

## Statistics 13, Lab 6

### Regression

#### 1. Getting started

The data for this lab come from a study initiated by the Tasmanian Aquaculture and Fisheries Institute to investigate the growth patterns of abalone living along the Tasmanian coastline. The harvest of abalone is subject to quotas that restrict both the number of abalone that can be caught as well as their size. The population of abalone can be regulated effectively if there is a simple way to tell the age of abalone based solely on their appearance. Hence, researchers are interested in relating abalone age to variables like length, height and weight of the animal (measurements that a diver can take before they harvest the animal). If a reasonably accurate model can be found, then the Tasmanian officials can develop rules that prevent overharvesting of young abalone.

Determining the actual age of an abalone is a bit like estimating the age of a tree. Rings are formed in the shell of the abalone as it grows, usually at the rate of one ring per year. Getting access to the rings of an abalone involves cutting the shell. After polishing and staining, a lab technician examines a shell sample under a microscope and counts the rings. Because some rings are hard to make out using this method, these researchers believed adding 1.5 to the ring count is a reasonable approximation of the abalones age.

The relationship between age and ring count, however, is somewhat controversial. Under certain conditions, abalone can grow more than one ring per year. These conditions relate to weather patterns and other environmental variables. These facts suggest that any relationship we find between ring count and size measurements taken from the animal is likely to involve a lot of error; there are plenty of variables that have been left out of the study

Let's start by loading your data set.

```
source("http://www.stat.ucla.edu/~cocteau/stat13/data/ab.R")
ls()
```

Among the various objects in your workspace, you should see a new dataset called ab (for abalone)

#### 2. Examining the data

With the discussion in the previous section as background, let's look at the variables that were collected for this study; presumably they are among the most important factors influencing ring count.

```
dim(ab)
names(ab)
```

```
ab[1:5,]
```

The last command returns the first five observations from the dataset; that is, data on 5 of the 2,500 abalone. We can approximately determine the age of an abalone in this study by adding 1.5 to the number of rings in the variable rings. How old are these 5 specimens? The other measurements recorded for each abalone are given in the table below.

Variable name	Units	Description
rings	count	number of rings
length	mm	longest shell measurement
diameter	mm	measured perpendicular to the length
height	mm	height of abalone with meat in the shell
whole	grams	weight of the whole abalone
shucked	grams	weight of just the meat
viscera	grams	gut weight after bleeding
shell	grams	weight of the shell after drying
infant	0/1	1 if the abalone is an infant, 0 otherwise

The 2,500 abalone in this dataset all had between 1 and 29 rings. If we use the approximate dating rule, that means they are between 2.5 and 30.5 years old. Examine the data for the oldest and youngest abalone in the dataset with the following two commands.

```
subset(ab, ab$rings==1)
subset(ab, ab$rings==29)
```

In the first command, recall that the == relation is asking for all the data in ab for abalone having just one ring; in the second command we are asking for the data for those abalone with 29 rings.

*Question 1. (a) Do the sets of measurements for abalone with 1 and with 29 rings make sense? That is, what aspects of the data agree with the fact that one of these abalone is very young and the other is very old? (Not more than four sentences, please.) (b) Briefly comment on the variables in your dataset. All but one is quantitative; use summary() to describe them. For the one categorical variable, use table(). Note: This is meant to document the fact that you looked at the data briefly. Do not spend a lot of time on this question. Good data analysis starts by looking at the data a little.*

### 3. Correlation

We now consider more formal descriptions of the relationships between two variables. In lecture 18 and in your text, you will find a description of the correlation coefficient. Under certain conditions, this quantity measures the degree to which the data in a simple scatterplot lie on a straight line. See your text or your

lecture notes for details on how you would compute the correlation coefficient from a sample of data. In R, we can use the command `cor`.

```
cor(ab$length, ab$rings)
```

What value do you get? Use the material in lecture to guess what the scatterplot of length and rings might look like. Now, let's make this plot and see if your guess was correct.

```
plot(ab$length, ab$rings)
```

In lecture we discussed the use of a scatterplot matrix to see the relationship between several variables at once. You can create this kind of plot with the R command `pairs`.

```
pairs(ab[, 1:3])
```

