Intro to R Programming: Class Test 1

Class Test Rules & Conditions

- The class test is taken under exam conditions.
- During the test you are not allowed to talk to or otherwise communicate with other students (email, instant messaging, etc.), or access the internet/course material. The only material you can use is the printed copy of R reference manual provided as well as the R help within RStudio.
- **DO NOT** open other web pages - you are only allowed two windows open (R studio and the Moodle Class Test 1 page)
- *Please note we use technological means to check compliance with the above!*
- Students who breech the rules above will be reported to the Clerk of Senate.

Starting the test ...

- You will be already logged in to the computer.
- Please log into Moodle and navigate to the Introduction to R Moodle page, Section **2019 Class Test 1.**
- Click on the link **Class Test 1 - Honours/DD80** to begin the test.
- You should use an R studio (or R) window to trial and test your answers. Once you have written and tested the code for a question, **copy the code into the corresponding answer field for the question.**
- In the answer field for each question, only include the code with answers for that specific question.

During the test ...

- You can move back and forward through questions during the class test period.
- **If you have any issues logging in to Moodle, or cannot locate the link to start the class test, please let one of the tutors know immediately.**
- **Once open, do not close the moodle browser window.
- The only external packages you are allowed to use (if you wish, they are not required) are `dplyr` and `ggplot2`.
- If you have experience any issues with the technology throughout the class test, please speak to a tutor immediately.
- For all parts in the test give the R code which can be used to answer the questions. All questions should be answered programatically and should not be hard coded.
- You should only include the code to answer the question, you do not need to include comments or output from the console window.

Important

All the graphical questions display the plots you should produce. You plots should look similar to the ones provided. Specifically, you should replicate: points/triangles/bar colors, axis labels, and titles.

Part 1

The data file `Boston.txt` contains information for housing values in suburbs of Boston. It has 14 columns, but we are only interested in the following:

Variable	Class	Description
medv	numeric	Median value of owner-occupied homes in $1000s.
lstat	numeric	Lower status of the population (percent).

Table 1: Variables for the `Boston` dataset.

1. [**2 marks**] Use R to read in the file `Boston.txt` correctly and save it as a data frame called `Boston`.
2. [**2 marks**] Update the `Boston` data frame by selecting only the columns corresponding to `lstat` and `medv`. The `Boston` data frame should now contain only two columns, `lstat` and `medv`. Sort the data frame in decreasing order according to the values of `lstat`.

We will model the relationship between `lstat` and `medv` using polynomial regression. For a covariates vector $\mathbf{x} = (x_1, ..., x_n)$ the design matrix for the polynomial regression of degree p takes the form

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ 1 & x_2 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^p \end{bmatrix}$$

3. [**2 marks**] Define a matrix `X` which is the design matrix for fitting a polynomial regression of order 3. For this matrix, the column `lstat` in the data frame `Boston` is the covariate **x**.
4. [**3 marks**] Define a vector `y.hat` which contains the fitted values for the polynomial regression of order 3 computed using the design matrix `X` from question 3 and `medv` as the vector of responses **y**. The fitted values can be computed using

$$\hat{y} = X(X^TX)^{-1}X^Ty. \tag{1}$$

5. [**2 marks**] Plot the data with `lstat` in the x-axis and `medv` in the y-axis. Use `pch=16`. Add the fitted polynomial regression line to the plot (use `lwd=3`). Your plot should look like the one in Figure 1.

Part 2

The data file `squirrels.csv` contains data from the New York city squirrel census in Central Park[1]. It contains 37 columns, but we are only interested in the ones listed in Table 2.

1. [**2 marks**] Use R to read in the file `squirrels.csv` correctly and save it as a data frame called `squirrels`.
2. [**2 marks**] Update the `squirrels` data frame by selecting only the variables we are interested in (defined in Table 2). The updated data frame should be called `squirrels`.

[1] Data originally from thesquirrelcensus.com. Raw data can be found in NYC OpenData.

