CS910 Exercise Sheet 3: Regression

From the previous set of exercises, we know that the abalone data set has a number of attributes that are well-correlated with each other. We will use regression to study models that predict other attributes.

The raw data is available from: `http://archive.ics.uci.edu/ml/datasets/Abalone`. A version in the Weka format is available at: `http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs910/exercises/abalone.arff`

Linear regression, multilinear regression and non-linear regression

1. Fit a simple linear regression model to give diameter as a function of length. Give the parameters of the model, and the correlation coefficient. Comment briefly on what the parameters of the model tell you about abalone (see `http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names` for a description of the attributes).

   Solution: We obtain Diameter = 0.8155 * Length - 0.0194, with correlation coefficient 0.9868 ($R^2 = 0.9738$).

   The model shows that the dependency is very close to linear. It indicates that abalone are all approximately the same shape: once one dimension is established, it provides the (approximate) value of the other two.

2. The dataset includes information about the total weight of each specimen, along with the weight of different pieces (e.g. shell). Fit a multilinear model to give the whole weight as a function of the shucked weight, viscera weight, and shell weight, and give its parameters.

   Common sense would suggest that the whole weight should be related to the sum of these weights. Looking at the model and the data, comment on whether this relation holds for the observations.

   Solution: The model is weight = 0.9366 * Shucked weight + 1.1116 * Viscera weight + 1.253 * Shell weight - 0.0078 (correlation coefficient 0.9954)

   This does not quite tally with expectations – some weights are multiplied by more than one, others are multiplied by less than one. A (small) negative amount of weight is added as a constant term. This suggests that the data doesn't quite fit with expectations. An explanation is that some liquid is included in the whole weight, and this varies from specimen to specimen, and is not counted.

There are also some potential problems with the data: there are some examples (about 153, roughly 3.7%) where the sum of the weights is more than the total weight!

3. There is a relationship between the whole weight (fifth attribute) and diameter (third attribute). Try plotting these two attributes to see the shape. Based on this, fit the following models to the data, and for each report the correlation coefficient.
   (a) A simple linear model, weight = $a\cdot$ diameter + b
   (b) A quadratic model, weight = $a\cdot$ diameter + $b\cdot$ diameter^2 + c
   (c) A cubic model without lower order or constant terms, weight = $a\cdot$ diameter^3
   (d) An exponential model, $\log$ (weight) = $a\cdot$ diameter + b

   Based on these results, and the meaning of the model for the data, which would you pick to model this dependency and why? You may find it useful to plot the models over the data.

Solution: (a) Linear model: $R^2 = 0.8564$, PMCC = 0.92541. In R:

```
modela <- lm(abalone$V5 ~ abalone$V3)
```

(b) Quadratic model: $R^2 = 0.9268$, PMCC = 0.9627. In R:

```
modelb <- lm(abalone$V5 ~ abalone$V3 + I(abalone$V3^2))
```

(c) Cubic model: $R^2 = 0.9809$, PMCC = 0.9904. In R:

```
modelc <- lm(abalone$V5 ~ I(abalone$V3^3) -1)
```

The -1 tells to not fit an intercept term.

(d) Exponential model: $R^2 = 0.9284$, PMCC = 0.9635. In R:

```
modeld <- lm(log(abalone$V5)~abalone$V3)
```

Note, it does not matter which base is used: the coefficient is the same for all.

Although the exponential model has a high regression coefficient, the cubic model is best, for a number of reasons. First, we know that a specimen of zero diameter should have zero weight: this model ensures it, while the other models give a non-zero weight. Second, since we know that the size grows proportionally, we would expect the weight to be approximately proportional to volume, which is approximately cubic in the diameter. Plotting the exponential curve shows that it rapidly shoots up as diameter grows, which is not followed by the data.

Continued overleaf.

Logistic Regression

4. The male and female abalone are quite hard to tell apart, so for this question we will try to build a model to tell whether a specimen is an infant (I) or an adult (M/F).

   Build a *logistic regression* model to predict this feature based on the following attributes:

   (a) Length only

   (b) Whole Weight only

   (c) Class Rings only

   (d) Length, whole weight, and class rings together.

   For each model, give the accuracy (percentage of training examples predicted correctly).

   Hint. You may find it helpful to modify the input dataset to recode the new class value.

   Solution: (a) Length only: about 78% accurate

   (b) Whole weight only: about 79% accurate

   (c) Class rings: also about 78% accurate

   (d) All four: about 82% accurate

   Note that the accuracy from guessing is about 68%, and from including all attributes is about 84%.

5. For the last question, we return to the familiar adult data set. (`http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs910/exercises/adult.arff`, or `http://archive.ics.uci.edu/ml/datasets/Adult`). Build a logistic regression model for the attribute sex (M/F) using combinations of attributes from adult.data (try adding and removing attributes to see what happens). The aim is to find a model that balances simplicity with accuracy, so try to include as few variables as possible while giving an accurate result. Describe the final model you obtain, the steps you followed to reach it, and its accuracy for the task.

   (a) Which attributes can be removed from the data set without affecting the accuracy of the resulting model significantly (say, by at most 1%)? Give an argument why this might be the case for the attributes in question.

   (b) Why is relationship-status helpful?

   (c) Why is country=Holland weighted heavily?

Solution: Many models are possible. Using Weka and throwing in all attributes obtains an accuracy of 84% (compared to just guessing 'male', which has accuracy of 67%)

Education-num can be removed, since it is duplicated by education. The 'fnlwght' attribute is also unhelpful. 'capital gain' and 'capital loss' can be removed without any problem. 'income' (class) 'native country' 'race' 'age' can also be removed.

'marital status' is not too important, nor is 'education'. 'workclass' is not too important: it is somewhat covered by 'occupation'

This leaves a model in terms of occupation, relationship, and hours-worked-per-week. This model achieves an accuracy of 83.6%.

(a) Attributes like Native country, race, and age can be removed: the sampling for this census seemingly ensured that there was uniform coverage and so these do not indicate anything about gender. It seems that income level does not help very much, as do capital gains/losses: in this data, money does not inform about gender. 'workclass' is not too important: it is somewhat covered by 'occupation'

(b) Relationship status include Wife or Husband, which are highly predictive for gender!

(c) Country=Holland corresponds to a single data point, which reveals the gender of that individual. It is a fluke of the data, and doesn't really help with the model.

Bonus question (for those wanting to explore further, not for credit):

What is the best regression model you can build to predict the number of rings (which is related to the age of the specimen) in the Abalone data set? Throwing all variables into multilinear regression obtains a regression coefficient of about 0.73: can you beat this? You may want to try transformations of some variables (logarithmic, polynomial) based on visualizing the data.

Solution: One approach to this problem in R is documented here: `http://scg.sdsu.edu/linear-regression-in-r-abalone-dataset/`.