Sample Exam Question

**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)

Extension for week 2:

Note: Doesn’t really make sense in real life because you don’t get claws without crabs, but let’s do it anyway:

Question 4

i) Divide the crab data into 90% 10% train and test data and build a your regression model from above (using **major claw size** and **claw ratio)** using 10 fold cross validation with 3 repetitions.

ii) Do R² and RMSE of this model differ from the one above? Why?

iii) Now apply your model to the 10% test data. How have R² and RMSE changed?

iv) Explain why you may want to split your data into training and test data.

v) Explain why you use cross validation for validating your model in the training phase

vi) 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.

vii) Perform the linear regression using the transformed variables (on the training data, using CV)

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

iv) now use this model to predict body size on the test data

5 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 major.side

|  |  |  |  |
| --- | --- | --- | --- |
| 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)

Actual Exam question from last year. I want you to divide data into test and training 90:10 from the start!

**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**. Use 10-fold cross validation with 3 reps for that.

(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)

x) added by me: predict death rate on the test data. Compare R² from applying model on training and testing data

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*). As usual with using CV**

(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)

iii) Get MSE and RMSE of this model

(2 marks)

MSE: RMSE:

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

iv) added by me: Now use forward stepwise selection until an optimal model is reached. Does it differ from the one chosen by regsubsets?

**Section (C)**

*Start a new R file for your answer to this question.*

Load the following file into R: starbucks\_data.csv

Write R code to find:

B1. Total Sales for each day recorded in the dataset.

(1 mark)

B2. The highest number of frappucinos sold in a single day.

(1 mark)

B3. The mean value of sales on Mondays

(1 mark)

B4. Perform a t-test to test the hypothesis that daily sales differ between a Monday and a Friday.

(1 mark)

B5. Interpret the result against the hypothesis based on a significance level of 0.05.

(1 mark)