Candidate Number:

ISSU2028 Data Science and Big Data Analytics

UCL International Summer School for Undergraduates 2018

**Assessment III Computer Practical under examination conditions (50%)**

**Timing Friday 10th August 2018 10am – 12:30pm**

**Examination conditions:**

* **The session is open book, so you can consult your notes, textbooks and programming websites as you work.**
* **You may work either on the UCL computers, or on your own laptop.**
* **All work submitted must be your own. All forms of communication and messaging with other students is strictly prohibited, and violations will be dealt with in accordance with UCL policy.**
* **At the end of the assessment you will have a short time to collate the files you have produced and upload them for marking.**

**Notes:**

All code used to complete the tasks must be submitted in R script files or R notebook files in Rmd format. Partial marks will be awarded for code sections that have been completed but are non-functional.

You should work in one file per section, so at the end of the exam upload either:

sectionA.R, sectionB.R and sectionC.R

or sectionA.Rmd, sectionB.Rmd and sectionC.Rmd

Use comments / notes in your files to indicate which part of the question you are attempting.

When answering questions you should include units where relevant and quote values to an appropriate number of significant figures. Plots should be displayed with appropriate axes labels.

Marks for questions are indicative, and a grading curve may be applied to generate a final grade.

**Section A**

**An analysis of Crab Body Dimensions.**

Male fiddler crabs have **one large claw** (used to attract females) and **one small claw**. These are classified as their **major** and **minor** claws respectively. Major claws can be either on the right or left.

The data in file crab.csv contains the following measurements taken from male crabs that have been sampled from a large crab colony.

**crab.id** an ID number for the measurement

**body.size** the size of the crabs body in cm

**major.claw** the size of the crabs major claw in cm

**minor.claw** the size of the crabs minor claw in cm

**claw.ratio** the ratio of the claw sizes (major/minor)

**major.side** which side the major claw is on (left or right)

We want to explore the data, and in particular be able to determine an expected body size given the values of the other measurements.

Q1i) Is this an example of constructing a model for inference or prediction?

(1 mark)

*inference / prediction* (delete as appropriate)

ii) Download the file crab.csv and write R code to load the data from the file into your environment.

(1 mark)

iii) How many rows are in the dataset?

(1 mark)

iv) What is the mean body size of the crabs sampled?

(1 mark)

v) What is the median size of the major claws recorded?

(1 mark)

vi) What percentage of the crabs in the dataset had their major claw on the right hand side?

(1 mark)

6 marks

Q2i) Write code that deletes the ID column from the loaded data.

(1 mark)

ii) Draw a scatter plot showing body size against major claw size.

Ensure that the axes are labelled as **body size (cm)** and **major claw size (cm)** and the plot has the title **Body measurements of crabs**.

(4 mark)

iii) Construct a linear model to fit **body size** using **major claw size**.

(1 mark)

iv) Discuss how we can interpret the value of the fitted coefficient for **major claw size**.

(1 marks)

v) Use the result of the fit to predict the mean body size of crabs that have major claw size equal to 10cm.

(1 mark)

/8 marks

Total 14

**Crabs are able to regrow a lost claws. It is proposed that this may lead to outliers as regrowing claws will not be at their final size.**

3i) On the scatter plot you have created there is one point (with claw size around 7cm) that appears to be an outlier to the fit .

Identify the row index for this outlier and use this to find its exact body size and claw size.

(2 marks)

Row index: Body size: Claw size:

ii) Is the position of this outlier consistent with the hypothesis that this crab is in the process of regrowing a lost claw? Explain your answer.

(2 mark)

4i) Calculate the coefficient of Pearson’s correlation r between the measurements of major and minor claw sizes.

(1 mark)

ii) Explain why including both these variables in a linear regression model to predict body size could be problematic.

(1 mark)

iii) It is proposed instead to use claw ratio in the model to avoid such problems.

Build a linear model of **body size** fitted against the two predictors **major claw size** and **claw ratio**.

(1 mark)

iv) Examine the information on the fit result. Look at the reported F-statistic.