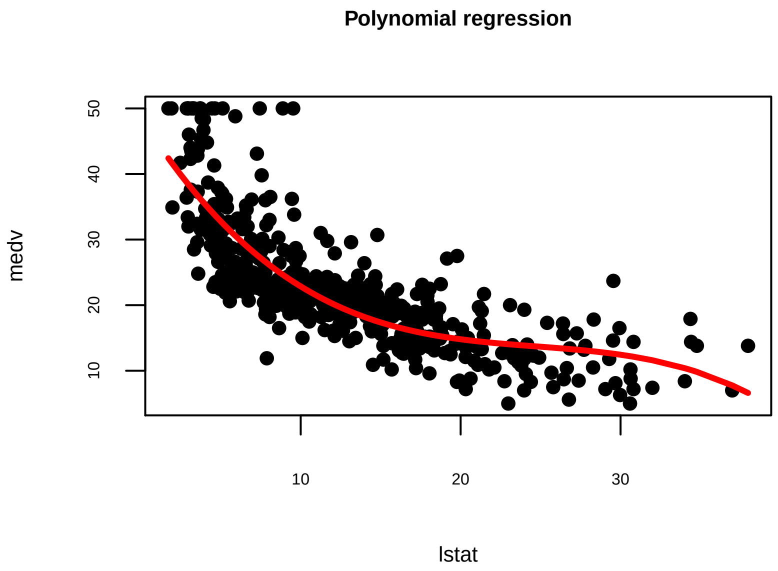

Figure 1: Polynomial regression for the Boston dataset.

Variable	Class	Description
long	numeric	Longitude
lat	numeric	Latitude
shift	character	Whether or not the sighting session occurred in the morning or late afternoon. Value is either "AM" or "PM."
date	integer	Concatenation of the sighting session day and month.
age	character	Value is either "Adult" or "Juvenile."
primary_fur_color	character	Value is either "Gray," "Cinnamon" or "Black."
location	character	Location of where the squirrel was when first sighted. Value is either "Ground Plane" or "Above Ground."
running	logical	Squirrel was seen running.
chasing	logical	Squirrel was seen chasing.
climbing	logical	Squirrel was seen climbing.
eating	logical	Squirrel was seen eating.
foraging	logical	Squirrel was seen foraging.
approaches	logical	Squirrel was seen approaching human, seeking food.
indifferent	logical	Squirrel was indifferent to human presence.
runs_from	logical	Squirrel was seen running from humans, seeing them as a threat.

Table 2: Variables for the `squirrels` dataset.

3. [**2 marks**] Update the `squirrels` data frame by removing all rows where the values of `primary_fur_color` are missing. The updated data frame should be called `squirrels`.

4. For the following questions, use `na.rm` if needed.

   i) ~~[**1 mark**] How many squirrels are seeing moaning during the survey?~~

   ii) [**1 mark**] How many black squirrels are observed climbing?

   iii) [**1 mark**] How many cinnamon squirrels are observed chasing?

   iv) [**2 marks**] How many juvenile gray squirrels are observed above ground?

5. [**2 marks**] Create a contingency table whose rows are the primary fur colours, and whose columns are the shift values.

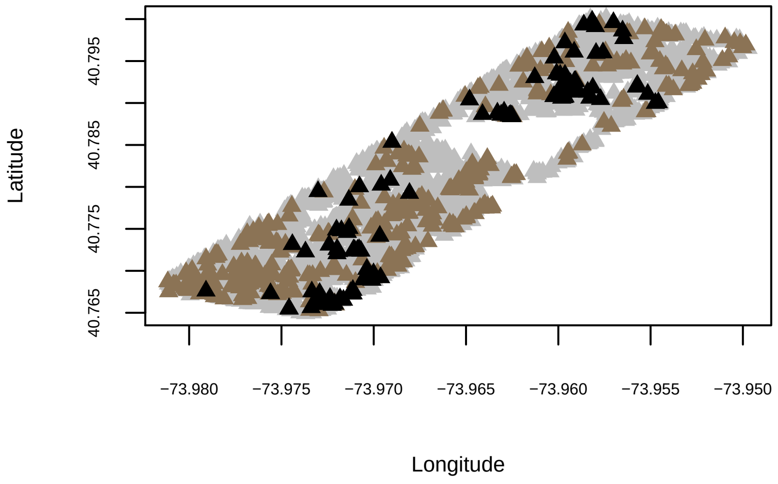

Figure 2: Location of squirrels and their primary fur color.

