Ridge 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 (2 or 3 pages) should include pieces of source code and graphical and numerical output.
Upload your answers in a .pdf document (use LaTeX or R Markdown, for instance), as well as the source code (*.R or *.Rmd, for instance). Your work must be reproducible.

1. Choosing the penalization parameter λ

1. Write an R function implementing the ridge regression penalization parameter λ choice based on the minimization of the mean squared prediction error in a validation set ($\text{MSPE}_{\text{val}}(\lambda)$).

   Input: Matrix x and vector y corresponding to the *training sample*; matrix x_{val} and vector y_{val} corresponding to the *validation set*; a vector `lambda.v` of candidate values for λ.

   Output: For each element λ in `lambda.v`, the value of $\text{MSPE}_{val}(\lambda)$.

   Additionally you can plot these values against $\log(1+\lambda)-1$, or against $\text{df}(\lambda)$.

2. Write an R function implementing the ridge regression penalization parameter λ choice based on k-fold cross-validation ($\text{MSPE}_{k-\text{CV}}(\lambda)$).

   Input, output and graphics as before (except that x_{val} and y_{val} are not required now as input).

3. Consider the prostate date used in class. Use your routines to choose the penalization parameter λ by the following criteria: behavior in the validation set (the 30 observations not being in the training sample); 5-fold and 10-fold cross-validation. Compare your results with those obtained when using leave-one-out and generalized cross-validation.

2. Ridge regression 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 at the library `MASS`,

```
library(MASS)
data(Boston)
help(Boston)
```

and also here:

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

This is the list of the 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 ridge regression to fit the regression model where the response is `MEDV` and the explanatory variables are the remaining 13 variables in the previous list. Try to provide an interpretation to the estimated model.