Biostat 625 Homework #1

Submit as a compressed file "hw1.tar.gz" containing the required computer code for the following problems. Note that:

- Follow exactly the requirements stated in the problems, as your code will be tested automatically by another program, which cannot accommodate human errors such as typos in file or function names. In particular, do not include any unnecessary test code or printout.
- Consider all possible test cases and make your code as robust as you can, as all potential input arguments are allowed as long as they are not in contradiction to the problem description.
- Make your code as efficient as you can, as some of the test cases may be computationally challenging and your code will be terminated if it does not finish after running for 10 seconds and you will lose the points for those test cases. It will also be terminated if it uses more than 1GB of memory. However, you should always turn in your code even if it is not very efficient so that you can at least get partial credits.
- All your code will be tested on the biostat cluster.

Problem 1 - Solve quadratic equation

Write an R function named "quadratic", which takes arguments of three numbers a, b and c, and returns the smallest possible solution of the equation $ax^2 + bx + c = 0$. Arguments a, b and c will be chosen so that there exists at least one solution. Save your function in a file named "quadratic.R". Example runs are given below.

```
> quadratic(1, 4, 1)
[1] -3.732051
> quadratic(1, -2, 1)
[1] 1
```

Hint: write your code to avoid issues during floating point calculations.

Problem 2 - Generalized logit model

In a generalized logit model (a.k.a. multinomial logistic regression), the outcome Y is a categorical variable with $K + 1$ levels: $0, 1, \ldots, K$, where 0 is taken as the reference level. The relationship between Y and covariates $\mathbf{X}$ is modeled as

$$\ln \frac{Pr(Y = k)}{Pr(Y = 0)} = \boldsymbol{\beta}_k^T \mathbf{X}$$

where $\mathbf{X} \in R^p$ and $\boldsymbol{\beta}_k \in R^p, k = 1, 2, \ldots, K$ are the regression coefficients.

Write an R function named "gen_logit", which takes arguments of a $K \times p$ matrix $\boldsymbol{\beta}$ and a length p vector $\mathbf{x}$, and returns the vector of probabilities of Y being in each level, i.e., $[Pr(Y = 0), Pr(Y = 1), \ldots, Pr(Y = K)]$. Save your function in a file named "gen_logit.R". Example runs are given below.

```
> gen_logit(matrix(0, 2, 2), c(1, 0))
[1] 0.3333333 0.3333333 0.3333333
> gen_logit(matrix(1:4, 2, 2), c(1, 1))
[1] 0.002178521 0.118943236 0.878878243
```

Hint: write your code to avoid issues during floating point calculations.

Problem 3 - Count the number of lines of R code

Write a bash shell script named `count_lines`, which takes an argument of an R code file name, and prints the number of lines of "real" R code in the file. That is, lines with only white spaces or comments should not be counted. An example run is given below.

```
$ cat test.R
n = 10

#this is a loop
for (i in 1:n) {
        print(i)
}
$ bash ./count_lines test.R
4
```

Hint: regular expression may be helpful.

Problem 4 - Is it a number

Write a bash shell script named `is_number`, which takes an argument of a string, and prints "YES" if the string is a number in canonical format, and "NO" if not. Here numbers in canonical format are those printed out by R function `print()`. You do not need to consider numbers in scientific notation such as `1e+10`. Example runs are given below.

```
$ bash ./is_number 123
YES
$ bash ./is_number 003
NO
```

Hint: regular expression may be helpful.