This will create a scatterplot matrix with the first three columns or variables in the dataset `ab`. (Note that plots involving rings will have stripes because the ring count from abalone shells is a discrete variable. It has a large number of levels, but ring count is still a whole number.) Just like `pairs`, the command `cor` can also act on a several variables and return their correlations in the form of a matrix. Try the following command.

```
cor(ab[, 1:3])
```

*Question 2.*

- (a) *Do the relationships in the scatterplots agree with your intuition about the correlation coefficients? Explain using a pair with high correlation and one with relatively low correlation. The diagonal elements in `cor(ab[, 1:3])` are all 1. Does that make sense?*
- (b) *Consider now the first four variables in `ab` and again form a scatterplot matrix and a matrix of correlation coefficients. The new variable introduced is `height`. How is it related to the other variables? How is it correlated with the other variables? (Have a look at the last row or column of the `pairs`-plot and the matrix of correlation coefficients.) Finally, what is odd about the variable `height`?*
- (c) *Now, remove the two outliers in `height`; they are observations 1381, 1504. You can remove them with the command*

```
ab2 <- ab[-c(1381, 1504), ]
```

*which creates a new dataset ab2. Again, we are indexing rows in this command, but the minus sign tells R that we want to leave these rows out. This new dataset should only have 2,498 rows. Now, remake the scatterplot matrix and the matrix of correlation coefficients, calling the commands above with ab2 rather than ab. What has happened to the correlation coefficients? Comment on the change in the plots.*

Notice that many of the relationships we have seen so far tend to exhibit unequal spread. The plot of height by diameter, for example, shows greater spread for larger values of diameter. We are not going to consider this problem in depth (this is your first taste of regression, after all), but it is something we should worry about - and will, in a more advanced class.

### 3. Regression with a single predictor

In the last section, we saw plots of rings against the first three predictor variables. There seemed to be a lot of spread in these data. As noted in the introduction, the investigators at the Marine Resources Division acknowledge that several environmental factors can influence the growth of rings in abalone. In other words, there are variables relating to the condition of the water off the coast of Tasmania as well as the weather patterns for the last 30 years that might help explain ring counts. None of these data are available to us at this point, and hence we expect a certain amount of error in any model we build with our 8 predictor variables. To begin the modeling process, we will just look at a simple linear model with only one predictor. In mathematical terms we consider the following description of the data:

$$\text{rings} = \beta_0 + \beta_1 \text{length} + \text{error} \quad (1)$$

where (error) is a random error. Here the error accounts for all the variables we haven't included in the model. We can compute least squares estimates for  $\beta_0$  and  $\beta_1$  with the command

```
fit <- lm(rings~length,data=ab)
```

Here, the argument `data = ab` tells R to look for the data on rings and length in the dataset `ab`. The summary table we studied in class is obtained with next command.

```
summary(fit)
```

We can extract just the coefficients from the model with the command

```
coefficients(fit)
```

*Question 3. (a) What are the least squares estimates  $\beta_0$  and  $\beta_1$ ? (b) Use these values to create an estimate of the conditional mean ring count for abalone that have length 0.4. Do the same for abalone that have length 0.7. (c) Assuming the 2,500 specimens referred to in your dataset are a random sample of abalone off the coast of Tasmania, explain how you would use the bootstrap to assess the uncertainty in the slope estimate  $\hat{\beta}_1$ .*

*Bonus. Implement the bootstrap procedure to estimate the standard error for  $\hat{\beta}_1$ . In the code below, we draw abalone from our sample with replacement from our original data set. For each bootstrap sample we fit a regression and record the slope estimate. In all we will have 5,000 bootstrap replicates of the slope. Describe the distribution of these numbers and use their standard deviation as an estimate of the standard error. How does it compare to the value in the regression table we generated with the command `summary` above?*

```
replicates <- rep(0,5000)

for(i in 1:5000){

  sample_points <- sample(1:nrow(ab),replace=T)
  bootsample <- ab[sample_points,]

  fit <- lm(rings~length,data=bootsample)
  replicates[i] <- coefficients(fit)[2]

  print(i)
}

hist(replicates)
sd(replicates)
```