contact: Ricardo de Matos Simoes; r.dematossimoes@qub.ac.uk
Peter Hamilton; P.Hamilton@qub.ac.uk

1 Row and Column Picture of a linear model

- solve the following linear system

$$\begin{aligned} 2x + y &= 7 \\ 2x + 3y &= 1 \end{aligned} \tag{1}$$

- Draw the row picture of the linear system
- Draw the column picture of the linear system

2 Data

All data is available at link `http://go.qub.ac.uk/toolkit/regularization`

3 Linear regression model

- Download and read the prostate cancer dataset *prostate.data* into a data matrix. The data set is taken from the free online book *The Elements of Statistical Learning* from Trevor Hastie, Robert Tibshirani and Jerome Friedman. `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data`

```
Prostate data info

Predictors (columns 1--8)

lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45

outcome (column 9)

lpsa
```

```
train/test indicator (column 10)
```

- Estimate and define a linear model using the *lm()* function.
- Split the prostate dataset into a test and training dataset (see column 10)
- Predict *lpsa* for the test examples

```
data<-read.table(file="...")
# split data
test.data<- ...
train.data<- ...
```

- Plot your results (true lpsa value y_i and predicted $\hat{y}_i$).
- Repeat the analysis using a random forest regression and plot your results (true lpsa value y_i and predicted $\hat{y}_i$).

4 Implement your own lm function

- Write a R function that estimates $\hat{\beta}$ coefficients for the linear model:

$$y = \beta_0 + \beta_{pred1}x_1 + \beta_{pred2}x_2 + \beta_{pred3}x_3 \tag{2}$$

  Remember to center scale the predictor variables and estimate β_0 separately. You can use the function *scale()*

- The coefficients of a linear model can be estimated by

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

5 Ridge Regression

- Write a function to estimate the coefficients using a ridge regression model.

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \tag{4}$$

  where I is the identity matrix and λ the penalty parameter.

- Implement a function that returns the optimal λ by a 10 fold cross validation. The function minimizes the prediction error measure (sum of squared error)

$$MSE = \sum (y - \hat{y})^2 \tag{5}$$

```
rss.error=sum((test$lpsa-pred)^2)
```

- Apply your function to the prostate dataset and report the model coefficients and optimal λ.