Bios 301: Assignment 2

Due Monday, 29 October, 12:00 PM

50 points total.

Submit a single knitr (either `.rnw` or `.rmd`) file, along with a valid PDF output file. Inside the file, clearly indicate which parts of your responses go with which problems (you may use the original homework document as a template). Raw R code/output or word processor files are not acceptable.

Question 1

20 points

Code a function that does golden section search, and use this function to find all of the local maxima on the following function:

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \\ |x| \log\left(\frac{|x|}{2}\right) e^{-|x|} & \text{otherwise} \end{cases}$$

on the interval [−10, 10].

To get an idea of what the function looks like, it might be helpful to plot it.

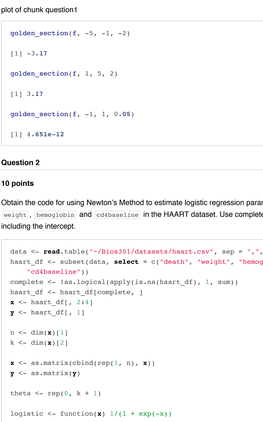
```
library(gqplot2)

Loading required package: methods

source("~/Bios301/exercises/gsection.r")

f <- function(x) ifelse(x != 0, abs(x) * log(abs(x)/2) * exp(-abs(x)),
    0)

x <- seq(-10, 10, 0.1)
qplot(x, f(x), geom = "line")
```



plot of chunk question1

```
golden_section(f, -5, -1, -2)

[1] -3.17

golden_section(f, 1, 5, 2)

[1] 3.17

golden_section(f, -1, 1, 0.05)

[1] 4.651e-12
```

Question 2

10 points

Obtain the code for using Newton's Method to estimate logistic regression parameters (`logtistic.r`) and modify it to predict `death` from `weight`, `hemoglobin` and `cd4baseline` in the HAART dataset. Use complete cases only. Report the estimates for each parameter, including the intercept.

```
data <- read.table("~/Bios301/datasets/haart.csv", sep = ",", head = T)
haart_df <- subset(data, select = c("death", "weight", "hemoglobin",
    "cd4baseline"))
complete <- !as.logical(apply(is.na(haart_df), 1, sum))
haart_df <- haart_df[complete, ]
x <- haart_df[, 2:4]
y <- haart_df[, 1]

n <- dim(x)[1]
k <- dim(x)[2]

x <- as.matrix(cbind(rep(1, n), x))
y <- as.matrix(y)

theta <- rep(0, k + 1)

logistic <- function(x) 1/(1 + exp(-x))

MAX_ITER <- 7
J <- rep(0, MAX_ITER)

for (i in 1:MAX_ITER) {

    # Calculate linear predictor
    z <- x %*% theta
    # Apply logit function
    h <- logistic(z)

    # Calculate gradient
    grad <- t((1/n) * x) %*% as.matrix(h - y)
    # Calculate Hessian
    H <- t((1/n) * x) %*% diag(array(h)) %*% diag(array(1 - h)) %*% x

    # Calculate log likelihood
    J[i] <- (1/n) %*% sum(-y * log(h) - (1 - y) * log(1 - h))

    # Newton's method
    theta <- theta - solve(H) %*% grad
}

print(theta)

                [,1]
rep(1, n)    3.576412
weight      -0.046211
hemoglobin  -0.350643
cd4baseline  0.002093


summary(glm(death ~ weight + hemoglobin + cd4baseline, data = haart_df,
    family = binomial(logit)))


Call:
glm(formula = death ~ weight + hemoglobin + cd4baseline, family = binomial(logit),
    data = haart_df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.237  -0.486  -0.329  -0.203   2.885

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.57641    1.22687    2.92  0.00356 **
weight      -0.04621    0.02256   -2.05  0.04049 *
hemoglobin  -0.35064    0.10506   -3.34  0.00085 ***
cd4baseline  0.00209    0.00181    1.15  0.24814
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 209.72  on 327  degrees of freedom
Residual deviance: 180.99  on 324  degrees of freedom
AIC: 189

Number of Fisher Scoring iterations: 6
```

Question 3

20 points

Consider the following very simple genetic model (*very* simple – don't worry if you're not a geneticist!). A population consists of equal numbers of two sexes: male and female. At each generation men and women are paired at random, and each pair produces exactly two offspring, one male and one female. We are interested in the distribution of height from one generation to the next. Suppose that the height of both children is just the average of the height of their parents, how will the distribution of height change across generations?

Represent the heights of the current generation as a dataframe with two variables, m and f, for the two sexes. The command *r*norm(100, 160, 20) will generate a vector of length 100, according to the normal distribution with mean 160 and standard deviation 20 (see Section 16.5.1). We use it to randomly generate the population at generation 1:

```
pop <- data.frame(m = rnorm(100, 160, 20), f = rnorm(100, 160, 20))
```

The command `sample(x, size = length(x))` will return a random sample of size `size` taken from the vector `x`. The following function takes the data frame `pop` and randomly permutes the ordering of the men. Men and women are then paired according to rows, and heights for the next generation are calculated by taking the mean of each row. The function returns a data frame with the same structure, giving the heights of the next generation.

```
next_gen <- function(pop) {
    pop$m <- sample(pop$m)
    pop$m <- apply(pop, 1, mean)
    pop$f <- pop$m
    return(pop)
}
```

Use the function `next_gen` to generate nine generations, then use the function `histogram` from the `lattice` to plot the distribution of male heights in each generation. The phenomenon you see is called regression to the mean.

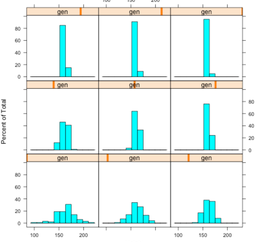
Hint: construct a data frame with variables height and generation, where each row represents a single man.

```
library(lattice)
pop <- data.frame(m = rnorm(100, 160, 20), f = rnorm(100, 160, 20))

next_gen <- function(pop) {
    pop$m <- sample(pop$m)
    pop$m <- apply(pop, 1, mean)
    pop$f <- pop$m
    return(pop)
}

generations <- cbind(pop, gen = 1)
for (i in 2:9) {
    pop <- next_gen(pop)
    generations <- rbind(generations, cbind(pop, gen = i))
}

histogram(~m | gen, generations, xlab = "Height")
```



plot of chunk question3