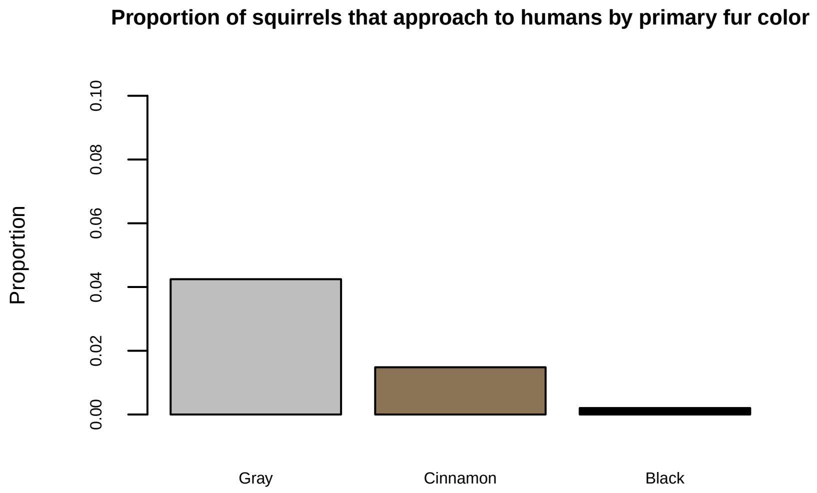

Figure 3: Proportion of squirrels that approach to humans by primary fur color.

6. [**3 marks**] In `R` you can find very sophisticated tools to create maps, but simpler maps can also be created using the function `plot`. Replicate the map in Figure 2 that displays the location of squirrels and their primary fur color. Use `pch=17` to create the triangle symbols, and the following colours: `black` for black squirrels, `gray` for gray squirrels, and `burlywood4` for cinnamon squirrels.

7. [**3 marks**] Create a barplot that shows the proportion of squirrels that approach to humans by primary fur color. For each bar, use the same colors defined in question 6. See Figure 3 for reference. If you want to check that the names below each bar are displayed correctly, just zoom in the plot.

8. [**3 marks**] To study the difference in behaviour between juvenile and adult squirrels, do the following:
   - Create a new variable called 'behaviour' which takes the value `Friendly` if the squirrel was seen approaching human, `Indifferent` if the squirrel was indifferent to human presence, and `Scared` if the squirrel was seen running from humans.
   - Use the function `table()` to create a contingency table of the counts at each combination between behaviour and age.
   - Use the function `prop.table(your.table, margin = 2)` to express your table as proportions with respect to the age.
   - Replicate the barplot in Figure 4.

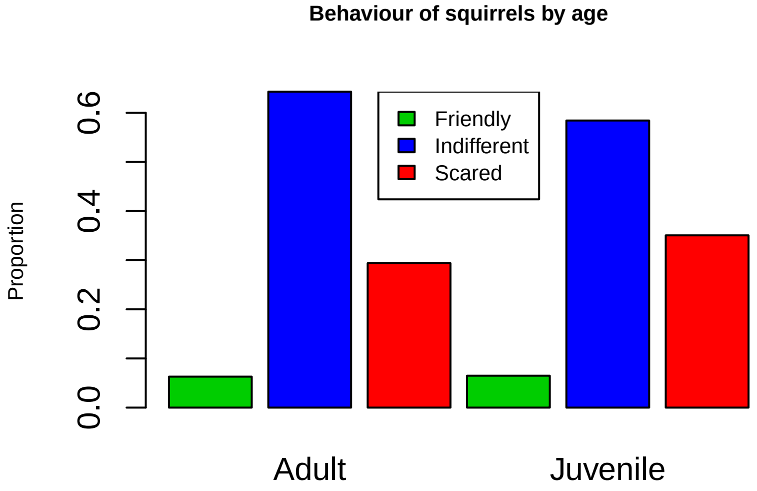

Figure 4: Behaviour of squirrels by age.

9. [**3 marks**] Freddy is a gray squirrel that enjoys learning statistics in his free time. Yesterday, while seated on a bench, Freddy counted the number of squirrels that arrived at the tallest tree in Central Park between 4 pm and 5 pm. During that time, he saw 15 squirrels arriving at the tree. Freddy knows that if the squirrel's arrivals are independent, then the distribution of the inter-arrival times is exponential (inter-arrival times are the time between arrivals). Freddy would like to compute the probability that the inter-arrival times are between 0.05 and 0.1 hours. To help Freddy, simulate the inter-arrival times of 100 squirrels and use your simulation to estimate the desired probability. Please, enter and run the line of code below **before** carrying out your simulation:

```
set.seed(123)
```

Hint: you can use the function `rexp(n, rate)` *to generate* `n` *draws from an exponential distribution with rate* `rate`.

10. Use your simulation from question 9 to answer the following:

    i) [**2 marks**] What is the average inter-arrival time (in minutes)?

    ii) [**2 marks**] How many squirrels took more than 30 minutes to arrive to the tree?