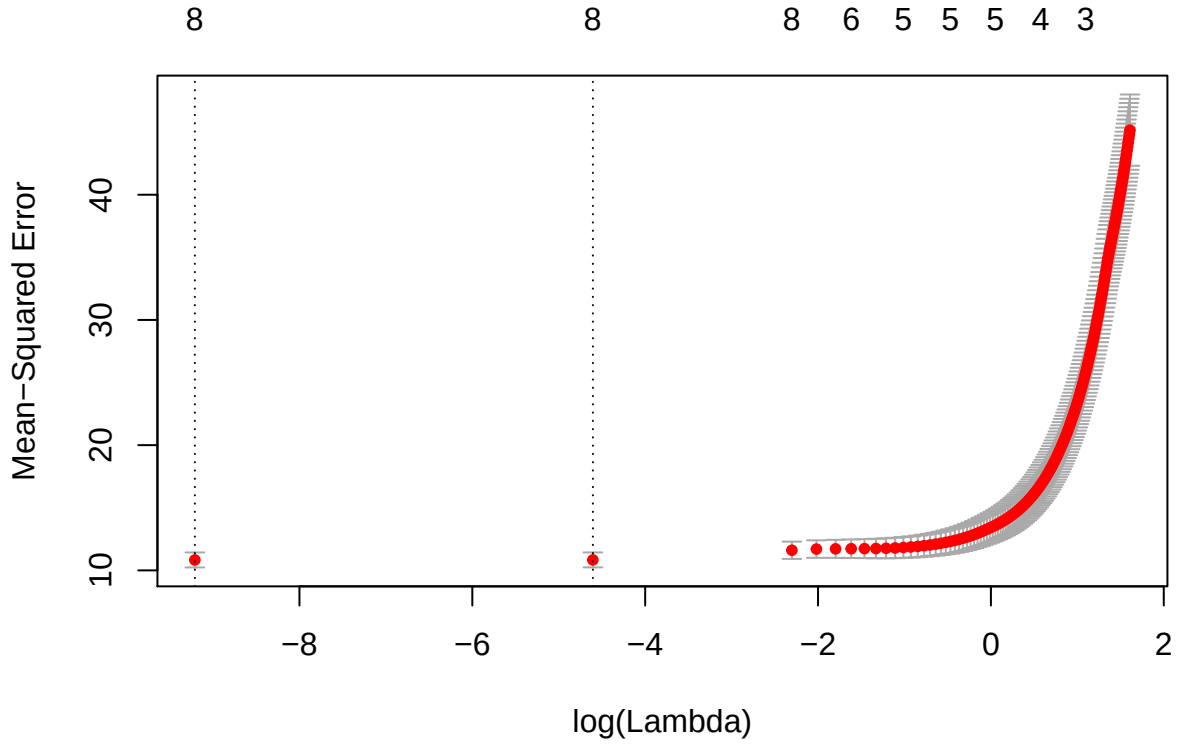
Figure 2: Figure 2.

```r
library(glmnet)
set.seed(4268)

x=model.matrix(mpg~.,ourAutoTrain)[,-1] #-1 to remove the intercept.
head(x)
y=ourAutoTrain$mpg

lambda=c(seq(from=5,to=0.1,length.out=150),0.01,0.0001) #Create a set of tuning parameters, adding low
cv.out=cv.glmnet(x,y,alpha=1,nfolds=10,lambda=lambda, standardize=TRUE) #alpha=1 gives lasso, alpha=0 g

plot(cv.out)
```




```
##   cylinders(5.5,8.01] displace horsepower weight acceleration year origin2
## 1                   1      307        130   3504         12.0   70       0
## 2                   1      350        165   3693         11.5   70       0
## 4                   1      304        150   3433         12.0   70       0
## 5                   1      302        140   3449         10.5   70       0
## 8                   1      440        215   4312          8.5   70       0
## 9                   1      455        225   4425         10.0   70       0
##   origin3
## 1       0
## 2       0
## 4       0
## 5       0
## 8       0
## 9       0
```

- Q14. Explain what the function `cv.glmnet` does. Hint: `help(cv.glmnet)`.
- Q15. Explain what we see in the above plot. How can it be used to identify the optimal λ? Remark: To find the optimal λ a popular choice is to choose the λ giving the lowest cross-validated MSE. Another choice is called the `1se`-rule. See `help(cv.glmnet)`.

- Q16. Use the output from `cv.glmnet` and the `1se-rule` to choose the "optimal"" λ.

c) Prediction [1 point]

- Q17. Use lasso regression to fit the model corresponding to the optimal λ from Q16. What are the coefficient estimates? Write down the model fit.
- Q18. Assume that a car has 4 cylinders, `displace=150`, `horsepower=100`, `weight=3000`, `acceleration=10`, `year=82` and comes from Europe. What is the predicted `mpg` for this car given the chosen model from Q17? Hint: you need to construct the new observation in the same way as observations in the model matrix `x` (the dummy variable coding for cylinders and origin) and `newx` need to be a matrix `newx=matrix(c(0,150,100,3000,10,82,1,0),nrow=1)`.

Problem 3 - Additive non-linear regression

In this exercise we take a quick look at different non-linear regression methods. We continue using the `ourAutoTrain` dataset from Problem 1 and 2.

a) Additive model [1 points]

- Q19: Fit an additive model using the function `gam` from package `gam`. Call the result `gamobject`.
  - `mpg` is the response,
  - `displace` is a cubic spline (hint: `bs`) with one knot at 290,
  - `horsepower` is a polynomial of degree 2 (hint: `poly`),
  - `weight` is a linear function,
  - `acceleration` is a smoothing spline with `df=3`,
  - `origin` is a step function (what we previously have called dummy variable coding). Plot the resulting curves (hint: first set `par(mfrow=c(2,3)` and then `plot(gamobject,se=TRUE,col="blue")`). Comment on what you see.
- Q20: Write down a basis for the cubic spline (`displace`).