STAT40730/STAT40620
Data Programming with R.
Assignment 2.

Instructions

- This assignment is due at the end of week 10, and is worth 5% of your final grade.
- You should submit a single R script file to the 'Assignment 2' assignment object in Blackboard.
- The marks available for each question are shown in brackets.
- You may have to learn and discover some new functions. Use `help()` and `help.search()` to find what you need.
- Make sure your R script file contains all the commands used to complete the tasks, and that it includes lots of comments (starting with #) so that the file is readable.
- Your R script file should include test code where appropriate.
- Assignment 2 is broken up into 3 tasks: statistical modelling, linear algebra and S3 classes. In the first task you are going to fit logistic regression models using 10-fold cross-validation. In the second task you will perform block-wise matrix inversion. In the third task you will write some methods for different S3 classes.

Task 1: statistical modelling (15 marks).

- We have met the technique of logistic regression in the extended examples in a few lectures. A logistic regression model can be fitted in R with the function `glm()`. One way of testing how well a model fits the data is to remove part of the data (i.e. remove some observations), fit the model to the remaining data and then use the model to 'predict' the left out data. If the model fits well then the predictions should match the real left out data. In the context of logistic regression 'predict' means predicting the binary response variable, based on the explanatory variable(s) of each left out observation. A popular version of this technique is known as *10-fold cross validation*, where the data are broken up into 10 parts. We remove one part, fit the model to the remaining 9 parts, then predict for the left out data, and repeat this for each part in turn. We can then compute a measure of our model fit e.g. by computing the *misclassification rate*.

- Your task is to write a function which takes a binary response variable y and a single explanatory variable x (of type `factor` or `numeric`), runs 10-fold cross validation and returns the proportion of the response variables y that are incorrectly classified (i.e. the misclassification rate).

- Use your function on the birthweight data to assess which of the explanatory variables (age, mother's weight, race, etc) performs best at predicting low birth weight (i.e. has the lower misclassification rate).
- Repeat your analysis for the South African heart data (found in the Blackboard folder).

Task 2: matrix inversion (20 marks).

- A useful technique for inverting matrices is that of the block-wise inversion method. A description can be found at the Wikipedia page: `https://en.wikipedia.org/wiki/Matrix_inversion#Blockwise_inversion`
- Your task is to write a recursive function which inverts any matrix using the block-wise inversion technique. Note that a 2×2 matrix can be inverted analytically; see the same Wikipedia page (inversion of 2×2 matrices). Your function must not use the `solve` function (or any other inbuilt R matrix inversion/decomposition function).
- Test your function on some simple matrices, such as:

```
M = matrix(rnorm(2^2),2,2)
M2 = matrix(rnorm(5^2),5,5)
M3 = matrix(rnorm(150^2),150,150)
```

  Your function should produce the same values as `solve(M)`, `solve(M2)`, etc.

Task 3: writing S3 methods (25 marks).

This task involves writing different methods for an object of S3 class. You should include suitable examples of your working code. This task is broken down into 3 parts:

- Write a print method for the function you created in Task 1. The output should include (at least) the misclassification rate and a misclassification table of predicted y versus true y.
- Write a summary method for the `findwords` function of lecture 3. The summary should include (at least) the total number of words and the top 5 words used in the paragraph.
- Write a plot method for the `polyreg` class created in lecture 8. The plot should show the data together with all the fitted lines and a legend.

Hints

A non-exhaustive list of some functions which might be useful are given below.

Function	Details
`glm(y ~ x,family="binomial")`	Fits a logistic regression model for response variable y and explanatory variable x (of type `factor` or `numeric`).
`predict`	Predicts new values of y from a statistical model such as a logistic regression. Note that the argument `type ="response"` gives estimated probabilities which can be rounded to the nearest whole number to give the predicted class.
`det`	Finds the determinant of a matrix. A matrix with determinant 0 is not invertible.
`system.time`	Reports how long a function takes to run.
`class`	Sets a class for an object.