What does this reported F-statistic relate to?

(1 mark)

v) In this case what does the result tell us? Explain your answer.

(1 mark)

vii) Construct a linear regression model that adds **major side** as a third predictor.

(1 mark)

viii) Is there evidence that including this predictor improves the model? Explain your answer.

(2 marks)

The following section considers the fit of **body size** using **major claw** **size** only.

5i) To estimate the range of R-squared values that might arise from fitting to different samples of the population, we will resample the dataset 100 times in a bootstrap analysis, and fit each bootstrapped dataset to collect a distribution of 100 measurements of the R-squared fit values.

Note. The R-squared value can be extracted directly from the summary of the fitted model

e.g. If model is lm.1 then rsq <- summary(lm.1)$r.squared gets the value we want.

i) Carry out this analysis, e.g. use a **for** loop to: generate a boostrapped dataset; then fit **body size** using **major claw size** and store the resulting value of R-squared, and repeat the loop until you have 100 bootstrapped measurements.

(4 marks)

ii) Make a histogram of the resulting distribution of the R-squared values. Use appropriate axes labels.

(2 marks)

iii) It is proposed that a fit of **log(body size)** against **log(major claw size)** will better describe the data in comparison to using the non-transformed measurements.

Create new columns in the data frame that store the log measurements of body and major claw size.

(2 marks)

iv) Perform the linear regression using the transformed variables and examine the fit results.

(1 mark)

v) Evaluate whether using log transformation makes the fit of **body size** using **major claw** **size** significantly better or worse.

(2 marks)

6 i) The file crabs.csv can be produced from the file raw.csv that stores left and right claw sizes in order of body size.

crab.id body.size right.claw left.claw

|  |  |  |  |
| --- | --- | --- | --- |
| 12 | 4.85 | 1.9 | 5.05 |
| 30 | 5.8 | 2.06 | 5.3 |
| 39 | 6 | 2.1 | 5.29 |
| 34 | 6.15 | 2.25 | 6.41 |
| 42 | 6.3 | 7.74 | 2.54 |

Write R code to demonstrate how this file can be loaded into R and converted into a data frame that contains only the data in the same format as it was when loaded from the crabs.csv file. i.e. your code should use this data file to create a data frame with only the following columns:

**body.size** stores body size

**major.claw** stores biggest claw size from the left and right claws

**minor.claw** stores smaller claw size from the left and right claws

**major.side** stores 'left' or 'right' as a factor

and it should be ordered by increasing crab id

(5 marks)

**Section B**

**This consists of an analysis of mortality (death rates) based on demographic information from metropolitan areas in the USA.**

The data file “mortality.csv” stores the following columns

Name Description

A1 average annual rainfall in inches

A2 average January temperature in degrees Fahrenheit

A3 average July temperature in degrees Fahrenheit

A4 percent of population 65 years old or older

A5 average household size, 1960

A6 average level of schooling for persons over 22

A7 percentage of households with full kitchens

A8 population per square mile in urbanized areas

A9 percent non-white population

A10 percent office workers

A11 poor families (annual income under $3000)

A12 relative pollution level of hydrocarbons

A13 relative pollution level of Nitrogen Oxides

A14 relative pollution level of Sulphur Dioxide

A15 percent relative humidity

B total annual age-adjusted mortality rate per 100,000

This was gathered as part of an investigation aiming to determine how air pollution and other factors (stored in the **A** columns) are related to death rates (stored in column **B**).

1i) Download the file mortality.csv and write R code to load the data into your environment.

(1 mark)

ii) Examine the correlation between the predictors and the response.

(2 marks)

1. List the predictor columns that have strong positive correlation ( above +0.5) to the mortality rate.
2. List the predictor columns that have strong negative correlation ( below -0.5) to the mortality rate.

/3 mark

Total 42

iii) Fit a multivariate linear model to predict the death rate **B** using all possible predictors **A1 to A15**.

