Part II, assignment 1

For this assignment you will use the dataset "prostate" (prostate2.Rdata). It contains data about prostate cancer patients with information on the size of the prostate, the age of the patient, a blood marker (lpsa) and so on. The response variable is a score (Cscore) on the progression of the cancer after detailed study of the tumor pathology. In this exercise you will **compare the performance of different linear modelling procedures** and compare the models they produce. You need to upload answers to the questions including **an explanation** of how you came to the answer, but also your R code so that the results can be reproduced. You can use labs of ISL chapter 6 for guidance. Some hints:

- For lasso/ridge use library `glmnet` in R, for PCR use library `pls`:

  `install.packages(c("glmnet", "pls")) library(glmnet) library(pls)`
- In the `glmnet` option alpha=0 corresponds to ridge and alpha=1 to lasso.
- `glmnet` requires the data in matrix format, pls doesn't
- Use the function predict to measure performance

  `Ypred = predict(model,newdata=data[test,],type="response")` for `glm`

  `Ypred = predict(model,newx=Xtest,s=lambda,type="response")` for `glmnet`

  `Ypred = predict(model,data[test,],ncomp=N,type="response")` for `pls`
- You can see the ridge/lasso coefficients with `predict(model,type="coefficients",s=lambda)` in `glmnet`

1. Evaluate and compare the variables. How many observations and predictor variables are there? What relevant characteristics do you note on the variables? For example, are there strong correlations between the variables? Do you see any issues with these variables that are relevant for modelling and making predictions?
2. Select an optimal linear model using Forward Selection. Implement algorithm 6.2 from the ISLR book yourself, and choose the appropriate performance measures to compare and choose models in the different stages of the algorithm. Report the result you obtain.
3. Make an appropriate Lasso model. Perform a cross validation to select an appropriate lambda value.
   - How do you interpret the plot of the lambda value vs cross validation error? Do you think the Lasso fit improves importantly over an Ordinary Least Squares linear model, with respect to the prediction performance? Is there a rationale for using Lasso in this context? Explain.
   - Write down the equation for the resulting Lasso model.
4. Repeat item 2 for Ridge.
5. Make a crossvalidated PCR model: How much of the variability is explained by each principal component?
   - How many principal components would you select for your PCR model?
   - How appropriate is PCR for this dataset?
6. Compare all the models above: forward selection, lasso, ridge, pcr.
   - How well do they perform in general?
   - Do they yield similar models?
   - Which variables are the most important in the prediction of the progression of prostate cancer, according to each model?
   - Which model do you think is the most appropriate for this dataset?