Homework Assignment 4: How the North American Mammalian Paleofauna Got a Crook in Its Regression Line

36-402, Advanced Data Analysis, Spring 2013

Due at 11:59 pm on Monday, 11 February 2013

> Turn in a single PDF file, with text and all figures, and a file name including your Andrew ID. Accompany this with a single plain-text file with all your R code, also named with your Andrew ID. Word files (`doc`, `docx`) will not be graded. As always, reporting more digits than is justified by statistical precision will cost you points.

Our problem set this week concerns an important question for evolutionary biology and paleontology. It has been argued[1] that larger organisms tend to have selective advantage over smaller ones of the same species, but larger bodies demand more specialized internal structure, more "division of labor", than small ones, indirectly driving the evolution of increased biological complexity. To evaluate this, it is important to know whether species tend to get larger over evolutionary time, and, if so, to characterize this accurately.

Our data set this week is taken from the North American Mammalian Paleofaunal Database, which contains information on the typical body mass of about 2000 living and extinct species of mammals native to North America. (You can find it on the website, `http://www.stat.cmu.edu/~cshalizi/uADA/13/hw/04/nampd.csv`.) Specifically, the columns of the data give: the scientific name of the species; the natural logarithm of its typical body mass (measured in grams); the natural logarithm of the mass of its ancestor (in grams); how long ago it first appeared in the fossil record (in millions of years); and how recently it last appeared (in millions of years; an NA in this column indicates the species is still alive). We will model how the *change* body mass is related to the body mass of the ancestral species. In particular, paleontologists have suggested that the correct model relating change in log mass to ancestral log mass should be piece-wise linear: a downward-sloping line for small ancestral log masses, and flat for larger ancestral masses. In this problem set, you will fit that model, and examine its predictions.

[1] For instance John Tyler Bonner, , *The Evolution of Complexity, by Means of Natural Selection* (Princeton University Press, 1988).

1. (10) *Basics*

   (a) (5) Load the data. Create a vector which gives each species' change in log body mass from its ancestor, and add it to the data frame as a new column. Explain, in your own words, what it would mean for a species to have a value of +0.7 in this column. Check that this column has NA values in the correct places. Explain how you know that those are the correct places. Remove all the rows with NA values for the change in log mass, and use this cleaned version of the data for all subsequent parts of the assignment.

   (b) (5) Plot the change in log body mass versus ancestral log body mass. Describe the plot briefly.

2. (10) *Linear model*

   (a) (2) Linearly regress the change in log body mass on the ancestral log body mass. Report the coefficients to reasonable precision.

   (b) (3) Create a new figure which is the scatter-plot from problem 1b, plus your fitted regression line.

   (c) (5) Based on the estimates 2a and the plot from 2b, does this model support or undermine the idea that new species tend to be larger than their ancestors? Explained.

3. (15) *Piecewise linear model*

   (a) (5) The piece-wise linear model predicts the following mean response as a function of the input x:

   $$\hat{y}(x) = \begin{cases} a + bx & \text{if } x \leq d \\ c & \text{if } x \geq d \end{cases}$$

   Assuming that this is continuous at d, solve for a in terms of b, c and d. Explain why, in this application, it is reasonable to assume continuity.

   (b) (10) Write a function in R, called[2] `deac`, that takes in a vector of numbers `x`, and three parameters `b`, `c`, and `d`, and returns the prediction of the model at each value of `x`.

   Check that your `deac` function is working properly by seeing that when $b = -1$, $c = 0.05$ and $d = 2$, giving `x=c(1,1.5,3)` outputs

   ```
   [1] 1.05 0.55 0.05
   ```

   Plot `deac`, with those parameters, as x goes over the range $(0, 4)$. Does it look right?

   Hints: `ifelse` for writing `deac`, `curve` for plotting.

[2] From the initials of the scientists who proposed this model; they didn't give it a name.

4. (15) Because `deac` varies nonlinearly with parameter d, we cannot estimate it by linear regression. However, we can still estimate the parameters by least squares. To do this, we need to write a function, make a starting guess about the parameters, and use the built-in optimization function `optim` (see recipe 13.2 in *The R Cookbook*).[3]. The following function fits the model to a data set by numerically minimizing the sum of squared errors:

```
my.start <- c(b=-1,c=0.2,d=10)
fit.a.deac <- function(data,start=my.start) {
  sse <- function(par) {
    preds <- deac(data$ln_old_mass,par[1],par[2],par[3])
    sum((data$delta_ln_mass - preds)^2)
  }
  fit <- optim(par=start,fn=sse,method="Nelder-Mead")
  coefficients <- fit$par
  fitted <- deac(data$ln_old_mass,coefficients[1],coefficients[2],
    coefficients[3])
  residuals <- data$delta_ln_mass - fitted
  mse <- mean(residuals^2)
  return(list(coefficients=coefficients,fitted=fitted,residuals=residuals,
    mse=mse,data=data))
}
```

(See online for the commented version; you'll want to source that, rather than typing this in and adding original errors.)

   (a) (7) Explain what the inner function, `sse`, does.

   (b) (8) What sort of output does `fit.a.deac` give — a vector, a list, an array, what? What do the various components of the output represent, in terms of the statistical problem?

5. (15) *Starting positions* The code given above looks for a vector of initial parameters called `my.start`, if no other starting point is supplied. The line before the function makes up some values for `my.start`; they are bad ones. We will see in a later problem set that a reasonable guess for d is about 5.

   (a) (5) Use this more-reasonable value of d to get a rough guess for c by taking the average change in log mass over all animals whose ancestral log mass exceeds d. Explain why this is a reasonable way to guess at c.

   (b) (5) Get a rough guess for b by linearly regressing the change in log mass on ancestral log mass for animals where the ancestral log mass is less than d. Explain why this is a reasonable way to guess at b.

[3] R has a built-in function, `nls`, for such "nonlinear least-squares" estimation, working more like `lm`. Unfortunately, `nls` can be flaky when the model doesn't have continuous derivatives, which is the case here. Besides, writing your own code builds character.

(c) (5) Re-define `my.start` to contain your improved guesses for b, c and d. Run `fit.a.deac` to get a fitted model, which you should call `nampd.deac`. Plot the fitted values as a function of log ancestral mass on a scatter-plot of change in log mass versus log ancestral mass.

6. (20) *Bootstrapping will continue until morale improves.* Use resampling of residuals, not cases, in both parts. *Note:* You can use the same resampled data-frames for both parts of this problem, but it needs more clever programming. 1000 bootstrap replicates takes 1–2 minutes on my computer.

   (a) (10) Find bootstrap standard errors, and 95% confidence intervals, for the parameters b, c and d. Report all these quantities.

   (b) (10) Find 95% bootstrap confidence bands for the fitted curve, and add them to your plot from problem 5c.

7. (15) *Linear vs. Piecewise Linear* One way to compare two models is to see which one can predict the other's parameter values. We will compare the simple linear model from problem 2a with the piecewise linear model `deac` model from problem 5c.

   (a) (5) Simulate the fitted `deac` model, using resampling of residuals, and estimate the linear model on the simulation. What coefficients do you estimate? Are they compatible with the ones you estimated from the data?

   (b) (5) Simulate the fitted linear model, using resampling of residuals, and estimate the `deac` model on the simulation. What coefficients do you get? Are they compatible with the ones you estimated from the data?

   (c) (5) Use five-fold cross-validation to compare the linear model from problem to the piecewise-linear `deac` model. Which one predicts mass changes better?