# Homework Assignment 3: How the *Hyracotherium* Got Its Mass

36-402, Advanced Data Analysis, Spring 2012

Due at the beginning of class, Tuesday 6 February 2012

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/12/hw/03/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.

As always, turn in a single PDF file, with text, figures, and with your R code in an appendix at the end. Word files (`doc`, `docx`) will not be graded.

1. (5 points) 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.

2. (5 points) Plot the change in log body mass versus ancestral log body mass. Describe the plot briefly.

3. (10 points) Linearly regress the change in log body mass on the ancestral body mass. Report the coefficients. Does this model support the idea that species tend to be larger than their ancestors?

[1] See for instance John Tyler Bonner, , *The Evolution of Complexity, by Means of Natural Selection* (Princeton University Press, 1988), a truly excellent book.

4. (10 points) Create a new figure which is the scatter-plot from problem 2, plus your fitted regression line, plus the conventional 95% intervals for the fitted values[2]. Explain why this qualifies or reinforces your qualitative conclusion from problem 3.

5. (10 points) Examine the residuals from the regression, and explain whether or not they make it reasonable to use the standard formulas for uncertainty.

6. (20 points) Use the bootstrap to re-sample data points or cases from the data frame, and re-fit the linear regression. Using 1000 bootstrap replicates, find 95% confidence intervals for the regression line, and add them to the figure from problem 4. Do they match the old prediction intervals? Explain why they should or should not.

7. (15 total)

   (a) (3 points) Explain what this does:

   ```
   npreg(y~x,data=df,regtype="ll",subset=s,tol=0.1,ftol=0.1)
   ```

   (b) (8 points) Fit a locally-linear kernel regression model for the expected change in log body mass as a function of ancestral log body mass. (If you cannot fit a locally-linear model, you can get partial credit for an ordinary kernel regression.) Add the fitted values to the plot.

   (c) (4 points) Is this non-parametric regression a good match to the linear regression? Is it inside the bootstrap confidence band for the linear regression? Explain what the bootstrap curve suggests both about the reliability of the linear model, and about the idea that species tend to be larger than their ancestors.

8. (20 points) Using the case-resampling bootstrap, with 1000 replicates, find a 95% confidence band for the kernel regression curve and add it to the plot.

   *Hint*: Doing this with a large number of replicates might take an hour or even more. Debug your code with a small number of replicates, then run the full version.

9. (5 points) Do these data support the idea that species tend, on average, to be larger than their evolutionary ancestors? Explain, referring to the regressions and confidence bands you calculated in earlier problems.

[2] I.e., intervals for "prediction of the mean response", not "prediction of an individual response".