(1 mark)

iv) Find the predicted death rates in each area according to the fitted model and store the values into a new column

(1 mark)

v) Make a plot showing **Actual Death Rate** on the y-axis and **Predicted Death Rate** on the x-axis. Add a line showing how a perfect prediction would perform.

(3 marks)

vi) Examine the diagnostic plots for the fit (call plot on the model). Identify the data row that has been flagged as having a high leverage.

(1 mark)

ix) Find the data associated with this point and suggest why this row produces an outlier when used to predict the death rate.

(1 mark)

7 marks

Total 49

2i) Construct a reduced model in which the fit on death rate **B** is made using columns A1, A2, A3, A4, A5, A6, A8, A9, A12, A13, A14, A15 **(*i.e. all except for A7 A10 and A11*).**

(1 mark)

ii) Discuss and explain the differences in the R-squared and Adjusted R-squared values for the reduced and full models.

(3 marks)

4 marks

Total 53

iii) Calculate the model performance of this reduced model as measured by MSE and RMSE calculated using the fitted **training** data.

(2 marks)

MSE: RMSE:

iv) Perform k-fold validation with 10 folds to cross validate the model performance.

(3 marks)

vi) Use the function regsubsets from the leaps library to test through all possible combinations and so find the subset of predictors found to optimise model performance.

(3 marks)

vii) Find the best performing of the selected subset models as measured by adjusted R-squared. Which predictors columns are **not included** in this model?

(1 mark)

viii) Make a plot to show how adjusted R-squared varies against number of predictors for the best subsets identified. Label your plot axes appropriately.

(3 marks)

Hint. You can find the set of adjusted R-squared values by using the summary command on the output of the regsubsets tool like this: summary(...)$adjr2

**Section C**

**An analysis aiming to classify two types of lily based on characteristics of stem length, petal width and petal length.**

The file "flowers.Rda" contains an R data frame called flowers. This contains 4 columns:

**petal.length** stores petal length in cm

**petal.width** stores petal length in cm

**stem.length** stores petal length in cm

**type**  stores flower type as an integer: 0 for pink lily and 1 for yellow lily

Our aim is to build a model that can classify the flower type using the predictor variables.

1i) Load in the data to your R environment.

(1 mark)

ii) Add a new column **type.factor** to the data frame that stores the flower type as a factor with appropriate level labels.

(2 marks)

iii) Make three figures that show boxplots to illustrate how the distribution of petal width, petal length and stem length differ for the two types of flowers.

(3 marks)

iv) Comment on which variable(s) seem most suited for discriminating between the two flower types.

(1 mark)

7 marks

Total 77

v) Perform a t.test to check if we can reject the null hypothesis that the populations of the yellow and pink lily distributions have the same mean stem length.

Comment on your results.

(2 marks)

vi) Make a scatter plot of petal length vs petal width, using different colours to distinguish the flower types. Label your plot axes appropriately.

(4 marks)

**We will now perform logistic regression to classify the flowers using petal width and petal length as our predictors.**

2i) Divide the data set into two data frames of equal length to create train and test data sets.

Ensure that the split is randomised as the loaded data frame is ordered by flower type.

(2 marks)

ii) Perform a Logistic regression on the **training** data to produce a fitted model that can classify the lily type based on petal width and length.

(3 marks)

iii) Make a new column in the **test** data frame that stores the flower type predicted from the fit.

(2 marks)

iv) Create the confusion matrix, that shows the result of the classification, in terms of the predicted and actual lily types on the **test** data.

(2 marks)

v) Calculate the misclassification rate on the test data and on the training data.

(3 marks)

**We will now use the method of KNN to classify the flowers.**

3i) Create the following data frames in preparation for performing KNN analysis:

1. marks)

- a dataframe containing all training rows, and columns for petal width and petal length only

- a dataframe containing all training rows, and factor column of petal type only

- a dataframe containing all test rows, columns for petal width and petal length only

ii) Carry out KNN analysis using K=10 and store the resulting predictions.

(3 marks)

iii) Calculate the misclassification rate of the resulting predictions.

(2 marks)