3. Linear Regression: We will now implement Linear Regression to predict the age of Abalone (a type of snail). The data set is made available as part of the provided zip archive (linregdata). You can read more about the dataset at the UCI repository link. We are interested in predicting the last column of the data that corresponds to the age of the abalone using all the other attributes.

(a) The first column in the data denotes the attribute that encodes-female, infant and male as 0, 1 and 2 respectively. The numbers used to represent these values are symbols and therefore should not be ordinal. Transform this attribute into a three column binary representation. For example, represent female as (1, 0, 0), infant as (0, 1, 0) and male as (0, 0, 1). [0.5 marks]

(b) Before performing linear regression, we must first standardize the independent variables, which includes everything except the last attribute (target attribute). Standardizing means subtracting each attribute by its mean and dividing by its standard deviation. Standardization will transform the attributes to possess zero mean and unit standard deviation. You can use this fact to verify the correctness of your code. [0.5 marks]

(c) Implement the following functions: (i) mylinridgereg(X, Y, λ) that calculates the linear least squares solution with the ridge regression penalty parameter (λ) and returns the regression weights; (ii) mylinridgeregeval(X, weights) that returns a prediction of the target variable given the input variables and regression weights; and (iii) meansquarederr(T, Tdash) that computes the mean squared error between the predicted and actual target values. [2 + 1 + 1 = 4 marks]

(d) Partition the dataset into 80% training and 20% testing (Let's call this the partition fraction, in this case 0.2). Now, use your mylinridgereg with different λ values to fit the penalized linear model to the training data and predict the target variable for both training and testing data. [1 mark]

(e) Identify the λ with the best performance and examine the weights of the ridge regression model. Which are the most significant attributes? Try removing two or three of the least significant attributes and observe how the mean squared errors change. [1 mark]

(f) We now would like to ask the question: Does the effect of λ on error change for different

partitions of the data into training and test sets? To do this, change the partition fraction (a value between 0 and 1, as defined earlier) with at least 4 other values. Repeat the following steps 25 times for each partition fraction:
• Randomly divide data into training and test sets.
• Standardize the training input variables.

• Standardize the testing input variables using the means and standard deviations from the training set.
• Follow step (d) for each such partition.
For each partition fraction, plot a figure with λ on the x-axis, and MSE on the y-axis. For each figure, include 2 graphs - one for the training MSE and one for the test MSE. (You should then have 5 figures in total, with 2 plots on each figure.) [3 marks]
(g) Do the above figures give you clarity? Also, plot two more figures. In the first graph, plot the minimum average mean squared testing error versus the partition fraction values. In the second graph, plot the λ value that produced the minimum average mean squared testing error versus the partition fraction. [1 mark]
(h) How good is your model? So far, we have been looking at only the mean squared error. We might also be interested in understanding the contribution of each prediction towards the error. Maybe the error is due to a few samples with large errors and others have tiny errors. One way to visualize this information is to a plot of predicted versus actual values. Use the best choice for the training fraction and λ, make two graphs corresponding to the training and testing set. The X and Y axes in these graphs will correspond to the predicted and actual target values respectively. If the model is good, then all the points will be close to a 45-degree line through the plot. [2 marks]
Include all the plots and your observations in your submission.