Lasso estimation in multiple linear regression

Write a report that contains the results of the computations that you are asked to carry out below, as well as the explanation of what you are doing. The main text should include pieces of source code and graphical and numerical output. Upload your answers in a .pdf or .html document (use R Markdown, for instance), as well as the source code (*.R or *.Rmd, for instance). Your work must be reproducible.

---

1. Lasso for the Boston Housing data

The Boston House-price dataset concerns housing values in 506 suburbs of Boston corresponding to year 1978. They are available here:

`https://archive.ics.uci.edu/ml/datasets/Housing`

This is the list of the available variables:

```
1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's
```

The Boston House-price corrected dataset (available in `boston.Rdata`) contains the same data (with some corrections) and it also includes the UTM coordinates of the geographical centers of each neighborhood.

For the Boston House-price corrected dataset, use Lasso estimation (in `glmnet`) to fit the regression model where the response is `CMEDV` (the corrected version of `MEDV`) and the explanatory variables are the remaining 13 variables in the previous list. Try to provide an interpretation to the estimated model.

2. A regression model with $p \gg n$

The following dataset is taken from the book of Hastie, Tibshirani and Friedman (2009, Section 18.6; see there for the original reference) and consists of 240 samples from patients with diffuse large B-cell lymphoma (DLBCL), with gene expression measurements for 7399 genes. The outcome is survival time, either observed or right censored.

Even if it is not the best way to analyze these data, we propose a multiple linear regression model for the $n = 138$ patients died before the end of the study, to explain the logarithm of the survival time as a linear function of the expressions of the $p = 7399$ genes. Here you have the code to read the data:

```
express <- read.csv("journal.pbio.0020108.sd012.CSV",header=FALSE)
surv <- read.csv("journal.pbio.0020108.sd013.CSV",header=FALSE)
death <- (surv[,2]==1)
log.surv <- log(surv[death,1]+.05)
expr <- as.matrix(t(express[,death]))
```

1. Use `glmnet` and `cv.glmnet` to obtain the Lasso estimation for regressing `log.surv` against `expr`. How many coefficient different from zero are in the Lasso estimator? Illustrate the result with two graphics.

2. Compute the fitted values with the Lasso estimated model (you can use `predict`). Plot the observed values for the response variable against the Lasso fitted values.

3. Consider the set S_0 of non-zero estimated Lasso coefficients. Use OLS to fit a regression model with response `log.surv` and explanatory variables the columns of `expr` with indexes in S_0. Plot the observed values for the response variable against the OLS fitted values.

4. Compare the OLS and Lasso estimated coefficient. Compare the OLS and Lasso fitted values. Do a plot for that.