Exercise 1

We consider the dataset diabetes (Efron, Hastie, Johnstone and Tibshirani (2003) "Least Angle Regression" Annals of Statistics) from the package `lars`. Ten baseline variables age, sex, body mass index (bmi), average blood pressure (map) and six blood serum measurements (tc, ldl, hdl, tch, ltg, glu), as well as disease progression one year after baseline (`y`), were obtained for n=442 diabetes patients. The baseline data is stored in `x` while a model matrix including interactions between baseline measurements is stored in `x2`. Here, we aim to predict the disease progression, one year after baseline based on the matrix x2. You can access the data via

```
# install.packages("lars")
library(lars)

## Warning:  package 'lars' was built under R version 3.4.4
## Loaded lars 1.2

data("diabetes")
```

(a) Split the data set into a training and a test set. Sample 70% of the data to the training set, 30% to the test set. Set the seed to 100 (`set.seed(100)`).

(b) Fit a linear regression model based on the training data and check the model assumptions. Is it important that all the assumptions are met? Now, use the model for prediction on the test data. Calculate the test error in terms of the mean squared error (MSE) and the mean absolute prediction error (MAPE) using OLS. Consider the predicted vs. the observed values on the test data.

(c) Now, we aim to predict the test data using ridge regression. Recall what ridge regression is doing and what's the impact of λ. Perform a cross validation to find the best parameter λ based on the training data. Is the model, fitted with the optimal parameter λ, a better prediction model for the test data compared to the linear regression model? Plot the predicted vs. the observed values.

(d) Now, we aim to predict the test data using a lasso regression. What's the difference to the ridge regression? Perform a cross validation on the training data to find the best parameter λ (`cv.glmnet(..., alpha=1)`). Is the model, fitted with the optimal parameter λ, a better prediction model than the linear and the ridge regression? Plot the predicted vs the observed values.

(e) Calculate the predictions on the training data for each of the three models. Which model fits best? Do the results make sense?