# STAT330/430 - Statistical Learning

**Assignment 1**

**Due date - 21st March**

**Note: You must submit two files for this assignment:**

- a pdf document of your solutions, and
- a complete and concisely annotated *R Script* file of all **R** analysis that was undertaken to produce your results.

In addition, to receive marks for Question 1 you are required to engage in the Topic 1 moodle discussion forum.

Please refer to the Assignment assessment criteria document for additional guidance.

## Question 1 [5 marks]

In each assignment you will be able to earn up to 5 marks based on your engagement on the moodle Discussion forums from the topics associated with the given assignment. See the Assignment assessment criteria document for more details.

## Question 2 [15 marks]

The Maximum Life Expectancy (MLE) of different animals is thought to be correlated to a variety of biological and physiological factors including those contained in the *Animals* data-set.

| Variable | Description |
|---|---|
| species | Species of animal |
| non_dreaming | Slow wave sleep (hrs/day) |
| dreaming | Paradoxical sleep (hrs/day) |
| MLE | Maximum life span (years) |
| gestation | gestation time (days) |
| predation | Relative risk of being preyed upon (Low, Med, and High) |
| exposure | Protection during sleep (Low, High) |
| log_body | Log body weight (in kg) |
| log_brain | Log brain weight (in g) |

(a) Use a combination of figures and summary statistics to explore the *Animals* data-set. Make sure you declare any factors and use only complete records. You should:

- visualise and quantify relationships between numerical variables,
- examine the relationship between the response variable and any categorical variables, and
- provide a general summary of your exploratory analysis.

(b) Fit a multiple regression model to predict the MLE of a given animal. Use all variables in the data-set with the exception of the ID variable, species. Provide a summary of your model, interpretation of output, and commentary on model assumptions.

(c) Is multicollinearity a problem for the model fitted in (b)? If so, take appropriate steps to address this issue, and present an updated model. Again, include a summary of your model, interpretation of output, and commentary on model assumptions.

(d) Use the model in (c) to provide insight into the following questions:

i. How does relative predation risk affect MLE?
ii. What effect does length of gestation have on MLE?

## Question 3 [25 marks]

A study on wrens was conducted both prior to and after a severe drought event. Physical characteristics (listed below) of all birds were measured along with a report of whether each bird survived the drought or not.

| Variable | Description |
|---|---|
| survive | Whether the bird survived the drought ("1") or did not ("0") |
| length | Body length (mm) |
| alar | Maximum wingspan (mm) |
| weight | Body weight (g) |
| lbh | Lenght of beak and head (mm) |
| lhum | Length of humerous (mm) |
| lfem | Length of femur (mm) |
| ltibio | Length of tibiotarsus (mm) |
| wskull | Width of skull (mm) |
| lkeel | Length of keel of sternum |

(a) Use a combination of figures and summary statistics to explore the *Wren* data-set. Make sure you declare any factors and use only complete records. You should:
- visualise and quantify relationships between numerical variables,
- examine the relationship between the response variable and the predictor variables, and
- provide a general summary of your exploratory analysis.

(b) Split the *Wren* data-set into 70:30 training:testing data-sets. *Note:* There are a wide variety of methods to undertake this task. You are welcome to use any method you would like as long as it is clearly documented in your R Script and you use the last three numbers of your Student ID number as the set.seed value so that the results can be replicated.

(c) Use all variables in the *Wren* data-set to fit a logistic regression model to predict Survival. Summarise the steps of your analysis including any testing of assumptions, clearly report and provide commentary on your results, and estimate the training and testing prediction accuracy of your model.

(d) Use all variables in the *Wren* data-set to fit a KNN model to predict Survival. Use a range of values for $k$ in order to present an optimised value of $k$. Summarise the steps of your analysis including any testing of assumptions, clearly report and provide commentary on your results, and estimate the training and testing prediction accuracy of your model.

(e) Compare and contrast the two different statistical analyses you conducted in (c) and (d). Which approach do you think is the best choice for this research question?