Exercises, Lecture 4

1. Logistic regression: Given a logistic model for lung cancer (yes/no) as a function of smoking (number of cigarettes per day) with β=0.02. Show that one units increase in smoking means an increase in lung cancer risk (odds-ratio) of $\exp(0.02)=1.02=2\%$.

2. We have a data material (Golub et al 1999) with gene expression levels from 72 patients with two forms of leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Gene expression levels (how actively the cells are using the information in different genes) are measured for 7128 genes. We would like to build a biomarker for classification of the two cancer forms. Ideally we would like to use only a few variables.

   a. How can you use logistic regression here?
   b. Build a classifier for training data in `GolubGXtrain.csv.` What regularization method do you prefer if you want to have few genes in the biomarker?
   c. How many variables do you end up with?
   d. Use the obtained model to calculate accuracy on the test data.

MATLAB:

`csvread` for reading files

`lassoglm` calculates regularized logistic regression

`glmval` for predictions

R:

```
library(glmnet) #logistic regression
read.csv('Name of file')
```

Python: