Univeristy of Cape Town – Department of Statistical Sciences
Data Science Masters Statistical Computing Exam
Examiners: Res Altwegg & Neil Watson
Internal moderator: Dr. Sebnem Er
External moderator: Prof Sugnet Lubbe
Total Marks: 50
Total time: 4 hours

Wednesday 23 June 1 pm - Wednesday 23 June 5 pm

Answer the questions for Part A in the “Part A.Rmd” file and the questions for Part B in the “Part B.Rmd” file. Please include the options ‘eval = T’ in all of your chunk options.

PART A

Question 1 [15]

The data you are going to use for this question are in the file called “Roberts_birddata1.xls”. The Roberts data base lists all bird species that occur in southern Africa but here, we are giving you a modified subset of the data; each row in the file is a different species. The columns contain information on each species.

Answer the following questions by filling in the code in the R Markdown template. The code should automate all data manipulation steps, including the inline code expressions. When I test your code on a different subset of the data, it should still produce the correct answers for that data set.

a. Read the rows you need (header row and data) into R, store the data in a data frame and use R commands to determine how many rows and columns the data frame contains. Remember to set the header correctly. The data set appears to have several rows that act as header; use only the first row as header. [2]
b. One of the variables in the data set is `Family`, i.e. which taxonomic family a particular species belongs to. Which family has the most species belonging to it? How many species does this family have? How many families only have a single species? [2]
c. Which family has the highest number of species that are endangered, i.e. score code = 2 on BirdLife International’s Red Data Status? [1]
d. Write a function that takes the name of a bird family as an argument and returns the number of species in that family. Then use this function to make a new data frame that contains only those rows of the original data frame that belong to families with more than 10 species. [3]

In this reduced data set, how many species are we left with?

e. Now, using this reduced data set, we want to examine the relationship between body mass and tarsus length (which is a measure of the size of the bird). The problem is that this information is scattered among several columns in the data set. For some species, we have data in the columns `Adult body mass` and `Adult tarsus length`. For other species, these measures are given separately for the two sexes. [3]

Create a new variable called `bodymass` that consists of the data in `Adult body mass` if this value is not missing, and of the data in column `Female body mass` otherwise.

Then create a new variable called `tarsuslength` that consists of the data in `Adult tarsus length` if this value is not missing, and of the data in column `Female tarsus length` otherwise.

Then exclude species for which either `bodymass` or `tarsuslength` is missing.

How many species are left now?

f. Now produce a scatterplot with the natural logarithm of bodymass on the x-axis and the natural logarithm of tarsuslength on the y-axis. Make sure the axes labels are informative. Also plot the y-axis annotation horizontally. Colour the points by bird family.

Write the figure to pdf and call it "fig1.pdf", then insert it below. [2]

g. Now fit a linear regression model to these data. Log tarsuslength should be the dependent (response) variable and log bodymass the independent (explanatory) variable. Re-plot the figure you made earlier. Then add the fitted regression line. [2]

What is the slope of the fitted regression line, rounded to two decimal places?

PART B

Question 2 [12]

Consider the following linear programming problem:

Maximize:

6x1 + 2x2 + 4x3 + 3x4

Subject to:

2x1 + x2 + 3x3 + 2x4 $\leq$ 4000
4x1 + 2x2 + x3 + 2x4 $\leq$ 6000
6x1 + 2x2 + x3 + 2x4 $\leq$ 10000
x1 $\leq$ 1000
x2 $\leq$ 2000
x3 $\leq$ 500
x4 $\leq$ 1000
x1, x2, x3, x4 $\geq$ 0

The optimal solution to the problem is:

Optimal value = 9200
x1 = 1000; x2 = 800; x3 = 400; x4 = 0

a. Use `constrOptim` to solve the linear program. Change the initial starting values - consider at least 5 sufficiently different starting value combinations within the bounds of each variable. Does the optimiser always reach the optimal solution? [4]

b. For each of the variables, create a sequence of length `nsims` of values within a range of 100 from the optimal value of that variable (but not including the optimal value). For example, generate a sequence of `nsims` values between 999 and 900 for `x1`,. . . , a sequence of `nsims` values between 1 and 100 for `x4`. Using these sequences as your initial values to the optimiser, write a function that solves the linear program for each set of these `nsims` initial values, and stores both the optimal value in an object `value` and the number of function counts in an object `counts`. [3]

c. Compute 5-number summaries of both `value` and `counts`. [1]

d. Plot `value` vs `counts`. Title your plot and label the axes appropriately. [1]

e. Create a vector `opt_val` that is a sequence of length `nsims` of optimal values between 9100 and 9199. For each value in this sequence, work out the proportion of times the algorithm converges to a greater optimal value. Save this in an object `opt_prop`. Plot `opt_prop` vs `opt_val`. Title your plot and label the axes appropriately. [2]

f. How effective is the `constrOptim` algorithm at solving this problem? Motivate your answer. [1]

Question 3 [13]

Assume that you have a response variable, y that is related to one explanatory variable, x. Further assume that $y_i = 5 + 10x_i + e_i$ where $e_i \sim N(0, \sigma = 5)$. Assume further that we have 100 observations such that $n = 100$. Let $x = (1 : 100)/100$ i.e. $x = 1/100, 2/100, 3/100, ..., 99/100, 100/100$.

a. Simulate the response variable and name the object `y`. [2]

b. Use the *lm* function to estimate the linear model using the 'x' and simulated 'y' data. Save the object as `mod1`. Create and save the parameter estimates in a vector `lm_params` [1]

The above model assumes that the linear model is of the form $\hat{y}_i = \hat{a} + \hat{b}x_i$. Two parameters, `a` and `b` are estimated by minimizing,

$$A = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \hat{a} - \hat{b}x_i\right)^2$$

Gradient descent is a numerical optimization routine that is often used to obtain the parameters that minimize a multi-parameter function. The algorithm is an iterative one. Using the above example, define

$$\beta_i = (a^{(i)}, b^{(i)})^T,$$

as the value of a 2 × 1 vector that contains the elements 'a' and 'b' at iteration i. The computer now updates the parameter vector using β_i as well as the gradient of the function `A` at β_i (i.e. $F(\beta_i)$) using the following update rule:

$$\beta_{i+1} = \beta_i - \alpha F(\beta_i),$$

where α is some constant.

c. Write an R function to calculate the gradients of `A`, i.e. you require the gradients $\frac{dA}{da}$ and $\frac{dA}{db}$. Return the answers as a two-dimensional vector. Name the function `gradA`. [3]

d. Now, write your own gradient descent algorithm. Name the function `grad_desc`. Set $\alpha = 0.01$. Your functions should contain the following arguments, `a`, `b`, `x`, `y`, `iters`, where `iters` is the total number of iterations used to obtain the parameter vector. Notice that the first two arguments could be parsed as a vector (say 'params'). Include a stopping criterion that terminates the algorithm once you are within `tol` $= 0.001$ of both of the parameters. [5]

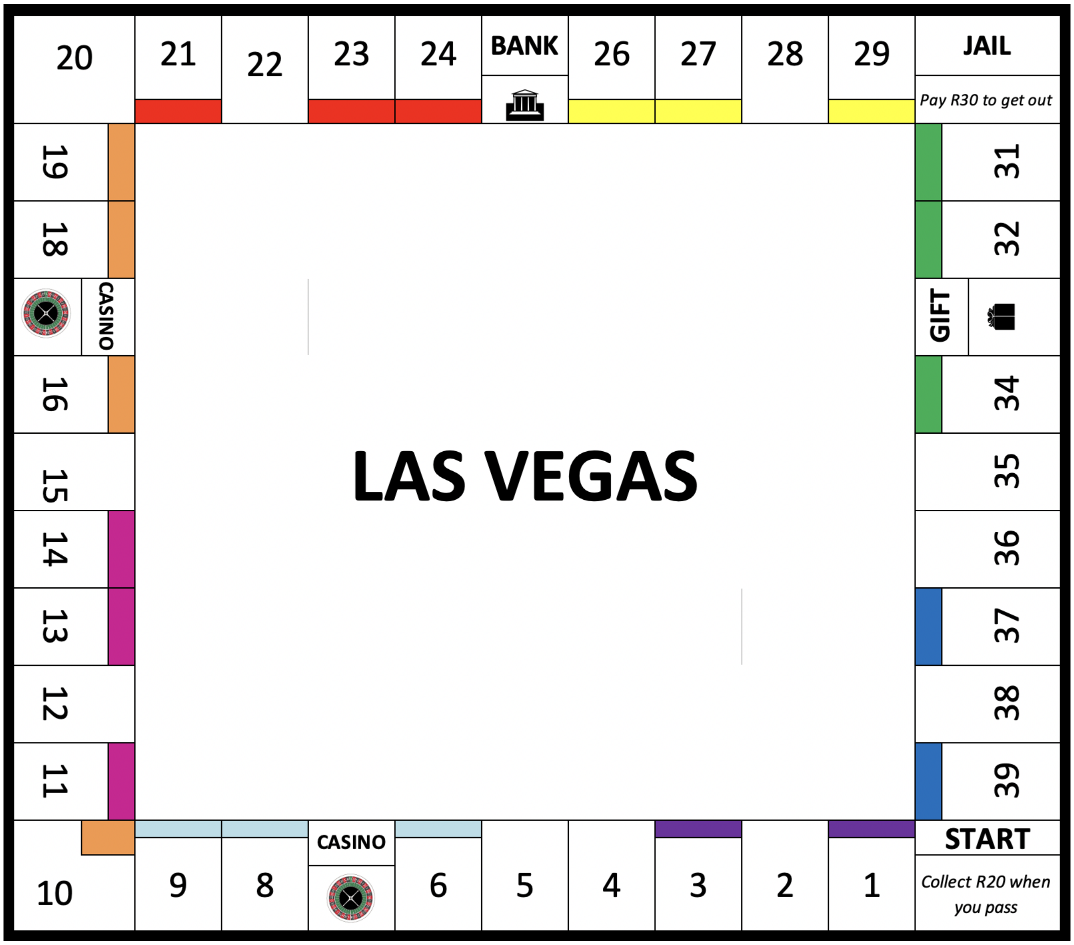
e. Print the value of the parameters after 50 iterations. [1]

f. If `tol` $= 0.0001$, how many iterations are required to reach the solution? [1]

Question 4 [10]

The figure below is of the game ‘Las Vegas’. The game is played by repeatedly rolling one six-sided die. The game ends when you run out of money. The objective of the game is to see how long (how many rounds = throws of the dice) you can last before your money runs out. Here are the rules of the game:

- A player begins the game at START with R100
- Each time a player passes over START, including landing on START, they are awarded an additional R20
- Each time a player lands on JAIL, they have to pay R40 to get out
- Each time a player lands on BANK, there is an equal chance that they have to pay the Bank R15 or R50
- Each time a player land on GIFT, they are gifted R10
- Each time a player lands on a CASINO, the following applies:
  - A 20% chance that you don’t play anything and hence no change to your current amount of money
  - A 40% chance that you play and lose R30
  - A 20% chance that you play and win R15
  - A 10% chance that you play and win R80
  - A 10% chance that you play and lose R100



Write an R program in order to simulate the above game 10000 times, and provide a histogram of the empirical distribution of how many rounds it will take before the player has no more money. Title the histogram appropriately. Clearly indicate the mean of the distribution on the histogram with a red line and print the value of the mean next to the line.